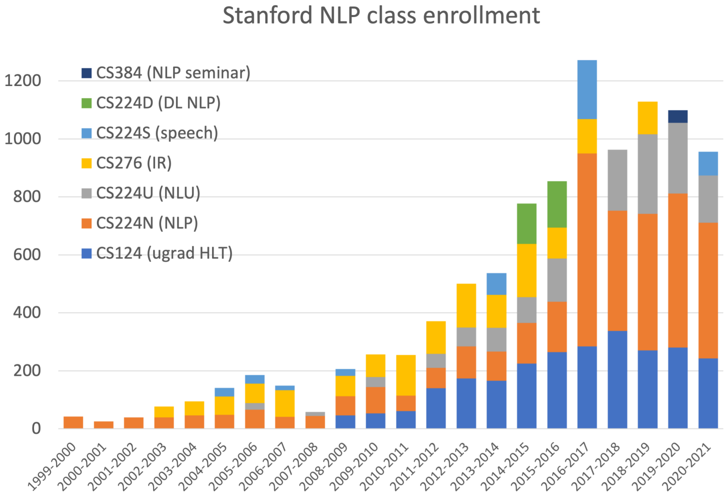
# Reliable characterizations of NLP systems as a social responsibility

Christopher Potts

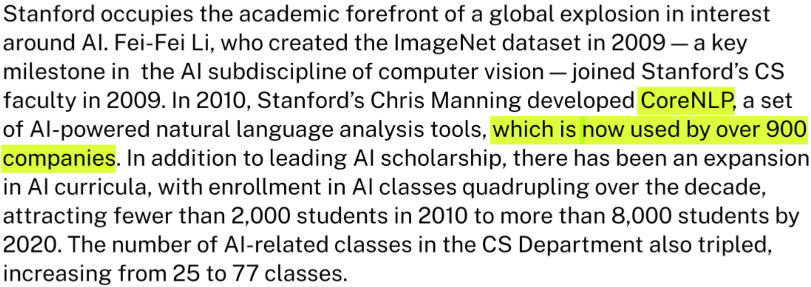
Stanford Linguistics and the Stanford NLP Group

ACL-IJCNLP 2021

# More impact than ever before



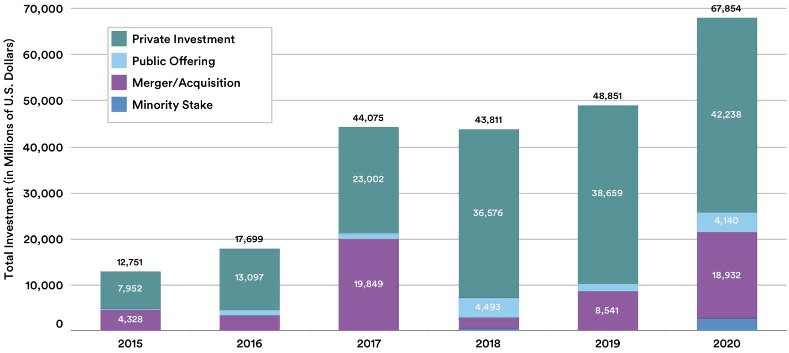
# More impact than ever before



Stanford occupies the academic forefront of a global explosion in interest around AI. Fei-Fei Li, who created the ImageNet dataset in 2009 — a key milestone in  the AI subdiscipline of computer vision — joined Stanford's CS faculty in 2009. In 2010, Stanford's Chris Manning developed CoreNLP, a set of AI-powered natural language analysis tools, which is now used by over 900 companies. In addition to leading AI scholarship, there has been an expansion in AI curricula, with enrollment in AI classes quadrupling over the decade, attracting fewer than 2,000 students in 2010 to more than 8,000 students by 2020. The number of AI-related classes in the CS Department also tripled, increasing from 25 to 77 classes.

Stanford Tech History Project

# More impact than ever before

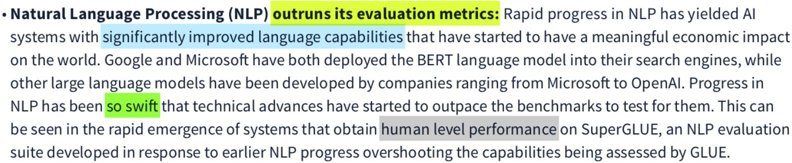


**GLOBAL CORPORATE INVESTMENT in AI by INVESTMENT ACTIVITY, 2015-20**

Source: CapIQ, Crunchbase, and NetBase Quid, 2020 | Chart: 2021 AI Index Report

2021 AI Index

# More impact than ever before



- **Natural Language Processing (NLP) outruns its evaluation metrics:** Rapid progress in NLP has yielded AI systems with significantly improved language capabilities that have started to have a meaningful economic impact on the world. Google and Microsoft have both deployed the BERT language model into their search engines, while other large language models have been developed by companies ranging from Microsoft to OpenAI. Progress in NLP has been so swift that technical advances have started to outpace the benchmarks to test for them. This can be seen in the rapid emergence of systems that obtain human level performance on SuperGLUE, an NLP evaluation suite developed in response to earlier NLP progress overshooting the capabilities being assessed by GLUE.

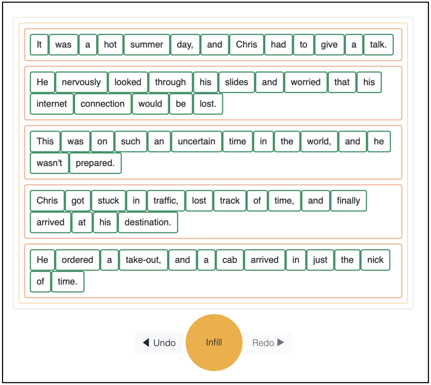
2021 AI Index

# Application areas

- Self-expression
- Language preservation
- Accessibility
- Community building
- Healthcare
- Fraud detection
- Securities trading
- Recommendations
- Advertising
- Surveillance
- Propaganda
- Disinformation

# Application areas

- Self-expression
- Language preservation
- Accessibility
- Community building
- Healthcare
- Fraud detection
- Securities trading
- Recommendations
- Advertising
- Surveillance
- Propaganda
- Disinformation



It was a hot summer day, and Chris had to give a talk.

He nervously looked through his slides and worried that his internet connection would be lost.

This was on such an uncertain time in the world, and he wasn't prepared.

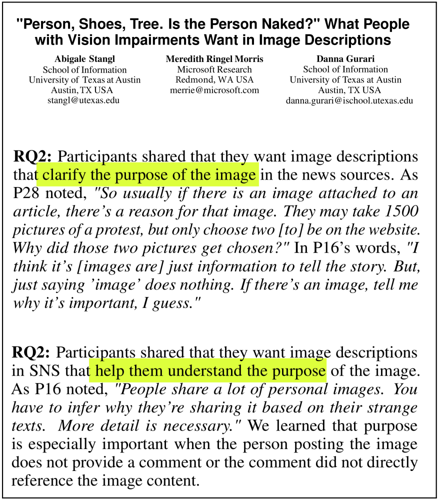
Chris got stuck in traffic, lost track of time, and finally arrived at his destination.

He ordered a take-out, and a cab arrived in just the nick of time.

Donahue et al. 2020

# Application areas

- Self-expression
- Language preservation
- Accessibility
- Community building
- Healthcare
- Fraud detection
- Securities trading
- Recommendations
- Advertising
- Surveillance
- Propaganda
- Disinformation



## "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions

**Abigale Stangl**
School of Information
University of Texas at Austin
Austin, TX USA
stangl@utexas.edu

**Meredith Ringel Morris**
Microsoft Research
Redmond, WA USA
merrie@microsoft.com

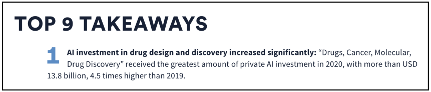
**Danna Gurari**
School of Information
University of Texas at Austin
Austin, TX USA
danna.gurari@ischool.utexas.edu

**RQ2:** Participants shared that they want image descriptions that clarify the purpose of the image in the news sources. As P28 noted, *"So usually if there is an image attached to an article, there's a reason for that image. They may take 1500 pictures of a protest, but only choose two [to] be on the website. Why did those two pictures get chosen?"* In P16's words, *"I think it's [images are] just information to tell the story. But, just saying 'image' does nothing. If there's an image, tell me why it's important, I guess."*

**RQ2:** Participants shared that they want image descriptions in SNS that help them understand the purpose of the image. As P16 noted, *"People share a lot of personal images. You have to infer why they're sharing it based on their strange texts. More detail is necessary."* We learned that purpose is especially important when the person posting the image does not provide a comment or the comment did not directly reference the image content.

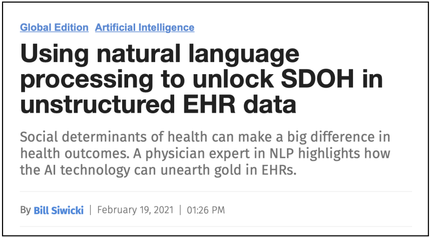
# Application areas

- Self-expression
- Language preservation
- Accessibility
- Community building
- Healthcare
- Fraud detection
- Securities trading
- Recommendations
- Advertising
- Surveillance
- Propaganda
- Disinformation



**TOP 9 TAKEAWAYS**

1 **AI investment in drug design and discovery increased significantly:** "Drugs, Cancer, Molecular, Drug Discovery" received the greatest amount of private AI investment in 2020, with more than USD 13.8 billion, 4.5 times higher than 2019.



Global Edition Artificial Intelligence

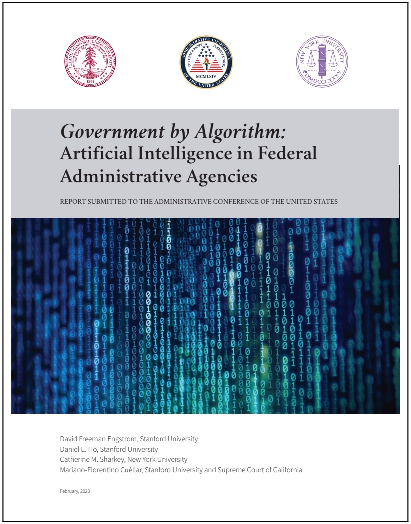
**Using natural language processing to unlock SDOH in unstructured EHR data**

Social determinants of health can make a big difference in health outcomes. A physician expert in NLP highlights how the AI technology can unearth gold in EHRs.

By Bill Siwicki | February 19, 2021 | 01:26 PM

# Application areas

- Self-expression
- Language preservation
- Accessibility
- Community building
- Healthcare
- Fraud detection
- Securities trading
- Recommendations
- Advertising
- Surveillance
- Propaganda
- Disinformation



*Government by Algorithm:*
Artificial Intelligence in Federal Administrative Agencies

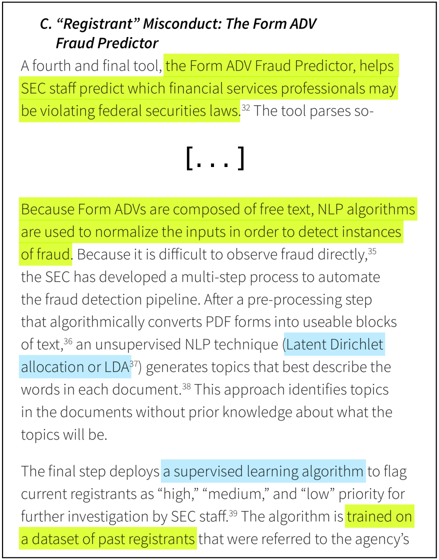
REPORT SUBMITTED TO THE ADMINISTRATIVE CONFERENCE OF THE UNITED STATES

David Freeman Engstrom, Stanford University
Daniel E. Ho, Stanford University
Catherine M. Sharkey, New York University
Mariano-Florentino Cuéllar, Stanford University and Supreme Court of California

February, 2020

# Application areas

- Self-expression
- Language preservation
- Accessibility
- Community building
- Healthcare
- Fraud detection
- Securities trading
- Recommendations
- Advertising
- Surveillance
- Propaganda
- Disinformation



***C. "Registrant" Misconduct: The Form ADV Fraud Predictor***

A fourth and final tool, the Form ADV Fraud Predictor, helps SEC staff predict which financial services professionals may be violating federal securities laws.[32] The tool parses so-

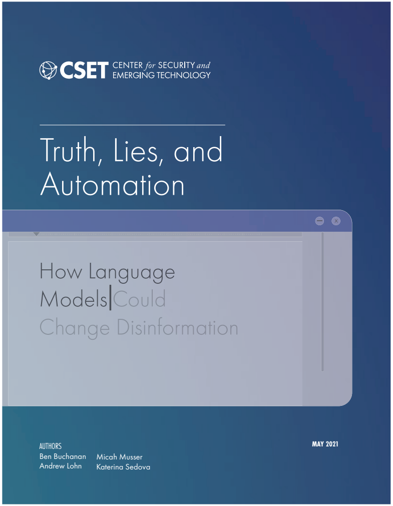
[. . .]

Because Form ADVs are composed of free text, NLP algorithms are used to normalize the inputs in order to detect instances of fraud. Because it is difficult to observe fraud directly,[35] the SEC has developed a multi-step process to automate the fraud detection pipeline. After a pre-processing step that algorithmically converts PDF forms into useable blocks of text,[36] an unsupervised NLP technique (Latent Dirichlet allocation or LDA[37]) generates topics that best describe the words in each document.[38] This approach identifies topics in the documents without prior knowledge about what the topics will be.

The final step deploys a supervised learning algorithm to flag current registrants as "high," "medium," and "low" priority for further investigation by SEC staff.[39] The algorithm is trained on a dataset of past registrants that were referred to the agency's

# Application areas

- Self-expression
- Language preservation
- Accessibility
- Community building
- Healthcare
- Fraud detection
- Securities trading
- Recommendations
- Advertising
- Surveillance
- Propaganda
- Disinformation



CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

Truth, Lies, and Automation

How Language Models Could Change Disinformation

AUTHORS
Ben Buchanan
Andrew Lohn
Micah Musser
Katerina Sedova

MAY 2021

Clark et al. 2021 🏆

# Notions of social responsibility

1. Pursuit of knowledge
2. Dissemination of knowledge
3. Utility
4. Consequences
    - for the planet
    - for study participants and subjects
    - for individuals and society

Edsall 1975, 1981

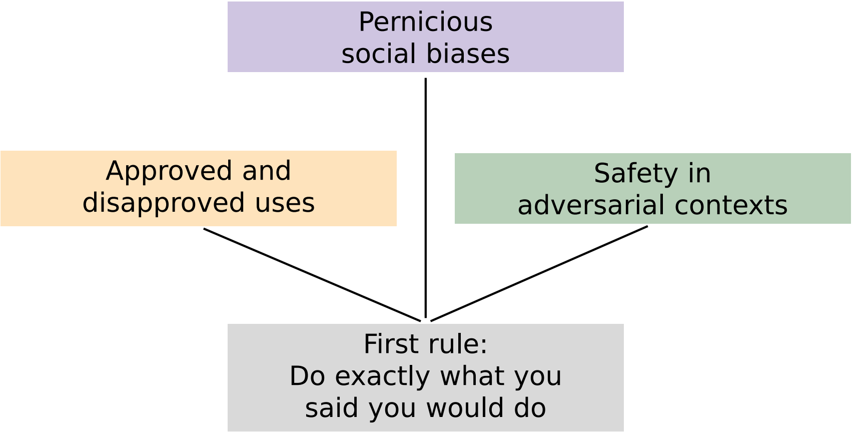
# First rule

Do exactly what you said you would do.

- Accurately charaterize what your dataset/model/system does and what it does not do.
- Disclosures (e.g., Model Cards, Datasheets)
- Effective communication about context

Raises a *different* set of challenging questions.

# Limited goals for today


Pernicious social biases
Approved and disapproved uses
Safety in adversarial contexts
First rule: Do exactly what you said you would do

# Roles to have in mind

First rule: Do exactly what you said you would do.

1. ~~Insider: ACL attendee~~
2. Practitioner: Informed and engaged engineer
3. Leader: Executive with technical training outside of AI

1. Media: “Robots are better at reading than humans” [link]
2. Insider: For SQuAD, a model has surpassed our estimate of human performance.
3. Practitioner: There might be value in QA models now.
4. Leader: Can we automate our question answering?

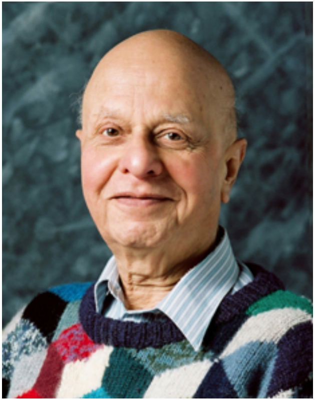
Schlangen 2020

# Overview

1. Benchmark datasets: Delimit responsible use
2. System assessment: Connect with real-world concerns
3. Discussion

# Benchmark datasets

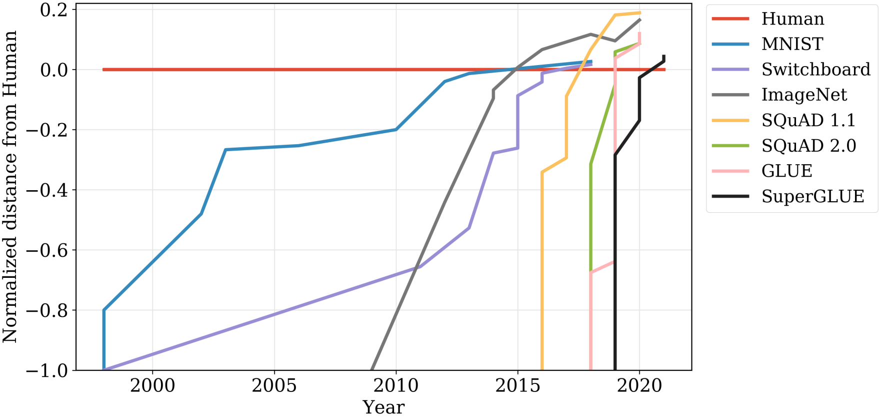
# Seeing farther than ever before



Aravind Joshi: Datasets as the telescopes of our field

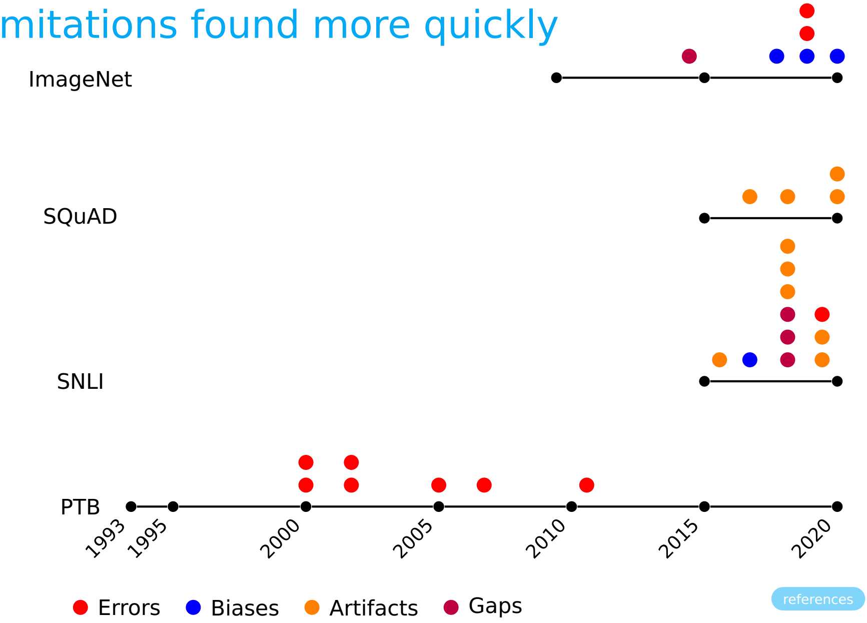
Photo credit: JoshiFest

# Benchmarks saturate faster than ever



Kiela et al. 2021

# Limitations found more quickly



references

# Two perspectives on dataset creation

## Fixed benchmarks

| Benefits | Drawbacks |
|---|---|
| Ease of measurement | Community-wide overfitting |
| Efficiency | Deficiencies inevitable |

Strathern’s Law: “When a measure becomes a target, it ceases to be a good measure.”

## Nie et al. (2020): “ ‘moving post’ dynamic target”

| Benefits | Drawbacks |
|---|---|
| Diversity | Expense |
| Evolving goals | Comparisons harder |

Can be responsive to evolving needs.

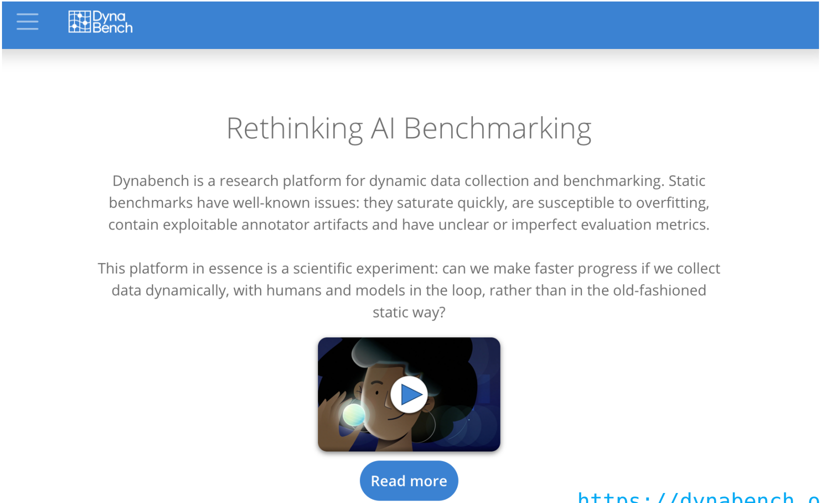
# Dynabench

## Dynabench: Rethinking Benchmarking in NLP

Douwe Kiela[†], Max Bartolo[‡], Yixin Nie[⋆], Divyansh Kaushik[§], Atticus Geiger[¶], Zhengxuan Wu[¶], Bertie Vidgen[‖], Grusha Prasad[⋆⋆], Amanpreet Singh[†], Pratik Ringshia[†], Zhiyi Ma[†], Tristan Thrush[†], Sebastian Riedel[†‡], Zeerak Waseem[††], Pontus Stenetorp[‡], Robin Jia[†], Mohit Bansal[⋆], Christopher Potts[¶] and Adina Williams[†]

[†] Facebook AI Research; [‡] UCL; [⋆] UNC Chapel Hill; [§] CMU; [¶] Stanford University
[‖] Alan Turing Institute; [⋆⋆] JHU; [††] Simon Fraser University
dynabench@fb.com

# Dynabench



## 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?

`https://dynabench.org`

# Dynamics of dynamic datasets

1. SWAG to BERT to HellaSWAG (Zellers et al. 2018, 2019)
2. Adversarial NLI (Nie et al. 2020)
3. Beat the AI (Bartolo et al. 2020)
4. Dynabench Hate Speech (Vidgen et al. 2020)
5. DynaSent (Potts et al. 2021)
6. Dynabench QA

# Dataset papers

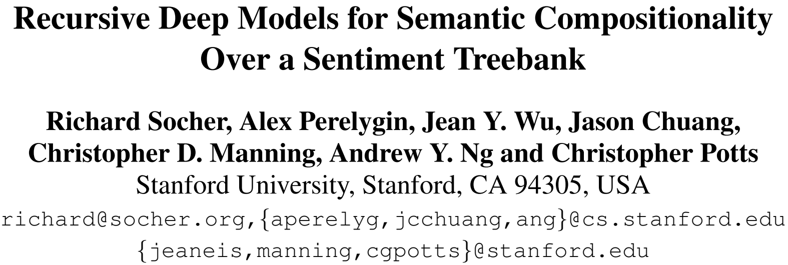
1. Standard: Motivation
2. Standard: Construction
3. Standard: Model evaluations
4. Proposed: Delimiting responsible use

   Datasheets: “Is there anything about the composition of the dataset [. . . ] that might impact future uses?”

   - Reaching the well-intentioned user

Gebru et al. 2018; NeurIPS Datasets & Benchmarks track

# Looking back on the SST



**Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank**

**Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang,**
**Christopher D. Manning, Andrew Y. Ng and Christopher Potts**
Stanford University, Stanford, CA 94305, USA
`richard@socher.org,{aperelyg,jcchuang,ang}@cs.stanford.edu`
`{jeaneis,manning,cgpotts}@stanford.edu`

Healthcare? Professional evaluations? Literary analysis?

Practitioner Leader

Socher et al. 2013

# Assessment

# Notions of assessment

- Our apparent relentness pursuit of F1 (and friends)
- Empowering users
- Estimating human performance

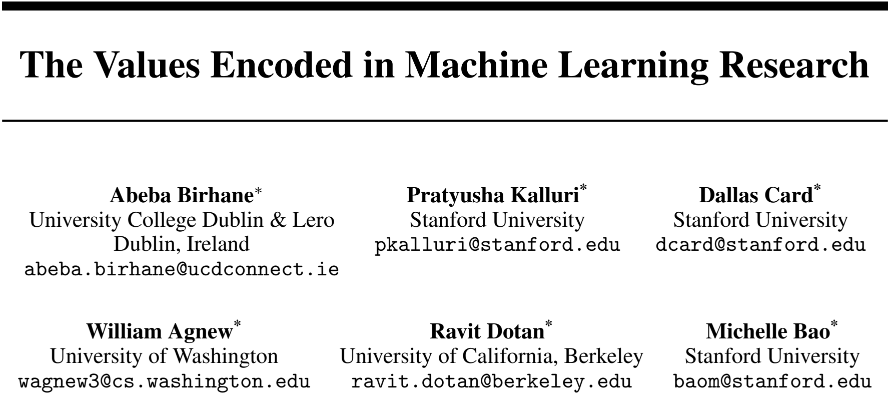
# Metrics and application areas

- Missing a safety signal costs lives; human review is feasible
- Exemplars need to be found in a massive dataset
- Specific mistakes are deal-breakers; others hardly matter
- Cases need to be prioritized
- The solution needs to work over an aging cell network
- The solution cannot provide worse service to specific groups
- Specific predictions need to be blocked

Our (apparent) answer: F1 and friends

Practitioner Leader

# What we seem to value



## The Values Encoded in Machine Learning Research

**Abeba Birhane***
University College Dublin & Lero
Dublin, Ireland
abeba.birhane@ucdconnect.ie

**Pratyusha Kalluri***
Stanford University
pkalluri@stanford.edu

**Dallas Card***
Stanford University
dcard@stanford.edu

**William Agnew***
University of Washington
wagnew3@cs.washington.edu

**Ravit Dotan***
University of California, Berkeley
ravit.dotan@berkeley.edu

**Michelle Bao***
Stanford University
baom@stanford.edu

# What we seem to value

Selected ‘Values encoded in ML research’ from Birhane et al. (2021):

Performance

Efficiency

Interpretability (for researchers)

Applicability in the real world

Robustness

Scalability

Interpretability (for users)

Benificence

Privacy

Fairness

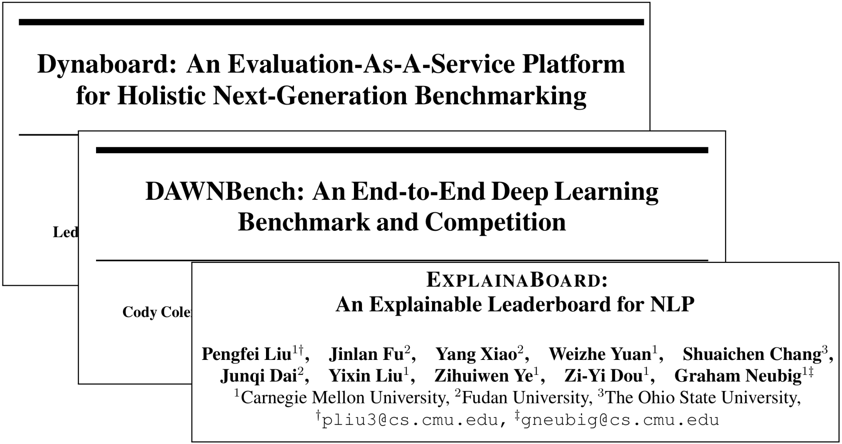
Justice

# What we seem to value

Selected ‘Values encoded in ML research’ from Birhane et al. (2021):

Performance

# Towards multidimensional leaderboards



**Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking**

**Led**

**DAWNBench: An End-to-End Deep Learning Benchmark and Competition**

**Cody Cole**

**EXPLAINABOARD: An Explainable Leaderboard for NLP**

**Pengfei Liu**[1†], **Jinlan Fu**[2], **Yang Xiao**[2], **Weizhe Yuan**[1], **Shuaichen Chang**[3], **Junqi Dai**[2], **Yixin Liu**[1], **Zihuiwen Ye**[1], **Zi-Yi Dou**[1], **Graham Neubig**[1‡]
[1]Carnegie Mellon University, [2]Fudan University, [3]The Ohio State University,
[†]pliu3@cs.cmu.edu, [‡]gneubig@cs.cmu.edu

Dodge et al. 2019; Ethayarajh and Jurafsky 2020

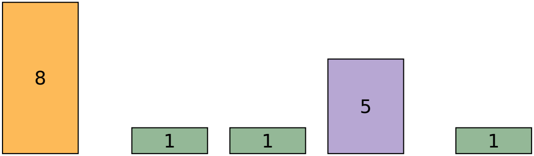
# Dynabench and Dynascore

| | 8 | 2 | 2 | 2 | 2 | |
|---|---|---|---|---|---|---|
| Model | Performance | Throughput | Memory | Fairness | Robustness | Dynascore |
| DeBERTa | 76.25 | 4.47 | 6.97 | 88.33 | 90.06 | 45.92 |
| ELECTRA-large | 76.07 | 2.37 | 25.30 | 93.13 | 91.64 | 45.79 |
| RoBERTa | 69.67 | 6.88 | 6.17 | 88.32 | 86.10 | 42.54 |
| ALBERT | 68.63 | 6.85 | 2.54 | 87.44 | 80.90 | 41.74 |
| BERT | 57.14 | 6.70 | 5.55 | 91.45 | 80.81 | 36.07 |
| BiDAF | 53.48 | 10.71 | 3.60 | 80.79 | 77.03 | 33.96 |
| Unrestricted T5 | 28.80 | 4.51 | 10.69 | 92.32 | 88.41 | 22.18 |
| Return Context | 5.99 | 89.80 | 1.10 | 95.97 | 91.61 | 15.47 |

## Question answering

Ma et al. 2021; https://dynabench.org

# Dynabench and Dynascore



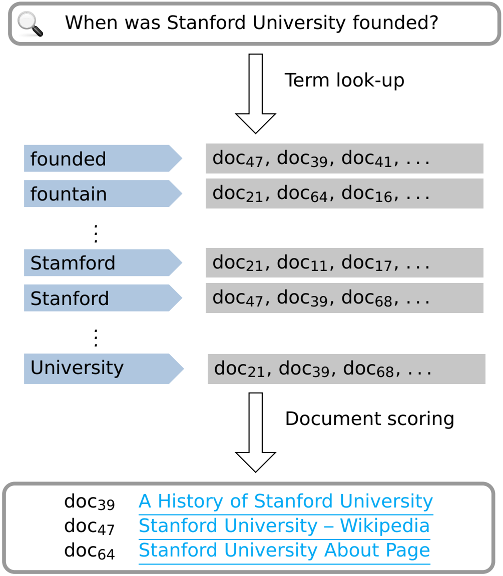
| Model | Performance | Throughput | Memory | Fairness | Robustness | Dynascore |
|---|---|---|---|---|---|---|
| DeBERTa | 76.25 | 4.47 | 6.97 | 88.33 | 90.06 | 46.70 |
| ELECTRA-large | 76.07 | 2.37 | 25.30 | 93.13 | 91.64 | 46.86 |
| RoBERTa | 69.67 | 6.88 | 6.17 | 88.32 | 86.10 | 43.37 |
| ALBERT | 68.63 | 6.85 | 2.54 | 87.44 | 80.90 | 42.66 |
| BERT | 57.14 | 6.70 | 5.55 | 91.45 | 80.81 | 37.17 |
| BiDAF | 53.48 | 10.71 | 3.60 | 80.79 | 77.03 | 34.62 |
| Unrestricted T5 | 28.80 | 4.51 | 10.69 | 92.32 | 88.41 | 23.19 |
| Return Context | 5.99 | 89.80 | 1.10 | 95.97 | 91.61 | 14.29 |

Question answering

Ma et al. 2021; `https://dynabench.org`

# New directions for neural IR – think of the User!

# New directions for neural IR – think of the User!



When was Stanford University founded?

Term look-up

| | |
|---|---|
| founded | doc$_{47}$, doc$_{39}$, doc$_{41}$, . . . |
| fountain | doc$_{21}$, doc$_{64}$, doc$_{16}$, . . . |
| ⋮ | |
| Stamford | doc$_{21}$, doc$_{11}$, doc$_{17}$, . . . |
| Stanford | doc$_{47}$, doc$_{39}$, doc$_{68}$, . . . |
| ⋮ | |
| University | doc$_{21}$, doc$_{39}$, doc$_{68}$, . . . |

Document scoring

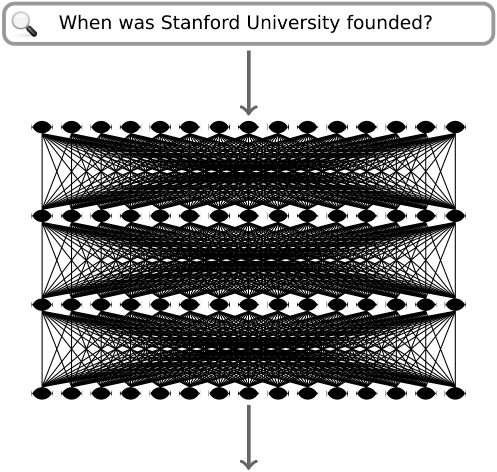
| | |
|---|---|
| doc$_{39}$ | A History of Stanford University |
| doc$_{47}$ | Stanford University – Wikipedia |
| doc$_{64}$ | Stanford University About Page |

✓ Provenance

✓ Updatability

✗ Synthesis

# New directions for neural IR – think of the User!

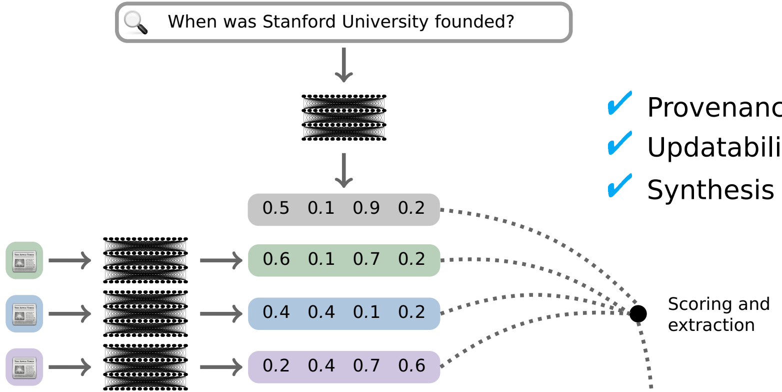


Stanford University was founded in 1891.

- ✗ Provenance
- ✗ Updatability
- ✓ Synthesis

Metzler et al. 2021

# New directions for neural IR – think of the User!



✓ Provenance

✓ Updatability

✓ Synthesis

“Stanford University was founded in 1885 by California senator Leland Stanford and his wife, Jane”
A History of Stanford University

“Stanford was founded in 1885 by Leland and Jane Stanford in memory of their only child, Leland Stanford Jr.”
Stanford University – Wikipedia

“Opened in 1891”
Stanford University About Page

Khattab et al. 2021

# Estimating human performance

| Premise | Label | Hypothesis |
|---|---|---|
| A dog jumping | neutral | A dog wearing a sweater |
| turtle | contradiction | linguist |
| A photo of a race horse | ? | A photo of an athlete |
| A chef using a barbecue | ? | A person using a machine |

Human response throughout: “Let’s discuss”

“Human performance” ≈ Average performance of harried crowdworkers doing a machine task repeatedly

Pavlick and Kwiatkowski 2019

# Summary

## Assessment today

- One-dimensional
- Largely insensitive to context (use-case)
- Terms set by the research community
- Opaque
- Tailored to machine tasks

## Assessments in the future

- High-dimensional and fluid
- Highly sensitive to context (use-case)
- Terms set by the stakeholders
- Judgments ultimately made by users
- Tailored to human tasks (?)

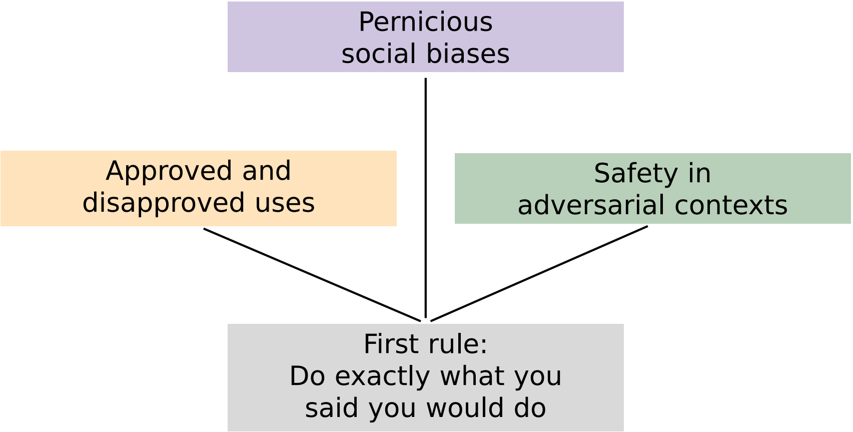
# Discussion

# Opportunities and social responsibilities

- Self-expression
- Language preservation
- Accessibility
- Community building
- Healthcare
- Fraud detection
- Securities trading
- Recommendations
- Advertising
- Surveillance
- Propaganda
- Disinformation

1. Insider: ACL attendee
2. Practitioner: Informed and engaged engineer
3. Leader: Executive with technical training outside of AI
4. User: Someone deriving value from an NLP-driven system

# First Rule . . . of many



Pernicious social biases

Approved and disapproved uses

Safety in adversarial contexts

First rule: Do exactly what you said you would do

# Translational research efforts

AI will call for unique solutions, but these examples might be inspiring:

- National Center for Advancing Translational Sciences
- The Translational Research Institute for Space Health
- Mapping Educational Specialist KnowHow (MESH)
- Nutrition labels on foods

(cf. https://datanutrition.org)

# Components and consequences

- Informing well-intentioned potential users of your ideas.
- Components:
  - Datasets
  - Assessment
  - Structural evaluation methods: Probing, feature attribution, causal abstraction, . . .
  - Licensing of data, code, models
  - Valuing tools as major contributions
  - Accurate naming of concepts (Mitchell 2021; Lipton and Steinhardt 2019)
  - . . .
- Consequences:
  - More multifaceted scientific goals
  - More success out in the wider world

Thanks!

# References I

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Stroudsburg, PA. Association for Computational Linguistics.

Adriane Boyd, Markus Dickinson, and Detmar Meurers. 2008. On detecting errors in dependency treebanks. *Research on Language and Computation*, 6(2):113–137.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. Truth, lies, and automation. Center for Security and Emerging Technology.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. 2017. Dawnbench: An end-to-end deep learning benchmark and competition. *Training*, 100(101):102.

Kate Crawford and Trevor Paglen. 2021. Excavating ai: The politics of images in machine learning training sets. *AI & SOCIETY*, pages 1–12.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S Bernstein, Alex Berg, and Li Fei-Fei. 2014. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3099–3102.

Markus Dickinson and W. Detmar Meurers. 2003a. Detecting errors in part-of-speech annotation. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Markus Dickinson and W. Detmar Meurers. 2005. Detecting errors in discontinuous structural annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 322–329, Ann Arbor, Michigan. Association for Computational Linguistics.

# References II

Markus Dickinson and Walt Detmar Meurers. 2003b. Detecting inconsistencies in treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

John T. Edsall. 1975. Scientific freedom and responsibility. *Science*, 188(4189):687–693.

John T. Edsall. 1981. Two aspects of scientific responsibility. *Science*, 212(4490):11–14.

David Freeman Engstrom, Daniel E Ho, Catherine M Sharkey, and Mariano-Florentino Cuéllar. 2020. Government by algorithm: Artificial intelligence in federal administrative agencies. *NYU School of Law, Public Law Research Paper*, (20-54).

Eleazar Eskin. 2000. Detecting errors within a corpus using anomaly detection. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hans van Halteren. 2000. The detection of inconsistency in manually tagged text. In *Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora*, pages 48–55, Centre Universitaire, Luxembourg. International Committee on Computational Linguistics.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. A moderate proposal for radically better AI-powered Web search. Stanford HAI Blog.

# References III

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Zachary Chase Lipton and Jacob Steinhardt. 2019. Troubling trends in machine learning scholarship. *Queue*, 17:45 – 77.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaicheng Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. Explainaboard: An explainable leaderboard forNLP. *arXiv preprint arXiv:2104.06387*.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. Ms., Facebook AI Research and Stanford University.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing – Part I*, number 6608 in Lecture Notes in Computer Science, pages 171–189. Springer, Berlin.

Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: Making experts out of dilettantes. *arXiv preprint arXiv:2105.02274*.

Melanie Mitchell. 2021. Why AI is harder than we think. *arXiv preprint arXiv:2104.12871*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

# References IV

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400, Long Beach, California, USA. PMLR.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

David Schlangen. 2020. Targeting the benchmark: On methodology in current natural language processing research. *arXiv preprint arXiv:2007.04792*.

Vincent Sitzmann, Martina Marek, and Leonid Keselman. 2016. Multimodal natural language inference. Final paper, CS224u, Stanford University.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA. Association for Computational Linguistics.

Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA. Association for Computing Machinery.

Pierre Stock and Moustapha Cisse. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

# References V

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv prerint arXiv:2012.15761*.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada. Association for Computational Linguistics.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

# References for the benchmark timeline

## Penn Treebank (Marcus et al. 1994)

| | | |
|---|---|---|
| 1. | van Halteren 2000 | E |
| 2. | Eskin 2000 | E |
| 3. | Dickinson and Meurers 2003a | E |
| 4. | Dickinson and Meurers 2003b | E |
| 5. | Dickinson and Meurers 2005 | E |
| 6. | Boyd et al. 2008 | E |
| 7. | Manning 2011 | E |

## SNLI (Bowman et al. 2015)

| | | |
|---|---|---|
| 1. | Sitzmann et al. 2016 | A |
| 2. | Rudinger et al. 2017 | S |
| 3. | Naik et al. 2018 | G |
| 4. | Glockner et al. 2018 | G |
| 5. | Naik et al. 2018 | G |
| 6. | Poliak et al. 2018 | A |
| 7. | Tsuchiya 2018 | A |
| 8. | Gururangan et al. 2018 | A |
| 9. | Belinkov et al. 2019 | A |
| 10. | McCoy et al. 2019 | A |

## SQuAD (Rajpurkar et al. 2016, 2018)

| | | |
|---|---|---|
| 1. | Weissenborn et al. 2017 | A |
| 2. | Sugawara et al. 2018 | A |
| 3. | Bartolo et al. 2020 | A |
| 4. | Lewis et al. 2021 | A |

## ImageNet (Deng et al. 2009)

| | | |
|---|---|---|
| 1. | Deng et al. 2014 | G |
| 2. | Stock and Cisse 2018 | B |
| 3. | Yang et al. 2020 | B |
| 4. | Recht et al. 2019 | E |
| 5. | Northcutt et al. 2021 | E |
| 6. | Crawford and Paglen 2021 | B |

timeline slide