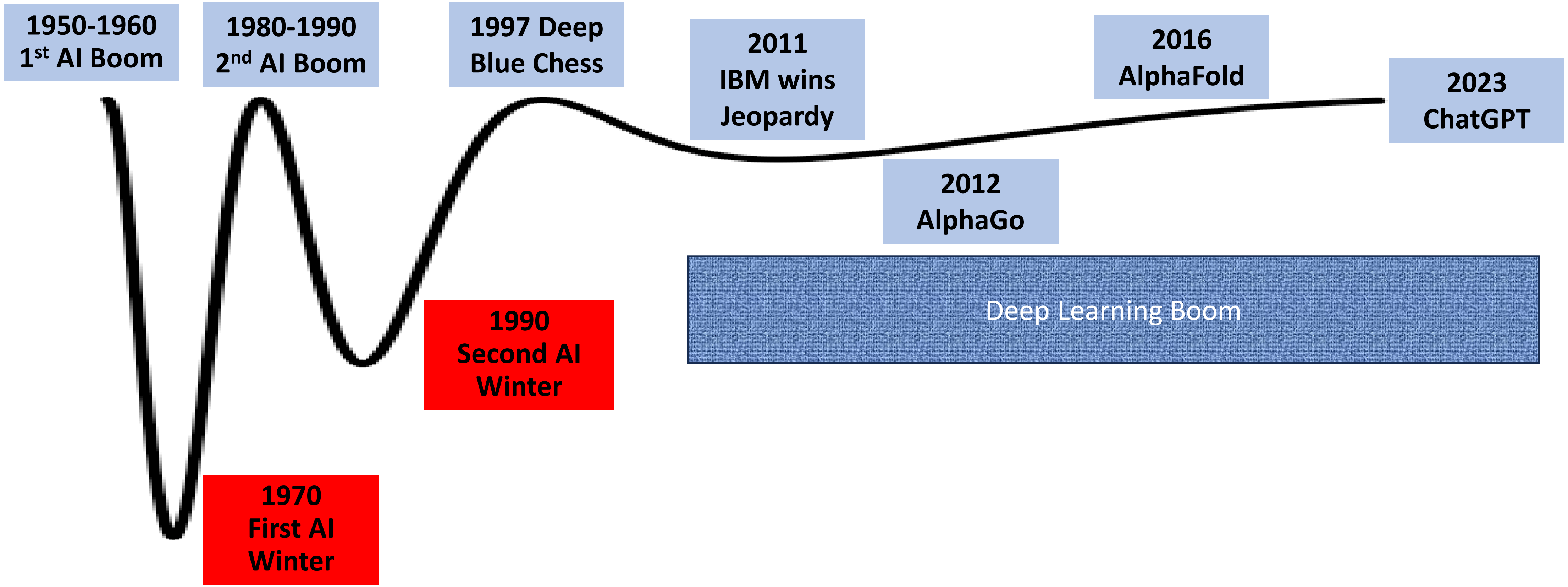
PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Will an AI win a chemistry Nobel Prize and replace us?

Simon Coles & Jeremy Frey
School of Chemistry
University of Southampton

@colessj 0000-0001-8414-9272



1950-1960 1st AI Boom

1980-1990 2nd AI Boom

1997 Deep Blue Chess

2011 IBM wins Jeopardy

2012 AlphaGo

2016 AlphaFold

2023 ChatGPT

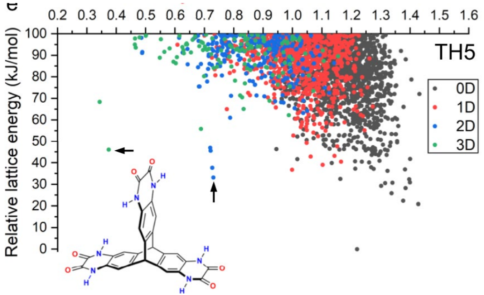
1970 First AI Winter

1990 Second AI Winter

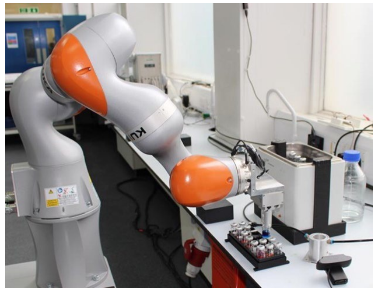
Deep Learning Boom

Pre-Digital Computers → Computers are Useful → Computers are Essential → Computers Take Over

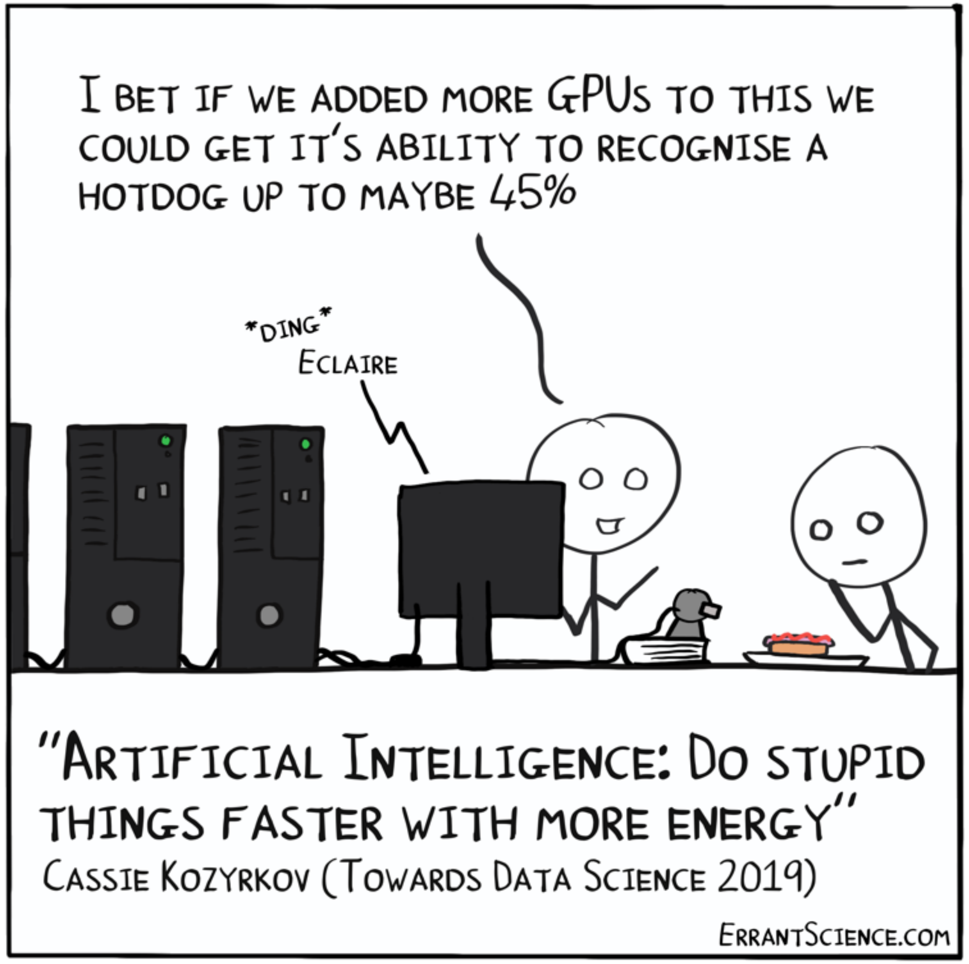
By Francois-Xavier Beckers, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=32926674



https://doi.org/10.1038/s41467-021-21091-w

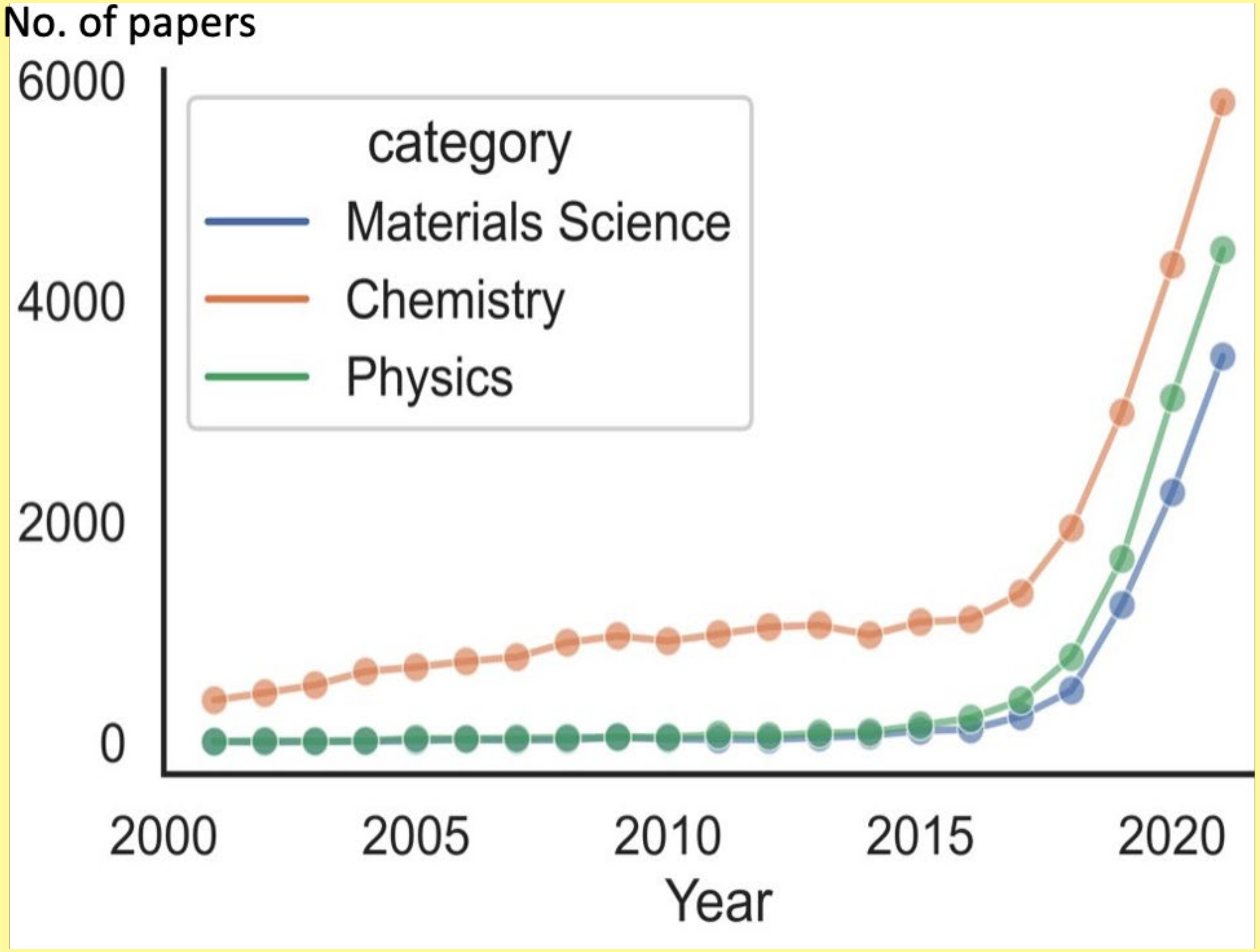


Augmented Chemical Intelligence?

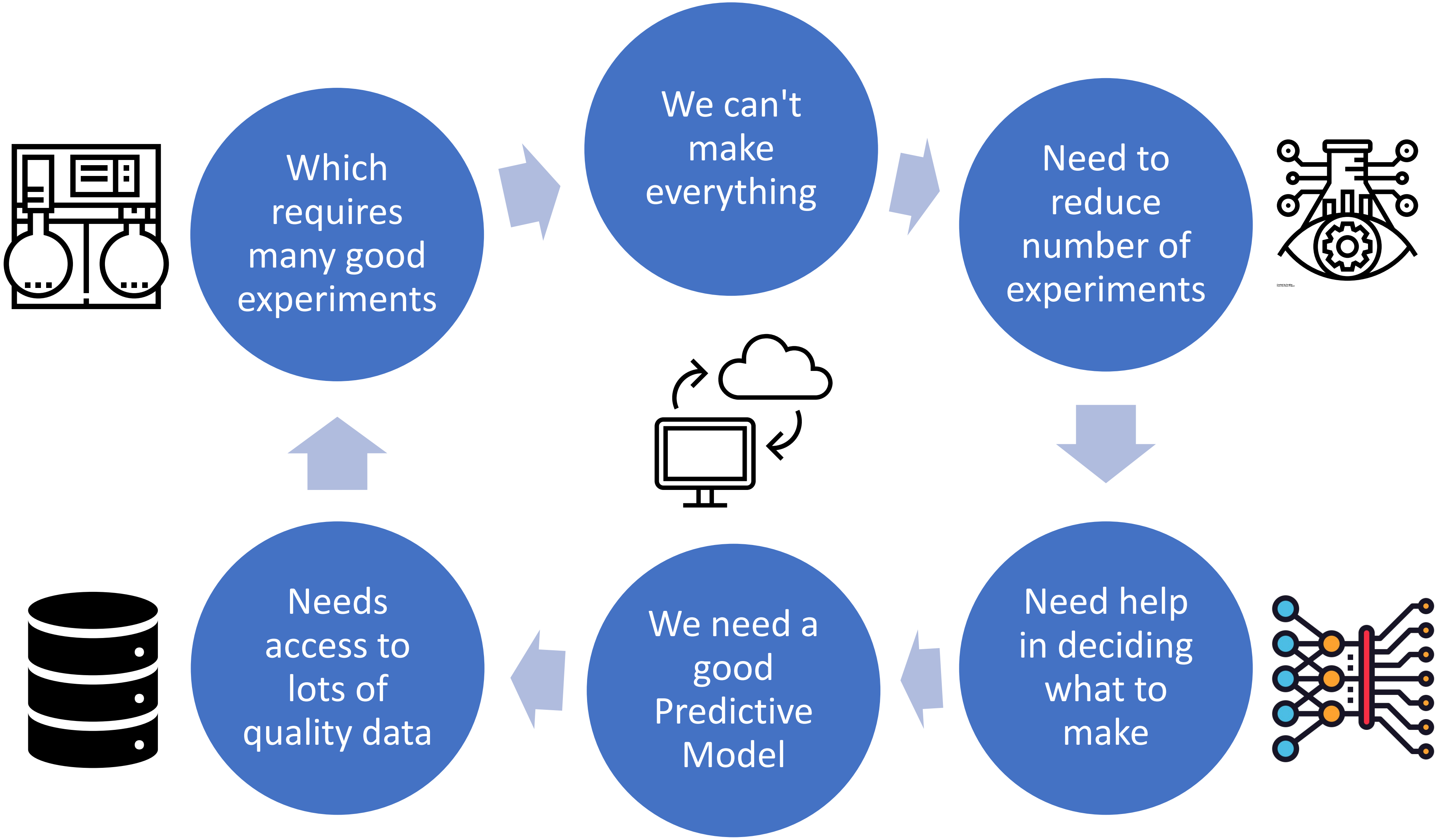

I BET IF WE ADDED MORE GPUs TO THIS WE COULD GET IT'S ABILITY TO RECOGNISE A HOTDOG UP TO MAYBE 45%
DING
ECLAIRE

"ARTIFICIAL INTELLIGENCE: DO STUPID THINGS FASTER WITH MORE ENERGY"
CASSIE KOZYRKOV (TOWARDS DATA SCIENCE 2019)

ERRANTSCIENCE.COM



Need to be more insightful, creative, efficient, environmentally conscious…



Which requires many good experiments

We can't make everything

Need to reduce number of experiments

Need help in deciding what to make

We need a good Predictive Model

Needs access to lots of quality data

From QSAR to Machine Learning

Relevant Data	Equation / Model	Name
Known	Known	Theory
Unknown	Known	Unproved theory
Known	Unknown	Statistical Modelling (QSAR)
Unknown	Unknown	Machine Learning

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

CrossMark
click for updates

Cite this: *Chem. Sci.*, 2016, 7, 1600

QSAR analysis of substituent effects on tambjamine anion transporters†‡

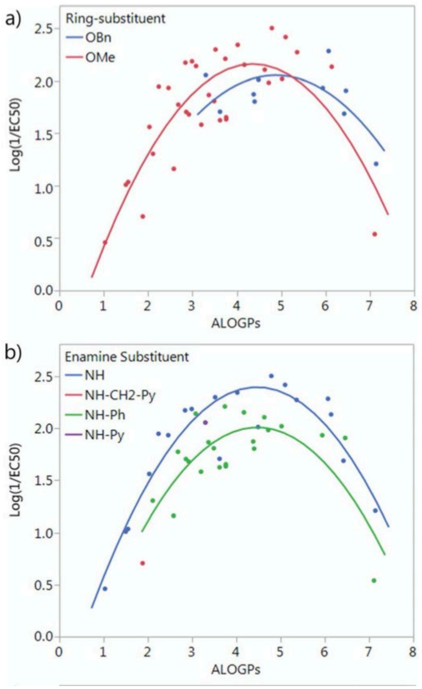
Nicola J. Knight,[a] Elsa Hernando,[b] Cally J. E. Haynes,§[a] Nathalie Busschaert,¶[a] Harriet J. Clarke,[a] Koji Takimoto,[c] María García-Valverde,[b] Jeremy G. Frey,*[a] Roberto Quesada*[b] and Philip A. Gale*[a]

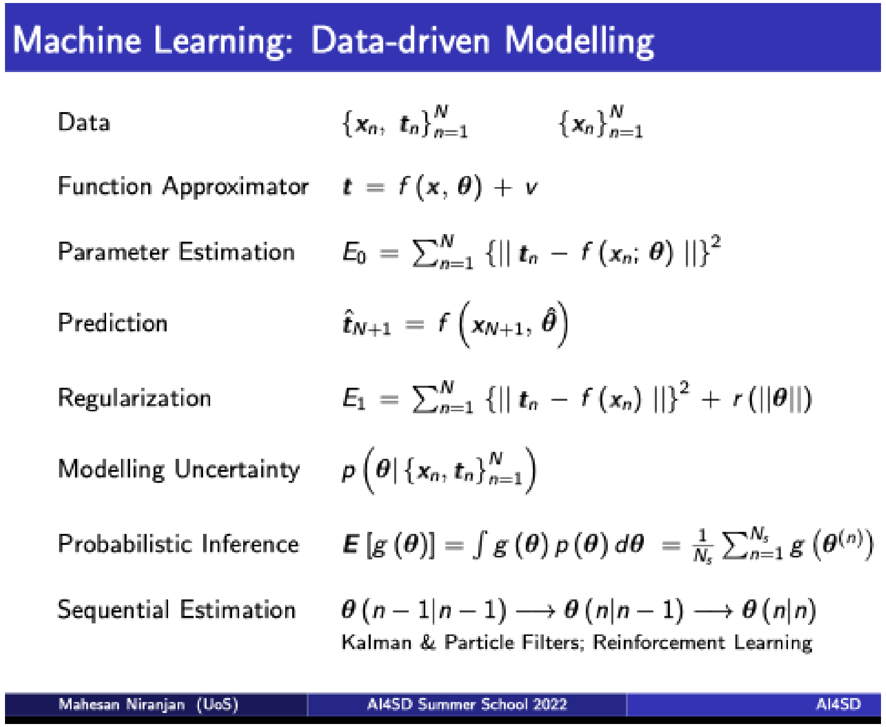
The transmembrane anion transport activity of 43 synthetic molecules based on the structure of marine alkaloid tambjamine were assessed in model phospholipid (POPC) liposomes. The anionophoric activity of these molecules showed a parabolic dependence with lipophilicity, with an optimum range for transport efficiency. Using a quantitative structure–transport activity (QSAR) approach it was possible to rationalize these results and to quantify the contribution of lipophilicity to the transport activity of these derivatives. While the optimal value of log *P* and the curvature of the parabolic dependence is a property of the membrane (and so similar for the different series of substituents) we found that for relatively simple substituents in certain locations on the tambjamine core, hydrophobic interactions clearly dominate, but for others, more specific interactions are present that change the position of the membrane hydrophobicity parabolic envelope.

Received 16th October 2015
Accepted 17th November 2015

DOI: 10.1039/c5sc03932k

www.rsc.org/chemicalscience





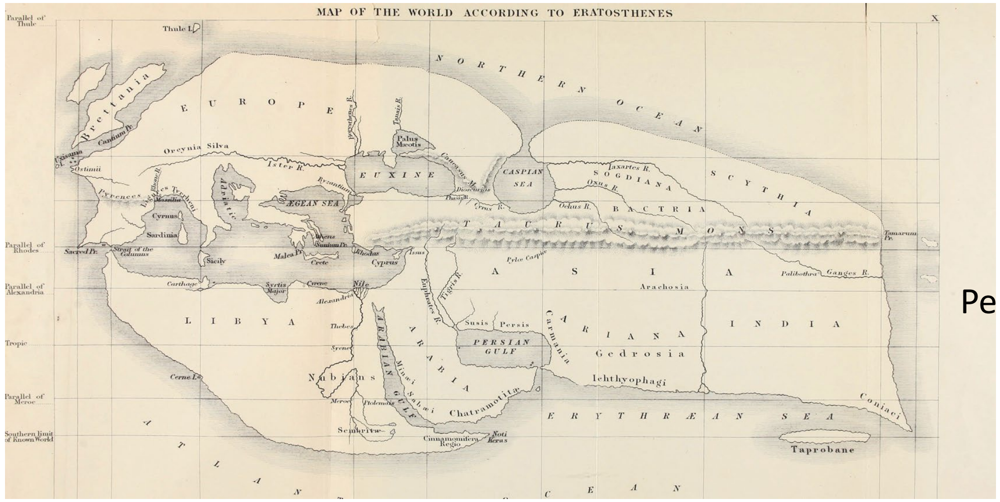
Machine Learning: Data-driven Modelling

Data	$\{x_n,\ t_n\}_{n=1}^N \qquad \{x_n\}_{n=1}^N$
Function Approximator	$t = f(x, \theta) + v$
Parameter Estimation	$E_0 = \sum_{n=1}^N \{\|\| t_n - f(x_n;\ \theta)\ \|\|\}^2$
Prediction	$\hat{t}_{N+1} = f\left(x_{N+1}, \hat{\theta}\right)$
Regularization	$E_1 = \sum_{n=1}^N \{\|\| t_n - f(x_n)\ \|\|\}^2 + r(\|\|\theta\|\|)$
Modelling Uncertainty	$p\left(\theta \| \{x_n, t_n\}_{n=1}^N\right)$
Probabilistic Inference	$E[g(\theta)] = \int g(\theta)\, p(\theta)\, d\theta = \frac{1}{N_s} \sum_{n=1}^{N_s} g\left(\theta^{(n)}\right)$
Sequential Estimation	$\theta(n-1\|n-1) \longrightarrow \theta(n\|n-1) \longrightarrow \theta(n\|n)$ Kalman & Particle Filters; Reinforcement Learning

Mahesan Niranjan (UoS) AI4SD Summer School 2022 AI4SD

The need to map chemical space?

The First Law of Geography:
”Everything is related to everything else, but near things are more related than distant things”.



- Not just big, high dimensional (need to consider Chemical Space-Time)
- What do we mean by related?
- What do we mean by near?
- In high dimensions ‘distance’ is less well defined

Structure

Characterisation

Performance

Property

Processing

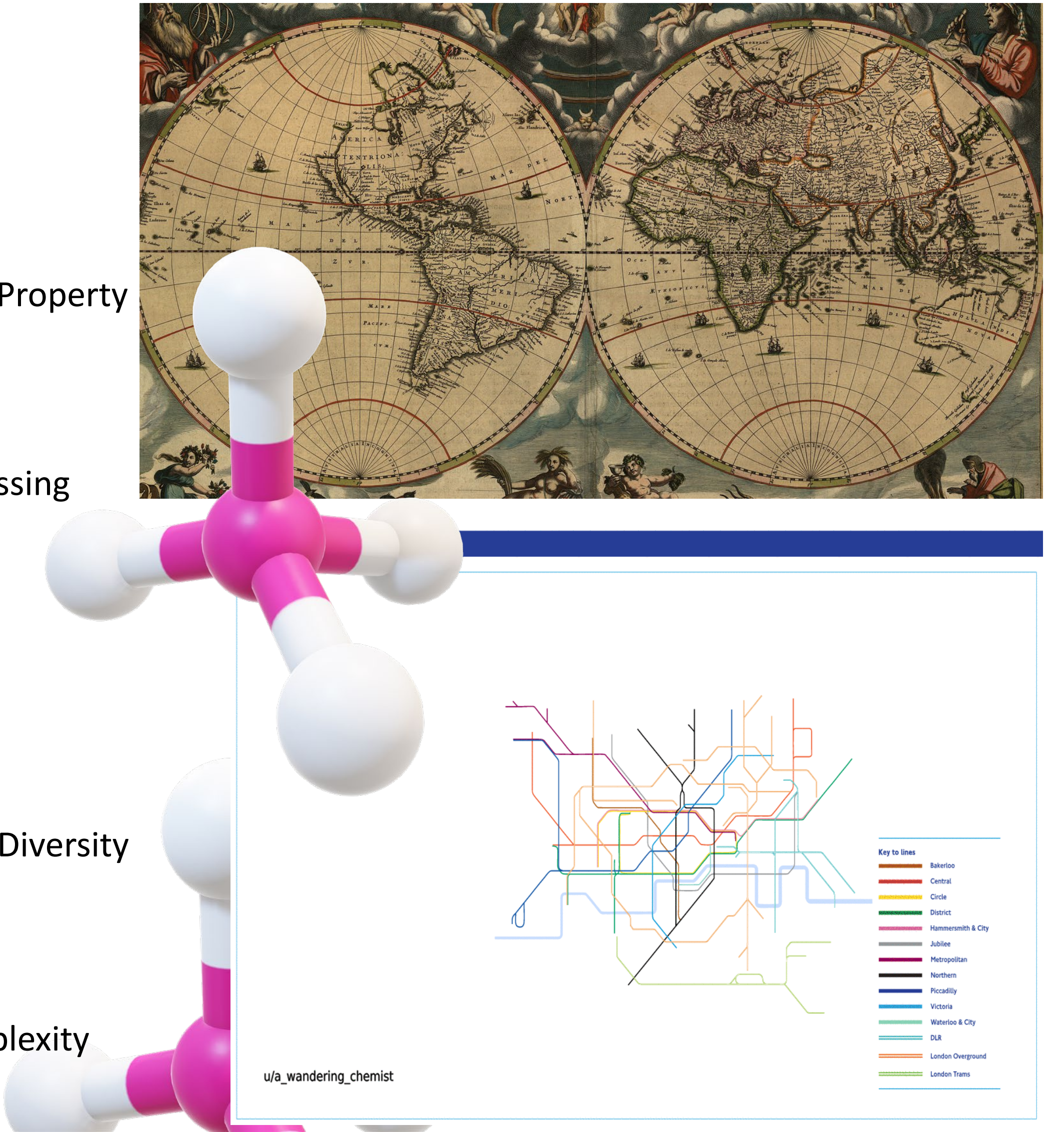
Dimension

Data

Diversity

Shape

Complexity



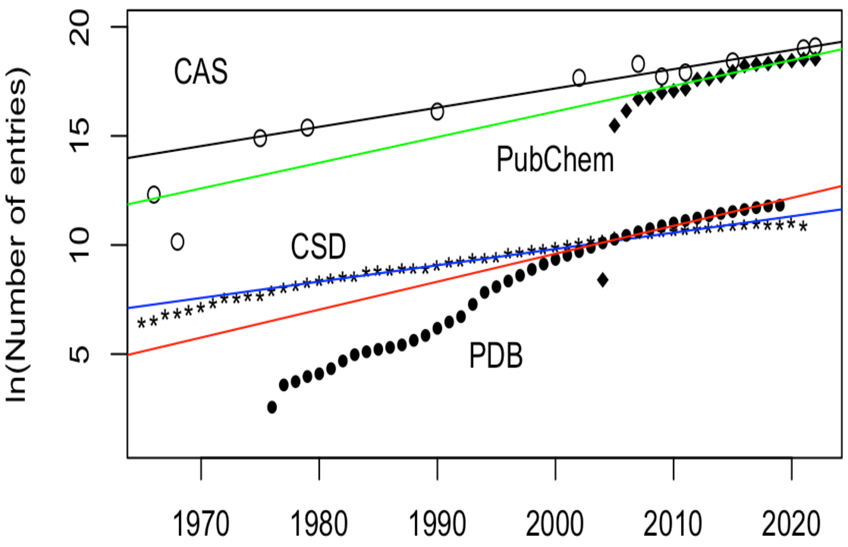
Data, Data everywhere - but not enough to model?

Some collections of trusted, curated data

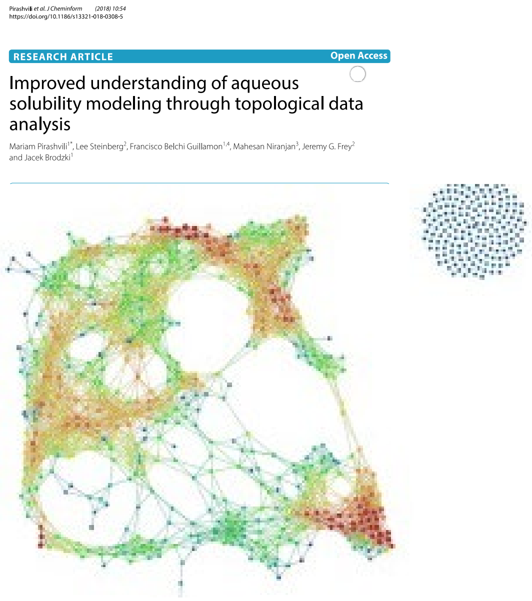
- Rapid increase in the number of *available* crystal structures
- By no means capturing all structures (being) determined

By far, the majority of data is *unstructured*

- Residing in articles, theses, patents (not particularly *accessible*)
- In difficult to process formats
- Never left the lab book… **UN** Findable Accessible Interoperable Reusable



- Beginning to move towards repositories, metadata standards, descriptors
- Need AI suitable ways to structure data e.g. Topology (shape of data)



Pirashvili *et al. J Cheminform* *(2018) 10:54*
https://doi.org/10.1186/s13321-018-0308-5

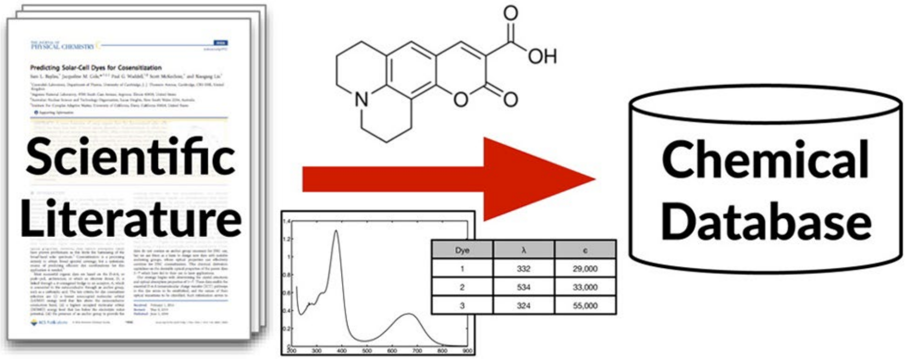
RESEARCH ARTICLE Open Access

Improved understanding of aqueous solubility modeling through topological data analysis

Mariam Pirashvili[1*], Lee Steinberg[2], Francisco Belchi Guillamon[1,4], Mahesan Niranjan[3], Jeremy G. Frey[2] and Jacek Brodzki[1]

Culture change required
e.g. compiling non-curated/unstructured data

- Generally only published, successful outcome, data is made available
- Extracting data is time consuming…

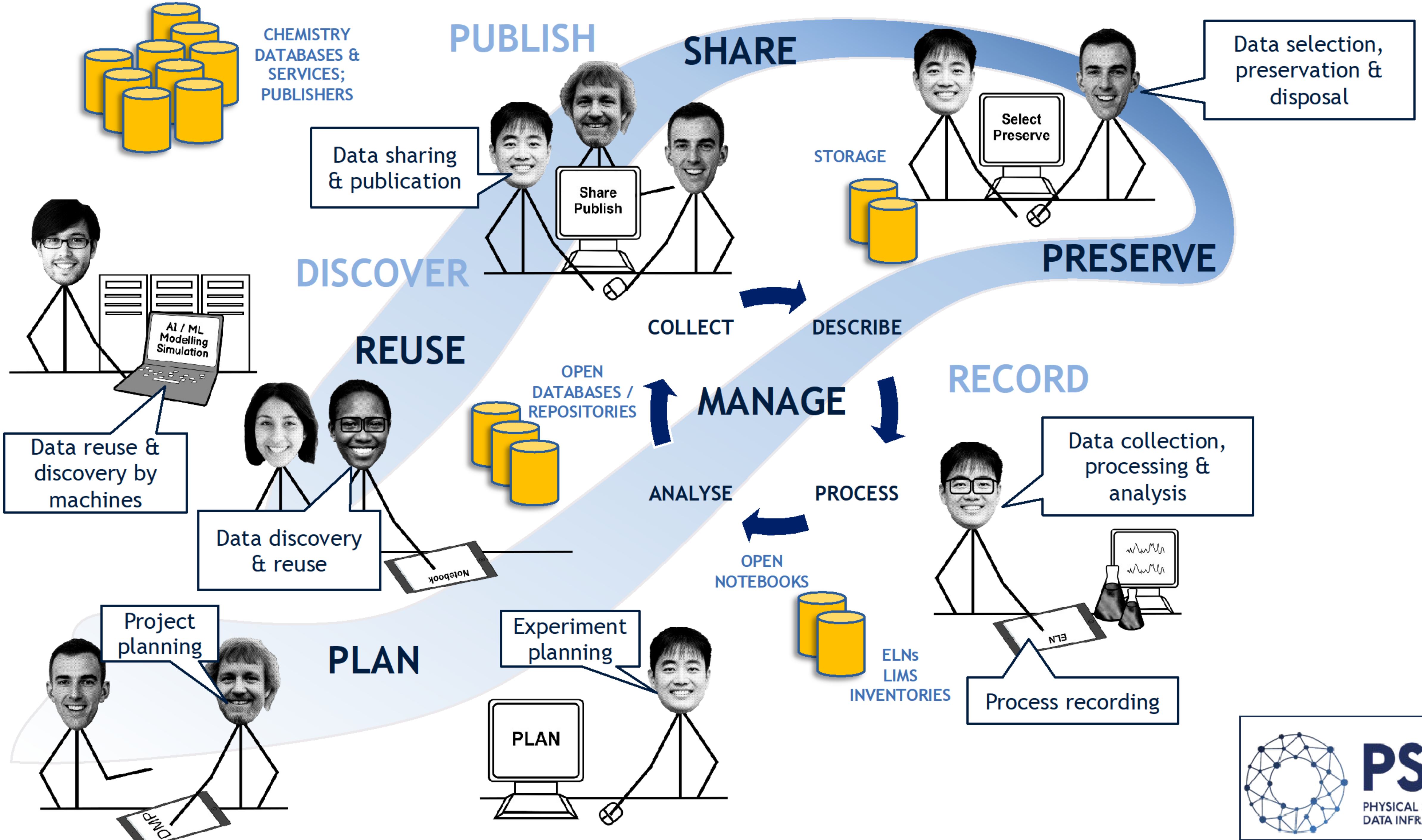


Dye	λ	ε
1	332	29,000
2	534	33,000
3	324	55,000

https://doi.org/10.1021/acs.jcim.6b00207

- It takes as much time to convert data to make it usable
- A clear need for data standards...

Desperate need for purpose-built, formal data infrastructures



PLAN

Project planning

Experiment planning

RECORD

Data collection, processing & analysis

Process recording

ELNs
LIMS
INVENTORIES

OPEN NOTEBOOKS

MANAGE

COLLECT

DESCRIBE

PROCESS

ANALYSE

PRESERVE

STORAGE

Data selection, preservation & disposal

SHARE

PUBLISH

Data sharing & publication

CHEMISTRY DATABASES & SERVICES; PUBLISHERS

OPEN DATABASES / REPOSITORIES

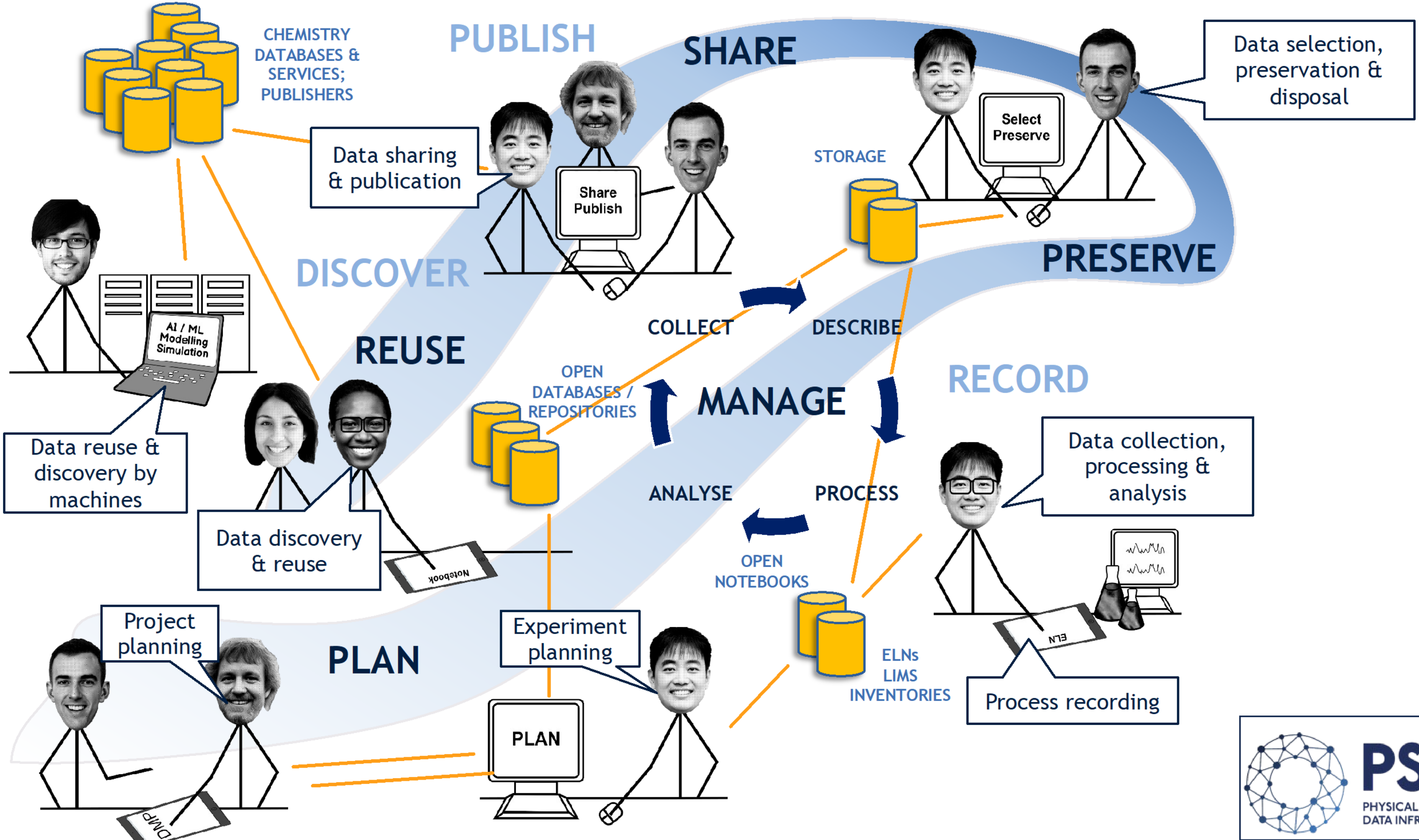
DISCOVER

REUSE

Data discovery & reuse

Data reuse & discovery by machines

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE



PUBLISH

SHARE

PRESERVE

DISCOVER

REUSE

RECORD

PLAN

MANAGE

COLLECT

DESCRIBE

PROCESS

ANALYSE

CHEMISTRY DATABASES & SERVICES; PUBLISHERS

STORAGE

OPEN DATABASES / REPOSITORIES

OPEN NOTEBOOKS

ELNs
LIMS
INVENTORIES

Data sharing & publication

Data selection, preservation & disposal

Data reuse & discovery by machines

Data discovery & reuse

Data collection, processing & analysis

Project planning

Experiment planning

Process recording

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

We are still in a liminal period!

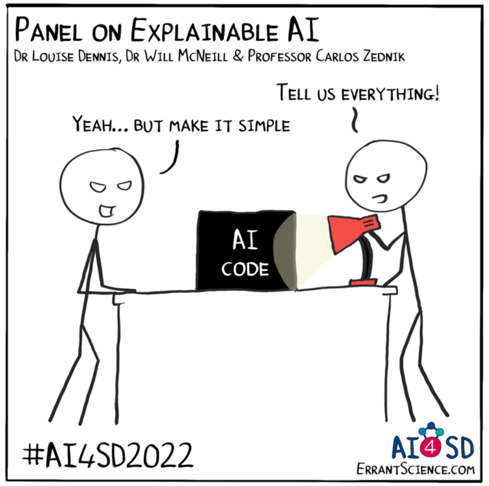
A key concern:

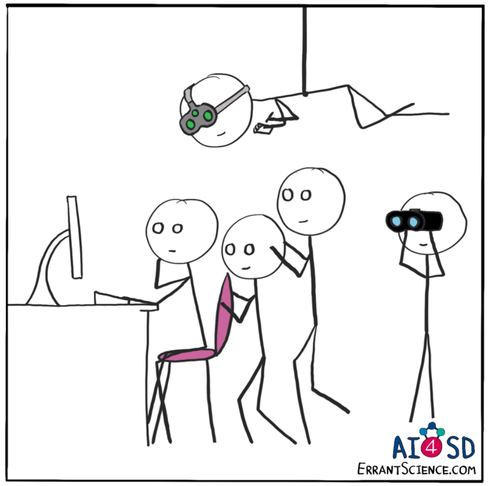
- e.g. Are GPS & Mapping leading to decline in spatial awareness?
- Overreliance on AI may put us in intellectual debt

Ethical & Explainable AI necessary for scientific discovery

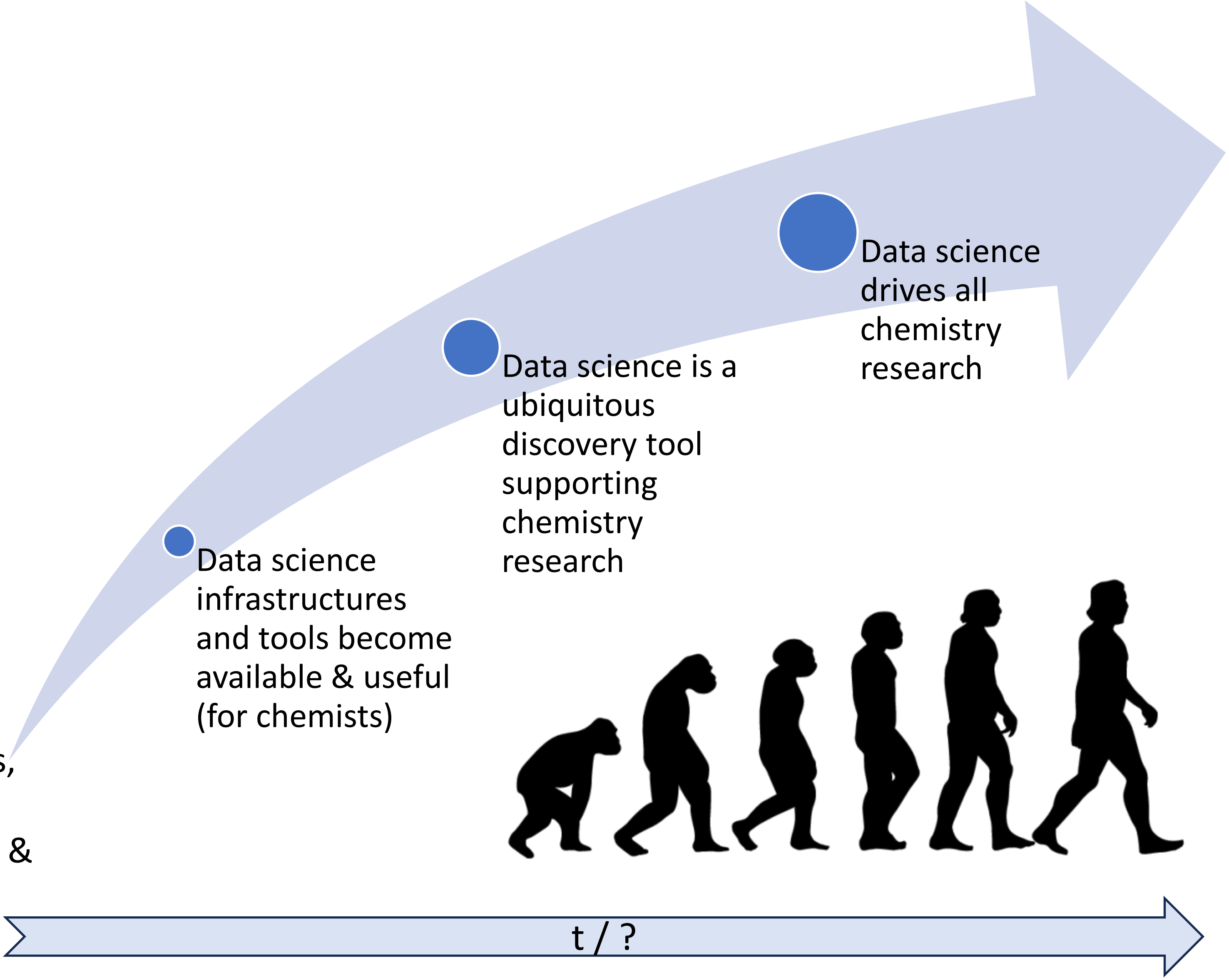
The need for trustworthiness

- Provenance of data / training sets
- Benchmark data
- Ability to scrutinise / understand models and methods


PANEL ON EXPLAINABLE AI
DR LOUISE DENNIS, DR WILL MCNEILL & PROFESSOR CARLOS ZEDNIK
TELL US EVERYTHING!
YEAH... BUT MAKE IT SIMPLE
AI CODE
#AI4SD2022
AI4SD
ERRANTSCIENCE.COM



Evolution of digital chemistry data



Pockets of experts – alongside enthusiasts, advocates, champions & principles

Data science infrastructures and tools become available & useful (for chemists)

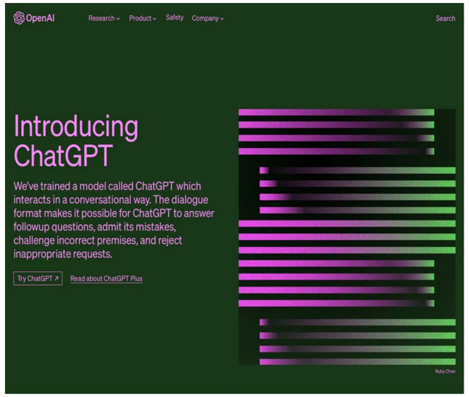
Data science is a ubiquitous discovery tool supporting chemistry research

Data science drives all chemistry research

‘Chemists’ design new materials with specific properties and tell engineers how to make them - without ever entering a laboratory.

Will an AI win the Nobel Prize and replace us?!

“As an AI language model, I cannot predict the future… However, it is possible for an AI to contribute significantly to chemistry research that could be awarded a Nobel Prize”



Introducing ChatGPT

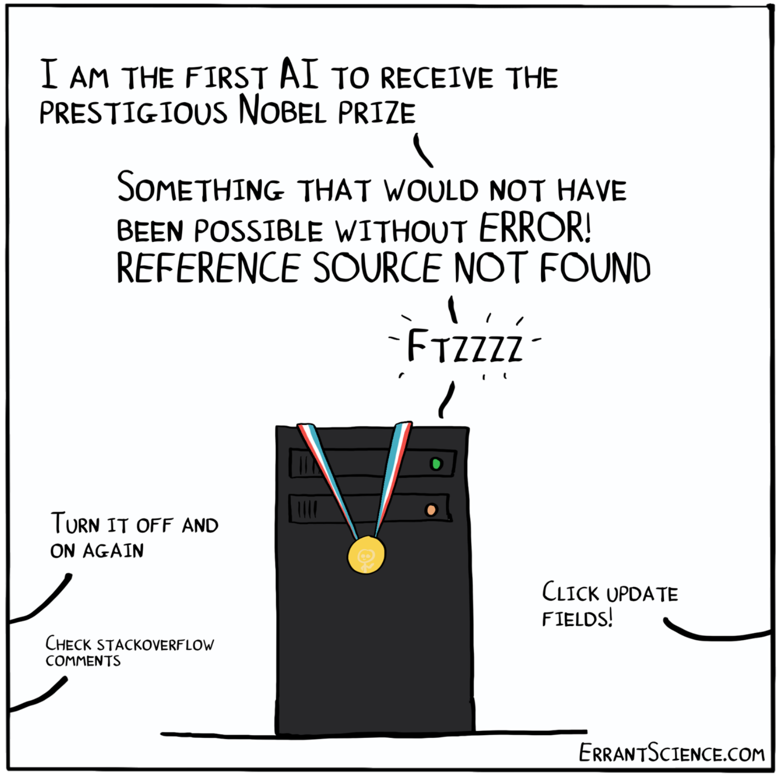
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

”However, it is important to note that the Nobel Prize is awarded to individuals or groups of individuals, not to machines or algorithms…. Even if AI plays a critical role… the prize would likely be awarded to the human scientists who developed and applied the AI methods.”

“…Do you know who will find these things out? Not our AI and ML systems, although I’m sure they’ll help whenever possible. No, it is going to be us. Just like it always has been. The law of conservation of data…”

Derek Lowe

In the Pipeline
Chemistry World


I AM THE FIRST AI TO RECEIVE THE PRESTIGIOUS NOBEL PRIZE
SOMETHING THAT WOULD NOT HAVE BEEN POSSIBLE WITHOUT ERROR! REFERENCE SOURCE NOT FOUND
FTZZZZ
TURN IT OFF AND ON AGAIN
CHECK STACKOVERFLOW COMMENTS
CLICK UPDATE FIELDS!
ERRANTSCIENCE.COM