Software Engineering for ML/AI

Claire Le Goues

Michael Hilton

Christopher Meiklejohn

institute for SOFTWARE RESEARCH

Carnegie Mellon University
School of Computer Science

Administrivia

- Homework 2 (Code Artifacts) due today.

Learning goals

- Identify differences between traditional software development and development of ML systems.
- Understand the stages that comprise the typical ML development pipeline.
- Identify challenges that must be faced within each stage of the typical ML development pipeline.

Participation Survey

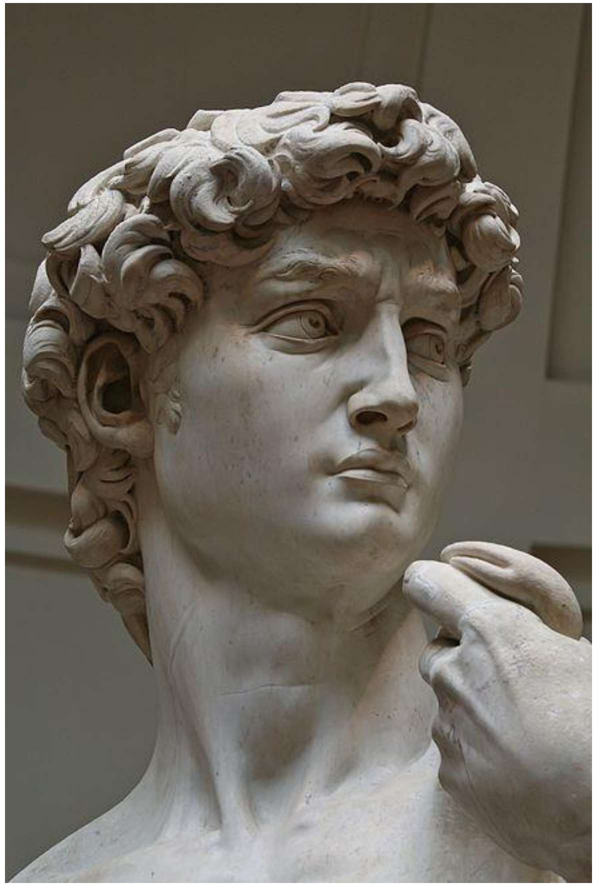
- **YES** in Zoom:
  "I've a taken machine learning course."

- **NO** in Zoom:
  "I have not taken a machine learning course."

Software Engineering and ML

Traditional Software Development

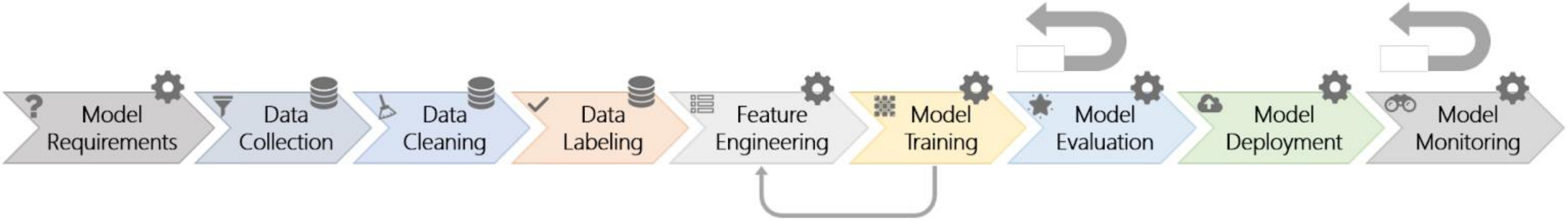
"It is easy. You just chip away the stone that doesn't look like David." – (probably not) Michelangelo



ML Development

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

Microsoft's view of Software Engineering for ML



Three Fundamental Differences:

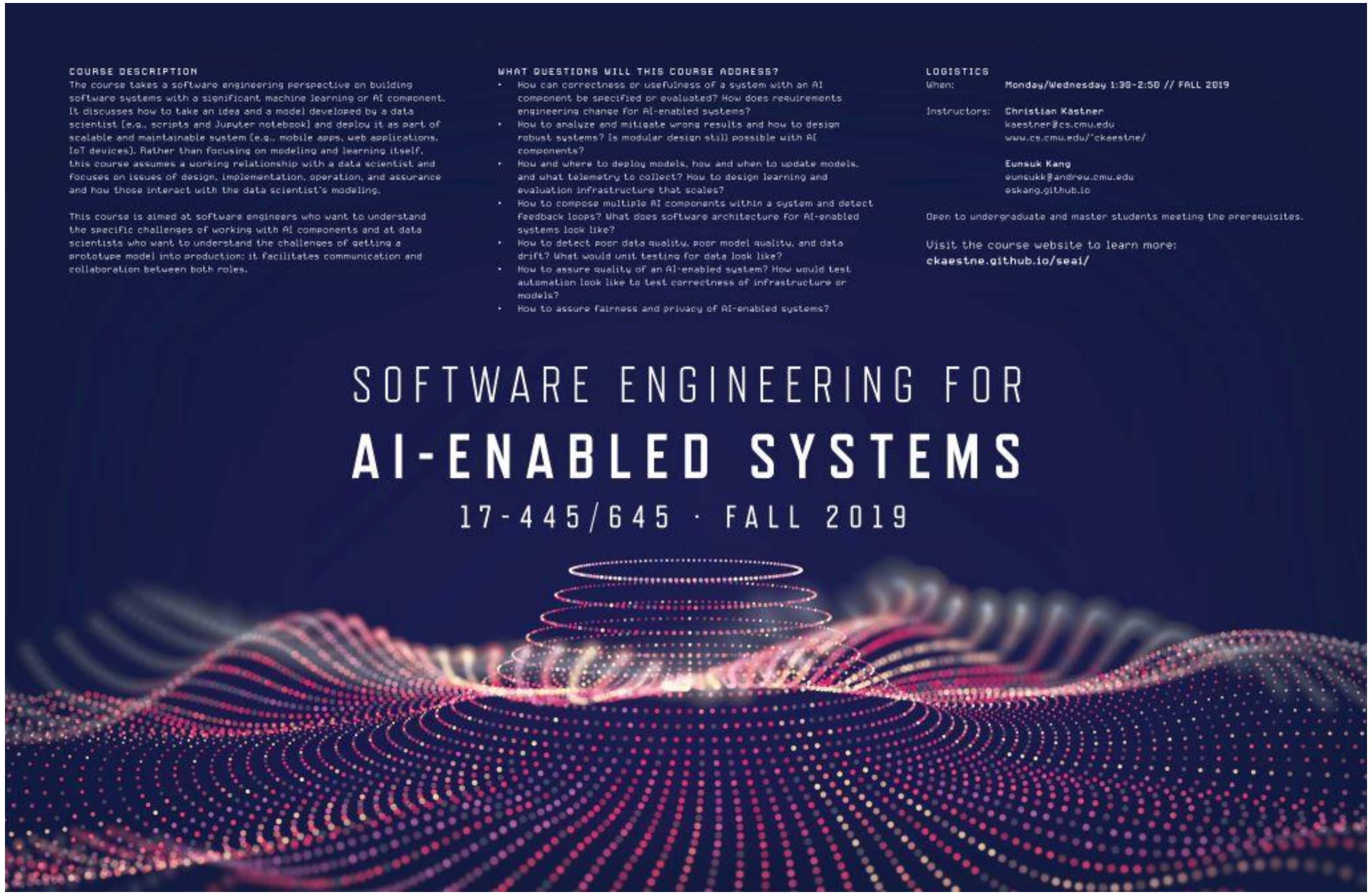
- Data discovery and management
- Customization and Reuse
- No incremental development of model itself

Case study developed by

Christian Kästner

https://ckaestne.github.io/seai/

CASE STUDY



WHAT CHALLENGES ARE THERE IN BUILDING AND DEPLOYING ML?


現金のみ
Japanese English
cash only

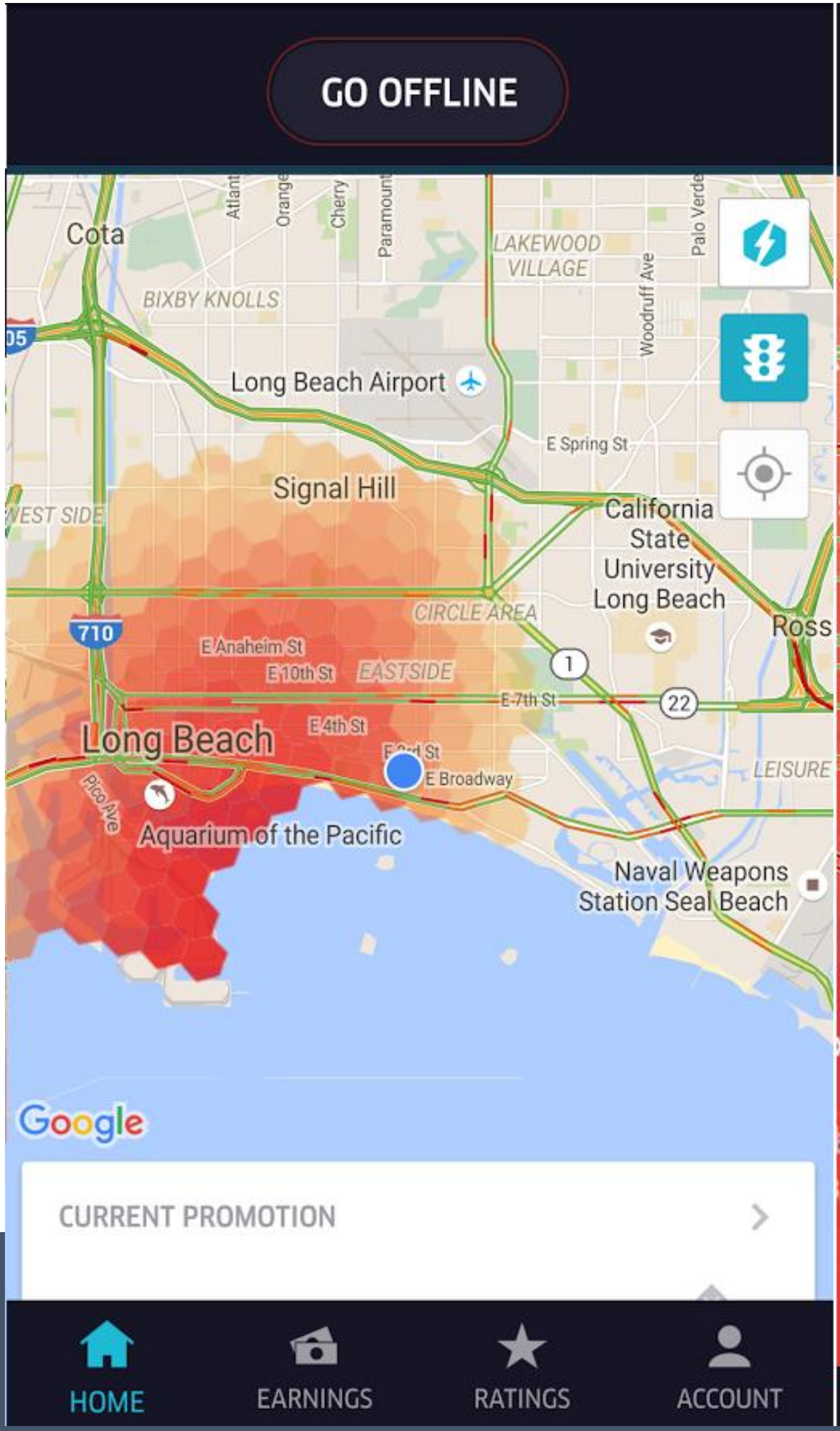

PhoDUCK
TIBETAN INDIAN NEPALI FOOD
POTALA
명동교자
아하 노래방
POTALA RESTAURANT
빤달빤달
鸡肋
タッカルビ
DAKGALBI
명동본점 2F
제임스 치즈 등갈비
MINI BEER PUB
LE PIANO
홍콩반점
부대찌개
정성본
김밥
명동교자
수제 돈까스
Homemade pork cutlet
SKIN
노래방

Qualities of Interest?

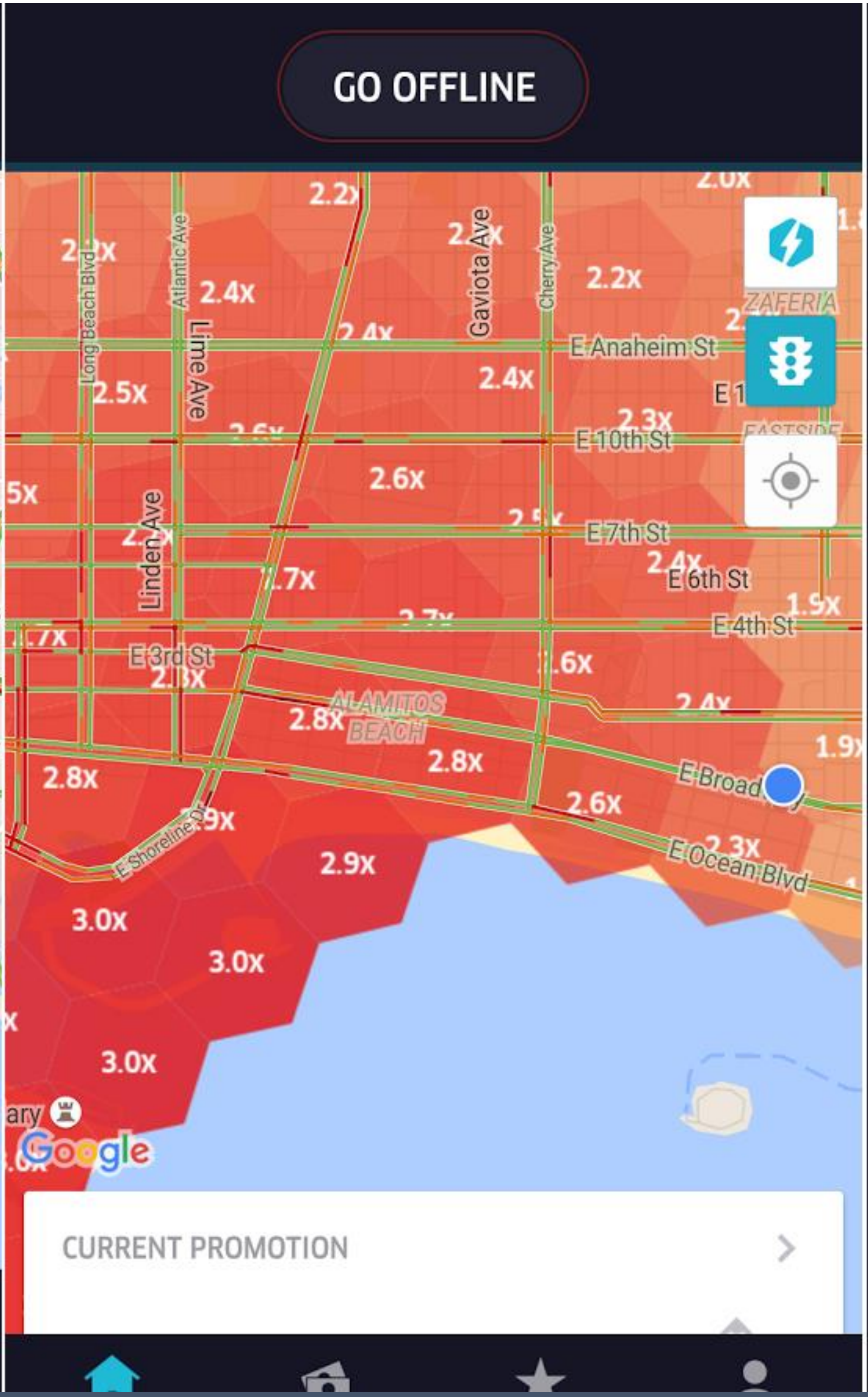
A

B

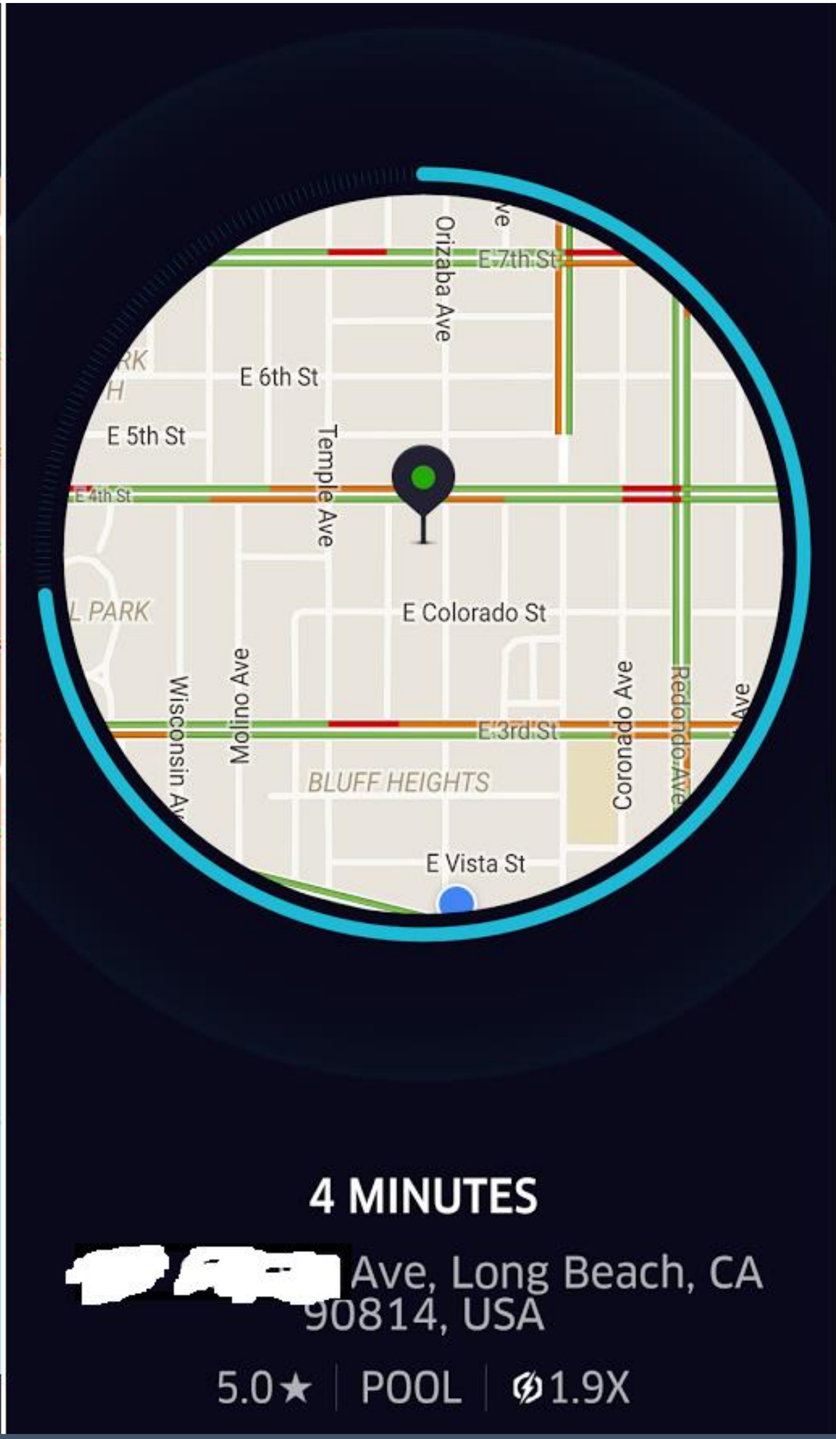
C



GO OFFLINE
Cota
BIXBY KNOLLS
LAKEWOOD VILLAGE
Long Beach Airport
E Spring St
Signal Hill
California State University Long Beach
CIRCLE AREA
Ross
E Anaheim St
E 10th St
EASTSIDE
E 7th St
E 4th St
Long Beach
E Broadway
LEISURE
Aquarium of the Pacific
Naval Weapons Station Seal Beach
Google
CURRENT PROMOTION
HOME
EARNINGS
RATINGS
ACCOUNT



GO OFFLINE
2.2x
2.4x
2.5x
2.6x
2.3x
2.7x
2.8x
2.9x
3.0x
1.9x
Gaviota Ave
Lime Ave
Linden Ave
E Anaheim St
E 10th St
E 7th St
E 6th St
E 4th St
E 3rd St
ALAMITOS BEACH
E Broadway
E Ocean Blvd
E Shoreline Dr
Google
CURRENT PROMOTION

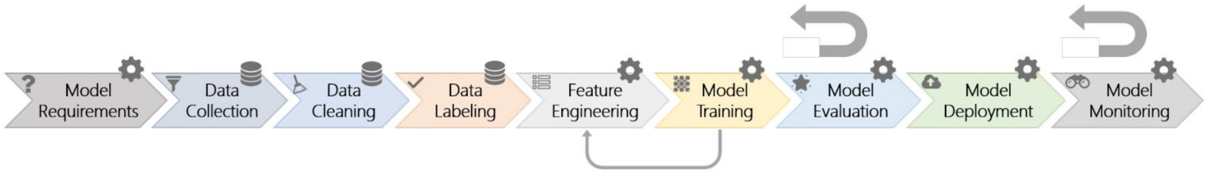


Orizaba Ave
E 7th St
E 6th St
E 5th St
Temple Ave
E Colorado St
Molino Ave
Wisconsin Ave
Coronado Ave
Redondo Ave
E 3rd St
BLUFF HEIGHTS
E Vista St
4 MINUTES
[redacted] Ave, Long Beach, CA 90814, USA
5.0★ | POOL | 1.9X

Qualities of Interest?

MACHINE LEARNING PIPELINE

Typical ML Pipeline

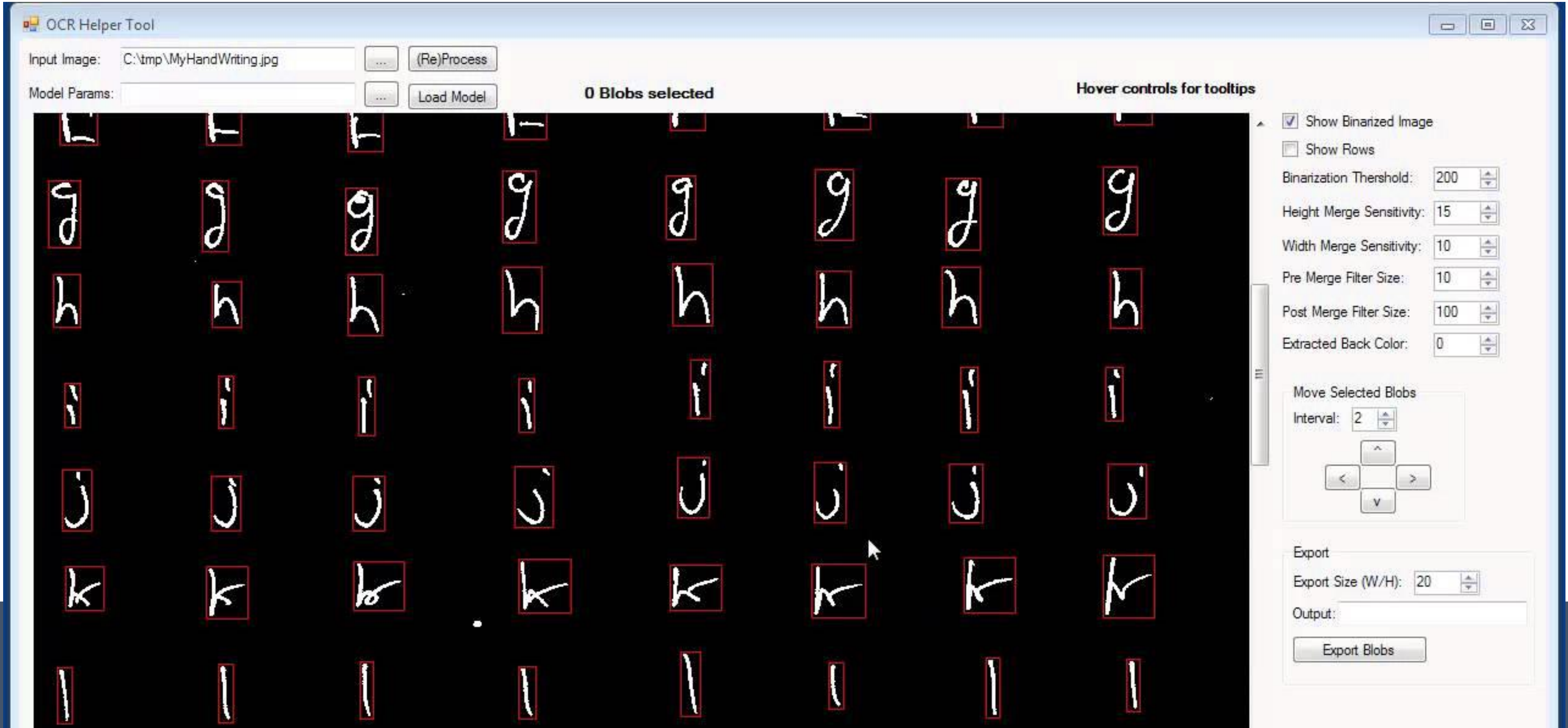


- Static
  - Get labeled data (data collection, cleaning and, labeling)
  - Identify and extract features (feature engineering)
  - Split data into training and evaluation set
  - Learn model from training data (model training)
  - Evaluate model on evaluation data (model evaluation)
  - Repeat, revising features
- with production data
  - Evaluate model on production data; monitor (model monitoring)
  - Select production data for retraining (model training + evaluation)
  - Update model regularly (model deployment)

Example Data

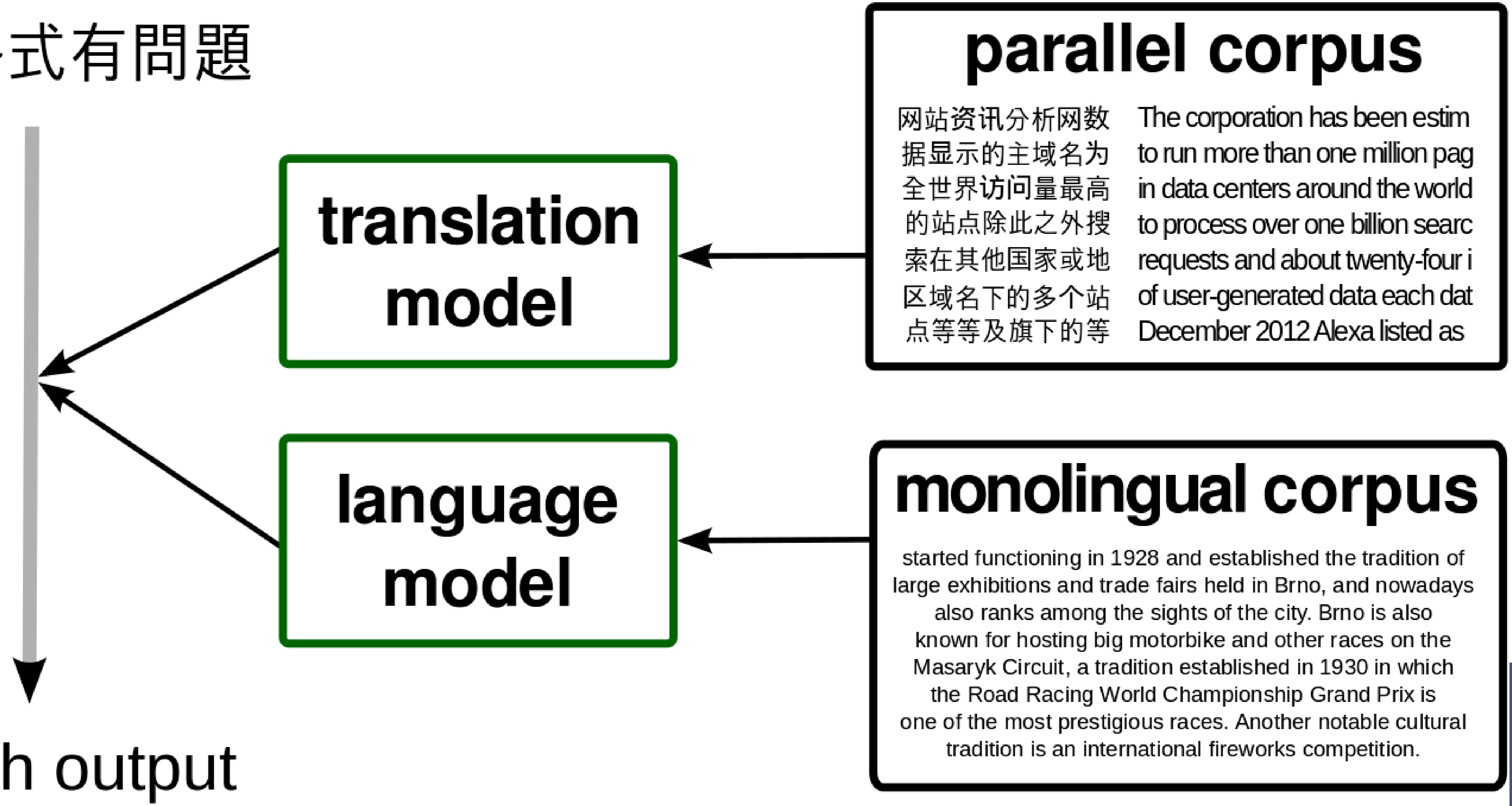
UserId	PickupLocation	TargetLocation	OrderTime	PickupTime
5	….	…	18:23	18:31
…				

Example Data



Learning Data

似乎格式有問題

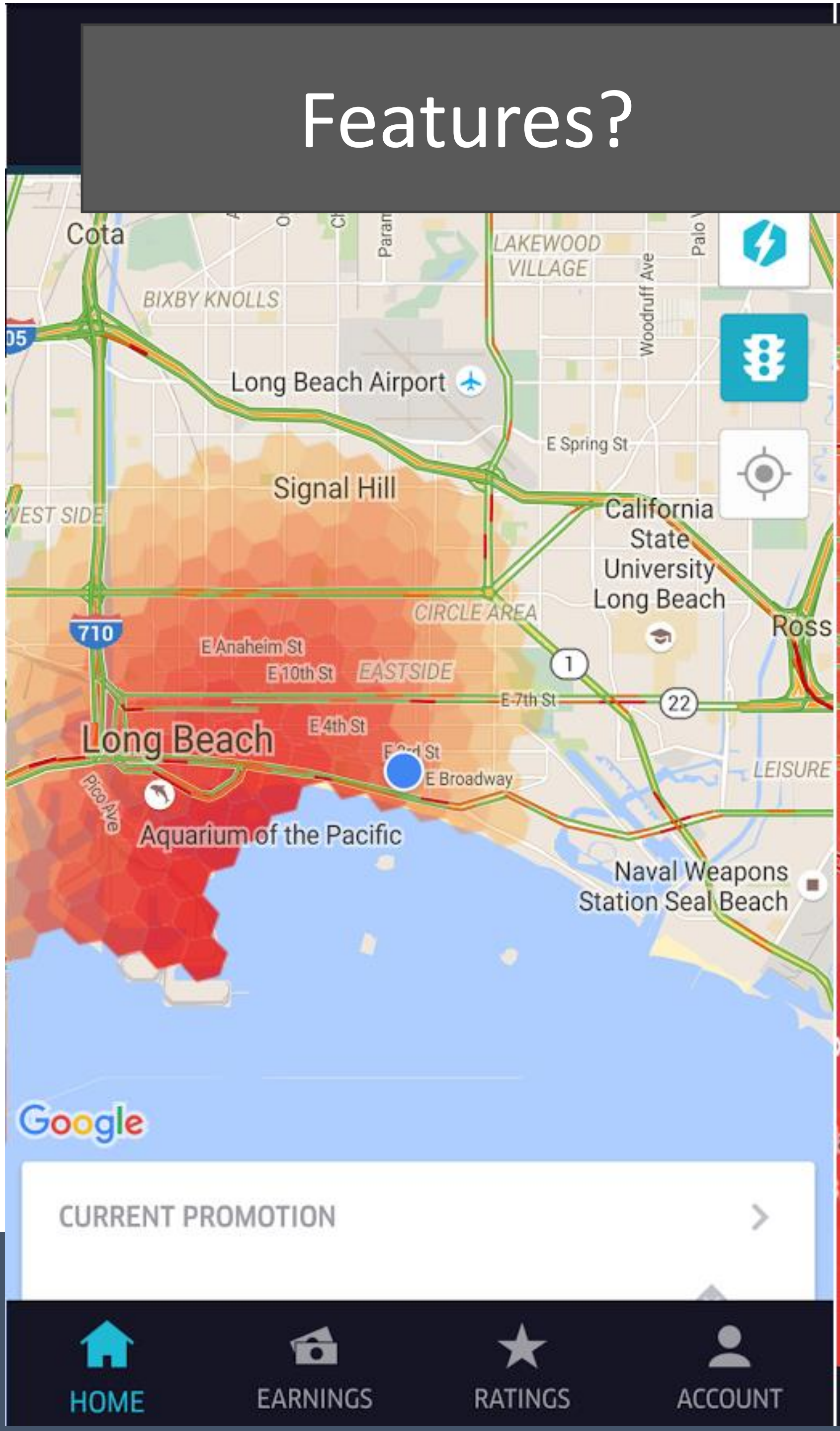


Feature Engineering

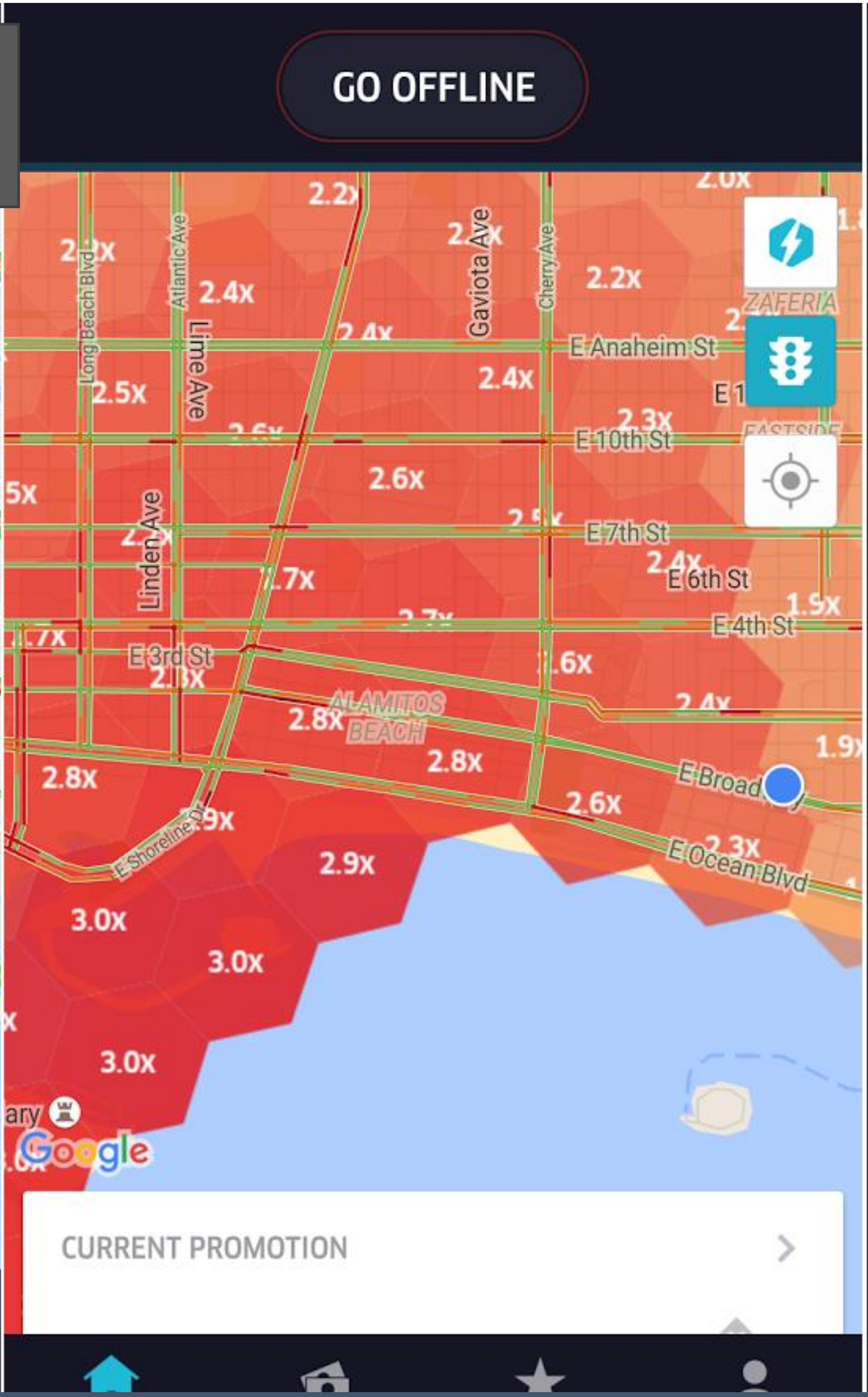
- Identify parameters of interest that a model may learn on
- Convert data into a useful form
- Normalize data
- Include context
- Remove misleading things
- In OCR/translation:

Features?

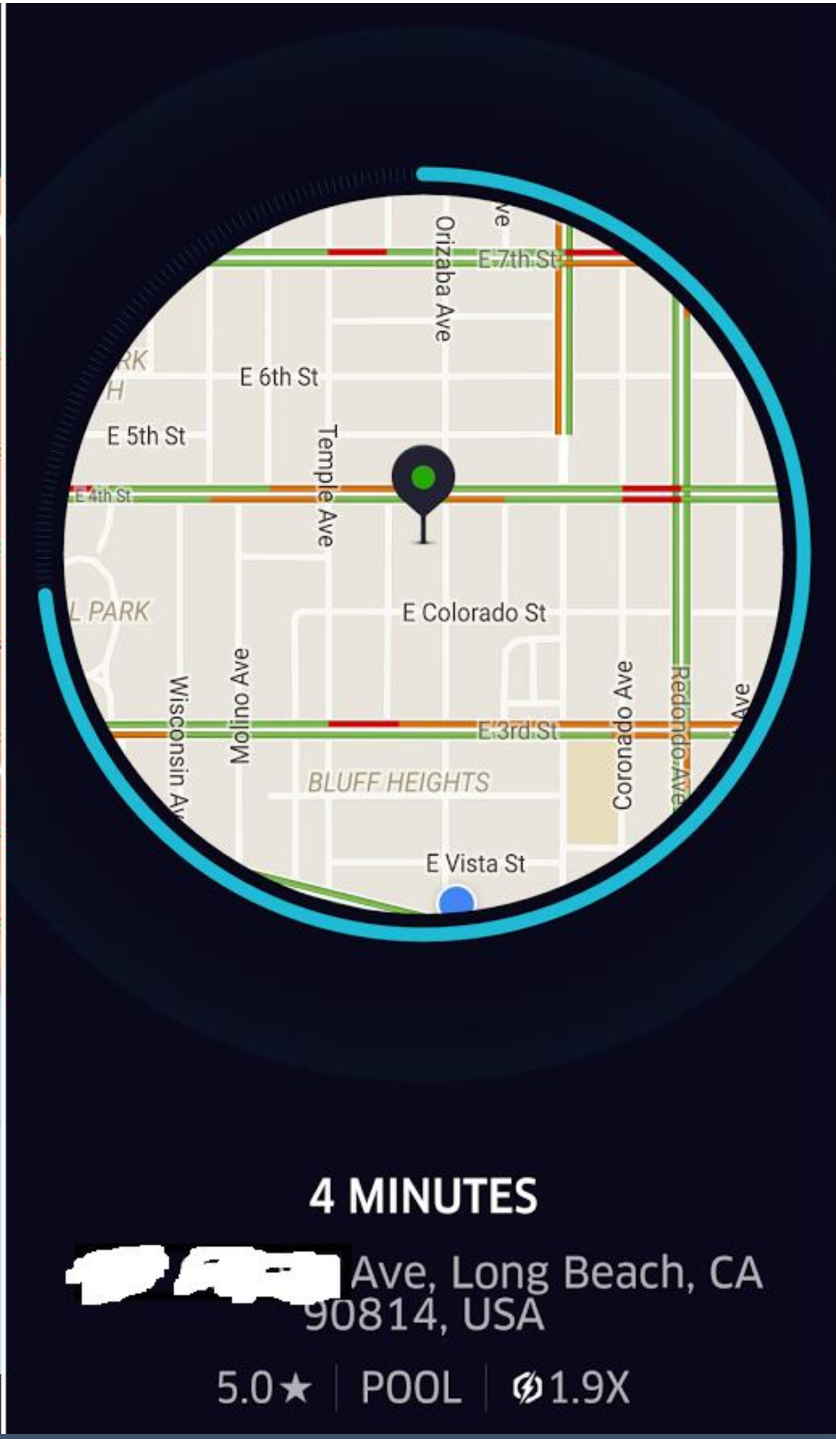
Features?



Cota
LAKEWOOD VILLAGE
BIXBY KNOLLS
Long Beach Airport
Signal Hill
E Spring St
California State University Long Beach
CIRCLE AREA
Ross
E Anaheim St
E 10th St
EASTSIDE
E 7th St
E 4th St
Long Beach
E Broadway
LEISURE
Aquarium of the Pacific
Naval Weapons Station Seal Beach
Google
CURRENT PROMOTION
HOME
EARNINGS
RATINGS
ACCOUNT



GO OFFLINE
2.2x
2.4x
2.5x
2.6x
2.7x
2.8x
2.9x
3.0x
2.3x
1.9x
E Anaheim St
E 10th St
E 7th St
E 6th St
E 4th St
E 3rd St
ALAMITOS BEACH
E Ocean Blvd
Linden Ave
Lime Ave
Gaviota Ave
Google
CURRENT PROMOTION



E 7th St
E 6th St
E 5th St
Orizaba Ave
Temple Ave
E Colorado St
Molino Ave
Wisconsin Ave
Coronado Ave
Redondo Ave
E 3rd St
BLUFF HEIGHTS
E Vista St
4 MINUTES
Ave, Long Beach, CA 90814, USA
5.0★
POOL
1.9X

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

Data Cleaning

- Removing outliers
- Normalizing data
- Missing values
- …

Learning

- Build a predictor that best describes an outcome for the observed features

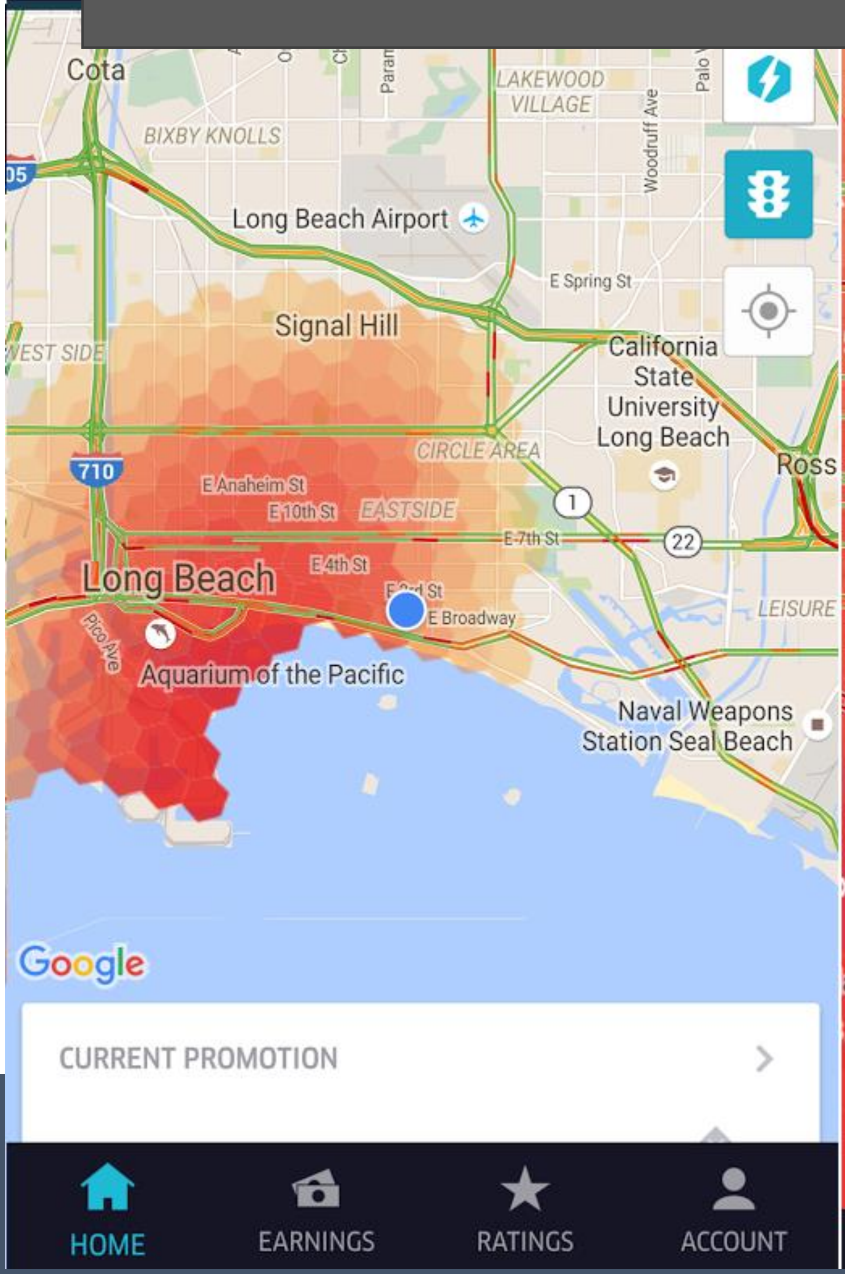
Evaluation

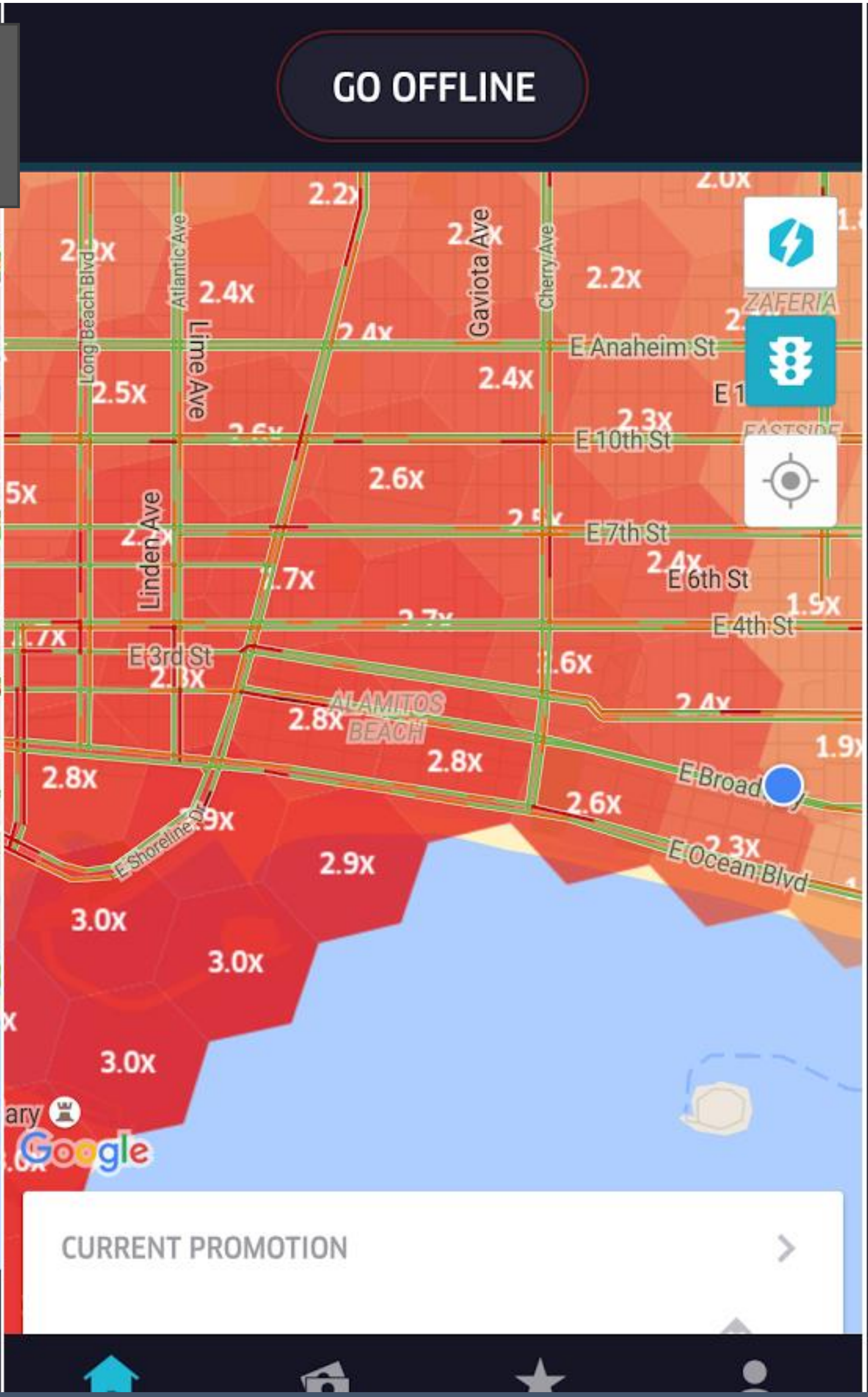
- Prediction accuracy on learned data vs
- Prediction accuracy on unseen data
  - Separate learning set, not used for training

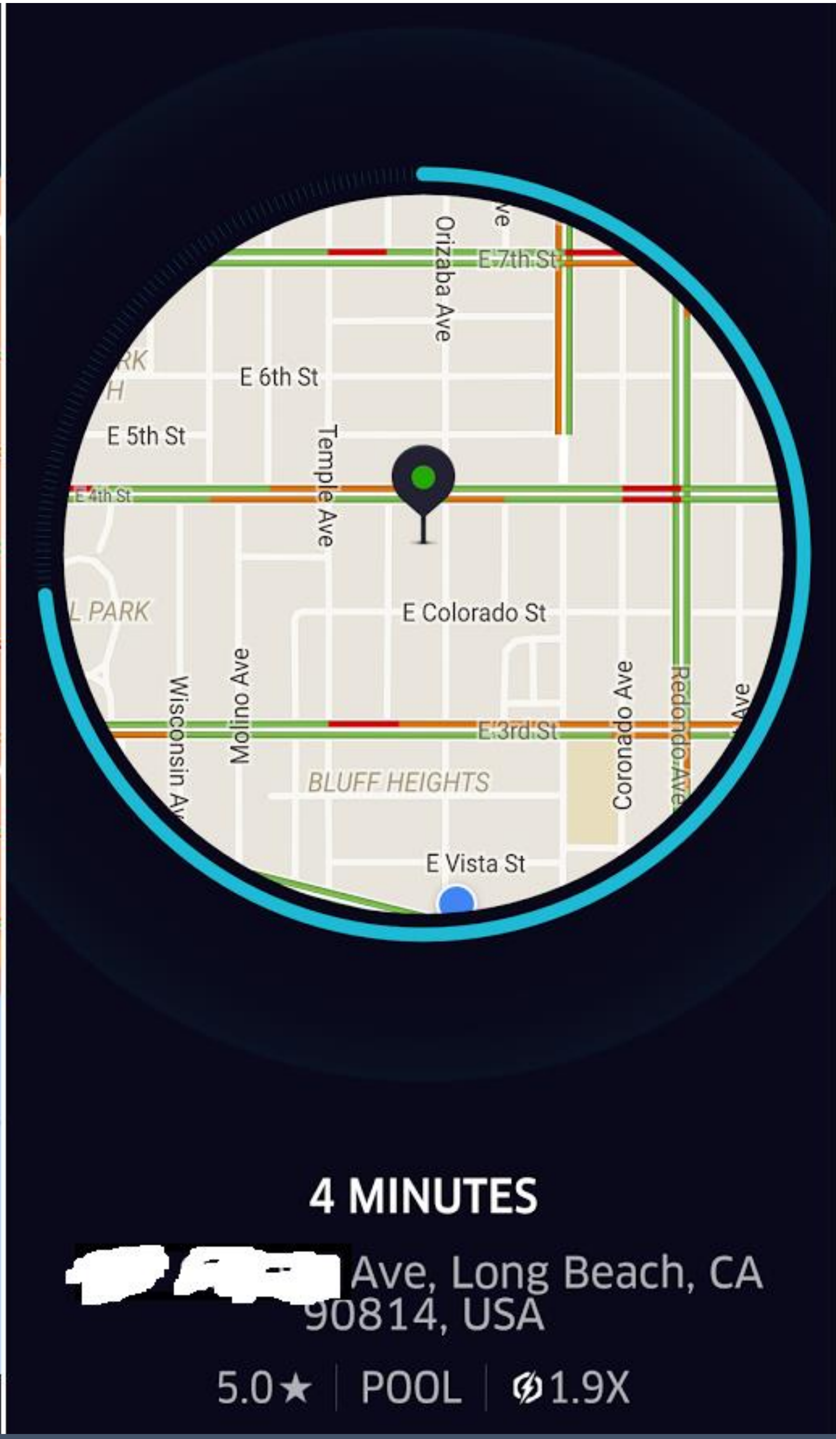
- For binary predictors: false positives vs. false negatives, precision vs. recall
- For numeric predictors: average (relative) distance between real and predicted value
- For ranking predictors: topK etc

Evaluation Data?

Evaluation Data?







Learning and Evaluating in Production

- Beyond static data sets, **build telemetry**
- Design challenge: identify mistakes in practice

- Use sample of live data for evaluation
- Retrain models with sampled live data regularly
- Monitor performance and intervene

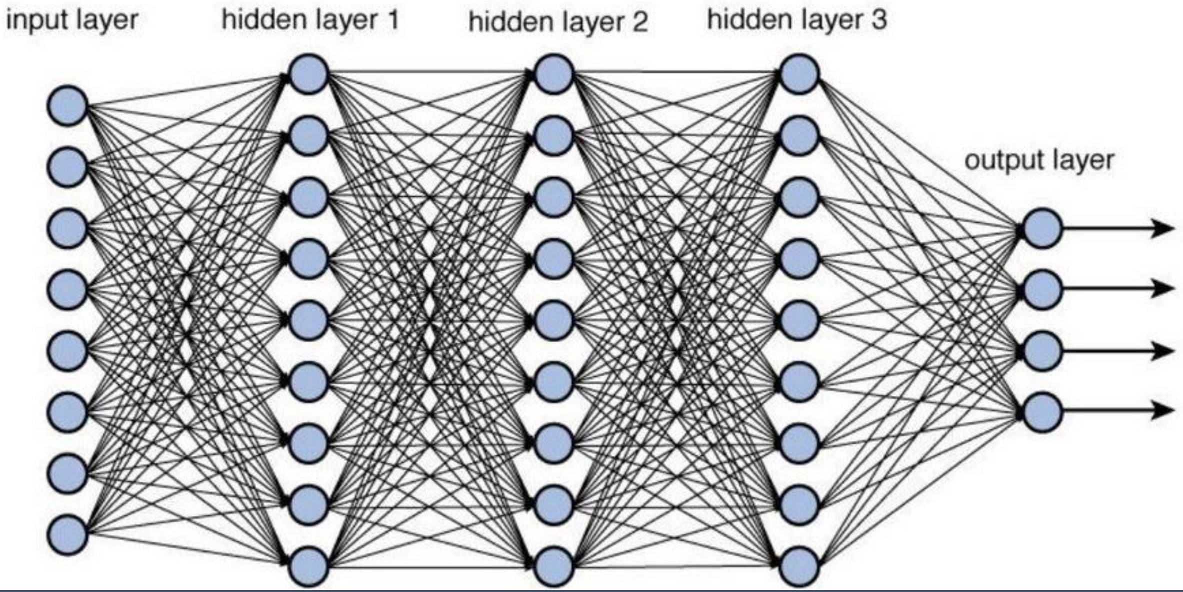
ML COMPONENT TRADEOFFS

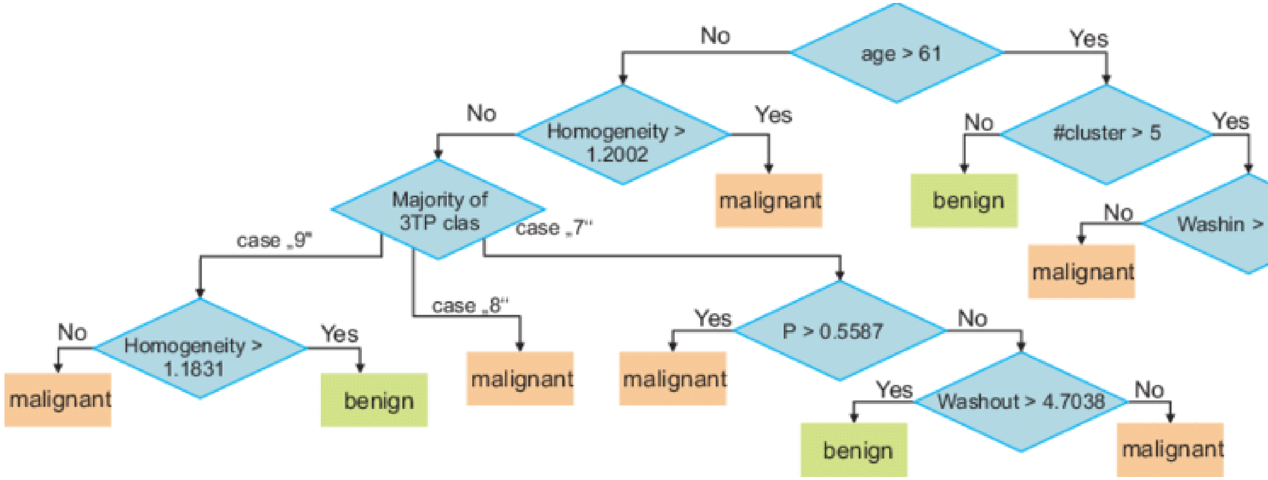
Qualities of ML Components

- 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





SYSTEM ARCHITECTURE CONSIDERATIONS

Where should the model live?

Glasses

Phone

Cloud

OCR Component

Translation Component

Where should the model live?

Car

Phone

Cloud

Surge Prediction

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?

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

Typical Designs

- Server-centric intelligence
  - latency in model execution (remote calls)
  - easy to update and experiment
  - operation cost
  - no offline operation
- Back-end cached intelligence
  - precomputed common results
  - fast execution, partial offline
  - saves bandwidth, complicated updates
- Hybrid models

Other Considerations

- Coupling of ML pipeline parts
- Coupling with other parts of the system
- Ability for different developers and analysists to collaborate
- Support online experiments
- Ability to monitor

Reactive System Design Goals

- Responsive
  - consistent, high performance
- Resilient
  - maintain responsive in the face of failure, recovery, rollback
- Elastic
  - scale with varying loads

Common Design Strategies

- Message-driven, lazy computation, functional programming
  - asynchronous, message passing style
- Replication, containment, supervision
  - replicate and coordinate isolated components, e.g. with containers
- Data streams, “infinite data”, immutable facts
  - streaming technologies, data lakes

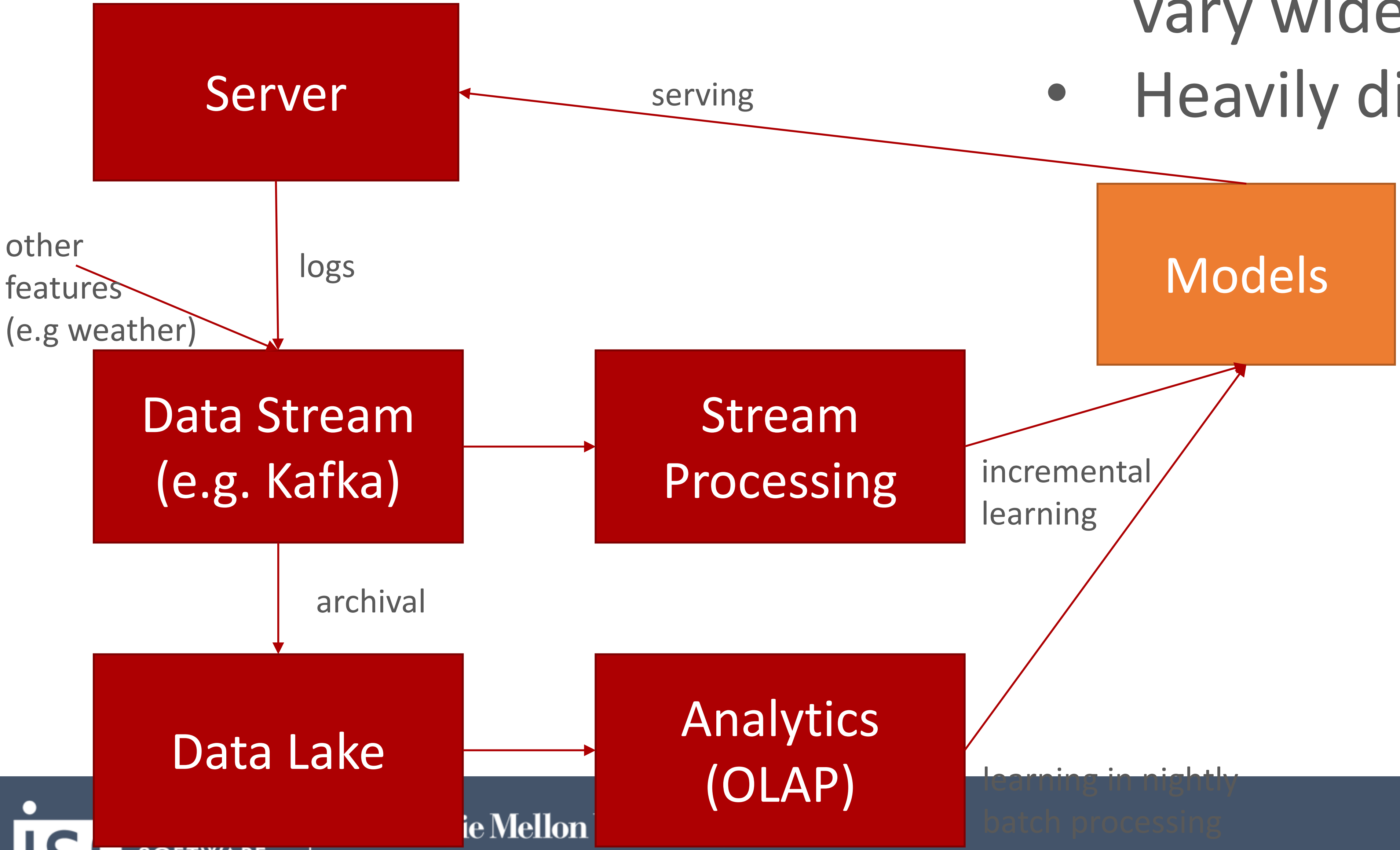
- See “big data systems” and “cloud computing”

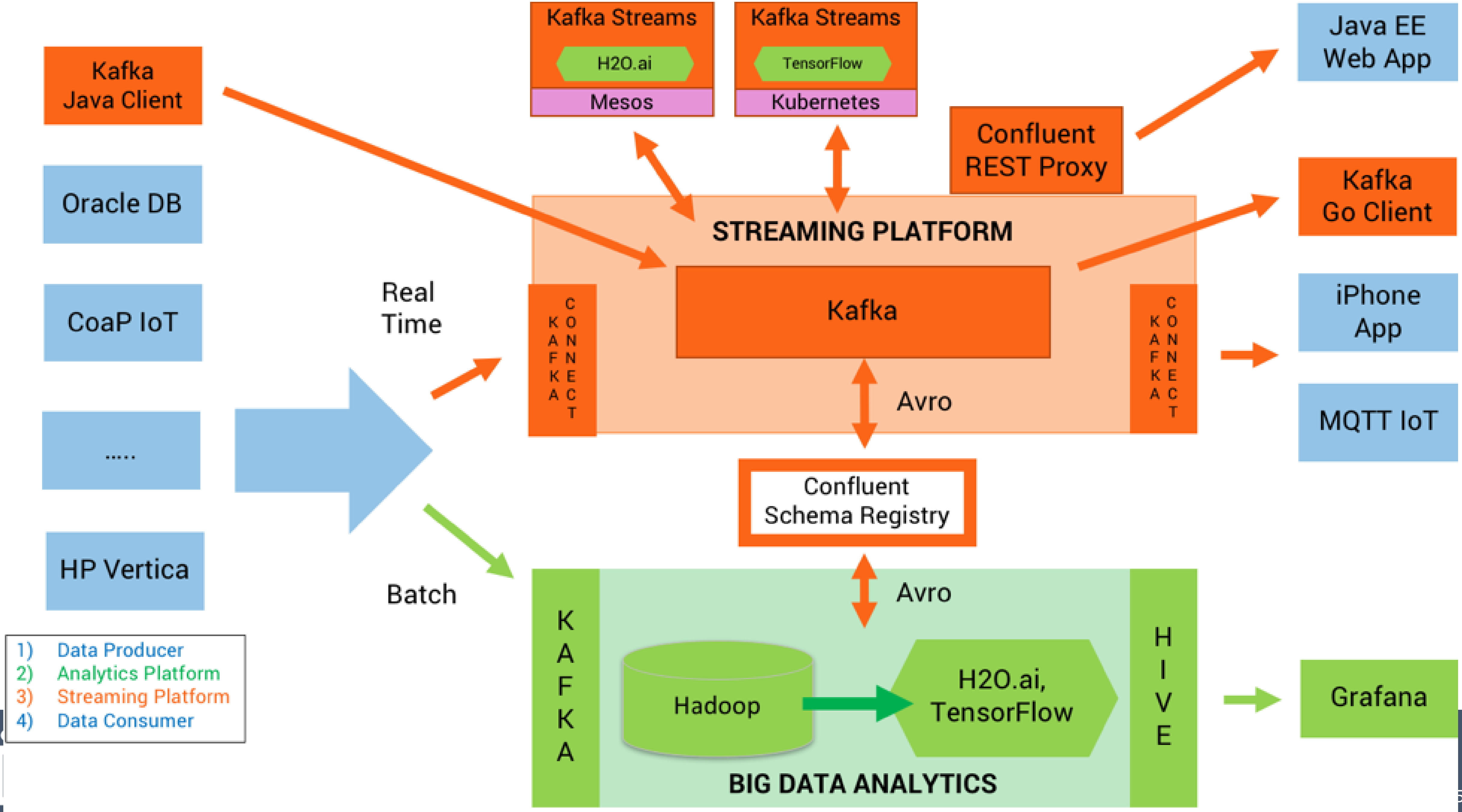
UPDATING MODELS

Updating Models

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

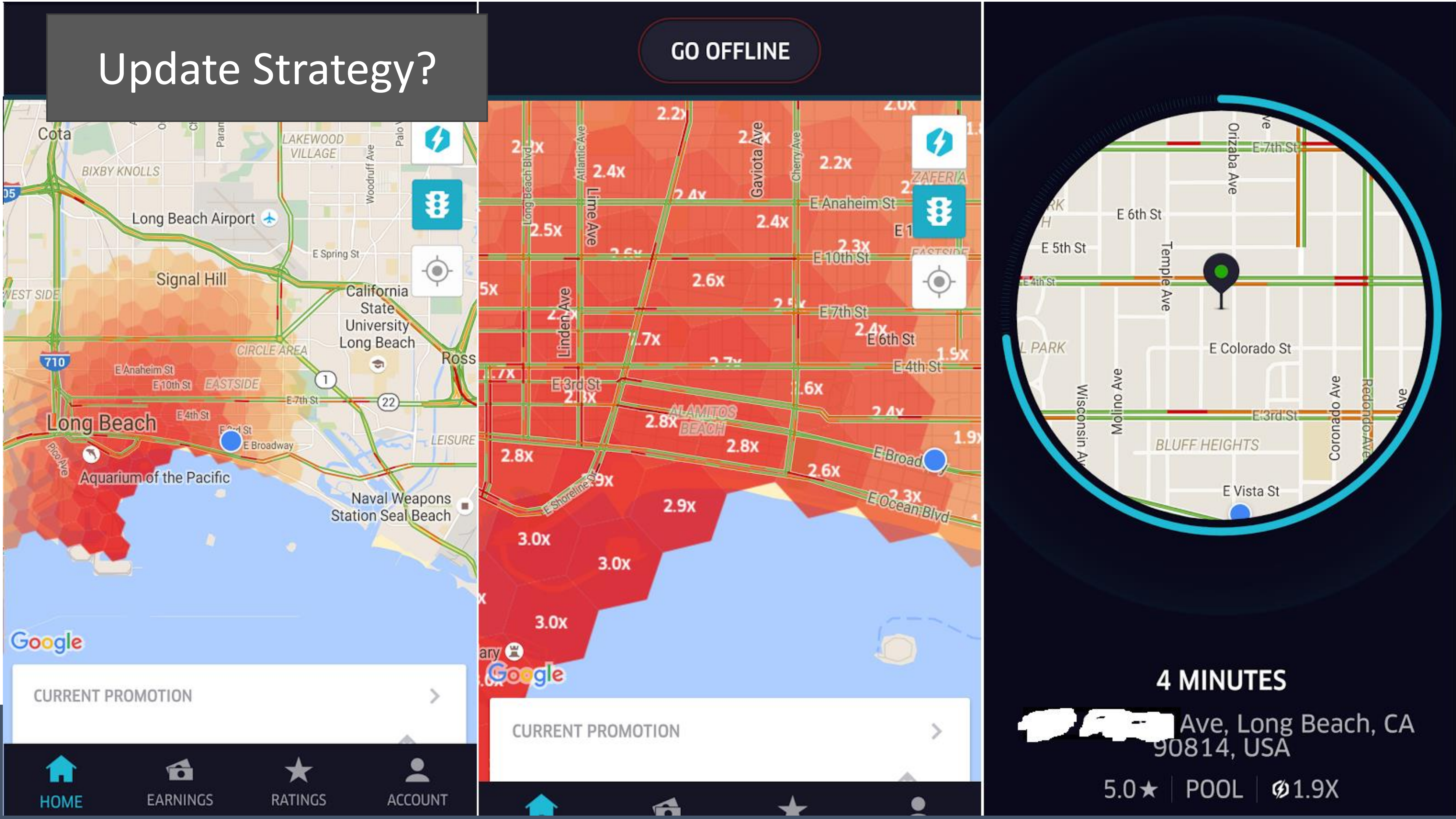
- Latency and automation vary widely
- Heavily distributed




Kafka Java Client
Oracle DB
CoaP IoT
.....
HP Vertica
Real Time
Batch
Kafka Streams
H2O.ai
Mesos
Kafka Streams
TensorFlow
Kubernetes
Confluent REST Proxy
STREAMING PLATFORM
KAFKA CONNECT
Kafka
Avro
KAFKA CONNECT
Confluent Schema Registry
Avro
KAFKA
Hadoop
H2O.ai, TensorFlow
HIVE
BIG DATA ANALYTICS
Java EE Web App
Kafka Go Client
iPhone App
MQTT IoT
Grafana
1) Data Producer
2) Analytics Platform
3) Streaming Platform
4) Data Consumer

Update Strategy?

Update Strategy?



PLANNING FOR MISTAKES

Mistakes will happen

- No specification
- ML components detect patterns from data (real and spurious)
- Predictions are often accurate, but mistakes always possible
- Mistakes are not predicable or explainable or similar to human mistakes
- Plan for mistakes
- Telemetry to learn about mistakes?

How Models can Break

- System outage
- Model outage
  - model tested? deployment and updates reliable? file corrupt?
- Model errors
- Model degradation
  - data drift, feedback loops

Hazard Analysis

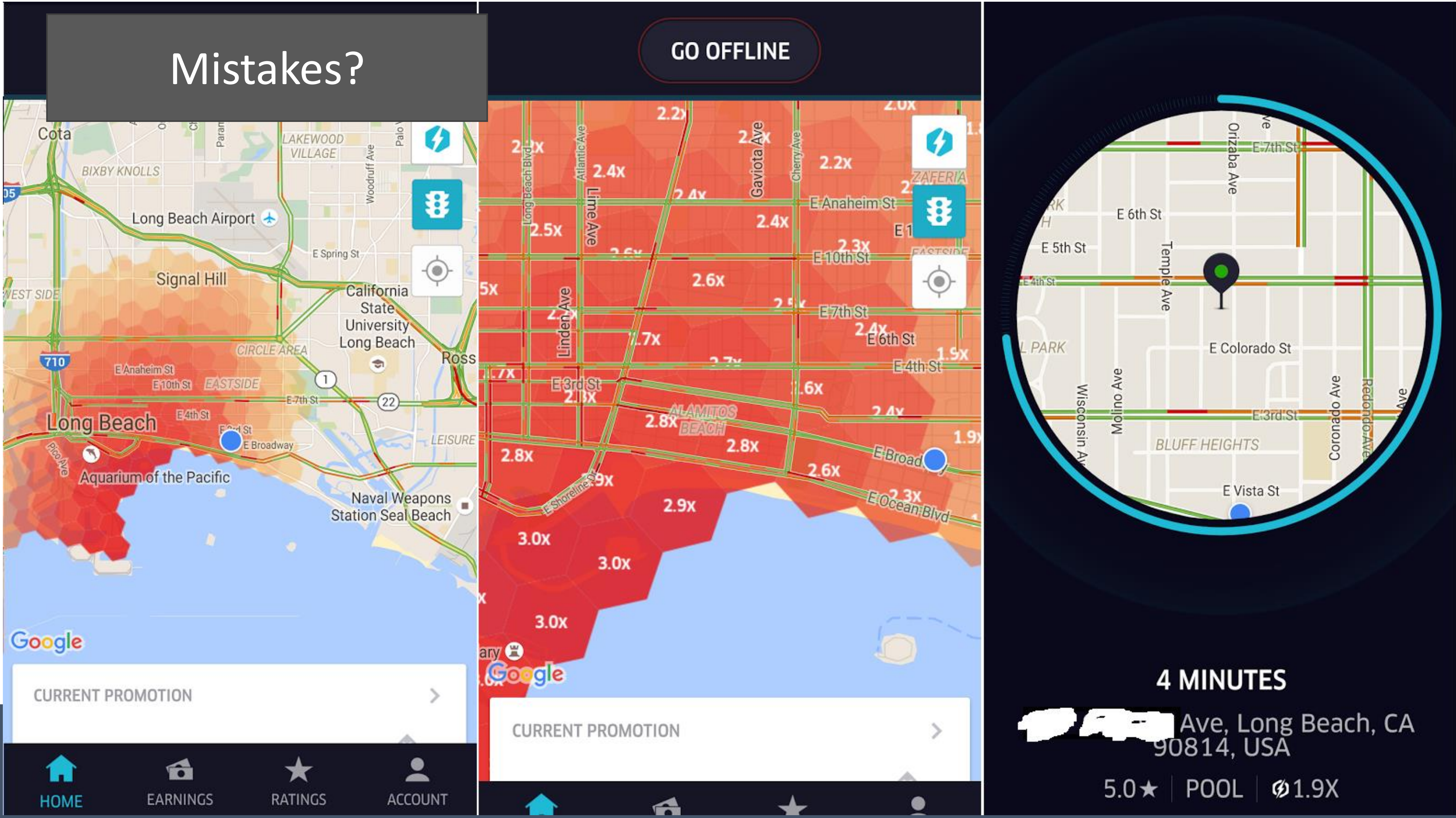
- Worst thing that can happen?
- Backup strategy? Undoable? Nontechnical compensation?

Mitigating Mistakes

- Investigating in ML
  - e.g., more training data, better data, better features, better engineers
- Less forceful experience
  - e.g., prompt rather than automate decisions, turn off
- Adjust learning parameters
  - e.g., more frequent updates, manual adjustments
- Guardrails
  - e.g., heuristics and constraints on outputs
- Override errors
  - e.g., hardcode specific results

Mistakes?

Mistakes?



Cota
LAKEWOOD VILLAGE
BIXBY KNOLLS
Long Beach Airport
Signal Hill
California State University Long Beach
CIRCLE AREA
E Anaheim St
E 10th St
EASTSIDE
E 7th St
E 4th St
Long Beach
E Broadway
LEISURE
Aquarium of the Pacific
Naval Weapons Station Seal Beach
Ross
Google
CURRENT PROMOTION
HOME
EARNINGS
RATINGS
ACCOUNT
GO OFFLINE
2.2x
2.4x
2.5x
2.6x
2.2x
2.4x
2.3x
2.7x
2.8x
2.9x
3.0x
3.0x
3.0x
1.9x
2.6x
2.3x
2.8x
E Anaheim St
E 10th St
E 7th St
E 6th St
E 4th St
E 3rd St
ALAMITOS BEACH
E Ocean Blvd
Linden Ave
Lime Ave
Gaviota Ave
Google
CURRENT PROMOTION
E 7th St
Orizaba Ave
E 6th St
E 5th St
Temple Ave
E Colorado St
E 3rd St
BLUFF HEIGHTS
Molino Ave
Coronado Ave
E Vista St
4 MINUTES
[redacted] Ave, Long Beach, CA 90814, USA
5.0★
POOL
1.9X

Telemetry

- Purpose:
  - monitor operation
  - monitor success (accuracy)
  - improve models over time (e.g., detect new features)
- Challenges:
  - too much data – sample, summarization, adjustable
  - hard to measure – intended outcome not observable? proxies?
  - rare events – important but hard to capture
  - cost – significant investment must show benefit
  - privacy – abstracting data

Summary

- Machine learning in production systems is challenging
- Many tradeoffs in selecting ML components and in integrating them in larger system
- Plan for updates
- Manage mistakes, plan for telemetry