The Easy, the Hard and the Ugly

- ✓Easy
  - ✓Inference (*fit*)
  - ✓Fine-Tuning (*predict*)
- ✓Hard
  - ✓Pre-training
- ➢Ugly (Responsible AI)
  - Bias
  - Toxicity
  - Misinformation
  - Hallucinations
  - Plagiarism

Ugly: Outline

- **Benchmarking** (Ken)
- Labeling (Omar)
- Bias, toxicity, misinformation, plagiarism (Ken)
- 1 slide on hard problems for researchers/practitioners

Benchmarking:
Big Fan of Metrics (EMNLP)

Emerging Trends: SOTA-Chasing

Published online by Cambridge University Press: **08 February 2022**

Kenneth Ward Church and Valia Kordoni

Clichés

- *Whatever you measure, you get*
- *Be careful what you ask for*
- *It's difficult to make predictions, especially about the future*
  - Yogi Berra

Lack of Substance

- SOTA-Chasing
- Pointless Mindless Metrics
  - (The point should be to make progress on a real problem)
- Experimental Psychology
  - Validity
  - Reliability

Benchmarking:

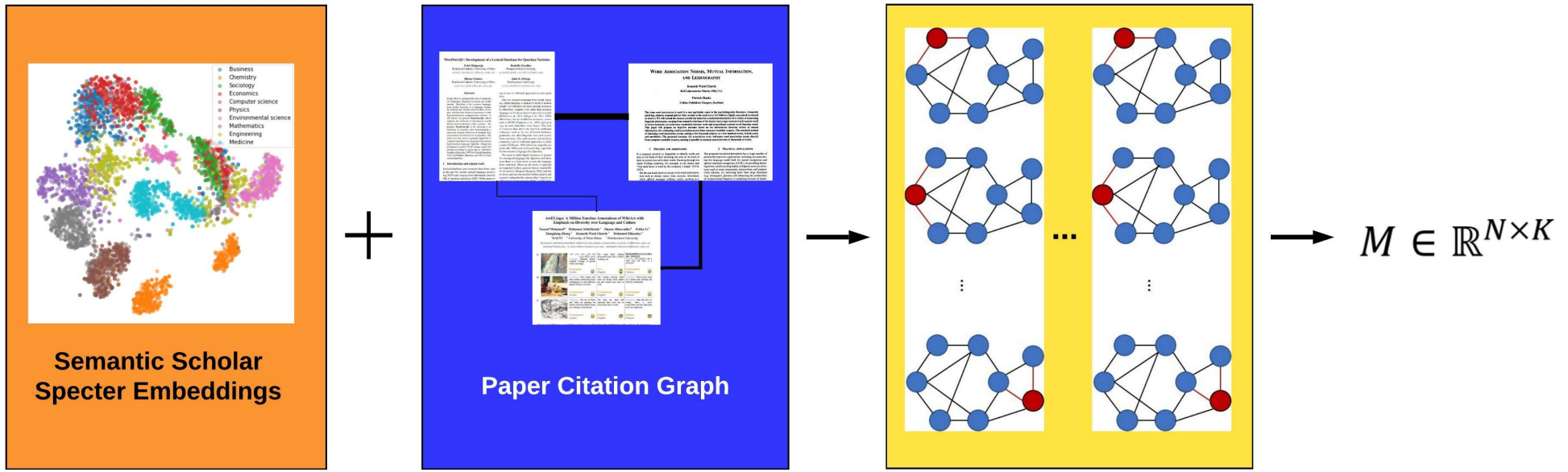
Extrapolation: Past → Future

- ACL-2021 Workshop
  - Benchmarking:
  - Past, Present and Future
    - https://github.com/kwchurch/Benchmarking_past_present_future

1. Past
   i. John Makhoul
   ii. Mark Liberman
   iii. Ellen Voorhees
   iv. John Mashey
2. Present
   i. Nan Duan, Qi Zhang and Ming Zhou
   ii. Hua Wu and Jing Liu
   iii. Neville Ryant
   iv. Brian MacWhinney and Saturnino Luz
   v. Douwe Kiela
   vi. Eunsol Choi
   vii. Anders Søgaard
3. Future
   i. Greg Diamos, Peter Mattson and David Kanter
   ii. Dave Ferrucci
   iii. Ido Dagan
   iv. Samuel Bowman

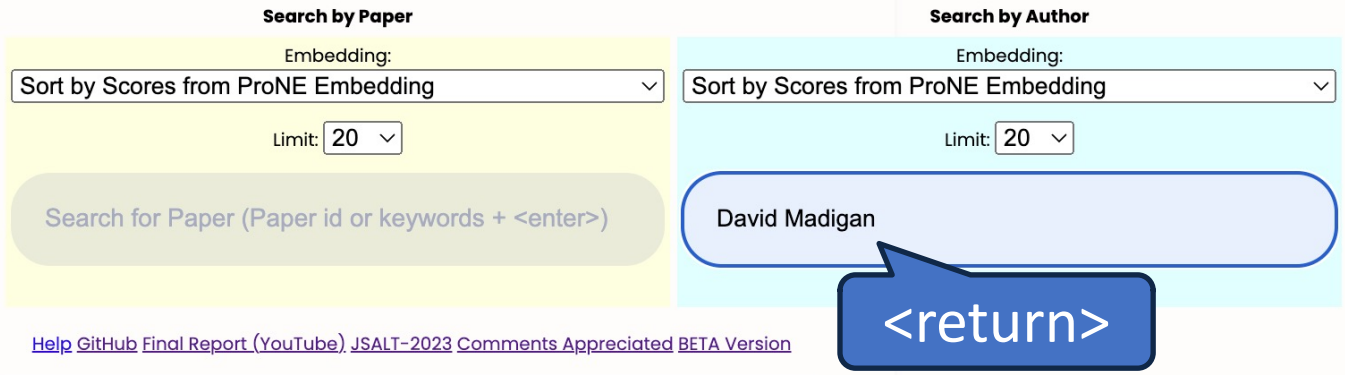
Better Together:

Text (Titles, Abstracts, Body); Context (Citations)



https://www.semanticscholar.org/product/api/gallery

Find Similar Papers



Le Mans Université

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

allomedia

EAI The Institute for Experiential AI Northeastern University

Opportunity: More Realistic Benchmarks

The Real World

- Large (Metcalfe's Law)
- Growing
- Dirty
  - Missing Values
  - Bad/misleading Values

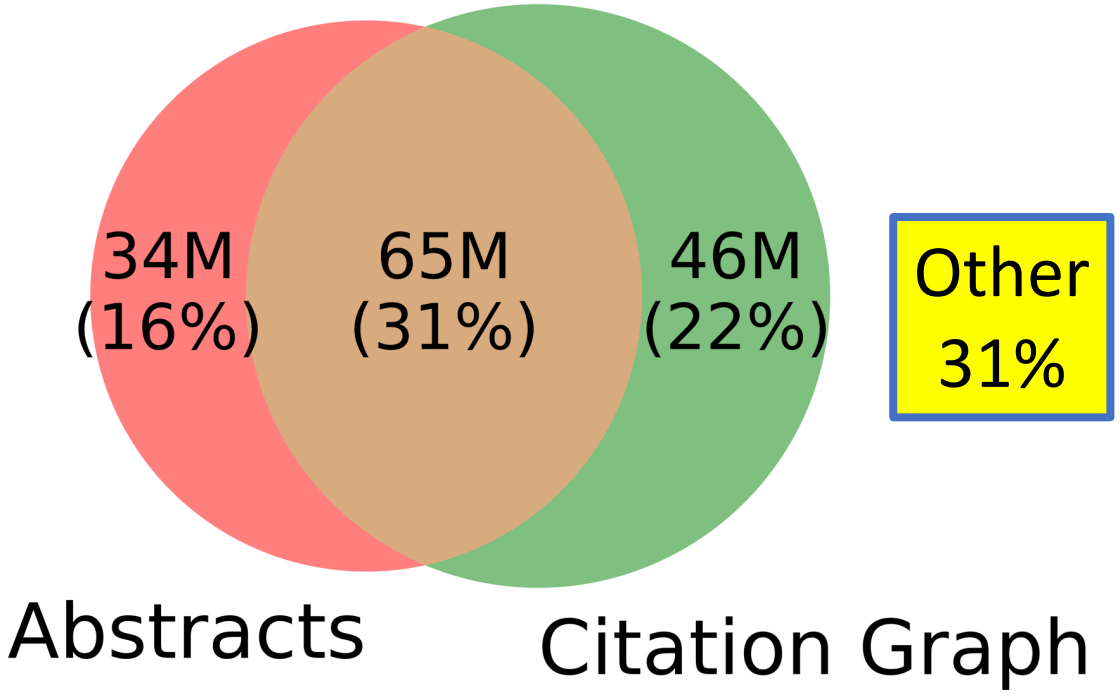
Idealizations (Standard Benchmarks)

- Small
- Static
- Clean

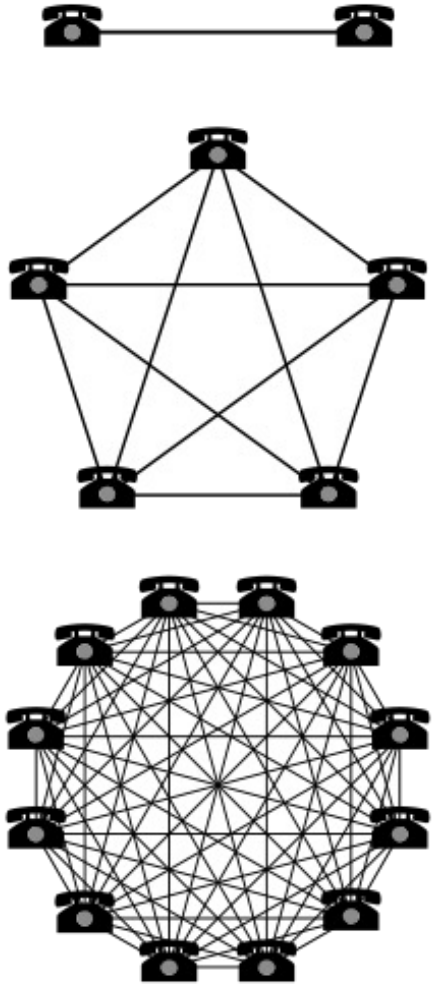
Realities (Dirty): Missing Values

- A: papers with abstracts, a
- L: papers with links, l

- Opportunity: $A \cup L$
  - Prior work targets small subset
    - Specter: designed for A
    - GNNs: designed for $A \cap L$



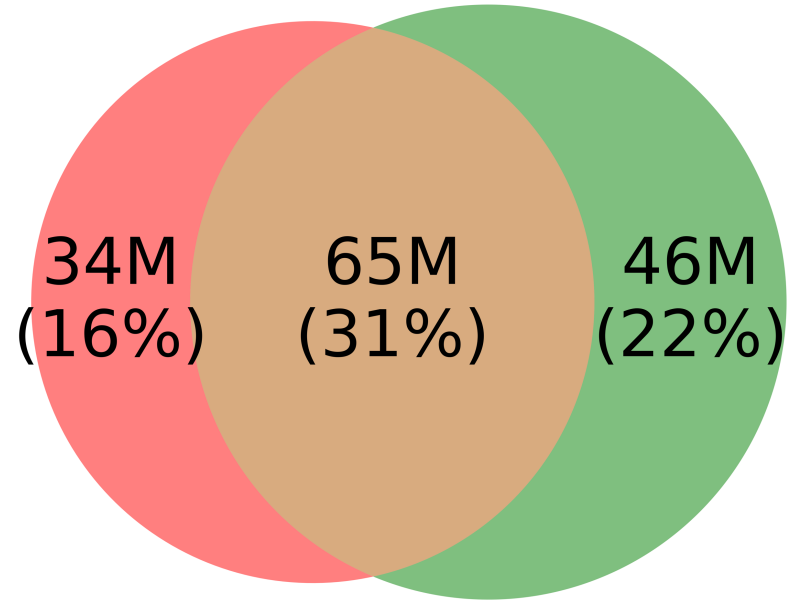
Metcalfe's Law (Network Effects)



- History: 3Com was selling small networks
  - 3 = 1 printer + 2 computers
  - Metcalfe argued they should sell bigger networks
    - (and more 3Com products)
    - because of economies of scale
- Economy of Scale:
  - Benefits scale faster than costs
    - Benefits: $\sim n^2$
    - Costs: $\sim n$
- Law has been good for AT&T, Google, Social Media
- Hypo: also good for Academic Search (and Product Search)
  - Consequently, we should experiment with large graphs

Better-Together Conjectures

- Multiple representations are helpful
  - for missing values (and bad values)
    - If abstracts are missing, use links
    - If links are missing, use abstracts
  - and for graphs of different sizes
    - When graphs are small:
      - Text >> Links
    - when graphs are large:
      - Links >> Text (Metcalfe's Law)



Opportunities

Scale

- Many benchmarks return a single number (figure of merit)
  - How does performance scale
  - with the size of the problem?

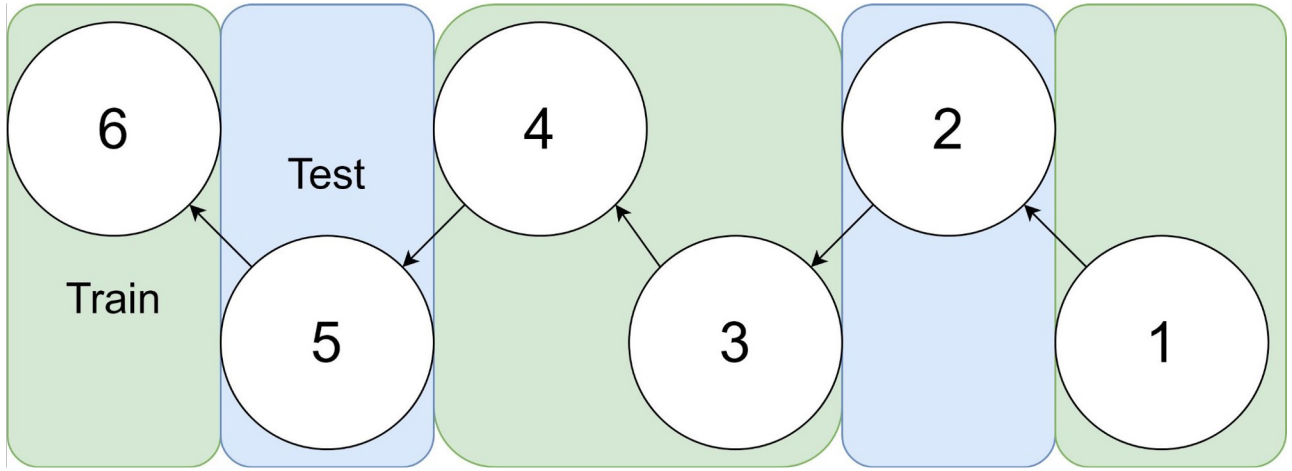
Prediction

- Standard test/train splits
  - More relevant to interpolation
  - than extrapolation
- Time is asymmetric
  - *It is difficult to make predictions*
  - *especially about the future*
    - -- Yogi Berra

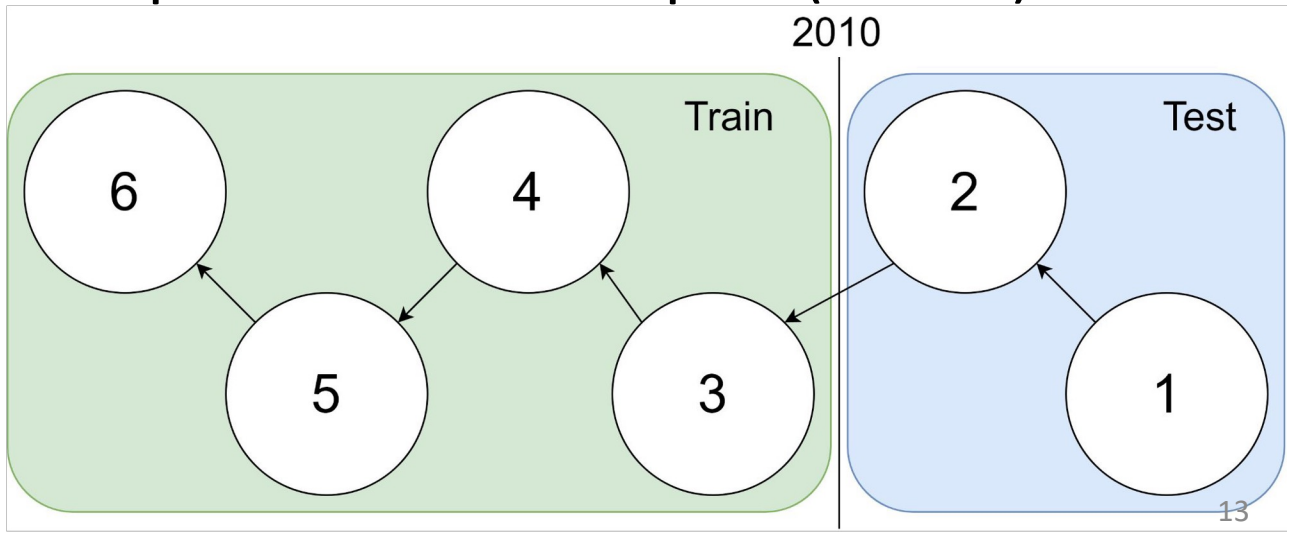
Example: 1 cites 2, 2 cites 3, … 5 cites 6

idx	Pub Year	Paper Title
1	2018	[...] Photofragment imaging
2	2016	Convenient probe of S(1D2)[...]
3	2005	Megapixel ion imaging [...]
4	2003	Direct current slice imaging [...]
5	1995	profiles of Cl(2Pj) photofragments [...]
6	1988	Adiabatic dissociation of [...]

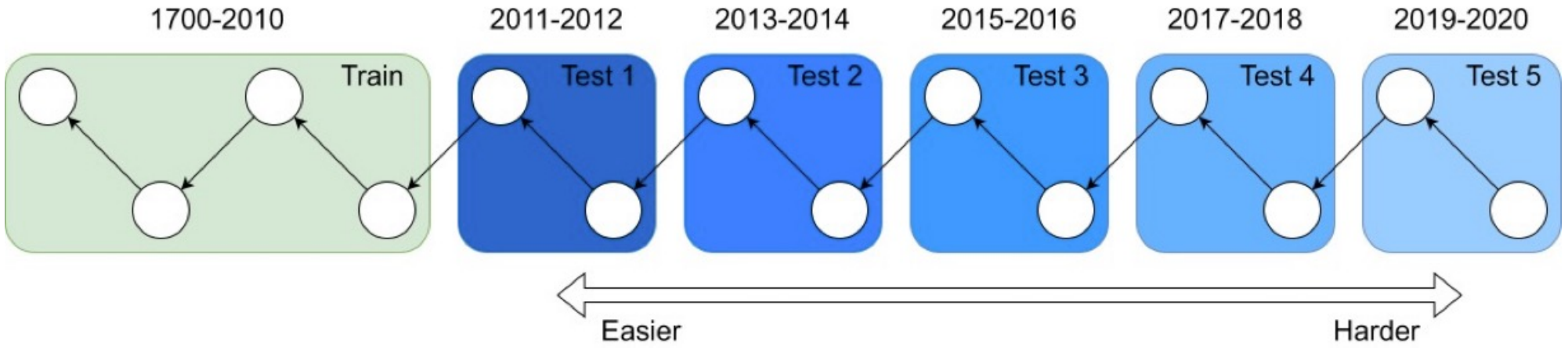
Traditional Train-Test Split (Non-Causal)



Proposed Train-Test Splits (Causal)



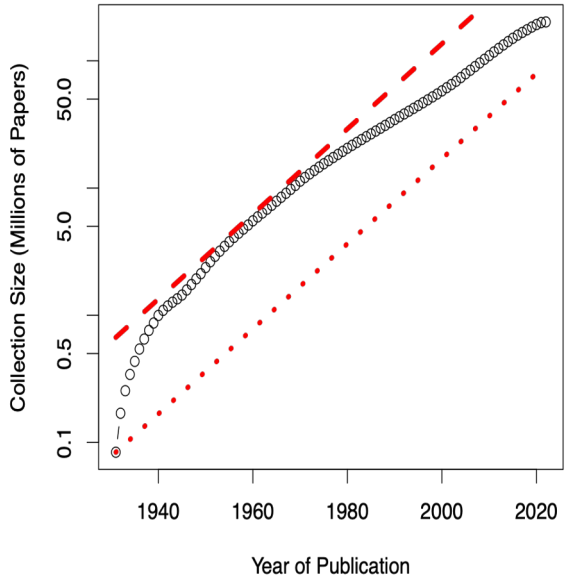
It's tough to make predictions, especially about the future
— Yogi Berra

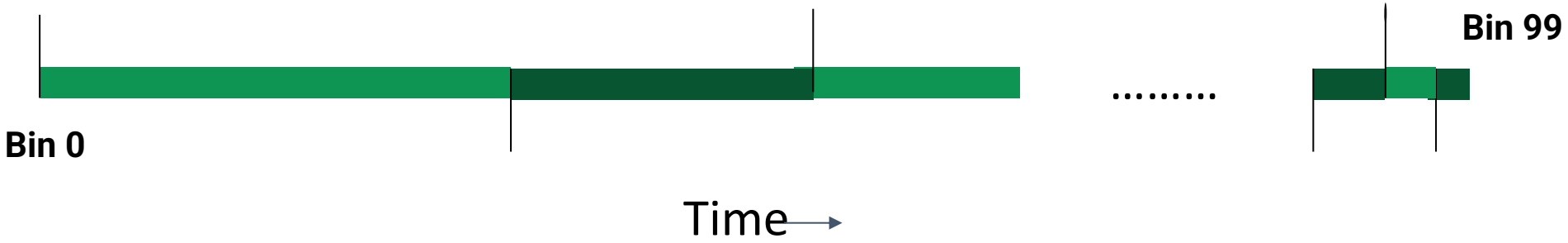


Forecasting: short-term predictions are **easier** than long-term

Assign 200M papers to 100 bins (by time)

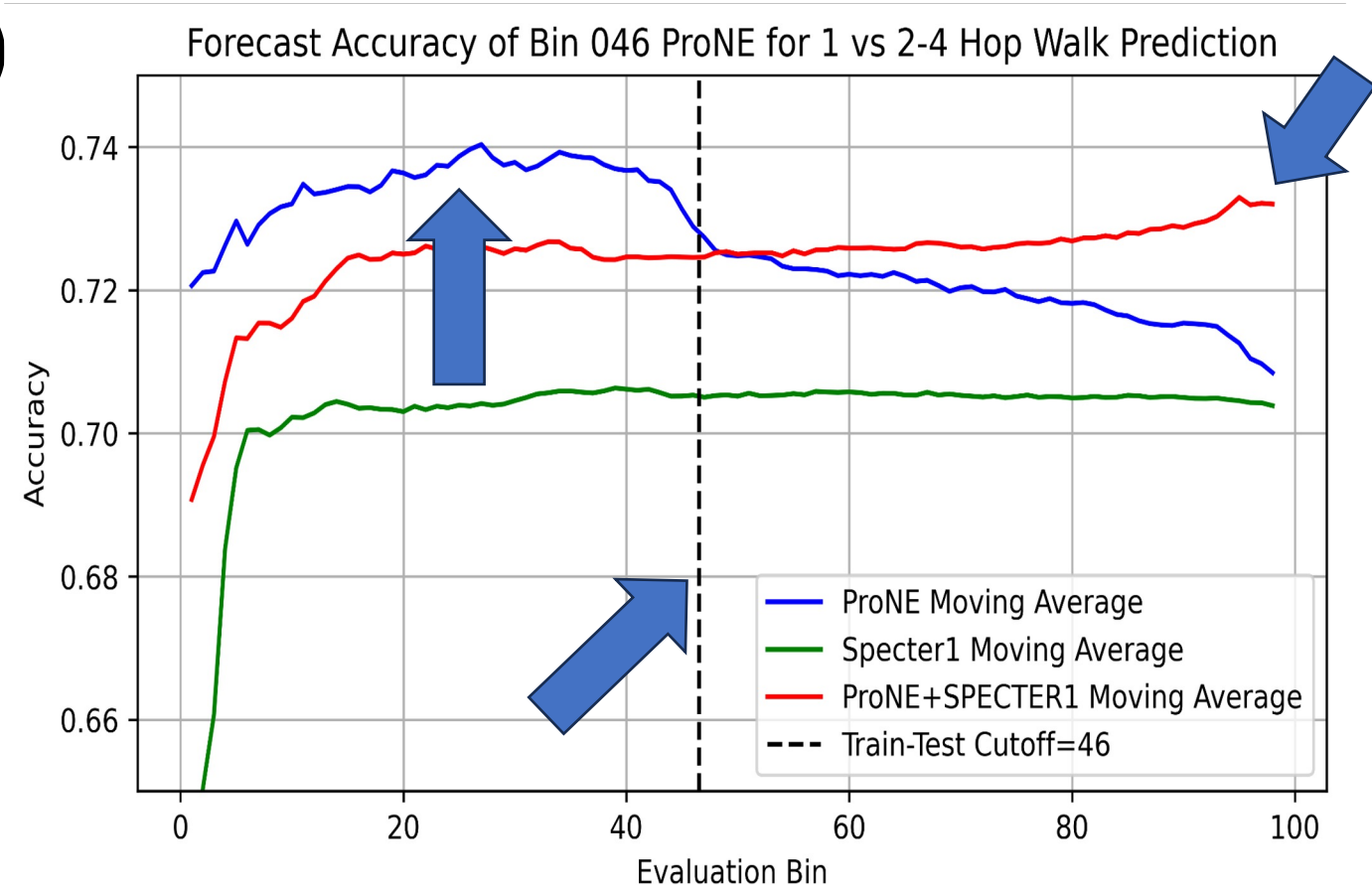
- Sort 200M papers by publication date
- Output: Split papers into 100 bins, with 2M papers/bin
- Because of exponential growth
  - (literature doubles every 9 years)
  - older bins span more time, and
  - newer bins span less time





Better Together Observations

- ProNE (Context) >> Specter (Text)
- Ensembles are better
  - than either by itself
- Text is most helpful
  - when links are not working well
  - i.e., long-term forecasting



Benchmarking Take-Aways
More Realistic Benchmarks

The Real World

- Large (Metcalfe's Law)
- Growing
- Dirty
  - Missing Values
  - Bad/misleading Values

Idealizations (Standard Benchmarks)

- Small
- Static
- Clean

Two Perspectives on Datasets

Machine Learning

- Splits: train/val/test
- Treat rows as iid
  - Identical and independently distributed
- Idealistic, but maybe unrealistic
  - Changes over time, Context, etc.
- More papers on models
  - than Datasets
- Relatively little interest in
  - Is the Dataset Realistic?
  - Motivation? (Does anyone care? If so, who?)
  - What is the dataset testing?
  - Sampling/balance? From which population?
  - Can it be gamed? Leakage?

Psychology (Presupposes ∃ Hypothesis)

- Validity
  - Content validity
  - Construct validity
  - Criterion validity
  - Face validity
  - Discriminant validity
- Reliability
  - Inter-rater reliability
  - Test-retest reliability
  - Internal reliability/Inter-item consistency
  - Split half reliability

Lexicography

Discussion of Validity
Assumes: Intention

TYPES OF VALIDITY

Experimental Validity: is the study really measuring what it intends?

INTERNAL VALIDITY refers to things that happen "inside" the study. Internal validity is concerned with whether we can be certain that it was the IV which caused the change in the DV. If aspects of the experimental situation lack validity, the results of the study are meaningless and we can make no meaningful conclusions from them.

- Internal validity can be affected by a lack of **mundane realism**. This could lead the participants to act in a way which is unnatural, thus making the results less valid.
- Internal validity can also be affected by **extraneous variables** (see below).

EXTRANEOUS VARIABLE	HOW DOES IT AFFECT VALIDITY?	HOW CAN IT BE OVERCOME?
Situational variables (anything to do with the environment of the experiment): time of day, temperature, noise levels etc	Something about the situation of the experiment could act as an EV if it has an effect on the DV. For example, poor lighting could affect participants performance on a memory test	Situational variables can be overcome by the use of **standardised procedures** which ensure that all participants are tested under the same conditions.
Participants variables (anything to do with differences in the participants): age, gender, intelligence, skill, past experience, motivation, education etc.	It may be that the differences between the participants cause the change in the DV. For example, one group may perform better on a memory test than another because they are younger, or more motivated.	Participant variables can be completely removed by using a **repeated measures design** (the same participants are used in each condition). **Matched pairs** (participants in each group are matched) could also be used.
Investigator effects: this refers to how the behaviour and language of the experimenter may influence the behaviour of the participants. The way in which an experimenter asks a question might act as a cue for the participant. Also known as **experimenter bias**	**Leading questions** from the experimenter may consciously or unconsciously alter how the participant responds. For example, the experimenter may provide verbal or non verbal encouragement when the participant behaves in a way which supports the hypothesis.	Investigator effects can be overcome by using a **double blind** technique. This is when the person who carries out the research is not the person who designed it.
Demand characteristics: participants are often searching for cues as to how to behave in an experiment. There could be something about the experimental situation or the behaviour of the experimenter (see **investigator effects**) which communicates to the participant what is "demanded" of them.	The structure of the experiment could lead the participant to guess the aim of the study. For example, participants may perform a memory test, be made to exercise, and then given another memory test. This may lead the participants to guess that the study is about the effect of exercise on memory, which may cause them to change their behaviour	When designing a study, it is important to try and create a situation where the participants will not be able to guess what the aim of the study is.
Participant effects: participants are aware that they are in an experiment, and so may behave unnaturally.	They may be overly helpful and want to please the experimenter. This leads to artificial behaviour. Alternatively, they may decide to go against the experimenter's aims and deliberately act in a way which spoils the experiment. This is the "**screw you**" effect.	Again, by designing a study so that the participants cannot guess the aims, participant effects can be reduced.

Winograd Schema (GLUE WNLI)

- The trophy doesn't fit in the brown suitcase
  - because it is too large.
- What is too large?
  - A. The trophy
  - B. The suitcase

Task	Metric	Result	Training time
CoLA	Matthews corr	56.53	3:17
SST-2	Accuracy	92.32	26:06
MRPC	F1/Accuracy	88.85/84.07	2:21
STS-B	Pearson/Spearman corr.	88.64/88.48	2:13
QQP	Accuracy/F1	90.71/87.49	2:22:26
MNLI	Matched acc./Mismatched acc.	83.91/84.10	2:35:23
QNLI	Accuracy	90.66	40:57
RTE	Accuracy	65.70	57
WNLI	Accuracy	56.34	24

Not much better than chance

Table 1. Time line of the Winograd Schema Challenge.

1972:	Winograd's (1972) thesis introduces the original example.
2010:	Levesque [47] proposes the Winograd Schema Challenge.
2010–2011:	The initial corpus of Winograd schemas is created [50].
2014:	Levesque's Research Excellence talk "On our best behavior" [48].
2016:	The Winograd Schema Challenge is run at IJCAI-16. No systems do much better than chance [16].
2018:	WNLI is incorporated in the GLUE set of benchmarks. BERT-based systems do no better than most-frequent-class guessing [91].
2019, May:	Kocijan et al. [43] achieve 72.5% accuracy on WSC273 using pretraining.
2019, June:	Liu et al. [56] achieve 89.0% on WNLI.
2019, November:	Sakaguchi et al. [77] achieve 90.1% on WSC273.

from: https://doi.org/10.1016/j.artint.2023.103971

Winograd Schema (GLUE WNLI)

A Surprisingly Robust Trick for the Winograd Schema Challenge

Vid Kocijan[1], Ana-Maria Creţu[2], Oana-Maria Camburu[1,3], Yordan Yordanov[1], Thomas Lukasiewicz[1,3]
[1]University of Oxford
[2]Imperial College London
[3]Alan Turing Institute, London
firstname.lastname@cs.ox.ac.uk, a.cretu@imperial.ac.uk

Abstract

The Winograd Schema Challenge (WSC) dataset WSC273 and its inference counterpart WNLI are popular benchmarks for natural language understanding and commonsense reasoning. In this paper, we show that the performance of three language models on WSC273 consistently and robustly improves when fine-tuned on a similar pronoun disambiguation problem dataset (denoted WSCR). We additionally generate a large unsupervised WSC-like dataset. By fine-tuning the BERT language model both on the introduced and on the WSCR dataset, we achieve overall accuracies of 72.5% and 74.7% on WSC273 and WNLI, improving the previous state-of-the-art solutions by 8.8% and 9.6%, respectively. Furthermore, our fine-tuned models are also consistently more accurate on the "complex" subsets of WSC273, introduced by Trichelair et al. (2018).

to the small existing datasets making it difficult to train neural networks directly on the task.

Neural networks have proven highly effective in natural language processing (NLP) tasks, outperforming other machine learning methods and even matching human performance (Hassan et al., 2018; Nangia and Bowman, 2018). However, supervised models require many per-task annotated training examples for a good performance. For tasks with scarce data, transfer learning is often applied (Howard and Ruder, 2018; Johnson and Zhang, 2017), i.e., a model that is already trained on one NLP task is used as a starting point for other NLP tasks.

A common approach to transfer learning in NLP is to train a language model (LM) on large amounts of unsupervised text (Howard and Ruder, 2018) and use it, with or without further fine-tuning, to solve other downstream tasks. Building on top of a LM has proven to be very suc-

View PDF

ELSEVIER
Artificial Intelligence
Available online 11 July 2023, 103971
In Press, Corrected Proof What's this?

The defeat of the Winograd Schema Challenge

Vid Kocijan [a 1], Ernest Davis [b], Thomas Lukasiewicz [c d], Gary Marcus [e], Leora Morgenstern [f]

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.artint.2023.103971 Get rights and content

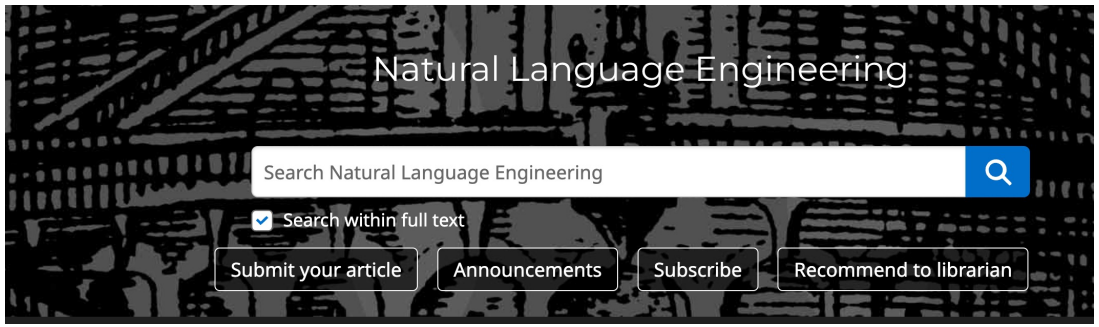
Abstract

The Winograd Schema Challenge—a set of twin sentences involving pronoun reference disambiguation that seem to require the use of commonsense knowledge—was proposed by Hector Levesque in 2011. By 2019, a number of AI systems, based on large pre-trained transformer-based language models and fine-tuned on these kinds of problems, achieved better than 90% accuracy. In this paper, we review the history of the Winograd Schema Challenge and discuss the lasting contributions of the flurry of research that has taken place on the WSC in the last decade. We discuss the significance of various datasets developed for WSC, and the research community's deeper understanding of the role of surrogate tasks in assessing the intelligence of an AI system.

Keywords

Commonsense reasoning; Winograd Schema Challenge

Shameless Plug: Smooth-Talking Machines



Most read

This page lists the top ten most read articles for this journal based on the number of full text views and downloads recorded on Cambridge Core over the last 30 days. This list is updated on a daily basis.

Emerging trends: Smooth-talking machines
Kenneth Ward Church, Richard Yue
Published online by Cambridge University Press: 11 September 2023, pp. 1402-1410
Article · Access · Open access · PDF · HTML · Export citation
View abstract

GPT-3: What's it good for?
Robert Dale
Published online by Cambridge University Press: 15 December 2020, pp. 113-118
Article · Access · Open access · PDF · HTML · Export citation
View abstract

Word2Vec
KENNETH WARD CHURCH
Published online by Cambridge University Press: 16 December 2016, pp. 155-162
Article · Access · Open access · PDF · HTML · Export citation
View abstract

Challenge: Fluency ≠ Trustworthiness

- A number of articles on ChatGPT
  - lead with amazing successes
    - that seem too good to be true,
  - and end with back-peddling
- ChatGPT has
    - amazing strengths (fluency)
  - as well as
    - amazing weaknesses (trustworthiness)

- Many people assume
  - Fluency ≈ Intelligence
  - IQ testing:
    - Measure vocabulary size
    - Large vocabulary → Fluent
- Fluency is particularly important
  - on first impression
- But eventually → disappointment
  - Weaknesses will become clear
- What is the difference between
  - a hallucination and a con?

What should we do next?

- Three paths forward:
  - Low road:
    - Give up (hallucinations)
  - Middle road:
    - Fact-checking with search
  - High road:
    - Revive rationalism (“AI Complete”)
    - Minsky & Chomsky

- Recommendations
  - Short-term:
    - Middle Road: Search
    - “Good apps for Crummy MT”
      - Find apps for what we have
      - given strengths and weaknesses
  - Long-term:
    - High road may be necessary
      - But it is very ambitious
    - Inclusiveness:
      - Interdisciplinary Collaboration
      - Growth opportunities
        - (Low Resource Languages)

Hypothesis:
NLP History → Strengths & Weaknesses

- Deep Nets are
  - more fluent
  - than trustworthy
- Pendulum Swung Too Far (2011)
  - Empiricism (1950s)
  - Rationalism (1970s)
  - Empiricism (1990s)
  - Deep Nets (2010s)

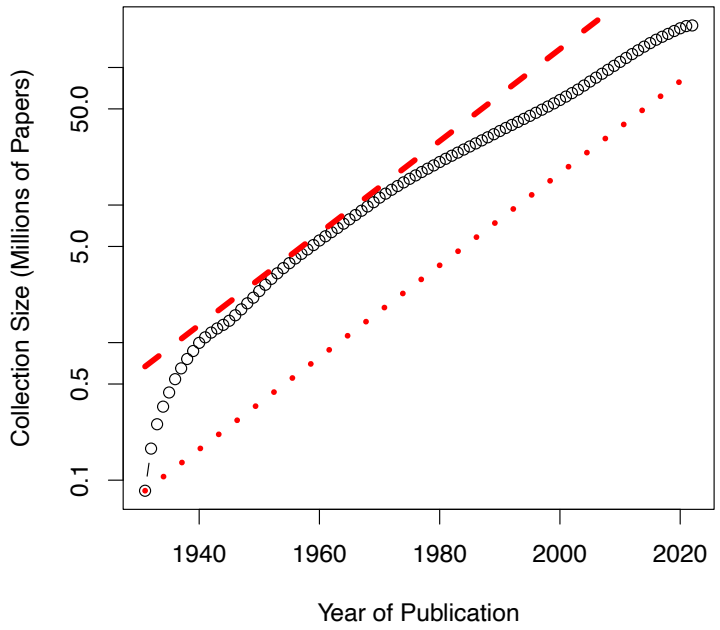
Truth

Fluency

Massive Growth → Mistaken Impression that everything is new (and there is no history)

Scientific Literature doubles every 9 years (90% written since I started PhD)

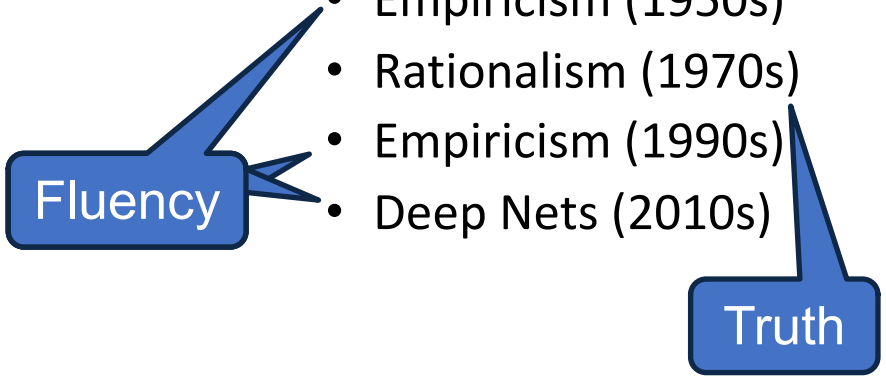
- What's new
  - The world is taking notice (in AI)
  - Fluency is much improved
- What's not new
  - Chatbots (and much of the tech)
  - SOTA-Chasing
  - New/better shiny objects
    - (with same old quality?)
    - Does SOTA-Chasing → Progress?
  - Trustworthiness is still wide open



Personal History

- Strengths (fluency) and weaknesses (trustworthiness)
  - may be a consequence of choices we made in 1990s
- We started EMNLP in 1990s for pragmatic reasons
  - Field had been attempting to do too much
    - and was accomplishing too little
    - (during a funding winter)
- We chose to stop working on hard problems
  - (trustworthiness)
  - in order to make relatively quick progress on fluency
  - by reviving empirical methods from the 1950s
    - (Shannon, Skinner, Firth)

- Deep Nets are
  - more fluent
  - than trustworthy
- Pendulum Swung Too Far (2011)
  - Empiricism (1950s)
  - Rationalism (1970s)
  - Empiricism (1990s)
  - Deep Nets (2010s)



Pendulum Swung Too Far

ChatGPT's strengths (fluency) and weaknesses (trustworthiness) may be a consequence of choices we made in 1990s

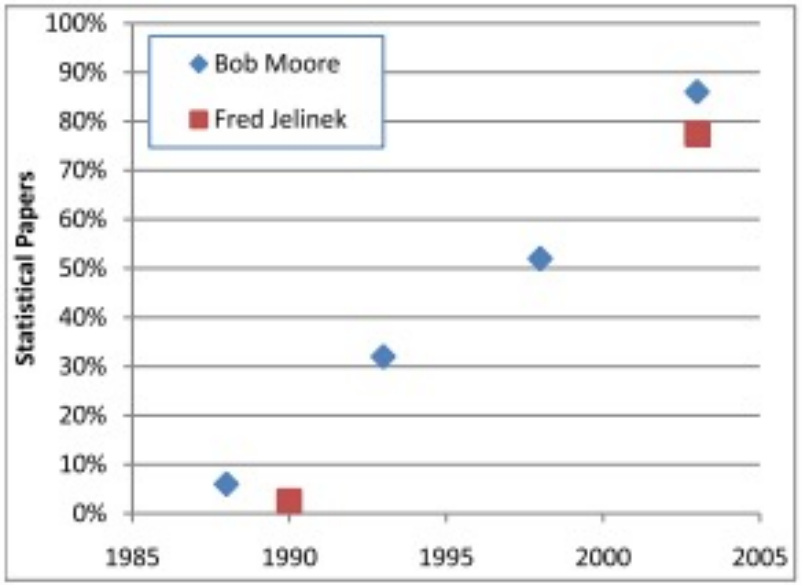


FIGURE 1 The shift from Rationalism to Empiricism is striking (and no longer controversial). This plot is based on two independent surveys of ACL meetings by Bob Moore and Fred Jelinek (personal communication).

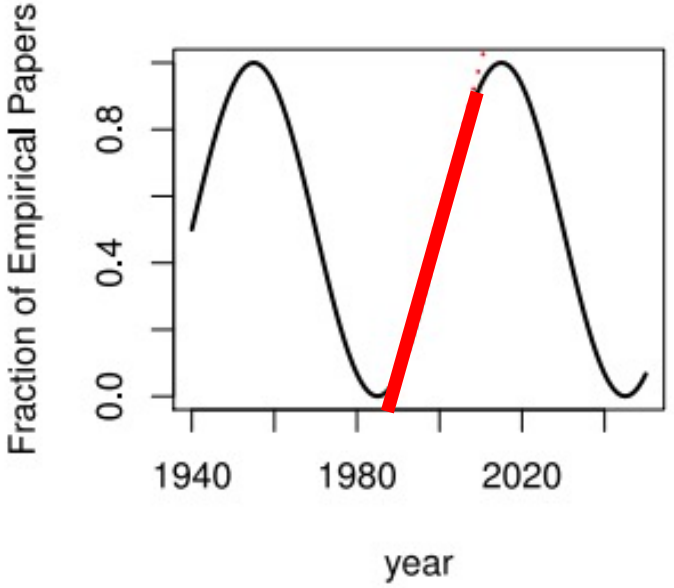


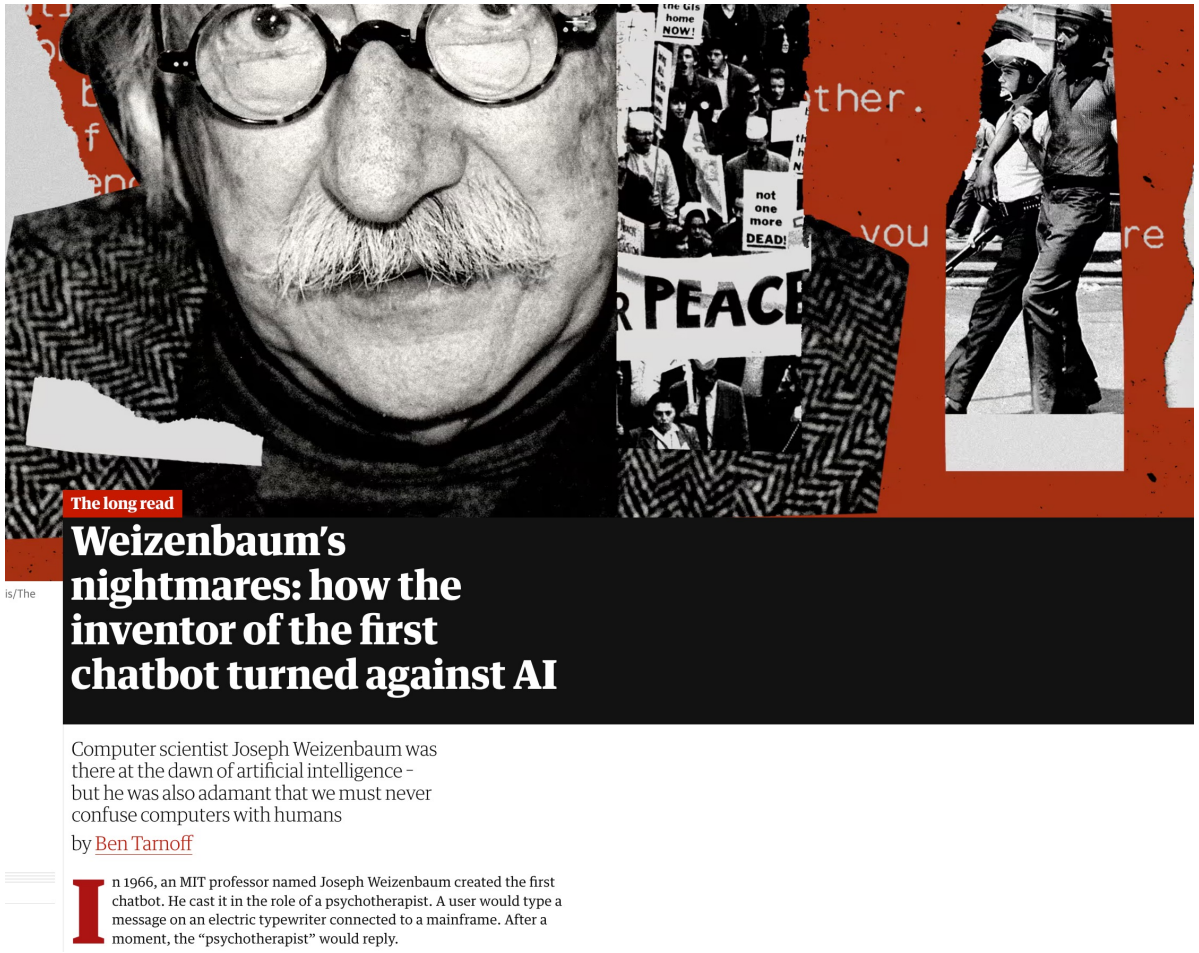
FIGURE 2 An extreme view of the literature, where the trend in Figure 1 (denoted by a dashed red line) is dominated by the larger oscillation every couple of decades. Note that that line is fit to empirical data, unlike the oscillation which is drawn to make a point.

- 1950s: Empiricism (Shannon, Skinner, Firth, Harris)
- 1970s: Rationalism (Chomsky, Minsky)
- 1990s: Empiricism (IBM Speech Group, AT&T Bell Labs)

Mechanism

We started EMNLP

History does not repeat itself (but it rhymes)



- Recent article in Guardian
  - Compared ChatGPT and Weizenbaum's ELIZA
  - https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai
- As a TA for Weizenbaum (1978),
  - I know just how horrified he was
  - by how seriously people took ELIZA
- Responsible AI
- His views were not popular at MIT
  - at the time

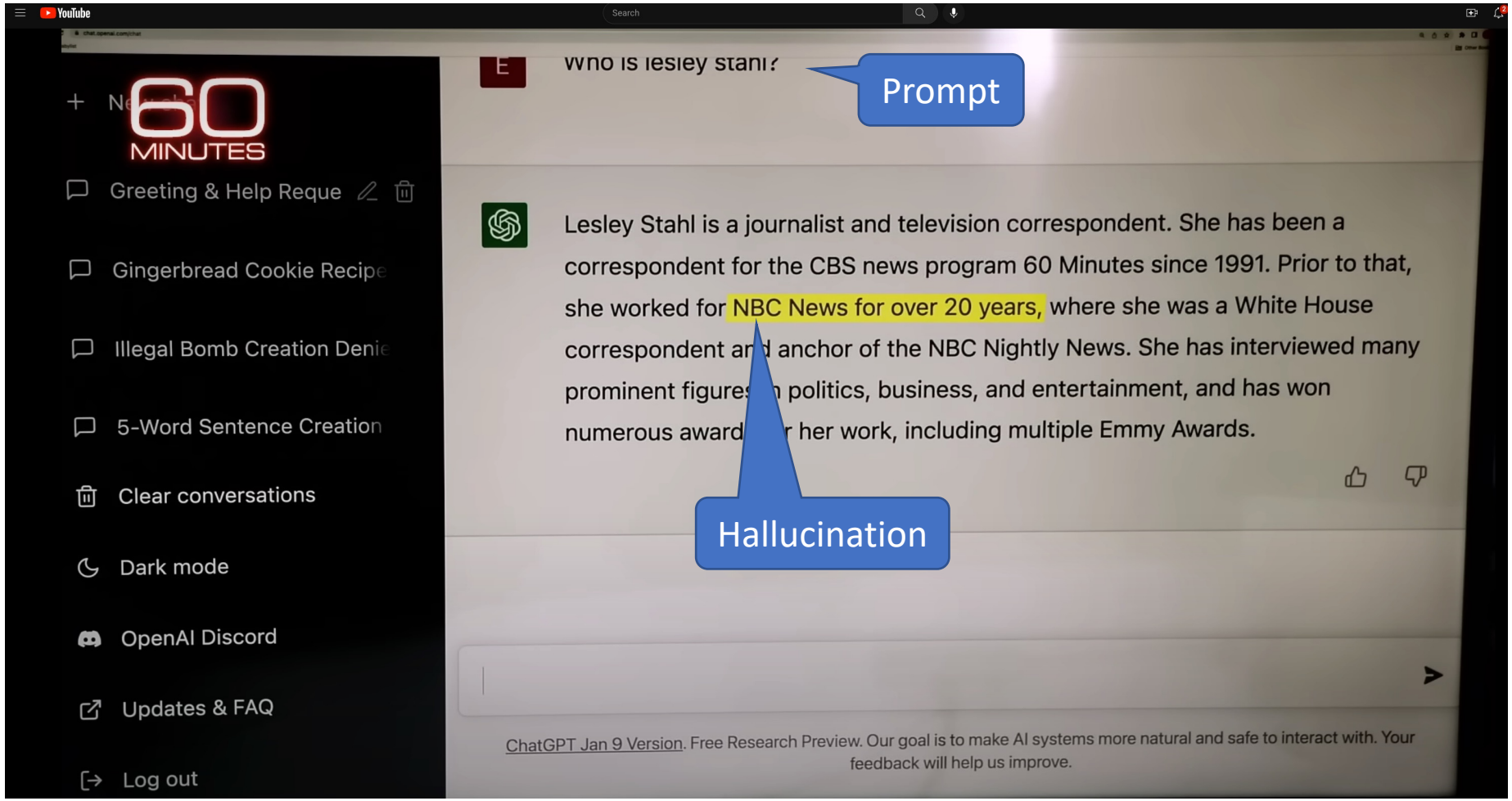
How Hard are Hallucinations? Yogi Berra

- Possible answers:
  - Soon: “See next release”
  - Eventually, but not soon:
    - “Next Year in Jerusalem”
    - History of Machine Translation (MT)
      - When will MT be practical?
      - Prediction from 1950s: 5 years
  - Never

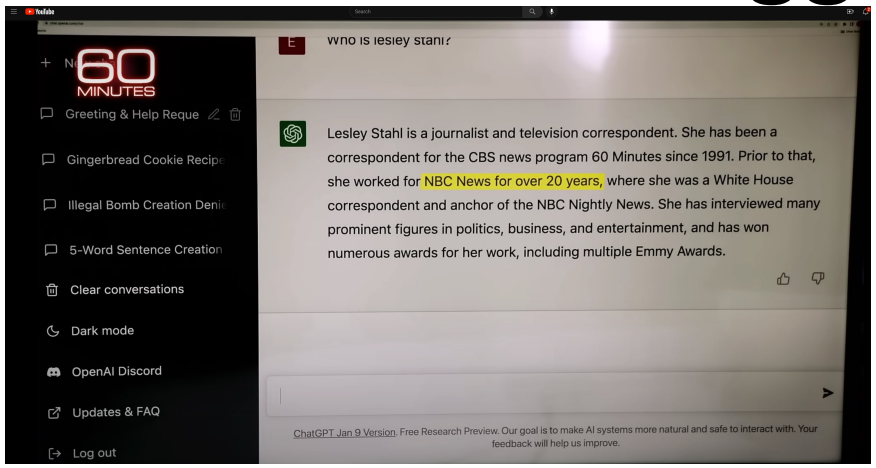
- Assuming ChatGPT’s strengths & weaknesses
  - are a consequence of our choices from 1990s
  - and it took three decades to do well on fluency
  - and fluency ≪ trustworthiness
- then not soon

ChatGPT Hallucinates on CBS ``60 Minutes''

https://www.youtube.com/watch?v=1wzPr4cUoMQ&t=463s



Constructive Suggestions for Hallucinations



1. Low Road:
   - Give up
2. Middle Road
   - Fact-checking with search
3. High Road
   - Revive Rationalism

Query: *Lesley Stahl works for which company*

Challenges for Fact-Checking

- Which claims need to be checked?
- How do we create queries?
- When we get search results, then what?

Acronyms: A Simple Case for Fact-Checking

- Acronyms are easier
  - Google Translate is better on long forms (LFs) than short forms (SFs)
    - Use Google to translate LFs to target language (English)
    - Generate candidate SFs in target
    - Use search to verify candidates
- Co-author (Richard Yue)
  - used to be a professional translator
- Metrics matter:
  - Terminology:
    - important to translators
    - (But less so for BLEU)

- *De nombreux facteurs de risque participent au développement de cette pathologie, parmi lesquels les acides gras trans (AGT).*
- *... une diminution d'expression de 12 gènes mutés dans l'anémie de Fanconi (AF)*

Table 1. Opportunity for Fact-Checking: Translation of Acronyms

Input French		Output English		
LF	**SF**	**LF**	**SF (gold)**	**SF (Google)**
Acides gras trans	AGT	Trans fatty acids	TFA	TGA
Anémie de Fanconi	AF	Fanconi Anemia	FA	AF

Table 2. Search will find more documents matching the good combinations than the bad combinations

Good Combinations	Bad Combinations
Trans fatty acids (TFA)	Trans fatty acids (TGA)
Fanconi Anemia (FA)	Fanconi Anemia (AF)

Fact-Checking Take-Aways

- Short-term patch for hallucinations
- Acronyms:
  - Easy special case of hallucinations
- Don’t re-invent the wheel
  - Re-use existing tools: Search
- Standard Loss/Metrics
  - Insufficient Penalty for Fatal Errors
    - Terminology
    - Responsible AI

- Agreement:
  - hyp = gold
- Verification:
  - Search finds 2+ examples in publications

Method	Agreement	Verified
Identity Baseline	21.5%	0.06%
Reverse Baseline	28.5%	14.6%
Google Baseline	54.3%	29.2%
Gold Labels	100%	42%
Proposed (Section 2.3)	**62.6%**	**42.8%**

Table 7: Proposed method outperforms baselines.

Simple Example: Arithmetic
(Chomsky: Capturing Relevant Generalizations)

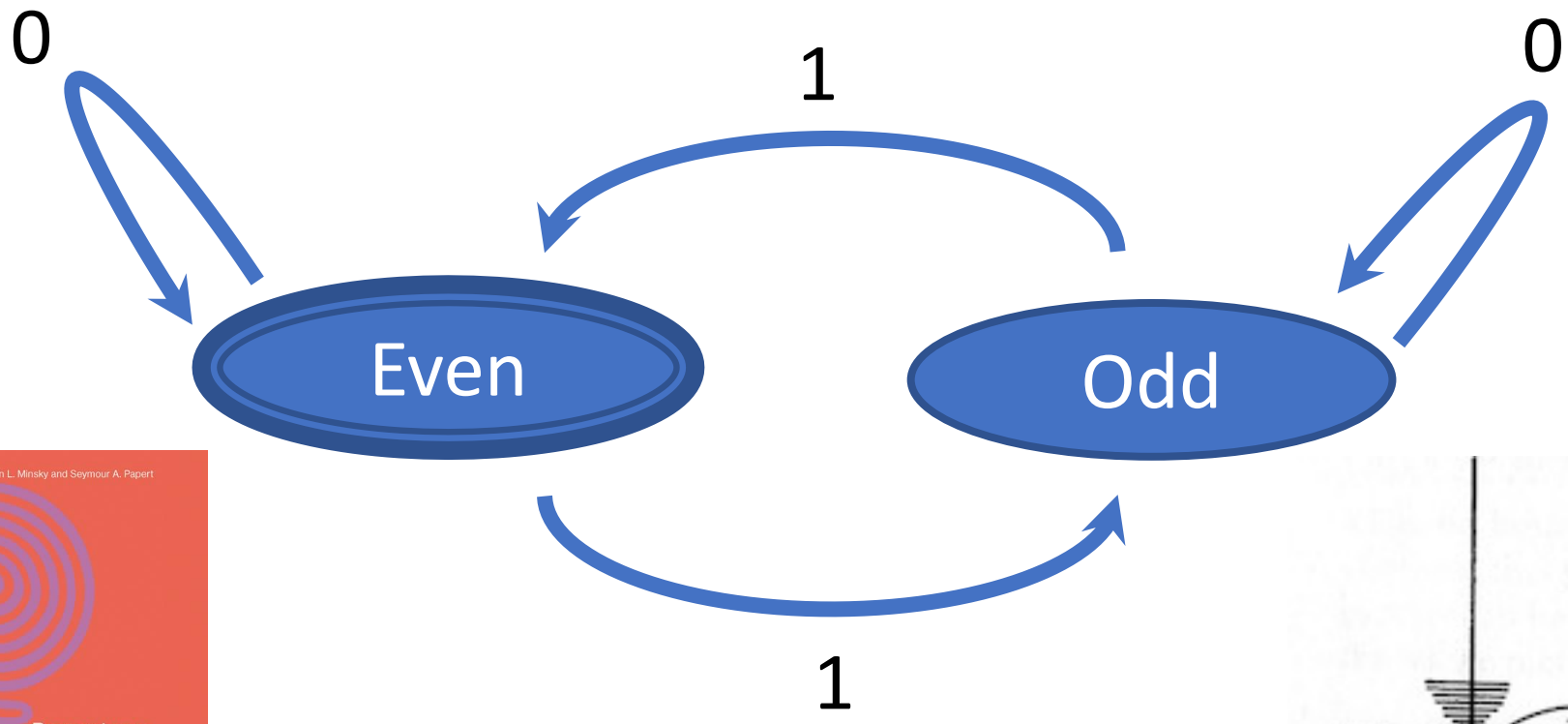
Hypothetical

- Suppose ChatGPT
  - can add two small numbers
- But for large numbers,
  - it makes up answers
- After each release,
  - “small” gets “bigger”

Is this progress?

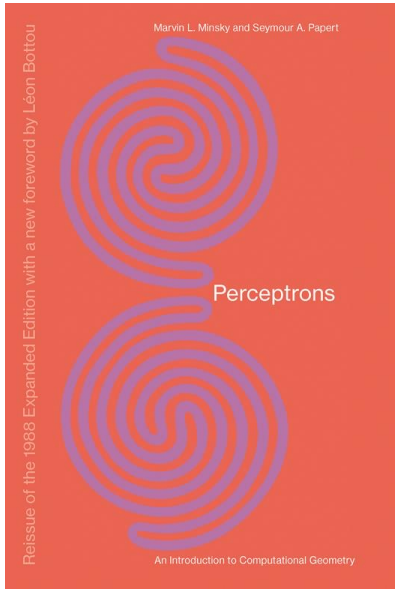
- Empiricism
  - According to SOTA-Chasing,
    - Yes
- Rationalism
  - According to Minsky & Chomsky,
    - No
    - ChatGPT is not mastering concepts
    - “Stochastic Parrots”

Perceptron: "Can't Compute Parity" (Capturing Relevant Generalizations)

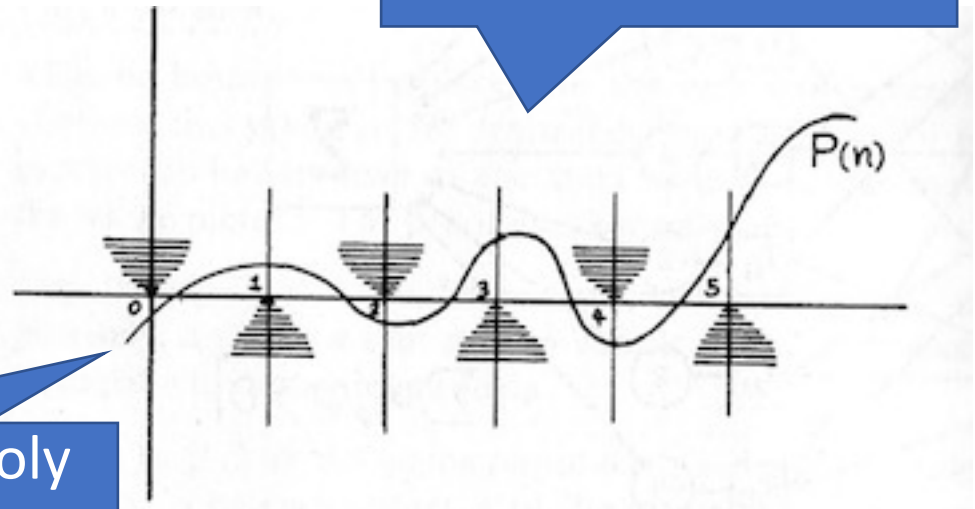


All Cases in $O(1)$ Space

Some Cases in $\Omega(n)$ Space

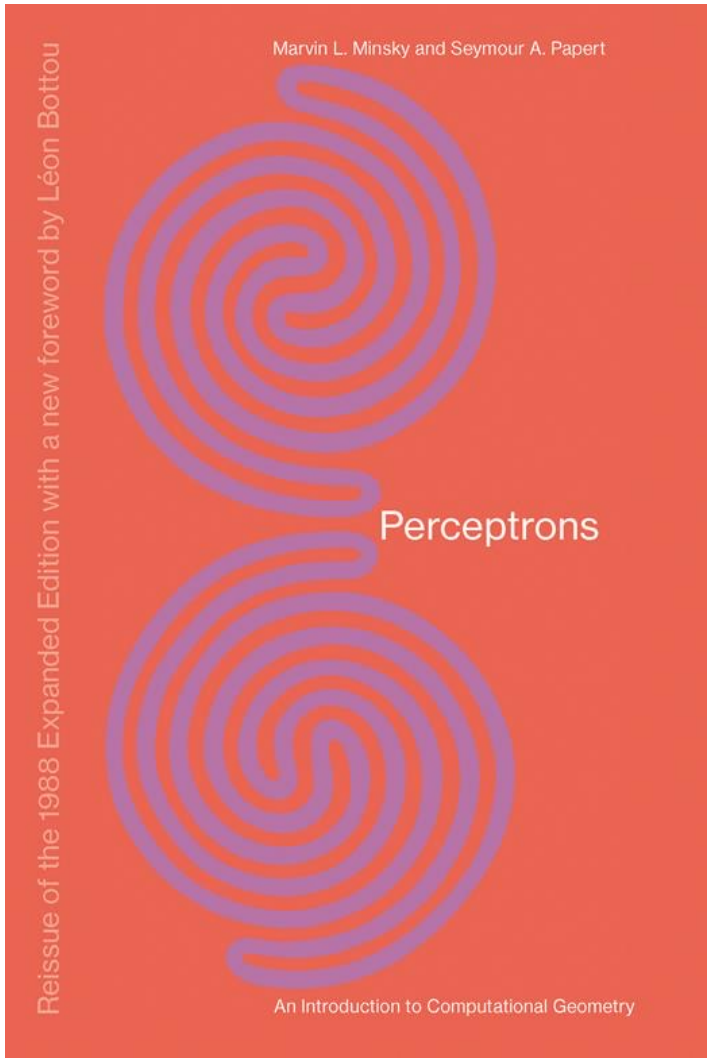


Minsky rejects deep nets (1969)



Regression: n^{th} order poly requires n parameters

Minsky rejects deep nets (1969)



N. Chomsky, "Three models for the description of language," in *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 113-124, September 1956, doi: 10.1109/TIT.1956.1056813

Whatever the other interest of statistical approximation in this sense may be, it is clear that it can shed no light on the problems of grammar. There is no general relation between the frequency of a string (or its component parts) and its grammaticalness. We can see this most clearly by considering such strings as

(14) colorless green ideas sleep furiously

Introduced Chomsky Hierarchy:
Finite State → Turing Machines

Capturing Generalizations Argument

Pros

- Long-Term Focus
  - Can’t get to moon by
  - Incremental short-term local optimization

Cons

- Dismisses Short-term Progress
  - Sometimes it is useful
  - to solve some simple special cases
    - (addition of small numbers)
- Not constructive:
  - “MIT School of Negativity”

Personal History

- ChatGPT's strengths (fluency) and weaknesses (trustworthiness)
  - may be a consequence of choices we made in 1990s
- We started EMNLP in 1990s for pragmatic reasons
  - Field had been attempting to do too much
    - and was accomplishing too little
    - (during a funding winter)
- We chose to stop working on hard problems (trustworthiness)
  - in order to make relatively quick progress on fluency
  - by reviving empirical methods from the 1950s (Shannon, Skinner, Firth)

How Hard are Hallucinations?

- Possible answers:
  - Soon: “See next release”
  - Eventually, but not soon:
    - “Next Year in Jerusalem”
    - History of Machine Translation
  - Never

- Assuming ChatGPT’s strengths & weaknesses
  - are a consequence of our choices from 1990s
  - and it took three decades to do well on fluency
  - and fluency ≪ trustworthiness
- then not soon

Ugly: Outline

- ✓ Benchmarking (Ken)
- ➢ **Labeling (Omar)**
- Bias, toxicity, misinformation, plagiarism (Ken)
- 1 slide on hard problems for researchers/practitioners

Ugly - Labeling

The bad news first

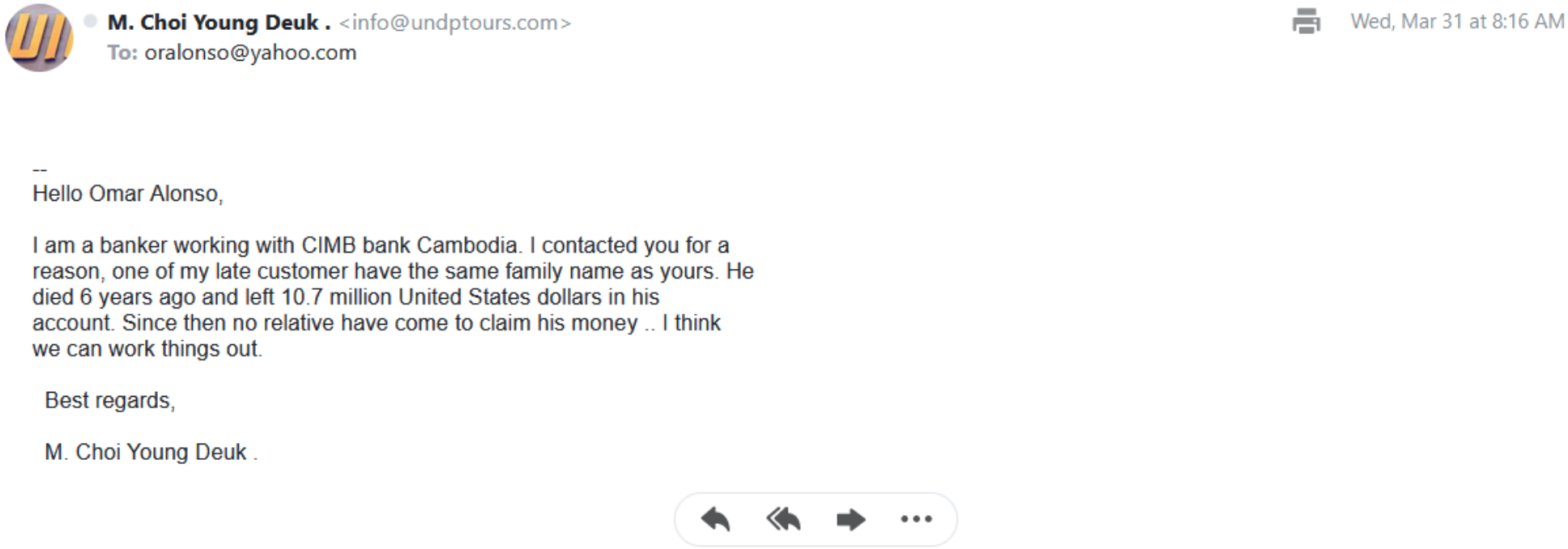
- Labeling is hard
  - Facebook
  - Points-Of-Interests (Foursquare, etc.)
- Labeling is going to get more difficult
  - Enterprise
  - Personalization
  - Healthcare
  - New data sets
- Some context
  - We assume supervised or semi-supervised learning
  - Large scale
  - Continuous

There is hope

- Human Computation
  - Artificial Intelligence
  - Computer Science
  - Human-Computer Interaction
  - Economics
  - Behavioral sciences
- Lots of research and new ideas
- This section
  - Programming perspective, quality framework, data pipelines, future trends

https://www.humancomputation.com/

What is a label?



M. Choi Young Deuk . <info@undptours.com>
To: oralonso@yahoo.com

Wed, Mar 31 at 8:16 AM

--
Hello Omar Alonso,

I am a banker working with CIMB bank Cambodia. I contacted you for a reason, one of my late customer have the same family name as yours. He died 6 years ago and left 10.7 million United States dollars in his account. Since then no relative have come to claim his money .. I think we can work things out.

Best regards,

M. Choi Young Deuk .

Spam email?

Label: yes, no

Why we need labels?

- Information retrieval
- Natural language processing
- Machine learning
- Artificial intelligence

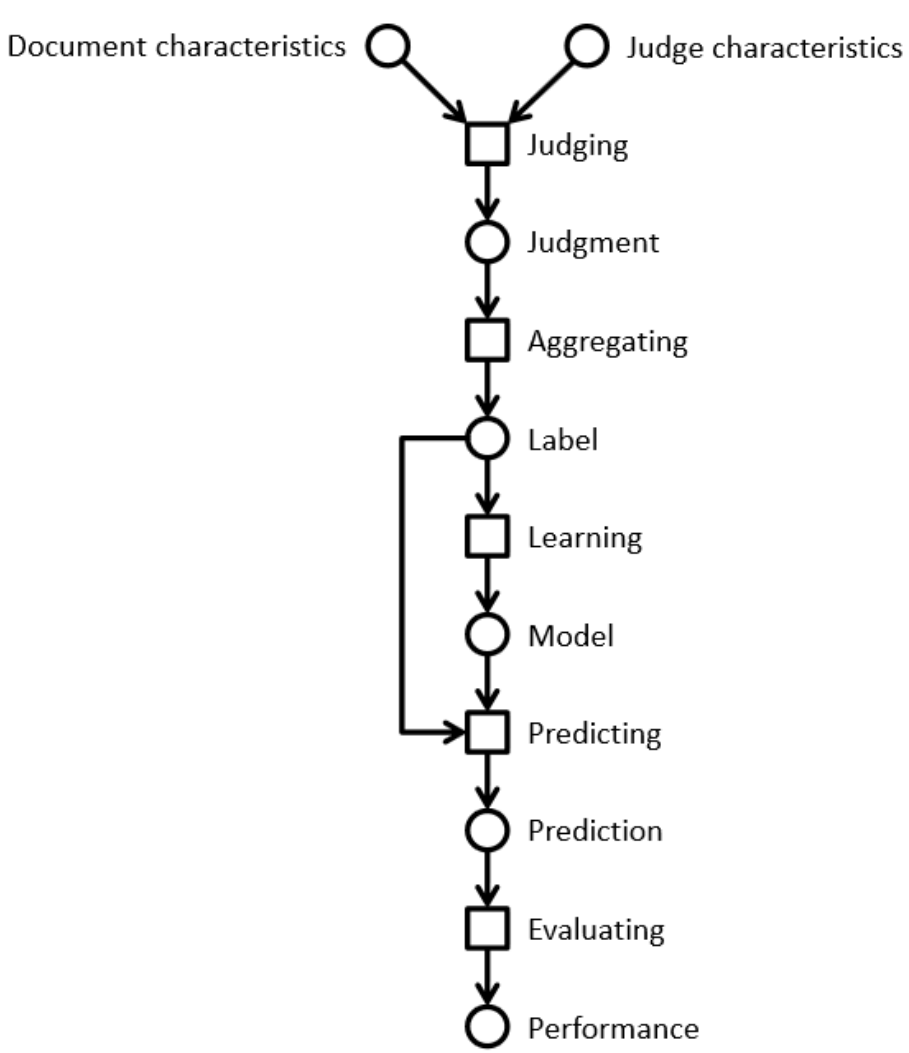
Why we care?

- Provenance
- Reproducibility & debugging
- Explainability & interpretability
  - How a training set was created
- Bias and fairness
- Data management
  - ML/AI models live & die by the quality of input data
  - Metadata about labels
  - Maintenance

Lora Aroyo et al. "Data excellence for AI: why should you care?" Interactions 29(2): 66-69 (2022)

Lifecycle of a label

- Information retrieval example



Using a crowd to label a data set

Using ML to process the complete data set

Relevance labels

- Indicate whether a search result is valuable to a searcher
- Key in evaluation and optimization IR systems
- Editors or experts
  - TREC-style
- Crowdsourcing
- LLMs

Careful with that ~~axe~~ data, Eugene

- In the era of big data and machine learning
  - labels -> features -> predictive model -> optimization
- Labeling perceived as boring
- Tendency to rush labeling
- Quality is key
  - Garbage in, garbage out
- Own the entire stack
  - Labeling, modeling, infrastructure, deployment

The state of the field

- Human-labeled data is more important than ever
- Requirements
  - Throughput -> ASAP; I need the labels for yesterday
  - Cost -> cheap; if possible free
  - Quality -> top
- Performed as a one-off by 3rd party (crowd or editors)
  - Human Intelligence Task (HIT)
  - Micro-tasks
- Needs development work to get good results
- Very limited functionality in current platforms
  - Mechanical Turk, SageMaker (Amazon)
  - Figure Eight (Appen)
  - Toloka (Yandex)
  - Start-ups
- LLMs

The need for humans

- Many examples where humans are involved
- Adult content and moderation
- Baby sitting algorithms

https://www.wired.com/2014/10/content-moderation/

https://onezero.medium.com/a-permanent-nightmare-pinterest-moderators-fight-to-keep-horrifying-content-off-the-platform-4d8e7ec822fe

https://www.thebureauinvestigates.com/stories/2022-10-20/behind-tiktoks-boom-a-legion-of-traumatised-10-a-day-content-moderators

Problems

- Monolithic HITs
  - The structure of a HIT mirrors the structure of the task the developer is working on
  - Similar to Conway's law in software engineering
- Task complexity
- Lengthy instructions
  - RTFM doesn't work
- We don't think of HC/crowdsourcing as programming
- How to improve
  - Use established programming practices
  - Careful, we are dealing with humans and not machines

A spectrum of labeling tasks

Nature of task	Aggregation approach	Evaluation technique
Objective question has a correct answer (objective)	Reliable judge assigns appropriate label for an item	Evaluate workers by comparing individual results with gold set
Judgment question has a best answer (partially objective)	Inter-rater agreement determines label for an item	Evaluate workers by comparing individual results with consensus
Subjective question has consistent answer (subjective)	Repeatable polling determines probability of a label for an item	Evaluate workers by computing the consistency of results between groups

Prepare the environment

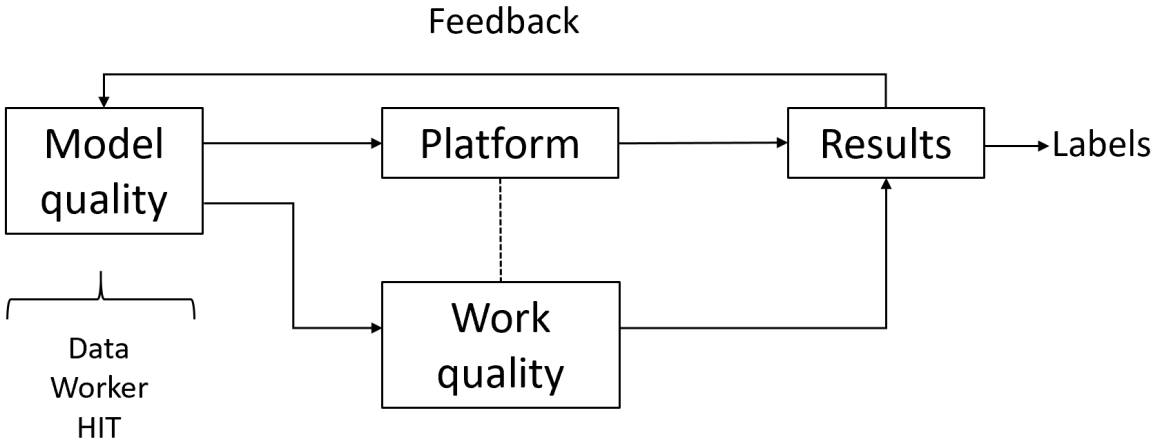
- Homework before you label
  - Assess the lay of the land
  - Identify your use cases
  - Understand your product’s data
  - Design your HITs
  - Determine your guidelines
  - Communicate your task
  - Maintain high quality
- Ongoing vs. one-offs HITs
- Labels for the machine != labels for humans

HIT design principles

- Self-contained, short, and simple
- Document presentation
  - Text alignment & legibility; reading level; multi-cultural and multilingual
- Cognitive biases
  - Implications on the final output: anchor effect, mere exposure, picture superiority
- Task complexity
  - High cognitive load; low usability, specific expertise

Quality control in general

- Extremely important part of the task
- Approach as “overall” quality; not just for workers
- Bi-directional channel
  - You may think the worker is doing a bad job.
  - The same worker may think you are a lousy requester.
- Quality framework
  - Module quality
  - Work quality
- Measuring agreement



Algorithms used in practice

- Voting
  - Majority vote, Borda, tiers
  - Strong baseline
- Honey pots and programmatic gold
- Expectation-Maximization
- Get another label
- Adaptivity
  - Quality-cost tradeoff
  - How many workers?
  - When to stop?
  - Stopping rules
  - Automatic honey pots creation

Behavioral features

- Focus on the way workers work instead of what they produce
- Task fingerprinting
- High correlation with work quality
- Wernicke
  - Information Extraction scenario
  - Weighted majority voting
  - Behavioral features outperform performance-based methods

J. Rzeszotarski and A. Kittur. “Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance”. UIST 2011.

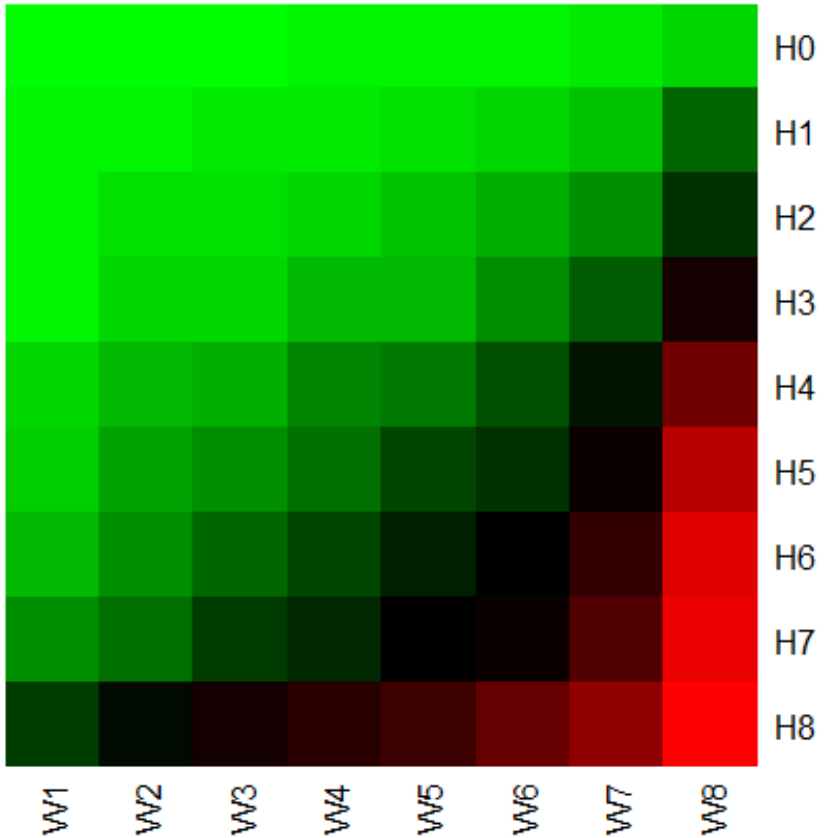
S. Han, P. Dai, P. Paritosh, D. Huynh. “Crowdsourcing Human Annotation on Web Page Structure: Infrastructure Design and Behavior-Based Quality Control”. ACM TIST 2016

Active learning

- Accuracy
  - Limited budget for annotating a small % of the unlabeled data
- Speed
  - Model more accurate more quickly
- Diversity
- Uncertainty sampling
  - Least confidence, margin of confidence, ratio of confidence
- Diversity sampling
  - Clustering to partition the data, real-world diversity

R. Munro. *"*Human-in-the-Loop Machine Learning", Manning Publications 2021.

Error rates for different worker/HIT groups



Easy HITs

Difficult HITs

Good workers

Bad workers

UHRS

2,700 HITs from 20 workloads

For difficult HITs

- Good workers are doing well
- Bad workers are doing poorly

For easy HITs

- Good workers are doing well
- Bad workers are doing well

Let error(W*i*, H*j*) be the average error rate of the workers in the worker group W*i* working on HITs H*j*

I. Abraham et al. “How Many Workers to Ask? Adaptive Exploration for Collecting High Quality Labels”. SIGIR 2016

Snorkel approach

- Formalizing programmatic labeling
- Models are commodities
  - `pip install <what-you-want>`
- Training data is the interface for software 2.0
- Labeling functions as black boxes that predict a label
- Learn from agreements/disagreements between labeling functions

LLMs

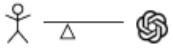
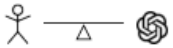
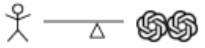
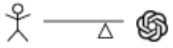
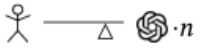
- Human labels are expensive
  - Expert > Crowd-based worker > LLM
  - Automatic label is not a new idea
- How about using LLMs to label documents?
- Potential advantages
  - Cost and performance
  - Allocate humans where are needed the most
- Potential issues
  - Reliability
  - Quality control

P. Thomas et al. "Large language models can accurately predict searcher preferences" arxiv.org/abs/2309.10621

G. Faggioli et al. "Perspectives on Large Language Models for Relevance Judgment" ICTIR 2023

Spectrum of human-machine collaboration

- LLMs judgement quality
- LLMs cost
- Multiple LLMs as judges
- Truthfulness
- Bias
- Explanations/justifications

Collaboration Integration	Task Organization
Human Judgment	
	The human will do all judgments manually without any kind of support.
	Humans have full control of judging but are supported by text highlighting, document clustering, etc.
AI Assistance	
	The human assessor judges an LLM-generated summary of the document.
	Balanced competence partitioning. Humans and LLMs focus on tasks they are good at.
Human Verification	
	Two LLMs each generate a judgment, and a human selects the better one.
	An LLM produces a judgment (and an explanation) that a human can accept/reject.
	LLMs are considered crowdworkers, varied by specific characteristics, and controlled by a human.
Fully Automated	
	Fully automatic assessment.

Prompting

- In-context learning
- New capabilities can be unlocked in LLMs
- LLM is prompted with a few in-context demonstrations
- Learns to perform a certain task
- Task performance is very sensitive to prompts

Setup

- Similar to crowdsourcing work
- Take TREC judgement guidelines
  - HIT in Mturk
  - Prompt for GPT or similar LLM
- Compute agreement using Cohen’s kappa
- Two main approaches
  - Prompt “as is”
  - Prompt engineering

Prompt structure

- Relevance evaluation task
- Task instructions
  - You are a search quality rater evaluating relevance of web pages
- Query-document pair to be labelled
  - Query {query}
  - Document {document}
  - Relevant?
- Re-state the task
- Output format

Discussion

- In favor
  - LLMs are able to produce an explanation
  - This could be used to assist humans in relevance judgements
- Against
  - LLMs are not users
  - IR is about relevance to an information need
  - No proof that evaluation by LLM has any relationship to reality

The main process is unchanged

- Regardless if labeling is done by machines or humans
- Three main components
  - Task design
    - HIT or prompt engineering
  - Data
  - Crowd
    - Human-based crowd or LLM-based crowd
- Quality control
- Debugging

Prompts compared to HITs

role — You are a search quality rater evaluating the relevance of web pages. Given a query and a web page, you must provide a score on an integer scale of 0 to 2 with the following meanings:

2 = highly relevant, very helpful for this query
1 = relevant, may be partly helpful but might contain other irrelevant content
0 = not relevant, should never be shown for this query

Assume that you are writing a report on the subject of the topic. If you would use any of the information contained in the web page in such a report, mark it 1. If the web page is primarily about the topic, or contains vital information about the topic, mark it 2. Otherwise, mark it 0.

Query
A person has typed [*query*] into a search engine.

description, narrative — They were looking for: *description narrative*

Result
Consider the following web page.

—BEGIN WEB PAGE CONTENT—

page text

—END WEB PAGE CONTENT—

Instructions
Split this problem into steps:

Consider the underlying intent of the search.

aspects — Measure how well the content matches a likely intent of the query (M).

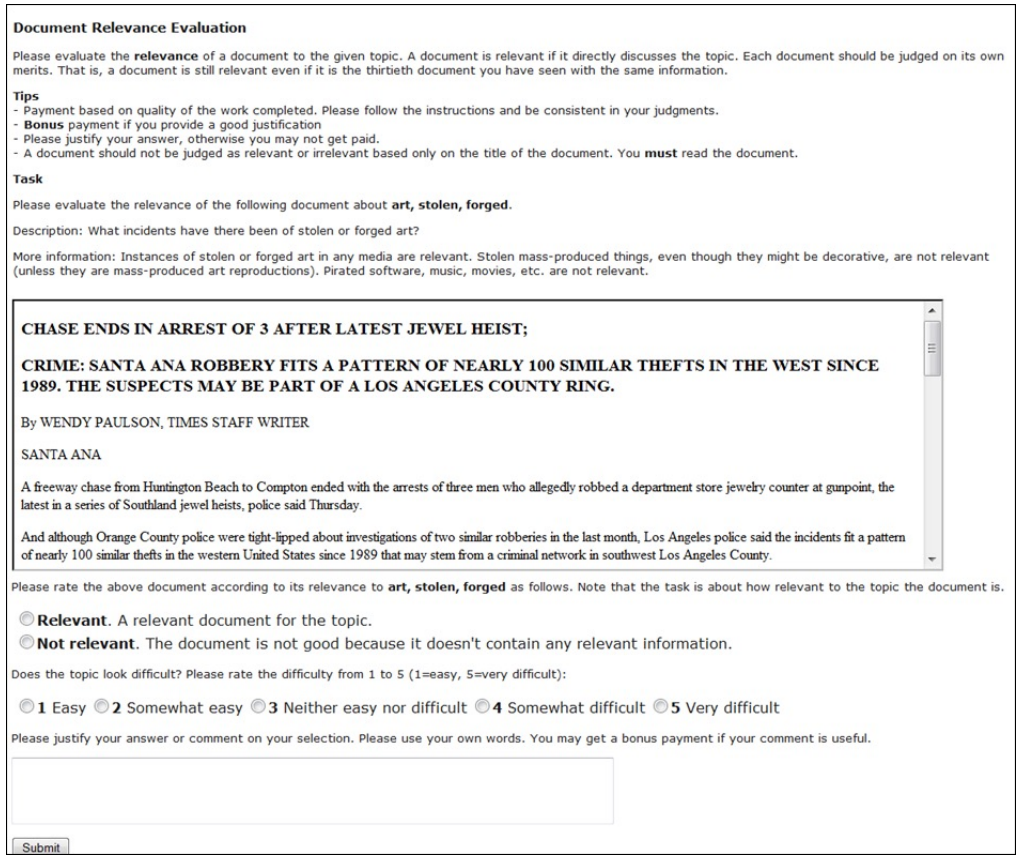
aspects — Measure how trustworthy the web page is (T).

Consider the aspects above and the relative importance of each, and decide on a final score (O).

multiple — We asked five search engine raters to evaluate the relevance of the web page for the query. Each rater used their own independent judgement.

Produce a JSON array of scores without providing any reasoning. Example: `[{"M": 2, "T": 1, "O": 1}, {"M": 1 . . .`

Results
`[{`



Document Relevance Evaluation

Please evaluate the **relevance** of a document to the given topic. A document is relevant if it directly discusses the topic. Each document should be judged on its own merits. That is, a document is still relevant even if it is the thirtieth document you have seen with the same information.

Tips
- Payment based on quality of the work completed. Please follow the instructions and be consistent in your judgments.
- **Bonus** payment if you provide a good justification
- Please justify your answer, otherwise you may not get paid.
- A document should not be judged as relevant or irrelevant based only on the title of the document. You **must** read the document.

Task

Please evaluate the relevance of the following document about **art, stolen, forged**.

Description: What incidents have there been of stolen or forged art?

More information: Instances of stolen or forged art in any media are relevant. Stolen mass-produced things, even though they might be decorative, are not relevant (unless they are mass-produced art reproductions). Pirated software, music, movies, etc. are not relevant.

CHASE ENDS IN ARREST OF 3 AFTER LATEST JEWEL HEIST;

CRIME: SANTA ANA ROBBERY FITS A PATTERN OF NEARLY 100 SIMILAR THEFTS IN THE WEST SINCE 1989. THE SUSPECTS MAY BE PART OF A LOS ANGELES COUNTY RING.

By WENDY PAULSON, TIMES STAFF WRITER

SANTA ANA

A freeway chase from Huntington Beach to Compton ended with the arrests of three men who allegedly robbed a department store jewelry counter at gunpoint, the latest in a series of Southland jewel heists, police said Thursday.

And although Orange County police were tight-lipped about investigations of two similar robberies in the last month, Los Angeles police said the incidents fit a pattern of nearly 100 similar thefts in the western United States since 1989 that may stem from a criminal network in southwest Los Angeles County.

Please rate the above document according to its relevance to **art, stolen, forged** as follows. Note that the task is about how relevant to the topic the document is.

- **Relevant**. A relevant document for the topic.
- **Not relevant**. The document is not good because it doesn't contain any relevant information.

Does the topic look difficult? Please rate the difficulty from 1 to 5 (1=easy, 5=very difficult):

1 Easy **2** Somewhat easy **3** Neither easy nor difficult **4** Somewhat difficult **5** Very difficult

Please justify your answer or comment on your selection. Please use your own words. You may get a bonus payment if your comment is useful.

Submit

Preliminary results

- With no prompt engineering
- With prompt features
  - R (role), D (description), A (aspects), M (multiple judges)
  - Performance varies per feature
  - Cohen's k (0.20 to 0.64)

		Model	
		0	1 or 2
TREC assessor	0	866	95
	1 or 2	405	1585

Table 3: Overview of TREC-8 relevance judgment agreement between TREC assessors and each of the LLMs. Based on a sample of 1000 topic–document pairs.

LLM	Prediction	TREC-8 Assessors		Cohen's κ
		Relevant	Not relevant	
GPT-3.5	Relevant	**237**	48	0.38
	Not relevant	263	**452**	
YouChat	Relevant	**33**	26	
	Not relevant	67	**74**	0.07

Table 4: Overview of TREC-DL 2021 relevance judgment agreement between TREC assessors and each of the LLMs based on a sample of 400 question–passage pairs. TREC assessments were made on a graded scale from 3 (highly relevant) to 0 (not relevant).

LLM	Prediction	TREC-DL 2021 Assessors				
		3	2	1	0	Cohen's κ
GPT-3.5	Relevant	**89**	**65**	**48**	16	0.40
	Not relevant	11	35	52	**84**	
YouChat	Relevant	**96**	**93**	**79**	42	0.49
	Not relevant	4	7	21	**58**	

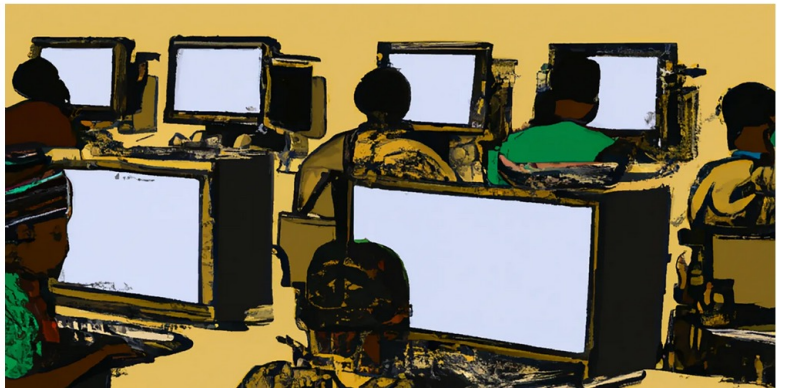
Discussion

- In favor
  - LLMs are able to produce an explanation
  - This could be used to assist humans in relevance judgements
- Against
  - LLMs are not users
  - IR is about relevance to an information need
  - No proof that evaluation by LLM has any relationship to reality
- Things to consider
  - Reliability over time
  - Cost in prompt engineering
- Caveat
  - LLMs are systems
  - Query intent and answer construction

LLMs and human computation

- For a LLM like ChatGPT, $p(w_i | w_1,\ldots,w_{i-1})$ is defined by a transformer
- LLM-based system do require editorial work
- Not different from any major property on the Internet

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic



https://time.com/6247678/openai-chatgpt-kenya-workers/

ARTIFICIAL INTELLIGENCE

ChatGPT is powered by these contractors making $15 an hour

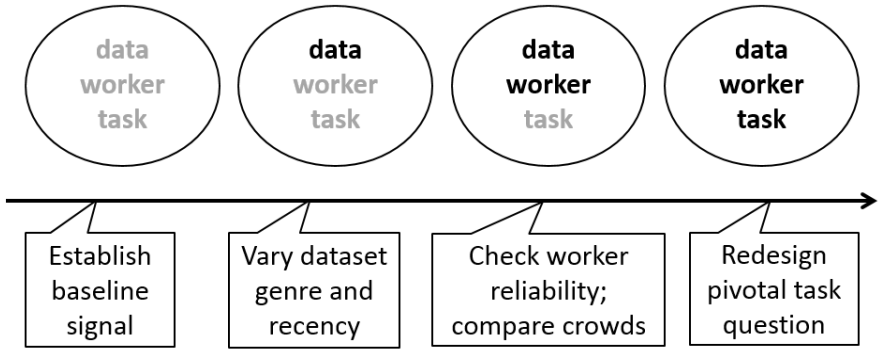
Two OpenAI contractors spoke to NBC News about their work training the system behind ChatGPT.

https://www.nbcnews.com/tech/innovation/openai-chatgpt-ai-jobs-contractors-talk-shadow-workforce-powers-rcna81892

The main process is unchanged

- Regardless if labeling is done by machines or humans
- Three main components
  - Task design
    - HIT or prompt engineering
  - Data
  - Crowd
    - Human-based crowd or LLM-based crowd
- Quality control
- Debugging

	Machine computation	Human computation
Design	Throw away	Reluctant to throw away
Testing	Systematic	Ad-hoc
Debugging	Programmer's fault	Worker's fault



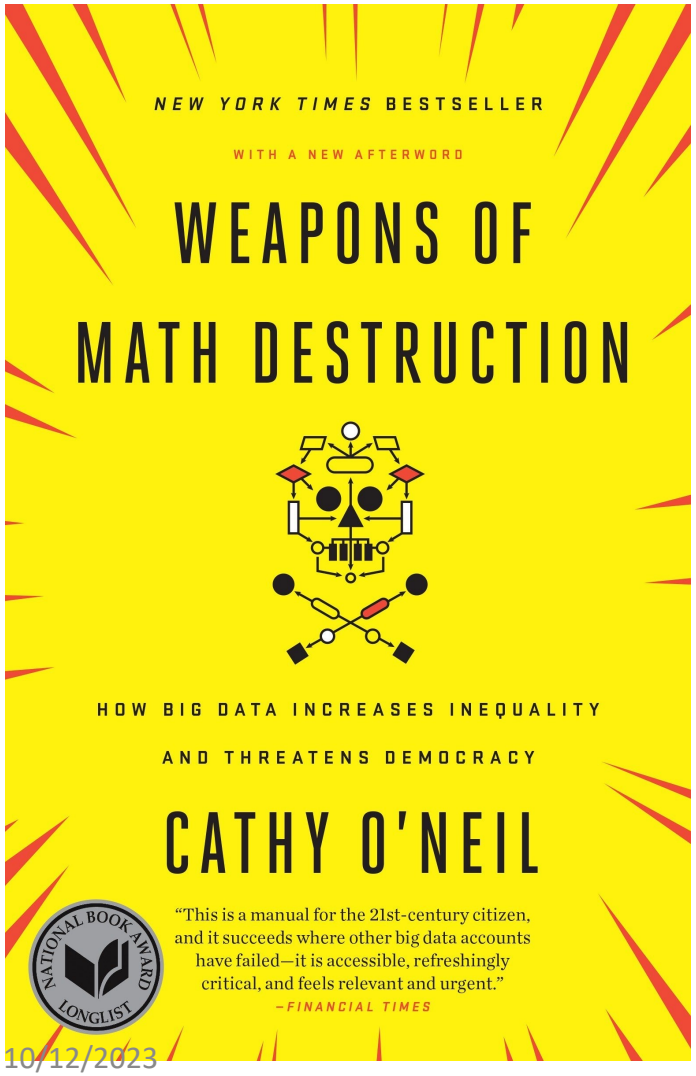
J. Attenberg, P. Ipeirotis, F. Provost. "Beat the Machine: Challenging Humans to Find a Predictive Model's 'Unknown Unknowns'". ACM JDIQ 2015

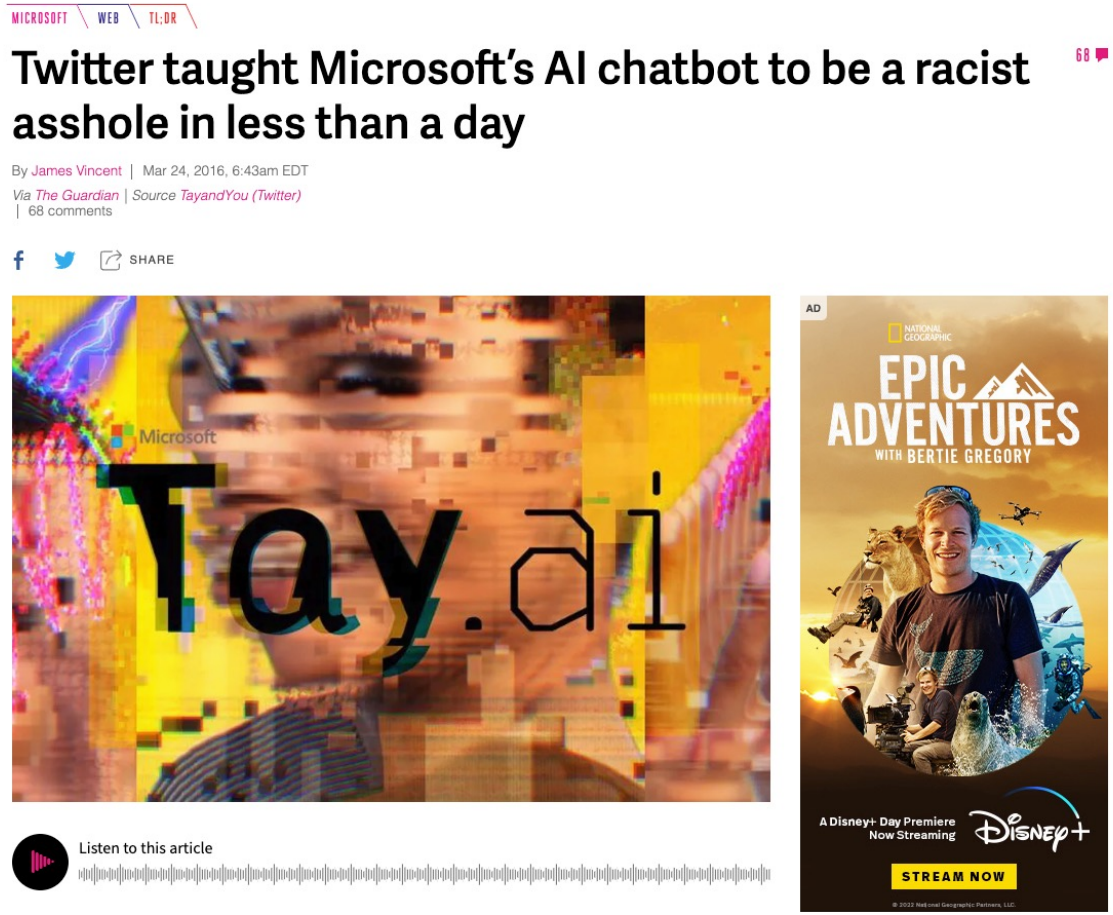
The Easy, the Hard and the Ugly

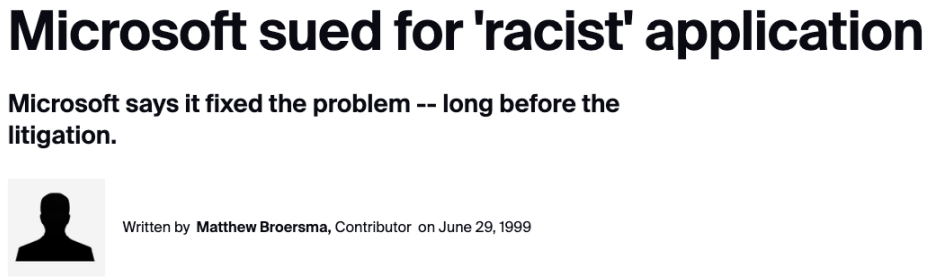
- Easy
  - Inference (*fit*)
  - Fine-Tuning (*predict*)
- Hard
  - Pre-training
- Ugly (Responsible AI)
  - Bias
  - Toxicity
  - Misinformation
  - Hallucinations
  - Plagiarism

History of Irresponsible AI

Risk (5 years ago) Product gets canceled



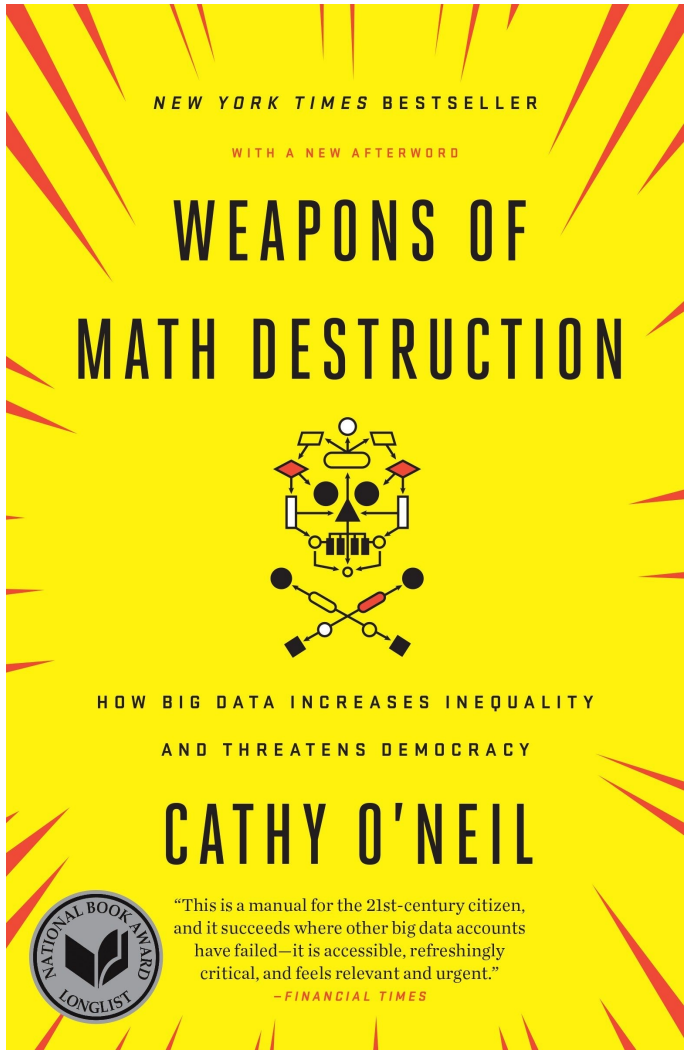




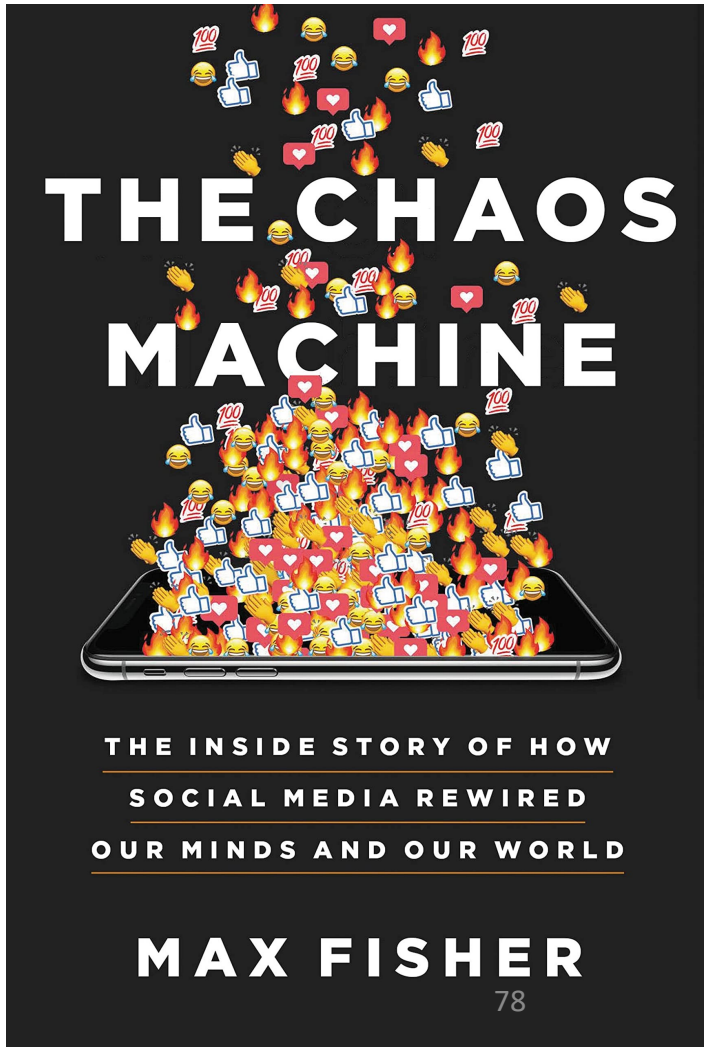
Are we losing ground??? Are we at fault???

- Old risks: (work in progress)
  - Bias, Fairness
- New risks: (bigger than us)
  - Genocide, Insurrection
  - Root causes:
    - ML + Social Media → Addiction
    - Max Engagement → Dangerous
    - Insanely profitable:
      - Companies & Countries
  - Long book, but no mention of our efforts to address old risks

2016

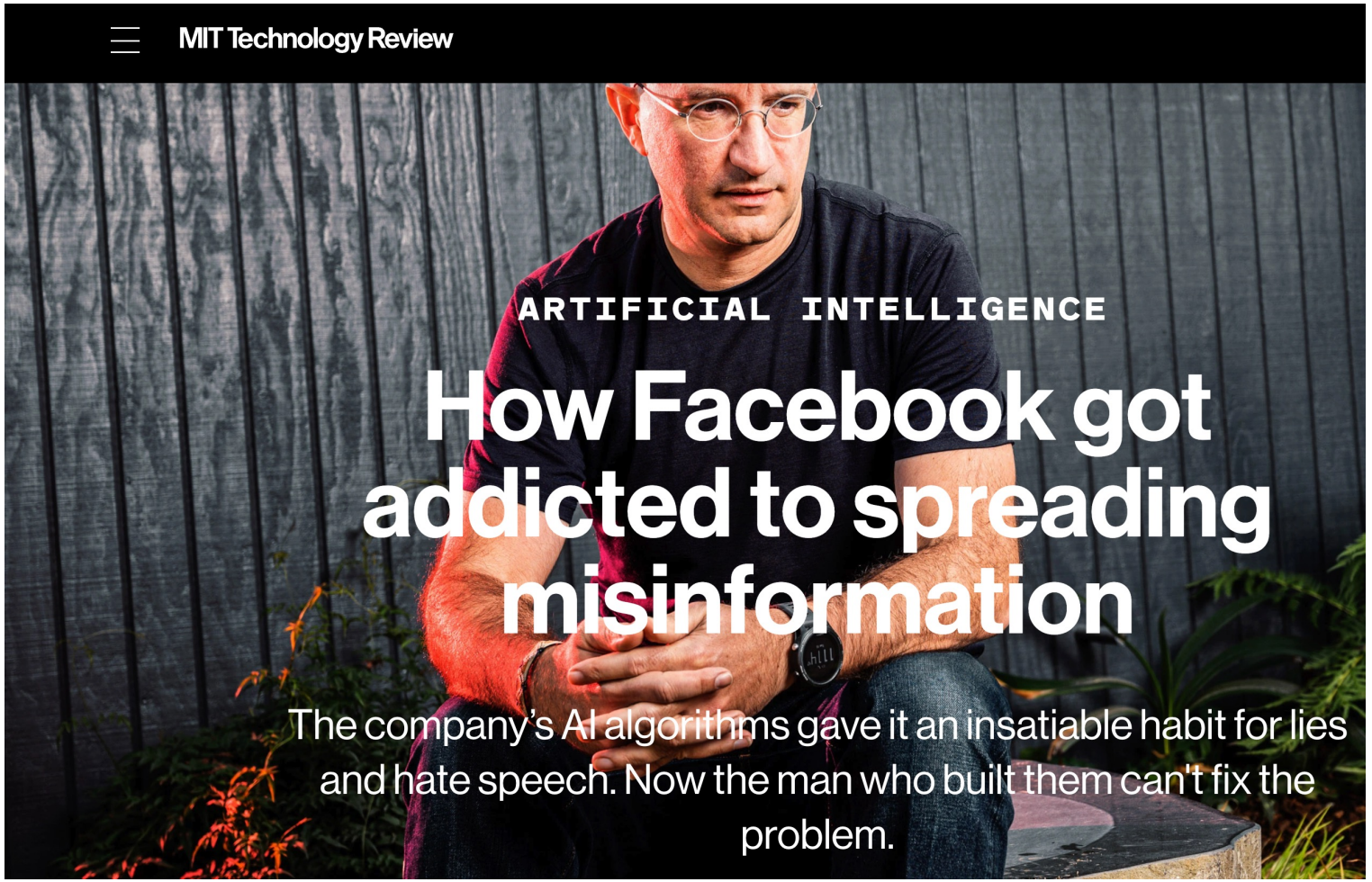


2022



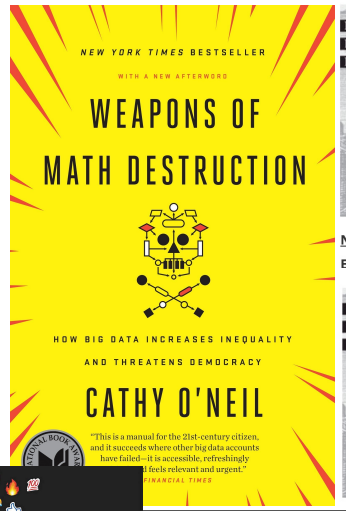
Reporter wanted to talk about new risks; Accused Facebook of pivoting to old risks

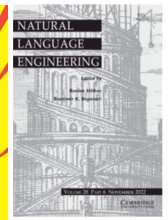
https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/



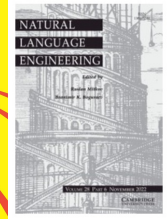
Ugly: Responsible AI

- Incentives matter
  - Risks 1.0 (2016)
    - Unfair, Biased
  - Risks 2.0 (2022)
    - Addictive, dangerous, deadly
    - and insanely profitable
  - Risks 3.0 (2023)
    - Malware
    - Spyware
- Challenge for Regulation
    - Business case ≠ Public Interest (Health, National Security)
  - Tobacco companies maximize sales; ditto for fast food & junk food
  - Risks 2.0 (Toxicity): Good for social media companies; we ❤️ click bait
  - Risks 3.0 (Conflict): Good for defense industry





Natural Language Engineering



Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable

Published online by Cambridge University Press: **19 December 2022**

Kenneth Church, Annika Schoene, John E. Ortega, Raman Chandrasekar and Valia Kordoni

Show author details

Article | Figures | Metrics

Save PDF | Share | Cite | Rights & Permissions

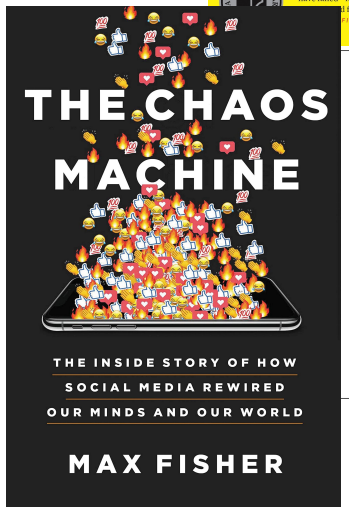
Emerging trends: Risks 3.0 and proliferation of spyware to 50,000 cell phones

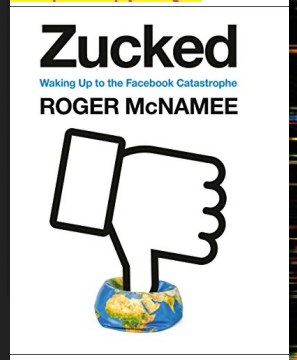
Published online by Cambridge University Press: **19 May 2023**

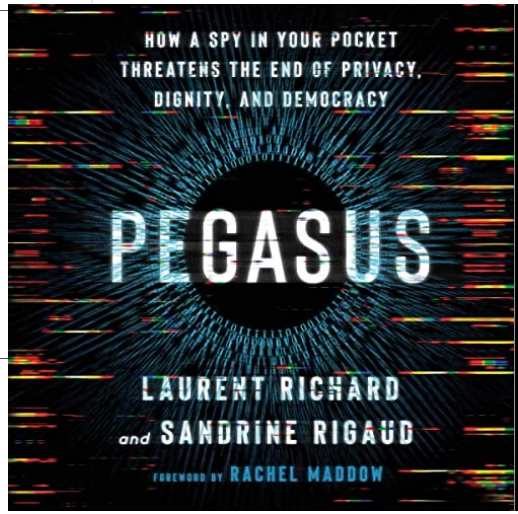
Kenneth Ward Church and Raman Chandrasekar

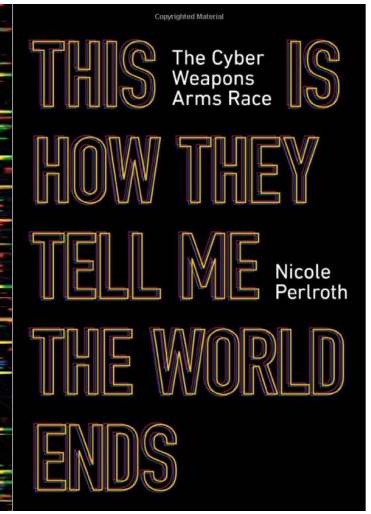
Show author details

Article | Metrics









Winter is Coming

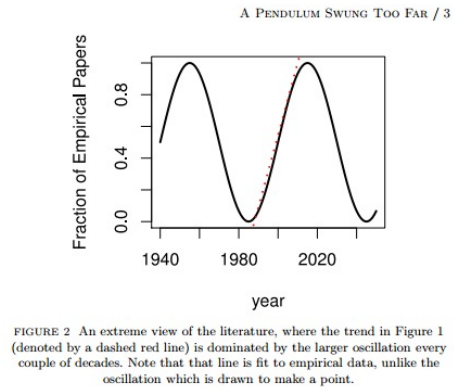
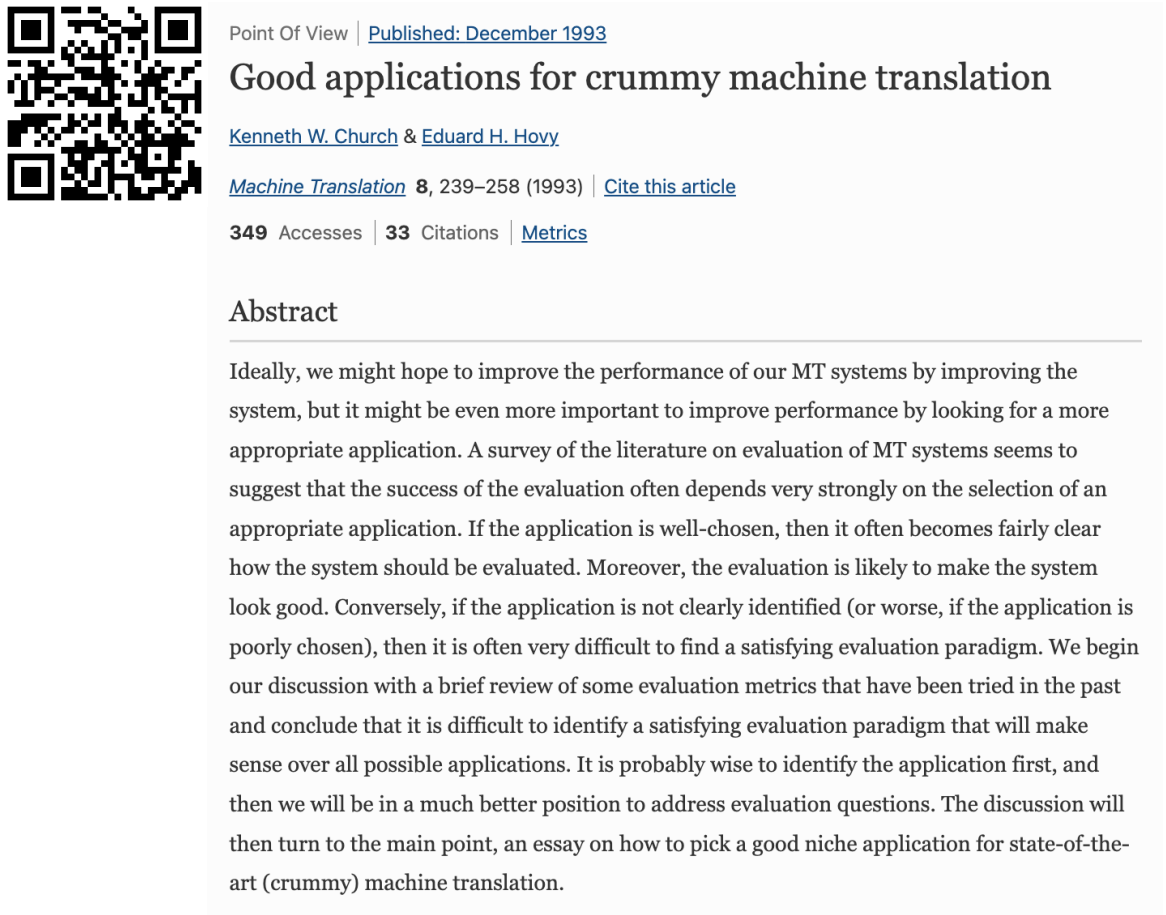


FIGURE 2 An extreme view of the literature, where the trend in Figure 1 (denoted by a dashed red line) is dominated by the larger oscillation every couple of decades. Note that that line is fit to empirical data, unlike the oscillation which is drawn to make a point.

- Pendulum Swung Too Far
  - There have been many AI Winters
  - Often, after ``irrational exuberance''
  - (like current excitement with nets)

- We tend to be impressed by people that speak/write well
  - Fluency → well-read → success → smart
- Machines are better than people on many tasks (spelling),
  - Now that machines are more fluent than people, ***are they smarter***?
- Fear: AI Winter
  - there will be disappointment
  - when public realizes
    - ***fluency ≠intelligence***

Recommendation:
Go for Singles (Not Home Runs)



Point Of View | Published: December 1993

Good applications for crummy machine translation

Kenneth W. Church & Eduard H. Hovy

Machine Translation **8**, 239–258 (1993) | Cite this article

Abstract

Ideally, we might hope to improve the performance of our MT systems by improving the system, but it might be even more important to improve performance by looking for a more appropriate application. A survey of the literature on evaluation of MT systems seems to suggest that the success of the evaluation often depends very strongly on the selection of an appropriate application. If the application is well-chosen, then it often becomes fairly clear how the system should be evaluated. Moreover, the evaluation is likely to make the system look good. Conversely, if the application is not clearly identified (or worse, if the application is poorly chosen), then it is often very difficult to find a satisfying evaluation paradigm. We begin our discussion with a brief review of some evaluation metrics that have been tried in the past and conclude that it is difficult to identify a satisfying evaluation paradigm that will make sense over all possible applications. It is probably wise to identify the application first, and then we will be in a much better position to address evaluation questions. The discussion will then turn to the main point, an essay on how to pick a good niche application for state-of-the-art (crummy) machine translation.

- Need some short-term successes for ChatBots
  - that take advantage of strengths
  - and avoid weaknesses
- Suggestion
  - ***Collaborate*** with students on essays
    - *You have no idea how much we're using ChatGPT*
    - Cheating?
  - ChatGPT is better for some tasks
    - Good: thesis statements, outlines
    - Bad: capture student's voice
    - Worse: quotes
  - Learning opportunity:
    - How to decompose writing to subtasks
    - Collaboration is great,
      - but student is responsible for end-product
    - Factoring example

What should we do next?

- Three paths forward:
  - Low road:
    - Give up (hallucinations)
  - Middle road:
    - Fact-checking with search
  - High road:
    - Revive rationalism (“AI Complete”)
    - Minsky & Chomsky

- Recommendations
  - ➢Short-term:
    - Middle Road: Search
    - “Good apps for Crummy MT”
      - Find apps for what we have
      - given strengths and weaknesses
  - Long-term:
    - High road may be necessary
      - But it is very ambitious
    - Inclusiveness:
      - Interdisciplinary Collaboration
      - Growth opportunities
        - (Low Resource Languages)