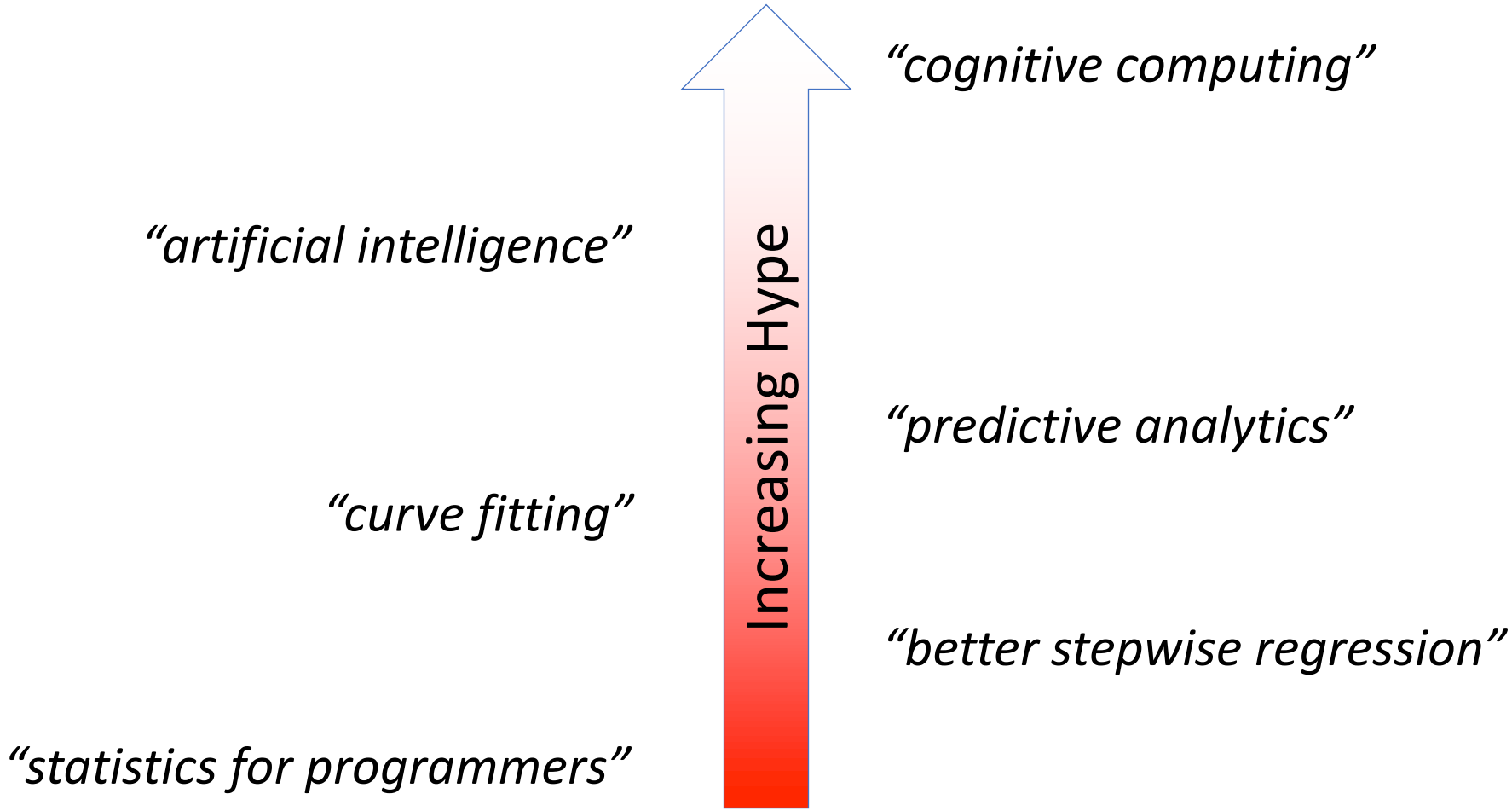
Principles and Culture of Data Science

CRTP Fall Term, Week 2

Matt Engelhard, Duke Biostatistics & Bioinformatics

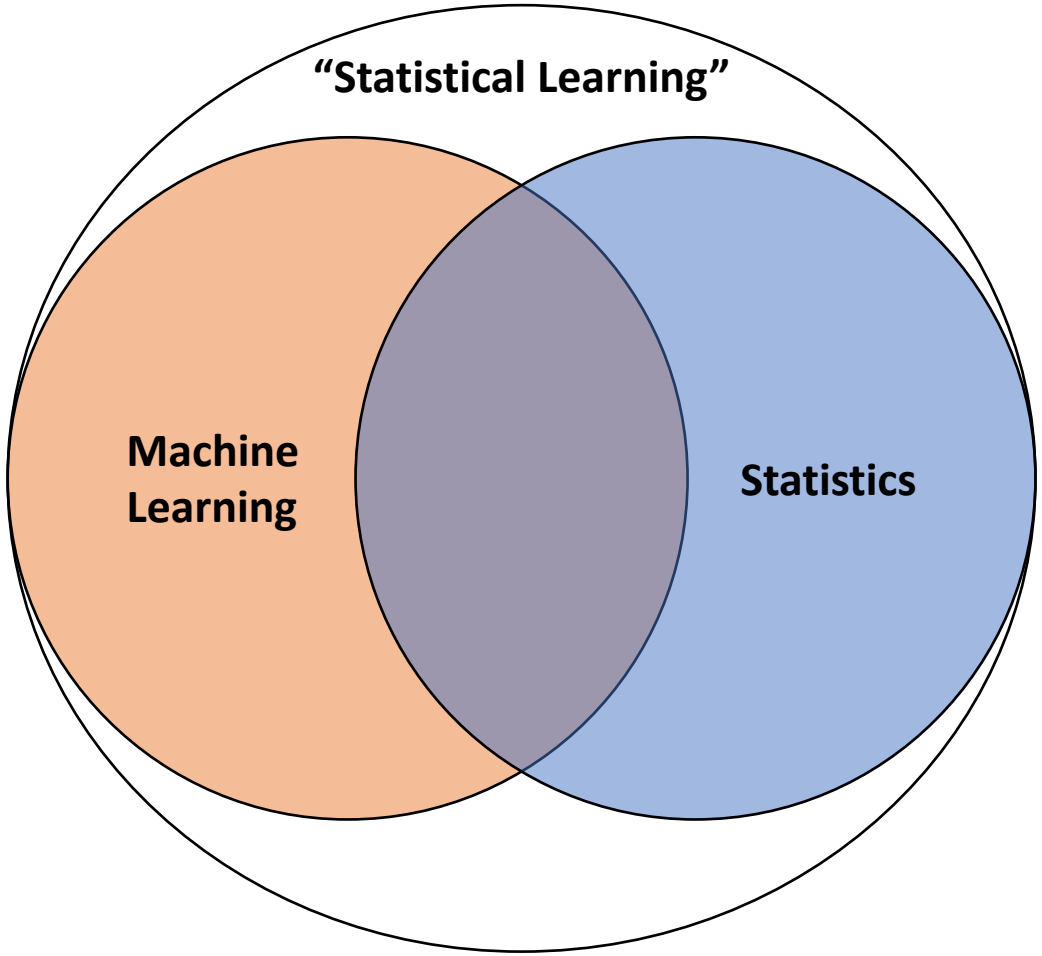
What are Machine Learning and Data Science?



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
BIG DATA

“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

Epic

Hardware and computational platforms for big data and deep learning

Edward

TensorFlow

Keras

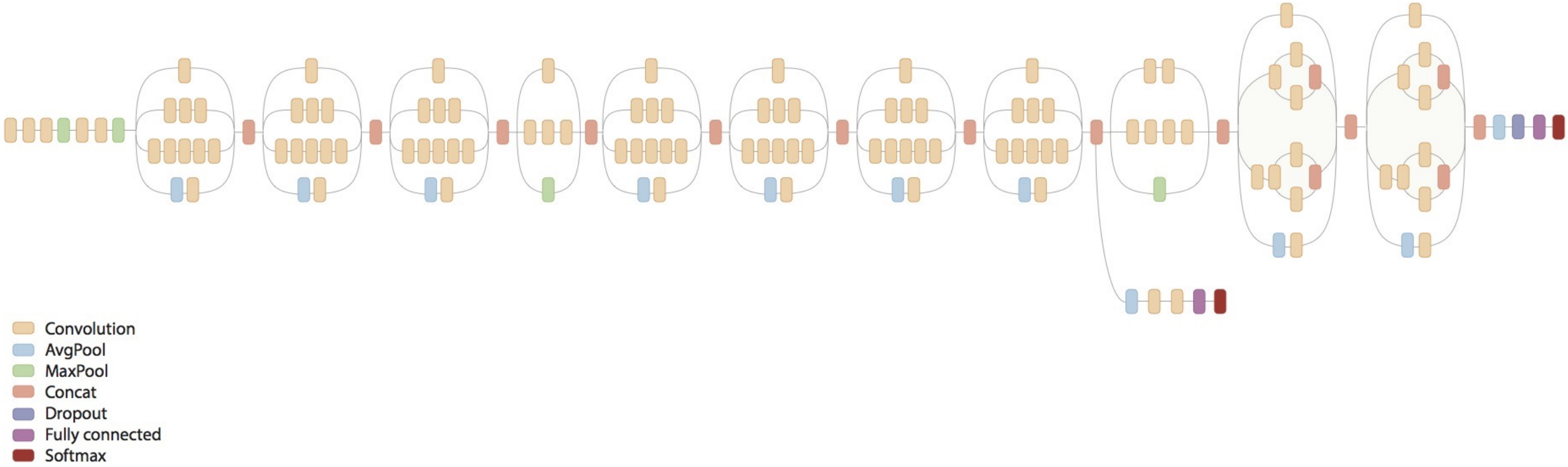
Long Data

- 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

Neural networks thrive on big data



Inception v3 Convolutional Neural Network

Predictive Modeling Culture

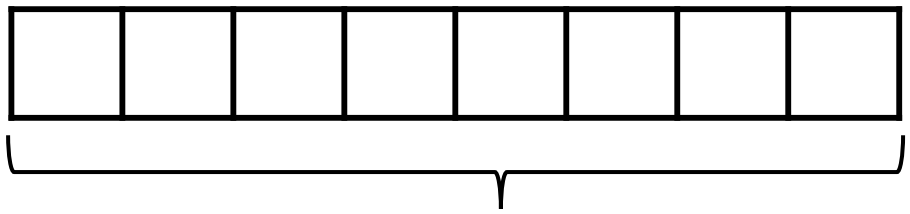
Previous Example: ICU Mortality Prediction

- Outcome:

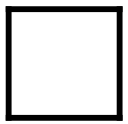
$$y_i = \begin{cases} 1, \text{patient } i \text{ dies} \\ 0, \text{patient } i \text{ lives} \end{cases}$$

- Features: On admission, what is patient i's

$$\{\text{age, sex, temperature, blood pressure, } \ldots\}$$



x_i, features for patient i



y_i, did patient i die

What's our goal?

<u>Answer 1</u>:

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

Prediction

<u>Answer 2</u>:

Identify predictors of mortality to further scientific understanding and design interventions

Interpretation/

Understanding

Machine Learning: A Black Box?

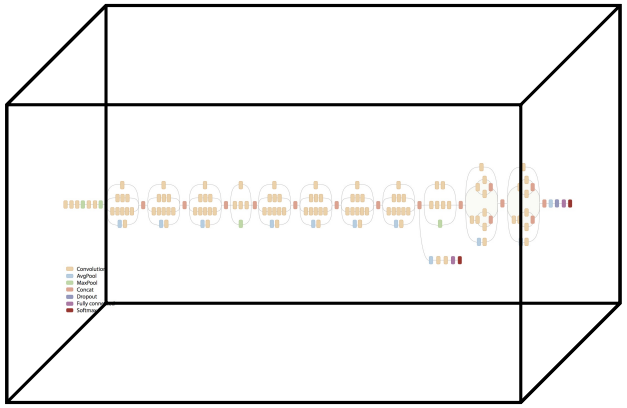
What’s our goal?

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?

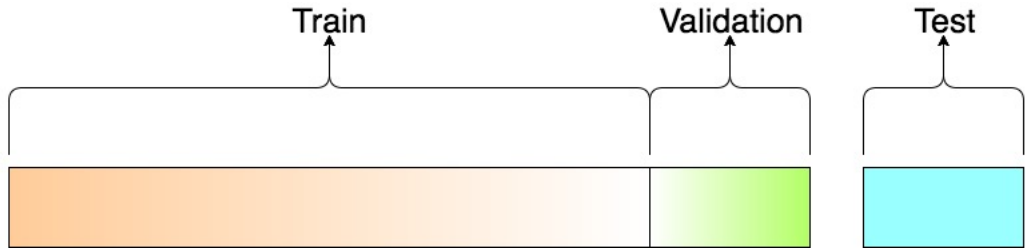
A: Prediction

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

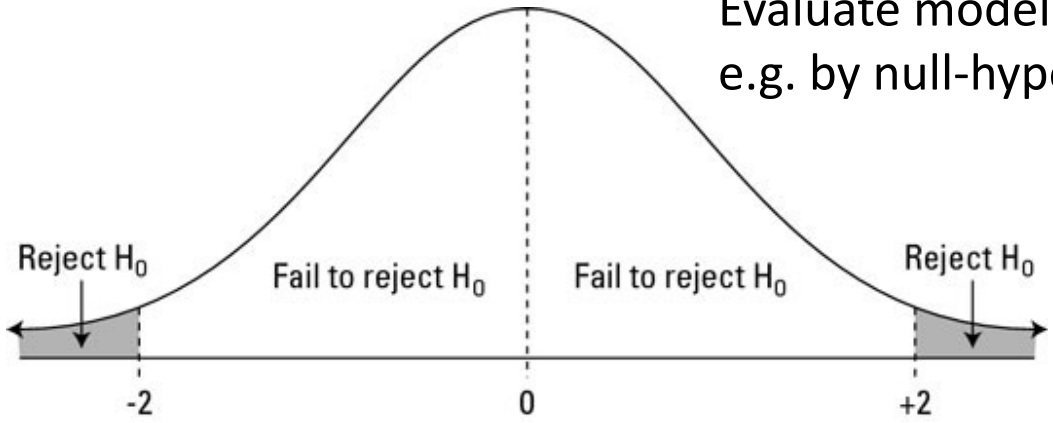
A: Interpretation

- We need to look inside the box.

DS evaluation:
Determine whether the model performs well on new data



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

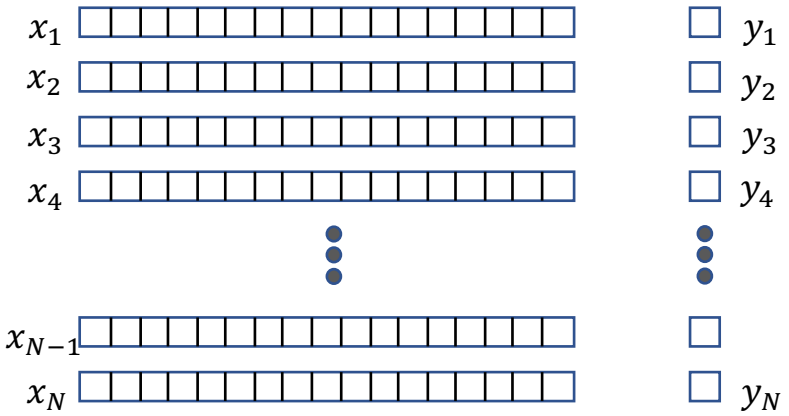


The Common Task Framework

- 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

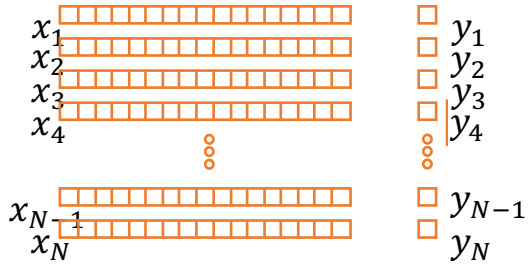
Standardized Evaluation Strategy

- 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?



Training Set

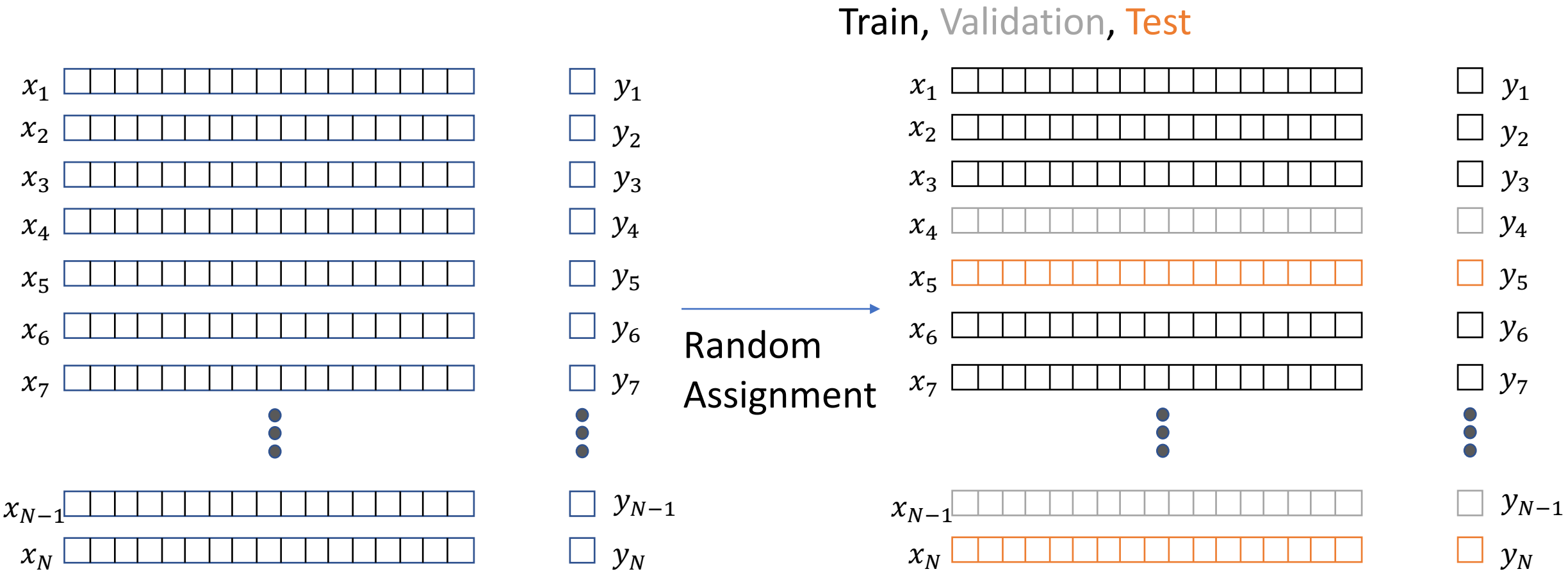
$(b_0, b_1, \ldots b_M)$



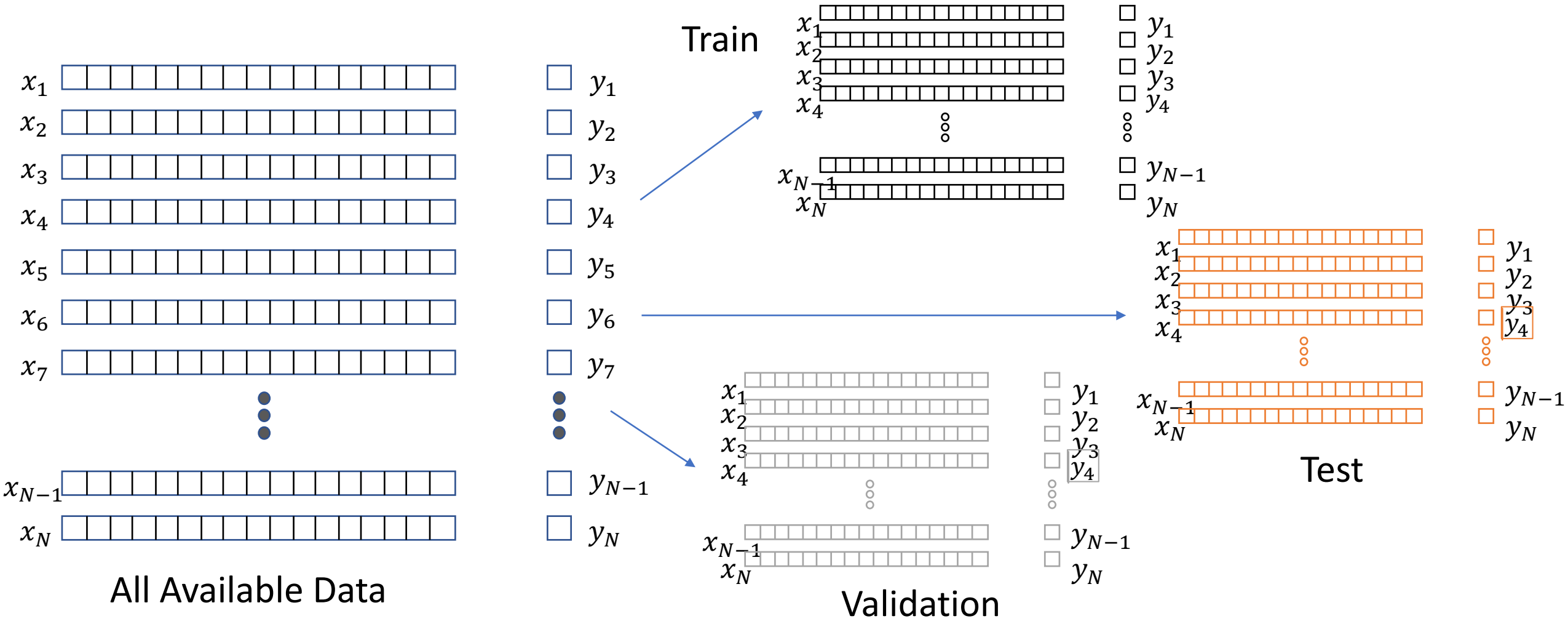
"New" Real-World Data

Estimate Real-World Performance

Split Data into Separate Groups



Split Data into Separate Groups



Test set

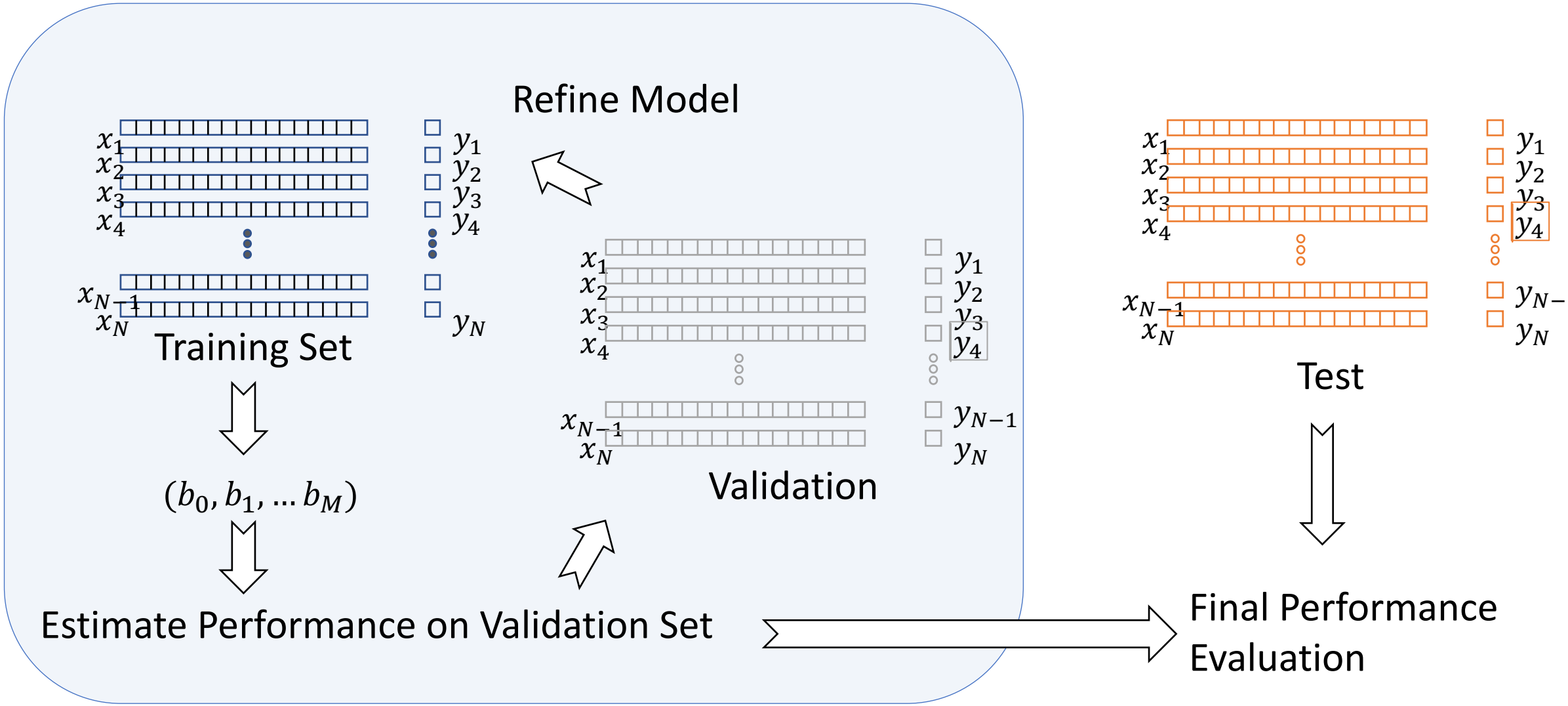
- This should be set aside prior to any analysis, and **will not be used to learn or fit any parameters**
- After learning the model, we evaluate its performance on the test set
  - This data was not included in the training/fitting, so it is analogous to running a new synthetic experiment
- Ideally, 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


Refine Model
x_1 x_2 x_3 x_4 x_{N-1} x_N
y_1 y_2 y_3 y_4 y_N
Training Set
$(b_0, b_1, \ldots b_M)$
x_1 x_2 x_3 x_4 x_{N-1} x_N
y_1 y_2 y_3 y_4 y_{N-1} y_N
Validation
Estimate Performance on Validation Set
x_1 x_2 x_3 x_4 x_{N-1} x_N
y_1 y_2 y_3 y_4 y_{N-1} y_N
Test
Final Performance Evaluation

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”

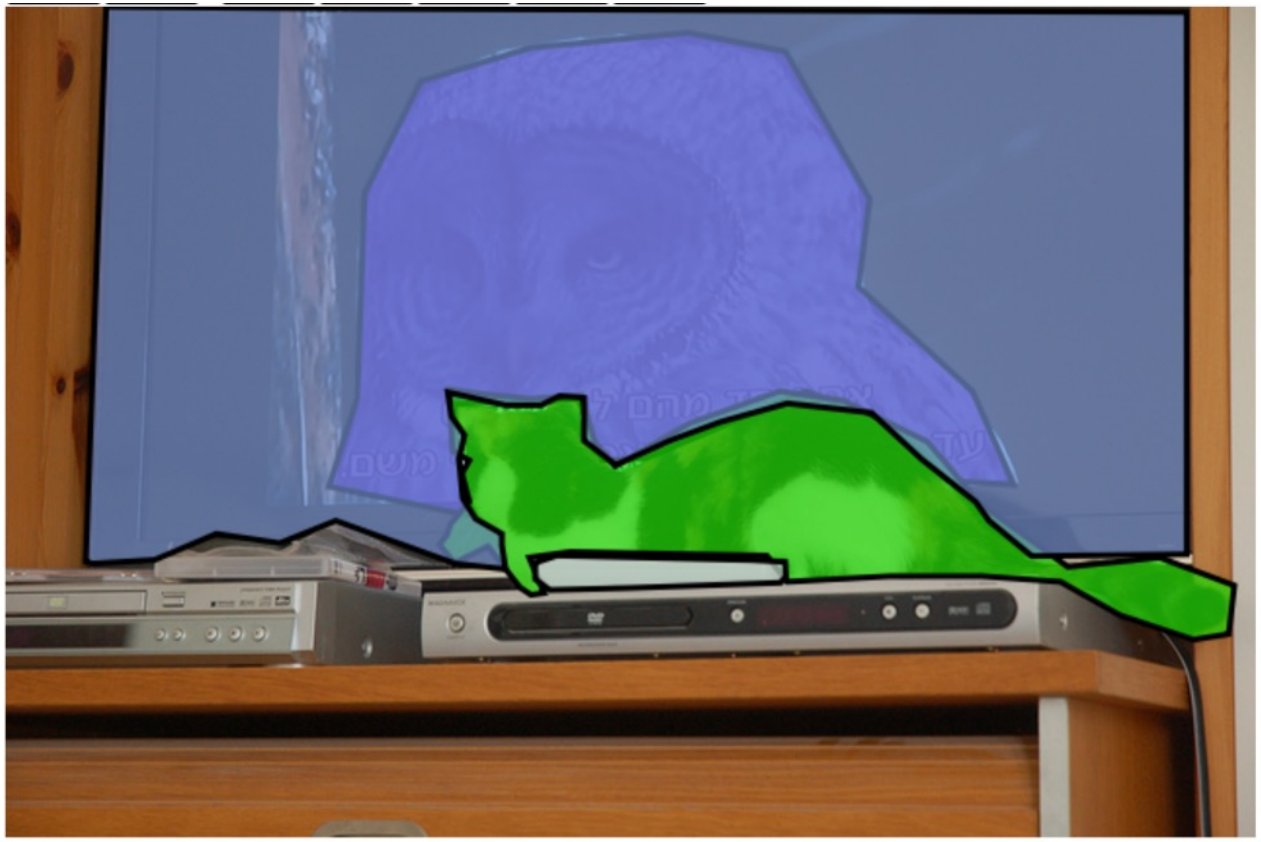
- Pitfall 3: Systematic differences between test, validation, and training sets

- 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

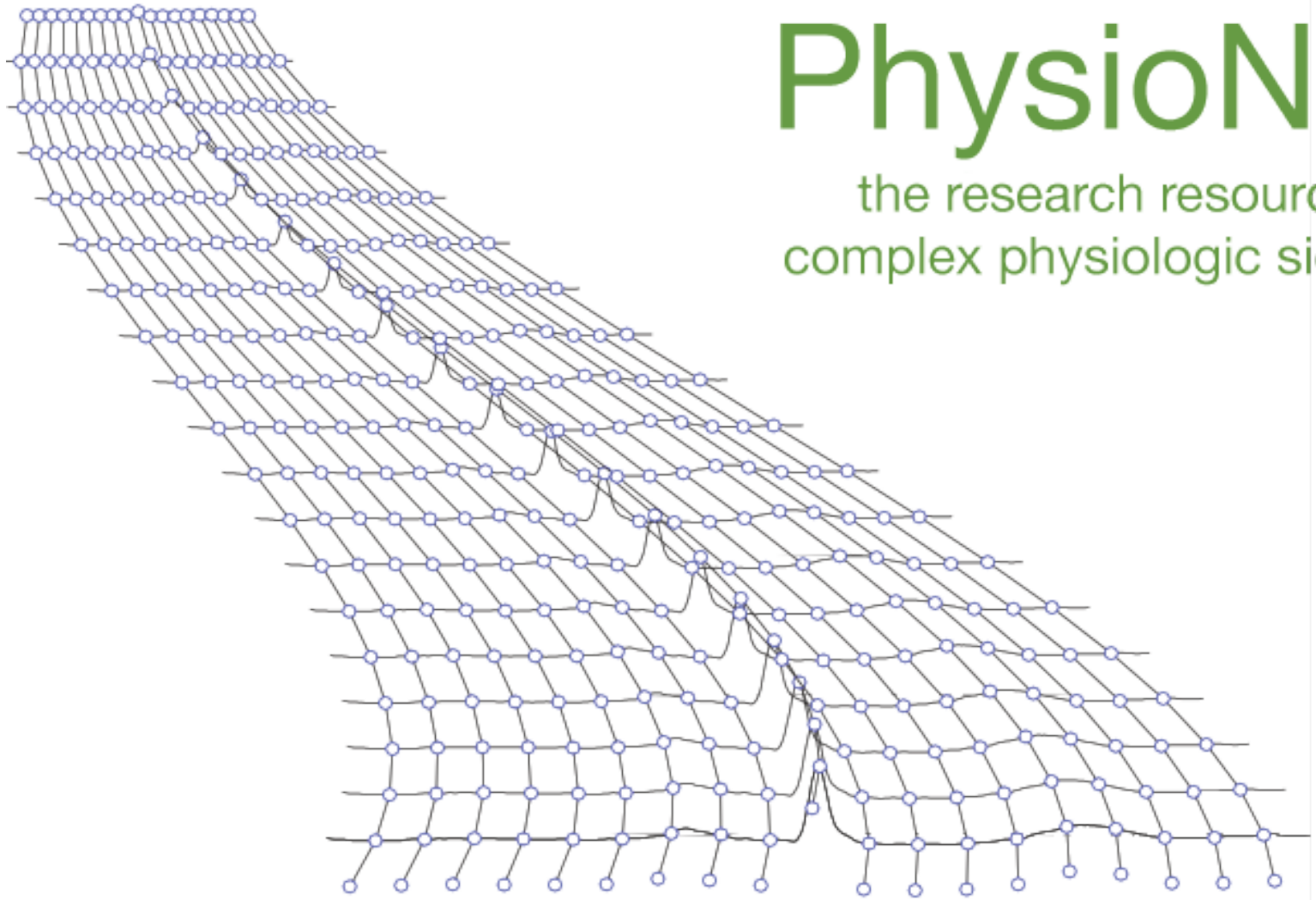
COCO

Common Objects in Context



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

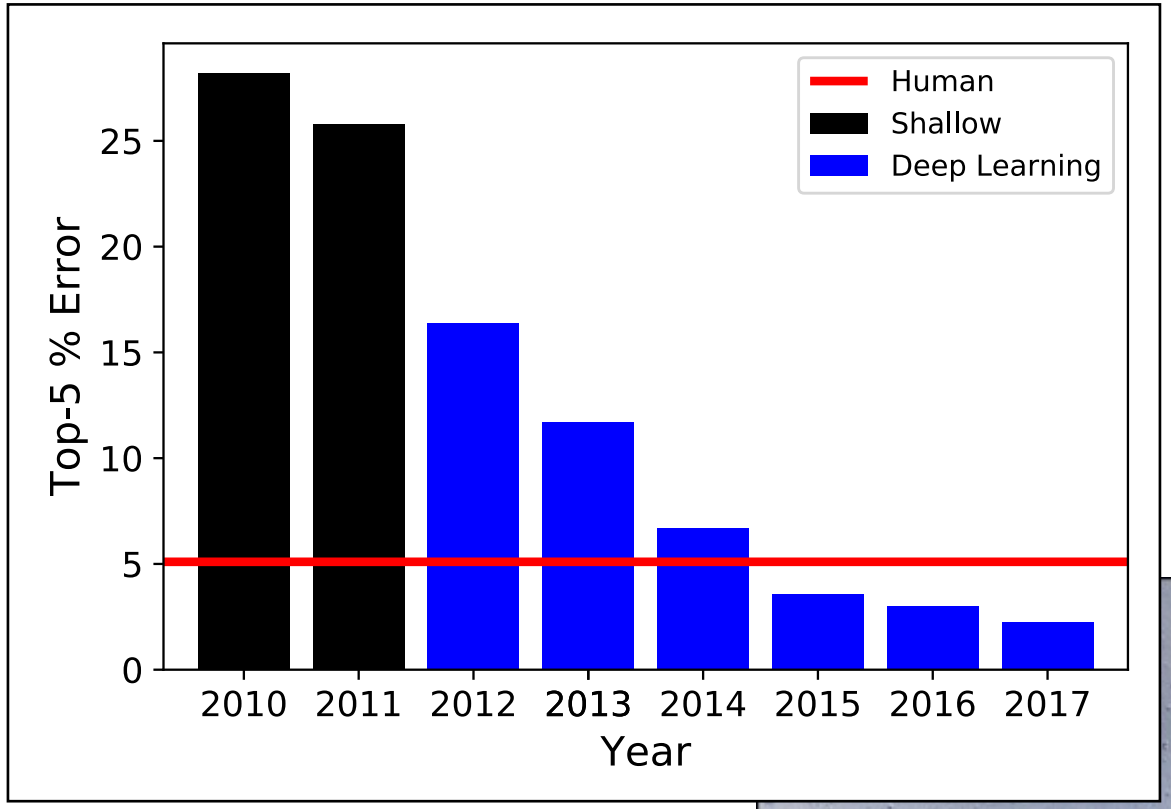
MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35.
Available from: http://www.nature.com/articles/sdata201635

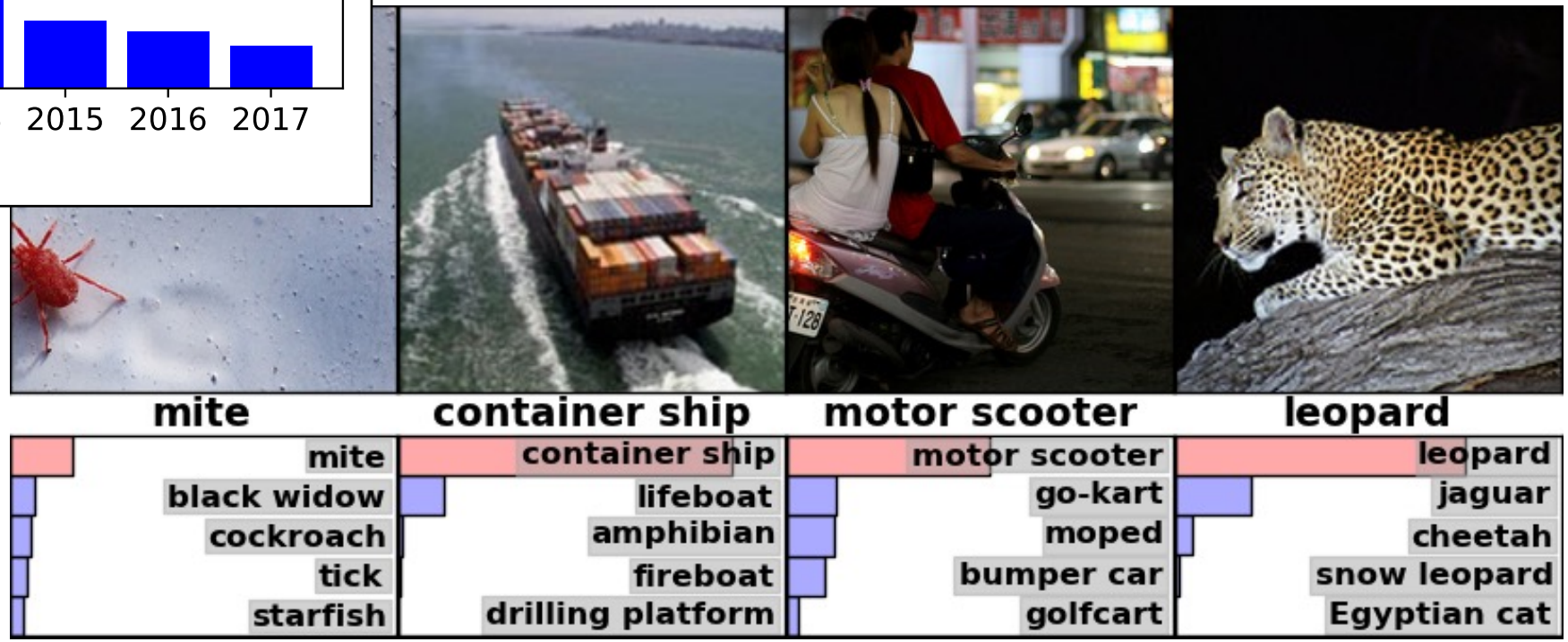


PhysioNet

the research resource for
complex physiologic signals

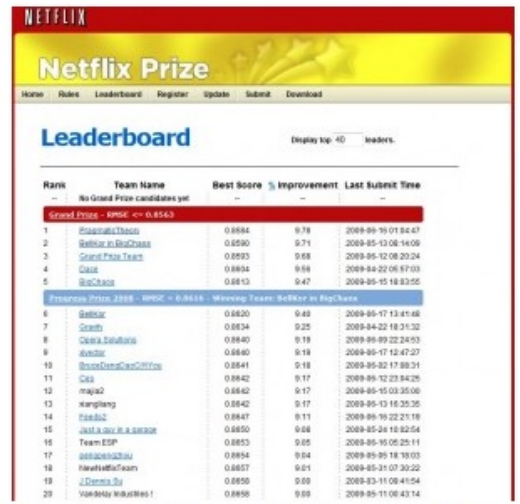
- Element 1: Standardized evaluation of model performance
- Element 2: Publicly-available training datasets
- Element 3: Competition with Impartial Scoring
- Element 4: Openness and code sharing


Human
Shallow
Deep Learning
Top-5 % Error
0
5
10
15
20
25
2010
2011
2012
2013
2014
2015
2016
2017
Year


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

CASEY JOHNSTON, ARS TECHNICA BUSINESS 04.16.12 08:20 AM

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



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


DeepMind
Workers: 14/16
AlphaStar
SC II HL


nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last — a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO

Duke TODAY

search

Topics

PUBLISHED JULY 9, 2018 IN RESEARCH, CAMPUS | UPDATED JULY 9, 2018

DUKE TEAMING UP WITH MICROSOFT IN DURHAM AND IN THE CLOUD

Microsoft Research facility to be added to Chesterfield Building

Stanford MEDICINE | News Center

FRONT PAGE | ALL NEWS | TOPICS | MULTIMEDIA

Email | Tweet

Stanford Medicine, Google team up to harness power of data science for health care

Stanford Medicine will use the power, security and scale of Google Cloud Platform to support precision health and more efficient patient care.

AUG 8
2016

Stanford Medicine and Google are working together to transform patient care and medical research through data science.

Lloyd Minor, dean of the School of Medicine, says the collaboration with Google marks a "milestone for the future of patient care and research."
Glenn Matsumura

The new collaboration combines Stanford Medicine's excellence in health-care research and clinical work with Google's expertise in cloud technology and data science. Stanford's forthcoming Clinical Genomics Service, which puts genomic sequencing into the hands of clinicians to help diagnose disease, will be built using Google Genomics, a service that applies the same technologies that power Google Search and Maps to securely store, process, explore and share genomic data sets.

Stanford Medicine includes the Stanford School of Medicine, Stanford Health Care and Stanford Children's Health. Together, Stanford Medicine and Google will build cloud-based applications for exploring massive health-care data sets, a move that could transform patient care and medical research.

"Stanford Medicine and Google are committing to major investments in preventing and curing diseases that afflict ordinary people worldwide. We're proud to be setting this milestone for the future of patient care and research," said Lloyd Minor, MD, dean of the School of Medicine.

The agreement — considered key to Stanford Health Care's development of the Clinical Genomics Service — makes Google Inc. a formal business associate of Stanford Medicine. As such, Google and Stanford will both comply with the Health Insurance Portability and Accountability Act, a federal law that regulates the privacy and security of medical information. HIPAA requires that Stanford Medicine patient data stored on Google Cloud Platform servers stay private. Patient information will be encrypted, both in transit and on servers, and kept on servers in the United States.

- 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

README.md

Tensorflow Object Detection API

Creating accurate machine learning models capable of localizing and identifying multiple objects in a sin
core challenge in computer vision. The TensorFlow Object Detection API is an open source framework bu
TensorFlow that makes it easy to construct, train and deploy object detection models. At Google we've c
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 AP
use the Tensorflow Object Detection API for a research publication, please consider citing:

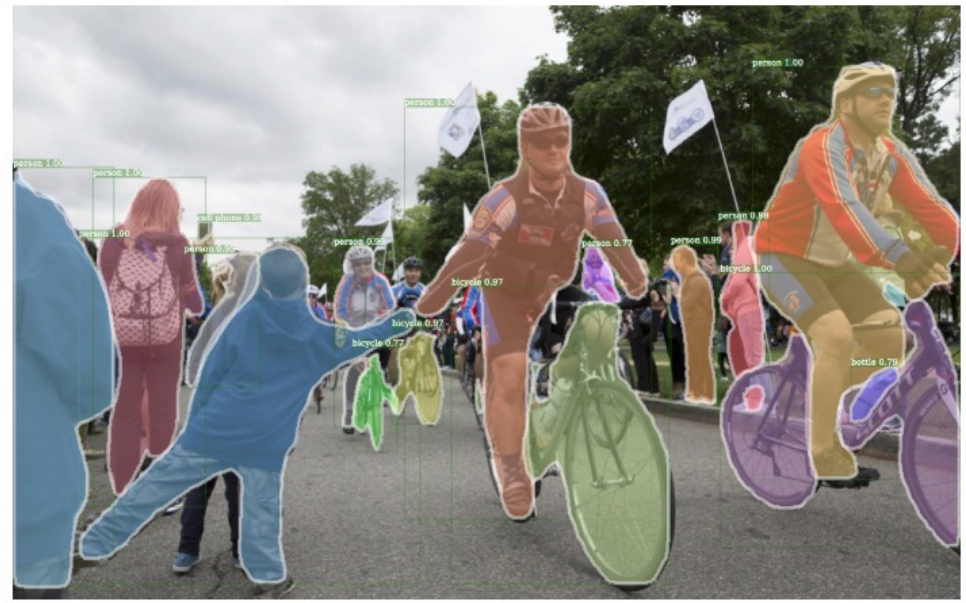
```
"Speed/accuracy trade-offs for modern convolutional object detectors."
Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z,
Song Y, Guadarrama S, Murphy K, CVPR 2017
```

README.md

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*