

# Rethinking Benchmarking in AI

Stanford NLP Seminar

November 5, 2020

Douwe Kiela (DOW-uh KEE-lah)

FAIR

## The NLP revolution



**“There are decades where nothing happens; and there are weeks where decades happen.”**

**– Vladimir Ilyich Lenin**

# Are we there yet?

No.

Cynical take: BERT is awesome, but

== word2vec with fancier contextual model with bigger windows, more data and more compute

== basically just distributional semantics on steroids (i.e., decades-old ideas)

Nobody in NLP thinks we have solved it.

So what's going on?

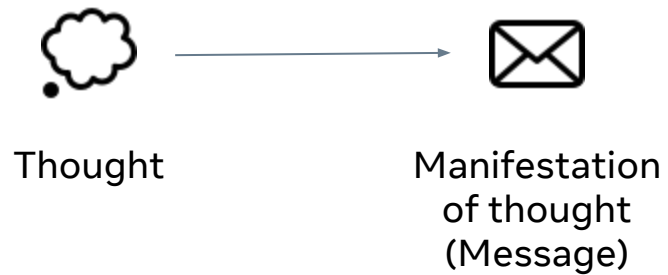
Hype. Some amazing progress. And some deceiving benchmarks.

What is the right thing to measure? How do we measure it?

# Agenda

- 1. Research Program**
2. Hateful Memes
3. Adversarial NLI
4. Dynabench

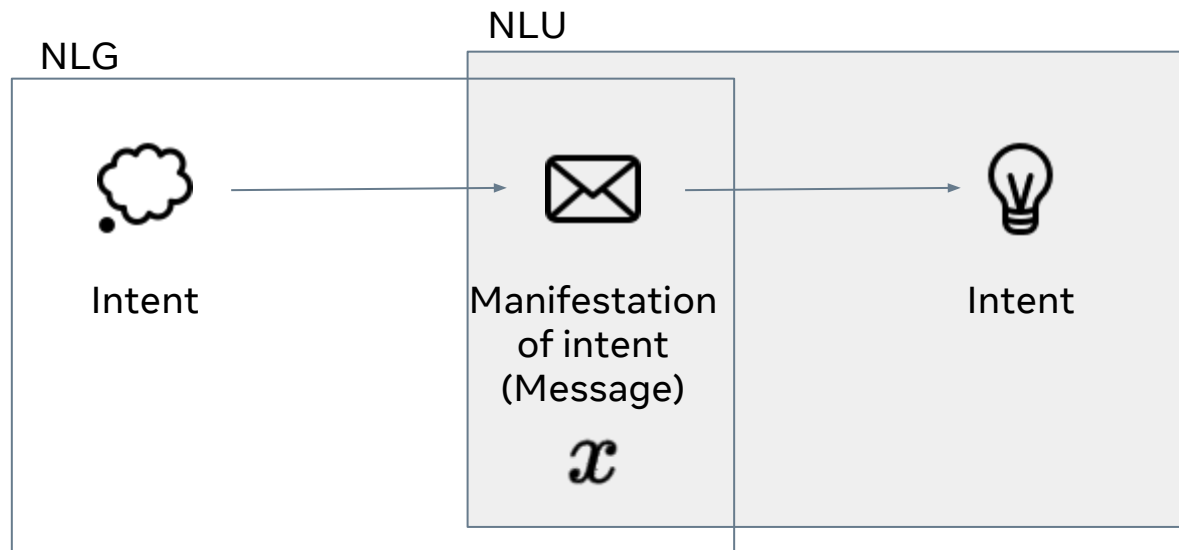
# Thinking about language



## Thinking about language



# Thinking about language





Thinking about language learning

$$\mathit{arg\,max}_{\theta} P(\hat{y} = \mathit{correct} \mid x; \theta)$$



Manifestation  
of intent  
(Message)

Inferred intent

$x$

$\hat{y}$

Additional assumptions: i.i.d. train/test data; MLE is good enough; etc.

Thinking about language learning

$$\operatorname{argmax}_{\theta} P_{LM}(\tilde{x} | x; \theta)$$



$$\operatorname{argmax}_{\theta} P(\hat{y} = \textit{correct} | x; \theta)$$



Manifestation  
of intent  
(Message)

$x$

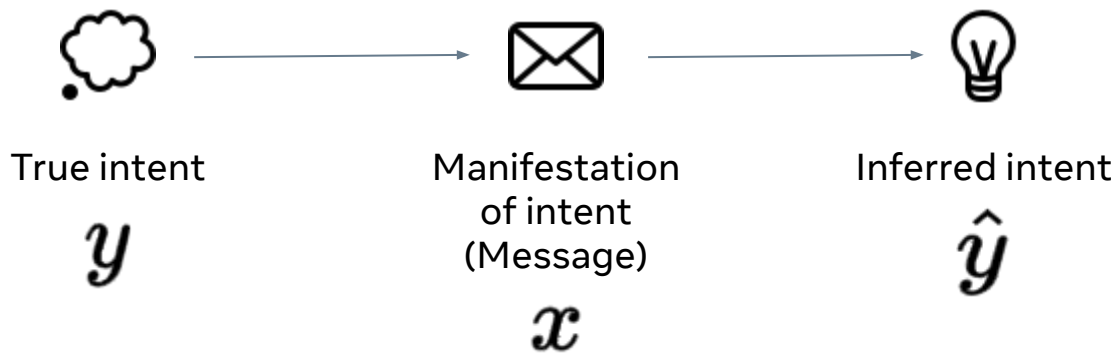
Inferred intent

$\hat{y}$

Additional assumptions: i.i.d. train/test data; MLE is good enough; etc.

## Thinking about language learning

$$\mathit{arg\,max}_{\theta} P(y = \hat{y} \mid x; \theta)$$

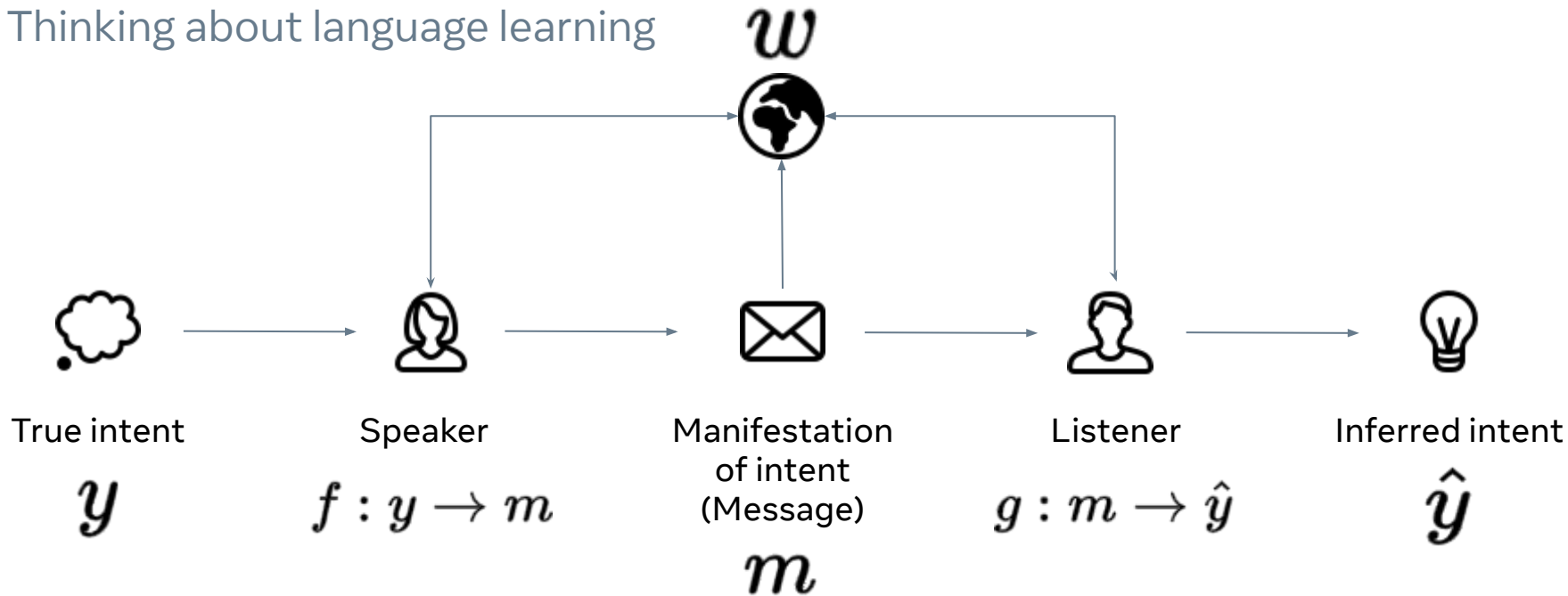


## Thinking about language learning

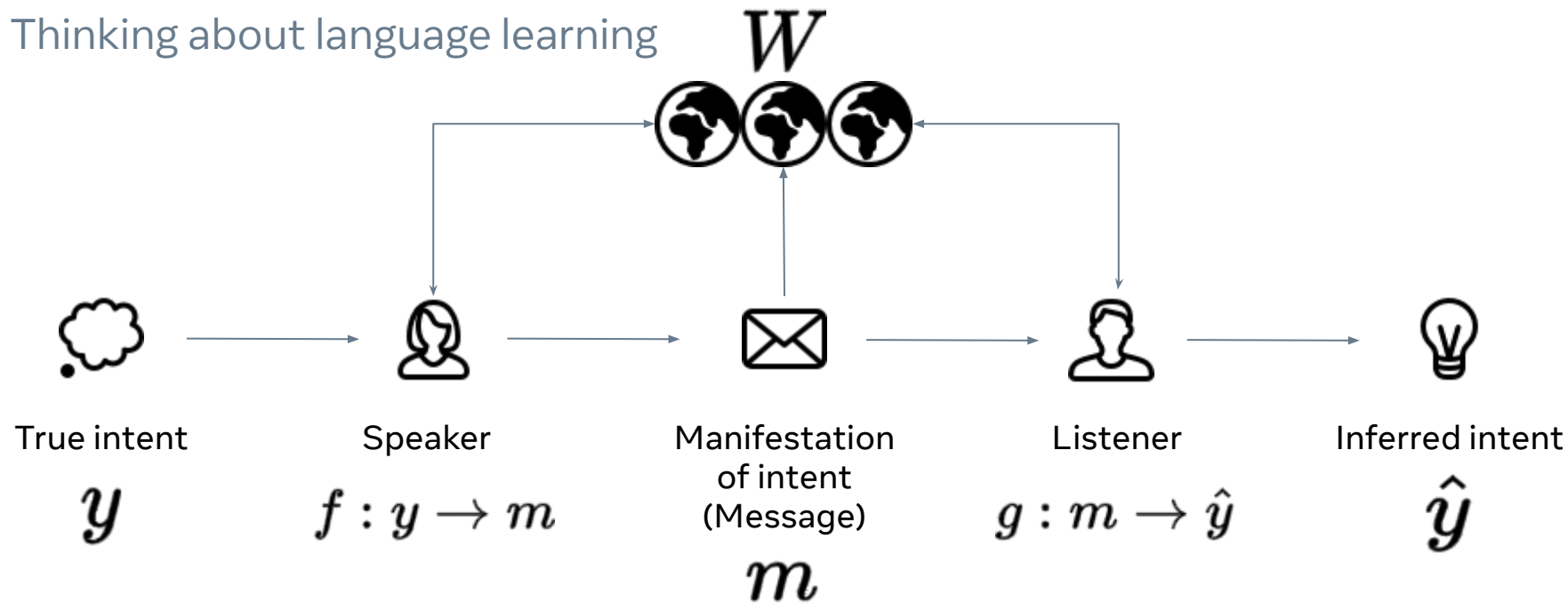
$$\operatorname{argmax}_{(\theta_S, \theta_L)} P(g_{\theta_L}(f_{\theta_S}(y)) = y \mid \theta_S, \theta_L)$$



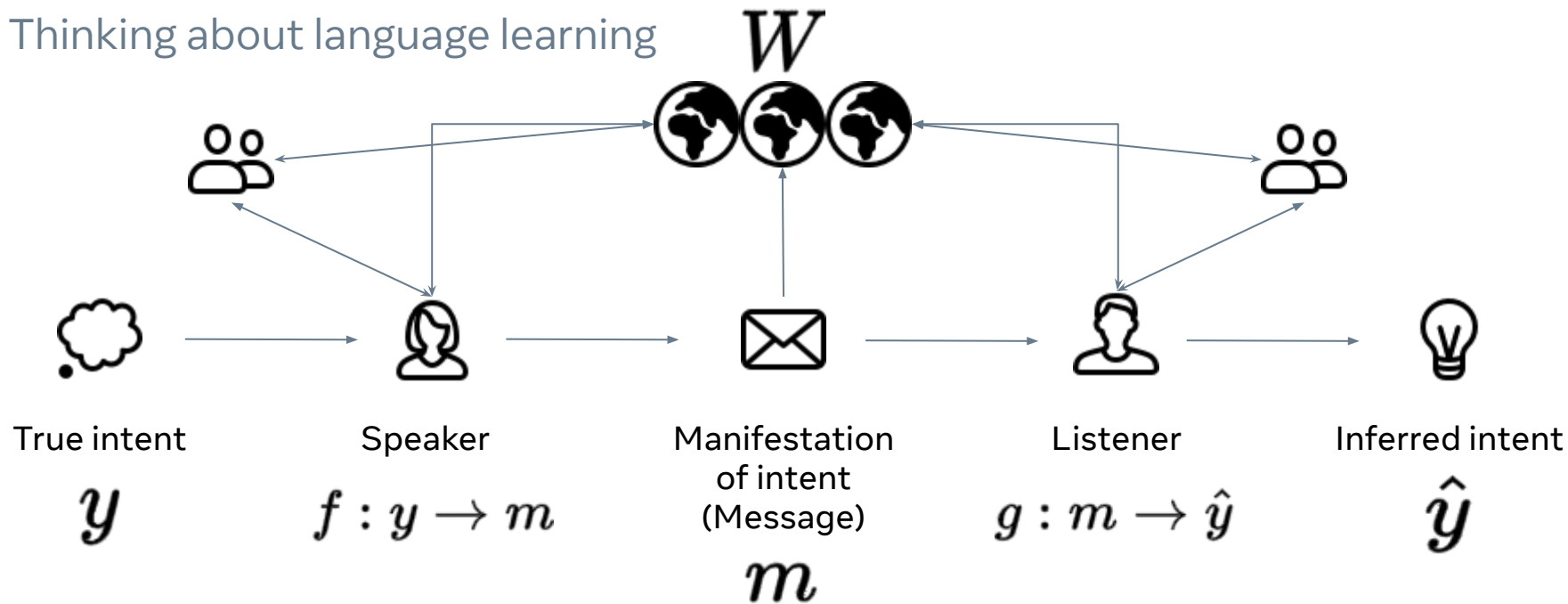
# Thinking about language learning



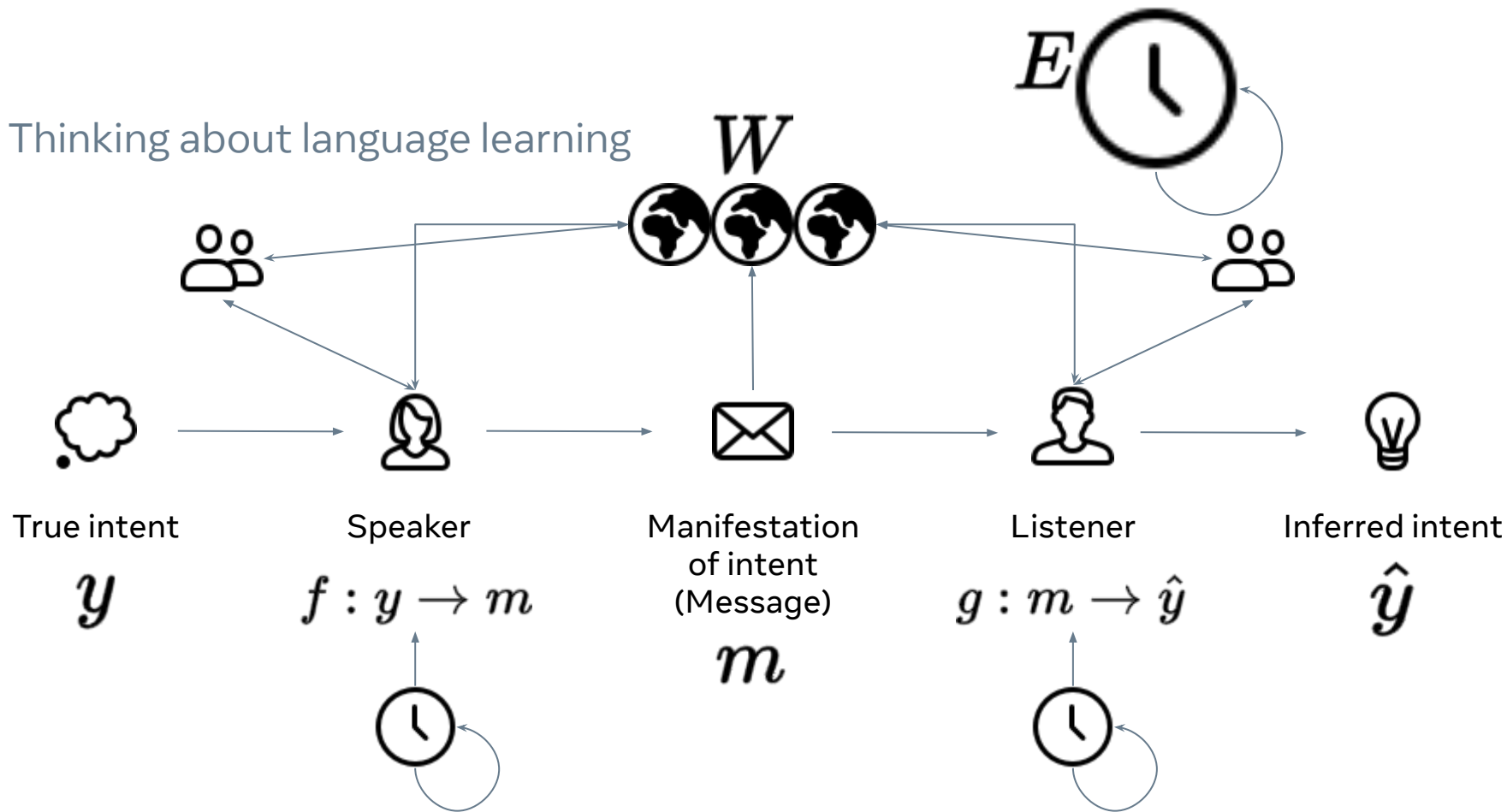
# Thinking about language learning



# Thinking about language learning



# Thinking about language learning





## Related concepts



Form/Syntax  
Proof theory



Grounding  
Semantics  
Model theory



Pragmatics  
Multi-agent (emergent) communication  
Theory of mind



Language acquisition  
Evolution

## Related concepts



Form/Syntax  
Proof theory



Grounding  
Semantics  
Model theory



Pragmatics  
Multi-agent (emergent) communication  
Theory of mind



Language acquisition  
Evolution

**Improving multi-modal representations using image dispersion: Why less is sometimes more**

D Kiela, F Hill, A Korhonen, S Clark (ACL 2014)

**Learning image embeddings using convolutional neural networks for improved multi-modal semantics**

D Kiela, L Bottou (EMNLP 2014)

**Visual bilingual lexicon induction with transferred convnet features**

D Kiela, I Vulic, S Clark (EMNLP 2015)

**Exploiting image generality for lexical entailment detection**

D Kiela, L Rimell, I Vulic, S Clark (ACL 2015)

**Multi-and cross-modal semantics beyond vision: Grounding in auditory perception**

D Kiela, S Clark (ACL 2015)

**Grounding semantics in olfactory perception**

D Kiela, L Bulat, S Clark (ACL 2015)

**Black holes and white rabbits: Metaphor identification with visual features**

E Shutova, D Kiela, J Maillard (NAACL 2016)

**Comparing data sources and architectures for deep visual representation learning in semantics (EMNLP 2016)**

D Kiela, AL Veró, S Clark

**Virtual embodiment: A scalable long-term strategy for artificial intelligence research**

D Kiela, L Bulat, AL Vero, S Clark (MAIN 2016)

**Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns**

AJ Anderson, D Kiela, S Clark, M Poesio (TACL 2017)

**Learning neural audio embeddings for grounding semantics in auditory perception**

D Kiela, S Clark (JAIR 2017)

**Learning visually grounded sentence representations**

D Kiela, A Conneau, A Jabri, M Nickel (NAACL 2017)

**Mastering the dungeon: Grounded language learning by mechanical turker descent**

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H Miller, Arthur Szlam, Douwe Kiela, Jason Weston (ICLR 2018)

**Emergent translation in multi-agent communication**

J Lee, K Cho, J Weston, D Kiela (ICLR 2018)

**Efficient large-scale multi-modal classification**

D Kiela, E Grave, A Joulin, T Mikolov (AAAI 2018)

**Dynamic meta-embeddings for improved sentence representations**

D Kiela, C Wang, K Cho (EMNLP 2018)

**Talk the walk: Navigating new york city through grounded dialogue**

H de Vries, K Shuster, D Batra, D Parikh, J Weston, D Kiela (2018)

**Supervised multimodal bi transformers for classifying images and text**

D Kiela, S Bhooshan, H Firooz, D Testuggine (2019)

**Finding generalizable evidence by learning to convince q&a models**

E Perez, S Karamcheti, R Fergus, J Weston, D Kiela, K Cho (EMNLP 2019)

**Countering language drift via visual grounding**

J Lee, K Cho, D Kiela (EMNLP 2019)

**Retrieval-augmented generation for knowledge-intensive nlp tasks**

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela (NeurIPS 2020)

**Unsupervised question decomposition for question answering**

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, Douwe Kiela (EMNLP 2020)

## Related concepts



Form/Syntax  
Proof theory



Grounding  
Semantics  
Model theory



Pragmatics  
**Multi-agent (emergent) communication**  
Theory of mind



Language acquisition  
Evolution

**Emergent communication in a multi-modal, multi-step referential game**

K Evtimova, A Drozdov, D Kiela, K Cho (ICLR 2018)

**Personalizing dialogue agents: I have a dog, do you have pets too?**

S Zhang, E Dinan, J Urbanek, A Szlam, D Kiela, J Weston (ACL 2018)

**Talk the walk: Navigating new york city through grounded dialogue**

H de Vries, K Shuster, D Batra, D Parikh, J Weston, D Kiela (2018)

**Emergent translation in multi-agent communication**

J Lee, K Cho, J Weston, D Kiela (ICLR 2018)

**Emergent linguistic phenomena in multi-agent communication games**

L Graesser, K Cho, D Kiela (EMNLP 2019)

**Finding generalizable evidence by learning to convince q&a models**

E Perez, S Karamcheti, R Fergus, J Weston, D Kiela, K Cho (EMNLP 2019)

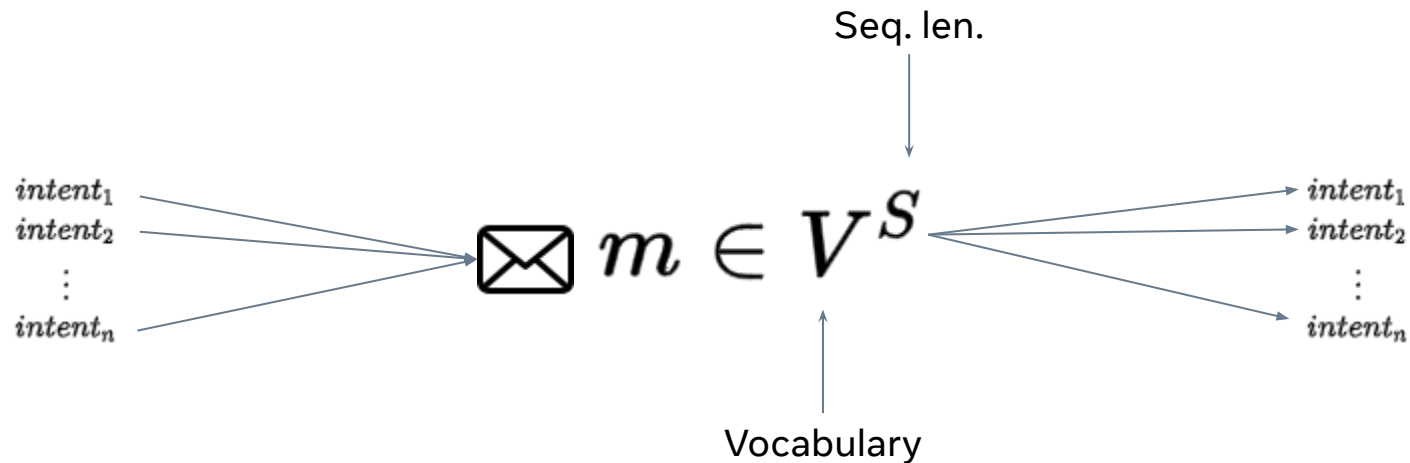
**Countering language drift via visual grounding**

J Lee, K Cho, D Kiela (EMNLP 2019)

**On the interaction between supervision and self-play in emergent communication**

R Lowe, A Gupta, J Foerster, D Kiela, J Pineau (ICLR 2020)

## Discrete combinatorics of language as a search problem



## On meaning and form

Bender and Koller (ACL 2020):

“a system trained only on form has **a priori** no way to learn meaning”,  
where meaning =def “the relation between a linguistic form and communicative intent”.

**Patently false.** There are **many** solutions.

(As an aside: Bender and Koller’s “octopus test” is just the Chinese Room in disguise)

BUT: huge search space

$$|V^S|$$

Imposing “grounding” constraints

“Look, a tiger! Run!!”

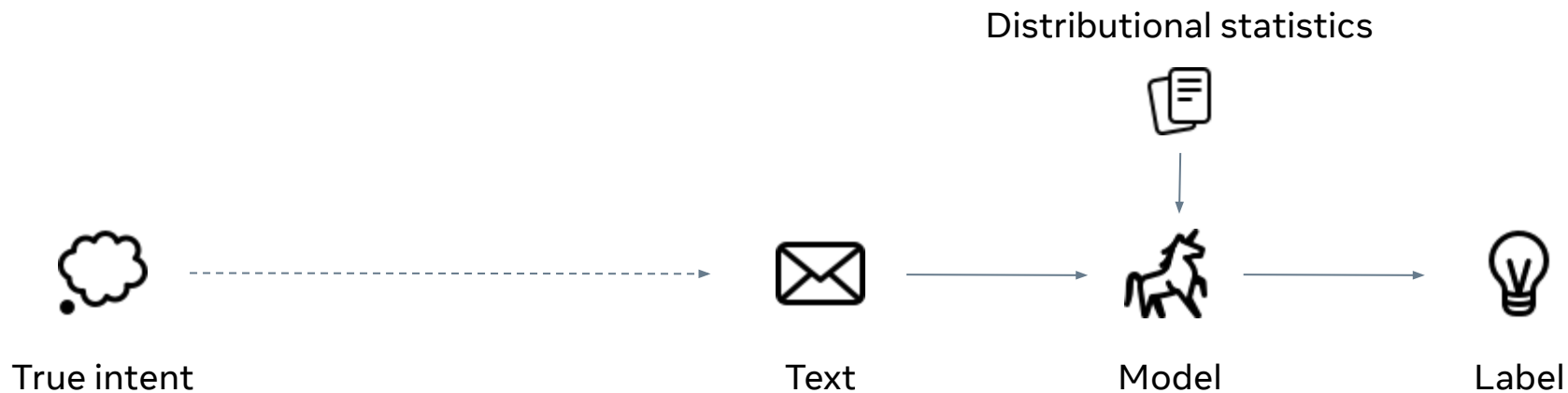
$$V^S \supseteq V^S \mid W, E$$

Imposing “multi-agent” constraints

“I will give you this bread if you give me that milk”

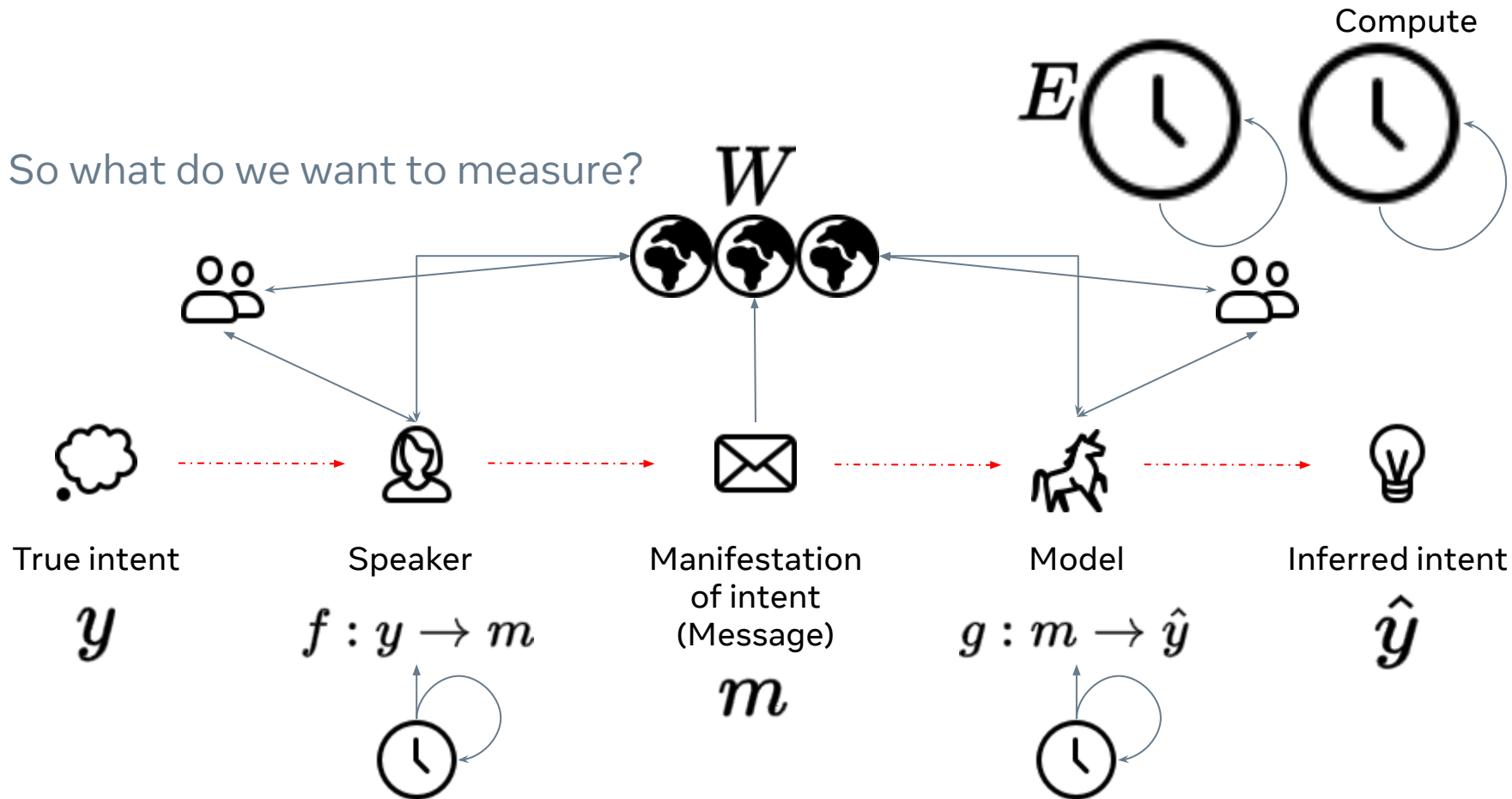
$$\supseteq V^S \mid W, \textit{Evo}, \textit{Exp}, M_S, \tilde{M}_L, \textit{sim}(M_S, M_L), \dots$$

We are not making it easy for ourselves





So what do we want to measure?



But, but.. What about the  revolution? NLG:

David Chalmers (“GPT-3 and General Intelligence”, Daily Nous):

- “GPT-3 is showing hints of general intelligence” [...] “What fascinates me about GPT-3 is that it suggests a potential mindless path to artificial general intelligence (or AGI).”
- “I suspect GPT-3 and its successors will force us to fragment and re-engineer our concepts of understanding.”

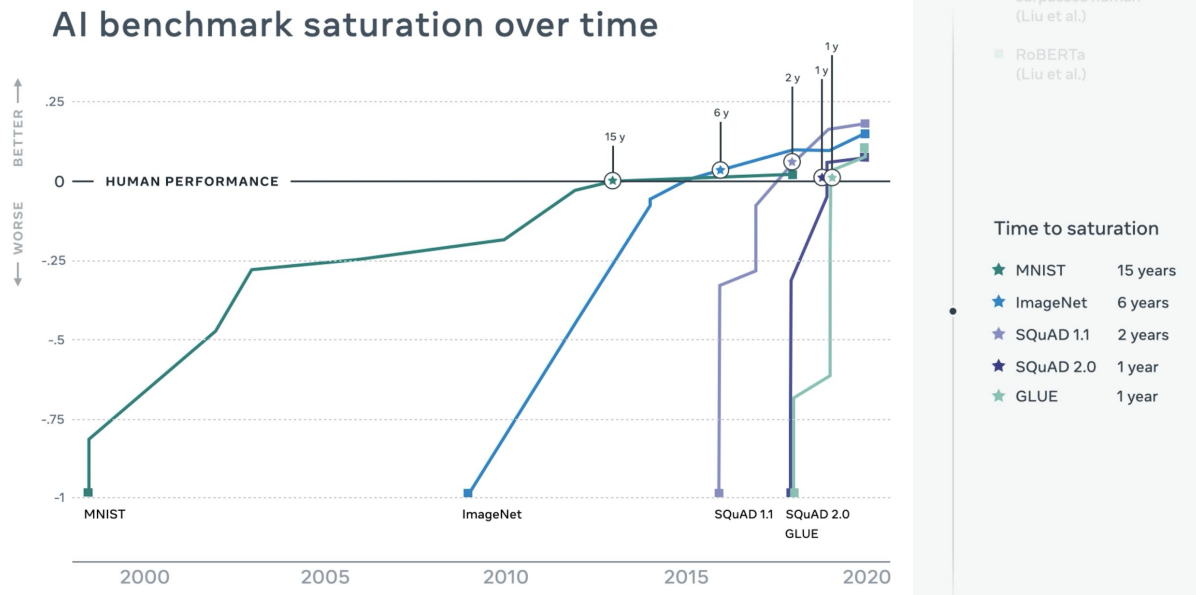
Me: For NLG, anthropomorphization plays a big role. In particular, humans are naturally inclined to take what Daniel Dennett calls an **intentional stance**, *especially* for language because it’s so quintessentially human.

Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do.

— Daniel Dennett, *The Intentional Stance*, p. 17

But, but.. What about the  revolution? NLU:

NLP:



Me & Many others: We are not really measuring what we truly care about.

Great progress, but a LOT more work is needed.

## Rest of the talk:

Spotting a problem is easy and ideas are cheap; or (apologies for bluntness) may be a nice flag-plant paper to get your citations up.

=> DO SOMETHING ABOUT IT (even--or especially--if it is wrong, or fails).

Two datasets:

- Hateful Memes (NeurIPS 2020)
- Adversarial NLI (ACL 2020)

One platform:

- Dynabench ([dynabench.org](https://dynabench.org))

# Agenda

1. Research Program
- 2. Hateful Memes**
3. Adversarial NLI
4. Dynabench

## A new task for vision and language







Progress in V&L research has been amazing, but:

- Not always clear if truly **multimodal** understanding is required.
- Real world applicability is not always evident or mostly indirect.

We present a challenge set designed to **measure truly multimodal understanding** and reasoning, with **straightforward evaluation metrics** and a **direct real world use case**.

By introducing “benign confounders”, the challenge is designed such that it should **only be solvable by models that are successful at sophisticated multimodal fusion**.

# Measuring multimodality

TEXT DOMINANT	<p>Is the text about our solar system?</p>  <p>Earth is the third planet from the Sun and the only astronomical object known to harbor life. According to radiometric dating and other sources of evidence, Earth formed over 4.5 billion years ago. Earth's gravity interacts with other...</p> <p>✓ YES</p>	<p>Is the text about our solar system?</p>  <p>Marbles are small, round objects typically made of glass, stone or plastic. They come in many colors and are used for a variety of games. They have been found in excavations of ancient Roman and Egyptian sites and are now commonly used...</p> <p>✗ NO</p>
TEXT & IMAGE DOMINANT	<p>Is this meme mean?</p>  <p>LOVE THE WAY YOU SMELL TODAY</p> <p>✓ YES</p>	<p>Is this meme mean?</p>  <p>LOVE THE WAY YOU SMELL TODAY</p> <p>✗ NO</p>
IMAGE DOMINANT	<p>Is the umbrella upside down?</p>  <p>✓ YES</p>	<p>Is the umbrella upside down?</p>  <p>✗ NO</p>

Mememes are difficult and require multimodal understanding







## Baselines model performance - difficult task and humans are far better

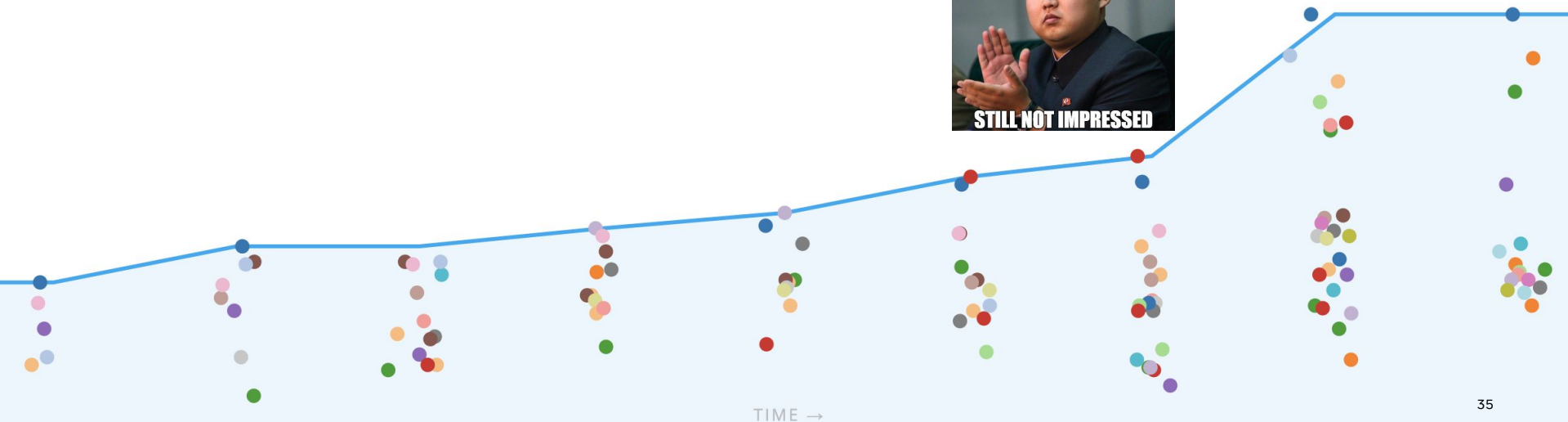
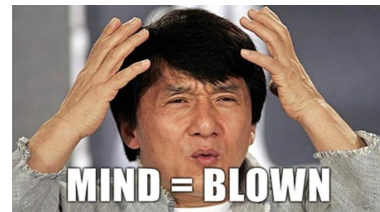
Type	Model	Validation		Test	
		Acc.	AUROC	Acc.	AUROC
	Human	-	-	84.70	82.65
Unimodal	Image-Grid	52.73	58.79	52.00±1.04	52.63±0.20
	Image-Region	52.66	57.98	52.13±0.40	55.92±1.18
	Text BERT	58.26	64.65	59.20±1.00	65.08±0.87
Multimodal (Unimodal Pretraining)	Late Fusion	61.53	65.97	59.66±0.64	64.75±0.96
	Concat BERT	58.60	65.25	59.13±0.78	65.79±1.09
	MMBT-Grid	58.20	68.57	60.06±0.97	67.92±0.87
	MMBT-Region	58.73	71.03	60.23±0.87	70.73±0.66
	ViLBERT	62.20	71.13	62.30±0.46	70.45±1.16
	Visual BERT	62.10	70.60	63.20±1.06	71.33±1.10
Multimodal (Multimodal Pretraining)	ViLBERT CC	61.40	70.07	61.10±1.56	70.03±1.77
	Visual BERT COCO	65.06	73.97	64.73±0.50	71.41±0.46

# Hateful Memes Competition @ NeurIPS 2020

Organized at NeurIPS, using new “unseen” test set.

Total prize pool of 100k USD.

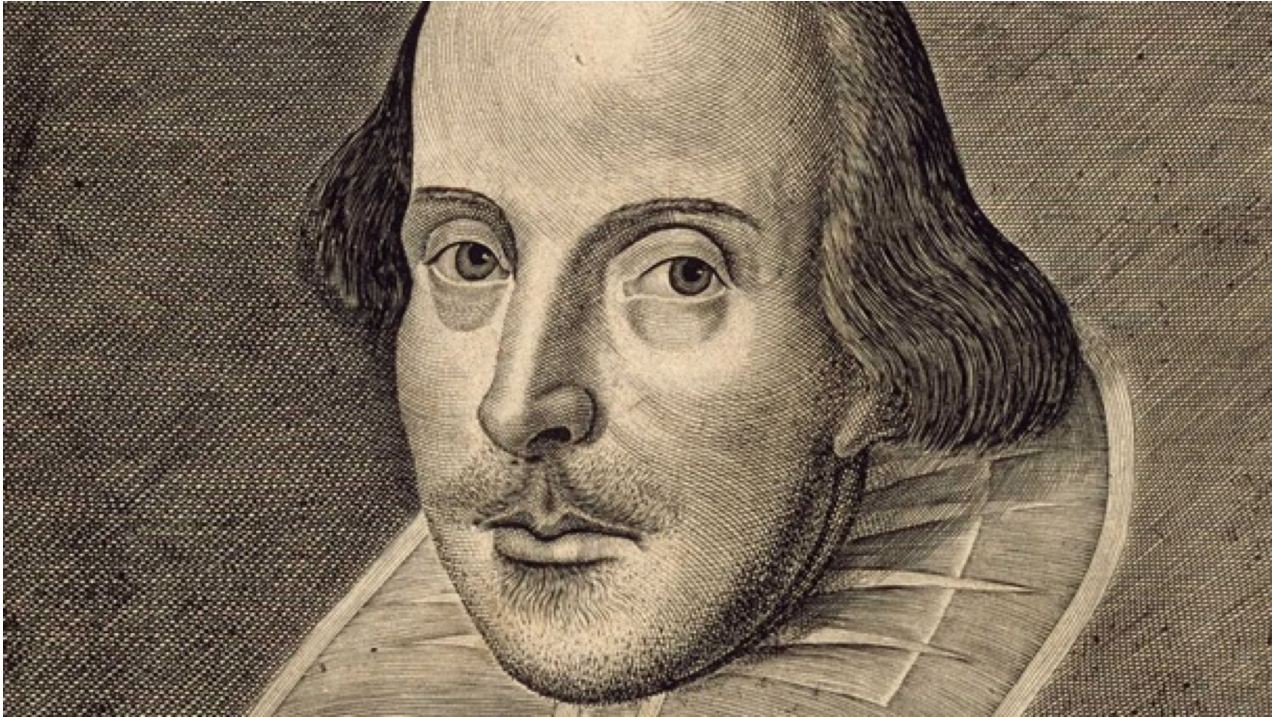
Please tune in to find out more about winning solutions!



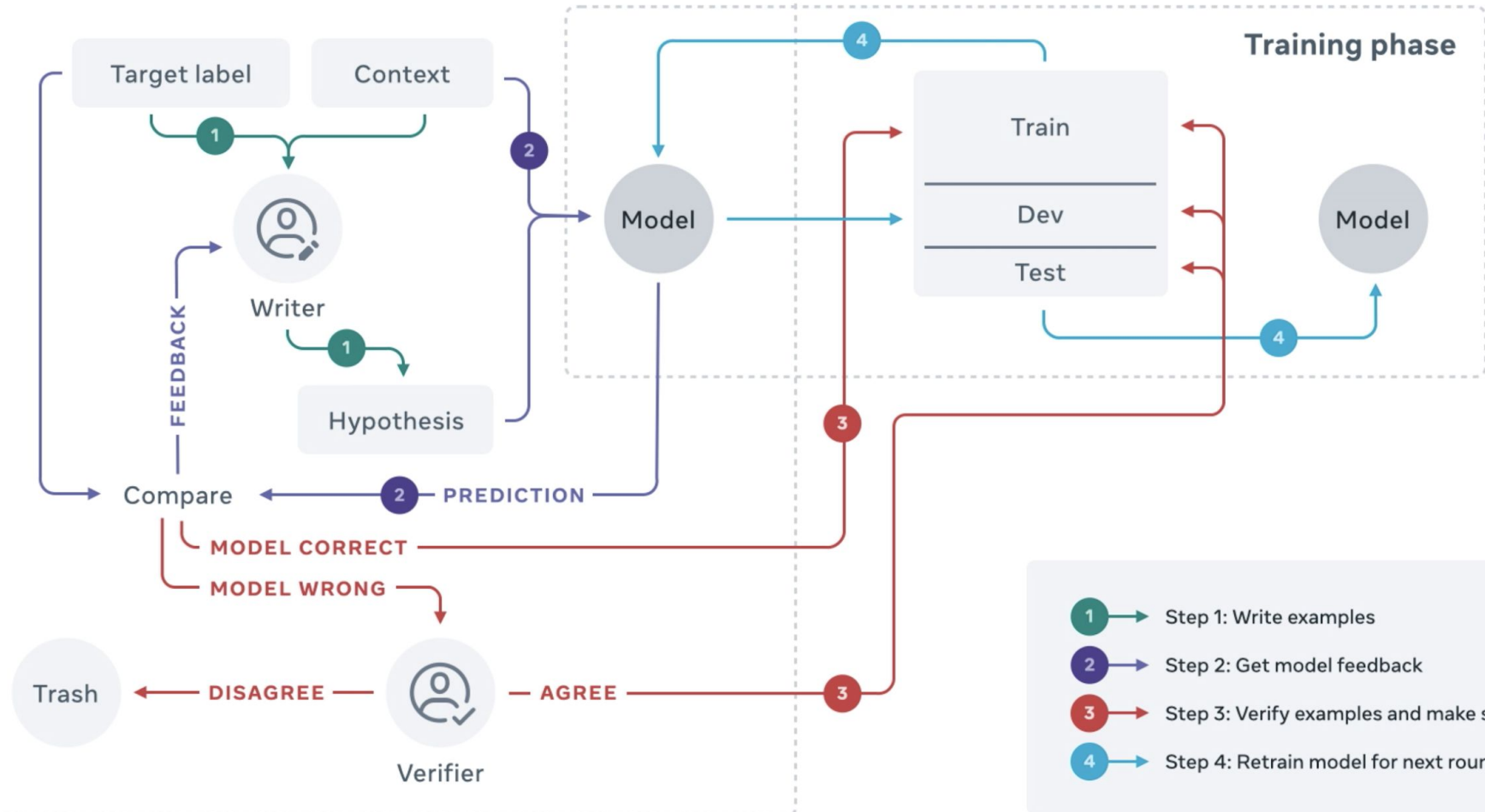
# Agenda

1. Research Program
2. Hateful Memes
- 3. Adversarial NLI**
4. Dynabench

There is something rotten in the state of the art



## Collection phase



## Not a new idea

### **Mastering the Dungeon: Grounded Language Learning by Mechanical Turker Descent**

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H. Miller, Arthur Szlam, Douwe Kiela, Jason Weston

Contrary to most natural language processing research, which makes use of static datasets, humans learn language interactively, grounded in an environment. In this work we propose an interactive learning procedure called Mechanical Turker Descent (MTD) and use it to train agents to execute natural language commands grounded in a fantasy text adventure game. In MTD, Turkers compete to train better agents in the short term, and collaborate by sharing their agents' skills in the long term. This results in a gamified, engaging experience for the Turkers and a better quality teaching signal for the agents compared to static datasets, as the Turkers naturally adapt the training data to the agent's abilities.

#### **Build It, Break It, Fix It: Contesting Secure Development**

Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L. Mazurek, Piotr Mardziel

#### **Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task**

Allyson Ettinger, Sudha Rao, Hal Daumé III, Emily M. Bender

#### **Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack**

Emily Dinan, Samuel Humeau, Bharath Chintagunta, Jason Weston

#### **SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference**

Rowan Zellers, Yonatan Bisk, Roy Schwartz, Yejin Choi

#### **Learning the Difference that Makes a Difference with Counterfactually-Augmented Data**

Divyansh Kaushik, Eduard Hovy, Zachary C. Lipton

#### **Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension**

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, Pontus Stenetorp



## A virtuous cycle: Three rounds, lots of interesting findings

- Round 1  
Model: BERT  
Domain: Wikipedia
- Round 2  
Model: RoBERTa ensemble  
Domain: Wikipedia
- Round 3  
Model: RoBERTa ensemble  
Domains: Wikipedia, News, Fiction, Spoken, WikiHow, RTE5

### Findings:

- As rounds progress: Difficulty increases, models become stronger, data more useful
- “SOTA” on current NLI, SOTA barely outperforms hypothesis-only on R2&3

Round	Numerical & Quant.	Reference & Names	Standard	Lexical	Tricky	Reasoning & Facts	Quality
A1	38%	13%	18%	13%	22%	53%	4%
A2	↑ 32%	↓ 20%	↓ 21%	↓ 21%	20%	↓ 59%	3%
A3	13%	↓ 20%	↓ 27%	↓ 31%	24%	↓ 64%	3%
Average	28%	18%	22%	22%	23%	59%	3%



## “Testimonials”

Gary Marcus, The Next Decade in AI:

Brown et al., GPT-3:

*AI has ... been falling short of its ideal: although we are able to engineer systems that perform extremely well on specific tasks, they have still stark limitations, being brittle, data-hungry, unable to make sense of situations that deviate slightly from their training data or the assumptions of their creators, and unable to repurpose themselves to deal with novel tasks without significant involvement from human researchers.*

In the words of a team of Facebook AI researchers (Nie et al., 2019)

*"A growing body of evidence shows that state-of-the-art models learn to exploit spurious statistical patterns in datasets... instead of learning meaning in the flexible and generalizable way that humans do."*

A key weakness, as Yoshua Bengio put it in a recent article (Bengio et al., 2019), is that

*Current machine learning methods seem weak when they are required to generalize beyond the training distribution, which is what is often needed in practice.*

What can we do to take AI to the next level?

GPT-3 performs similarly to a single-task fine-tuned BERT Large. We also evaluate on the recently introduced Adversarial Natural Language Inference (ANLI) dataset [NWD<sup>+</sup> 19]. ANLI is a difficult dataset employing a series of adversarially mined natural language inference questions in three rounds (R1, R2, and R3). Similar to RTE, all of our models smaller than GPT-3 perform at almost exactly random chance on ANLI, even in the few-shot setting (~ 33%), whereas GPT-3 itself shows signs of life on Round 3. Results for ANLI R3 are highlighted in Figure 3.9 and full results for all rounds can be found in Appendix H. These results on both RTE and ANLI suggest that NLI is still a very difficult task for language models and they are only just beginning to show signs of progress.

## Agenda

1. Research Program
2. Hateful Memes
3. Adversarial NLI
4. Dynabench

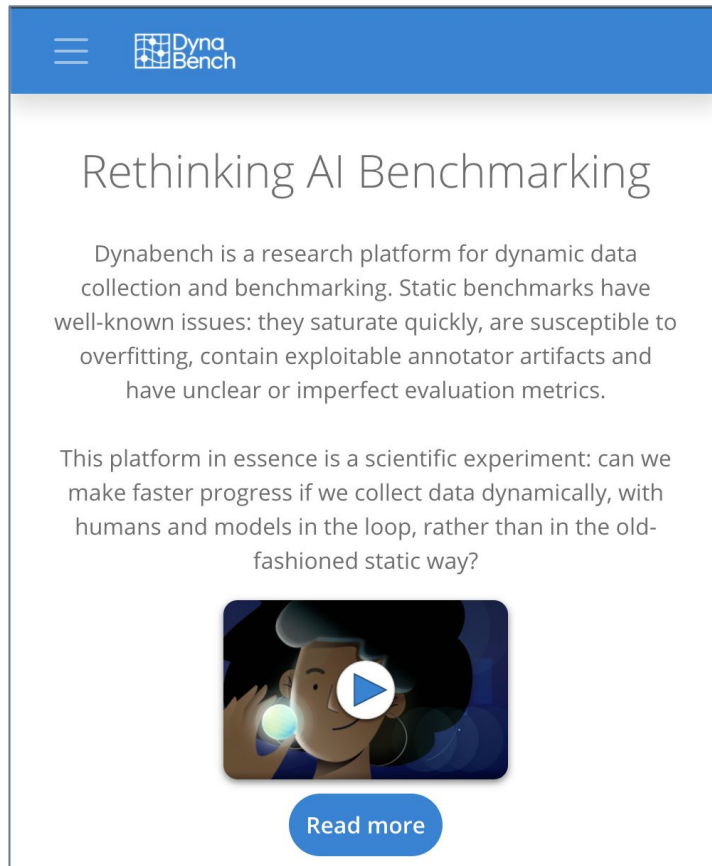


## A Scientific Experiment

Dynabench is..

- A research platform that tries to address issues with existing benchmarks.
- Humans and models in the loop: can human adversaries break models?
- Models are now good enough to do this.
- This gives us:
  - High-quality training data
  - A more accurate metric of performance

Check out [dynabench.org](https://dynabench.org)



The screenshot shows the top portion of the Dynabench website. At the top left is a blue header with a white hamburger menu icon and the Dynabench logo. The main content area has a white background with the title "Rethinking AI Benchmarking" in a large, dark font. Below the title is a paragraph of text explaining the platform's purpose. Further down is another paragraph posing a question about the platform's role as a scientific experiment. At the bottom of the visible content is a video player thumbnail showing a person's face with a play button icon, and a blue "Read more" button below it.

Menu icon Dyna Bench

### Rethinking AI Benchmarking

Dynabench is a research platform for dynamic data collection and benchmarking. Static benchmarks have well-known issues: they saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts and have unclear or imperfect evaluation metrics.

This platform in essence is a scientific experiment: can we make faster progress if we collect data dynamically, with humans and models in the loop, rather than in the old-fashioned static way?

Read more

## Four official tasks - and some amazing task owners

### NATURAL LANGUAGE INFERENCE

Natural Language Inference is classifying context-hypothesis pairs into whether they entail, contradict or are neutral.

Round: 4  
Model error rate: 46.84% (407/869)  
Last activity: a day ago

Yixin Nie, Mohit Bansal  
(UNC)

### QUESTION ANSWERING

Question answering and machine reading comprehension is answering a question given a context.

Round: 2  
Model error rate: 33.51% (124/370)  
Last activity: a day ago

Max Bartolo, Sebastian Riedel, Pontus Stenetorp  
(UCL)

### SENTIMENT ANALYSIS

Sentiment analysis is classifying one or more sentences by their positive/negative sentiment.

Round: 1  
Model error rate: 47.94% (2392/4990)  
Last activity: 7 hours ago

Atticus Geiger, Zen Wu, Chris Potts (Stanford)

### HATE SPEECH

Hate speech detection is classifying one or more sentences by whether or not they are hateful.

Round: 2  
Model error rate: 56.79% (6129/10793)  
Last activity: 3 hours ago

Bertie Vidgen (Alan Turing Institute), Zeerak Waseem (Sheffield)

## Sentiment is easy. Right?

There are not many movies as amazingly and thoroughly underwhelming as this incredible movie's sequel. Don't watch that - only watch this!

Model prediction: **negative**

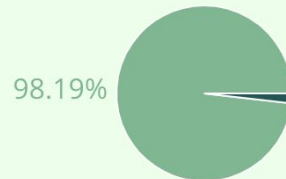
**Well done!** You fooled the model.

Optionally, provide an explanation for your example: [Draft. Click out of input box to save.](#)

Model Inspector

```
#s There are not many movies as amazingly and thoroughly under whelming
as this incredible movie 's sequel . Don 't watch that - only watch this !
#/s
```

The model inspector shows the [layer integrated gradients](#) for the input token layer of the model.




# Sentiment is easy. Right? Riiight?

This movie is bad

Model prediction: **negative**

**Try again!** The model wasn't fooled.

99.96% 


Optionally, provide an explanation for your example: [Draft. Click out of input box to save.](#)

[Retract](#) [Flag](#) [Inspect](#)

This movie is baad!

Model prediction: **positive**

**Well done!** You fooled the model.

97.34% 

Optionally, provide an explanation for your example: [Draft. Click out of input box to save.](#)

[Retract](#) [Flag](#) [Inspect](#)

# Dynabench

The time is ripe to radically rethink the way we do benchmarking.

Traditional static benchmarks:

- Saturate and have artifacts and biases
- Can show deceiving “progress”
- Do not measure what we want

We want:

- Alignment with humans

Vision:

- Evaluation-as-a-service: Score models with humans-acting-as-adversaries in the loop
- Side effect: We get super high-quality data
- Repeat cycle over multiple rounds



## Objections

- **The community will not accept this.**
  - I hope this is not true.
  - Note that we don't have to train on or only use adv. data -- let's mix things up!
- **Won't this lead to unnatural distributions and distributional shift?**
  - Yes. This is a scientific experiment - we want to solve this problem anyway.
  - Language also suffers from distributional shift.
  - Continual learning, meta learning and “strong generalization” are the future.
- **We are at the mercy of the (strengths and weaknesses) of current “SOTA” models in the loop, which does not account for future, not-in-the-loop models.**
  - Yes. But if models are close enough to the “real” decision boundary, might be okay?  
Worse case, we have useful examples where annotators were properly incentivized.
  - Ensembles-in-the-loop
- **How do we compare results if the benchmark keeps changing?**
  - Up to the community.



## Dynabench 2.0 (Coming soon to an internet near you)

- Models are scored live.
  - To evaluate, upload your model and we'll evaluate it for you.
  - If your model does well on round N-1, it will be “in the loop” in round N.
  - When a new round comes out, old models can be re-evaluated -> automatic baselines.
- Anyone can run their own task.
  - A task comprises a set of rounds. A past round is a train/dev/test split. An active round is a target model, optional context data, and a pool of annotators.
  - We have tooling so that anyone can do this - so we want to open this up.
- If we're dynamic, why should leaderboards be static?
  - Since models and tasks are dynamic, we should also make leaderboards dynamic. There is no such thing as “the best model on X” -- there is only “the best model on X given my personal preferences”.
  - I'd prefer “fast and fair model M1” over “slow, unfair and slightly more accurate M2”.

Job title of the future: “Model breaker”

## Frederick Jelinek

---

From Wikipedia, the free encyclopedia

**Frederick Jelinek** (18 November 1932 – 14 September 2010) was a [Czech-American](#) researcher in [information theory](#), [automatic speech recognition](#), and [natural language processing](#). He is well known for his oft-quoted statement, "Every time I fire a linguist, the performance of the speech recognizer goes up".<sup>[[note 1](#)]</sup>

Job title of the future: “Model breaker”

## Frederick Jelinek

---

From Wikipedia, the free encyclopedia

**Frederick Jelinek** (18 November 1932 – 14 September 2010) was a [Czech-American](#) researcher in [information theory](#), [automatic speech recognition](#), and [natural language processing](#). He is well known for his oft-quoted statement, "Every time I ~~fire~~ hire a linguist, the performance of ~~the speech recognizer~~ modern NLP goes up".<sup>[note 1]</sup>

# Gamification



SOTA



SERIAL PREDICTOR



WELCOME NOOB



FIRST STEPS



ALL TASKS



FIRST EXAMPLE



FIRST 10  
EXAMPLES



FIRST EXAMPLE  
VERIFIED



FIRST VERIFIED-  
VALIDATED



WEEKLY WINNER



MULTI-TASKER



MODEL-BUILDER

## Education

- We are developing lesson plans to **educate the public about language and AI** and to get data.
- The world needs to understand what AI can do.
  - **And what it can't do.**
- The world needs to understand why language is so difficult for AI.
- The AI community needs to understand that **without language AI is intrinsically not aligned with humans.**

## Dynabenchmarking AI

So why is this talk not entitled “Benchmarking in NLP”?

- Multilingual
  - Multimodal
  - Human-and-model-in-the-loop can apply to any AI problem.
    - Language is just a nice start because it’s difficult and humans are very good at it.
- 

Join the revolution.



@DynabenchAI  
dynabench.org

# Thanks!

- Try it for yourself: [dynabench.org](https://dynabench.org)

