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
- Illustrate the challenges in engineering an ML-enabled system beyond accuracy
- Illustrate the challenges in engineering an AI-enabled system beyond accuracy

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**

Why ML/AI projects fail?

Why ML/AI projects fail?

NATIONAL HARBOR Md., June 7, 2022

Gartner Predicts Half of Finance AI Projects Will Be Delayed or Cancelled By 2024

FORBES > INNOVATION

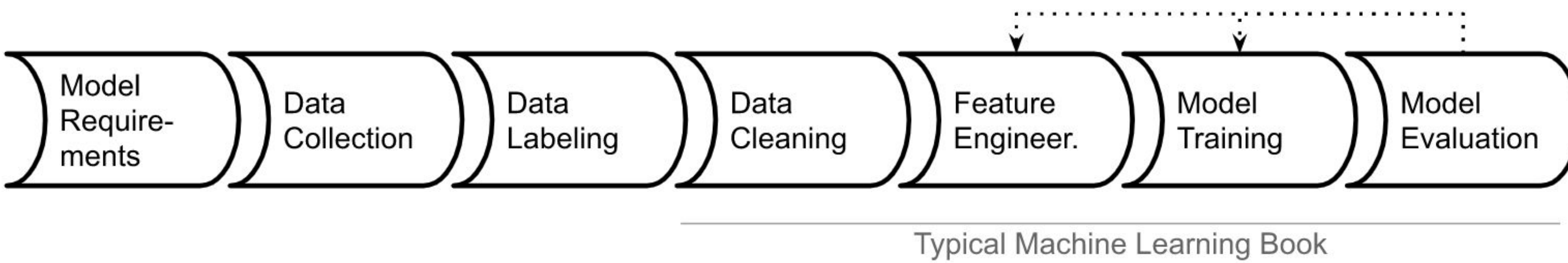
Why Most Machine Learning Applications Fail To Deploy

Usama Fayyad Forbes Councils Member
Forbes Technology Council COUNCIL POST | Membership (Fee-Based)

Apr 10, 2023, 08:45am EDT

Model-centric vs system-wide focus

- Traditional Model Focus (data science)



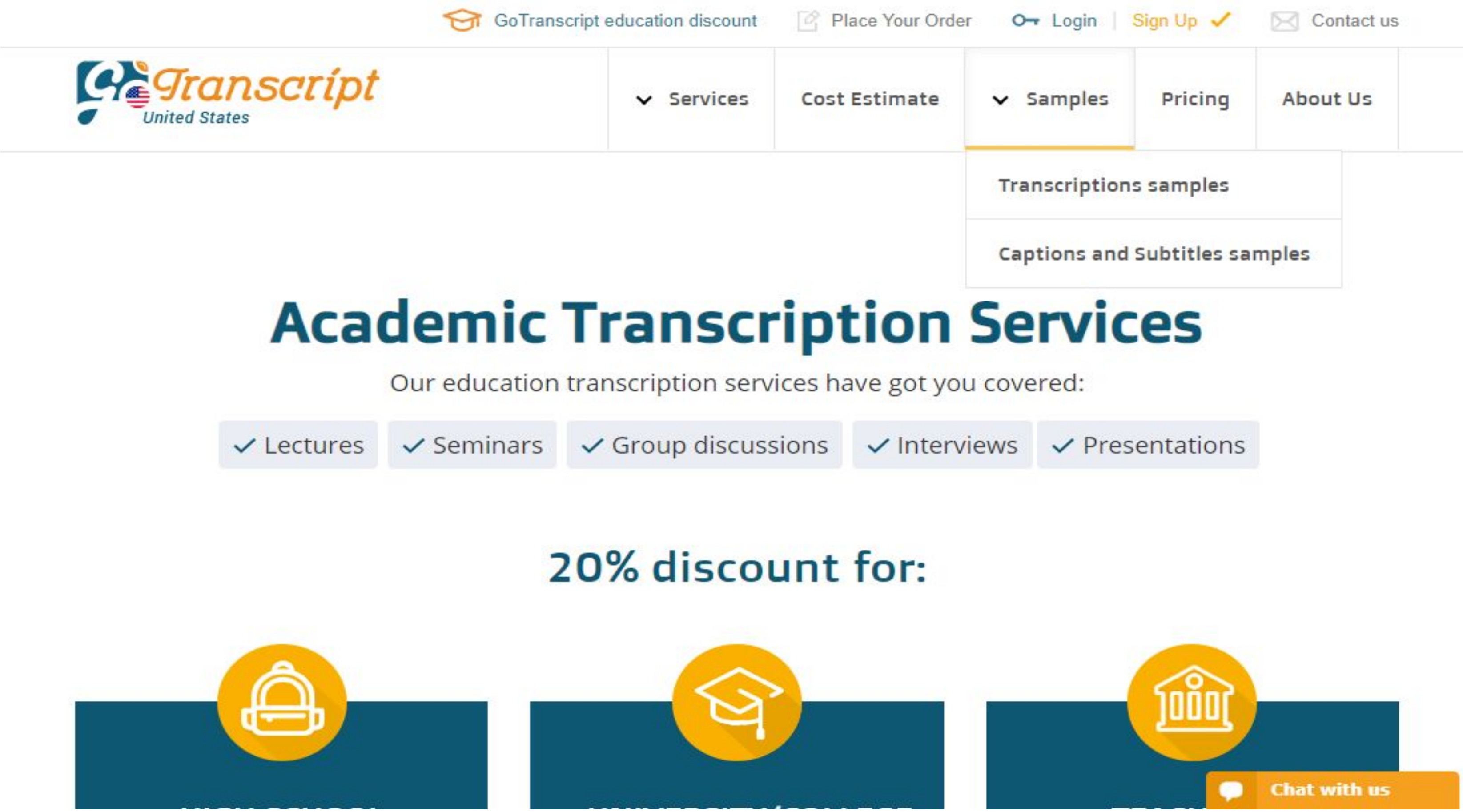
What's wrong with the model-centric pipeline?

Insufficient (relevant) data

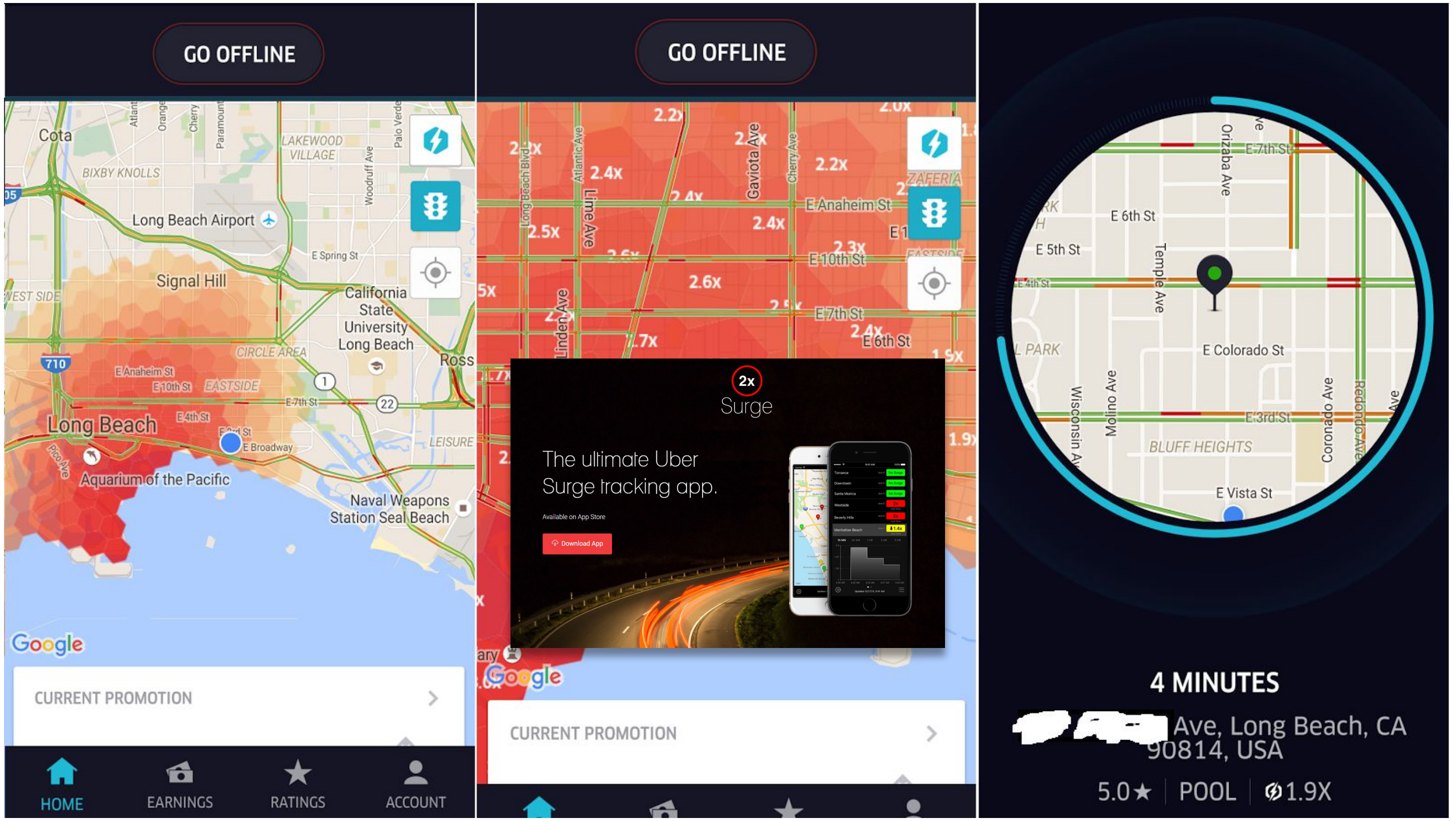
- “Little attention is paid at the one end to how data is collected and labeled.”
- “Paradoxically, data is the most undervalued and de-glamorized aspect of AI.”
- Understand the Data Requirements

World is not static

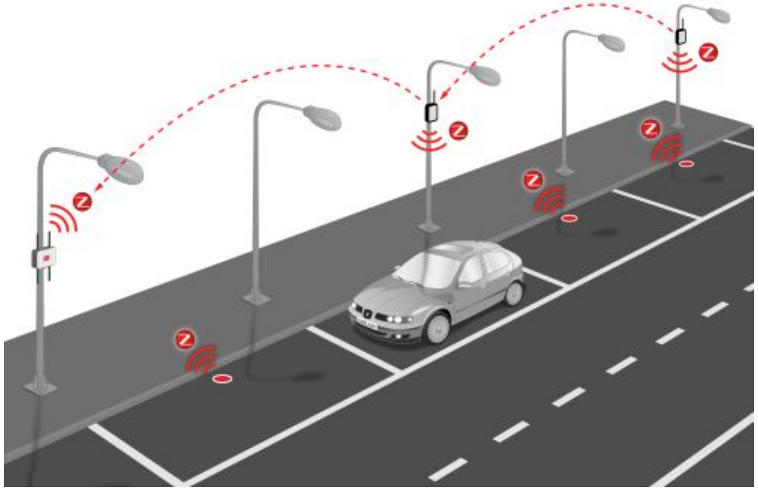
- Concepts drift
  - ML estimates `f(x) = y`
  - What if the relationship between `x` & `y` changes over time?



Reasons for change?



Reasons for change?



Reasons for change?

World is not static

- Newer better models released
  - Better model architectures
  - More training data

ML makes mistakes

NeuralTalk2: A flock of birds flying in the air
Microsoft Azure: A group of giraffe standing next to a tree
Image: Fred Dunn, https://www.flickr.com/photos/gratapictures - CC-BY-NC

Mitigation strategies?

Collecting feedback

Updating Models

- Models are rarely static outside the lab
- Data drift, feedback loops, new features, new requirements
- When and how to update models?

Human in the loop

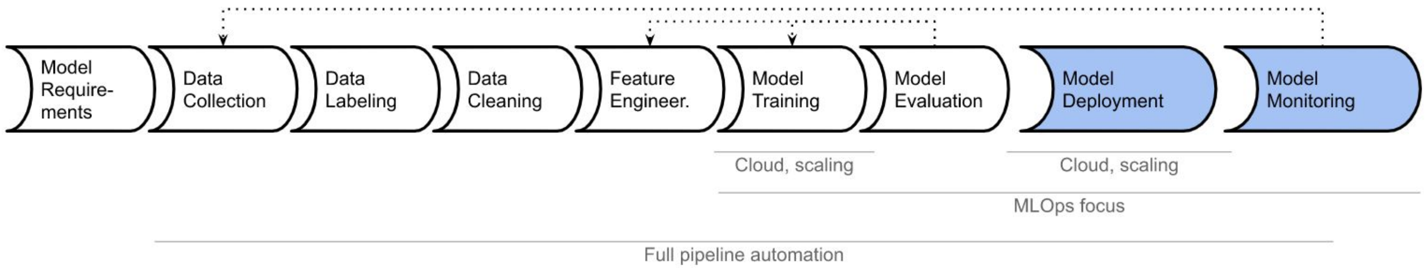
Design for failures/mistakes

- Human-AI interaction design (human in the loop):
- Guardrails
- Mistakes detection and correction
- Undoable actions

https://ckaestne.medium.com/safety-in-ml-enabled-systems-b5a5901933ac

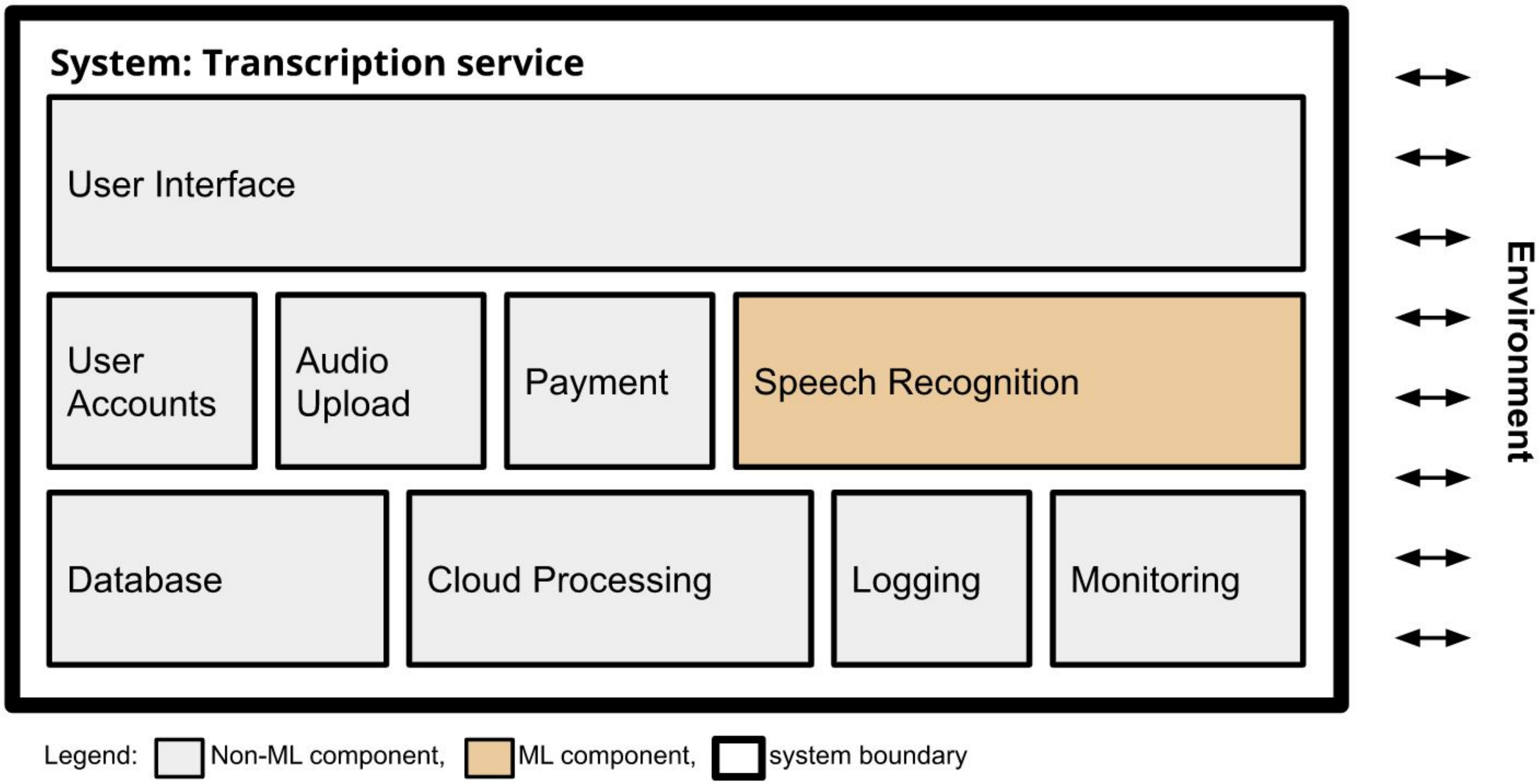
How to implement these mitigation strategies?

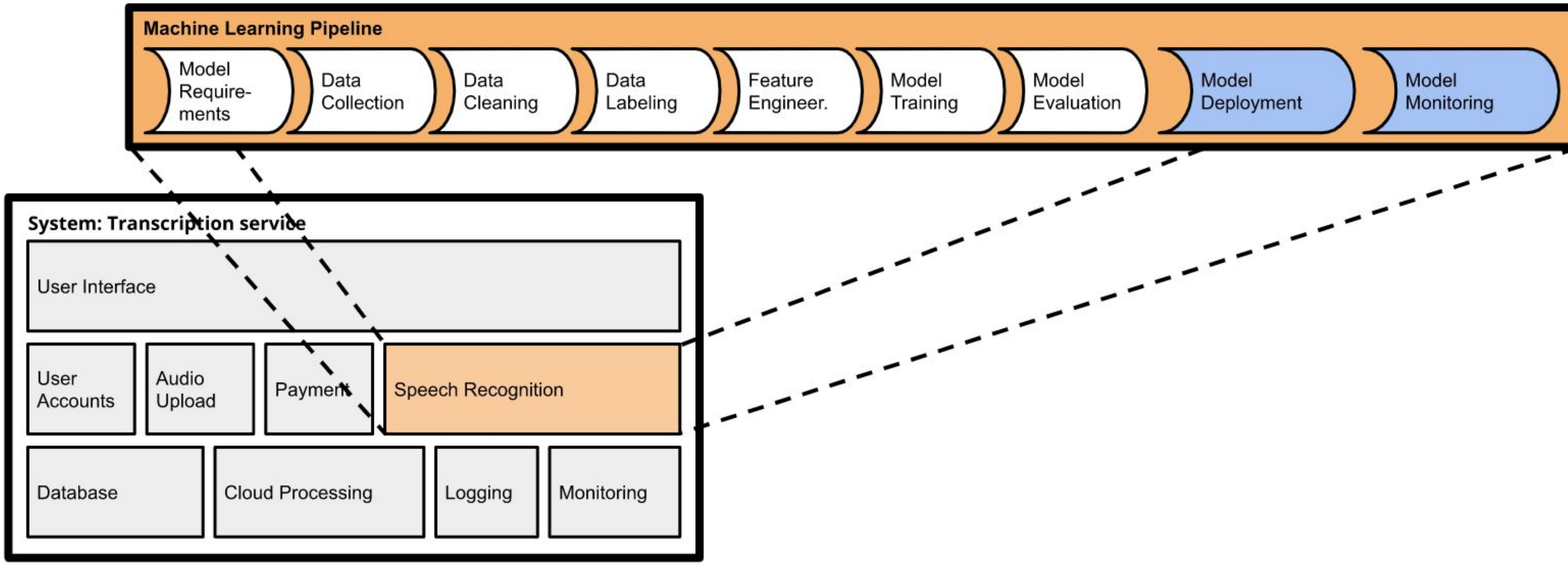
System-wide pipeline: MLOps



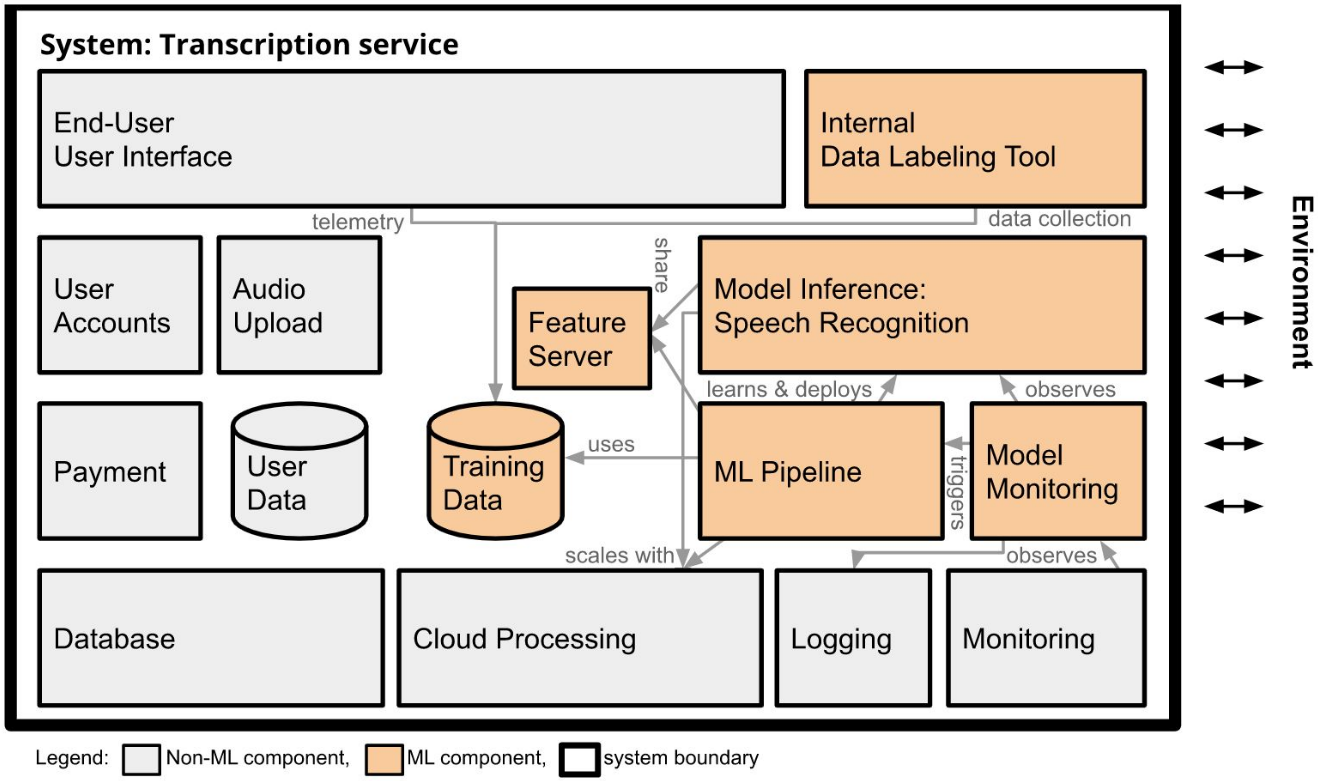
Focus: experimenting, deploying, scaling training and serving, model monitoring and updating

ML models as part of a system





Machine Learning Pipeline
Model Require-ments
Data Collection
Data Cleaning
Data Labeling
Feature Engineer.
Model Training
Model Evaluation
Model Deployment
Model Monitoring
System: Transcription service
User Interface
User Accounts
Audio Upload
Payment
Speech Recognition
Database
Cloud Processing
Logging
Monitoring


System: Transcription service
End-User
User Interface
Internal
Data Labeling Tool
telemetry
data collection
User
Accounts
Audio
Upload
share
Feature
Server
Model Inference:
Speech Recognition
learns & deploys
observes
Payment
User
Data
Training
Data
uses
ML Pipeline
triggers
Model
Monitoring
scales with
observes
Database
Cloud Processing
Logging
Monitoring
Environment
Legend: Non-ML component, ML component, system boundary

Traditional vs. System-wide ML Pipeline

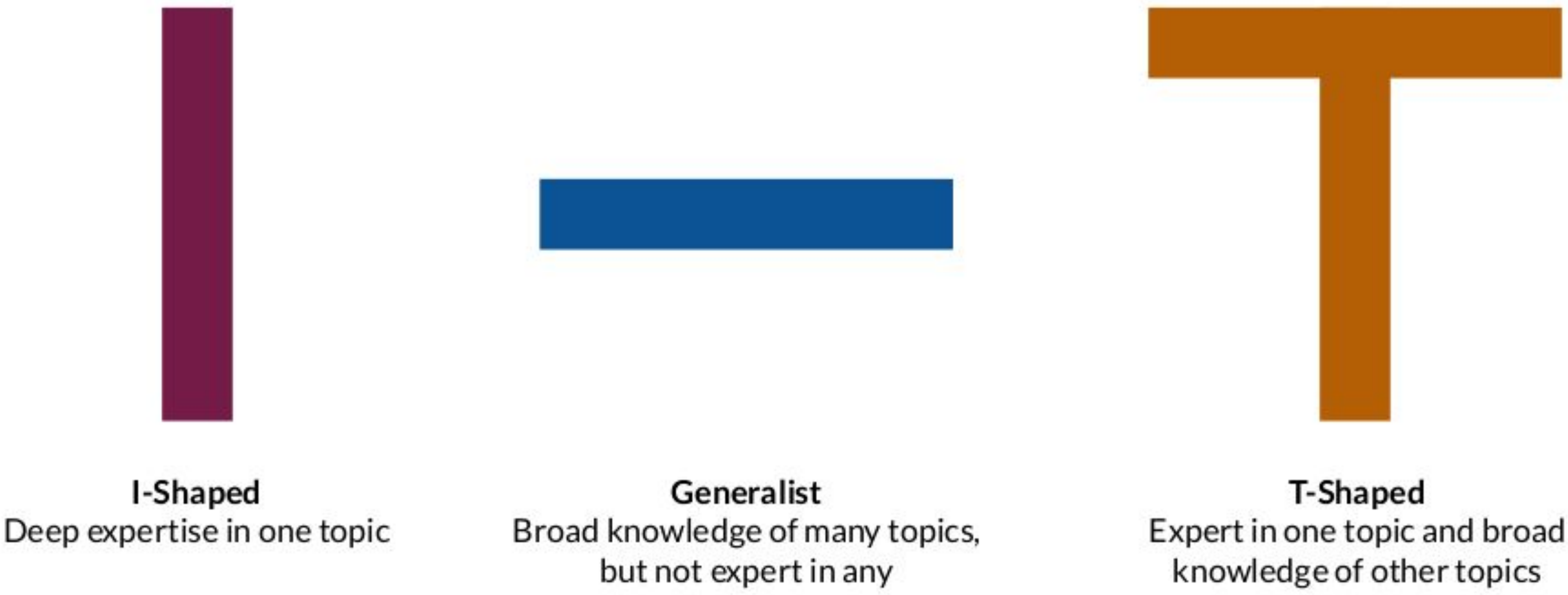
- Traditional
  - Get labeled data
  - Identify and extract features
  - Split data into training and evaluation set
  - Learn model from training data
  - Evaluate model on evaluation data
  - Repeat, revising features
- With production data
  - Evaluate model on production data; monitor
  - Select production data for retraining
  - Update model regularly

What (real) challenges are there in building and deploying systems with ML?

The road to production: a paradigm shift

S3D

T-shaped professionals



What makes software with ML challenging?

- Lack of specification (unreliability, uncertain output, mistakes)?

Lack of specification

```
/**
  Return the text spoken within the audio file
  ????
*/
String transcribe(File audioFile);
```

What makes software with ML challenging?

- Lack of specification (unreliability)?
- Complexity?

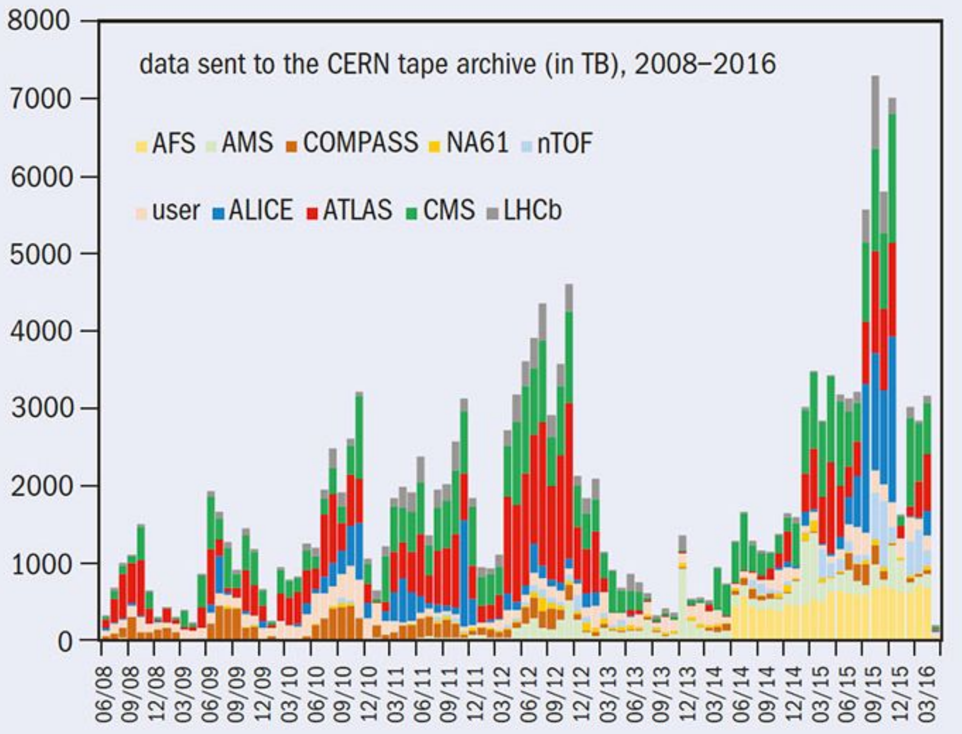
Complexity in Engineering Systems

- Automobile ~30K parts
- Airplane ~3M parts
- MS Office ~40M LOC
- Debian ~400M LOC

What makes software with ML challenging?

- Lack of specification (unreliability)?
- Complexity?
- Big Data?

Big Data?

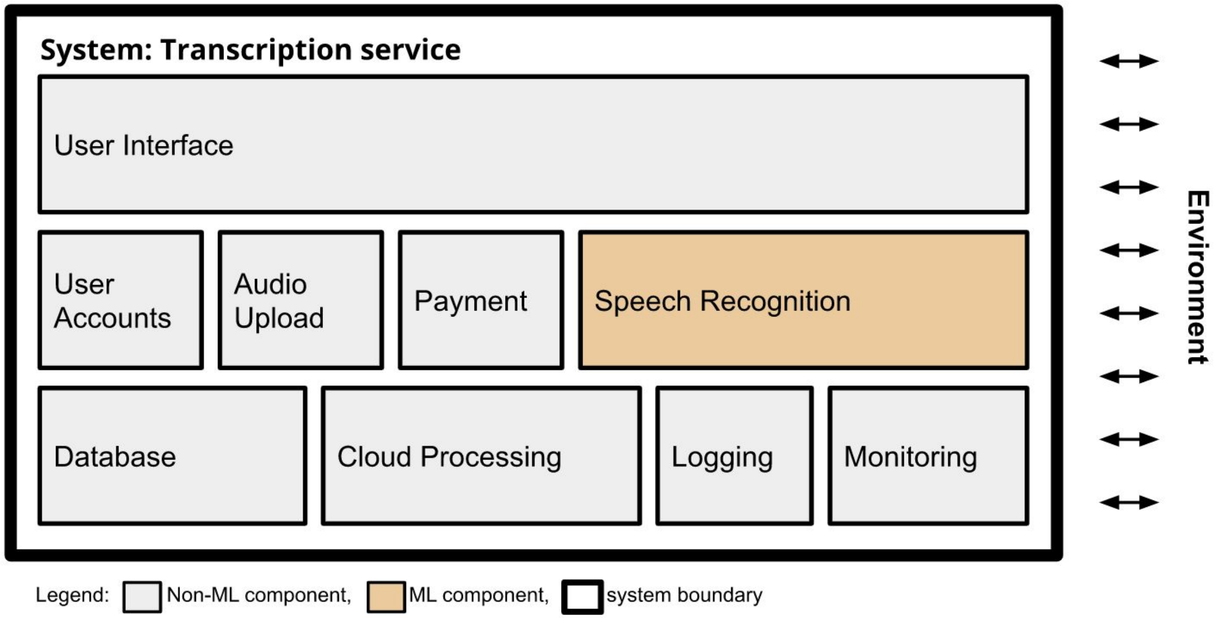


This plot represents the amount of data, in TB, being sent to the CERN archive between 2008 and 2016. The yearly amount of LHC data has gradually increased since 2010 (Run 1, 2010: 12.5 PB, 2011: 19.1 PB, 2012: 27 PB) and during Run 2 (31.5 PB).
Image credit: CERN.

What makes software with ML challenging?

- Lack of specification (unreliability)?
- Complexity?
- Big Data?
- Interaction with the environment?

Interaction with the environment



What challenges could be new? What challenges could be magnified?

Safety?

https://www.alphr.com › review › smart-toaster

The Highest-Rated Smart Toasters in 2022 - Alphr Reviews

Aug 19, 2022 — Works on **artificial intelligence** (**AI**). A **smart toaster** operates on **artificial intelligence** to detect and control the whole toast-making process, ...

Safety risks?

How can you mitigate these risks?

Interaction with the environment: safety

The Daily Star

Microwave attempts to murder owner after gaining artificial intelligence 'demon soul'

A YouTuber who tried to resurrect his childhood imaginary friend by giving a microwave artificial intelligence says it tried to kill him.

Apr 28, 2022

S3D

Safety Assurance in/outside the Model

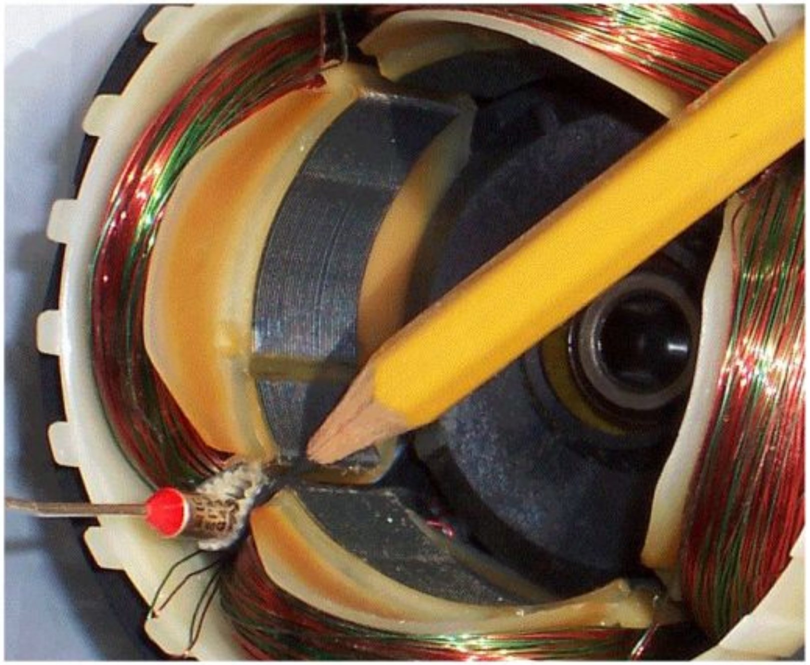
In the model

- Ensure maximum toasting time
- Use heat sensor and past outputs for prediction

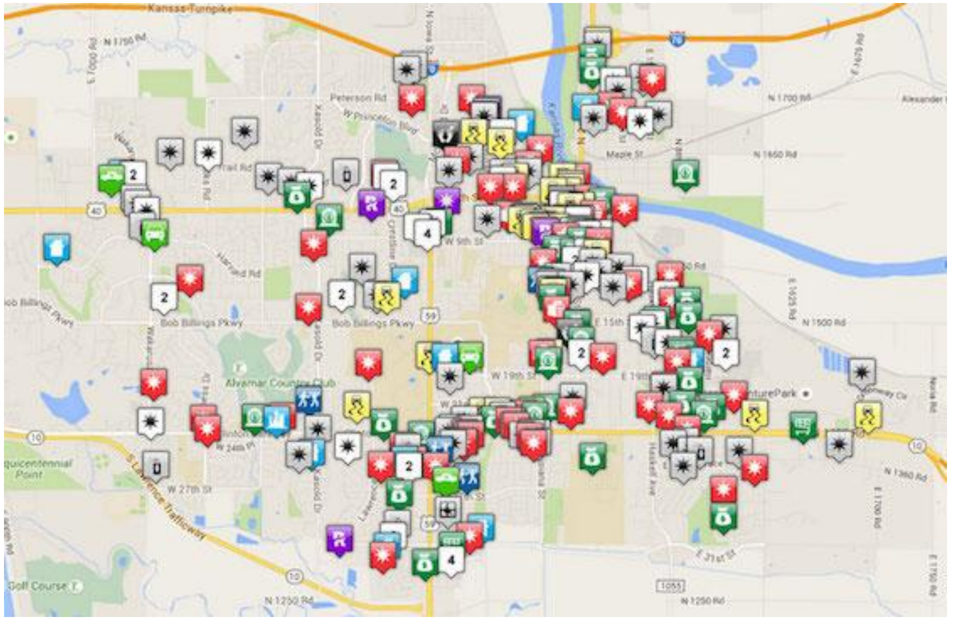
Hard to make guarantees

Outside the model

- Simple code check for max toasting time
- Non-ML rule to shut down if too hot
- Hardware solution: thermal fuse

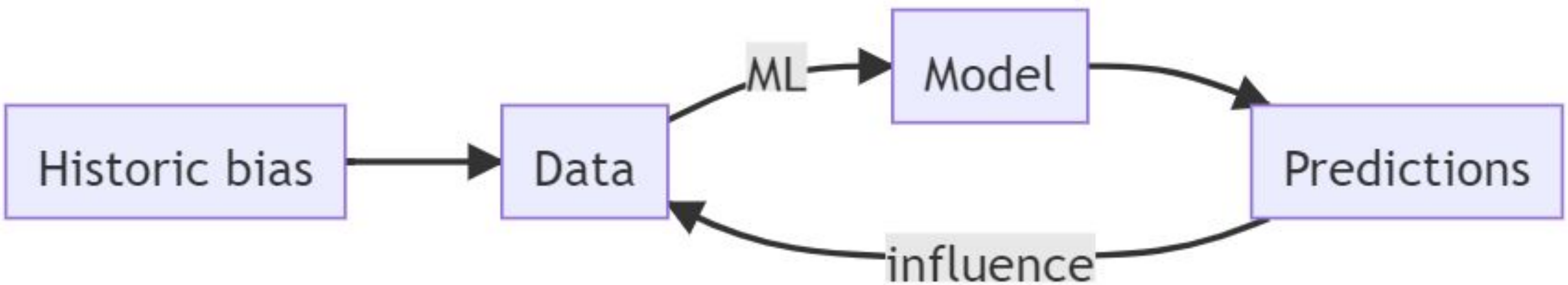


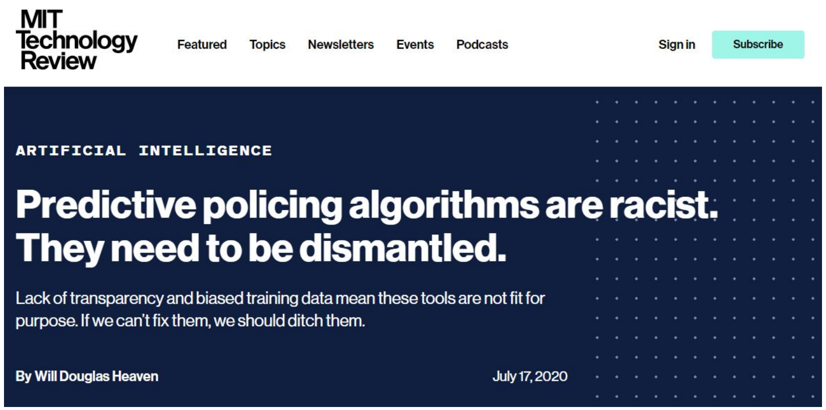
Interaction with the environment: feedback loops



ML Model: Use historical arrest records to predict crime rates by neighborhoods
Used for predictive policing: Decide where to allocate police patrol

Feedback loops




MIT Technology Review

Featured Topics Newsletters Events Podcasts

Sign in Subscribe

ARTIFICIAL INTELLIGENCE

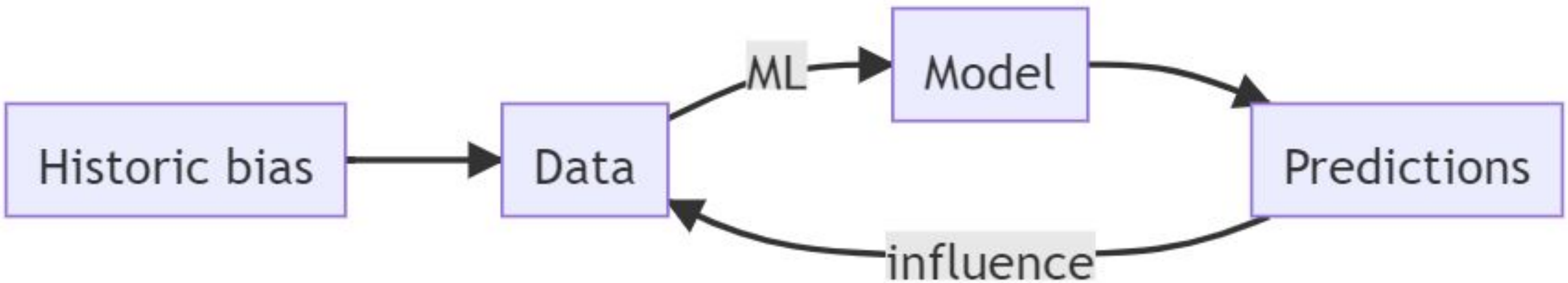
Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

By Will Douglas Heaven

July 17, 2020

Feedback loops



The New York Times

THE SHIFT

YouTube Unleashed a Conspiracy Theory Boom. Can It Be Contained?

What makes software with ML challenging?

- Lack of specification (unreliability)
- Complexity
- Big Data
- Interaction with the environment

What makes software (systems) with ML challenging?

- **It’s not all new**
- Safe software with unreliable components
- Cyber-physical systems
- Non-ML big data systems, cloud systems
- "Good enough" and "fit for purpose" not "correct"
- **We routinely build such systems**
- **ML intensifies our challenges**

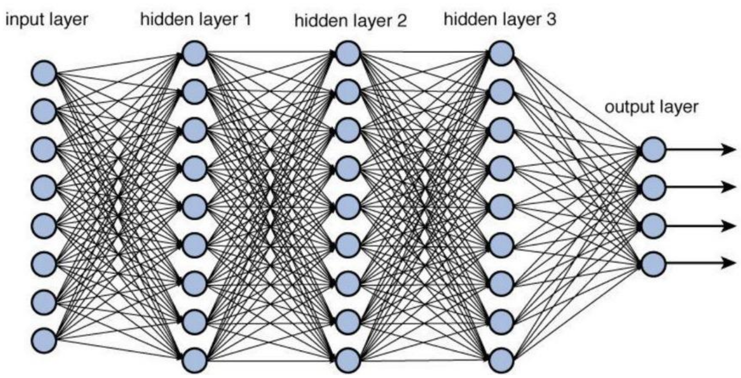
ML COMPONENT TRADEOFFS

Qualities of ML Components

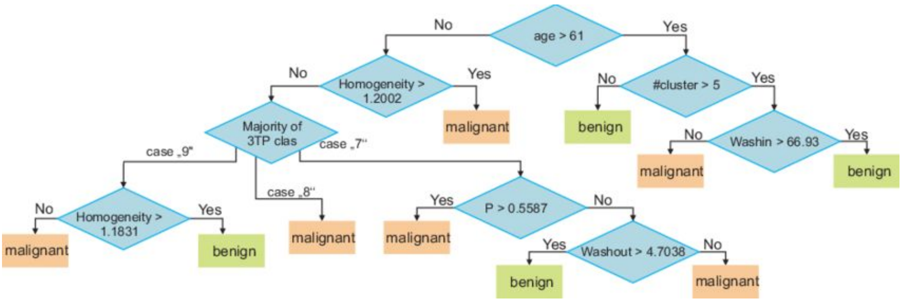
- Accuracy
- Capabilities (e.g. classification, recommendation, clustering…)
- Amount of training data needed
- Inference latency
- Learning latency; incremental learning?
- Model size
- Explainable? Robust?
- …

Understanding Capabilities and Tradeoffs

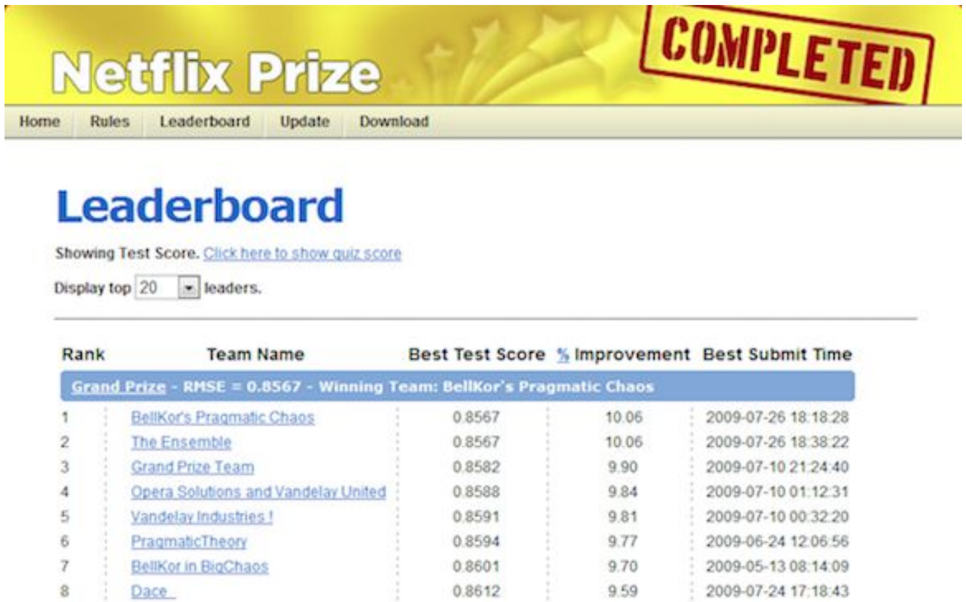
- Deep Neural Networks



- Decision Trees



Trade-offs: Cost vs Accuracy



Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update | Download

Leaderboard

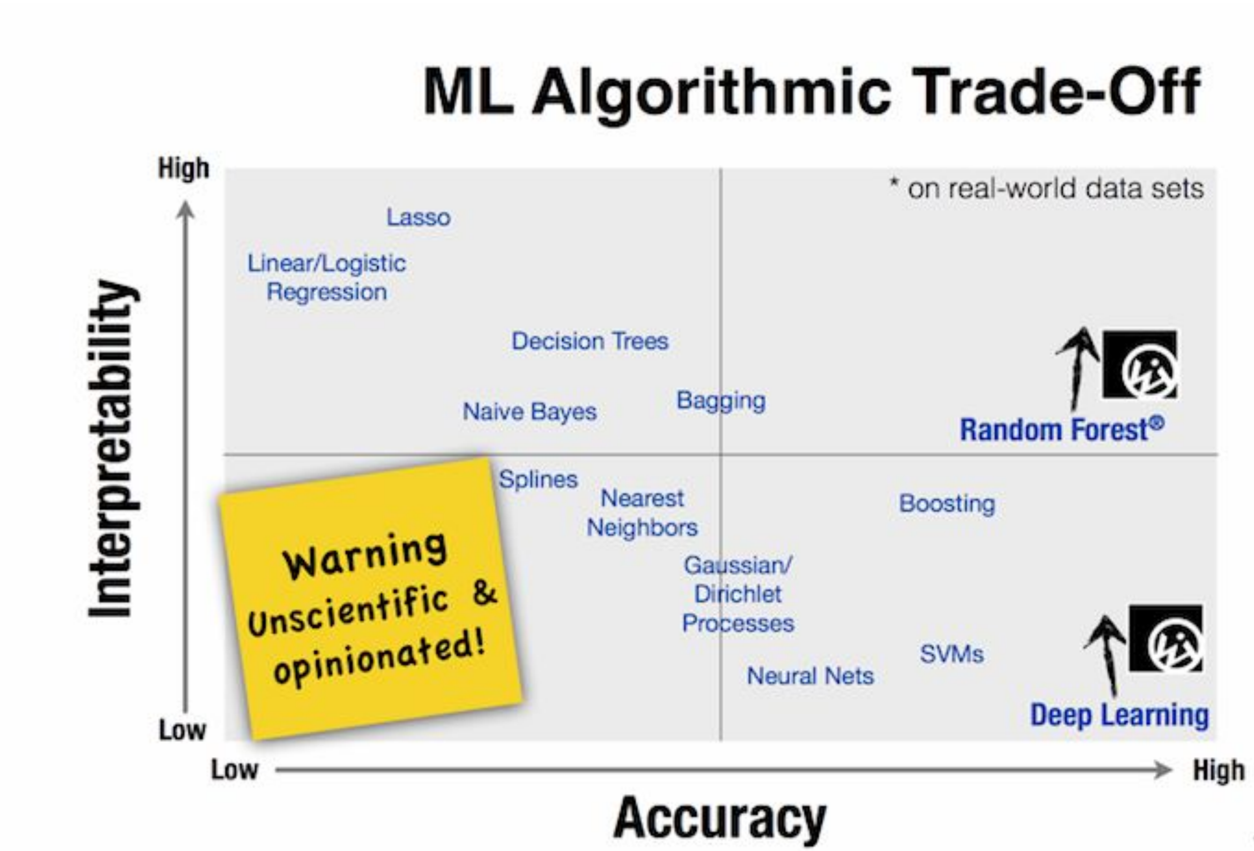
Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43

"We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment."

Trade-offs: Accuracy vs Interpretability



SYSTEM ARCHITECTURE CONSIDERATIONS

Case Study: Augmented reality translation

Where should the model live?

Glasses

Phone

Cloud

OCR Component

From images to text

Translation Component

From text to text

Where should the model live?

Car

Phone

Surge Prediction

Cloud

S3D

Where should the model live?

Pod

Gateway

Cloud

Car Detector

Typical Designs

- Static intelligence in the product
  - difficult to update
  - good execution latency
  - cheap operation
  - offline operation
  - no telemetry to evaluate and improve
- Client-side intelligence
  - updates costly/slow, out of sync problems
  - complexity in clients
  - offline operation, low execution latency

Considerations

- How much data is needed as input for the model?
- How much output data is produced by the model?
- How fast/energy consuming is model execution?
- What latency is needed for the application?
- How big is the model? How often does it need to be updated?
- Cost of operating the model? (distribution + execution)
- Opportunities for telemetry?
- What happens if users are offline?

Summary

- Production AI-enabled systems require a *whole system perspective* beyond just the model or the pipeline

- Machine learning brings new challenges and intensifies old ones

- Building ML systems need team efforts
  - Collaborative culture among Software Engineers, Data Scientists, Stakeholders is necessary