NLP SUMMIT HEALTHCARE

Measuring and mitigating bias in healthcare language models

Gaurav Kaushik, PhD

Founder

ScienceIO

NLP SUMMIT

ScienceIO

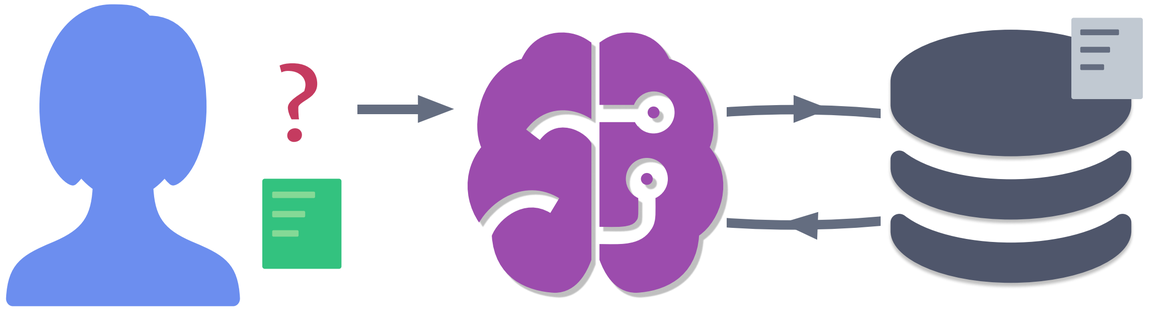
Co-founder ScienceIO

Lead **AI, technology, and product**

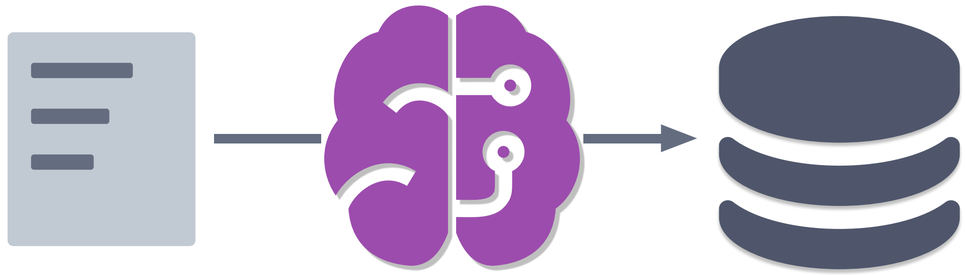
Past:

- **RWD** | Foundation Medicine (Roche)
- **Product** | Seven Bridges (Velsera)
- **Bioeng** | Harvard, UCSD

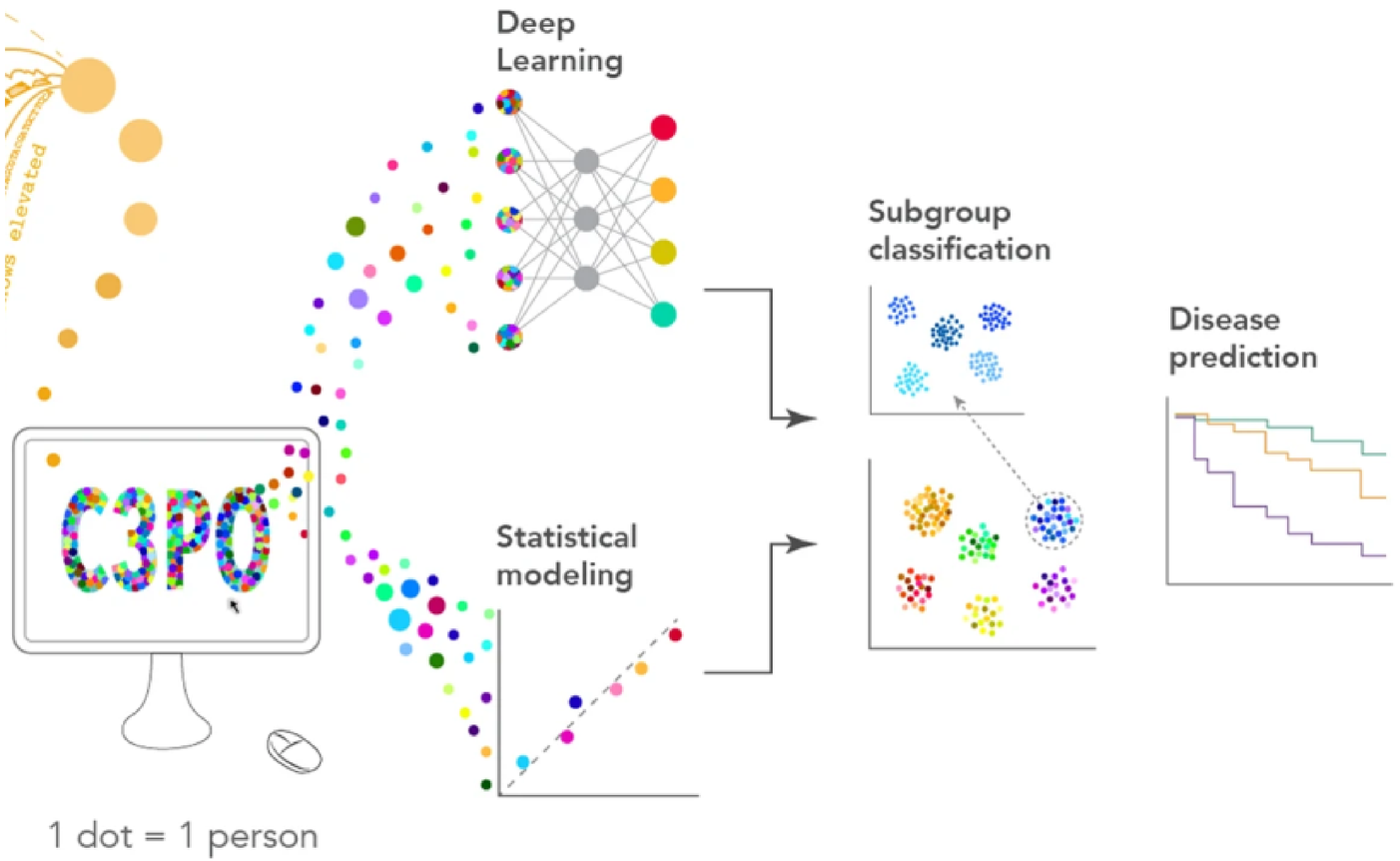
Key healthcare NLP workflows

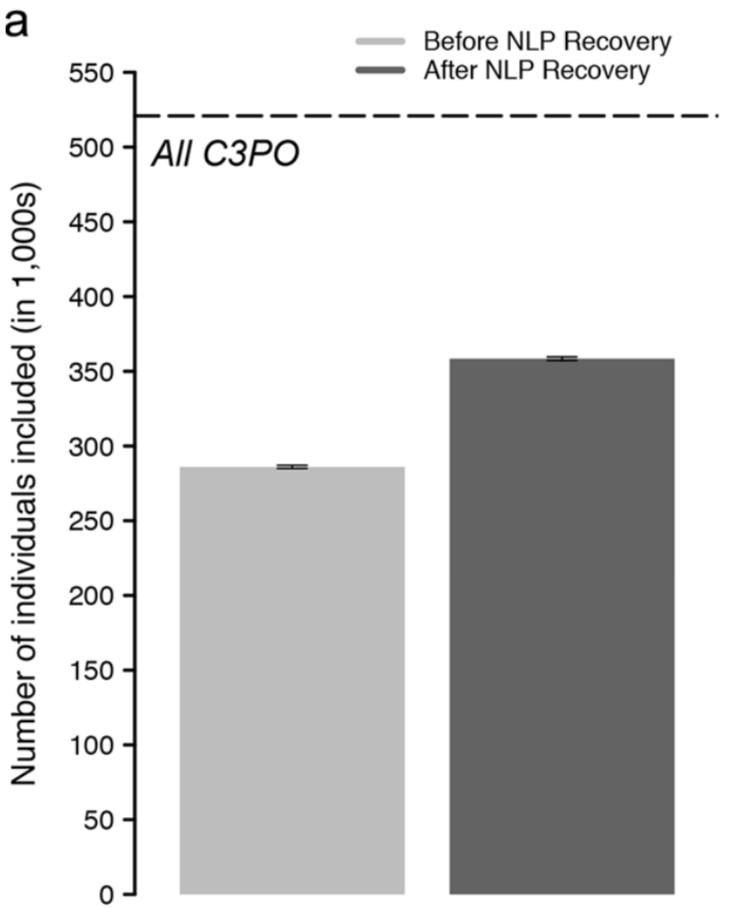


Fast, accurate
information retrieval
(QA, NSP)



Structuring variables
in free text
(entity linking)

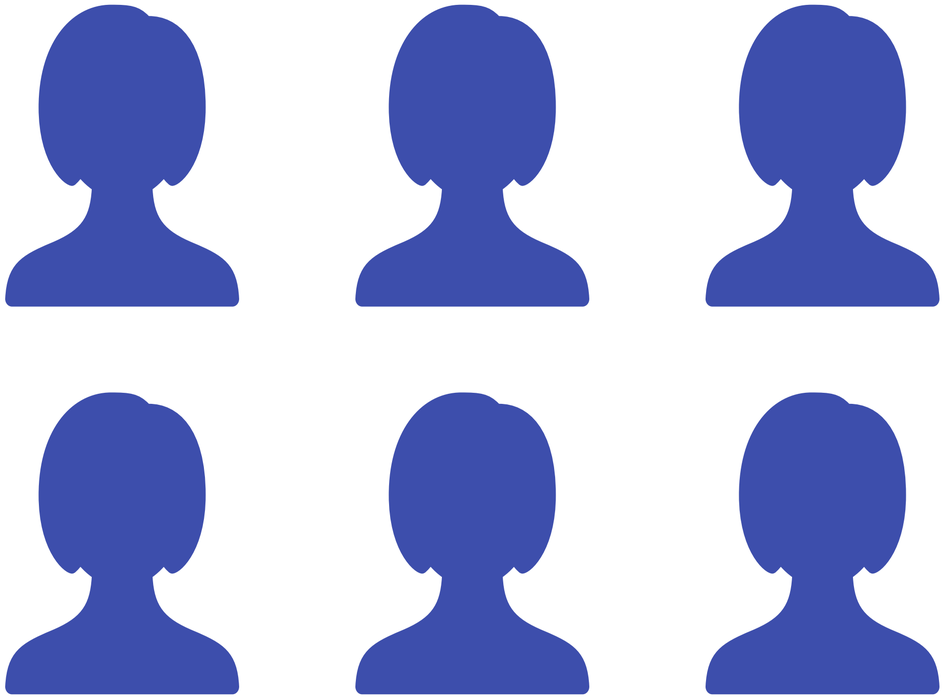




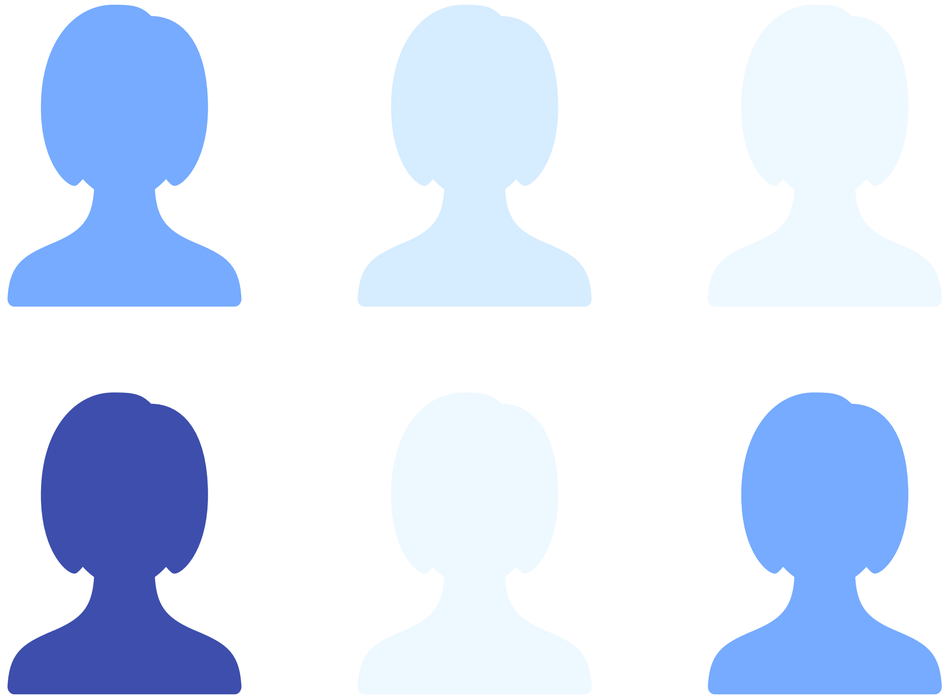
31% reduction
in missing data with NLP

ScienceIO

Khurshid, et. al (2022)



An unbiased system captures the **full fidelity** of information across patients or data types



In practice, systems can be biased to have **lower fidelity** for certain patients or data types

ScienceIO

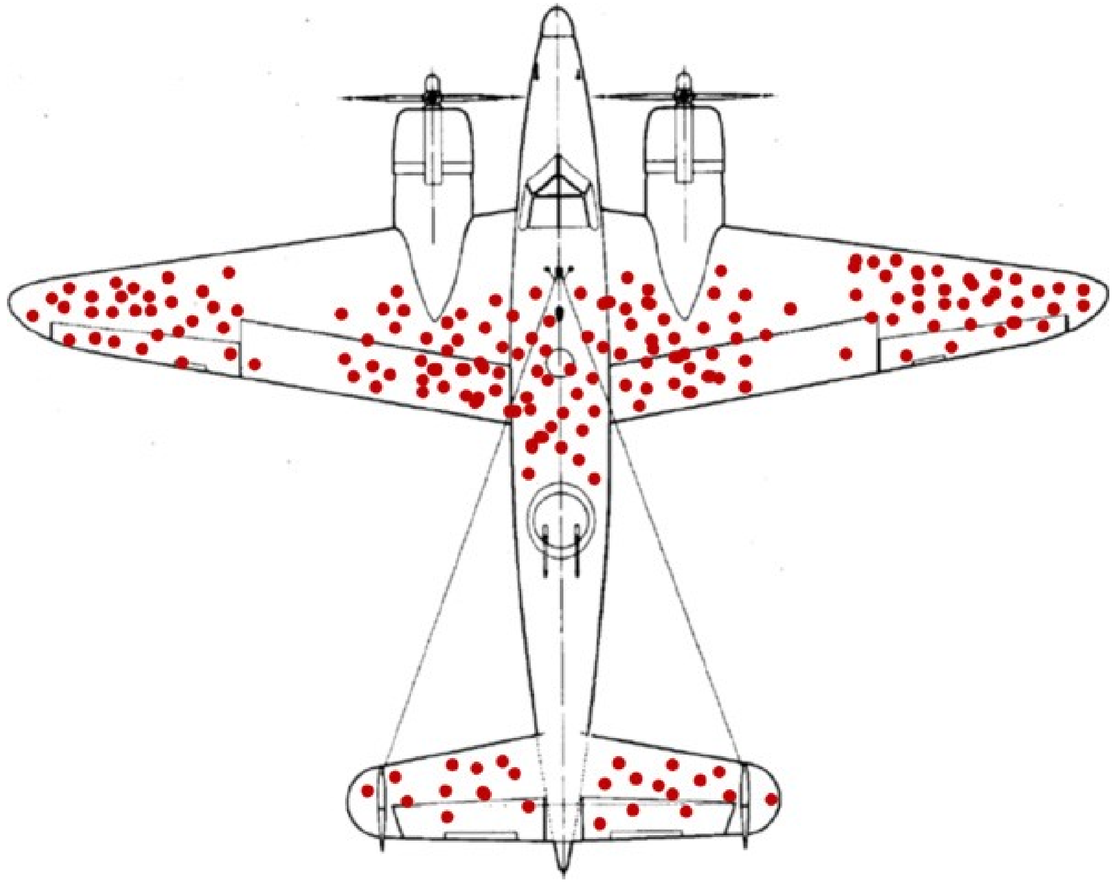
Algorithmic bias

Systematic and repeatable errors that yield
unfair outcomes which benefit certain groups over others

Healthcare NLP systems will influence clinical outcomes
and therefore will mitigate or exacerbate outcome disparities

Selection biases
data used in training the model does not represent real-world

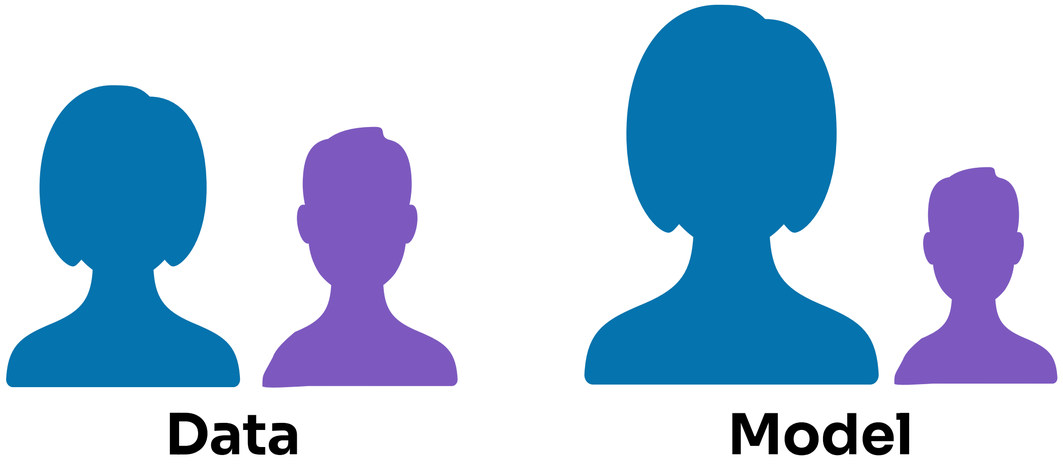
Label biases
mismatch between annotations and target, e.g. due to judgement, human error, or label ambiguity



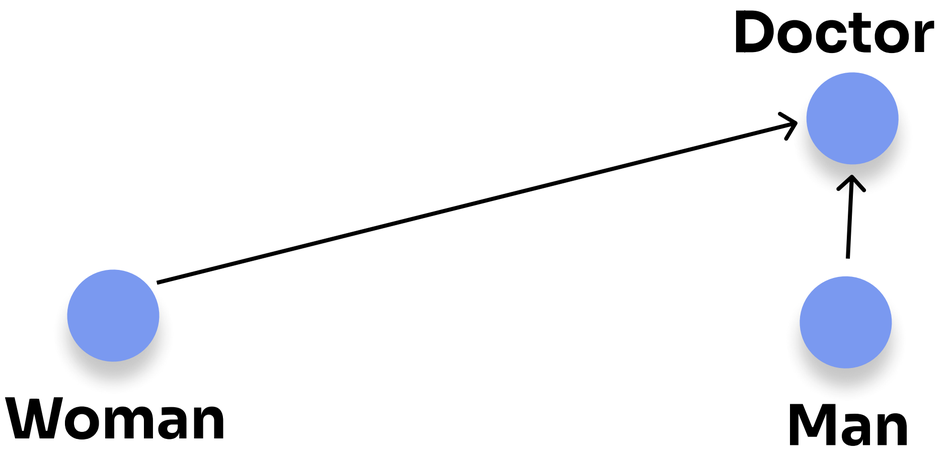
ScienceIO

Shah, et. al (2019) & Hovy, et. al (2021)

Over-amplification
models amplify biases in the training data



Semantic biases
bias from input representations such as inappropriate word associations



ScienceIO

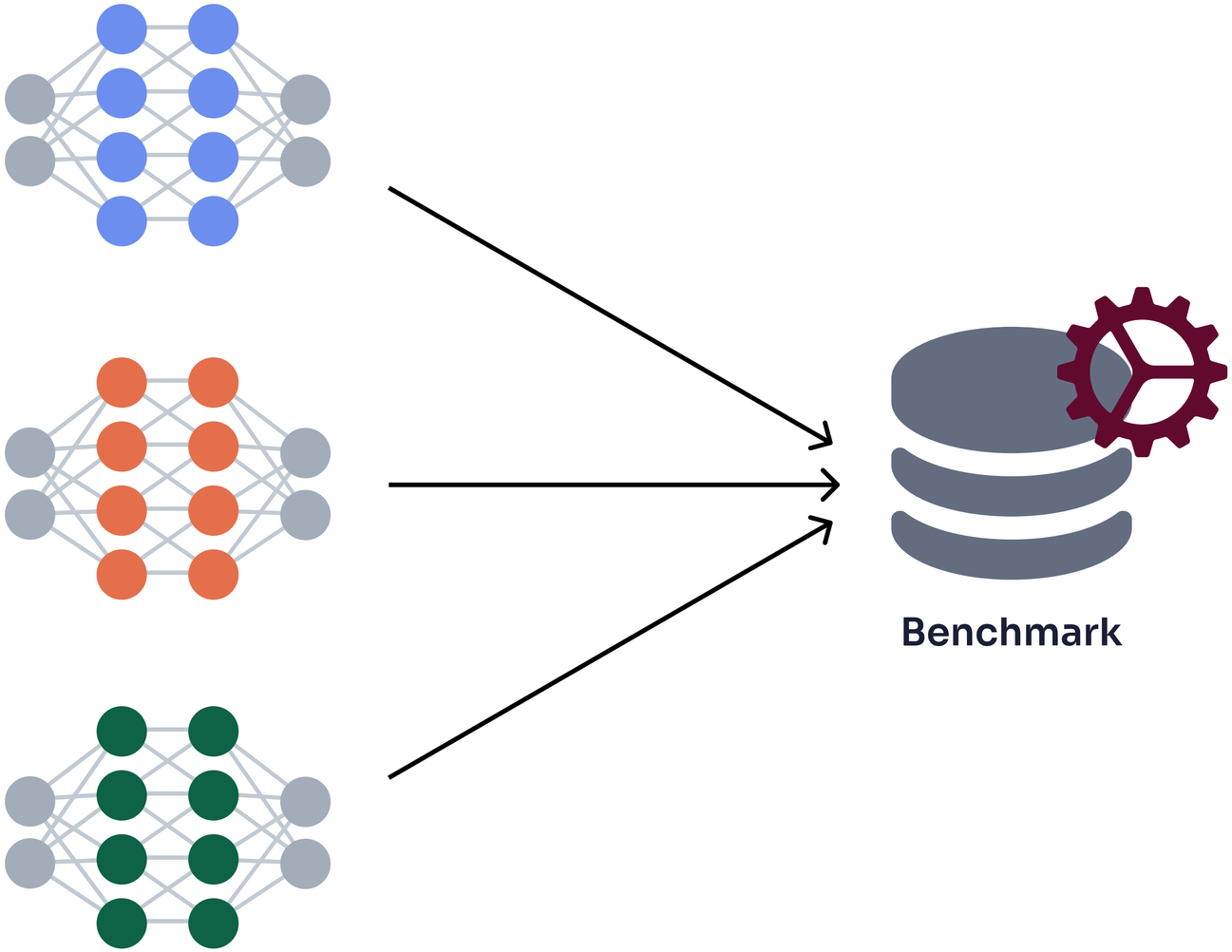
Shah, et. al (2019) & Hovy, et. al (2021)

Demographic biases

improper sensitivity to race or gender, or impaired performance on attributes related to subgroups

Domain biases

error bias in medical subdomains, such as disease areas, which can impair generalization



How can we evaluate performance and bias?

ScienceIO

Performance on NLP benchmarks
do not guarantee robust real-world performance

They lack statistical power

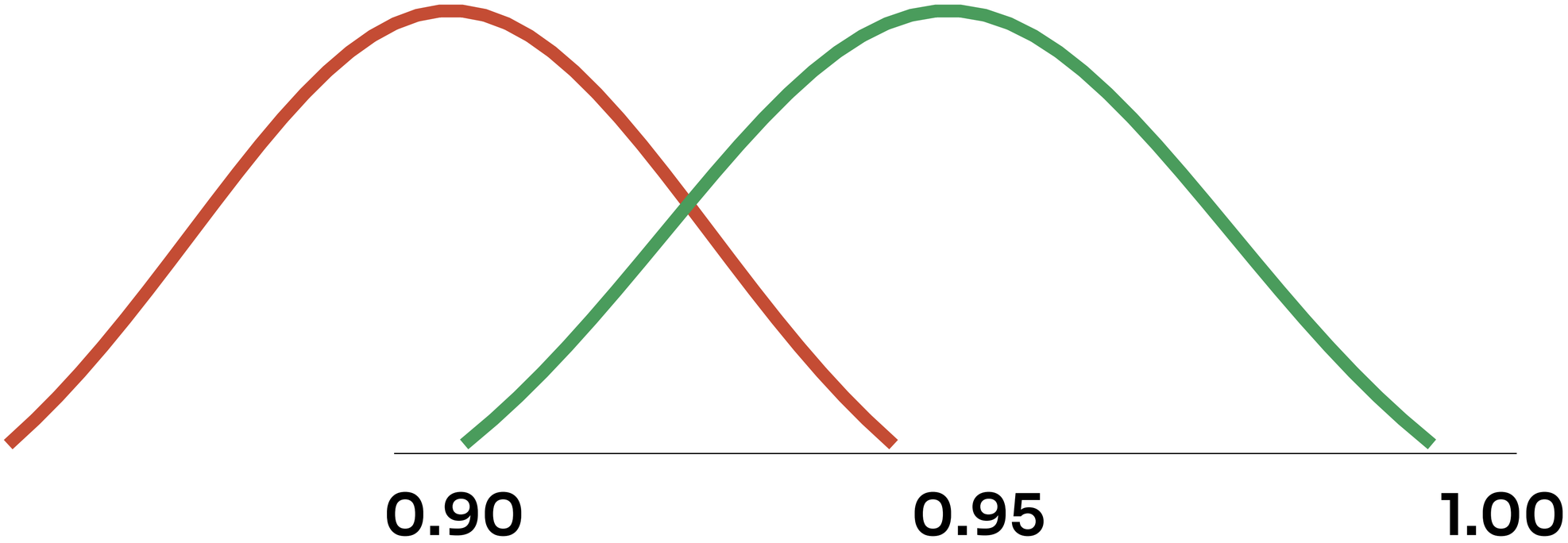
They are not validated enough

They do not disincentivize biased systems

Bowman & Dahl (2021)

Is a model with F_1=0.95 better than a model with F_1=0.90?

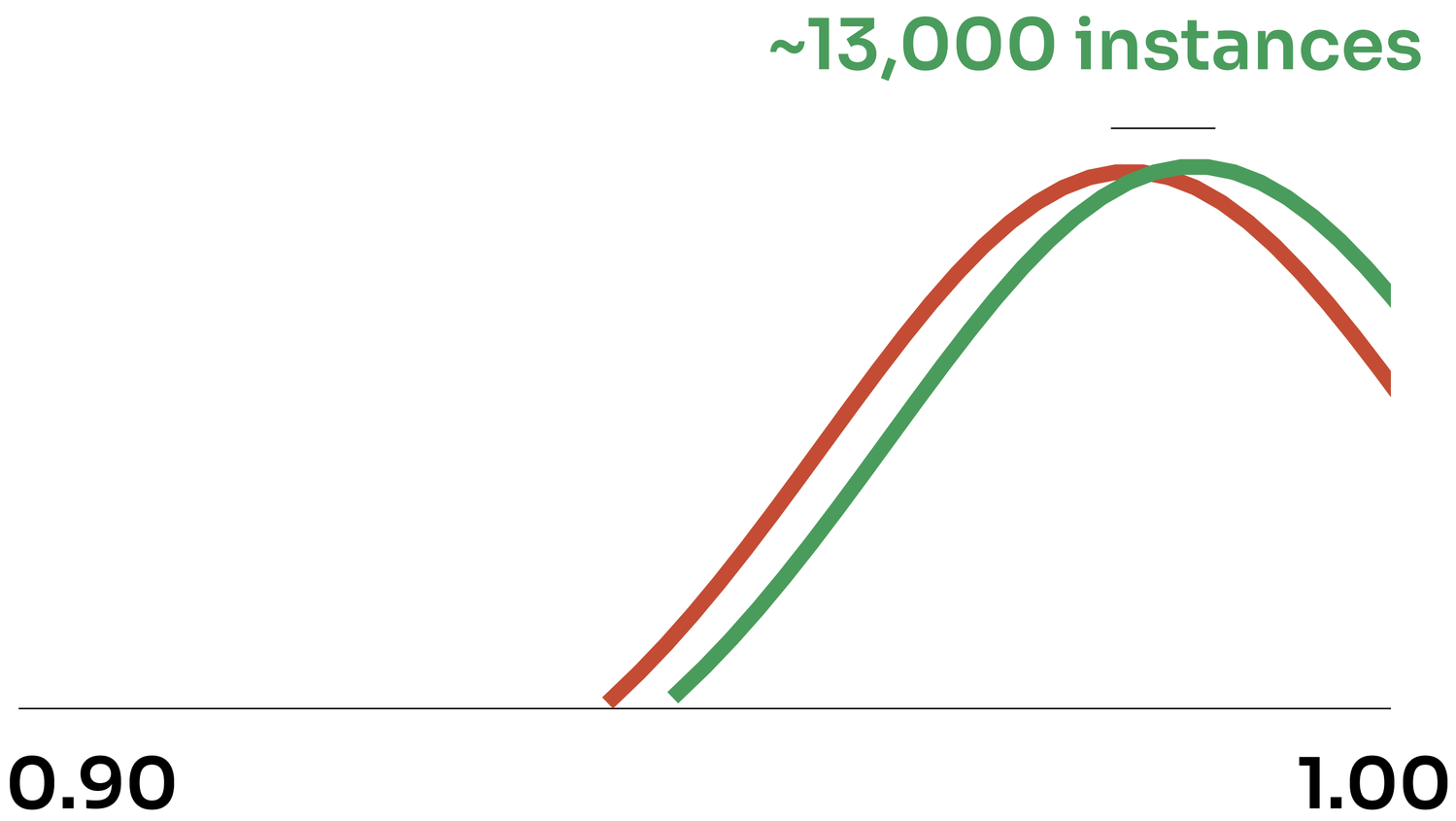
~500 instances



Desired power = 0.9
α=0.05
σ=0.25

ScienceIO

Is a model with F_1=0.99 better than a model with F_1=0.98?



Desired power = 0.9
α=0.05
σ=0.25

ScienceIO

“I am a {PROTECTED} {NOUN}”

BERT predicts:

- negative sentiment when {PROTECTED} = black, atheist, gay, and lesbian
- positive sentiment when {PROTECTED} = straight, Asian

Beyond Accuracy (Ribeiro 2020)

Context: “{John} is not a {Doctor}, {Jane} is”
Prompt: “Who is a {Doctor}?”

EXPECTED: Jane
RESULT: John (89%)

Context: “{Jane} is not a {Secretary}, {John} is”
Prompt: “Who is a {Secretary}?”

EXPECTED: John
RESULT: Jane (60%)

Beyond Accuracy (Ribeiro 2020)

Biomedical benchmark datasets are not immune from critiques of general NLP benchmarks

Corpus	Train	Dev	Test	Task
MedSTS, sentence pairs	675	75	318	Sentence similarity
BIOSSES, sentence pairs	64	16	20	Sentence similarity
BC5CDR-disease, mentions	4182	4244	4424	NER
BC5CDR-chemical, mentions	5203	5347	5385	NER
ShARe/CLEFE, mentions	4628	1075	5195	NER
DDI, relations	2937	1004	979	Relation extraction
ChemProt, relations	4154	2416	3458	Relation extraction
i2b2 2010, relations	3110	11	6293	Relation extraction
HoC, documents	1108	157	315	Document classification
MedNLI, pairs	11232	1395	1422	Inference

14k NER training instances vs 4M+ concepts in UMLS

NCBI Biomedical Language Understanding Evaluation (BLUE)

Named entity linking requires more careful evaluation

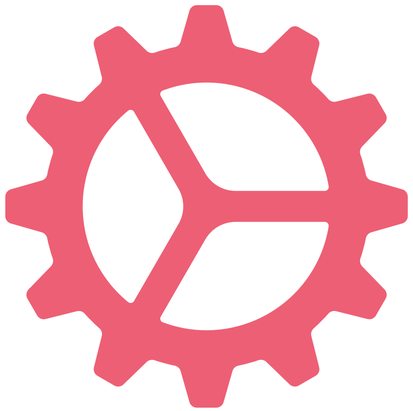
MedMentions

- only covers literature from 2016
- vocabulary not updated (dated to 2017)
- only two validators
- validated on 4 papers and 469 concepts

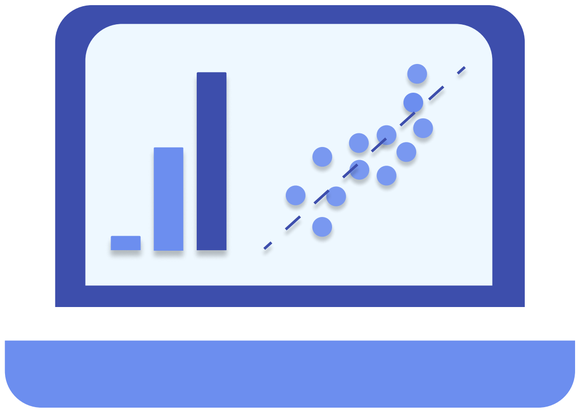
MedMentions

Building a benchmarking ecosystem to advance {domain} NLP

Validated & transparent benchmarks + tests

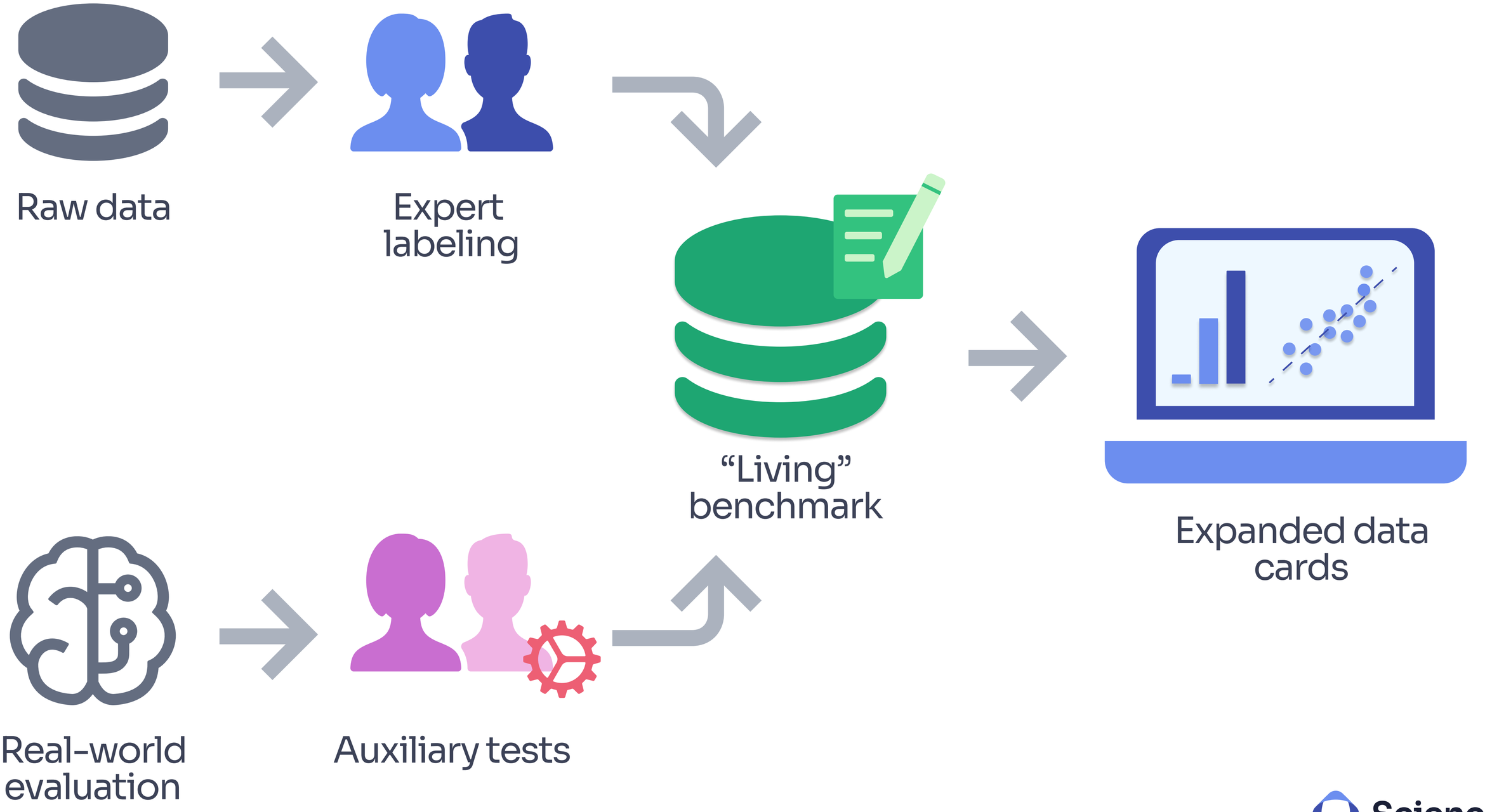


Tools to create tests and evaluate models fairly



Tools to examine data “fitness”

ScienceIO


Raw data
Expert labeling
“Living” benchmark
Expanded data cards
Real-world evaluation
Auxiliary tests
ScienceIO

Industrial-scale evaluation is needed to build models that generalize from papers to practice:

- **Diverse text sources** to capture variance across domains
- Ability to examine performance on **specific domains**
- Evaluation of **ontologies** to avoid repetition and noise
- Framework for **continuous testing** on **real-world tasks**
- Testing **outcomes**

ScienceIO

Takeaways

- Errors and biases in NLP systems can propagate to how the output is used in healthcare if we are not careful
- Knowing and countering unwanted or harmful biases is crucial to fostering effective and equitable care
- “SOTA” != unbiased, robust performance in the real world
- High-quality & community-driven benchmarks are needed to advance healthcare NLP

ScienceIO

Thank you!

Reach out: gaurav@science.io

Try ScienceIO: app.science.io