Carnegie Mellon University Qatar

Software Engineering for ML/AI

17-313, Foundations of Software Engineering, Fall 2023

(based on the slides of 17-445 by Christian Kästner)

S3D Software and Societal Systems Department

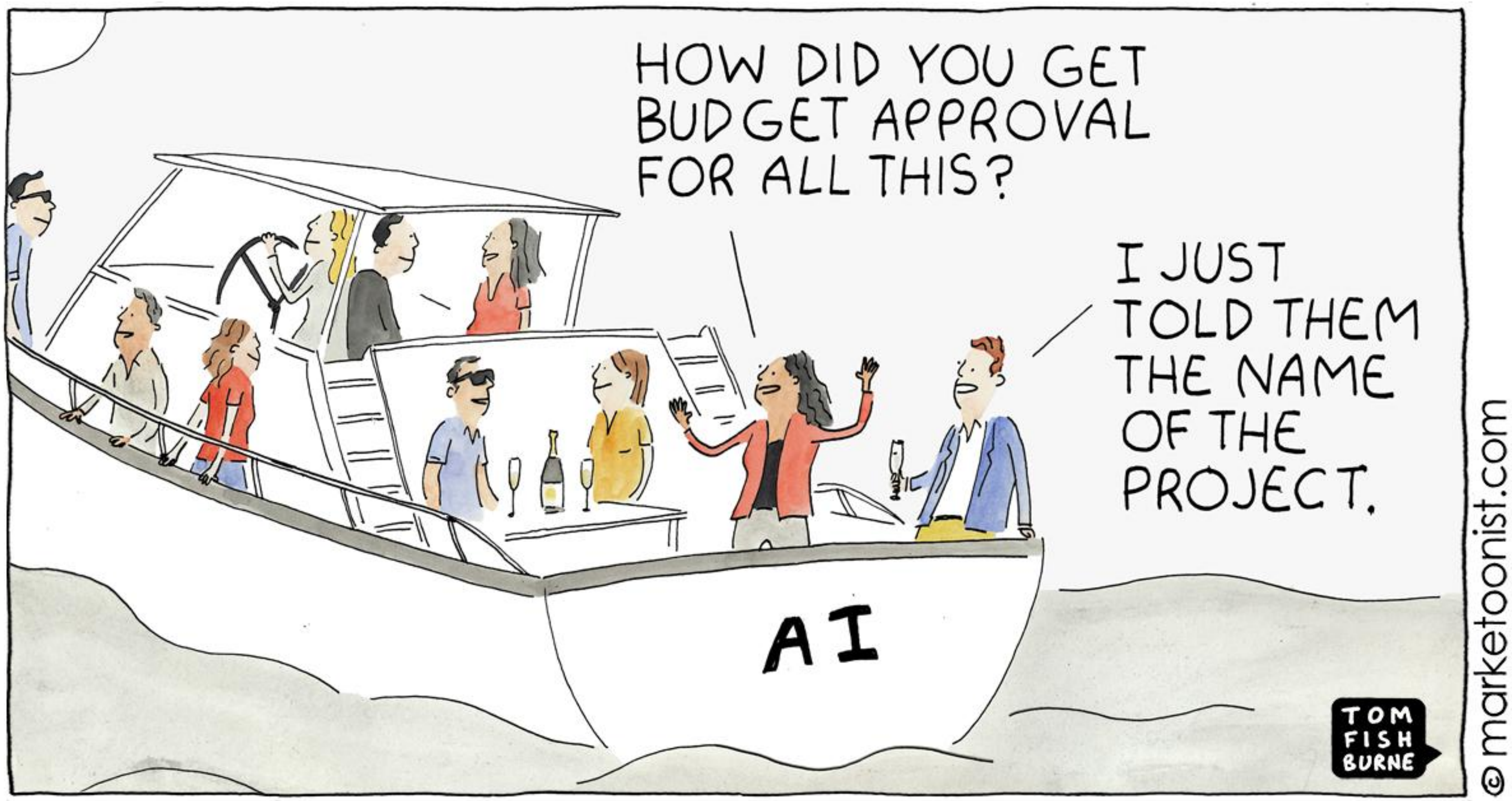
Learning goals

- Understand how machine learning (ML) components are parts of larger systems
- Explain the typical machine learning process
- Illustrate the challenges in engineering an ML-enabled system beyond accuracy
- Key differences from traditional software
- Illustrate the challenges in engineering an AI-enabled system beyond accuracy

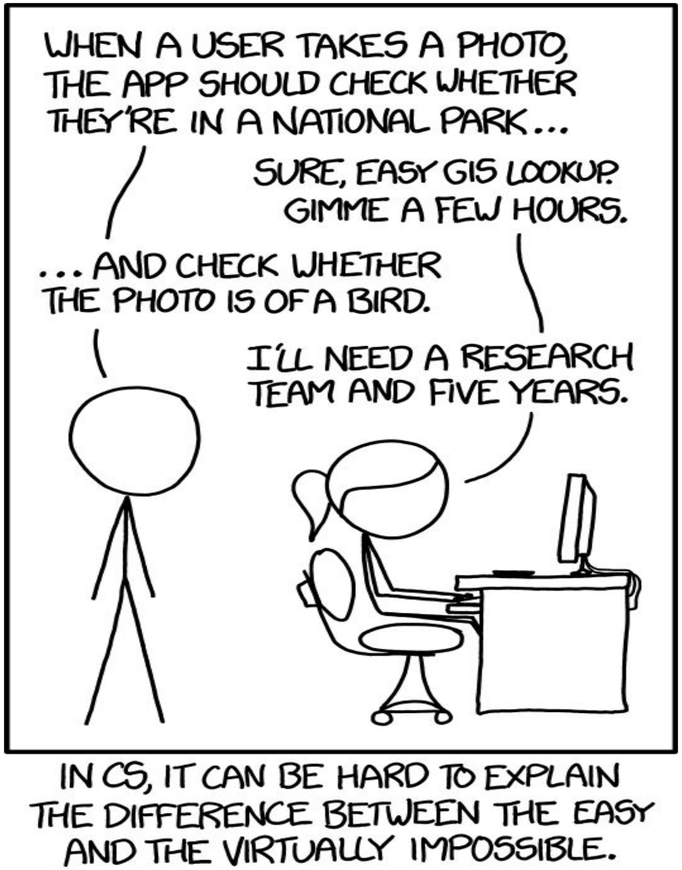
Software Engineering For AI/ML

Why?

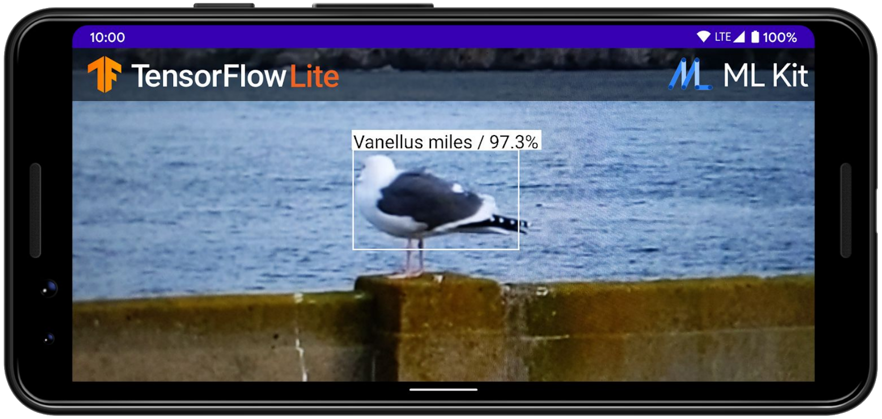
Navigating the AI Hype



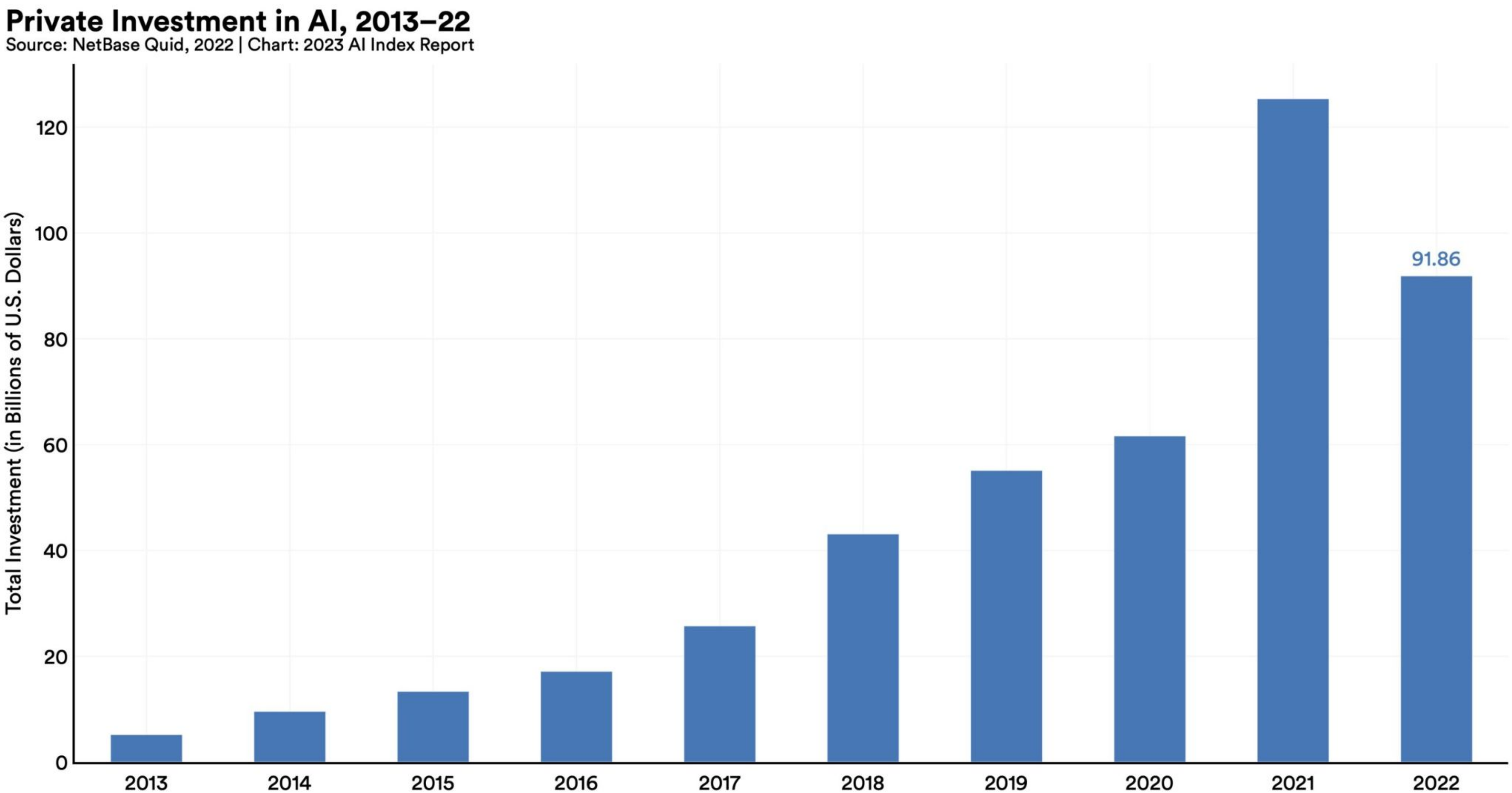
2014



... a few years later



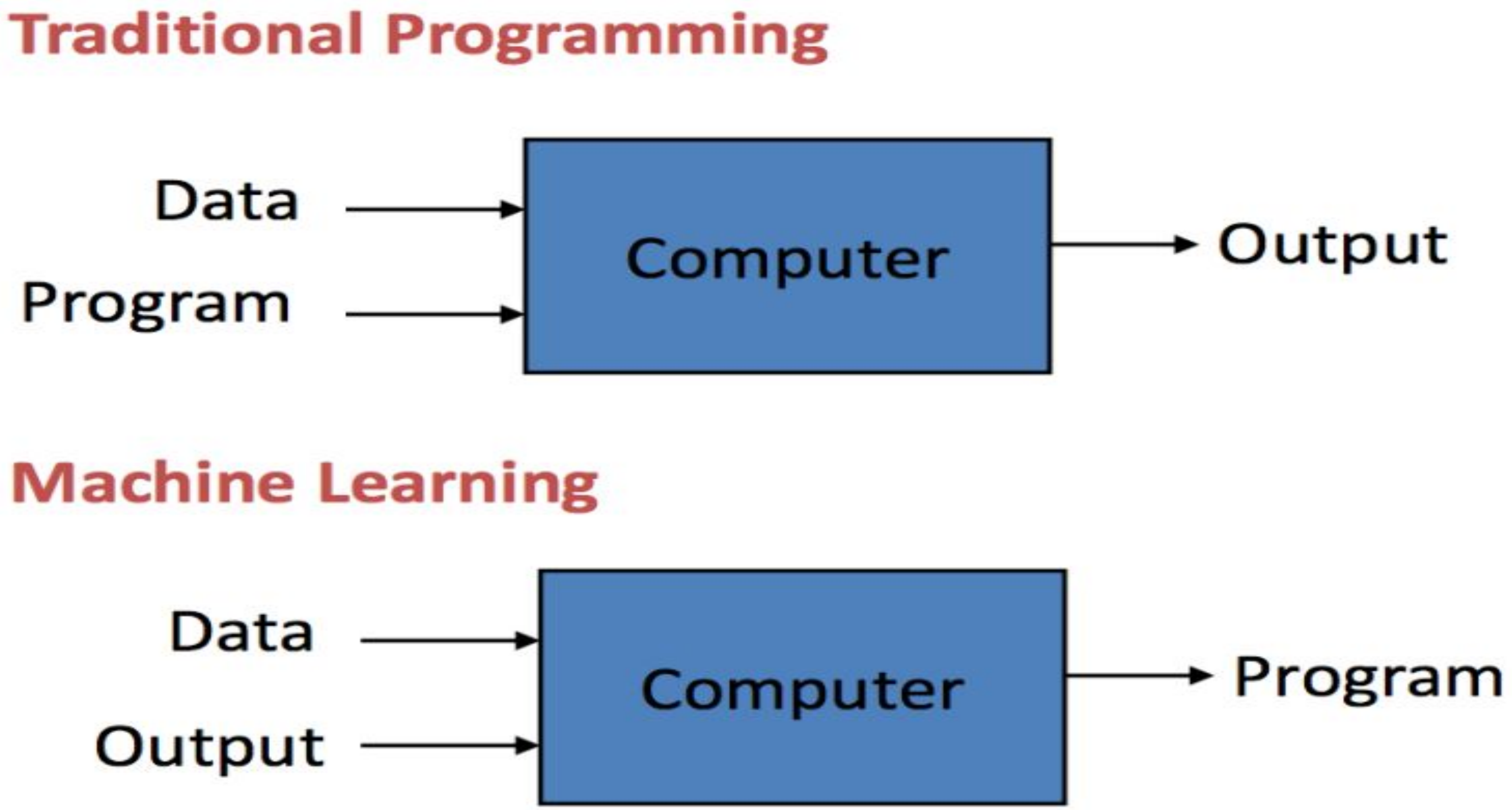
S3D


Private Investment in AI, 2013–22
Source: NetBase Quid, 2022 | Chart: 2023 AI Index Report
Total Investment (in Billions of U.S. Dollars)
0
20
40
60
80
100
120
91.86
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022

Outline

- Traditional Programing vs. ML
- Case Studies
- Model-Centric Pipeline: ML Basics
  - Features
  - Model Building
  - Evaluation
- Using ML as part of a system
  - Are there any new challenges?
  - ML Ops

Traditional Programming vs ML



Traditional Programming

“It is easy. You just chip away the stone that doesn’t look like David.” –(probably not) Michelangelo

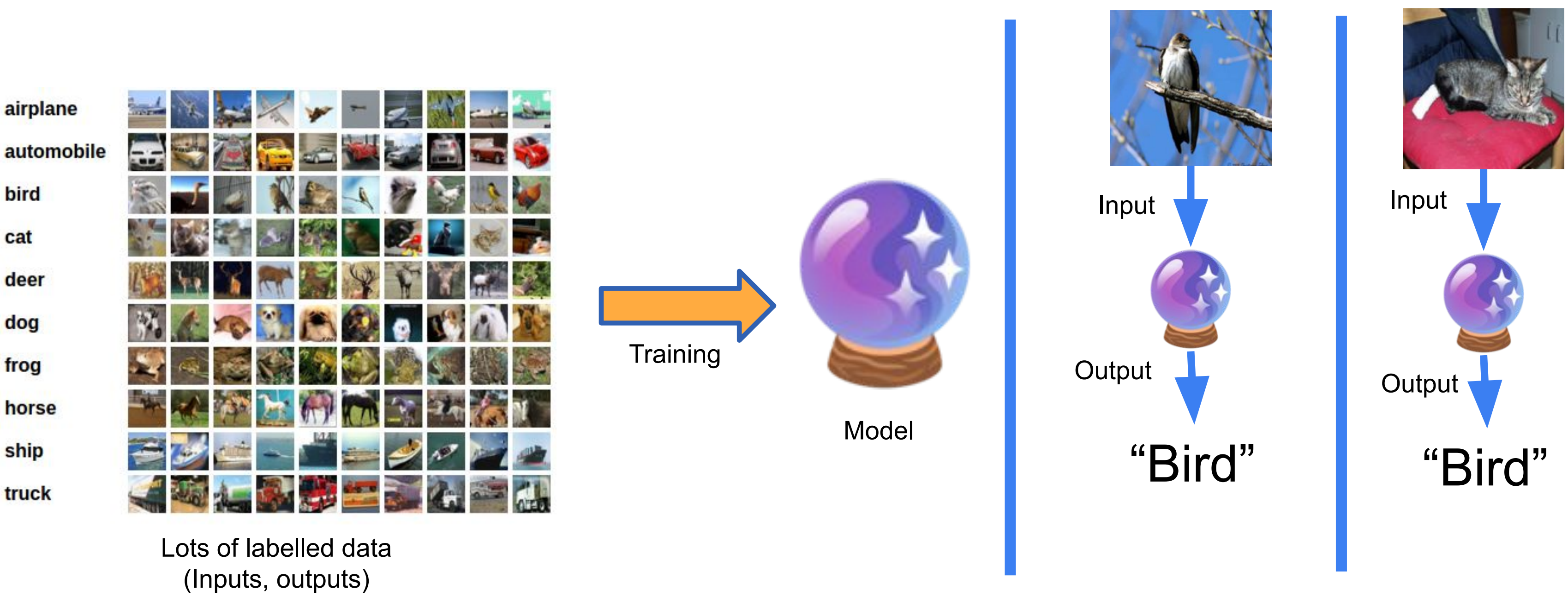
ML Development

- Observation
- Hypothesis
- Predict
- Test
- Reject or Refine Hypothesis

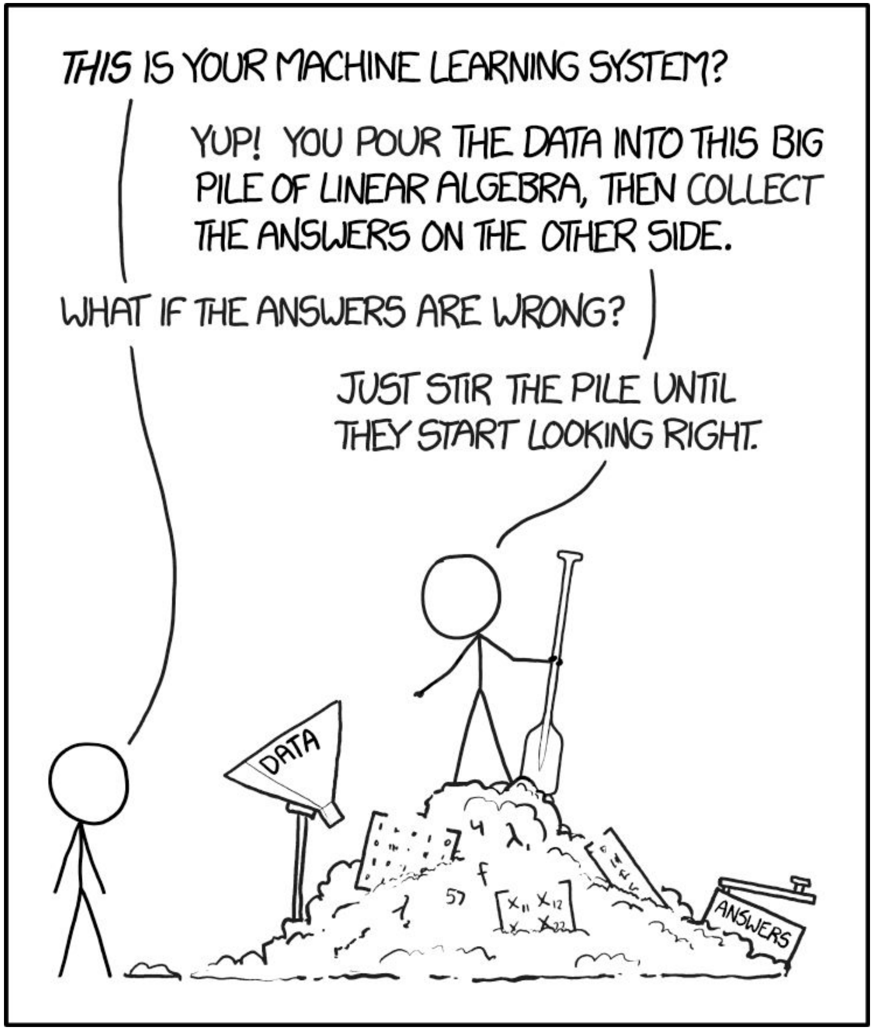
S3D

Machine Learning in One Slide

(Supervised)



Black-box view of ML



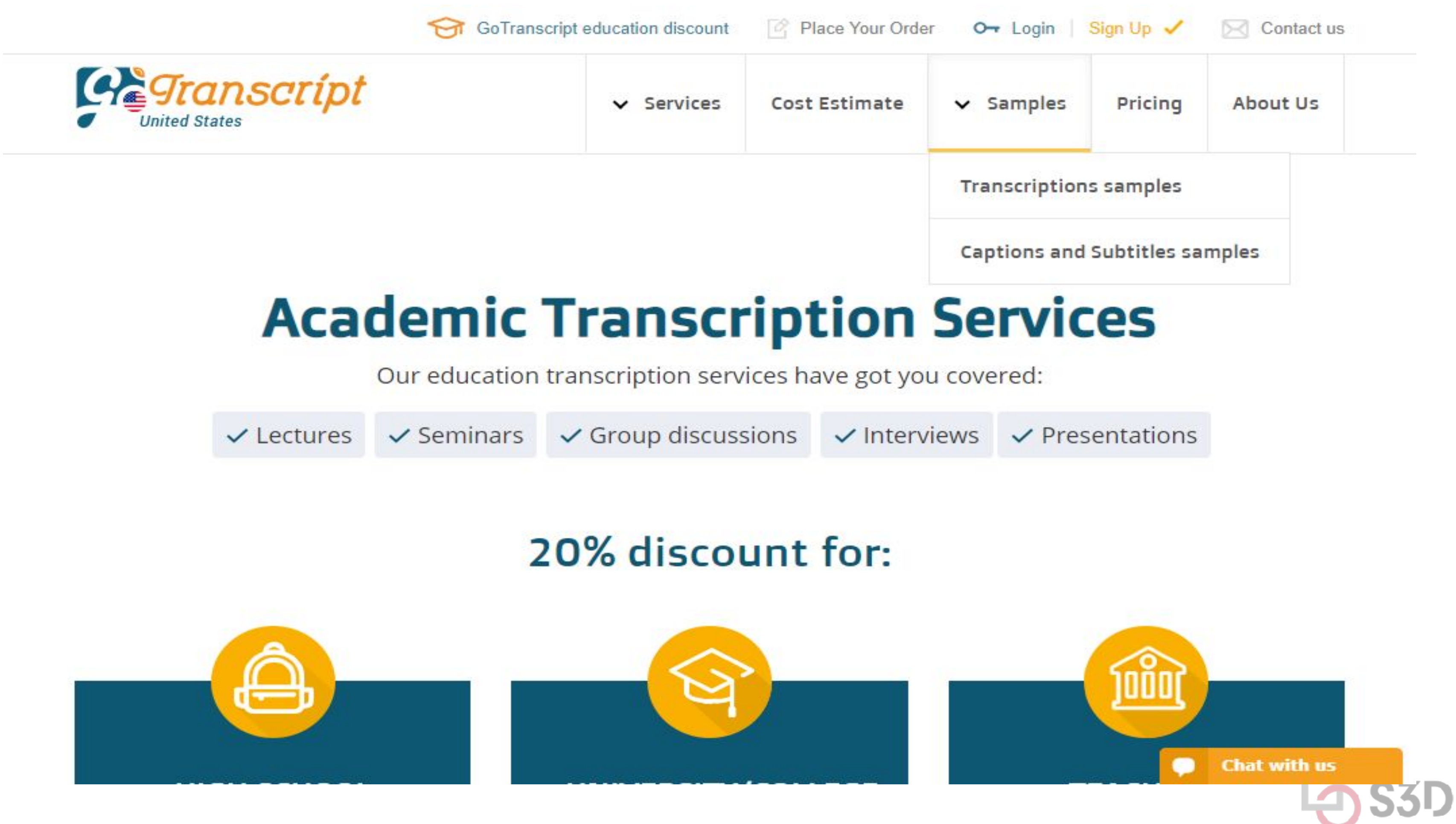
Outline

- Traditional Programing vs. ML
- Case Studies
- Model-Centric Pipeline: ML Basics
  - Features
  - Model Building
  - Evaluation
- Using ML as part of a system
  - Are there any new challenges?
  - ML Ops

Outline

- Traditional Programing vs. ML
- **Case Studies**
- Model-Centric Pipeline: ML Basics
  - Features
  - Model Building
  - Evaluation
- Using ML as part of a system
  - Are there any new challenges?
  - ML Ops

Case Study: The Transcription Service Startup



Transcription Services

- Take audio or video files and produce text.
- Used by academics to analyze interview text
- Podcast show notes
- Subtitles for videos
- State of the art: Manual transcription, often mechanical turk (1.5 $/min)

The Startup Idea

- Ph.D. research on domain-specific speech recognition that can detect technical jargon
- DNN trained on public PBS interviews + transfer learning on smaller manually annotated domain-specific corpus
- Research has shown excellent accuracy for talks in medicine, poverty and inequality research, and talks at Ruby programming conferences; published at top conferences
- Idea: Let’s commercialize the software and sell it to academics and conference organizers

Team Discussion: Likely challenges in building commercial product?

- Think about challenges that the team will likely focus on when turning their research into *a product*:
  - One machine-learning challenge
  - One engineering challenge in building the product
  - One challenge in operating and updating the product
  - One team or management challenge
  - One business challenge
  - One safety or ethics challenge

Transcription Services

- What qualities are important for a good commercial transcription product?

Case study: Smart IoT Parking Sensor

Case Study: Surge Prediction App

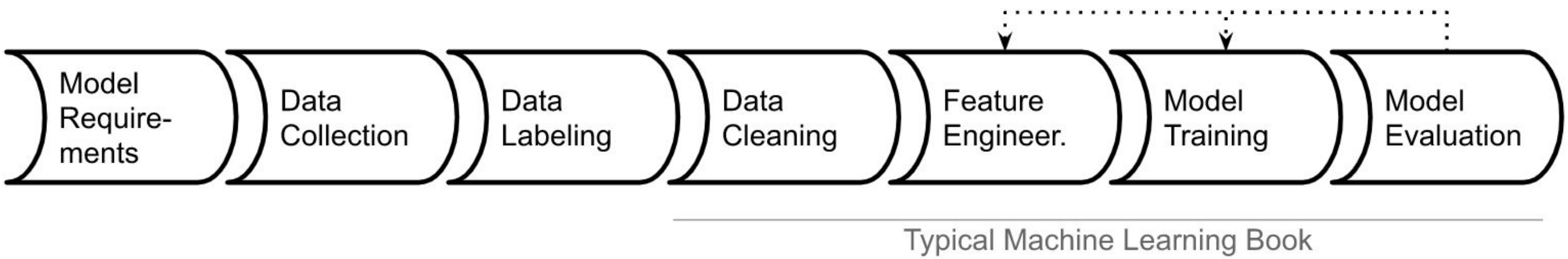
S3D

Outline

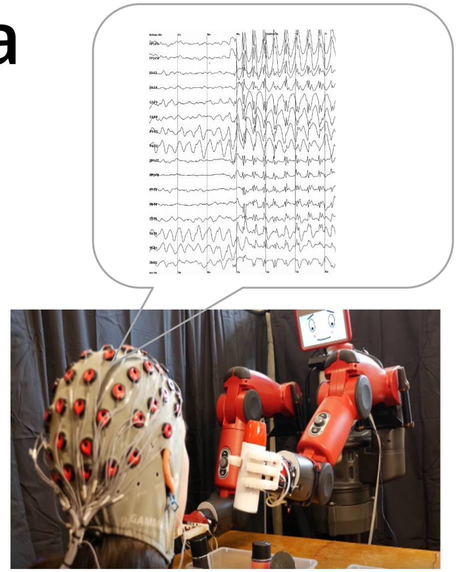
- Traditional Programing vs. ML
- Case Studies
- **Model-Centric Pipeline: ML Basics**
  - **Features**
  - **Model Building**
  - **Evaluation**
- Using ML as part of a system
  - Are there any new challenges?
  - ML Ops

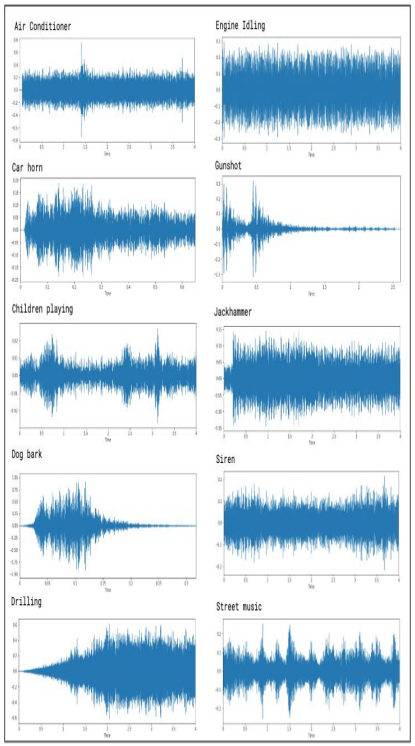
ML 101: Model-centric pipeline

- Traditional Model Focus (data science)



Example Data





UserId	PickupLocation	TargetLocation	OrderTime	PickupTime
5	....	...	18:23	18:31
...				

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	disease
63	M	3	145	233	1	0	150	0	2.3	0	0	1	Y
37	M	2	130	250	0	1	187	0	3.5	0	0	2	Y
41	F	1	130	204	0	0	172	0	1.4	2	0	2	Y
56	M	1	120	236	0	1	178	0	0.8	2	0	2	N
57	F	0	120	354	0	1	163	1	0.6	2	0	2	Y
57	M	0	140	192	0	1	148	0	0.4	1	0	1	Y
56	F	1	140	294	0	0	153	0	1.3	1	0	2	N
44	M	1	120	263	0	1	173	0	0	2	0	3	N
52	M	2	172	199	1	1	162	0	0.5	2	0	3	N
57	M	2	150	168	0	1	174	0	1.6	2	0	2	Y
54	M	0	140	239	0	1	160	0	1.2	2	0	2	N
48	F	2	130	275	0	1	139	0	0.2	2	0	2	Y
49	M	1	130	266	0	1	171	0	0.6	2	0	2	Y

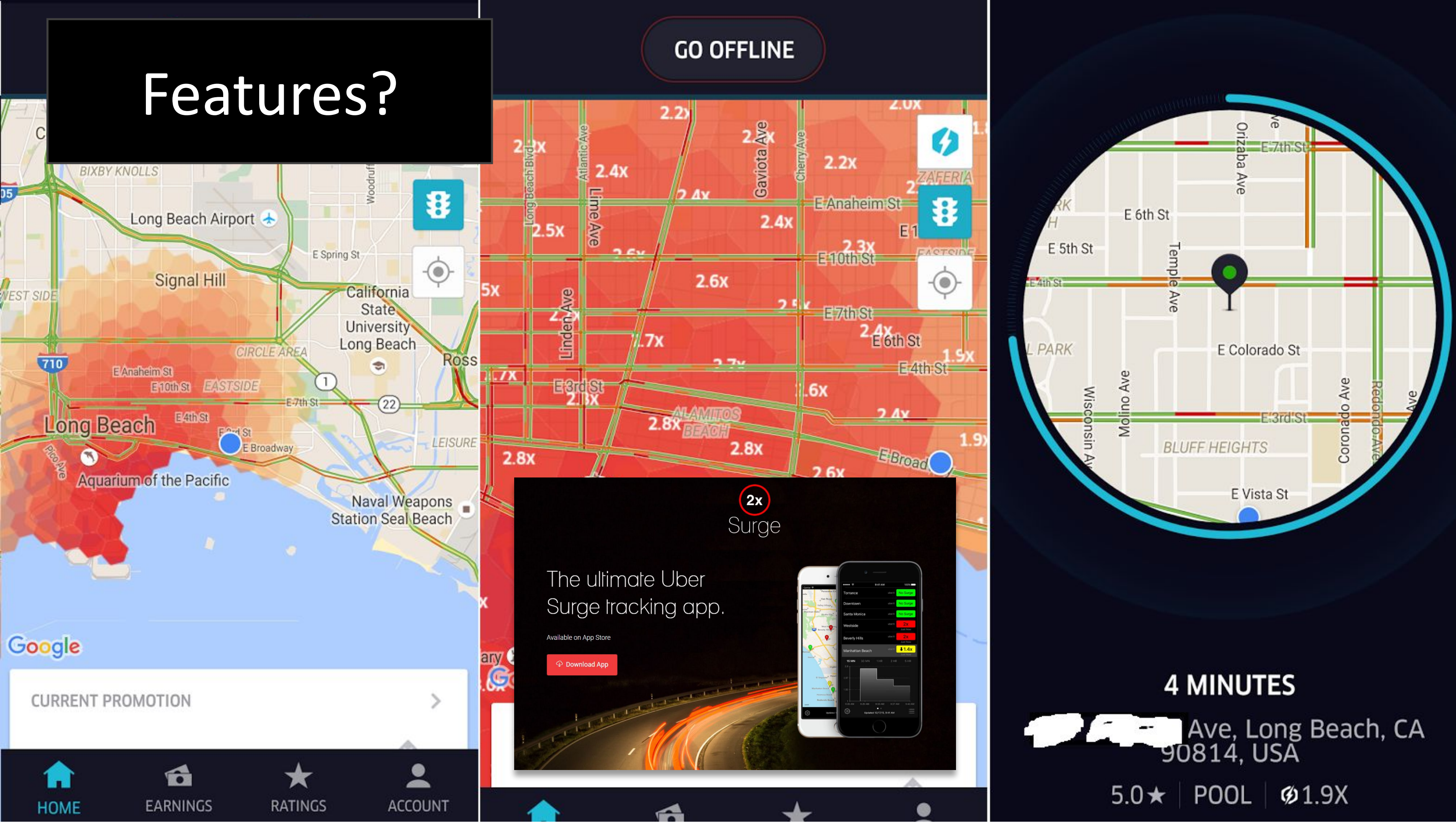
Data Cleaning

- Removing outliers
- Normalizing data
- Missing values
- ...

Feature Engineering

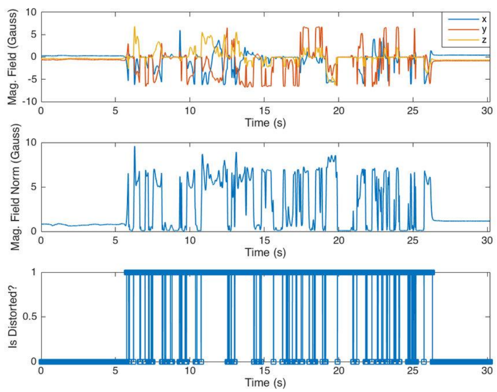
- Convert raw data into a functional form
    - Transform raw data into a more compact representation that captures the most important information in the data.
- Improve the performance of models by focusing on the most relevant information in the data
    - Remove misleading things

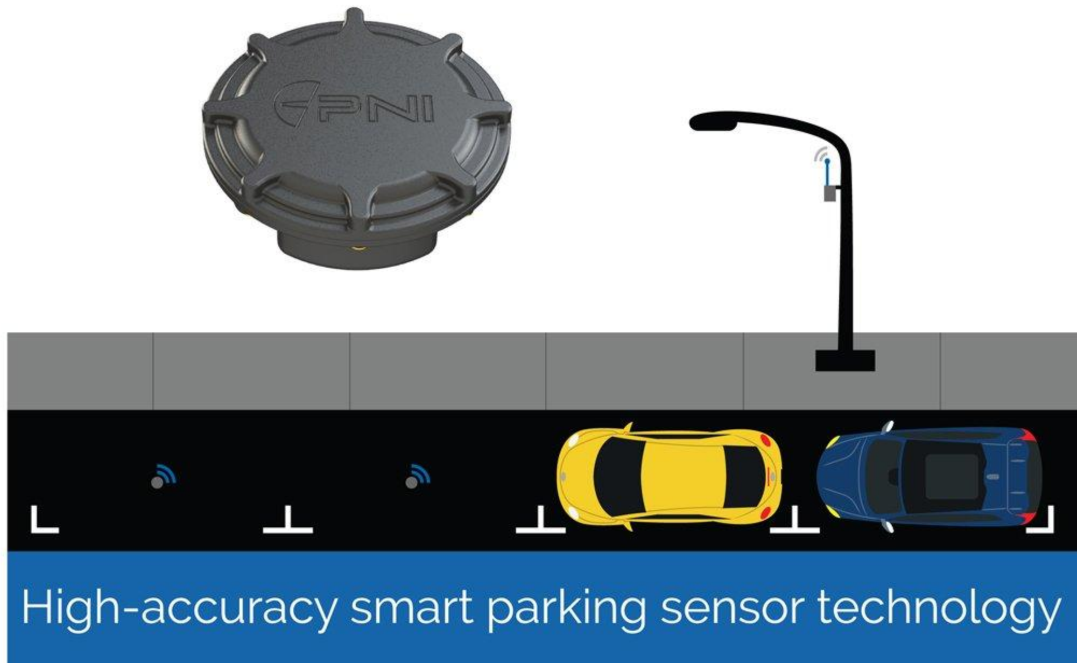
Features?



Long Beach Airport
Signal Hill
California State University Long Beach
Long Beach
Aquarium of the Pacific
Naval Weapons Station Seal Beach
CURRENT PROMOTION
HOME
EARNINGS
RATINGS
ACCOUNT
GO OFFLINE
2x
Surge
The ultimate Uber Surge tracking app.
Available on App Store
Download App
4 MINUTES
Ave, Long Beach, CA 90814, USA
5.0★
POOL
1.9X

Features?


Mag. Field (Gauss)
Time (s)
x
y
z
Mag. Field Norm (Gauss)
Time (s)
Is Distorted?
Time (s)


High-accuracy smart parking sensor technology

Feature Extraction

- In surge prediction:
  - Location and time of past surges
  - Events
  - Number of people traveling to an area
  - Typical demand curves in an area
  - Demand in other areas
  - Weather

Feature Extraction

- In vehicle detection:
  - Moving window averages
  - Peaks
  - Weather
  - Sound
  - Time of day
  - Current parking lot occupancy
  - Weather

Model Building

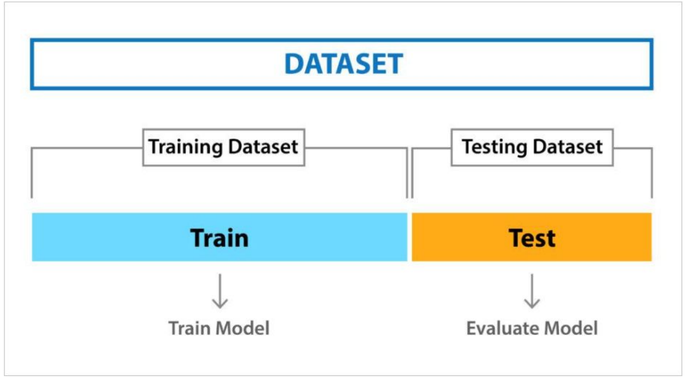
- Build a predictor that best describes an outcome for the observed features
- Many algorithms: Linear Regression, Logistic Regression, Probabilistic models, Decision Trees, Support Vector Machines, Gaussian Processes, Neural Networks

Evaluation

- Prediction accuracy on learned data vs. Prediction accuracy on unseen data
- Why?

Evaluation

- Prediction accuracy on learned data vs. Prediction accuracy on unseen data
  - Separate learning set, not used for training



Evaluation

- Binary classification: Positive / Negative
- Possible classification outcomes:
  - TN: True Negatives
  - TP: True Positives
  - FN: False Negatives
  - FP: False Positives

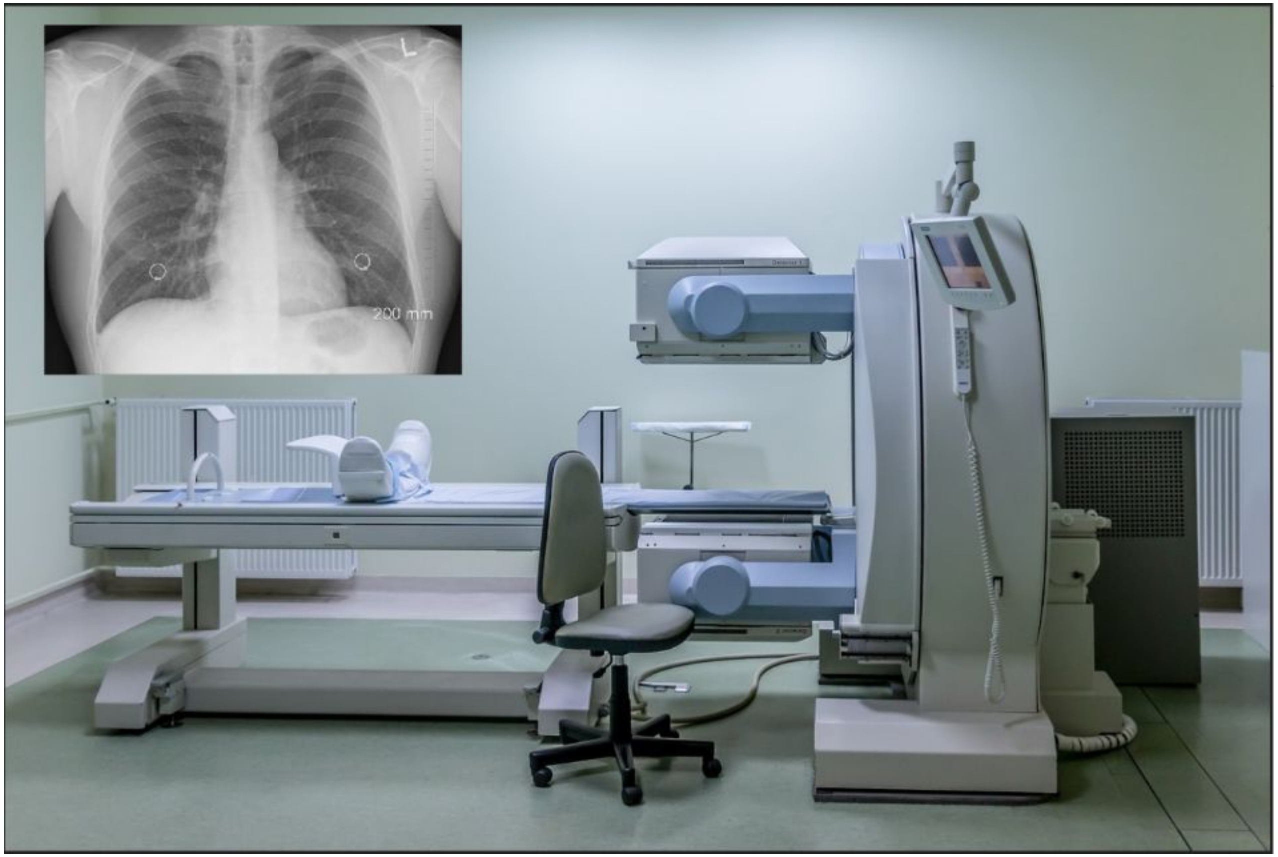
Actual Class	Predicted Class	
	Negative	Positive
Negative	**TN**	**FP**
Positive	**FN**	**TP**

Accuracy is calculated as the total number of two correct predictions (TP + TN) divided by the total number of a dataset (TP + TN + FP + FN).

S3D

Evaluation: is model accuracy enough?

Are false positives and false negatives equally bad?



False positives and false negatives, equally bad?

- Recognizing cancer
- Suggesting products to buy on e-commerce site
- Identifying human trafficking at the border
- Predicting high demand for ride sharing services
- Predicting recidivism chance
- Approving loan applications

Evaluation

- Prediction accuracy on learned data vs. Prediction accuracy on unseen data
  - Separate learning set, not used for training
- For binary predictors: false positives vs. false negatives, recall, precision
- For numeric predictors: average (relative) distance between actual and predicted value

Outline

- Traditional Programing vs. ML
- Case Studies
- Model-Centric Pipeline: ML Basics
  - Features
  - Model Building
  - Evaluation
- **Why ML/AI projects fail?**
  - **What's wrong with the model-centric pipeline?**
  - **Are there any new challenges?**
  - **ML Ops**