nature

https://doi.org/10.1038/s41586-019-1725-y

Supplementary information

Quantifying the dynamics of failure across science, startups and security

In the format provided by the authors and unedited

Yian Yin, Yang Wang, James A. Evans & Dashun Wang

Supplementary Information for Quantifying dynamics of failure across science, startups, and security

Yian Yin,[1,2,3] Yang Wang,[1,2,4] James A. Evans,[5,6] Dashun Wang[1,2,3,4]

[1]*Center for Science of Science and Innovation, Northwestern University, Evanston, IL 60208, USA*

[2]*Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208, USA*

[3]*McCormick School of Engineering, Northwestern University, Evanston, IL 60208, USA*

[4]*Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA*

[5]*Department of Sociology, University of Chicago, Chicago, IL 60637, USA*

[6]*Santa Fe Institute, Santa Fe, NM 87501, USA*

Contents

S1 Data description

In this project, we compiled a comprehensive database consisting of three large-scale datasets across three different domains: Dataset D_1 contains submission histories of individual scientists in the US National Institutes of Health (NIH) grant system. D_2 contains profiles of innovators together with their startup ventures recorded in the VentureXpert investment database. D_3 records terrorist organizations and attacks retrieved from the Global Terrorism Database.

S1.1 NIH grant application dataset

Our first dataset contains all R01 grant applications (776,721 in total) that have been ever submitted by 139,091 scientists to NIH from 1985 to 2015. For each grant application, we obtained its evaluation score (if reviewed on a panel), a unique identifier for the PI, the PI's name, and the application outcome (funded/not funded).

The NIH grant application dataset represents an excellent setting to study dynamics of failure for several reasons. First, it contains ground-truth information for both successes and failures. Second, as the world's largest public funder for biomedical research, NIH is the dominant funding source for biomedical scientists in the US[29,51]. Indeed we tracked funding acknowledgment information cited within biomedical research papers, finding among all PubMed papers published in the US (2008 to 2015), NIH represents the majority of funding sources (81% out of top 10 agencies).

R01 is the most common research funding mechanism within the NIH[27–29], accounting for the

majority of the total funding. To compare the dynamical pattern between R01 and other granting mechanisms, we downloaded successful NIH grants from other mechanisms from NIH Research Portfolio Online Reporting Tools (RePORT), finding R01 grants are uniformly distributed within all NIH grants one obtains throughout a career.

Here we extract all new grant applications (excluding renewals, revisions and resubmissions) to reconstruct sequences of attempts. We truncate each sequence if (i) the individual gets one grant (successful group); or (ii) the individual has been inactive for a long period (unsuccessful group). We show results using all failure samples in the main text. We also repeated our results using just the first sequence of failures—failure streak without prior success, finding our conclusions remain the same (Extended Data Fig. 6). We also find that the observed patterns are not affected by potential periodicity of grant applications, and the results are robust against such variations. Indeed, we find the results remain the same if we add to timestamp of each attempt an artificial random noise at the scale of review cycles ($\sim$ 120 days).

S1.2 VentureXpert investment dataset

Our second dataset traces start-up investment records from the VentureXpert (SDC Platinum) database, including 58,111 startup companies and 163,106 investment rounds from 1970 to 2016. For each investment we obtained information on investment amount, funding date, company name and a full list of innovators involved. We then link these records with company information on Initial Public Offering and Merger & Acquisitions as outcome variables. Following the entrepreneur-

ship literature[31,32,52], we match individual entrepreneurs and startup ventures by linking each company with people listed as executives or board members at the first funding round. One advantage of this dataset is that 98.7% records have complete information of first and last names rather than initials, allowing us to construct career trajectories of 253,579 innovators.

Among the existing datasets capturing startups, the VentureXpert database, the official database of the National Venture Capital Association is among the most comprehensive and authoritative databases[30]. To further explore the coverage of the database, we compare the number of IPOs within our data versus US total counts, finding our dataset captures a significant fractions of IPOs, with the ratio between the two statistics remaining stable over time, documenting the reliability of this dataset. We also cross-validated individual entrepreneurs coverage with Crunchbase. We select top 1000 serial executives and board members ranked by the number of different jobs in Crunchbase, finding more than 70% of the profiles are included in VentureXpert.

Another challenge in modeling dynamics of failure in startup datasets is the ambiguity of 'failures'[53], which could include bankruptcy, termination to prevent future losses, and deviation from desired results. Recognizing the complexity of this issue, here we closely follow existing literature on venture capital and serial entrepreneurship[31,32]. We focus on all portfolio companies that have received at least one round of funding, and define those who went public or got acquired or merged at high values (percentile as compared with all M&As in the same year) as successes. We performed different measurement variations by changing the percentile threshold (1% and 5%) and also by only including IPOs (Extended Data Fig. 7). We find our results remain the same. If

a company obtained its first investment but did not succeed within a certain period, this venture is marked as a failure. In this dataset we treat each new venture as an attempt, starting at the date of first round investment. Similar to D_1, sequences of attempts by each individual are collected into a sequence, where the stopping criterion is defined by either (i) the individual is involved in one company that eventually achieved IPO or high-value M&As (successful group); or (ii) the individual has been inactive for a long period without success (unsuccessful group).

S1.3 GTD terrorism attack dataset

Our third dataset contains 170,350 terrorist attacks by 3,178 organizations from 1970 to 2017, collected by the Global Terrorism Database, one of the most systematic databases on domestic and transnational terrorist events[33]. For each attack we obtain information on its date, type, location, and consequences in terms of the number of people killed and wounded. Some records in this corpus are based on speculation or dubious claims of responsibility, which are discarded in our analysis to ensure the data quality.

There lacks a clear definition of 'success' for terrorist attacks, partly due to their diverse intents and consequences. To be consistent with our empirical steps in D_1 and D_2, here we treat an attack as successful if it killed at least one victim. To this end, we collected sequences of attacks of each terrorist organization, and classify the samples as (i) the organization killed at least one people (successful group); or (ii) the organization has been inactive for a long period without success (unsuccessful group).

One potential concern with this definition is that goals of terrorist attacks differ, and not all attacks are aimed at killing victims. This concern is somewhat alleviated since (1) 84.7% the attacks were targeted at human beings (i.e. assassination, bombing/explosion and assault) and (2) human-targeted attacks were uniformly distributed within full attack history of terrorist organizations. To rule out the possibility that samples in unsuccessful group are simply those who do not aim for killing victims, we further remove samples from the unsuccessful group if more than half of the events in this sample are not human-targeted. We also performed robustness checks by performing the same operation on successful group or using the full sample in unsuccessful group (Extended Data Fig. 8), finding our results remain robust. Although these checks do not necessarily account for the diverse goals of terrorist attacks, they do consistently show no evidence of systematic bias.

S1.4 Data limitations

Although the three datasets are among the largest in their respective domains, there are limitations of our data that readers should keep in mind.

First, despite the scale of our data, it remains difficult to obtain the full coverage of all attempts. For example, one might apply for grants from other funding agencies, found startup ventures without VC investments, or stop launching terrorist attacks for other activities. Further, agents who failed may also change their goals and subsequently transfer to other systems.

Second, data may contain missing values, resulting in false negatives. For example, terrorist

groups may not claim the events, especially when the events are small. For startups, not all M&As have the dollar amount associated with them, and we can only look at those who have M&A values.

Third, while individuals recorded in D_1 are uniquely identified with IDs, individuals and organizations in D_2 and D_3 are recorded and identified by full name, which may be affected by name ambiguity or name-changing issues.

Fourth, while grant applications in D_1 are binary events–either funded or unfunded, startup funding in D_2 and terror attacks in D_3 have varied definitions for success and failure. Take startup ventures as an example, and consider the growing trend of large startups to remain private despite having large valuations and funding. Unicorn ventures like Airbnb or WeWork could already be considered successful despite not having had an IPO or being acquired (see S7.3). For terrorist organizations, their intents differ by group ideology, and vary over time, hence their goals may not always be related to lethality.

While our systematic validation efforts in S1.1-S1.3 and robustness checks in S7.1-7.5 have not uncovered any potential biases, readers should keep in mind of the existence of such factors.

S2 Related work and models

S2.1 Learning literature

This paper is closely related to the rich literature on learning and failures. Canonical frameworks in understanding how people react to failures[20,39,53–57] have identified several key factors that could impact learning, including individual characteristics and organizational structures and strategies. These findings have also prompted quantitative studies using failure records across different industries, ranging from entrepreneurship[31,32] to commercial banking[58], from healthcare[59] to coal mining[60] to trains[61], and airlines[62] to orbital launch vehicles[63].

Another relevant line of inquiry is in psychology and organization behavior, which concerns learning curves from both theoretical[19–24,34,35,38–40,43,44,64–66] and empirical[22,38–41,45,67] perspectives, quantifying how performance and efficiency improve with experience. One key result is the famous Wright's law[45], i.e. the power law form of cost reduction.

Next we review a series of major models and compare key predictions with our empirical results. We summarize all these models in Table S2.

S2.2 Stochastic models with memory

One school of thought can be viewed as modeling the dependence structure among failures. Indeed, the failure of the chance model suggests that non-trivial dependence may be essential for modeling

the fat-tailed length distribution of failure streaks, which raises an important question: Could other stochastic processes (Markov process, random walk, autoregressive model, etc.) account for our observations? Indeed, if we consider a general framework of fixed dependence as follows

$$S_n = f_n(S_1, S_2, \cdots, S_{n-1}), \tag{S1}$$

where S_n denotes the performance at the n-th attempt and f_n can be a deterministic or stochastic non-decreasing mapping. This framework covers a wide range of stochastic processes, e.g. $f_n(S_1, \cdots, S_{n-1}) = f_n(S_{n-1})$ for a discrete space of S_n leads to Markov process, $f_n(S_1, \cdots, S_{n-1}) = S_{n-1} + \epsilon_n$ leads to random walk, $f_n(S_1, \cdots, S_{n-1}) = \sum_{i=1}^{p} \phi_i S_{n-i} + \epsilon_n$ leads to autoregressive model. We note that if this is true, we can obtain

$$S_n = f_n(S_1, f_1(S_1), \cdots, f_{n-1}(S_1, f_1(S_1), \cdots)) \equiv g_n(f_1, \cdots, f_n)(S_1) \tag{S2}$$

Hence, S_n can be formulated as a non-decreasing function of S_1, indicating that there should be detectable 'fitness' differences in the first attempt. Indeed, these results indicate that if there exists no difference in the dependency structure f_n, the differences in outcomes should be at least partly contributed by performance at the first attempt, which contradicts with our data. This hypothesis also cannot explain the fat-tail length distribution of failure streaks (S3.8).

S2.3 Adaptation models

The evolutionary perspective for individual and organizational learning assumes that the agent improves through updating information and belief on different alternatives. Here we discuss three representative models, each assuming a finite pool of available options.

S2.3.1 Crossman's model

Crossman's model, first proposed in[68], aims to explain the temporal dynamics observed in individual tasks. The model suggests a process from r methods M_i ($1 \leq i \leq r$), each with a time cost t_i. The individual improves operation strategy through changing probabilities for using different methods, i.e. p_i where $\sum_{i=1}^{r} p_i = 1$. At the n-th trial, the expected time cost can be formulated as

$$T(n) = \sum_{i=1}^{r} t_i p_i(n) \tag{S3}$$

The change of probability for choosing method M_i is proportional to the difference between its time cost and current average time cost, i.e.

$$p_i(n+1) - p_i(n) = -k(t_i - T(n)) \tag{S4}$$

Therefore, the time cost decays as

$$T(n+1) = T(n) - k \sum_{i=1}^{r} p_i (t_i - T(n))^2 \tag{S5}$$

S2.3.2 NK model

NK model, initially proposed by Kauffman[69] is a canonical model in organizational learning[70]. Consider a rugged fitness space of N dimensions $X = (x_1, \cdots, x_N)$, where $x_i \in \{0, 1\}$. The fitness score of each possibility is the summation of interaction among K adjacent dimensions, that writes

$$\phi(x) = \sum_{i=1}^{N} \phi_i(x_i, \cdots, x_{i+K}) \tag{S6}$$

One heuristic searching strategy in this rugged landscape concerns two options:

(1) Local search, i.e., walk to a neighbor, y, which satisfies $|y - x| = 1$.

(2) Global search, i.e., jump to a new node randomly.

S2.3.3 Denrell and March's model

Denrell and March proposed a simple adaptation model to understand the interplay between information and adaptation, explaining why people have bias against novel and risky choices[71]. In this model, P_t, defined as the probability for the first option to be chosen at time t, depends on its past probability P_{t-1} and current performance. If the option leads to better outcome compared with the other, one updates

$$P_{t+1} = P_t + a(1 - P_t) \tag{S7}$$

otherwise,

$$P_{t+1} = (1 - a)P_t \tag{S8}$$

All three models presented here can mimic specific performance or efficiency trajectory as one tries repeatedly. The main issue with these models is that they all base on a finite space of possible options, which leads to a limit in performance and efficiency improvement that one cannot overcome, which contradicts with our data.

S2.4 Search models

Search models assume an iterative process, where one decides whether to use existing components or try new ones based on component quality. Such models are often characterized by an improvement in the objective performance function because of the extreme values theory, i.e. as one always selects the best version from experimentation, she will eventually arrive at the version that is reasonably good.

S2.4.1 Roberts' model

Robert proposed a model based on greedy algorithms[65]. To understand the universal learning process, the model assumes production efficiency p as lognormal, following

$$x = b \ln p \tag{S9}$$

where x follows the standard normal distribution $N(0, 1)$. Each time the agent randomly selects a sample x' and compares it with current efficiency x, adopting the new method when $x' < x - a$. The model predicts

$$\ln p \sim \ln N/ab \tag{S10}$$

S2.4.2 Muth's model

Muth's model[43] builds on a simple assumption: the individual tries a new method at each trial and uses the new method if it costs less. The model further assumes appropriate regularity conditions

for the cumulative distribution function (CDF) of cost F, e.g.

$$\lim_{x \to x_0} \frac{F(x)}{(x - x_0)^k} = c \tag{S11}$$

where x_0 is the limiting cost of production. The model predicts the expected cost $E[X_n]$ of the n-th production as

$$E[X_n] = x_0 + \Gamma(1 + 1/k)(cn)^{-1/k} \tag{S12}$$

Muth's model is an elegant model explaining the emergence of power law scaling and can be extended to dependent component cases.

S2.4.3 McNerney's model

McNerney et al further extended Muth's model by assuming a power law distribution of costs of each component ($f(c_i) \sim x_i^{\gamma - 1}$) and using design structure matrix to characterize the dependency among different components[44]. The model predicts the cost y decreases as a function of productions n following

$$y(n) \sim n^{-1/\gamma d^*} \tag{S13}$$

where d^* is the design complexity and equals to 1 when all components are independent.

Search models successfully explain the emergence of power-law scaling in repeated attempts and serve as the basis of our frameworks (e.g. $k \to \infty$ limit). Yet they cannot account for the co-existence of two groups and their diverging patterns.

S2.5 Individual learning models

There has also been an active line of inquiry in explaining practice curves in individual tasks[41,42,46,72]. These models use psychology models as well as cognitive theories to explain 'practice makes perfect'.

S2.5.1 Newell and Rosenbloom's chunking model

To explain the power-law scaling observed in human task performance, e.g. inverted text reading and ten-finger game, Newell and Rosenbloom modeled the learning process using chunking theory[41]. In this model, there is a tree structure for goal hierarchies of height H and the speed-up of task completing is due to the emergence of higher-order chunks. The current highest order of chunk is denoted as η, leading to

$$\frac{dT}{dN} = \frac{dT}{d\eta}\frac{d\eta}{dN} \tag{S14}$$

The model further assumes each non-terminal goal has β non-terminal subgoals and ω terminal subgoals. As one constructs chunks of higher levels, the corresponding time to perform a new attempt decreases exponentially following

$$\frac{dT}{d\eta} \sim \beta^{H-\eta} \tag{S15}$$

If we also assume the chunking rate is linear with respect to time and the birth of a single level-h chunk requires time $s(h)$, we have

$$\frac{d\eta}{dN} \sim \frac{\beta^{\eta-H}}{s(\eta)}T \tag{S16}$$

Therefore, if $s(\eta)$, the number of possible states for goals at level η (complexity at this level), takes an exponential form as $s(\eta) \sim e^{\alpha\eta}$, which is consistent with the tree structure, we have

$$\frac{dT}{dN} \sim \frac{(T+E)^{-x}}{T} \tag{S17}$$

which follows a power law scaling. Hence, by combining two exponential forms in a tree structure, the model successfully derives the power law scaling.

S2.5.2 Anderson's model

Based on ACT's strengthening process, Anderson developed a model explaining cost decay[42]. The model assumes the amount of practice as S and the production execution in ACT takes the form

$$T = c + aS^{-1} \tag{S18}$$

The amount of past practice also decays as a power law of practice time:

$$S = \sum_{i=0}^{P-1} s(i,P) \sim \sum_{i=1}^{P} i^{-d} \sim P^{1-d} \tag{S19}$$

Therefore, we have

$$T = C' + A'P^{d-1} \tag{S20}$$

The two models are very relevant to our settings and can predict power law temporal scaling in the successful group. They represent two fundamental classes of cognitive architectures in related studies: ACT and Soar (and their variants)[50], highlighting the role of memory and chunks in learning process. Yet such mechanisms are more appropriate for modeling simple tasks rather than complex innovative ones and cannot account for the co-emergence of success and unsuccessful groups.

S2.6 Urn models

Urn model and its variants are among the canonical models in social physics as well as innovation process[73]. This model family is closely related to the famous Heaps' law[74], originally predicting that the number of distinct words S in a paragraph of length n scales as

$$S(n) \sim n^{\beta}, \;\; 0 < \beta \leq 1 \tag{S21}$$

Note that if we assume generating a new word costs unit time, we know the expected time spent on the n-th 'word' follows

$$t_n \sim n^{\beta-1} \equiv n^{-\gamma}, \;\; 0 < \gamma \leq 1 \tag{S22}$$

which recovers our empirical findings. Here we review several generative models explaining this scaling.

S2.6.1 Simon's model

Simon's model is among the earliest frameworks modeling 'cumulative advantages'[75]. It assumes that (1) There is always constant probability p for an agent to take a new word for the next element; (2) Otherwise (with probability $1-p$) the agent reuses past words based on frequency, i.e. randomly select a word from the past sequence. This model predicts a linear scaling between S and n i.e. $\beta = 1$, which can only explain the emergence of the unsuccessful group.

S2.6.2 Tria's model

By extending studies on urn model, Tria et al[76] assume an urn U of ideas and a sequence of S to generate. Every time an element is sampled from U to S, ρ copies are put back to U. Further, if this sampled idea is new in S, it triggers ν adjacent new ideas, hence the number of different ideas in a sequence follows the master equation

$$\frac{dD}{dt} \approx \frac{\nu D}{\rho t + (\nu + 1)D} \tag{S23}$$

The solution reveals that D grows linearly with t for $\nu > \rho$, but follows Heaps' law $D \sim t^{\nu/\rho}$ for $\nu < \rho$. These predictions are similar to the first phase transition point k^* in our model.

S2.6.3 Iacopini's model

To further document the impact of past transition sequence in innovative attempts, a recent paper[77] proposed a network-based model, where ideas are represented as nodes, and one can travel from one idea to another when they are linked by a weight. The process is set to be a weighted random walk on networks, following

$$P^t(i \to j) = \frac{w_{ij}^t}{\sum_k w_{ik}^t} \tag{S24}$$

When a specific path $i \to j$ is traveled, the weight of this edge is updated

$$w_{ij}^{t+1} = w_{ij}^t + \delta w \tag{S25}$$

Depending on different network structures, the model can lead to scaling $S \sim n^\beta$ with varying β.

While this class of models does not capture the performance dynamics underlying failures,

they are highly relevant to our study in that their predictions are consistent with the temporal patterns observed in our data.

S2.7 Other models

S2.7.1 Levy's model

Levy modeled the improvement of productivity based on the limited range of output denoted as P[40]. Given the current rate of production after producing q items, $Q(q)$, the improvement of production rate is proportional to the amount that the process can improve, i.e.

$$\frac{dQ(q)}{dq} = \mu[P - Q(q)] \tag{S26}$$

leading to

$$Q(q) = P[1 - e^{a+\mu q}] \tag{S27}$$

Levy's model captures a kind of production process where the final plateau part is significant, but it fails to predict the power-law form of productivity improvement.

S2.7.2 Shrager's model

By collecting and analyzing data of path length in the bit game, Shrager et al developed a graph-dynamic model for route-finding in ER networks $G(n, p)$[23]. The authors proposed a strategy where the individual randomly selects an edge after deleting the ones moving away from the destination with probability r. The number of trials increases the network density p linearly and the cost is the

path length of the whole process s. For r near 0, the model predicts

$$s \sim \frac{2}{p}(1-r)^{\ln n/\ln(np)} \tag{S28}$$

while for r near 1, the model predicts

$$s \sim \ln n/\ln(np) \tag{S29}$$

S2.7.3 Sahal's model

Sahal explains the progress function in industry productions through probabilistic and deterministic models[64]. The model assumes different manpower levels and $X(s,t)$ to be the number of product quantities requiring s amount of manpower at time t. If we assume the improvement across u manpower levels does not depend on the current level and can be formulated as $p(u)$, yielding

$$X(s,t+1) = \sum_{u=-n}^{1} X(s-u,t)p(u) \tag{S30}$$

If we define $X(s) = \lim_{t\to\infty} X(s,t)$, the solution of this equation can be formulated as

$$X(s) = b^s,\ 0 < b < 1 \tag{S31}$$

The model further assumes levels manpower are distributed on a logarithmic scale with width h, obtaining

$$F(Y) \sim Y^{-\log b/h} \tag{S32}$$

where $F(Y)$ is the number of product quantities requiring manpower greater than Y.

S2.7.4 Johnson's model

Johnson et al reported a similar scaling from the time interval of terrorist attacks and other human confrontations[35]. An illustrative model for this scenario considers confrontation between 'Red Queen' and 'Blue King', and the advantage of Red Queen after n events, $R(n)$, can be formulated as

$$R(n) = \sum_{i=1}^{n} x_i \tag{S33}$$

where x_i takes value $+d$ or $-d$ with probability 1/2. Depending on the auto-correlation of x_i, one can get

$$R(n) \sim n^b d, 0 \leq b \leq 1 \tag{S34}$$

Taking the inverse of the advantage, we get the attack rate scales as a negative power law of n, i.e.

$$\tau_n \sim n^{-b}, 0 \leq b \leq 1 \tag{S35}$$

S2.7.5 Clauset's model

Clauset's model[34] also predicts the temporal pattern of terrorist attacks, but in a very different way from Johnson's model[35]. Indeed, if we assume that the size of terrorism organizations scales linearly with its past attacks, i.e.

$$s(n+1) = s(n) + \eta \tag{S36}$$

The model further assumes a new takes time as the inverse of organization size, i.e.

$$\Delta t \sim 1/s \tag{S37}$$

Taken together, we have

$$\Delta t \sim 1/n \tag{S38}$$

This model successfully links group size to temporal dynamics, predicting a power law scaling. Yet it only applies to group dynamics and the exponent of power law in the original linear assumption is restricted to be -1.

One commonality among these models is that they lack predictions of the interplay between performance and time. By contrast, our data show that the temporal scaling cannot be simply explained by agents optimizing time cost t_n since the performance also improves for the successful group. These models also cannot explain the co-existence of success and unsuccessful groups observed in our data.

S2.8 Summary of contributions

To sum up, despite the ubiquity of power laws across a wide variety of settings[14,15,17,47–49] and the foundational literature on learning curves[19,21,41–46], none of the existing models, to our knowledge, anticipated the existence of the early signals documented in the paper (Table S2). As such, the paper makes several contributions which we next summarize in terms of its empirical measurements, theoretical contributions, and predictive signals:

1. Empirical contributions: Our quantitative understanding of the dynamics of failure is important, but has remained limited, due to difficulties in collecting large-scale datasets that

capture failures. This highlights the first contribution of our paper – to be able to assemble large datasets from three disparate domains that contain records of both success and failure cases.

2. Theoretical contributions: These new datasets allow us to derive among the first empirical evidence about the dynamics of failure to test existing models. In particular, Figure 1 highlights the fundamental tension with existing modeling frameworks, and the simplicity of measurements in Fig. 1 highlights the paucity of quantitative approaches thus far to model failures, highlighting the key contribution of our paper – By establishing a new theoretical basis for understanding failures, our paper not only explains empirical patterns that existing models cannot capture (Fig. 1), but also predict new patterns that existing models did not anticipate (Fig. 3). As such, the model is unique in its ability to (i) predict two fundamentally different behaviors simultaneously at two extremes (e.g., $k = 0$ and $k = \infty$), hence serving as the first model to unify existing paradigms; and (ii) reveal a highly discontinuous pattern between progression and stagnation regimes. This further leads to four new predictions, all of which are tested and validated across our three datasets. This was only possible thanks to the new theoretical insights, and in particular the novel predictions that our model offers.

3. Predictive signals: Our findings unveil identifiable yet previously unknown early signals that allow us to distinguish failure dynamics that will lead to ultimate victory or defeat. Traditionally the primary distinction between ultimate victory and defeat has been attributed to differences in luck, learning strategies or individual characteristics, but here our model offers an important new explanation with crucial implications: Even in the absence of dis-

tinguishing initial characteristics, agents may experience fundamentally different outcomes. As such, our model shows that the success and unsuccessful groups may be initially similar, but each follows their respective, highly predictable patterns, distinguishable long before the eventual outcome becomes apparent. Specifically, we show that observing the timing of each attempt alone can help us identify those more likely to succeed. Considering the myriad factors related to success in a grant proposal/startup company/terrorist attack, this level of predictive power achieved by a singular, simple predictor is somewhat unexpected.

S3 Modeling failure dynamics

S3.1 The k model

In order to formulate a new attempt, the individual needs to go through every component, and decide what to do next. For a past attempt j, each component i is characterized by an evaluation score $x_j^{(i)}$, which falls between 0 and 1. The agent can either create a new version (with probability p), or with probability $1-p$ reuse an old one by choosing among past versions. The main cost of creating a new version is time. Here we assume each new version takes one unit of time, and upon creation takes up an evaluation score, drawn randomly from a fixed distribution $\rho(x)$. Real systems are likely to differ in their specific score distributions. Here for simplicity, we assume $\rho(x)$ follows a uniform distribution on $[0,1]$, approximating the percentile of any underlying score distributions real systems may follow. One difference between our model and canonical learning curve models[44] is that one has little information on the new versions until it gets implemented and

evaluated, hence new versions are not guaranteed to increase or decrease their score.

Of the many factors that may influence p, one key factor is the quality of existing versions. Denoting with x^* the best score among past versions, we expect p to be a function of x^*. Indeed, consider the two extreme cases. If $x^* \to 0$, existing versions of this component have among the worst scores hence a high potential to be improved upon with a new version. Therefore the likelihood of creating a new version is high, i.e., $p \to 1$. On the other hand, $x^* \to 1$ indicates an already excellent version, corresponding to a decreased incentive to create a new one ($p \to 0$). Reusing the existing best version allows the particular component to retain its score x^* and also avoids incurring additional time cost the individual can avoid spending time working on. To this end, considering $P(x \geq x^*) = 1 - x^*$ as the potential to improve on existing versions, we assume $p = (1 - x^*)^\alpha$, where $\alpha > 0$ characterizes an individual's propensity to create new versions given the quality of existing versions. The higher this potential, the more likely one may create a new version[71].

The dynamics of quality score, x_n, can be captured by a higher-order Markov process of memory length k, following

$$x_n^* = \max\{x_{n-k}, \cdots, x_{n-1}\} \tag{S39}$$

$$x_n \sim \begin{cases} U[0,1], & w.p.\ (1 - x_n^*)^\alpha \\ \delta(x - x_n^*), & w.p.\ 1 - (1 - x_n^*)^\alpha \end{cases} \tag{S40}$$

where we assume $x_n = 0$ for all $n < 0$.

The parameter k in our model can be viewed as approximating the 'memory' of past ver-

sions. The rationale of using k for the model is rooted in the learning literature, showing that the general notion of 'forgetting' takes multiple forms, often representing a combination of individual, organizational and environmental factors. Indeed, several relevant factors may be at play, which can generate patterns similar to 'forgetting'. For example, in rapidly shifting innovation domains, not all past failures remain useful over time, and some become obsolete. Consider the concept of 'knowledge depreciation'[78], which could also apply in our settings as environments (scientific knowledge/capital markets/security situations) evolve over time, such that past experience could become useless even if memorized. For example, an NIH proposal four failures ago may become irrelevant as the ideas proposed have been dispositively proven wrong, or published by the PI or another research group[79,80]. Similarly, startup ideas from the dot com era may be irrelevant in the era of AI and Blockchain[32]. Terrorist tactics can also depreciate over time, as past strategies attracted media coverage and gave rise to tighter security measures defending against them[35]. This line of reasoning supports the idea that recent attempts are most relevant. It is also consistent with the learning literature, which suggests knowledge 'forgetting' can happen in distinct ways, either voluntarily or involuntarily[81]. Motivated by these reasons, here we select a single parameter k to encapsulate a variety of potential contributing factors.

To solve the model, let's first look at two extreme cases.

S3.2 Independent model ($k = 0$)

Here we first consider a simple case when $k = 0$, i.e. there lacks any reusable materials in memory as one tries again. In this case, one creates a new version every time, hence for all n we have

$$x_n \sim U[0, 1] \tag{S41}$$

and

$$t_n \equiv 1 \tag{S42}$$

S3.3 Learning from all failures ($k \to \infty$)

We now turn to another extreme: learn from all past failures. We can rewrite the process as

$$x_n^* = \max\{x_0, \cdots, x_{n-1}\} \tag{S43}$$

$$x_n \sim \begin{cases} U[0, 1], \;\; w.p. \; (1 - x_n^*)^\alpha \\ \\ \delta(x_n^*), \;\; w.p. \; 1 - (1 - x_n^*)^\alpha \end{cases} \tag{S44}$$

Here we focus on the dynamics of x^*, obtaining

$$x_{n+1}^* \sim \begin{cases} U[x_n^*, 1], \;\; w.p. \; (1 - x_n^*)^{\alpha+1} \\ \\ \delta(x_n^*), \;\; w.p. \; 1 - (1 - x_n^*)^{\alpha+1} \end{cases} \tag{S45}$$

where $x_1^* \sim U[0, 1]$. To this end, let us denote f_n as the probability density function of x_n^*, obtainng

$$f_{n+1}(x) = f_n(x)(1 - (1 - x)^{\alpha+1}) + \int_0^x f_n(y)(1 - y)^\alpha dy \tag{S46}$$

with $f_1(x) \equiv 1$ for $x \in [0,1]$. By induction we obtain

$$f_n(x) \sim [1-(1-x)^{\alpha+1}]^{n-1} \tag{S47}$$

The normalization constant equals to

$$\int_0^1 [1-(1-x)^{\alpha+1}]^{n-1} dx = \int_0^1 x^{-\alpha}(1-x^{\alpha+1})^{n-1} dx^{\alpha+1}/(\alpha+1) = B(n, 1/(\alpha+1))/(\alpha+1)$$

Therefore we have

$$\begin{aligned} t_n &= \frac{\int_0^1 (1-x)^\alpha f_n(x) dx}{\int_0^1 f_n(x) dx} \\ &= \frac{B(n,1)}{B(n,1/(\alpha+1))} \\ &\sim \frac{\Gamma(1) n^{-1}}{\Gamma(1/(\alpha+1)) n^{-1/(\alpha+1)}} \\ &\sim \Gamma\left(\frac{1}{\alpha+1}\right)^{-1} n^{-\frac{\alpha}{\alpha+1}} \end{aligned} \tag{S48}$$

and

$$\begin{aligned} 1-x_n &= \frac{\int_0^1 \{(1-x)[1-(1-x)^\alpha] + (1-x)^\alpha/2\} f_n(x) dx}{\int_0^1 f_n(x) dx} \\ &= \frac{B(n, 2/(\alpha+1)) - B(n, 1+1/(\alpha+1)) + B(n,1)/2}{B(n, 1/(\alpha+1))} \\ &\sim \frac{\Gamma(2/(\alpha+1)) n^{-2/(\alpha+1)} - \Gamma(1+1/(\alpha+1)) n^{-1-1/(\alpha+1)} + \Gamma(1) n^{-1}/2}{\Gamma(1/(\alpha+1)) n^{-1/(\alpha+1)}} \\ &\sim \Gamma\left(\frac{1+\min\{\alpha,1\}}{\alpha+1}\right) \Gamma\left(\frac{1}{\alpha+1}\right)^{-1} n^{-\frac{\min\{\alpha,1\}}{\alpha+1}} \end{aligned} \tag{S49}$$

Therefore, both efficiency and quality scales with n, following $\gamma = 1 - 1/(\alpha+1)$ and $\eta = \min\{\gamma, 1-\gamma\}$.

S3.4 Solving the general model

We note that the previous two cases are tractable because either x_n or x_n^* can be formulated into a simple Markov process without higher-order dependencies. However, such techniques cannot be applied directly solving the general cases. As we next discuss, using renewal process theories[37] can help us obtain estimations of scaling exponents. More specifically, we first note that

$$|\{n_1 \le n \le n_2 : x_n = x_m^*\}| \le n_2 - n_1 + 1 \tag{S50}$$

$$|\{n_1 \le n \le n_2 : x_n = x_m^*\}| \ge \sum_{i=0}^{[(n_2-n_1)/k]-1} \sum_{j=0}^{k-1} I(x_{n_1+ki+j} = x^*_{n_1+ki+i}) \ge [(n_2-n_1)/k] \tag{S51}$$

Hence to calculate the length of a sequence, we only need to estimate the number of versions that are once baseline versions (i.e. n such that $x_n = x_m^*$ for some $n+1 \le m \le n+k$).

Denote $z_m = 1 - x_n^*$ as all such baseline scores. We now calculate for a specific z_m to be taken by a new one, the number of attempts it takes. Indeed, given a score z_m and assuming that it has been reused as $z_m = z_{m-1}$, we have

$$z_{m+1} = \begin{cases} z_m & w.p.\ \frac{[1-z_m^{k\alpha}(1-z_m)^k](1-z_m^\alpha)}{1-z_m^\alpha(1-z_m)} \sim O(1) \\ U[0, z_m] & w.p.\ \frac{[1-z_m^{k\alpha}(1-z_m)^k]z_m^{\alpha+1}}{1-z_m^\alpha(1-z_m)} \sim O(z_m^{\alpha+1}) \\ \min\{U_1[0,1], \cdots, U_k[0,1]\} & w.p.\ z_m^{k\alpha}(1-z_m)^k \sim O(z_m^{k\alpha}) \end{cases} \tag{S52}$$

Here we use the big-O notation to find the asymptotic case for $z_m \to 0$. This equation shows two important insights:

(1) If we calculate the number of iterations that z_m gets reused, it should be in the order of $O(z_m^{-\min\{k\alpha,\alpha+1\}})$, leading to two cases that will be discussed in detail.

(2) There exist two different ways for the substitution of baseline versions: *Quality* ($w.p.\ O(z^{k\alpha})$) and *Recency* ($w.p.\ O(z^{\alpha+1})$). For $k\alpha < \alpha + 1$, the recency mechanism dominates for small z, i.e. produces a worse succeeding score. Hence, it keeps a stable score distribution of new baseline scores as n increases. However, once $k\alpha > \alpha + 1$, quality mechanism takes over for small z, characterizing a continuous path of improvement.

Here, we first derive our results for the regime $k\alpha < \alpha + 1$, and then extend the obtained results to the other regime.

S3.4.1 Case 1: $k\alpha < \alpha + 1$

When $z_{m+1} \neq z_m$, our previous results show that with high probability, z_m is the extreme value among k i.i.d. random variables on $U[0,1]$, hence the pdf of z_m, $f(z_m) \sim const$ as $z_m \to 0$. Below we offer a more rigorous proof: Take all the different z_m as $\tilde{z}$ and consider a limiting distribution of $f(\tilde{z})$. We have

$$f(\tilde{z}) \sim \int_0^1 f(\tilde{z}')O(1)d\tilde{z}' + \int_{\tilde{z}}^1 f(\tilde{z}')O(\tilde{z}'^{\alpha+1-k\alpha})/\tilde{z}'d\tilde{z}' \tag{S53}$$

Assuming $f(\tilde{z}) \sim \tilde{z}^{\beta_1}$ and consider $\tilde{z} \to 0$ one gets

$$\beta_1 = \min\{0, 1, \beta_1 + \alpha + 1 - k\alpha\} = \min\{0, \beta_1 + \alpha + 1 - k\alpha\} \tag{S54}$$

Since $k\alpha < \alpha + 1$, we get $\beta_1 = 0$. Hence, as we generate a new baseline score satisfying $z_m \neq z_{m-1}$, we approximate the number of iterations it will be retained as $u \sim z^{-k\alpha}$. Let $z_m = z_{m+1} = \cdots = z_{m+u}$. For z_{m+u+1} we take a new random variable from a fixed distribution on $[0,1]$

whose probability density does not diverge near 0. If we consider the change of baseline scores as a 'jump' and number of iterations of repeated reuse as the length of this jump (u), we eventually arrive at a Levy flight[82].

We can define $u_i \equiv z_i^{-k\alpha}$, following asymptotically power law pdf $P(u) \sim u^{-1/k\alpha-1} \equiv u^{-\mu-1}$, and $m(N) \equiv \min_m\{u_1 + \cdots u_m \geq N\}$. Next we solve $\langle u^{\lambda}_{m(N)}\rangle$ for some λ. We first calculate $P(u_{m(N)})$, which equals to

$$\begin{aligned} P(u_{m(N)} = u) &= P(u) \int_{\max\{N-u,0\}}^{N} \sum_{k=0}^{\infty} P_k(v)dv \\ &= P(u) \int_{\max\{N-u,0\}}^{N} G(v)dv \end{aligned} \tag{S55}$$

where $P_k(v) \equiv P(v_1 + \cdots + v_k = v)$ and $G(v) \equiv \sum_{k=0}^{\infty} P_k(v)$. P_k can be obtained analytically by induction, following

$$P_k = \begin{cases} P_{k-1} \circ P, & k \leq 1 \\ \\ \delta(0), & k = 0 \end{cases} \tag{S56}$$

Hence we have

$$G = \sum_{k=0}^{\infty} P_k = G \circ P + \delta(0) \tag{S57}$$

Taking the Laplace transformation we obtain

$$\tilde{G} = \tilde{G}\tilde{P} + 1 \tag{S58}$$

leading to

$$\tilde{G} = \frac{1}{1 - \tilde{P}} \tag{S59}$$

The quantity of interest, $M(N) \equiv \langle u^{\lambda}_{m(N)} \rangle$, can be formulated as

$$\begin{aligned} M(N) &= \int_0^{\infty} P(u_{m(N)} = u) u^{\lambda} \\ &= \int_0^N P(u) u^{\lambda} \int_{N-u}^N G(v) dv du + \int_N^{\infty} P(u) u^{\lambda} \int_0^N G(v) dv du \\ &= \int_0^N Q(u)[H(N) - H(N-u)] du + \int_N^{\infty} Q(u) H(N) du \\ &= H(N) \int_0^{\infty} Q(u) du - \int_0^N Q(u) H(N-u) du \\ &= H(N) \int_0^{\infty} Q(u) du - (Q \circ H)(N) \end{aligned} \tag{S60}$$

where $H(N) = \int_0^N G(v) dv$ and $Q(u) = u^{\lambda} P(u)$. Performing again the Laplace transformation, we obtain

$$\begin{aligned} \tilde{M} &= \tilde{H}(\int_0^{\infty} Q(u) du - \tilde{Q}) \\ &= \tilde{G}(\int_0^{\infty} Q(u) du - \tilde{Q})/s \\ &= \frac{\int_0^{\infty} Q(u) du - \tilde{Q}}{s(1 - \tilde{P})} \end{aligned} \tag{S61}$$

Assuming

$$P(x) = \mu x^{-\mu-1} I(x \geq 1) \tag{S62}$$

we obtain

$$\tilde{P}(s) = \mu s^{\mu} \Gamma(-\mu, s) \tag{S63}$$

$$\tilde{Q}(s) = \mu s^{\mu-\lambda} \Gamma(\lambda - \mu, s) \tag{S64}$$

$$\int_0^{\infty} Q(u) du = \frac{\mu}{\mu - \lambda} \tag{S65}$$

where $\Gamma(a, s) = \int_s^{\infty} t^{a-1} e^{-t} dt$ is the upper incomplete Gamma function. Inserting these results

into the previous function we arrive at

$$\tilde{M} = \frac{\mu/(\mu-\lambda) - \mu s^{\mu-\lambda}\Gamma(\lambda-\mu, s)}{s[1-\mu s^{\mu}\Gamma(-\mu, s)]} \tag{S66}$$

To obtain asymptotic results for $M(N)$ as $N \to \infty$, we approximate $\tilde{M}(s)$ as $s \to 0^+$. Here we use the following expansion

$$\Gamma(a, s) = \Gamma(a) - \frac{s^a}{a} + \frac{s^{a+1}}{a+1} + O(s^{a+2}) \tag{S67}$$

The previous equation hence writes

$$\begin{aligned}
\tilde{M} &\approx \frac{\mu/(\mu-\lambda) - \mu s^{\mu-\lambda}\Gamma(\lambda-\mu) + \mu s^{\mu-\lambda}s^{\lambda-\mu}/(\lambda-\mu) - \mu s^{\mu-\lambda}s^{\lambda-\mu+1}/(\lambda-\mu+1)}{s[1-\mu s^{\mu}\Gamma(-\mu) + \mu s^{\mu}s^{-\mu}/(-\mu) - \mu s^{\mu}s^{-\mu+1}/(1-\mu)]} \\
&= \frac{-\mu s^{\mu-\lambda}\Gamma(\lambda-\mu) - \mu s/(\lambda-\mu+1)}{s[-\mu s^{\mu}\Gamma(-\mu) - \mu s^{\mu}s^{-\mu+1}/(1-\mu)]} \\
&= \frac{s^{\mu-\lambda}\Gamma(\lambda-\mu) + s/(\lambda-\mu+1)}{s[s^{\mu}\Gamma(-\mu) + s/(1-\mu)]} \sim s^{\min\{\mu-\lambda,1\}-\min\{\mu,1\}-1}
\end{aligned} \tag{S68}$$

Hence we obtain

$$M = L^{-1}(\tilde{M}) \sim n^{-\min\{\mu-\lambda,1\}+\min\{\mu,1\}} \tag{S69}$$

Let us consider the two specific cases:

Case 1: $\lambda = -1/k$, we have $M \sim n^{\min\{1/(k\alpha),1\}-1}$, hence

$$\langle (1-x^*)^{\alpha} \rangle \approx M = \begin{cases} \text{const.}, & k\alpha \leq 1 \\ n^{-1+1/(k\alpha)}, & k\alpha > 1 \end{cases} \tag{S70}$$

Case 2: $\lambda = -1/(k\alpha)$, we have $M \sim n^{\min\{1/(k\alpha),1\}-\min\{2/(k\alpha),1\}}$, hence

$$\langle 1-x^* \rangle \approx M = \begin{cases} \text{const.}, & k\alpha \leq 1 \\ n^{-1+1/(k\alpha)}, & 1 < k\alpha \leq 2 \\ n^{-1/(k\alpha)}, & k\alpha > 2 \end{cases} \tag{S71}$$

This eventually leads to

$$\langle 1-x\rangle = \langle z\rangle = \langle z^* + z^{*\alpha}/2 - z^{*(\alpha+1)}\rangle \approx \langle z^* + z^{*\alpha}/2\rangle \sim n^{-\min\{1,k\alpha-1\}/k\alpha} \sim n^{-\min\{\gamma,1-\gamma\}} \quad \text{(S72)}$$

S3.4.2 Case 2: $k\alpha > \alpha + 1$

As we discussed, in this regime the quality dynamics is dominated by the second mechanism, which does not depend on k, and asymptotically follows the same mechanism as learning from all failures model ($k = \infty$). Indeed, if we expand our solution and take $k \to (1 + 1/\alpha)^-$, we obtain $\gamma = 1 - 1/(k\alpha) \to \alpha/(\alpha + 1)$ and $\eta = \min\{\gamma, 1 - \gamma\} \to \min\{1, \alpha\}/(\alpha + 1)$, which are the same as $k = \infty$. Hence, the regime lying between $k = 1+1/\alpha$ and $k = \infty$ should have the same scaling behaviors.

Taken together, we obtain

$$\gamma = \begin{cases} 0, & k < k^* \\ 1 - k^*/k, & k^* \le k < k^* + 1 \\ 1/(k^* + 1), & k \ge k^* + 1 \end{cases} \quad \text{(S73)}$$

$$\eta = \min\{\gamma, 1 - \gamma\} \quad \text{(S74)}$$

where $k^* = 1/\alpha$.

S3.5 Connections with canonical ensembles

The piecewise function in our solutions raises an interesting question: What characterizes the discontinuous pattern at $k = k^*$ and $k = k^* + 1$? In this section, we establish a mapping between our model and a canonical ensemble system, showing that the observed critical points can be phenomenologically linked to phase transitions (Extended Data Fig. 1).

For simplicity, we rescale this system through

$$\begin{aligned} K &= k - k^* \\ \Gamma &= k^* \gamma / (1 - \gamma) \end{aligned} \tag{S75}$$

obtaining

$$\Gamma = \begin{cases} \Gamma_a(K) \equiv 0, & K < 0 \\ \Gamma_b(K) \equiv K, & 0 \leq K < 1 \\ \Gamma_c(K) \equiv 1, & K \geq 1 \end{cases} \tag{S76}$$

Note that all smoothness conditions are preserved since the transformations in S75 are infinitely differentiable. Here we consider a system with three different states a, b, c with corresponding energy density $E_a(h), E_b(h), E_c(h)$. Its partition function can be written as

$$Z = e^{-NE_a(h)} + e^{-NE_b(h)} + e^{-NE_c(h)} \tag{S77}$$

where N is the total number of particles and h is external field. We further assume that $E_a(h) = (2\epsilon h - 1)^2$, $E_b(h) = (2h - 1)^2$, and $E_c(h) = [2\epsilon(1 - h) - 1]^2$ where $\epsilon \to 0^+$. The introduction of ϵ is to distinguish state a from state c, and we approximate this with limiting condition $E_a(h) = E_c(h) = 1$.

Next, we map $f \to (2\Gamma - 1)^2$, $N \to \ln n$, $h \to K$, and $E_i(h) = [2\Gamma_i(K) - 1]^2$. Hence, the two transition points k^* and $k^* + 1$ corresponds to $h = 0$ and 1 in the canonical ensemble systems. To explore the nature of discontinuity at k^* and $k^* + 1$, we now turn back to the analytical solutions of the mapped system.

As $N \to \infty$, the free energy density $f = \ln Z/N$ converges to the minimal energy $f = \min(E_a(h), E_b(h), E_c(h))$. Hence, the magnetization density $m = \frac{df}{dh}$ is discontinuous at the boundary of two $E_i(h)$. In particular, the differences across the boundary is caused by changes in base states, i.e. the mechanisms that dominate the current system. Therefore, there exists phase transitions at h^* if $E_i(h^*) = E_j(h^*)$ for $i \neq j$. Indeed, we obtain phase transition at $h^* = 0$ and $h^* = 1$, respectively, which correspond to the two transition points at k^* and $k^* + 1$ in our model.

To unveil the origin of these transitions, here we inspect $u(z)$, defined as the number of attempts where a version of high score $x \to 1$ (i.e. potential $z \equiv 1 - x \to 0$) is retained. We can analytically derive its asymptotic distribution as

$$P_z(u) \sim \left\{ \frac{(1 - z^{1/k^*})[1 - z^{k/k^*}(1 - z)^k]}{1 - z^{1/k^*}(1 - z)} \right\}^{-Au} \sim [1 - z^{\min\{k/k^*, 1/k^*+1\}}]^{-Au} \tag{S78}$$

where A is a constant independent of z and u. Eq (5) enables us to calculate the expected life span of a high-quality version $\langle u(z) \rangle \sim \langle z^{-\min\{k/k^*, 1/k^*+1\}} \rangle$. The first critical point $k = k^*$ hence corresponds to the finiteness of this first moment $\langle u \rangle$. When k is small ($k < k^*$), $\langle u \rangle$ is finite. In this region, although new versions build on past k attempts, good versions will only be reused for a limited number of attempts. This is similar to an asymmetric (super-)*diffusive* random walk where the step size has finite expectation (renewal process), predicting a linear relationship

between number of attempts and time cost. Once k passes the critical threshold k^*, we find $\langle u \rangle = \infty$, hence a good version may be retained for an unlimited long period. This is similar to a *ballistic* random walk where the expectation of step size is infinite, and the scaling behavior between steps (time cost) and distance (number of attempts) begins to emerge. The second phase transition originates from the competition between two dynamical forces: (a) the k/k^* term represents the chance that the current best version gets forgotten due to k consecutive attempts in creating new versions; (b) the $1/k^* + 1$ term captures the chance that the current best version is substituted by a better one. Comparing the dominance of the two mechanisms points to the second transition point $k^* + 1$, beyond which k plays no major role.

S3.6 Functional forms of $\rho(x)$ and $p(x)$

Two important quantities in our model are $\rho(x)$, the score distribution for a new version, and $p(x)$, the probability to create a new version given reference score x. For simplicity we assume $\rho(x) \equiv 1$ and $p(x) = (1 - x)^\alpha$ in the main text. Here we show that similar results can be obtained as we consider a general class of functional forms of $\rho(x)$ and $p(x)$.

Indeed, since both quantities depend on the scoring system, we may fix one to a specific form. Consider two score systems x and y that can be derived through $y = c(x)$. We can derive the transformations as

$$\rho_X(x) = \rho_Y(c(x))c'(x) \tag{S79}$$

$$p_X(x) = p_Y(c(x)) \tag{S80}$$

Combining the two equations we find the quantities can be connected through

$$\rho_X/(p_Y^{-1} \circ p_X)' = \rho_Y \circ p_Y^{-1} \circ p_X \tag{S81}$$

Indeed, selecting appropriate transformations one can apply the derived protocols for other existing models in learning curve studies. To demonstrate this, let us consider a selection model documented in[43]. Here we define $\rho_X = x^{\beta-1}, \rho_Y = 1, p_X = 1$, we obtain $c(x) = x^\beta/\beta$ and $p_Y = 1$ (i.e. $\alpha = 0$), assuming $k = \infty$ we have

$$\langle x_n \rangle \sim \langle y_n^{*(1/\beta)} | k = \infty, \alpha = 0 \rangle \sim n^{-1/\beta} \tag{S82}$$

In this way we arrive at a system y that is mathematically equivalent to existing model systems[43], where one has power law cost distribution, try new versions at every attempt and learns from all past experiences. Hence our approach is also able to recover this $n^{-1/\beta}$ scalings ($n^{-1/k}$ using notations in original paper[43]) documented in learning curve models through mathematical transformations. For following discussions we always assume $\rho \equiv 1$ and consider different forms of $p(x)$.

Our previous results have shown solutions for $p(x) = (1-x)^\alpha$, prompting us to consider a more general form using expansion

$$\ln(p(x^*) - p(1)) = \alpha \ln(1 - x^*) + o(\ln(1 - x^*)), \;\; x \to 1 \tag{S83}$$

where $\alpha \equiv \lim_{x^* \to 1} \frac{\ln p(x^*)}{\ln(1-x^*)} \geq 0$ captures the asymptomatic behavior of p near $x^* \to 1$. If $p(1) > 0$, there is certain positive probability that one will create a new one, no matter how good

she did, which will cause both t_n and x_n converging to positive limit. On the other hand, when $p(1) = 0$, we can approximate $p(x^*) \sim (1 - x^*)^\alpha$, hence we should observe the same scaling as $p(x^*) \sim (1 - x^*)^\alpha$. Indeed, all our previous derivations only rely on the power law tail of $x^{-k\alpha}$ rather than a precise power law form.

Despite its simplicity, the assumption enables us to work with a broad range of functions, including all functions that are analytic at $x^* = 1$ (e.g. $p = e^{c(1-x^*)} - 1 \sim c(1 - x^*)$) as well as many that are not (e.g. $p = (1 - x^*)^c$ with non-integer c) through a single parameter α. Note that this relaxation in the functional form of $p(x)$ is again closely related to the relaxation in $\rho(x)$ documented in[43] due to the relationship between the two quantities we discussed before.

S3.7 Null models

Our model demonstrates that both experience and evaluations play an important role in dynamics of failure. To verify that both ingredients are necessary, we investigate two variants of the model.

To understand the role of experience, we explore a model (a) assuming that an individual does not reuse past versions. We find model (a) reduces to the case of $k = 0$, where each attempt is made independently. Again, we recover results from S3.2, predicting constant efficiency $t_n = 1$ and quality $x_n = 0.5$.

We then keep the experience mechanism, but eliminate the role of evaluations by assuming that one chooses to reuse past version regardless of its score. In other words, model (b) assumes

that the probability to create a new version, p, is constant, independent of past scores. This allows us to write the master equation as

$$x^*_{n+1} \sim \begin{cases} U[0,1], & w.p.\ p \\ \delta(x^*_n), & w.p.\ 1-p \end{cases} \tag{S84}$$

By induction one has $x_n \sim U[0,1]$ for any n, again predicting constant efficiency $t_n = p$ and quality $x_n = 0.5$. This indicates that in the absence of evaluations the model fails to reproduce the observed scaling behavior. Indeed, the improvement in the original model is mainly driven by reuse preference towards version with higher-scores, explaining why it does not exist in this null model.

Together, the predictions of these two alternative models indicate that a combination of the two ingredients is essential for the emergence of scaling observed in Fig. 3. One may also hypothesize that the uncovered patterns are affected if we define the finite capacity using the unit of time (t) rather than trials (n), prompting us to consider a model (c): Here we assume that individuals at time t consider all past failures that occurred during a time window τ, i.e. individuals at time t consider all past failures that occurred during a time interval $(t-\tau, t]$, where the window size τ, instead of our previous parameter k, measures how long one looks back upon past failures. We further assume that the number of components equals to one for simplicity. The previous master equation can be written as

$$x^*_n = \max_{t_m+\cdots+t_n \le \tau} \{x_m\} \tag{S85}$$

$$x_n \sim \begin{cases} U[0,1], & w.p.\ (1-x_n^*)^\alpha \\ \delta(x_n^*), & w.p.\ 1-(1-x_n^*)^\alpha \end{cases} \tag{S86}$$

To solve this model, we note that the following equations hold.

$$|\{n_1 \le n \le n_2 : x_n = x_m^*\}| \le n_2 - n_1 + 1 \tag{S87}$$

$$|\{n_1 \le n \le n_2 : x_n = x_m^*\}| \ge \sum_{i=0}^{[(n_2-n_1)/(\tau+1)]-1} \sum_{j=0}^{\tau} I(x_{n_1+(\tau+1)i+j} = x^*_{n_1+(\tau+1)i+i}) \ge [(n_2-n_1)/(\tau+1)] \tag{S88}$$

This is because, if we consider $\tau+1$ versions $(x_i, \cdots, x_{i+\tau})$, we should find (1) at least two of the versions are the same or (2) these are $\tau+1$ different versions. If (1) is true, i.e. $x_j = x_k$ for some $i \le j < k \le i+\tau$, we have $x_j = x_k^*$, i.e. the duplicated version is a baseline version. Otherwise, (2) means that there are at least τ new versions, covering all versions over the last τ time units. Hence we have $x^*_{i+\tau+1} \in \{x_i, \cdots, x_{i+\tau}\}$.

Using the notations in previous derivations, we can also recover the master equation as

$$z_{m+1} = \begin{cases} z_m & w.p.\ \frac{[1-z_m^{\tau\alpha}(1-z_m)^\tau](1-z_m^\alpha)}{1-z_m^\alpha(1-z_m)} \sim O(1) \\ U[0,z_m] & w.p.\ \frac{[1-z_m^{\tau\alpha}(1-z_m)^\tau]z_m^{\alpha+1}}{1-z_m^\alpha(1-z_m)} \sim O(z_m^{\alpha+1}) \\ \min\{U_1[0,1], \cdots, U_\tau[0,1]\} & w.p.\ z_m^{\tau\alpha}(1-z_m)^\tau \sim O(z_m^{\tau\alpha}) \end{cases} \tag{S89}$$

To this end, we find that this variant of the model is asymptotically similar to our original model, with $\tau^* = k^* = 1/\alpha$. Indeed, when a baseline version is out of date and gets replaced, the recency mechanism happens after k^* (τ^*) new versions have been created without reuse, explaining

why τ^*, the critical number of different versions to look back, equals to k^*, the critical number of versions to look back.

S3.8 Failure streak length

To understand the fat-tailed distribution documented in Fig. 1, let us consider a single-component case of our model for simplicity. We assume that q, the probability for a new version to success, is independent of its score. We denote N as the number of failures before success.

Assume $N \geq n$, i.e. one has not achieved success in the first n attempts. For one to succeed in the $(n+1)$-th attempt, she needs to (1) create a new version at this time, corresponding to probability $t_n \sim n^{-\gamma}$ and (2) succeed for this new version, which has probability q. Together we obtain

$$P(N = n | N \geq n) \sim q n^{-\gamma} \tag{S90}$$

Note that this form is closely related with Lindy's law[83,84]. Here the right hand side of the equation is decreasing, since a long failure streak indicates the existence of an (unsuccessful) version that has been used for a long period. Therefore, the same version is more likely reused again in the future, reducing the chance to create a new, successful version at the next step.

If we define the survival function $S(n) = P(N \geq n)$, this equation is equivalent to

$$1 - S(n+1)/S(n) \sim q n^{-\gamma} \tag{S91}$$

Using a continuous approximation we obtain

$$-\frac{dS}{S} \sim qn^{-\gamma}dn \tag{S92}$$

leading to the solution

$$P(N \geq n) = S(n) \sim e^{-cn^{1-\gamma}} \tag{S93}$$

Hence, it predicts that the length distribution follows the well-known Weibull distribution.

To further understand the Weibull form, here we recognize that it is closely related to Heaps' law[74] caused by the reuse mechanism. Indeed, given that one needs to create M different versions before success, the distribution can be formulated as an exponential model

$$P(M \geq m) = (1-q)^m \tag{S94}$$

However, repeated reuse leads to a sub-linear scaling between N and M, following the Heaps' law with exponent $1-\gamma$:

$$M(N) = \sum_{n=1}^{N} t_n \sim \sum_{n=1}^{N} n^{-\gamma} \sim N^{1-\gamma} \tag{S95}$$

Combining the two equations one can obtain the same Weibull model

$$P(N \geq n) = S(n) \sim e^{-cn^{1-\gamma}} \tag{S96}$$

We can further relax our assumption by considering success probability q as a function of evaluation score x. As long as (i) $q(x)$ is non-decreasing with x, hence a better score corresponds to a higher probability of success, and (ii) $q(x) < 1$ for all x, we have

$$P(M \geq m) = \left(1 - \int_0^1 q(x)dx\right)^m \tag{S97}$$

Using the sub-linear scaling between M and N, the failure streak length is found to be again captured by the Weibull distribution. An interesting insight from these results is that all quantities of interest exhibit scale-free properties. This means if we consider different criteria of success that are organized into hierarchal structures, our results are robust against the selection of criterion.

One assumption in this analysis is that eventual success comes from creation of new versions rather than simple reuse. Hence it also predicts that the last inter-event time (time between penultimate failure and eventual success) has a lower bound and in empirical settings may appear longer than expected , especially for domains with higher learning rate. This is consistent with our observations on D_3, where the penultimate inter-event time could be higher than previous ones. This selection issue can be resolved by calculating $T_n = t_n/t_1$ for all samples with at least n data points, as we did in Fig. 3.

Another possibility that can lead to fat-tailed distributions is fitness heterogeneity[17]. Indeed, since different individuals may have different fitness, it might be possible that the fat tail of failure streaks can emerge without the reuse mechanism. To test if this is sufficient for modeling dynamics of failure, here we compare it with other observations, finding the fitness hypothesis cannot account for the observed patterns for a series of reasons:

1. Initial performance fails to predict eventual outcome. One direct prediction of the fitness hypothesis is the predictive power of initial performance for the eventual outcome. However, as shown in Figs. 4g-i, we find that for large n, the success and unsuccessful group show no

statistical differences at the first attempt, which is in strong contrast with our prediction.

2. Weak correlation between initial performance and failure streak. Assuming performance dynamics is mostly driven by fitness heterogeneity, those who succeeded with fewer failures should show better performance at the very beginning. Hence, one would expect a strong correlation between initial performance and failure streak. However, we find that across the three datasets, the correlation between the two is weak (Extended Data Fig. 10).

3. Fat tail remains as we control fitness. If the fat-tail is caused by a broad distribution of fitness, we should observe a narrower tail as we control the fitness through conditioning on initial performance. Our results show that, as we conditional on samples with top/bottom performance at the beginning, $P(N)$ still distinguishes from the exponential model (Extended Data Fig. 10).

4. Failure dynamics. Most importantly, the fitness hypothesis states that success and unsuccessful groups lie on a continuous spectrum, hence we should not expect fundamental differences in their temporal patterns. To this end, it fails to account for the observed patterns documented in Figs. 3d-f.

S4 Generalized models

The one-parameter k-model discussed above offers a simple framework to quantify the complex dynamics underlying failures. It can be generalized into richer frameworks by taking into account more realistic assumptions. Here we present two variants of the model. While the key predic-

tions of our original model (i.e. the stagnation and progression regimes) remains intact, the new frameworks add more flexibility to the model, exhibiting richer mathematical properties.

S4.1 $k - \alpha$ model

The original model has one parameter k, measuring how memory length affects failure dynamics. Yet agents may differ in the judgment of their own work or incentives to change given feedback. Here we consider $p = 1-(1-x^*)^\alpha$, where α quantifies probability to create a new version p given score x (Extended Data Fig. 2). Indeed, $\alpha = 0$ indicates that no matter what evaluation one gets, the agent will always create a new version (thrash around with new versions). On the other hand, $\alpha \to \infty$ points to the other extreme where one does not create a new version unless it is extremely bad.

To explore the role that α plays, let us revisit the analytical results and substitute k^* back with $1/\alpha$, obtaining solutions to this two-parameter model

$$\gamma = \begin{cases} 0, & k < 1/\alpha \\ 1 - k^*/k = 1 - 1/(k\alpha), & 1/\alpha \le k < 1/\alpha + 1 \\ \alpha/(\alpha+1), & k \ge 1/\alpha + 1 \end{cases} \tag{S98}$$

We discover a two-dimensional phase diagram with three different phases (Extended Data Fig. 2b). The boundaries separating different phases are $k\alpha = 1$ and $(k-1)\alpha = 1$, respectively. This means, if we fix α, the two boundaries reduce back to the two critical points k^* and $k^* + 1$

($\alpha = 1/k^*$), consistent with all our findings for the previous k-model. On the other hand, if we fix k and vary α, there always exists a critical point that separates the stagnation and progression regimes at $\alpha = 1/k$. This result predicts that agents who base their decisions more carefully on feedback evaluations (higher α) can have a higher scaling exponent γ even holding k constant, allowing us to incorporate alternative explanations into a more general modeling framework.

We can also find the location of the second critical point by solving the condition $k\alpha = \alpha+1$, obtaining $\alpha = 1/(k-1)$. This point is well-defined only if $k > 1$, which is consistent with predictions from the previous k model (second critical point at $k^* + 1 > 1$). This suggests that for agents who only learn from last failure ($k = 1$), although the scaling exponent $\gamma(1, \alpha) = \max(0, 1 - 1/\alpha)$ can be relatively high by selecting a high α, this strategy cannot enter the third regime. Hence it is always sub-optimal as we compare it with $\gamma(\infty, \alpha) = 1 - 1/(\alpha + 1)$.

Our model further offers a quantitative approach to understand the interplay between learning and incentive (i.e. parameters k and α) – the preference to borrow from prior attempts and sensitivity to ongoing feedback (Extended Data Fig. 2b). Given that both k and α are important factors that could affect outcomes, the phase diagram presented here allows us to quantify their joint effect. Here we find that, the two parameters jointly define an 'effective' $K \equiv k - k^* = k - 1/\alpha$ (S75). The critical boundaries reduce into two simple equations: $K = 0$ and $K = 1$. According to (S76), the scaling relationship collapses into a simple equation: $\Gamma = \min\{\max\{0, \Gamma\}, 1\}$. As we inspect this effective parameter more, we find that those who have higher α have a larger effective K.

In the previous k model, the scaling exponent γ is upper bounded by $\gamma(k) \leq \gamma(\infty) =$

$\alpha/(\alpha+1)$. Canonical studies on learning have reported varying scaling exponents ranging between 0 and 1[22]. These two results are reconciled following the introduction of α, the incentive parameter. Indeed, as α takes different values, going from 0 to ∞, all possible values that γ can take are exactly captured within the interval $[0, 1]$. This further makes an important prediction for future studies to test: although the co-emergence of stagnation and progression regimes can happen for agents with different incentive levels, those with higher incentive for success may be identified by a higher upper bound for their learning rate. This opens up a new avenue to diagnose the productive and pathological implications of incentives for successful productivity.

S4.2 $k - \alpha - \delta$ model

The original model also assumes that one has perfect inference of past feedback and always selects the best among last k versions as the baseline. Yet individuals or organizations may not always chhose the best version, motivating us to frame the selection of baseline versions in a probabilistic fashion for $k > 0$, i.e.

$$P(i) = \frac{f(x_i)1_{n-k\leq i\leq n-1}}{\sum_{j=n-k}^{n-1} f(x_j)} \tag{S99}$$

where f is a function that maps real quality of a version to an individual's inference of its quality. One common way to formulate f in a model like this is by assuming $f(x) = (1-x)^{-\delta}$ with $\delta \geq 0$. In this formulation, $\delta = 0$ means one cannot differentiate quality between past versions and selects randomly among different versions. By contrast, $\delta \to \infty$ means one always chooses the prior version with the highest quality, converging back to the case we considered in the $k - \alpha$ model. To this end, the δ parameter generalizes our previous model to situations where one has imperfect

recognition of quality.

This generalization leads to a more practical interpretation of the model, yet at the same time poses a technical challenge to solve the model. Indeed, the baseline version under this model may not be a high-quality one. Rather, it is now possible that one repeatedly reuses low-quality versions, even though better ones are also available. This means we have to develop a new theoretical strategy to solve this model. Here we track the duration of a specific version with score $x \equiv 1 - z(z \to 0)$ and calculate the score composition of the most recent k versions. We observe the state once every k versions, i.e., if we count the state based on versions $(x_{n-k}, \cdots, x_{n-1})$, next time we count the state based on versions $(x_n, \cdots, x_{n+k-1})$. Here we introduce a new notation for different versions:

1. $\circ$ representing there is at least one version with the same score (x) in recent k versions.
2. $\uparrow$ representing there is at least one version with a higher score ($> x$) in recent k versions.
3. $\downarrow$ representing there is at least one version with a lower score ($< x$) in recent k versions.

There are in total 7 states as we inspect the composition of k consecutive versions: $[\circ]$, $[\downarrow \circ]$, $[\circ \uparrow]$, $[\downarrow \circ \uparrow]$, $[\downarrow]$, $[\uparrow]$, and $[\downarrow\uparrow]$. We are mostly interested in the system state where $\circ$ is no longer available, which prompts us to consider the last three states as absorbing states. We next calculate the leading factor for state transition probabilities between the transient states.

1. If we start from $[\circ]$:

$$P([\circ] \to [\circ]) = O(1 - z^{\alpha})^{k} \approx 1$$

$$P([\circ] \to [\downarrow \circ]) \sim z^{\alpha}(1 - z)O(1) \sim O(z^{\alpha})$$

$$P([\circ] \to [\circ \uparrow]) \sim z^{\alpha} z O(1) \sim O(z^{\alpha+1})$$

To calculate $P([\circ] \to [\downarrow \circ \uparrow])$, we recognize that the key of this transition is to create at least two different new versions, with at least one of which being a higher-quality version. Starting from $[\circ]$, the probability of creating a new version is z^{α}. We can calculate the probability of creating the second new version, which can be approximated as

$$\int_0^1 \frac{y^{\alpha-\delta} + z^{\alpha-\delta}}{y^{-\delta} + z^{-\delta}} dy = \int_0^1 \frac{z^{\alpha-\delta}}{y^{-\delta} + z^{-\delta}} dy + \int_0^1 \frac{y^{\alpha-\delta}}{y^{-\delta} + z^{-\delta}} dy$$

The first term on the right-hand side (RHS) estimates the probability of creating a new version when the original version is used as the baseline, which can be estimated by

$$\int_0^1 \frac{z^{\alpha-\delta}}{y^{-\delta} + z^{-\delta}} dy = z^{\alpha+1} \int_0^{1/z} \frac{u^{\delta}}{u^{\delta} + 1} du \sim z^{\alpha+1} \left[\int_0^1 u^{\delta} du + \int_1^{1/z} 1 du \right] \sim z^{\alpha}$$

The second term on the RHS estimates the probability of creating a new version when the first new version is used as the baseline, which can be estimated by

$$\int_0^1 \frac{y^{\alpha-\delta}}{y^{-\delta} + z^{-\delta}} dy = z^{\alpha+1} \int_0^{1/z} \frac{u^{\alpha}}{u^{\delta} + 1} du \sim z^{\alpha+1} \left[\int_0^1 u^{\alpha} du + \int_1^{1/z} u^{\alpha-\delta} du \right] \sim z^{\min\{\alpha+1, \delta\}}$$

Given the new versions, the probability for at least one of them having a score higher than x is z, helping us calculate the state transition probability:

$$P([\circ] \to [\downarrow \circ \uparrow]) \sim O(z^{\alpha}) O(z^{\alpha} + z^{\min\{\alpha+1, \delta\}}) z O(1) \sim O(z^{\min\{\alpha, \delta\}+\alpha+1})$$

2. If we start from $[\downarrow \circ]$: Going to $[\circ]$ means one need to repeatedly reuse the original version for k times, each time happening with probability

$$\int_0^1 \frac{z^{-\delta}(1-z^{\alpha})}{y^{-\delta}+z^{-\delta}}dy \sim z\int_0^{1/z}\frac{u^{\delta}}{u^{\delta}+1}du = 1 - z\int_0^{1/z}\frac{1}{u^{\delta}+1}du \approx 1$$

Together we have

$$P([\downarrow \circ] \to [\circ]) \approx 1$$

Approaching $[\downarrow \circ]$ mostly needs a $\downarrow$ version, which is with probability

$$P([\downarrow \circ] \to [\downarrow \circ]) \sim O(1)O(z^{\min\{\alpha,\delta\}} + z^{\min\{1,\delta\}}) \sim O(z^{\min\{\alpha,\delta,1\}})$$

To arrive at $[\circ \uparrow]$ we need to create a new version and make sure it has high quality, i.e.

$$P([\downarrow \circ] \to [\circ \uparrow]) \sim zO(z^{\min\{\alpha,\delta\}}) \sim O(z^{\min\{\alpha,\delta\}+1})$$

Combining these two conditions we can also derive

$$P([\downarrow \circ] \to [\downarrow \circ \uparrow]) \sim O(z^{\min\{\alpha,\delta,1\}+\min\{\alpha,\delta\}+1})$$

3. If we start from $[\circ \uparrow]$: Note that the probability to choose between $\circ$ or $\uparrow$ versions when both are available can be approximated by non-zero constant. This is because the score of a $\uparrow$ version can be well approximated by $[x, 1]$, hence the score of $\circ$ and $\uparrow$ are comparable. Using this fact and results from case 1 we have

$$P([\circ \uparrow] \to [\circ]) \sim O(1)$$

The probability is approaching a non-zero constant, but it is strictly smaller than 1, since we also have

$$P([\circ \uparrow] \to [\circ \uparrow]) \sim O(1)$$

If one creates a version that is $\downarrow$, we have

$$P([\circ\uparrow] \to [\downarrow\circ]) \sim O(z^{\alpha})$$

$$P([\circ\uparrow] \to [\downarrow\circ\uparrow]) \sim O(z^{\alpha})$$

4. If we start from $[\downarrow\circ\uparrow]$, derivations of transition probabilities are similar to case 2, following

$$P([\downarrow\circ\uparrow] \to [\circ]) \sim O(1)$$

$$P([\downarrow\circ\uparrow] \to [\downarrow\circ]) \sim O(z^{\min\{\alpha,\delta,1\}})$$

The only difference is that we do not need to create new versions to obtain $\uparrow$. To the contrary, there are existing $\uparrow$ to be reused, leading to

$$P([\downarrow\circ\uparrow] \to [\circ\uparrow]) \sim O(1)$$

$$P([\downarrow\circ\uparrow] \to [\downarrow\circ\uparrow]) \sim O(z^{\min\{\alpha,\delta,1\}})$$

$$P([\downarrow\circ\uparrow] \to [\downarrow\circ\uparrow]) \sim O(z^{\alpha})$$

	$[\circ]$	$[\downarrow \circ]$	$[\circ \uparrow]$	$[\downarrow \circ \uparrow]$
$[\circ]$	0	0	0^*	0^*
$[\downarrow \circ]$	α	$\min\{\alpha, \delta, 1\}$	α	$\min\{\alpha, \delta, 1\}$
$[\circ \uparrow]$	$\alpha + 1$	$\min\{\alpha, \delta\} + 1$	0^*	0^*
$[\downarrow \circ \uparrow]$	$\min\{\alpha, \delta\} + \alpha + 1$	$\min\{\alpha, \delta, 1\} + \min\{\alpha, \delta\} + 1$	α	$\min\{\alpha, \delta, 1\}$

Table S1: **Scaling behavior of the model with δ parameter.** We calculate the approximate state transition matrix for the score composition of recent k versions. $\circ$: a same-score version, $\uparrow$: a higher-score version, $\downarrow$: a lower-score version. Numbers are exponents of the probability $\lim_{z \to 0} \frac{\ln P(1-z, s' \to s)}{\ln z}$, with 0 denoting $P \to 1$ and 0^* denoting $P \to c$ for some $c \in (0, 1)$.

What is the limiting distribution of these four states conditional on the system has not been absorbed? To solve this we again represent scaling forms as the leading factor, i.e. for a state s, one has $\mathrm{P}(s) \sim z^{\beta_s}$. Recognizing that $P(s) = \sum_{s'} P(s')P(s' \to s)$ and $\sum_s P(s) = 1$, we can solve the equations

$$\begin{cases} \beta_s & = \min_{s'}\{\beta_{s'} + \lim_{z\to 0} \frac{\ln P(s'\to s)}{\ln z}\} \\ \min_s \beta_s & = 0 \end{cases} \tag{S100}$$

Substituting our previous results into the equation we need to solve

$$\begin{cases} \beta_{[\downarrow\circ]} = \min\{\beta_{[\circ]} + \alpha, \beta_{[\circ\uparrow]} + \alpha, \beta_{[\downarrow\circ\uparrow]} + \min(\alpha, \delta, 1)\} \\ \beta_{[\circ\uparrow]} = \min\{\beta_{[\circ]} + \alpha + 1, \beta_{[\downarrow\circ]} + \min(\alpha, \delta) + 1, \beta_{[\downarrow\circ\uparrow]}\} \\ \beta_{[\downarrow\circ\uparrow]} = \min\{\beta_{[\circ]} + \alpha + \min(\alpha, \delta) + 1, \beta_{[\downarrow\circ]} + \min(\alpha, \delta, 1) + \min(\alpha, \delta) + 1, \beta_{[\circ\uparrow]} + \alpha\} \\ \min\{\beta_{[\circ]}, \beta_{[\downarrow\circ]}, \beta_{[\circ\uparrow]}, \beta_{[\downarrow\circ\uparrow]}\} = 0 \end{cases} \tag{S101}$$

The only solution to this set of equations is

$$(\beta_{[\circ]}, \beta_{[\downarrow\circ]}, \beta_{[\circ\uparrow]}, \beta_{[\downarrow\circ\uparrow]}) = (0, \alpha, \alpha + 1, \min\{\alpha, \delta\} + \alpha + 1)$$

This solution allows us to calculate the state after $\circ$ is eventually abandoned (before normalization).

We first calculate the probability for at least one better version to exist:

$$P([\circ] \to [\uparrow]) \sim O(z^{\alpha})$$

$$P([\circ\uparrow] \to [\uparrow]) \sim P([\downarrow\circ\uparrow] \to [\uparrow]) \sim O(1)$$

$$P([\downarrow\circ] \to [\uparrow]) \sim O(z^{\min\{\alpha,\delta\}+1})$$

Together we have

$$P([\uparrow]) = \sum_{s\in\{[\circ],[\downarrow\circ],[\circ\uparrow],[\downarrow\circ\uparrow]\}} P(s)P(s \to [\uparrow]) \sim O(z^{\alpha+1})$$

Next then consider $P([\circ] \to [\downarrow])$, to achieve k versions other than $1-z$ one need to (1) create a lower-scored version based on $1-z$ (with probability $z^\alpha(1-z) \sim z^\alpha$) and (2) generate other $k-1$ versions, either by creating new ones or by reusing the ones just created. Each step in (2) happens with probability $O(z^{\min\{1,\alpha,\delta\}})$, allowing us to calculate the leading term as

$$P([\circ] \to [\downarrow]) \sim O(z^{\alpha+(k-1)\min\{1,\alpha,\delta\}})$$

Similarly we have

$$P([\circ\uparrow] \to [\downarrow]) \sim O(z^{\alpha+(k-1)\min\{1,\alpha,\delta\}})$$

Calculating the remaining two probabilities are more complicated, here we take $[\downarrow\circ] \to [\downarrow]$ as a example. Indeed, we find this probability largely depends on the concrete composition. Consider $k=3$, if we start from the state $(\downarrow,\circ,\circ)$ then the probability is in the order of $O(z^{3\min\{\alpha,\delta,1\}})$, yet if we start from $(\circ,\circ,\downarrow)$ the probability is in the order of $O(z^{2\min\{\alpha,\delta,1\}})$. To this end, we index each state (consecutive k versions, i.e. $(x_{n-k},\cdots,x_{n-1})$) by a state m, defined as

$$m = \min_{1\le i\le k}\{x_{n-i} = 1-z\}$$

We find the transition probability from state i to state j follows $P_{i\to j} \sim O(z^{(j-1)\min\{\alpha,\delta,1\}})$, which leads to $P_i \sim O(z^{\alpha+\max\{i-2,0\}\min\{\alpha,\delta,1\}})$. Also note that $P_{i\to[\downarrow]} \sim O(z^{(k-i+1)\min\{\alpha,\delta,1\}})$. Together we have the probability that one starts from $[\downarrow\circ]$ and reaches $[\downarrow]$ at the next step follow

$$\sum_{i=1}^{k} P_i P_{i\to[\downarrow]} \sim \sum_{i=1}^{k} O(z^{\alpha+\max\{i-2,0\}\min\{\alpha,\delta,1\}})O(z^{(k-i+1)\min\{\alpha,\delta,1\}}) \sim O(z^{\alpha+(k-1)\min\{\alpha,\delta,1\}})$$

We can obtain similar results for $[\downarrow \circ \uparrow] \to [\downarrow]$. Together we have

$$P([\downarrow]) \sim z^{(k-1)\min\{\alpha,\delta,1\}+\alpha}$$

The other possibility, $P([\uparrow\downarrow])$ involves higher-order terms and can be therefore neglected in this derivation.

Together, these results allow us to map an analogy between this model and the $k-\alpha$-model. Indeed, the three important indexes that determine the separation of regimes used to be 1, $k\alpha$ and $\alpha+1$. Yet here it becomes 1, $(k-1)\min\{\alpha,\delta,1\}+\alpha$ and $\alpha+1$. Using a similar technique in S3 we obtain

$$\gamma = \begin{cases} 0, & k < 1/\alpha \\ 1 - 1/[(k-1)\min\{\alpha,\delta\}+\alpha], & 1/\alpha \le k < 1/\min\{\alpha,\delta\}+1 \\ \alpha/(\alpha+1), & k \ge 1/\min\{\alpha,\delta\}+1 \end{cases} \tag{S102}$$

Or equivalently

$$\gamma(k,\alpha,\delta) = 1 - \{\max[\min(\alpha+(k-1)\min\{\alpha,\delta\}, \alpha+1), 1]\}^{-1}$$

Here we simplified $\min\{\alpha,\delta,1\}$ into $\min\{\alpha,\delta\}$. This is because, for 1 to be taken into consideration we need $(k-1)+\alpha < \alpha+1$, i.e. $k<2$. While for $k=1$, $(k-1)\min\{\alpha,\delta,1\}+\alpha = (k-1)\min\{\alpha,\delta\}+\alpha = \alpha$.

Most interestingly, with the addition of the δ parameter, this three-parameter model is found to induce four different phases (Extended Data Fig. 10d). Three of the regimes are generalizations of those found in the $k-\alpha$ model, where the scaling exponent γ does not depend on δ, following

$\gamma(k, \alpha, \delta) = \gamma(k, \alpha, \infty)$ (Extended Data Fig. 10c). The fourth one, however, is a new phase and only exists for small δ. In this regime, the inability to select a high-quality version (small δ) dominates the scaling behavior, with exponent $\gamma(k, \alpha, \delta) = 1 - \frac{1}{(k-1)\delta+\alpha}$. The phase diagram reveals novel mathematical properties with the introduction of the δ parameter. Here we summarize three key, novel insights offered by this model.

- The (α, δ) phase diagram shows two triple points, located at $(k, 1/k, 1/k)$ and $(k, 1/(k-1), 1/(k-1))$ respectively (Extended Data Fig. 10d). The existence of these triple points indicates that if we fix (k, α) or (k, δ), the four regimes cannot exist simultaneously. For example, if we tune δ for a given pair of k and α, we can find one (if $\alpha < 1/k$), two (if $\alpha > 1$), or three (if $1/k < \alpha < 1$) different regimes. The most relevant case is $1/k < \alpha < 1$, where δ induces a transition across scaling and non-scaling regimes at $\delta^* = \frac{1-\alpha}{k-1}$. Hence, in this regime we observe a phase transition in δ. Yet outside this interval, changing δ would not reproduce the transition between stagnation and progression regimes.

- We can further combine the three parameters and extend the effective K though $K \equiv 1 - 1/\alpha + (k-1)\min(1, \delta/\alpha)$. The scaling and non-scaling regimes are again separated by $K = 0$, manifesting its consistency with predictions of the k-model. This formulation allows us to understand the role each parameter plays in determining the learning rate. For example, when one has low δ ($\delta < \alpha$), looking back on more failures (increasing k) can increase the effective K. But this strategy is less effective in this case because of the low return on effective K, i.e. the slope δ/α is smaller than 1.

- The phase transitions induced by δ degenerate when $k \to \infty$. Indeed, the critical points δ are upper bounded by $1/(k-1)$, which goes to 0 in the case of $k \to \infty$. This means that as long as one has some ability of calibration on quality ($\delta > 0$) and a large number of failures (large k) to consider, increasing δ alone does not help. This result lends further support for our previous findings, showing that for a wide range of practically reasonable regimes, we arrive at the same conclusion by only considering k and without considering δ. This also demonstrates that while the full model at this point has three parameters, k remains as the most fundamental parameter, which by itself can generate various scaling behaviors, especially being the key to account for the separation of stagnation and progression regimes.

S5 Empirical measurements

S5.1 Quantifying performance dynamics

Here we leverage our three datasets and compile three different measurements for performance.

For the NIH grant application dataset, we make use of the percentile scores assigned by NIH review panels. NIH uses a two-step peer review mechanism: Roughly half of the proposals are selected for the second round discussion, where each proposal is given a percentile score based on their percentile ranking among its peers. Percentile score has been widely used to measure the quality of R01 grant applications[28,85], reflecting judgment of expert reviewers. Although reviewers score are necessarily imperfect, there is growing evidence for strong correlations between percentile score and subsequent successes of the project[51,86]. One disadvantage of using the per-

centile score is that undiscussed proposals (those get rejected in the first round) do not have such scores. Moreover, since there exist differences concerning the discussion rate, applications lying on the boundary of discussion can have either marginal scores or no scores. Indeed, here we calculate the proportion of having a percentile score around 57% and plot the score distribution. We find as score exceeds 50, there are much fewer samples, since many proposals at this rank did not even get discussed and assigned a score. To avoid discrimination across study sections, here we take score below 50 and regard the remaining proposals as undiscussed. We also vary the threshold to 55, finding results remain the same. Lower percentile scores indicate better performance. To be consistent with other measures (higher the better) we rescale the percentile scores using 1-0.01$\times$original score, so the values reported in main text are bigger the better.

To measure the performance in startup ventures, we leverage the investment amount in the first funding round as a proxy. Although there are a series of firm-level statistics that could potentially measure the quality of a venture, investment amount stands out as a preferred choice of representing investor evaluations. This definition does not account for geographical and industrial factors, as such information is not available to us, but it serves as a reasonable index of startup companies potentials in achieving their eventual goals (IPO or high-value M&As).

Similar to other frequently used measures in economics, investment amount follows a fat-tailed distribution and exhibits time-dependent properties. To address the two challenges, we take logarithmic of the investment amount and calculate z-score within each year. Denoting the amount of all investments made in year t as $\{s_1^t, \cdots, s_n^t\}$, here we rescale the values into the performance

score z through

$$z_i^t = \frac{\log(s_i^t) - \mathrm{E}[\log(s^t)]}{\sqrt{\mathrm{Var}[\log(s^t)]}}$$

Once rescaled, we find z_i^t approximately follow the standard normal distribution $N(0,1)$ independent from t, allowing us to directly compare attempts made in different years. We then compare first-round investment amounts for successful and failed attempts, finding the two samples are clearly separated.

Similarly, for terrorist attacks, one measure for performance is the number of individuals wounded, which is reported for more than 91% of the attacks recorded in the database. To this end, we collect wound statistics as our performance measure. Indeed, fatal (successful) attacks also lead to a higher number of wounded individuals than others, validating the effectiveness of using wounded statistics as performance measurements. Related studies of terrorist attacks suggest the outcome of attacks follow a power law distribution, which is also confirmed in our dataset. To this end, we rescale the original values by $\log(wounded + 1)$ in our analysis.

Note that although the overall coverage of performance measures is high (94% for D_2 and 91% for D_3), in both datasets there are missing values. To ensure that they do not affect our results, we also label these missing values as NA and exclude them as we analyze performance dynamics. Analyses that do not require performance information are measured on the full data sample.

Note that while statistical tests in performance dynamics consistently show a lack of improvement between first and second attempts for the unsuccessful group, they do not rule out the

possibility that this group may have decreased performance. Such performance dynamics are supported by some of our observation (Extended Data Figure. 6i, 8gi, 9jlm) and may be associated with competitions within the system. For example, performance of NIH proposals is evaluated by a percentile score, hence simply retaining the same performance may result in a worse score as his/her peers improve systematically.

S5.2 Length distribution of failure streaks

The length distribution of the failure streak, defined as the number of failures before success, is measured directly from data and fitted using maximum likelihood estimation techniques[47]. We fit empirical data with discrete version of Weibull (stretched exponential) form using maximum likelihood estimation with parameters $x_{\min} = 2$ and calculate uncertainty from bootstrapping over 100 simulations, yielding $\beta_1 = 0.666 \pm 0.017$, $\beta_2 = 0.566 \pm 0.086$, and $\beta_3 = 0.129 \pm 0.033$. Comparing this with γ estimated from temporal dynamics, two-sided t-tests indicate that none of the three datasets can reject the validity of the scaling identity $\beta + \gamma = 1$ ($P = 0.176,\ 0.421,\ 0.141$). We further compare the fitting results from alternative models, i.e. lognormal, power law, and truncated power law using likelihood ratio test[47], finding that Weibull distribution is consistently among the best functional forms (Table S3). To quantify the uncertainty of parameter estimations, we performed bootstrapping technique (100 times) to calculate optimal estimation for each round, and obtained standard error of parameter estimators. We also repeated the results for $x_{\min} = 3$, obtaining $\beta_1 = 0.592 \pm 0.032$, $\beta_2 = 0.513 \pm 0.175$, and $\beta_3 = 0.139 \pm 0.060$, which again statistically supports $\beta + \gamma = 1$.

To further test these results, we perform two randomization processes. We performed our first randomization operation, by keeping the timing and outcome of each attempt but changing the individual/organization associated with the attempt via random selection. The null model leads to exponentially distributed failure streaks (Fig. 1). We then performed a second randomization procedure by taking the samples used in Fig. 1 and shuffling the success/failure label from each attempt. This operation keeps constant both the overall success rate and the total number of attempts for each individual (Extended Data Fig. 4c-e). The two versions of randomization both lead to exponential like distributions, showing clear deviation from data.

Note that Fig. 1h-j and 3a-c only show results for less than 21 consecutive failures prior to the eventual outcome, accounting for 99.99%, 100%, 99.35% for the successful group and 99.99%, 100%, 99.60% for the unsuccessful group. All statistical tests are performed on the full data (100%).

S5.3 Measuring failure dynamics

Given the highly skewed distributions of N and t_n, to measure $T_n = t_n/t_1$ we first performed log transformation to calculate the mean and variance of $\log(T_n)$ from

$$\mathrm{E}[\log(T_n)] = \langle \log(t_n/t_1) \rangle \tag{S103}$$

$$\mathrm{Var}[\log(T_n)] = \langle [\log(t_n/t_1)]^2 \rangle - \langle \log(t_n/t_1) \rangle^2 \tag{S104}$$

where we take $t_n = \max\{t_n, 1\}$ when necessary. We have also checked the robustness of this operation by trying to replace 1 with 0.5, finding the results remain similar. As the number of

samples decreases dramatically with n, here we focus on $n \leq 10$ for D_1, $n \leq 7$ for D_2, and $n \leq 4$ for D_3.

The two equations immediately give us mean $\mathrm{E}[\log(T_n)]$ and standard error of the mean $\sqrt{\mathrm{Var}[\log(T_n)]/\text{sample size}}$, as plotted in Fig. 3. The divergence between the two groups can be detected as early as the second attempt. Although $T_1 \equiv 1$ by construction, Student's t-test rejects the hypothesis that $\log(T_2)$ between success and unsuccessful groups are the same ($P = 0.000457$, 0.00773, and 0.0992, respectively).

To calculate the temporal scaling exponent γ, here we run linear regressions between $\log(n)$ and $\log(T_n)$ and take the negative slope as γ, i.e.

$$\log(t_n/t_1) = -\gamma \log(n) + c, \tag{S105}$$

yielding $\gamma_1 = 0.361 \pm 0.010$, $\gamma_2 = 0.509 \pm 0.036$ and $\gamma_3 = 0.640 \pm 0.153$ for successful group, with $P < 0.001$ for all three datasets. We also performed individual fixed effect linear models using samples with at least three data points, i.e.

$$\log(t_{n,j}) = -\gamma \log(n) + c_j + \epsilon_{n,j}, \tag{S106}$$

where j is the index for different samples and c_j is the fixed effect term for each agent j. We obtain similar results $\gamma_1 = 0.372 \pm 0.017$, $\gamma_2 = 0.431 \pm 0.077$ and $\gamma_3 = 0.685 \pm 0.182$. For unsuccessful group there exists no significant relationships between $\log(n)$ and $\log(T_n)$ since the second failure (i.e. excluding T_1), with $P = 0.450$, $P = 0.884$ and $P = 0.957$ respectively. Together, these results offer strong empirical support for the diverging temporal patterns predicted by our model.

S5.4 Quantifying component dynamics

In our modeling attempts, we treat components as purely abstract properties of a grant proposal, fledgling company, or terrorist campaign. Here we further consider if we can empirically measure or approximate components, thereby better estimating and understanding their dynamics and validating the descriptive power of our model. The difficulty of this measurement stems from the fact that the existing datasets obtained above, while extensive, are nevertheless inadequate in this respect. Indeed, unlike scientific papers, which have reference information that can approximate the units of knowledge they piece together, grant proposals are largely isolated documents, making it difficult to infer the 'substance' of each proposal. Furthermore, while some metadata are associated with each proposal, such as funding institute and PI affiliation, these data are typically constant for each individual applicant and hence useless for evaluating the dynamics of components across different attempts by the same individual.

To tackle these challenges, we acquired a new data corpus from the NIH that contains abstract information for all R01 proposals submitted after 2008 (both funded and unfunded). Since the abstract data is only available after 2008, and the definition of the unsuccessful group requires five years of inactivity, so there's not enough data for us to measure the unsuccessful group. Nevertheless, the new data does offer a possibility for us to empirically measure the component dynamics for the successful group.

Our hypothesis here is that if we can perform content analysis on abstracts, it may allow us to measure components embedded in each new attempts. To achieve this, we applied a natural lan-

guage processing (NLP) technique to NIH abstracts that estimated MeSH (Medical Subject Headings) terms associated with each proposal. Note that MeSH terms are one of the most commonly used classification codes for biomedical research[87], and this operation is only possible thanks to recent advances in NLP classification, allowing us to automatically and accurately infer MeSH terms from abstract texts. Specifically, we applied NLM Medical Text Indexer, an official protocol developed by US National Library of Medicine Indexing Initiative, to extract a list of MeSH terms given abstract texts.

While the obtained MeSH terms are necessarily imperfect and may not directly correspond to distinct components of the proposal, they capture information that reflects different facets of the proposal, including methods and experimental techniques (e.g., genomic screens), objects of analysis (e.g., breast cancer), research design (e.g., genome-wide association study), and physical phenomena (e.g., estradiol). Here we approximate the creation of new versions by the number of new MeSH terms (terms that did not appear in the previous k submissions), defined as m_n. For example, to measure the dynamics under $k = 1$, we count m_n as the number of Mesh terms that appear in the n-th attempt but not in the $(n-1)$-th attempt. More generally, if we define S_n as the set of all Mesh terms associated with the n-th attempt, our definition can be formulated as $m_n \equiv |S_n - (S_{n-1} \cup \cdots \cup S_{n-k})|$, where $|A|$ denotes the size of a set A (Extended Data Fig. 4a).

According to our model, the time cost comes from creating new versions, traced by the introduction of additional components. Hence, our model suggests that given k, we can use $M_n \equiv \langle m_n \rangle / \langle m_1 \rangle$ to mimic the temporal dynamics of $T_n \equiv \langle t_n \rangle / \langle t_1 \rangle$. More precisely, for the successful

group, we should expect to observe that for large k ($k > k^*$), M_n and T_n should be similar. Yet for small k ($k < k^*$), the two quantities should be quite different. This means that in the same way faster resubmissions (T_n) predict ultimate success, so do shrinking sets of new components (M_n).

We set out to test this new prediction by calculating M_n for different k. We find that the two curves follow different dynamics ($k \leq 3$). Yet the dynamics of M_n and T_n cannot be statistically distinguished for $k > 3$ (from 4 to ∞), both following a power law with $\gamma \sim 0.35$ (Extended Data Fig. 4b). Both findings appear consistent with model predictions. Given that Mesh terms are merely a rough estimate of idea combination in NIH proposals, this degree of agreement seems unexpected.

S5.5 Learning by organizational vs. individual

One aspect of our paper is that here we study learning processes at three different levels, ranging from individual attempts (PIs) to individuals in teams (entrepreneurs) to larger-scale organizations (terrorist groups). The patterns we uncovered reveal that all three levels follow similar statistical patterns governing failure dynamics. But beyond the universality, what differences should we expect across different levels? To answer this question, we contextualize our paper in the literature it builds upon.

The organizational learning literature has identified several factors for the emergence of learning within organizations, with some arguing that individual learning is just one factor in how and why organizations may learn. For example, knowledge gained from past experience can be

embedded within both individual habits and organizational routines (including the idea of transactive memory)[38,88,89]. These suggest that organization-level learning, compared with individual learning, should be characterized by higher learning rate on average. There is also evidence that organizational learning tends to be conservative due to inflexible routines. For example, given versions with the same quality, organizations may have higher probability to reuse rather than create a new one.

Together, these theories predict that of the three domains studied, those closer to organizational learning (such as terrorists) should correspondingly have higher learning rates than those closer to individual learning (such as NIH PIs). We can test this hypothesis by calculating the average learning rate for our samples. We find that our estimations appear consistent with the hypotheses outlined above: For NIH PIs, the average learning rate γ is around ~ 0.361; The learning rate for the entrepreneurship case is higher, around ~ 0.509, and terrorist groups have the highest rate on average ~ 0.640. While these differences could be due to inherent domain-specific differences, they do show consistency with the theories from the organizational learning literature.

S5.6 Scientific achievements and learning rate

Existing literature has also highlighted a series of factors related to why one learns more than others[90], including individual ability, motivation and opportunity to learn. These factors may play a role, manifested in the k parameter. One empirical challenge here is that it remains unclear how to infer k directly from data. But we also realize we can relax the assumption to infer a weak

form of the parameters by inferring γ, and correlate individual characteristics (y) with γ. Indeed, according to our model, if y correlates with k, it may not correlate with γ (if it's in the third phase ($k > k^* + 1$)), but if y correlates with γ, then it must correlate with k.

High achieving scientists are more visible, better recognized, and have access to more resources (Matthew effect in science)[91–93], suggesting that individual prior achievement may manifest in a higher learning rate[90]. Here we test this hypothesis from our data, by collecting additional datasets that allow us to identify individual characteristics and achievements.

Here we extend our analysis of individual characteristics by linking NIH data to the Web of Science citation database. This procedure involved systematic effort in paper matching and author name disambiguation. In this revision, we began from a list of NIH supported publications in PubMed and selected those authored by the same PI. Then we use a WoS-PubMed crosswalk file to locate these papers in WoS and treated them as 'seed' papers. We then expand this initial set to other publications by the same-name author in WoS by tracing the citation relationships and following standard name disambiguation procedures[9,94]: If a paper was contributed by an author with the same name and had citation/reference/co-reference relationships with the initial set, we included it into the PI publication list as well. Implementing this method iteratively allowed us to construct a comprehensive publication list for each PI in our sample.

We then calculated the learning rate γ by regression for all samples with at least three failures before eventual success (i.e., more than two inter-event time periods). Based on the learning literature, we hypothesize that the learning rate may be related to experiences both within and

outside the task of producing an NIH proposal[38]. To this end, we calculated the total number of citations of a PI for all his/her papers published before the first failure (logged), approximating his/her overall 'status' and accomplishments. We find that it is significantly, positively correlated with the learning rate γ ($P < 0.001$, after controlling for the first inter-event time). We further test this correlation by including the number of prior successes and application year as control variables, finding that although past funding success is also correlated with higher learning rate ($P < 0.001$), the relationship between citations and γ remains robust ($P = 0.014$). Although it may seem intuitive that citations and grant applications are correlated, note that the samples studied here include PIs who all failed at least three times before eventually being awarded the grant (i.e., similar success rate). In this respect, it is somewhat unexpected to observe that the speed with which scientists learn from failures can be anticipated by measuring prior achievements. This is consistent with the hypothesis that prior attention and success may provide scientists with greater confidence and resources that allow them to persist and refine rather than abandon and replace the components from an initial, failed proposal.

S5.7 Gender and learning rate

The results presented above offer support for the notion that individual characteristics can indeed affect learning. Here we further anchor other individual characteristics that may distinguish learning. The literature suggests gender could be a potential robust factor that applies across domains, especially in science and entrepreneurship, which are characterized by persistent gender inequality[95–99]. It thus suggests that, if we can separate individuals by gender, we may detect

differential learning rates as well.

To test this relationship in our data, we use a gender detector algorithm to infer gender information from person's first name. We find that gender indeed plays an important role, after we control for all other factors. Our regression analysis shows significant correlation between gender and learning rate. All other factors being equal, the learning rate γ of a male PI in NIH system exceeds that of a female PI by 0.14 ($P = 0.001$). That is, male PIs fail faster than their female colleagues. This difference appears substantial, considering that the average learning rate is centered around 0.35. Note that here we do not essentialize these gender differences, and recognize that they may flow from institutional as well as individual causes, such as a culture that discourages women from persistence and encourages oversensitivity to feedback. Furthermore, such correlations cannot fully account for the discovered signals, as a substantial amount of predictive power by our model remains (AUC higher than 0.7) after we separate our samples by gender.

We further test this relationship on startup dataset, finding a similar gap of 0.10 in the same direction between male and female innovators, though the result is not significant, possibly due to a smaller sample size. These insights are consistent with existing literature on gender inequality in science and entrepreneurship[95–98]. They also highlight the fact that our paper offers a new theoretical framework to systematically study learning, failures and the factors that may influence them.

S6 Prediction task

S6.1 Predicting ultimate success

Our model uncovers time as an early signal for predicting eventual success and failure. This prediction is somewhat unexpected, since individuals through failures are aimed at improving their performance, rather than saving the time, hence we should expect the two groups have similar temporal patterns. To test this, we use D_1 to set up a simple prediction task (Extended Data Fig. 3a). The goal of this task is not to design state-of-art classifiers for predicting success/failure. Rather, to test the predictive power of the uncovered temporal regularity. From this respect, our results offer a lower bound for the predictability of failure dynamics.

To this end, here we first assume a logistic classification model (Model 1) to predict the eventual success following N consecutive failed attempts. For each N, we collect positive samples as individuals succeeded after N failures versus negative samples as individuals dropped out after the same number of consecutive failures. Each sample has a $N-1$-dimensional predictor t_n ($1 \leq n \leq N-1$). The classifier writes as

$$\frac{\log(success)}{1-\log(success)} = \beta_0 + \sum_{n=1}^{N-1} \beta_n \log(t_n) \tag{S107}$$

To evaluate the performance of our predictions, we calculate the AUC curve (average area under the receiver operating characteristic) over 10-fold cross validation for different N.

Our model further predicts that the inter-event time sequence follows a power law decay, suggesting that we can further simplify the prediction model. Indeed, the power-law form means that

we can rescale the $N-1$-dimensional feature $(\log(t_1), \cdots, \log(t_{n-1}))$ into two simple parameters by calculating the slope $-\gamma$ and intercept θ in the log-log plot, i.e.

$$\log(t_n) = -\gamma \log(n) + \theta \tag{S108}$$

Our prediction model 2 is based on the two variables γ and θ to train a simpler classifier for eventual success, following

$$\frac{\log(success)}{1-\log(success)} = \beta_0 + \beta_1 \gamma + \beta_2 \theta \tag{S109}$$

This simplification is expected to be inaccurate since it reduces a feature with high dimensions to data points into a 2-dimension feature. However, to our surprise, we find that a similar prediction accuracy can be achieved as the previous model 1 across different N (Extended Data Fig. 4), accounting for more than 95% of accuracy in terms of additional predictive power (AUC-0.5).

Model 2 also offers additional evidence that is consistent with model predictions. First, the coefficient for γ, β_1 remains positive, demonstrating that escalations in failure dynamics are related to eventual success. Our previous results also suggest that the membership of the two groups are mainly determined by the learning process (different k) rather than the initial advantage (score/time at the first attempt). If so, we would expect the increasing majority of predictive power coming from information encoded in the parameter γ, especially for individuals with large N. To test this hypothesis, we apply an ad-hoc approach for variable importance in logistic regressions on D_1. We calculate the ratio coefficient for normalized input, i.e.

$$R = \frac{|\beta_1|[var(\gamma)]^{1/2}}{|\beta_2|[var(\theta)]^{1/2}} \tag{S110}$$

R measures the ratio of coefficients once the two variables are normalized to have identical variance. We find that R increases systematically with n, suggesting that the variable γ contributes increasingly to the predictive power as one fails more, supporting the hypothesis that the dynamic process itself, rather than the starting point, has a xlarger impact on the eventual outcome following failures.

S6.2 Testing power law model

Despite long history in using power law forms to model learning curves, the literature has also suggested other functional forms[50]. One of the frequently used alternatives is exponential function, predicting

$$t_n \sim ab^{-n} \tag{S111}$$

Indeed, recent studies have also suggested that the observed power law could be an artifact by average different samples, and may be better characterized by an exponential decay[100].

The difficulty in testing different hypotheses in our datasets comes from the small sample sizes: in contrast to industrial production or simple individual tasks, it is hard to observe empirically a large number of failures for a given individual. Hence directly comparing the fitting of different models would suffer from overfitting issues. To this end, here for each individual sample, we take all but the last inter-event time for model fitting, comparing model predictions for the last inter-event time. This out-of-sample testing technique helps to alleviate the issue of overfitting.

Using this method we compare the performance of power law, exponential and linear models

in characterizing t_n for each individual, measuring their prediction error (Extended Data Fig. 2). We find that across the three datasets, the power law model yields the smallest error in most cases.

S7 Robustness checks

S7.1 Definition of success and failure

We vary our definition of success and failure across different datasets. For D_1 we remove all renewal/resubmission successes and only focus on new applications, finding our conclusions are not affected by resubmissions (Extended Data Fig. 6).

For D_2 we vary the definition of success for a startup. Previously we have considered IPO and high-value M&A as success. Similar with hit papers defined in science of science, we define high-value M&As as those with transaction value ranking top 1% among all transactions in the same year. We vary this definition to top 5% transactions or exclude all M&As (Extended Data Fig. 7), finding our conclusions still hold. One problem with our definition for success is that it does not include ventures that could already be considered successful despite not having had an IPO or being acquired. To this end, we collected a list of unicorn companies, defined as privately held startup companies valued at over 1 billion, from CB Insights website, yielding 121 companies in our sample, which can be linked through company names. Overall we find such cases are relatively rare. We also test our conclusions by removing these cases from the unsuccessful group, or re-defining them as successful attempts. In both cases we find our results remain the same (Extended Data Fig. 7).

For D_3 we tried variants by expanding unsuccessful groups to all samples or restricting successful groups to human-target samples only (Extended Data Fig. 8). Both variants yield similar results. We also vary the threshold in our data, changing our definition of successful group as organizations that killed at least 5, 10 and 100 people in a terrorist attack (Extended Data Fig. 8). We find the patterns hold the same.

S7.2 Threshold for being inactive in the system

The definition of unsuccessful group depends on the threshold for inactive in the system. In main text we set up the threshold as 5 years, i.e. if one does not appear in the system for the last 5 years, we consider such cases as drop-out samples. To test the effect of this threshold, here we repeat our main results for 3 years and 7 years (Extended Data Fig. 5), respectively. We find all our results are robust as we tune this criterion.

S7.3 Effect of overall success rate

It is also important to keep in mind that the success rate may go up and down over time. Here we control for the overall success rates across our three datasets and test its potential impact on our results. More specifically, we renormalize our empirical data by weighing different samples by success rate to ensure that each year has effectively the same success rate. For example, for samples from the successful group ending in year y, we count the total number of successes and failures in that year, defined as $S(y)$ and $F(y)$. We then calculate the weight of each sample as $w \equiv (F+S)/S$, i.e., the inverse of the overall success rate. This is equivalent to resampling within

all successful cases, with the sampling probability proportional to the inverse of the success rate. To this end, the weighted sum of each year's success should be $S/w = (F + S)$, or proportional to the total number of samples in the same year.

We then repeat all of our main measurements using the renormalized samples. As shown in Extended Data Fig. 9, all of the main predictions made in our paper hold the same. This suggests that even though intelligence agencies may improve their ongoing detection of terror attacks, congress may decrease (or increase) its annual budget for science, and economic cycles may increase or reduce the companies with successful exits, these changes are smooth in time, and do not affect the conclusions drawn in the paper.

S7.4 Comparing first failures versus halfway/penultimate failures

Figure 3 showed performance divergence patterns in two groups using first and second failures. Here we also compares the first failures versus halfway or penultimate failures, recovering the same patterns (Extended Data Fig. 9).

S7.5 Other checks

For D_1 we further confirmed that only focusing on failures before the first success yield similar results. Indeed, as we plot T_n for samples with and without prior success, we find the dynamical patterns remain the same. Lastly, we check the threshold of discussion score, considering original percentile score higher than 55, rather than 50, as undiscussed. All these variants show results

consistent with Fig. 3 (Extended Data Fig. 6).

For D_3, 5.7% of the records contain vague numbers of killed people despite the evidence of fatalities, which we discarded in our original analysis. We also consider these events as successful attempts and repeated our results, finding the patterns remain the same (Extended Data Fig. 8).

Reference

1. Fortunato, S. *et al.* Science of science. *Science* **359**, eaao0185 (2018).

2. Azoulay, P. *et al.* Toward a more scientific science. *Science* **361**, 1194–1197 (2018).

3. Harford, T. *Adapt: Why success always starts with failure* (Farrar, Straus and Giroux, 2011).

4. Fleming, L. Recombinant uncertainty in technological search. *Management science* **47**, 117–132 (2001).

5. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).

6. Jones, B. F. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies* **76**, 283–317 (2009).

7. Petersen, A. M., Riccaboni, M., Stanley, H. E. & Pammolli, F. Persistence and uncertainty in the academic career. *Proceedings of the National Academy of Sciences* **109**, 5213–5218 (2012).

8. Clauset, A., Arbesman, S. & Larremore, D. B. Systematic inequality and hierarchy in faculty hiring networks. *Science advances* **1**, e1400005 (2015).

9. Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).

10. Liu, L. *et al.* Hot streaks in artistic, cultural, and scientific careers. *Nature* **559**, 396 (2018).

11. Hu, Y., Havlin, S. & Makse, H. A. Conditions for viral influence spreading through multiplex correlated social networks. *Physical Review X* **4**, 021031 (2014).

12. Jara-Figueroa, C., Jun, B., Glaeser, E. L. & Hidalgo, C. A. The role of industry-specific, occupation-specific, and location-specific knowledge in the growth and survival of new firms. *Proceedings of the National Academy of Sciences* **115**, 12646–12653 (2018).

13. Hidalgo, C. *Why information grows: The evolution of order, from atoms to economies* (Basic Books, 2015).

14. Barabasi, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211 (2005).

15. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).

16. Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462–465 (2006).

17. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Reviews of modern physics* **81**, 591 (2009).

18. Malmgren, R. D., Stouffer, D. B., Campanharo, A. S. & Amaral, L. A. N. On universality in human correspondence activity. *Science* **325**, 1696–1700 (2009).

19. Argote, L. & Epple, D. Learning curves in manufacturing. *Science* **247**, 920 (1990).

20. Sitkin, S. B. Learning through failure: the strategy of small losses. *Research in organizational behavior* **14**, 231–266 (1992).

21. Yelle, L. E. The learning curve: Historical review and comprehensive survey. *Decision sciences* **10**, 302–328 (1979).

22. Dutton, J. M. & Thomas, A. Treating progress functions as a managerial opportunity. *Academy of management review* **9**, 235–247 (1984).

23. Shrager, J., Hogg, T. & Huberman, B. A. A graph-dynamic model of the power law of practice and the problem-solving fan-effect. *Science* **242**, 414–416 (1988).

24. Levitt, B. & March, J. G. Organizational learning. *Annual review of sociology* **14**, 319–338 (1988).

25. Huber, G. P. Organizational learning: The contributing processes and the literatures. *Organization science* **2**, 88–115 (1991).

26. Edmondson, A. Psychological safety and learning behavior in work teams. *Administrative science quarterly* **44**, 350–383 (1999).

27. Gross, C. P., Anderson, G. F. & Powe, N. R. The relation between funding by the national institutes of health and the burden of disease. *New England Journal of Medicine* **340**, 1881–1887 (1999).

28. Ginther, D. K. *et al.* Race, ethnicity, and nih research awards. *Science* **333**, 1015–1019 (2011).

29. Li, D. & Agha, L. Big names or big ideas: Do peer-review panels select the best science proposals? *Science* **348**, 434–438 (2015).

30. Kaplan, S. N. & Lerner, J. Venture capital data: Opportunities and challenges. In *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges* (University of Chicago Press, 2016).

31. Eggers, J. & Song, L. Dealing with failure: Serial entrepreneurs and the costs of changing industries between ventures. *Academy of Management Journal* **58**, 1785–1803 (2015).

32. Gompers, P., Kovner, A., Lerner, J. & Scharfstein, D. Performance persistence in entrepreneurship. *Journal of Financial Economics* **96**, 18–32 (2010).

33. National Consortium for the Study of Terrorism and Responses to Terrorism (START). *Global Terrorism Database [Data file]* (2018).

34. Clauset, A. & Gleditsch, K. S. The developmental dynamics of terrorist organizations. *PloS one* **7**, e48633 (2012).

35. Johnson, N. *et al.* Pattern in escalations in insurgent and terrorist activity. *Science* **333**, 81–84 (2011).

36. Durrett, R. *Probability: theory and examples* (Cambridge university press, 2010).

37. Bass, R. F. *Stochastic processes*, vol. 33 (Cambridge University Press, 2011).

38. Argote, L. *Organizational learning: Creating, retaining and transferring knowledge* (Springer Science & Business Media, 2012).

39. Dahlin, K. B., Chuang, Y.-T. & Roulet, T. J. Opportunity, motivation, and ability to learn from failures and errors: Review, synthesis, and ways to move forward. *Academy of Management Annals* **12**, 252–277 (2018).

40. Levy, F. K. Adaptation in the production process. *Management Science* **11**, B–136 (1965).

41. Newell, A. & Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition* **1**, 1–55 (1981).

42. Anderson, J. R. Acquisition of cognitive skill. *Psychological review* **89**, 369 (1982).

43. Muth, J. F. Search theory and the manufacturing progress function. *Management Science* **32**, 948–962 (1986).

44. McNerney, J., Farmer, J. D., Redner, S. & Trancik, J. E. Role of design complexity in technology improvement. *Proceedings of the National Academy of Sciences* **108**, 9008–9013 (2011).

45. Wright, T. P. Factors affecting the cost of airplanes. *Journal of the aeronautical sciences* **3**, 122–128 (1936).

46. Snoddy, G. S. Learning and stability: a psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology* **10**, 1 (1926).

47. Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009).

48. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

49. Bettencourt, L. M., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences* **104**, 7301–7306 (2007).

50. Ritter, F. E. & Schooler, L. J. The learning curve. *International encyclopedia of the social and behavioral sciences* **13**, 8602–8605 (2001).

51. Stephan, P. E. *How economics shapes science*, vol. 1 (Harvard University Press Cambridge, MA, 2012).

52. Paik, Y. Serial entrepreneurs and venture survival: Evidence from us venture-capital-financed semiconductor firms. *Strategic Entrepreneurship Journal* **8**, 254–268 (2014).

53. Walsh, G. S., Cunningham, J. A. *et al.* Business failure and entrepreneurship: emergence, evolution and future research. *Foundations and Trends® in Entrepreneurship* **12**, 163–285 (2016).

54. McGrath, R. G. Falling forward: Real options reasoning and entrepreneurial failure. *Academy of Management review* **24**, 13–30 (1999).

55. Edmondson, A. C. Strategies for learning from failure. *Harvard business review* **89**, 48–55 (2011).

56. Shepherd, D. A. Learning from business failure: Propositions of grief recovery for the self-employed. *Academy of management Review* **28**, 318–328 (2003).

57. Denrell, J. Vicarious learning, undersampling of failure, and the myths of management. *Organization Science* **14**, 227–243 (2003).

58. Kim, J.-Y. & Miner, A. S. Vicarious learning from the failures and near-failures of others: Evidence from the us commercial banking industry. *Academy of Management Journal* **50**, 687–714 (2007).

59. Edmondson, A. C. Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *The Journal of Applied Behavioral Science* **40**, 66–90 (2004).

60. Madsen, P. M. These lives will not be lost in vain: Organizational learning from disaster in us coal mining. *Organization Science* **20**, 861–875 (2009).

61. Baum, J. A. & Dahlin, K. B. Aspiration performance and railroads' patterns of learning from train wrecks and crashes. *Organization Science* **18**, 368–385 (2007).

62. Haunschild, P. R. & Sullivan, B. N. Learning from complexity: Effects of prior accidents and incidents on airlines' learning. *Administrative science quarterly* **47**, 609–643 (2002).

63. Madsen, P. M. & Desai, V. Failing to learn? The effects of failure and success on organizational learning in the global orbital launch vehicle industry. *Academy of Management Journal* **53**, 451–476 (2010).

64. Sahal, D. A theory of progress functions. *AIIE Transactions* **11**, 23–29 (1979).

65. Roberts, P. A theory of the learning process. *Journal of the Operational Research Society* **34**, 71–79 (1983).

66. Kluger, A. N. & DeNisi, A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin* **119**, 254 (1996).

67. Asher, H. *Cost-quantity relationships in the airframe industry*. Ph.D. thesis, The Ohio State University (1956).

68. Crossman, E. A theory of the acqusition of speed-skill. *Ergonomics* **2**, 153–166 (1959).

69. Kauffman, S. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology* **128**, 11–45 (1987).

70. Levinthal, D. A. Adaptation on rugged landscapes. *Management science* **43**, 934–950 (1997).

71. Denrell, J. & March, J. G. Adaptation as information restriction: The hot stove effect. *Organization Science* **12**, 523–538 (2001).

72. Laird, J., Rosenbloom, P. & Newell, A. *Universal subgoaling and chunking: The automatic generation and learning of goal hierarchies*, vol. 11 (Springer Science & Business Media, 2012).

73. Loreto, V., Servedio, V. D., Strogatz, S. H. & Tria, F. Dynamics on expanding spaces: modeling the emergence of novelties. In *Creativity and universality in language*, 59–83 (Springer, 2016).

74. Heaps, H. S. *Information retrieval, computational and theoretical aspects* (Academic Press, 1978).

75. Simon, H. A. On a class of skew distribution functions. *Biometrika* **42**, 425–440 (1955).

76. Tria, F., Loreto, V., Servedio, V. D. P. & Strogatz, S. H. The dynamics of correlated novelties. *Scientific reports* **4**, 5890 (2014).

77. Iacopini, I., Milojević, S. & Latora, V. Network dynamics of innovation processes. *Physical review letters* **120**, 048301 (2018).

78. Argote, L., Beckman, S. L. & Epple, D. The persistence and transfer of learning in industrial settings. *Management science* **36**, 140–154 (1990).

79. Kuhn, T. S. *The structure of scientific revolutions* (University of Chicago press, 2012).

80. Merton, R. K. Singletons and multiples in scientific discovery: A chapter in the sociology of science. *Proceedings of the American Philosophical Society* **105**, 470–486 (1961).

81. Holan, P. M. d. & Phillips, N. Remembrance of things past? the dynamics of organizational forgetting. *Management science* **50**, 1603–1613 (2004).

82. Zaburdaev, V., Denisov, S. & Klafter, J. Lévy walks. *Reviews of Modern Physics* **87**, 483 (2015).

83. Mandelbrot, B. B. *The fractal geometry of nature*, vol. 1 (WH freeman New York, 1982).

84. Taleb, N. N. *The black swan: The impact of the highly improbable*, vol. 2 (Random house, 2007).

85. Jacob, B. A. & Lefgren, L. The impact of research grant funding on scientific productivity. *Journal of public economics* **95**, 1168–1177 (2011).

86. Li, D., Azoulay, P. & Sampat, B. N. The applied value of public investments in biomedical research. *Science* **356**, 78–81 (2017).

87. Lipscomb, C. E. Medical subject headings (mesh). *Bulletin of the Medical Library Association* **88**, 265 (2000).

88. Wegner, D. M. Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior*, 185–208 (Springer, 1987).

89. Liang, D. W., Moreland, R. & Argote, L. Group versus individual training and group performance: The mediating role of transactive memory. *Personality and social psychology bulletin* **21**, 384–393 (1995).

90. Argote, L., McEvily, B. & Reagans, R. Managing knowledge in organizations: An integrative framework and review of emerging themes. *Management science* **49**, 571–582 (2003).

91. Merton, R. K. *et al.* The matthew effect in science. *Science* **159**, 56–63 (1968).

92. Petersen, A. M., Jung, W.-S., Yang, J.-S. & Stanley, H. E. Quantitative and empirical demonstration of the matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences* **108**, 18–23 (2011).

93. Azoulay, P., Stuart, T. & Wang, Y. Matthew: Effect or fable? *Management Science* **60**, 92–109 (2013).

94. Huang, J., Ertekin, S. & Giles, C. L. Efficient name disambiguation for large-scale databases. In *European conference on principles of data mining and knowledge discovery*, 536–544 (Springer, 2006).

95. Shen, H. Inequality quantified: Mind the gender gap. *Nature News* **495**, 22 (2013).

96. Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. Bibliometrics: Global gender disparities in science. *Nature News* **504**, 211 (2013).

97. Yang, T. & Aldrich, H. E. Who's the boss? explaining gender inequality in entrepreneurial teams. *American Sociological Review* **79**, 303–327 (2014).

98. Argote, L., Insko, C. A., Yovetich, N. & Romero, A. A. Group learning curves: The effects of turnover and task complexity on group performance. *Journal of Applied Social Psychology* **25**, 512–529 (1995).

99. Bailey, C. D. Forgetting and the learning curve: A laboratory study. *Management science* **35**, 340–352 (1989).

100. Heathcote, A., Brown, S. & Mewhort, D. The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review* **7**, 185–207 (2000).

Category	Reference	Time	Performance	Power law	Coexistence
Adaptation	Crossman[68]	✓	✗	✗	✗
	Kauffman & Levin[69]	✗	✓	✗	✗
	Denrell & March[71]	✗	✓	✗	✗
Search	Roberts[65]	✗	✓	✓	✗
	Muth[43]	✗	✓	✓	✗
	Mcnerney *et al*[44]	✗	✓	✓	✗
Individual learning	Newell *et al*[41]	✓	✗	✓	✗
	Anderson[42]	✓	✗	✓	✗
Urn	Simon[75]	✗	✗	✓	✗
	Tria *et al*[76]	✓	✗	✓	✓
	Iacopini *et al*[77]	✓	✗	✓	✓
Other	Levy[40]	✗	✓	✗	✗
	Shrager *et al*[23]	✗	✓	✗	✗
	Sahal[64]	✓	✗	✓	✗
	Johnson *et al*[35]	✓	✗	✓	✗
	Clauset & Gleditsch[34]	✓	✗	✓	✗

Table S2: **Literature review of relevant models**. We test whether the models listed can predict (1) Time: time reduction; (2) Performance: performance improvement (or reduction in any cost other than time); (3) Power law: analytical form of power law scaling; (4) Coexistence: coexistence of two groups with different dynamics (success and unsuccessful groups in this paper). We find that none of the existing models can predict all the observations in our paper.

	Exponential	Lognormal	Power law	Truncated power law
NIH grants	0.0	0.154	7.01×10^{-4}	2.33×10^{-159}
Startups	7.01×10^{-5}	0.723	2.48×10^{-6}	0.953
Terrorist attacks	0.0	0.822	0.566	0.221

Table S3: **Comparing different functional forms of distributions with Weibull distributions.** All P-values terms denote the degree that Weibull distribution is compared over the other in loglikelihood ratio tests ($n = 20427,\ 667,\ 233$). Among all alternatives, only lognormal models show comparable fitting performance. Yet lognormal model uses two free parameters while the shape parameter of Weibull distribution is constrained by the scaling identity (Eq. 4 in main text).

	NIH grants	Startups	Terrorist attacks
γ	0.361 ± 0.010	0.509 ± 0.036	0.640 ± 0.153
β	0.666 ± 0.017	0.566 ± 0.086	0.129 ± 0.033
P	0.176	0.421	0.141

Table S4: **Parameter estimates (mean$_{\pm \text{s.e.m}}$)**. γ corresponds to the temporal scaling exponent uncovered in Eq. (2) in the main text (sample size is the same as in Fig. 3 d-f) and β is the shape parameter of the Weibull distribution (s.e.m. estimated from bootstrapping over 100 simulations), characterizing the length distribution of failure streaks. Two-sided t-tests indicate that none of the three datasets can reject the validity of the scaling identity $\beta + \gamma = 1$.