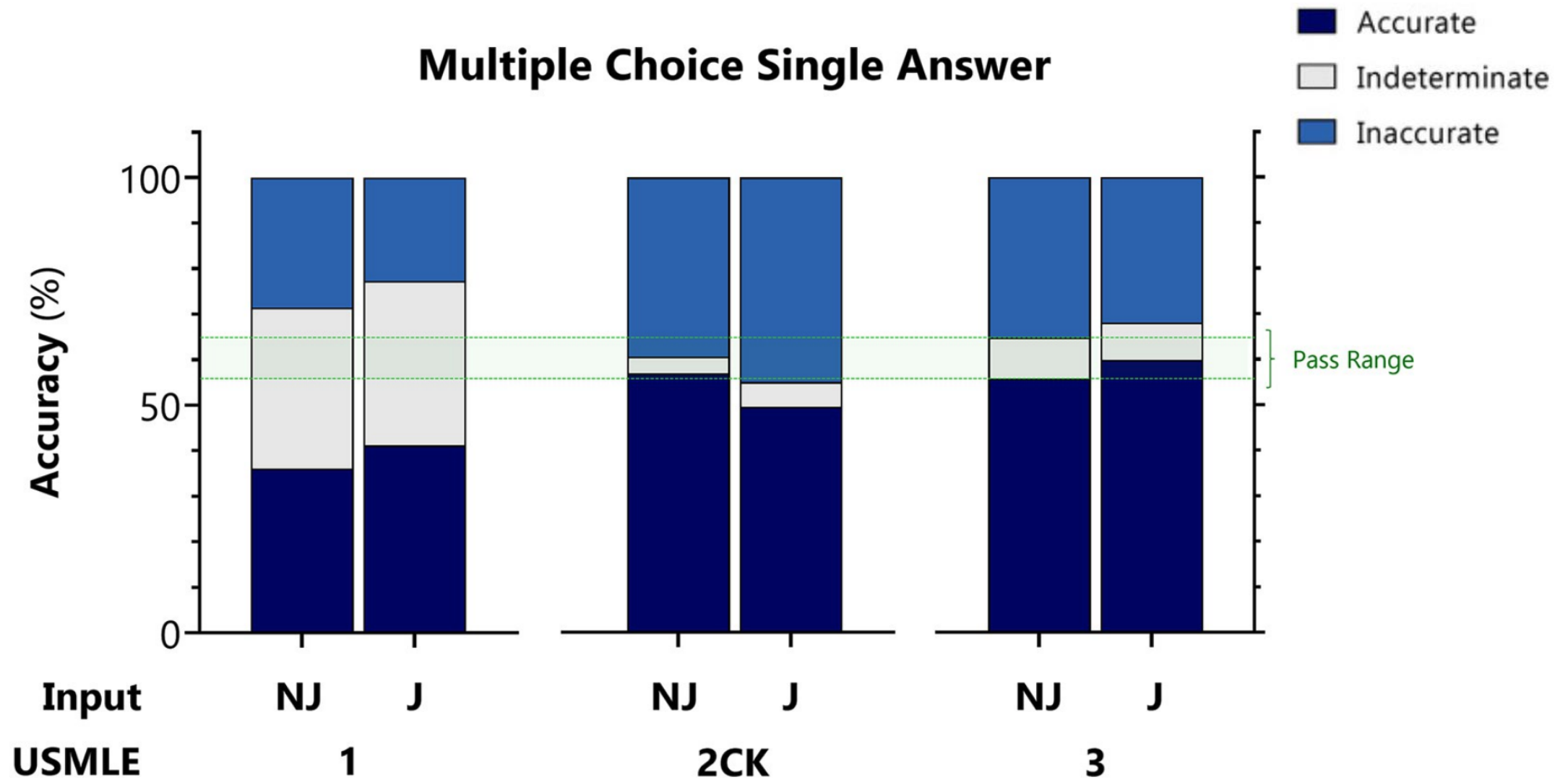




Data 146: Foundations for CPH **Benchmarking**

Irene Y. Chen



What makes a "good" benchmark?

What makes a “good” benchmark?

1. Publicly available datasets
2. Clear and relevant metrics
3. Easily automated for quick evaluations
4. Investigate reliability, generalizability, and fairness of models
5. Multimodal or incorporating different data types
6. Dynamic to evolving AI models
7. Source of centralized resources to annotate and create tasks

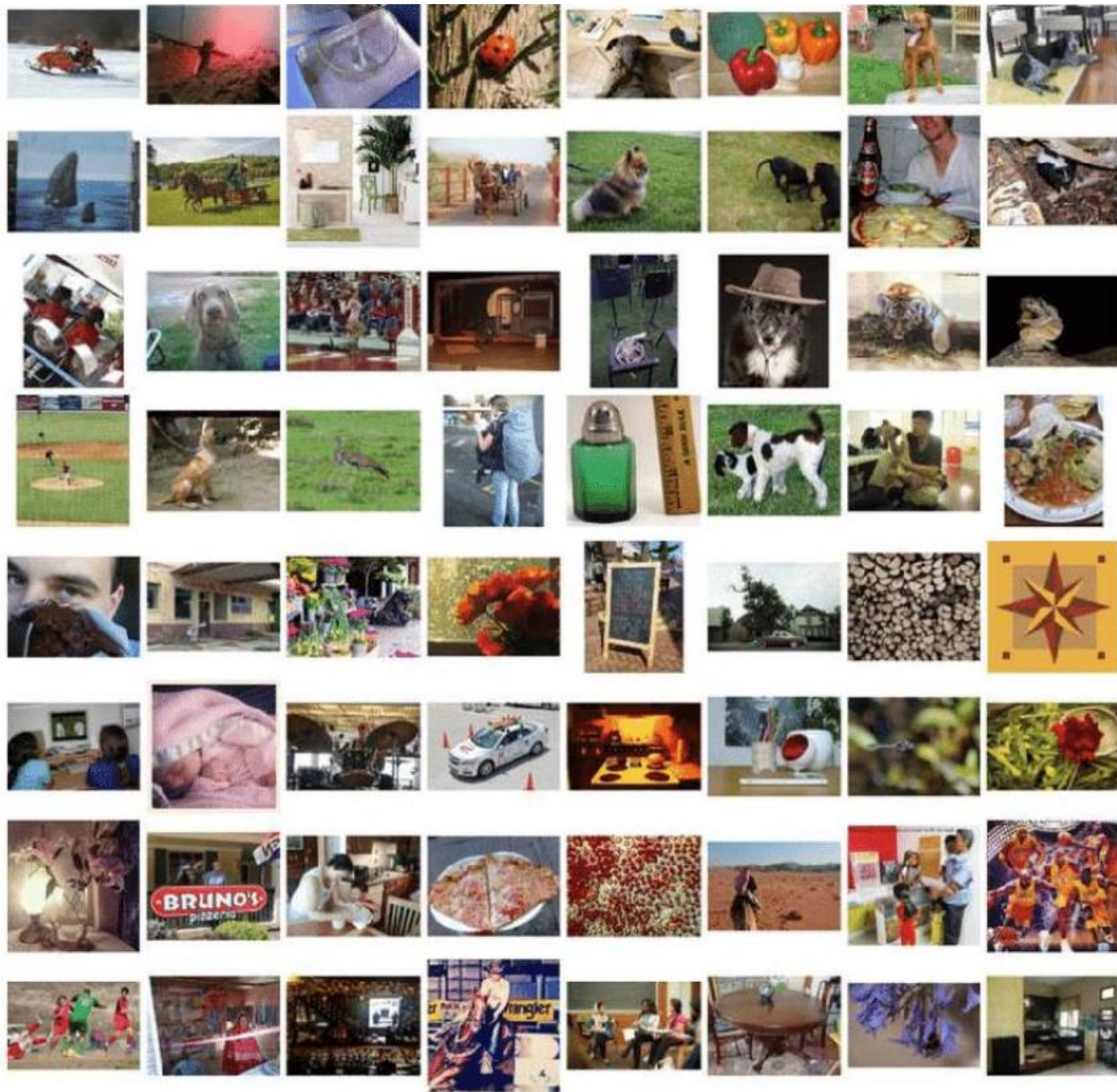
Outline

- **Lessons from ImageNet** (20 mins)
- **Health benchmarks** (10 mins)
- **LLM benchmarks** (15 mins)
- **Discussion** (5 mins)



How can we make Data
146 better for you?

Learning Objective: Understand criteria for health AI benchmarks

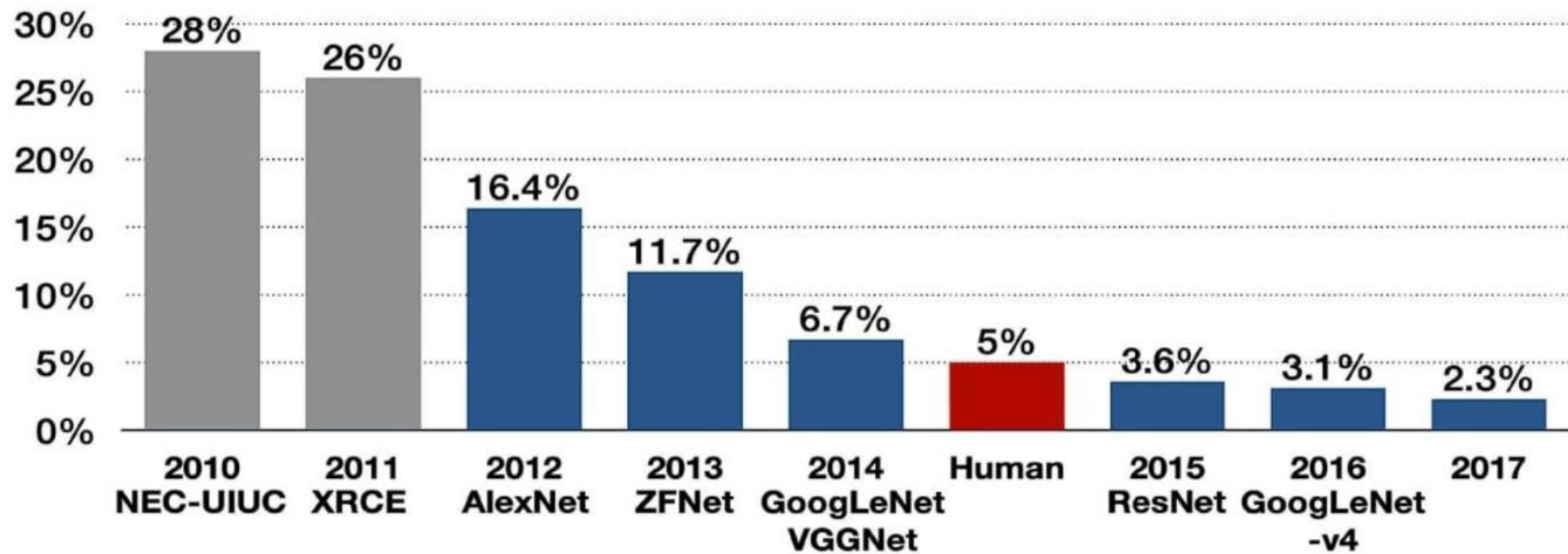


ImageNet

- Created by Fei-Fei Li in 2006
- Contains 14 million images, 20k groups
- 120 categories of dog breeds
- Images scraped from online image search (Google, Flickr, Yahoo)
- Labeled with Amazon Mechanical Turk



Top-5 error

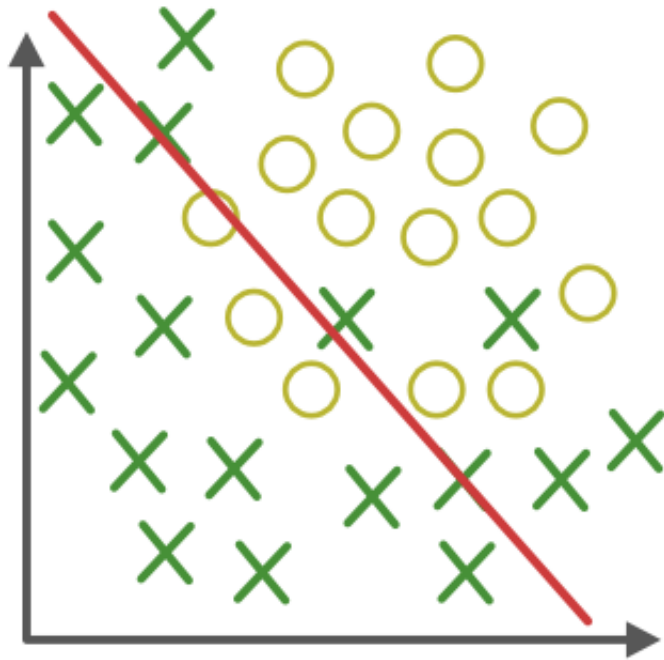


Do ImageNet Classifiers Generalize?

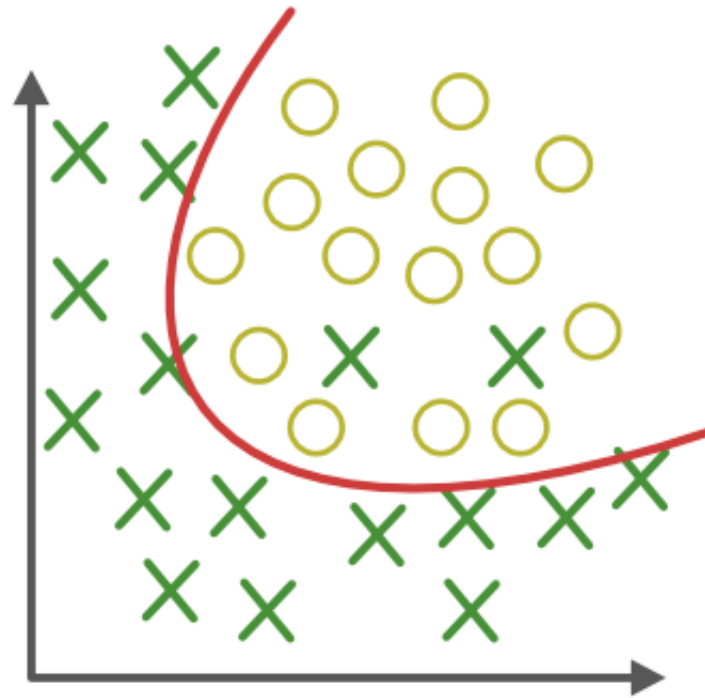
- Classifiers trained and tested on ImageNet may be overfitting
- Increases in performance may only be learning about internet photos from these specific sources (e.g., camel on sand vs grass)
- When does the benchmark stop serving the purpose?

What is overfitting?

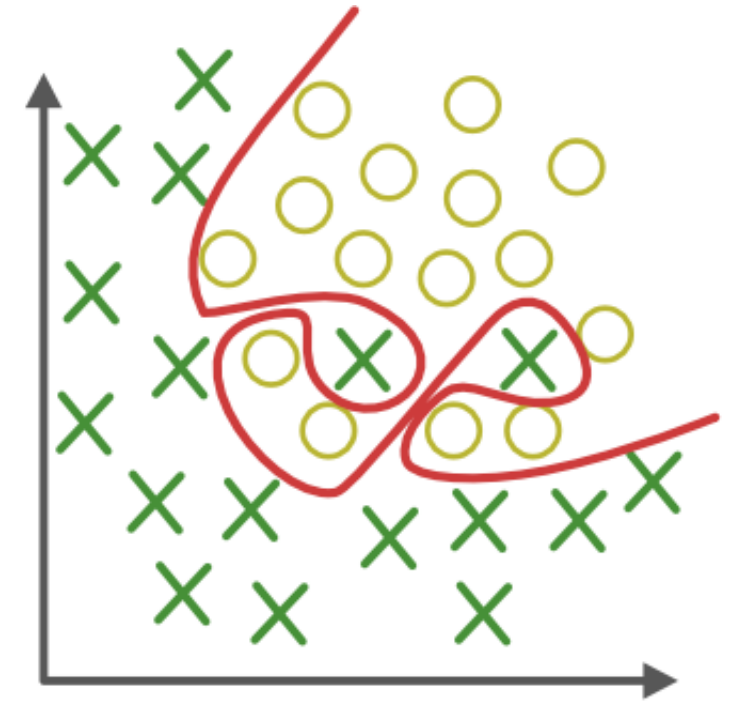
What is overfitting?



Under-fitting
(too simple to
explain the variance)



Appropriate-fitting



Over-fitting
(forcefitting--too
good to be true) 

What is overfitting in the age of deep learning?

1. In higher dimensional data or with models with many parameters, do we always overfit? Not necessarily!
 - Example: neural networks
2. Instead, think of “overfitting” as a form of generalizing from source to target dataset

What is overfitting?

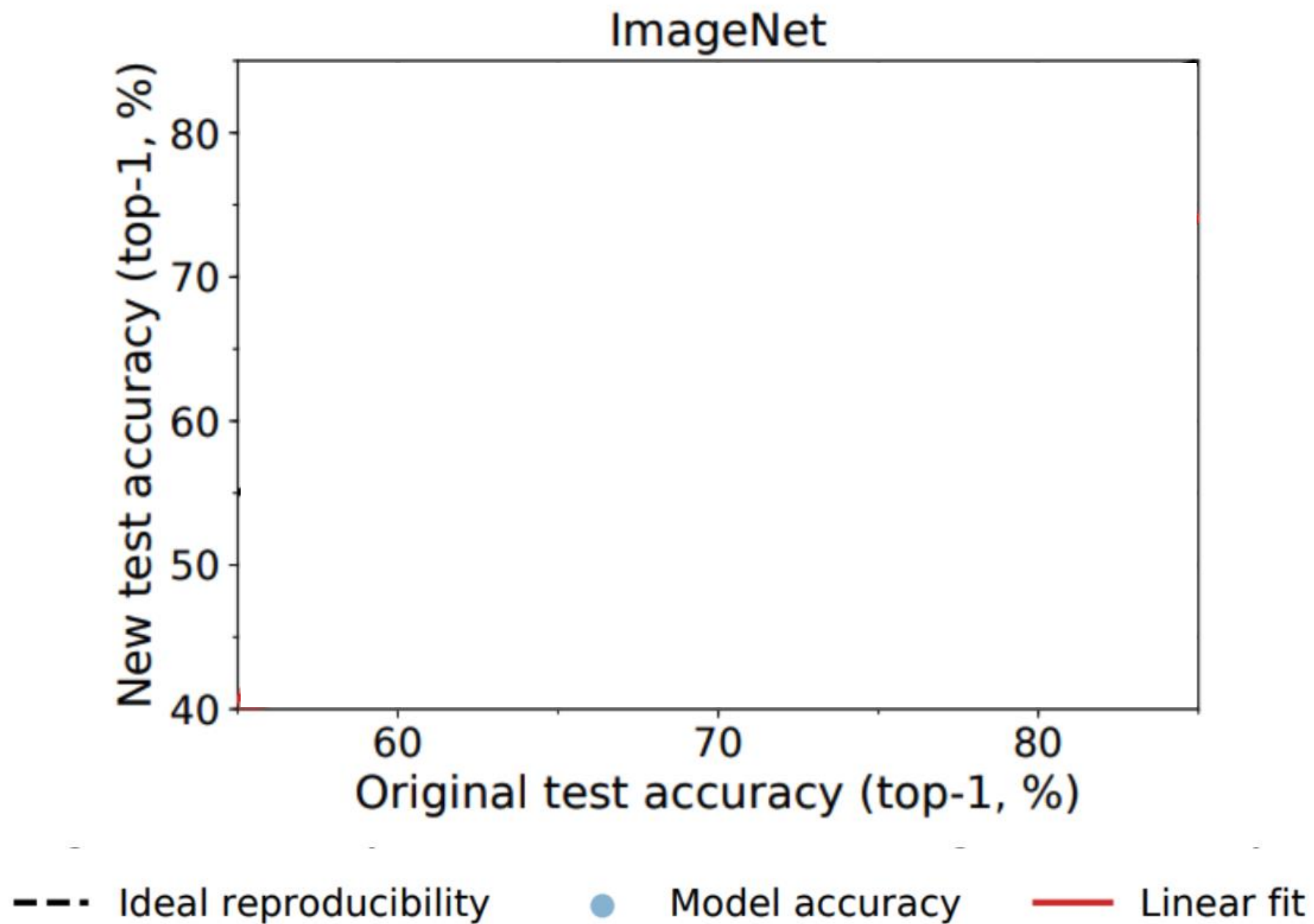
- Given data points (x_i, y_i) drawn i.i.d from an unknown distribution $P(X, Y)$
- A model $f_\theta(x)$ with parameters θ predicts y from x
- If we define loss function $L(f_\theta(x), y)$:
 - Expected risk is $R(\theta) = E_{(x,y) \sim P}[L(f_\theta(x), y)]$
 - Empirical risk is $\hat{R}_n(\theta) = \frac{1}{n} \sum L(f_\theta(x_i), y_i)$
- Overfitting happens when $\hat{R}_n(\theta) \ll R(\theta)$

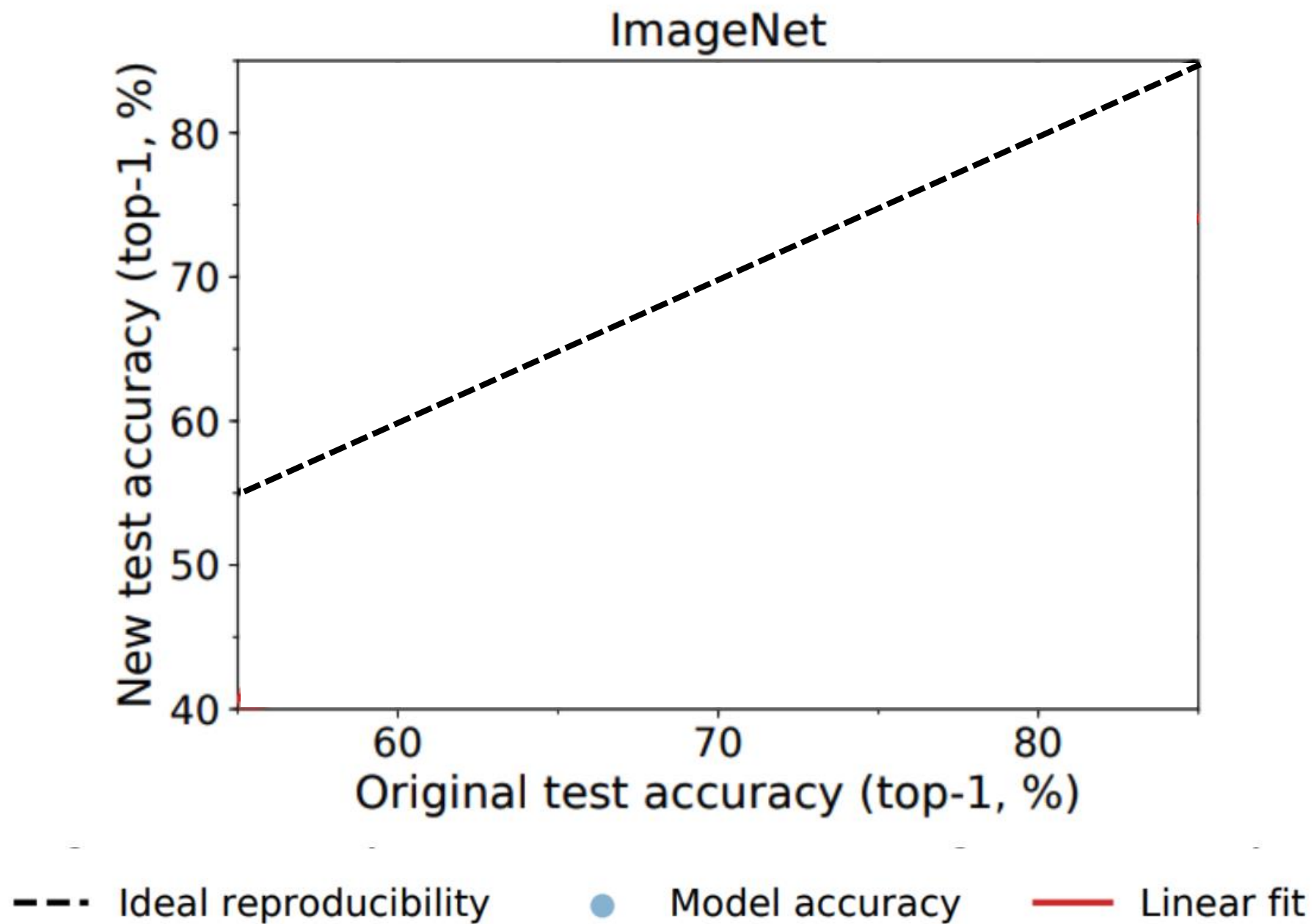
Do ImageNet Classifiers Generalize?

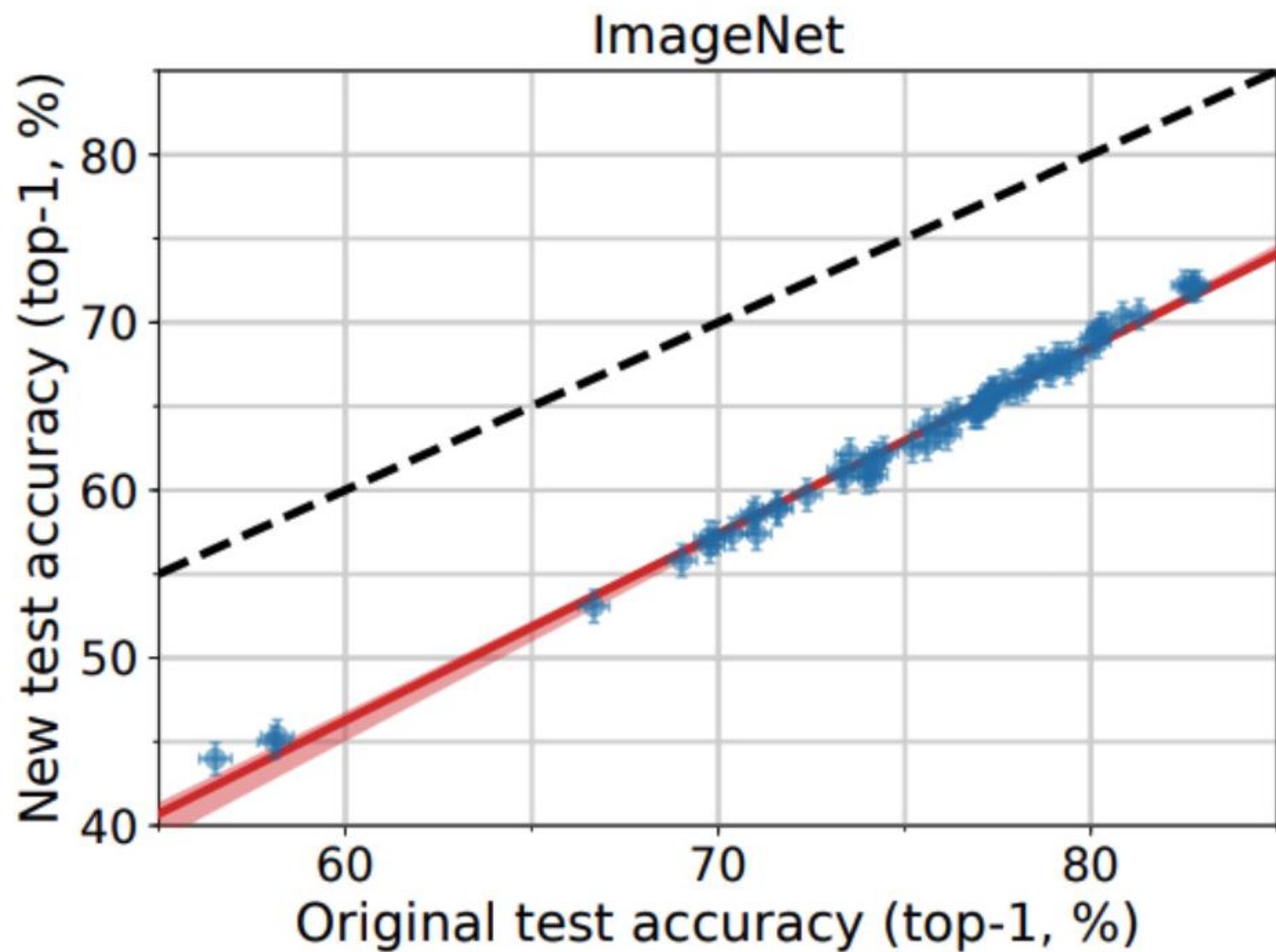
Do ImageNet Classifiers Generalize **to** **ImageNet**?

Do ImageNet Classifiers Generalize **to ImageNet?**

- Researchers were interested in generalization
 - Do overuse of a benchmark dataset and benchmark task lead to models that generalize?
 - **Corollary: Do models perform as well on the same dataset?**
- Researchers recreated data collection using the protocols of CIFAR-10 (Tiny Images) and ImageNet (Flickr)
 - Used Mechanical Turk labelers







--- Ideal reproducibility ● Model accuracy — Linear fit

Do ImageNet Classifiers Generalize to ImageNet?

1. Overall performance goes down
2. Relative ordering of models stays consistent
3. Models do NOT seem to be overfitting

What makes a “good” benchmark?

1. Publicly available datasets
2. Clear and relevant metrics
3. Easily automated for quick evaluations
4. Investigate reliability, generalizability, and fairness of models
5. Multimodal or incorporating different data types
6. Dynamic to evolving AI models
7. Source of centralized resources to annotate and create tasks

What makes a “good” **healthcare** benchmark?

1. Publicly available datasets
2. Clear and relevant metrics
3. Easily automated for quick evaluations
4. Investigate reliability, generalizability, and fairness of models
5. Multimodal or incorporating different data types
6. Dynamic to evolving AI models
7. Source of centralized resources to annotate and create tasks

What makes a “good” **healthcare** benchmark?

1. Publicly available datasets
2. **Clear and relevant metrics**
3. Easily automated for quick evaluations
4. Investigate reliability, generalizability, and fairness of models
5. Multimodal or incorporating different data types
6. Dynamic to evolving AI models
7. Source of centralized resources to annotate and create tasks

What makes a “good” **healthcare** benchmark?

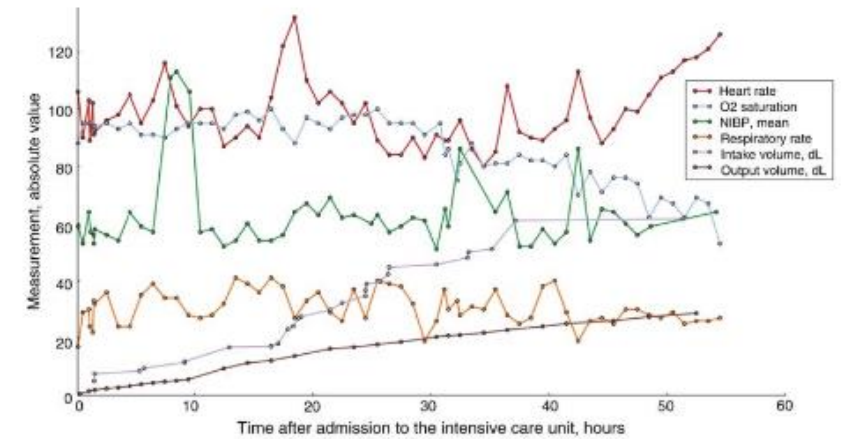
1. Publicly available datasets
2. Clear and relevant metrics
3. **Easily automated for quick evaluations**
4. Investigate reliability, generalizability, and fairness of models
5. Multimodal or incorporating different data types
6. Dynamic to evolving AI models
7. Source of centralized resources to annotate and create tasks

Healthcare Benchmark Datasets

- MIMIC-III/MIMIC-IV (ICU data)
- eICU Collaborative Research Database
- CheXpert, ChestX-ray14 (Medical imaging)
- PhysioNet Challenges (ECG, sepsis prediction)
- USMLE, NEJM Clinical Questions, MedQA (Natural language)

MIMIC

- Developed in 2016 by a team at MIT
- Includes 53k adult patients in ICU at Beth Israel Deaconess Medical Center (BIDMC) from 2001-2012 and 7800 neonates
- Includes clinical notes, signals, diagnoses, and full ICU stay information



MIMIC Benchmarking

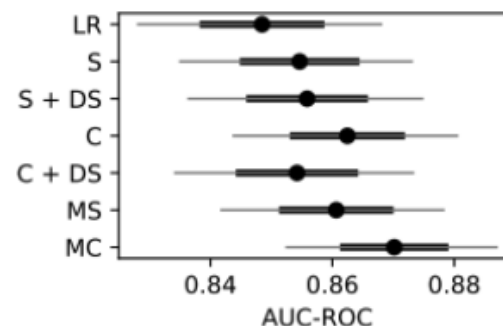
LR – logistic regression
S – standard LSTM

C – channel-wise LSTM
DS – deep supervision

MS – multitask standard LSTM
MC – multitask channel-wise LSTM

In-hospital Mortality

Model	AUC-ROC	AUC-PR
SAPS	0.720 (0.720, 0.720)	0.301 (0.301, 0.302)
APS-III	0.750 (0.750, 0.750)	0.357 (0.356, 0.357)
OASIS	0.760 (0.760, 0.761)	0.311 (0.311, 0.312)
SAPS-II	0.777 (0.776, 0.777)	0.376 (0.376, 0.377)
LR	0.848 (0.828, 0.868)	0.474 (0.419, 0.529)
S	0.855 (0.835, 0.873)	0.485 (0.431, 0.537)
S + DS	0.856 (0.836, 0.875)	0.493 (0.438, 0.549)
C	0.862 (0.844, 0.881)	0.515 (0.464, 0.568)
C + DS	0.854 (0.834, 0.873)	0.502 (0.447, 0.554)
MS	0.861 (0.842, 0.878)	0.493 (0.439, 0.548)
MC	0.870 (0.852, 0.887)	0.533 (0.480, 0.584)



	LR	S	S + DS	C	C + DS	MS	MC
LR	-	15.8	12.9	1.9	18.1	4.1	0.0
S	84.2	-	38.5	3.4	53.8	13.7	0.2
S + DS	87.1	61.5	-	10.7	63.4	19.2	0.6
C	98.1	96.6	89.3	-	95.5	62.3	4.7
C + DS	81.9	46.2	36.6	4.5	-	12.9	0.1
MS	95.9	86.3	80.8	37.7	87.1	-	0.9
MC	100.0	99.8	99.4	95.2	99.9	99.1	-

A 57-year-old man develops increasing anxiety, hand tremor, and nausea four days following an uncomplicated total knee replacement surgery. The patient is currently undergoing rehabilitation at a nearby facility and was transferred to the emergency department. Upon initial assessment, the patient states “There goes a bear!” The patient is currently receiving a continuous infusion of intravenous morphine. Past medical history includes advanced liver disease. The patient drinks seven cans of beer daily. He does not use tobacco or illicit drugs. Temperature is 37.5°C (99.5°F), pulse is 106/min, respirations are 19/min and blood pressure is 177/60 mmHg. Physical examination shows a confused middle-aged man with slight tremulousness and diaphoresis. Which of the following examination findings may also be present in this patient?

- A. Miosis**
- B. Nasal septum perforation**
- C. Tongue fasciculations**
- D. Vertical nystagmus**
- E. Conjunctival injection**



A 57-year-old man develops increasing anxiety, hand tremor, and nausea four days following an uncomplicated total knee replacement surgery. The patient is currently undergoing rehabilitation at a nearby facility and was transferred to the emergency department. Upon initial assessment, the patient states “There goes a bear!” The patient is currently receiving a continuous infusion of intravenous morphine. Past medical history includes advanced liver disease. The patient drinks seven cans of beer daily. He does not use tobacco or illicit drugs. Temperature is 37.5°C (99.5°F), pulse is 106/min, respirations are 19/min and blood pressure is 177/60 mmHg. Physical examination shows a confused middle-aged man with slight tremulousness and diaphoresis. Which of the following examination findings may also be present in this patient?

- A. Miosis
- B. Nasal septum perforation
- C. Tongue fasciculations**
- D. Vertical nystagmus
- E. Conjunctival injection



Are medical exams good benchmarks?

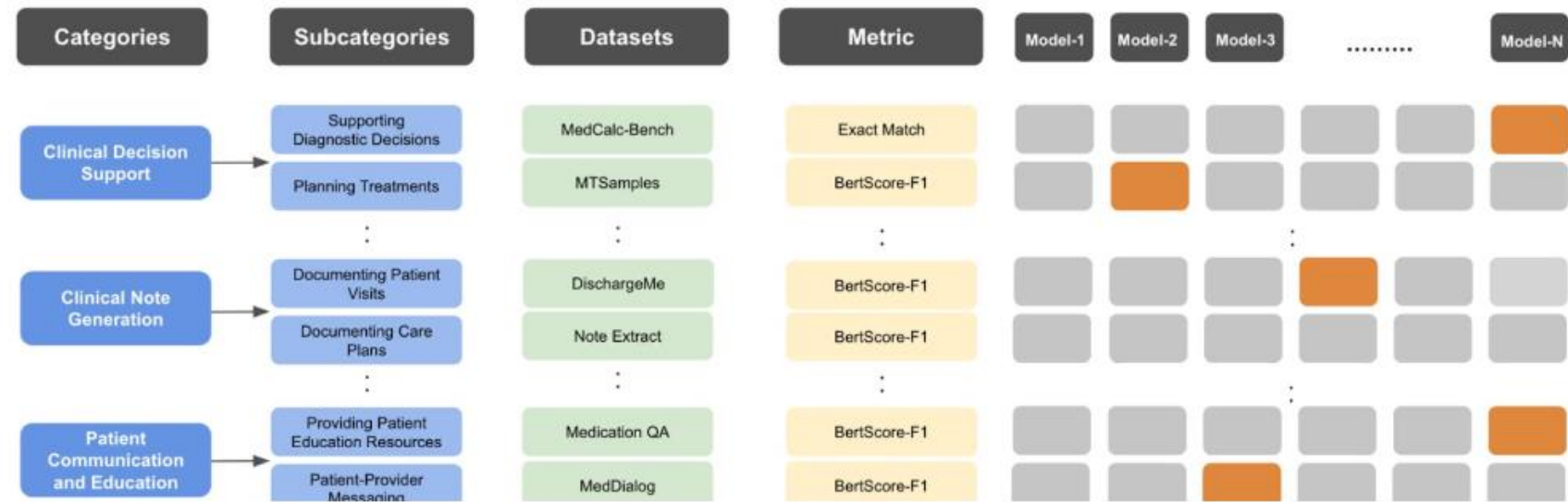
Are medical exams good benchmarks?

1. Ignores complexity of real-world clinical practice
 - Contains complete data, non-iterative
 - Well-written narrative
 - Humans who score better on USMLE don't make better doctors
2. LLMs are used for wider array of clinical tasks
 - Different formats of real clinical data
 - LLMs are sensitive to answer choice order
3. Medical exams are for clinicians, not all healthcare workers
 - Wide range of skills and expertise needed

MedHELM

Medical and AI experts build a benchmark for evaluation of LLMs grounded in real-world healthcare needs.

We introduce **MedHELM**, a comprehensive healthcare benchmark to evaluate language models on real-world clinical tasks using real electronic health records. Building on the HELM framework, **MedHELM** comprises a structured taxonomy with 5 categories, 22 subcategories, and 121 distinct clinical tasks as well as 31 diverse datasets (12 private, 6 gated-access, and 13 public). The datasets represent a spectrum of healthcare scenarios, from diagnostic decision-making to patient communication, providing a more nuanced and clinically relevant assessment of AI capabilities in healthcare settings.



Model	Mean win rate
GPT-4o (2024-05-13)	0.72 🔗
GPT-4o mini (2024-07-18)	0.623 🔗
Llama 3.3 Instruct (70B)	0.582 🔗
Gemini 1.5 Pro (001)	0.398 🔗
Qwen2.5 Instruct (7B)	0.296 🔗
Phi-3.5 Mini	0.27 🔗
SEE MORE	

MedHELM Metrics

- For exact match datasets, check against “correct answer”
- For open-ended text generation, use either:
 - String-based metrics (BLEU, ROUGE, METEOR)
 - Semantic similarity (BERTScore)

BLEU Evaluation Metrics

Reference (Human) translation:

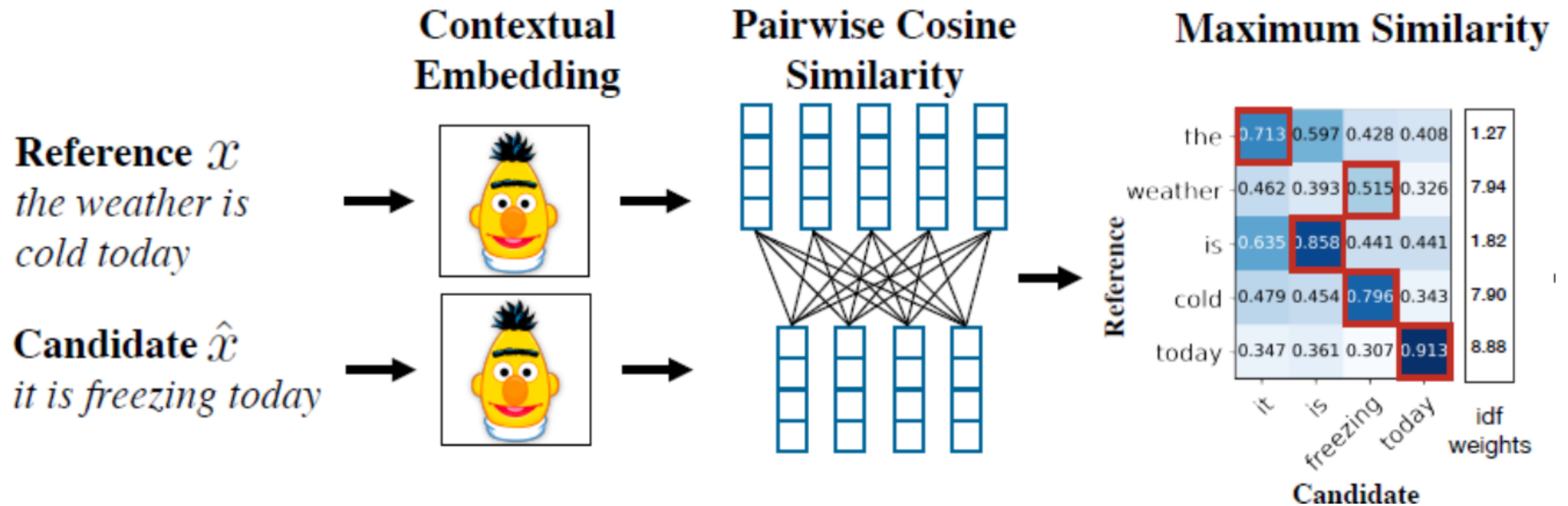
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- Score is between 0 and 1 (sometimes normalized to a number between 0 and 100)
- **What percentage of MT output n-grams (text string clusters) can be found in the reference translation?**
- Usually calculated on ~1000 test sentences.
- Important to reward the right things and there is brevity penalty
- Getting larger word clusters to match provides better scores

Example from C. Wayne



Medical Summarization Tasks

- Given the original note, how good are the medical summaries?
- If two summaries (LLM-generated and gold-standard) are deemed “close” by BERTScore, are they also deemed “close” by humans?
- Pearson correlation:
 - 0 = no correlation
 - 1 = perfect correlation

Metric Name	Dimension
BERTScore	Completeness
	Correctness
BLEU	Completeness
	Correctness
ROUGE	Completeness
	Correctness

Metric Name	Dimension	Pearson Correlation
BERTScore	Completeness	0.28 ³² , 0.44 ³³
	Correctness	0.23 ³² , 0.52 ³³ , 0.022 ³⁶
BLEU	Completeness	0.22 ³²
	Correctness	0.13 ³²
ROUGE	Completeness	0.30 ³² , 0.42 ³³ , 0.479 ³⁷
	Correctness	0.16 ³² , 0.53 ³³ , -0.01 ³⁶ , 0.416 ³⁷

Discussion: How would you develop a benchmark?

Summary

- Lessons from ImageNet
- Health benchmarks
- LLM benchmarks
- Discussion



How can we make Data
146 better for you?

Next class: Bias and fairness