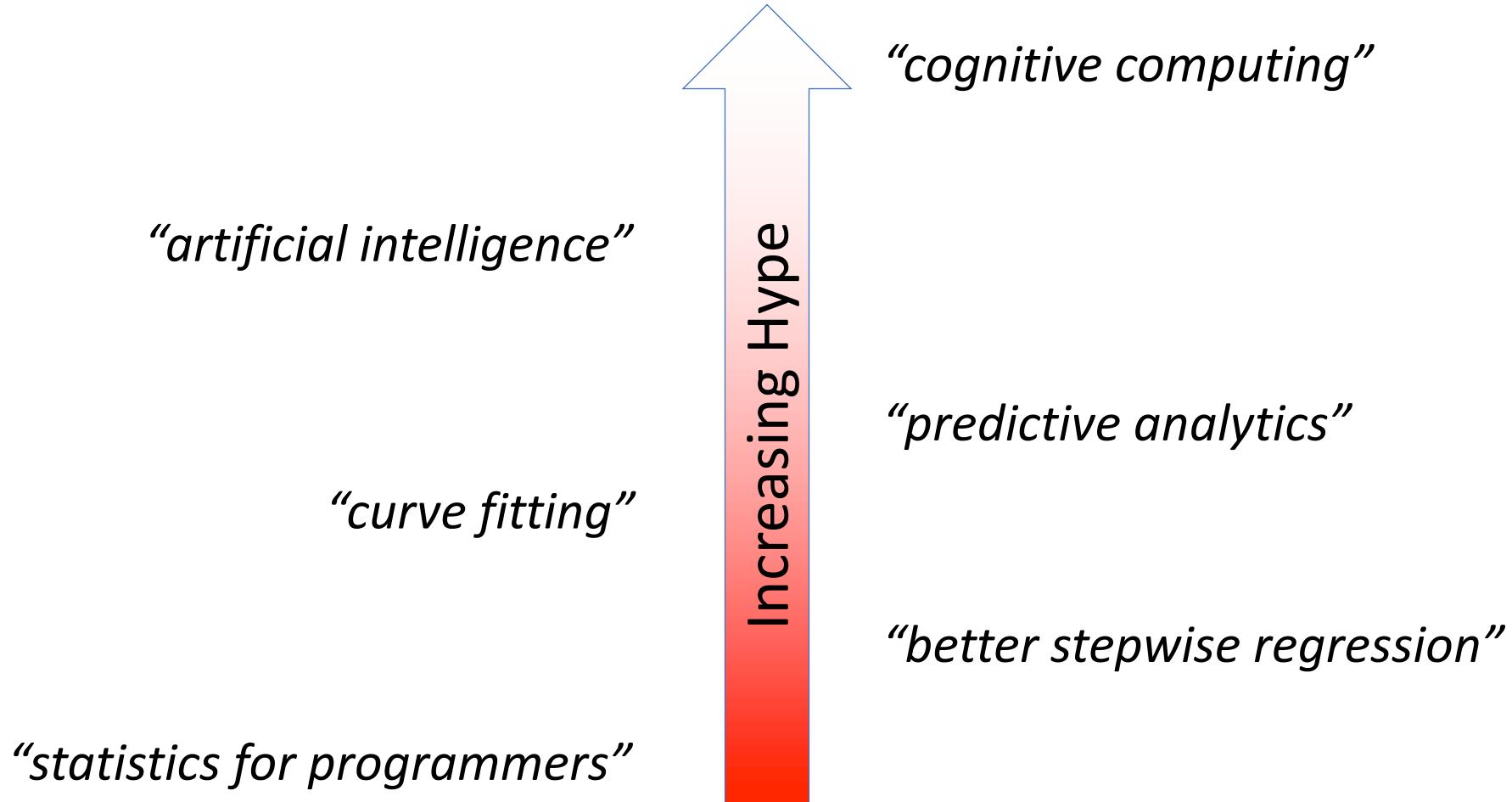


Principles and Culture of Data Science

MMCi Block 1
Matthew Engelhard

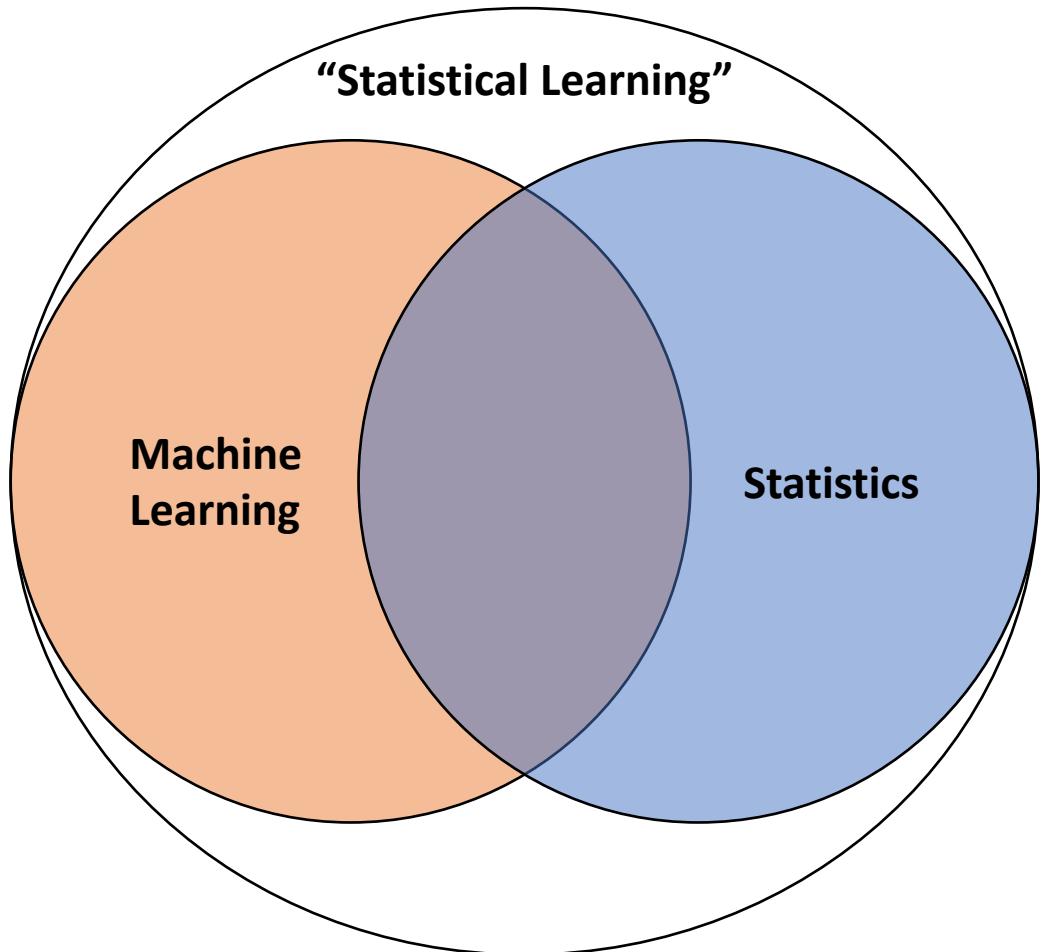
What are Machine Learning and Data Science?



Q: What is Machine Learning?

Q: What is Machine Learning?

“ML is an algorithmic field that blends ideas from statistics, computer science and many other disciplines to design algorithms that process data, make predictions and help make decisions” – M. I. Jordan



- Can't squeeze information that doesn't exist out of a dataset
- So, scientific and clinical judgement remain critical to success

“Big Data”: Distinction or Meme?

4 September 2008 | www.nature.com/nature | £10

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

nature

THE BITER BIT

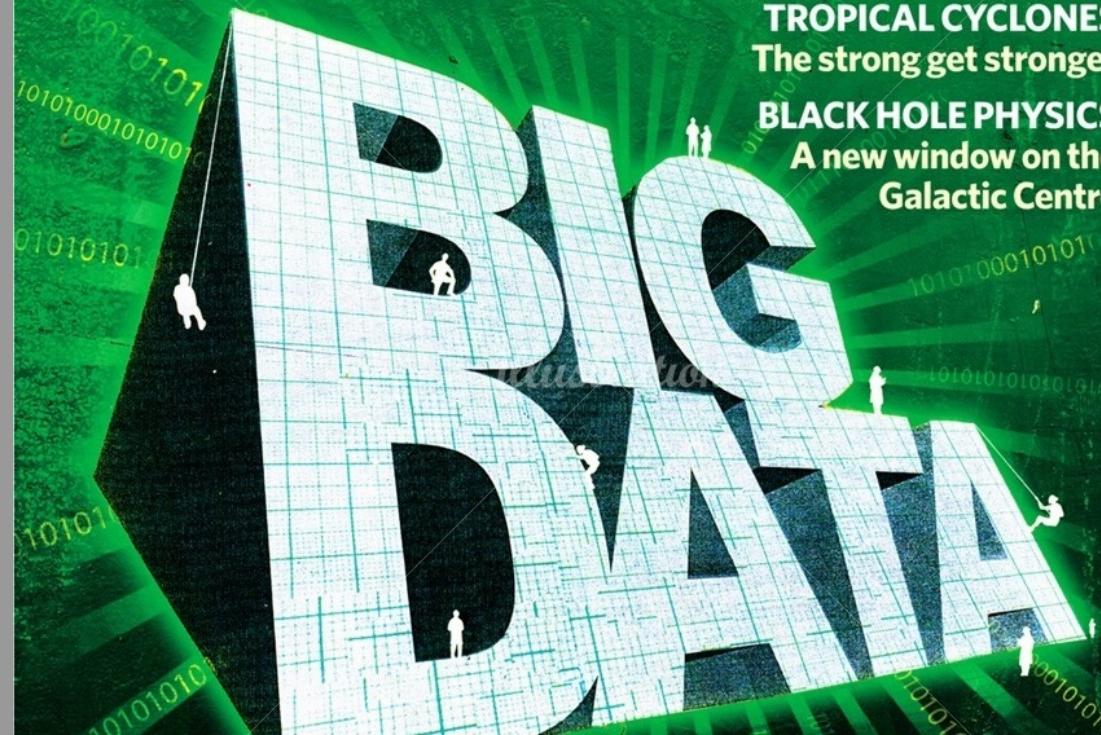
Viral infections for viruses

TROPICAL CYCLONES

The strong get stronger

BLACK HOLE PHYSICS

A new window on the
Galactic Centre



Epic



Long Data

- Large N
- A clinical study with 50k participants
- A speech database with 10M short recordings

Wide Data

- Large M
- A mHealth study with 6 months of wearables data in 20 participants
- 10 minutes of raw fMRI data in 8 participants

“Big Data”: Distinction or Meme?

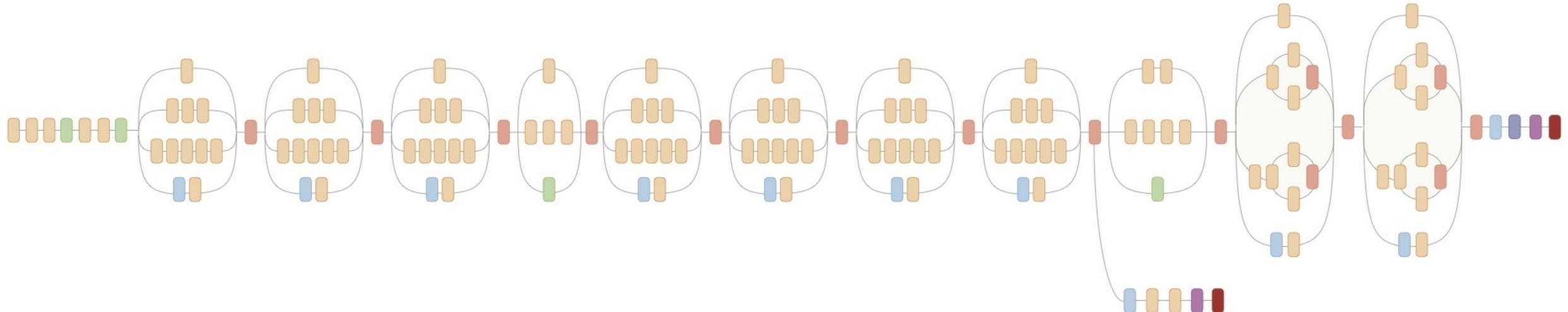
“We can immediately reject ‘big data’ as a criterion for meaningful distinction between statistics and data science”

-- David Donoho, *50 years of Data Science*

His points:

- Statisticians have been looking at big data, e.g. census data, for >200 years.
- Statisticians have long studied sampling and sufficient statistics, which allow them to work with big datasets

Neural networks thrive on big data



- Convolution
- AvgPool
- MaxPool
- Concat
- Dropout
- Fully connected
- Softmax

Inception v3 Convolutional Neural Network

Hardware and computational platforms for big data and deep learning

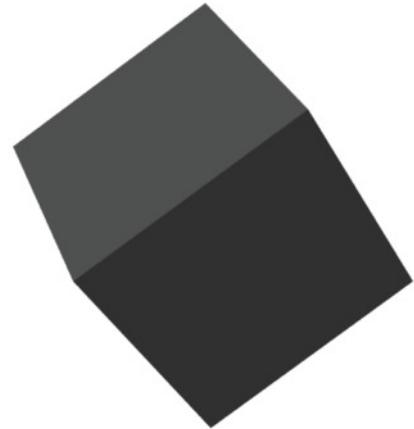


TensorFlow



Keras

Edward



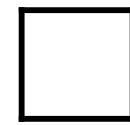
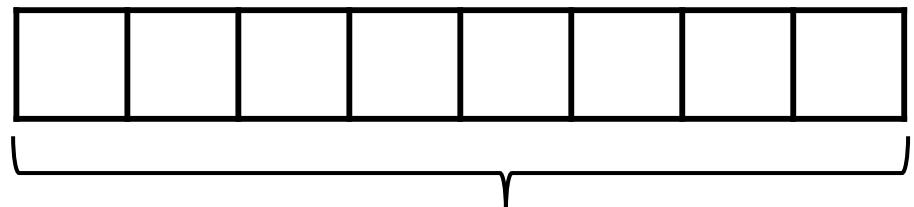
Predictive Modeling Culture

Previous Example: ICU Mortality Prediction

- Outcome:

$$y_i = \begin{cases} 1, & \text{patient } i \text{ dies} \\ 0, & \text{patient } i \text{ lives} \end{cases}$$

- Features: On admission, what is patient i 's {age, sex, temperature, blood pressure, ... }



y_i , did patient i die

What's our goal?

Answer 1:

Determine who's at highest risk of death so we can plan or manage resources accordingly

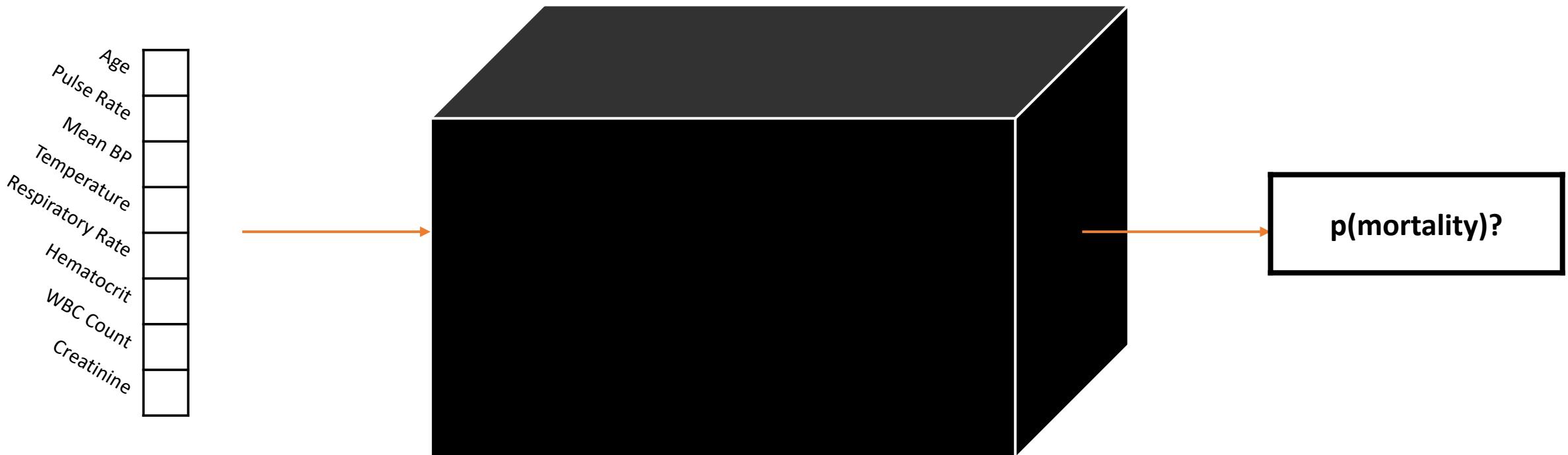
Prediction

Answer 2:

Identify predictors of mortality to further scientific understanding and design interventions

*Interpretation/
Understanding*

Machine Learning: A Black Box?



What's our goal?

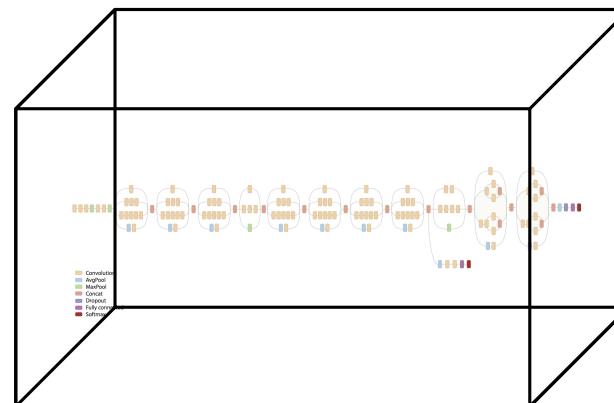
A: Prediction

- Black box may be OK as long as we're sure it works well.



A: Interpretation

- We need to look inside the box.



What's our goal?

A: Prediction

- Black box may be OK as long as we're sure it works well.

DS evaluation:

Determine whether the model performs well on new data

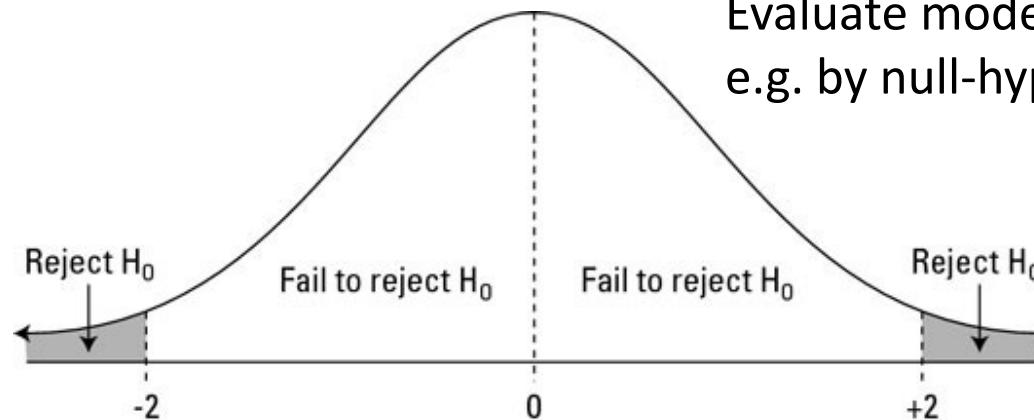


A: Interpretation

- We need to look inside the box.

Explanatory Approach:

Evaluate model parameters,
e.g. by null-hypothesis testing

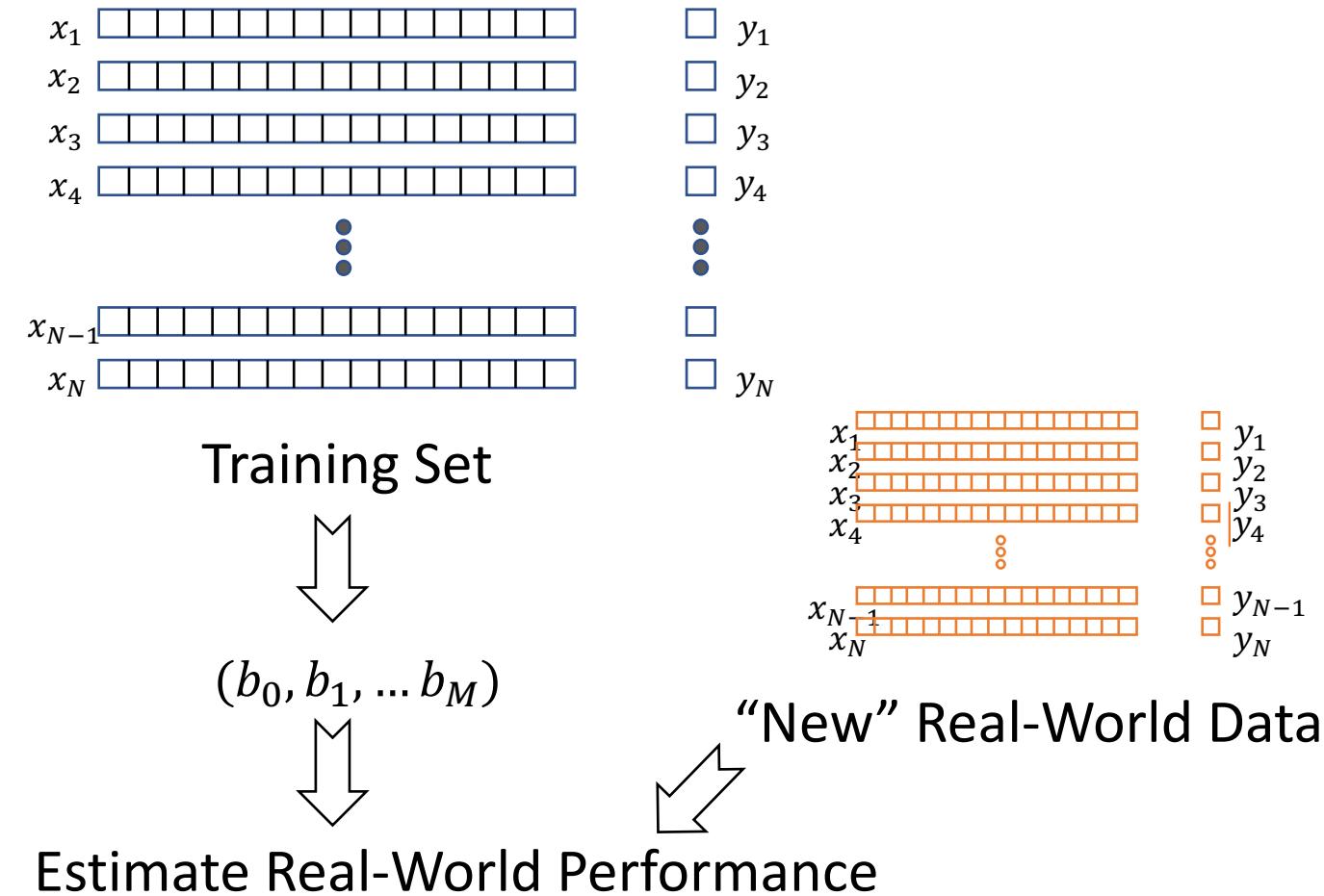


The Common Task Framework

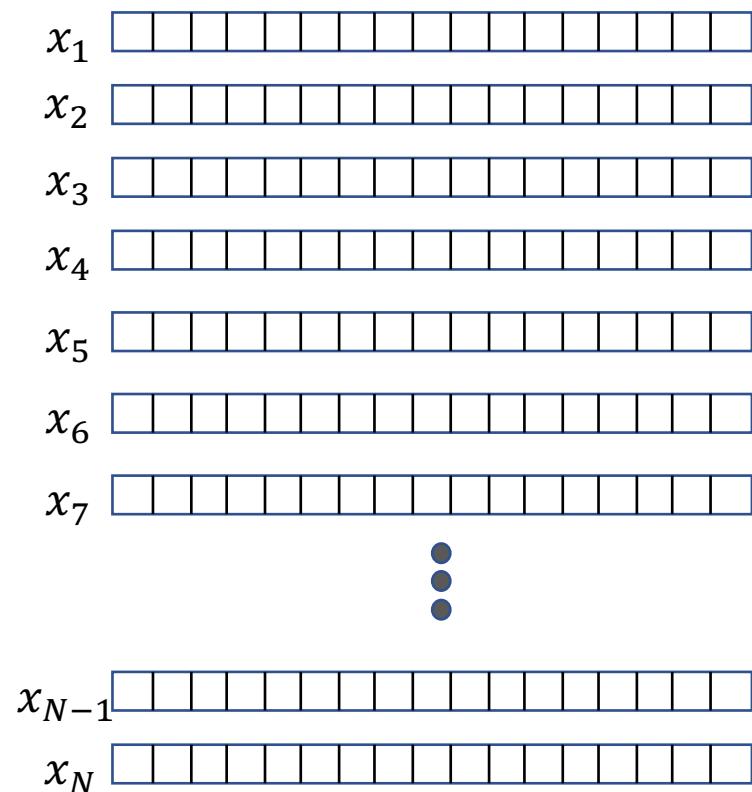
- Element 1: Standardized evaluation of model performance
- Element 2: Publicly-available training datasets
- Element 3: Competition with Impartial Scoring
- Element 4: Openness and code sharing

Standardized Evaluation Strategy

- We want to know how the network will perform *in the real world*
- So, we try it with new data
- This is costly; instead, can we use existing data to estimate performance?



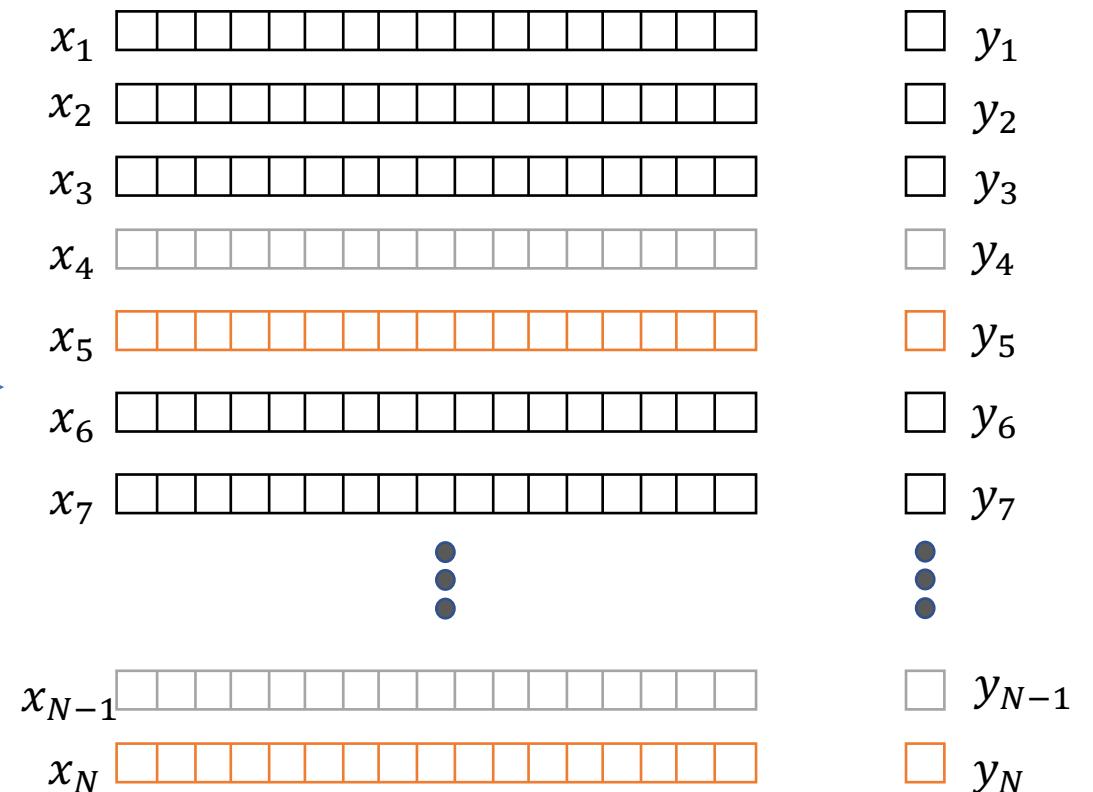
Split Data into Separate Groups



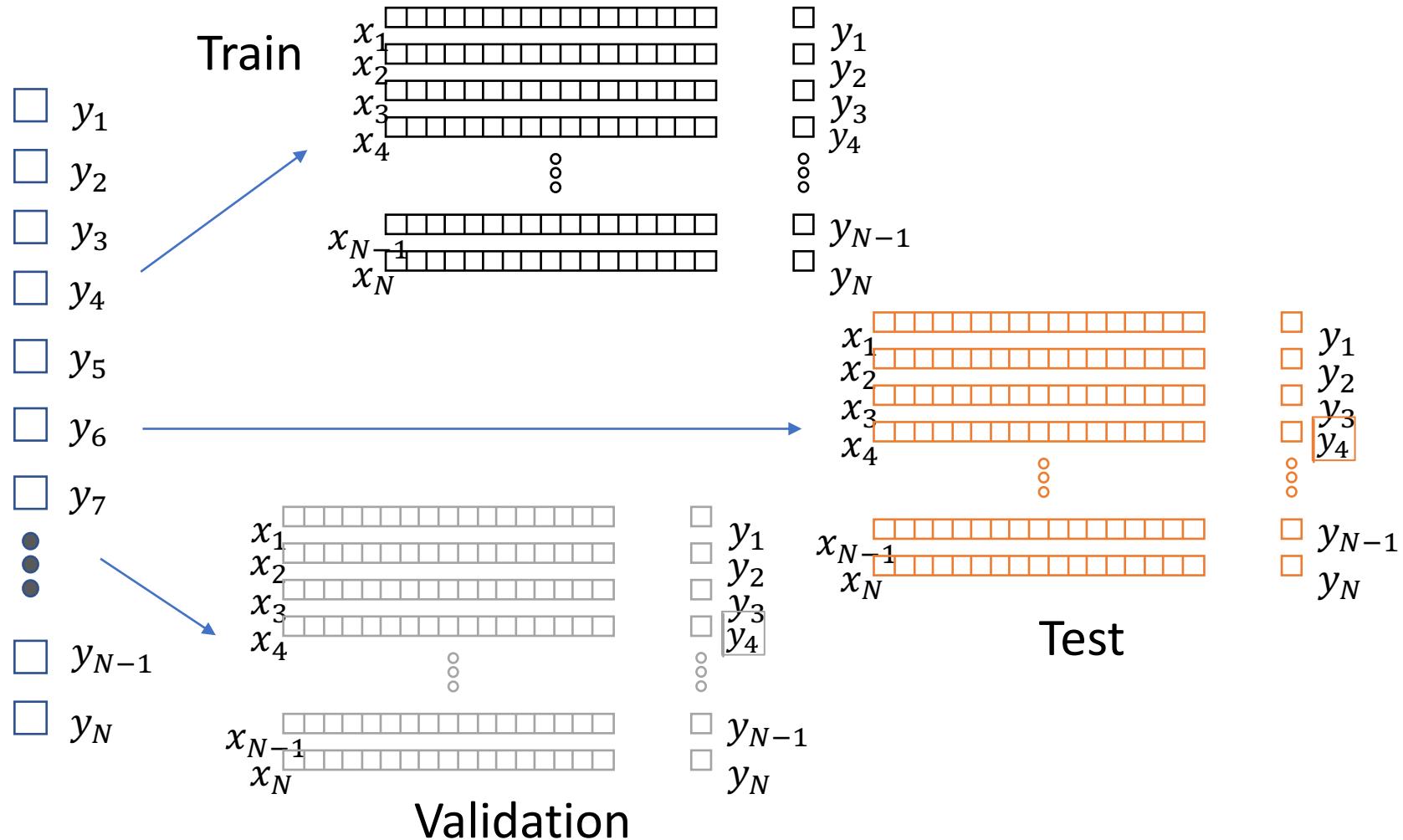
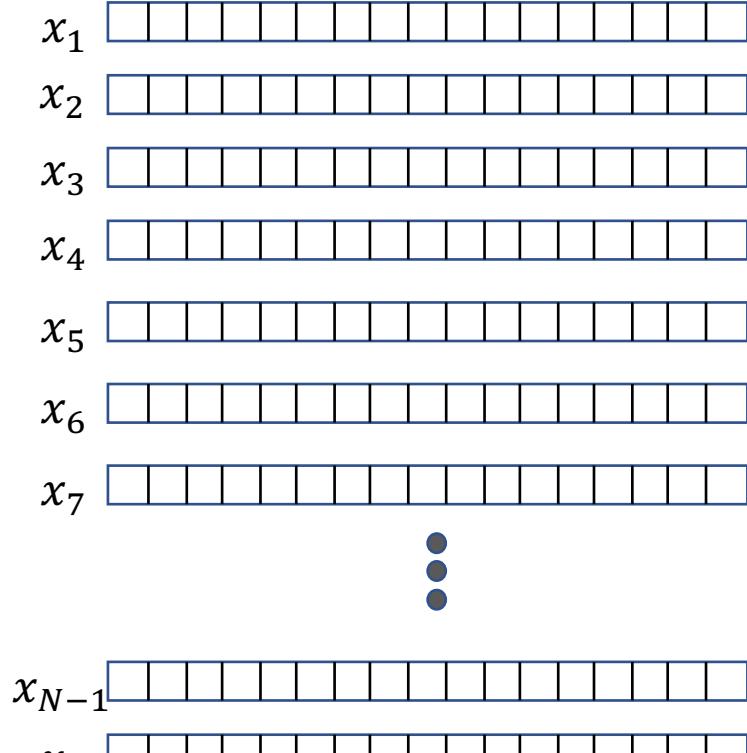
Train, Validation, **Test**

Random
Assignment

A blue arrow points from the 'All Available Data' section to the 'Train, Validation, Test' section. Below the arrow, the text 'Random Assignment' is written vertically.



Split Data into Separate Groups

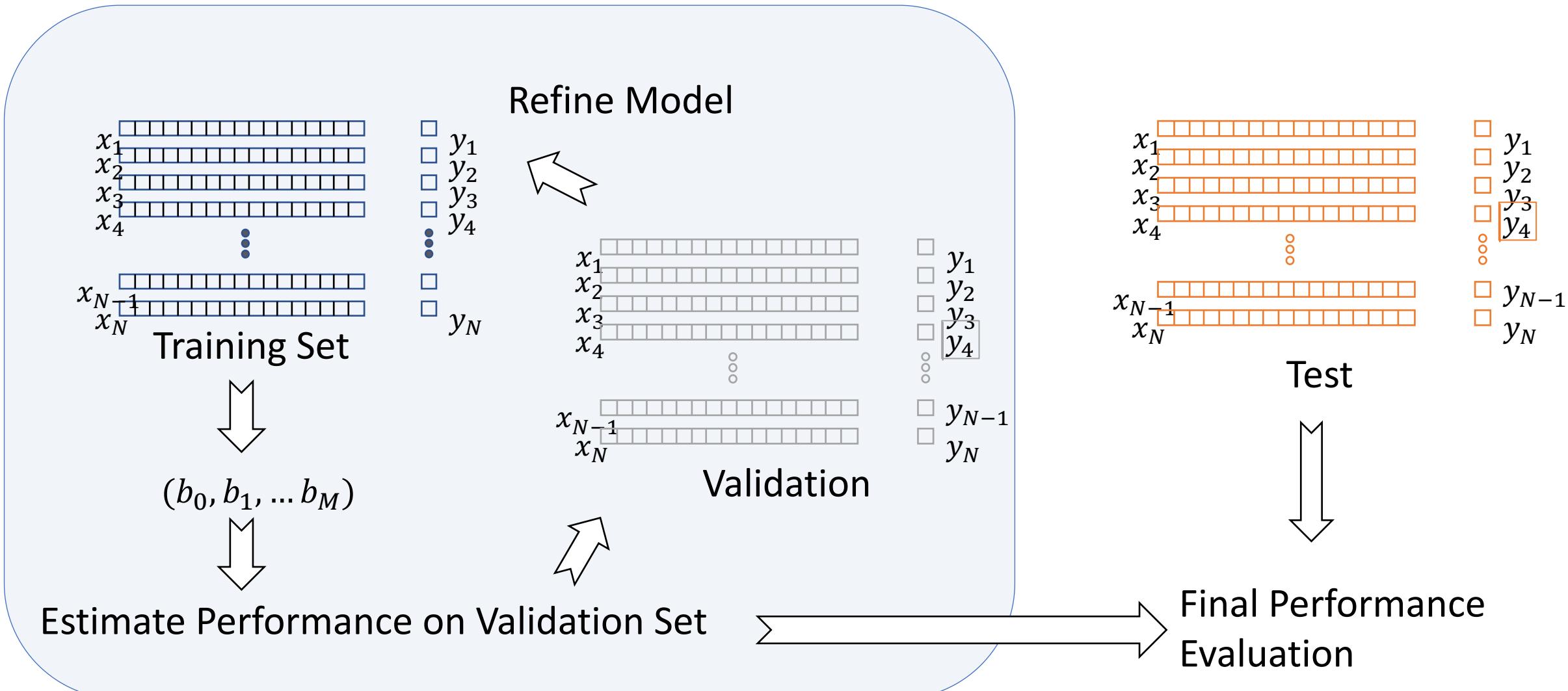


Test set

- This should be set aside prior to any analysis, and **will not be used to learn or fit any parameters**
- After learning the model, we evaluate its performance on the test set
 - This data was not included in the training/fitting, so it is analogous to running a new synthetic experiment
- Ideally, the test set will be used *once*.
 - Reusing the test set leads to bias; performance estimates will be optimistic
- So, how do we compare different models?

Validation (or *Tuning*) Set

- Want to be able to compare which approach is best
 - Problematic if we only want to use a test set once
 - Can create a second held-out dataset
- The validation data is not used for learning parameters, but can be used repeatedly to estimate performance of a model
- We can pick the model with the best performance on the validation set, and run a final evaluation on the test data



Pitfalls: Training, Validation, and Test

**This can be more nuanced than it might seem,
and is a common source of methodological errors!**

- **Pitfall 1:** Multiple data points from the same individual or source
- **Pitfall 2:** Utilizing data from the test or validation set prior to “modeling”
 - feature selection using the whole dataset
 - other information “leaks”
- **Pitfall 3:** Systematic differences between test, validation, and training sets

Pitfalls: Training, Validation, and Test

**This can be more nuanced than it might seem,
and is a common source of methodological errors!**

- **PITFALL 0:** Repeatedly evaluating on the test set!
- Pitfall 1: Multiple data points from the same individual or source
- Pitfall 2: Utilizing data from the test or validation set prior to “modeling”
 - feature selection using the whole dataset
 - other information “leaks”
- Pitfall 3: Systematic differences between test, validation, and training sets

- Element 1: Standardized evaluation of model performance
- Element 2: Publicly-available training datasets
- Element 3: Competition with Impartial Scoring
- Element 4: Openness and code sharing



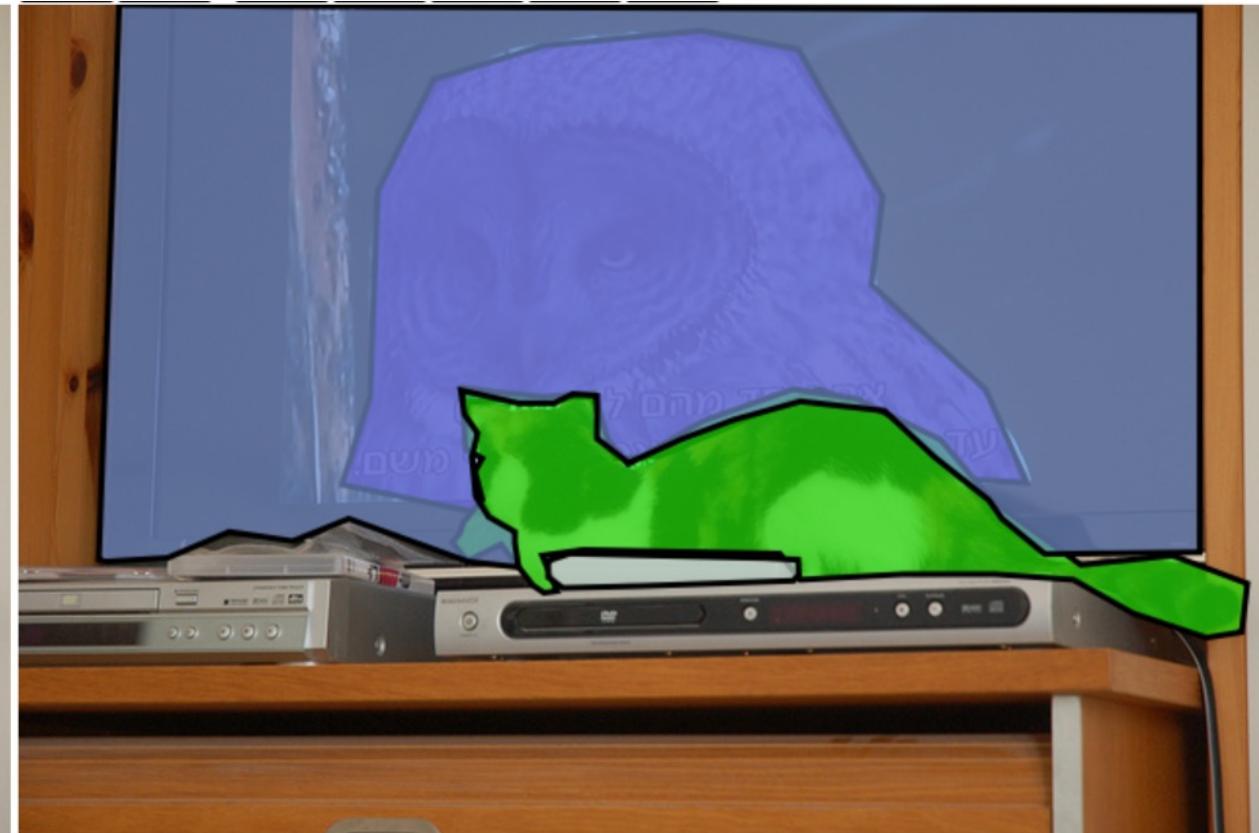
IMAGENET

The IMAGENET logo consists of the word "IMAGENET" in a large, light gray sans-serif font. The letter "A" is unique, featuring three colored squares (green, orange, and red) connected by thin lines, where each square is positioned above or below one of the vertical stems of the letter.



coco

Common Objects in Context



Stanford Question Answering Dataset (SQuAD 2.0)

Microorganisms or toxins that successfully enter an organism encounter the cells and mechanisms of the innate immune system. The innate response is usually triggered when microbes are identified by pattern recognition receptors, which recognize components that are conserved among broad groups of microorganisms, or when damaged, injured or stressed cells send out alarm signals, many of which (but not all) are recognized by the same receptors as those that recognize pathogens. Innate immune defenses are non-specific, meaning these systems respond to pathogens in a generic way. This system does not confer long-lasting immunity against a pathogen. The innate immune system is the dominant system of host defense in most organisms.

What part of the innate immune system identifies microbes and triggers immune response?

Ground Truth Answers: pattern recognition receptors receptors cells
Prediction: pattern recognition receptors

For most organisms, what is the dominant system of defense?

Ground Truth Answers: innate immune system innate immune system The innate immune
Prediction: The innate immune system

Pattern recognition receptors recognize components present in broad groups of what?

Ground Truth Answers: microorganisms microorganisms microorganisms
Prediction: microorganisms

The innate immune system responds in a generic way, meaning it is what?

Ground Truth Answers: non-specific non-specific non-specific
Prediction: non-specific

Amazon Reviews Dataset

```
{"reviewerID": "A2SUAM1J3GNN3B",
"asin": "0000013714",
"reviewerName": "J. McDonald",
"helpful": [2, 3],
"reviewText":
    "I bought this for my husband who plays the
    piano. He is having a wonderful time playing these
    old hymns. The music is at times hard to read
    because we think the book was published for singing
    from more than playing from. Great purchase
    though!",
"overall": 5.0,
"summary": "Heavenly Highway Hymns",
"unixReviewTime": 1252800000,
"reviewTime": "09 13, 2009"}
```



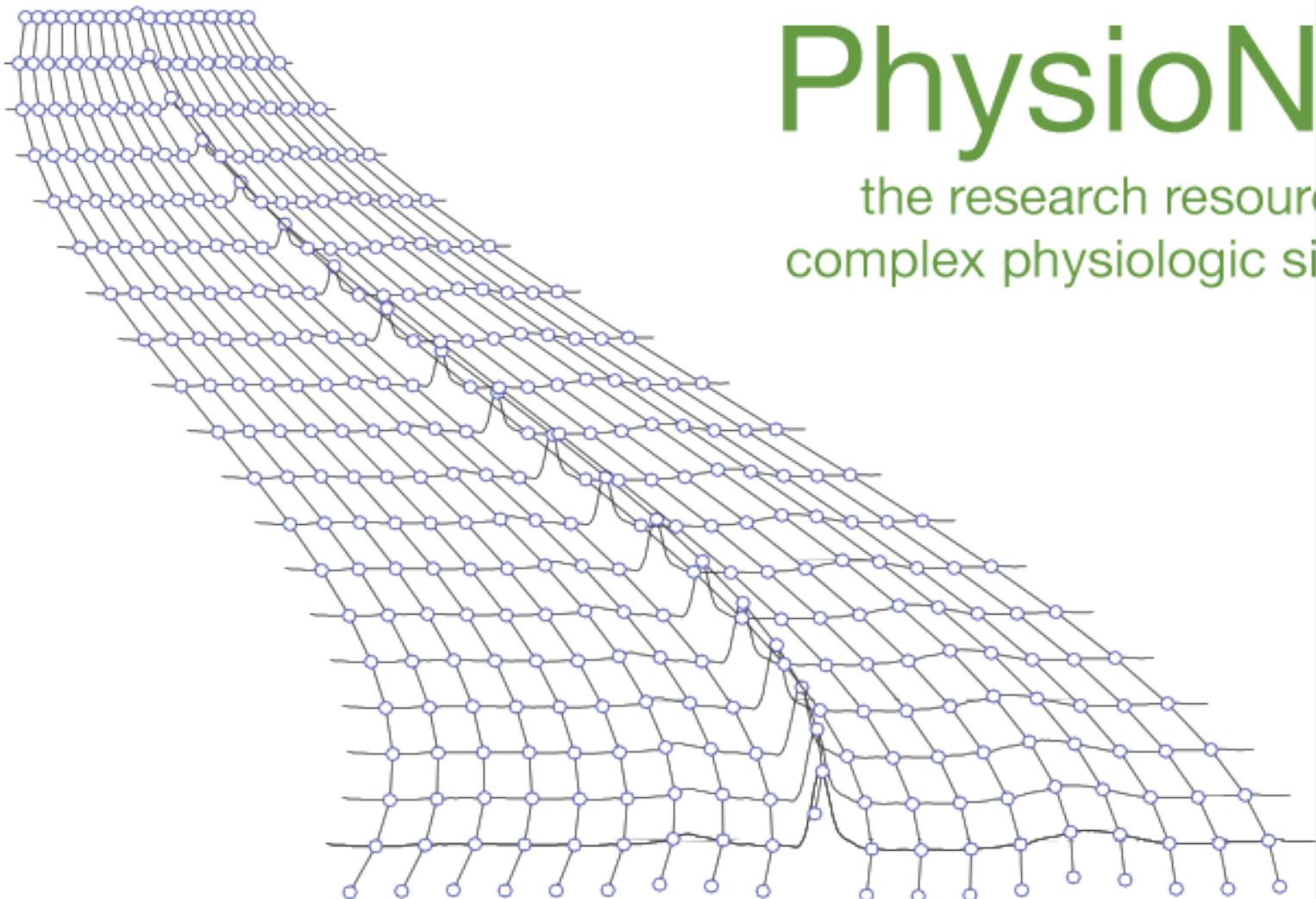
If you use MIMIC data or code in your work, please cite the following publication:

MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. *Scientific Data* (2016). DOI: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35).

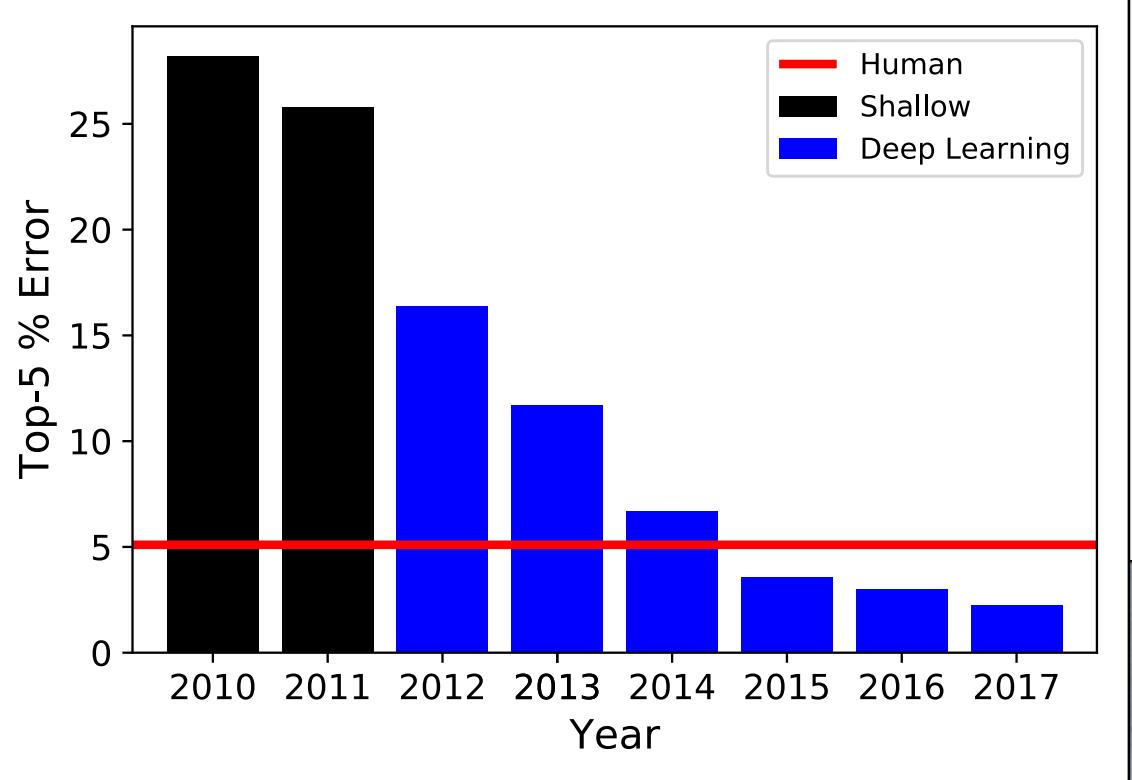
Available from: <http://www.nature.com/articles/sdata201635>

PhysioNet

the research resource for
complex physiologic signals



- Element 1: Standardized evaluation of model performance
- Element 2: Publicly-available training datasets
- Element 3: Competition with Impartial Scoring
- Element 4: Openness and code sharing



mite	container ship	motor scooter	leopard
mite	container ship	motor scooter	leopard
black widow	lifeboat	go-kart	jaguar
cockroach	amphibian	moped	cheetah
tick	fireboat	bumper car	snow leopard
starfish	drilling platform	golfcart	Egyptian cat

Leaderboard				
Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	-	-	-
Grand Prize - RMSE <= 0.8563				
1	PragmaticTheory	0.8584	9.78	2009-06-16 01:04:47
2	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
3	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
4	Dace	0.8604	9.56	2009-04-22 05:57:03
5	BigChaos	0.8613	9.47	2009-06-15 18:03:55
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
6	BellKor	0.8620	9.40	2009-06-17 13:41:48
7	Gravity	0.8634	9.25	2009-04-22 18:31:32
8	Opera Solutions	0.8640	9.19	2009-06-09 22:24:53
9	xlvector	0.8640	9.19	2009-06-17 12:47:27
10	BruceDengDaoCiYiYou	0.8641	9.18	2009-06-02 17:08:31
11	Ces	0.8642	9.17	2009-06-12 23:04:25
12	majia2	0.8642	9.17	2009-06-15 03:35:00
13	xiangliang	0.8642	9.17	2009-06-13 16:35:35
14	Feeds2	0.8647	9.11	2009-06-16 22:21:19
15	Just a guy in a garage	0.8650	9.08	2009-05-24 10:02:54
16	Team ESP	0.8653	9.05	2009-06-16 05:25:11
17	pengpengzhou	0.8654	9.04	2009-05-05 18:18:03
18	NewNetflixTeam	0.8657	9.01	2009-05-31 07:30:22
19	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
20	Vandelay Industries !	0.8658	9.00	2009-05-11 00:43:14

Display top leaders.

SHARE

SHARE



TWEET



COMMENT



EMAIL

NETFLIX NEVER USED ITS \$1 MILLION ALGORITHM DUE TO ENGINEERING COSTS

Rank	Team Name	Best Score	% Improvement	Last Submit Time
Grand Prize - RMSE <= 0.8543				
1	Pearl's Choice	0.8894	9.78	2009-09-19 01:04:47
2	Bellkor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
3	Grand Prize Team	0.8593	9.68	2009-05-12 08:20:24
4	Caser	0.8894	9.58	2009-04-22 06:57:03
5	BigChaos	0.8613	9.47	2009-05-15 18:32:55
Prizes Prize = 2000 - RMSE = 0.8543 - Missing Team: Bellkor in BigChaos				
6	Bellkor	0.8620	9.49	2009-05-17 13:41:48
7	Scat	0.8634	9.25	2009-04-22 18:31:32
8	Caser's solution	0.8640	9.18	2009-05-09 22:24:53
9	Asteric	0.8640	9.18	2009-05-17 12:47:27
10	BigDataCzarCff1024	0.8641	9.18	2009-05-02 17:09:31
11	Casa	0.8642	9.17	2009-05-12 23:34:25
12	mag2	0.8642	9.17	2009-05-15 05:35:26
13	karlspeng	0.8642	9.17	2009-05-15 18:02:25
14	tsudic	0.8647	9.11	2009-05-19 22:21:18
15	Just a guy in a garage	0.8650	9.08	2009-05-24 18:02:54
16	Team ESP	0.8653	9.08	2009-05-16 05:25:11
17	ajayachandru	0.8654	9.04	2009-05-05 18:18:03
18	ReuthefotTeam	0.8657	9.01	2009-05-31 07:39:22
19	Dinner 89	0.8658	9.00	2009-03-11 08:41:54
20	Vanderbilt University	0.8658	9.00	2009-05-11 08:43:14

Netflix awarded a \$1 million prize to a developer team in 2009 for an algorithm that increased the accuracy of the company's recommendation engine by 10 percent. But it doesn't use the million-dollar code, and has no plans to implement it in the future, Netflix announced on its blog Friday. The post goes on to explain why: a combination of too much engineering effort for the results, and a shift from movie recommendations to the "next level" of personalization caused by the transition of the business

Deep Learning is Approaching Human Performance in Language Understanding Tasks

Microorganisms or toxins that successfully enter an organism encounter the cells and mechanisms of the innate immune system. The innate response is usually triggered when microbes are identified by pattern recognition receptors, which recognize components that are conserved in microorganisms, or when damaged, injured signals, many of which (but not all) are recognized by those that recognize pathogens. Innate immunity is meaning these systems respond to pathogen signals, but do not confer long-lasting immunity against them. Innate immunity is the dominant system of host defense.

What part of the innate immune system identifies microbes and triggers immune response?

Ground Truth Answers: pattern recognition receptors receptors cells

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) Google AI Language https://github.com/google-research/bert	86.673	89.147
2	BERT + N-Gram Masking + Synthetic Self-Training (single model) Google AI Language https://github.com/google-research/bert	85.150	87.715





kaggle

Welcome to Kaggle Competitions

Try your skills on real-world machine learning problems

NO CHEATING!

You submit your model,
and the organizers apply it
to the test set, which is
not publicly-available

Announcing the 2018 BHI & BSN Data Challenge

In collaboration with the IEEE Conference on Biomedical and Health Informatics (BHI) 2018 and the IEEE Conference on Body Sensor Networks (BSN), we are hosting a challenge to explore real clinical questions in critically ill patients using the MIMIC-III database. Participants in the challenge will be invited to present at the BHI & BSN Annual Conference in Las Vegas, USA (4-7 March 2018): <https://bhi-bsn.embs.org/2018/>

Will there be prizes?

Yes, prizes are sponsored by Google Cloud and Georgia Tech.

Google Cloud

Georgia Tech 

- Element 1: Standardized evaluation of model performance
- Element 2: Publicly-available training datasets
- Element 3: Competition with Impartial Scoring
- Element 4: Openness and code sharing

Tensorflow Object Detection API

Creating accurate machine learning models capable of localizing and identifying multiple objects in a single image is a core challenge in computer vision. The TensorFlow Object Detection API is an open source framework built on top of TensorFlow that makes it easy to construct, train and deploy object detection models. At Google we've open-sourced our codebase to be useful for our computer vision needs, and we hope that you will as well.



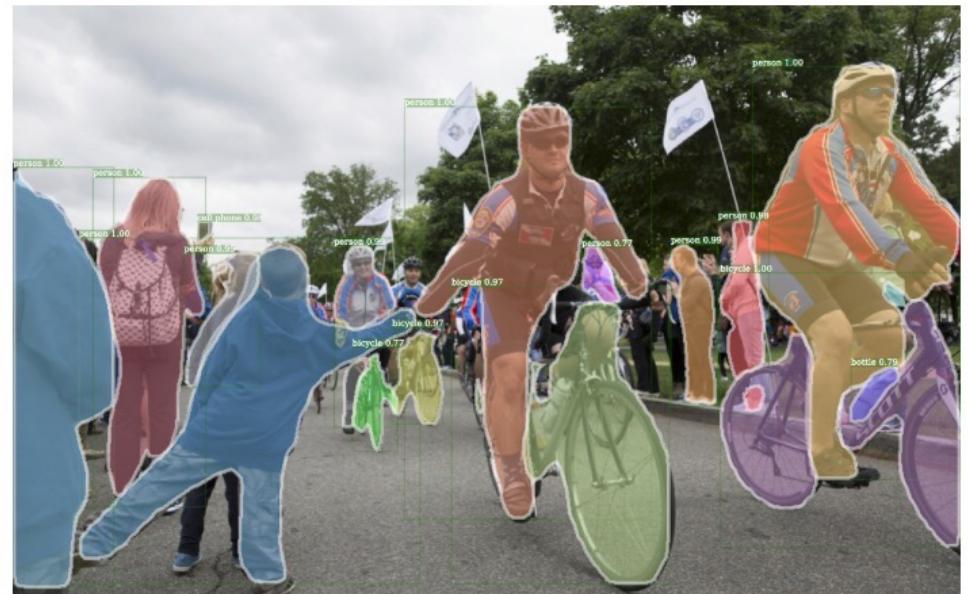
Contributions to the codebase are welcome and we would love to hear back from you if you find this API useful. If you use the Tensorflow Object Detection API for a research publication, please consider citing:

"Speed/accuracy trade-offs for modern convolutional object detectors."
Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S, Murphy K, CVPR 2017

Detectron

Detectron is Facebook AI Research's software system that implements state-of-the-art object detection algorithms, including Mask R-CNN. It is written in Python and powered by the Caffe2 deep learning framework.

At FAIR, Detectron has enabled numerous research projects, including: Feature Pyramid Networks for Object Detection, Mask R-CNN, Detecting and Recognizing Human-Object Interactions, Focal Loss for Dense Object Detection, Non-local Neural Networks, Learning to Segment Every Thing, Data Distillation: Towards Omni-Supervised Learning, DensePose: Dense Human Pose Estimation In The Wild, and Group Normalization.



Example Mask R-CNN output.

Less Stats Theory, More CS Theory and Culture

In 2001, William S. Cleveland (of Bell Labs) wrote that:

...[results in] data science should be judged by the extent to which they enable the analyst to learn from data... Tools that are used by the data analyst are of direct benefit. Theories that serve as a basis for developing tools are of indirect benefit.

He proposed 6 foci of activity, even suggesting allocations of effort.

- Multidisciplinary investigations (25%)
- Models and Methods for Data (20%)
- Computing with Data (15%)
- Pedagogy (15%)
- Tool Evaluation (5%)
- Theory (20%)

Several academic statistics departments that I know well could, at the time of Cleveland's publication, fit 100% of their activity into the 20% Cleveland allowed for Theory. Cleveland's paper was republished in 2014. I can't think of an academic department that devotes today 15% of its effort on pedagogy, or 15% on Computing with Data. I can think of several academic statistics departments that continue to fit essentially all their activity into the last category, Theory.

-- David Donoho, *50 years of Data Science*