Software Engineering for ML Applications (I)

17-313 Fall 2024

Foundations of Software Engineering

https://cmu-17313q.github.io

Eduardo Feo Flushing

These “AI start-ups” are getting out of hand

Cell: 8897603002
8106765920
ChaiGPT
Chai GPT (Genuinely Pure Tea)
Enhanced with AI (Adrak & ilaichi)
@tendies_wsb
UB
UB

Outline

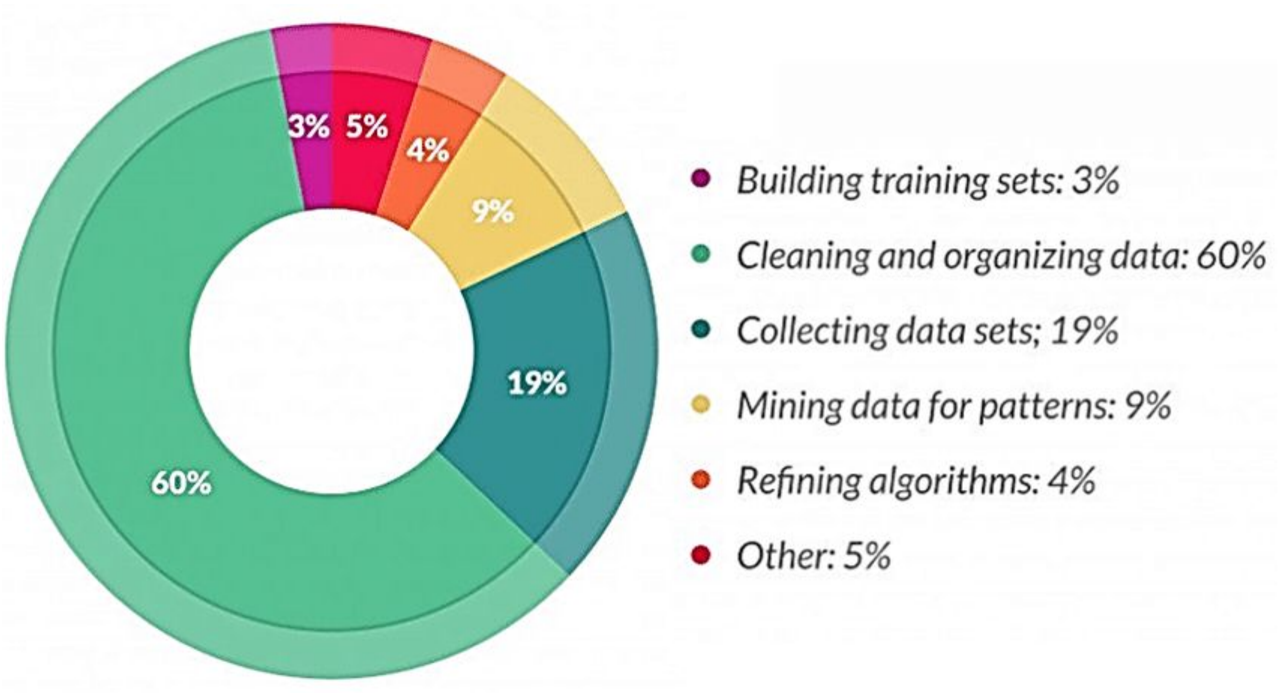
- Why ML/AI projects fail?
  - Data quality
  - Fairness issues
- What’s wrong with the model-centric pipeline?
- Are there any new challenges?
- What is ML Ops?

Why ML/AI projects fail?

“***Data***” science

When you find out Machine Learning really means endless data cleaning

"Data cleaning and repairing account for about 60% of the work of data scientists."


3%
5%
4%
9%
19%
60%
● Building training sets: 3%
● Cleaning and organizing data: 60%
● Collecting data sets; 19%
● Mining data for patterns: 9%
● Refining algorithms: 4%
● Other: 5%

What makes good quality data?

- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - The entries are recorded once.
- Consistency
  - Format, units, data agrees with itself
- Timeliness
  - The data is kept up to date.

Data is noisy

- Multiple sources
- Unreliable sensors or data entry
- Wrong results and computations, crashes
- Duplicate data, near-duplicate data
- Out of order data
- Data format invalid

Data quality and ML

- More data -> better models (up to a point)
- Noisy data (imprecise) -> less confident models
  - Some ML techniques are more or less robust to noise
- Inaccurate data: misleading models, biased models
- Need the "right" data
- Invest in data quality, not just quantity

Dirty data: Example

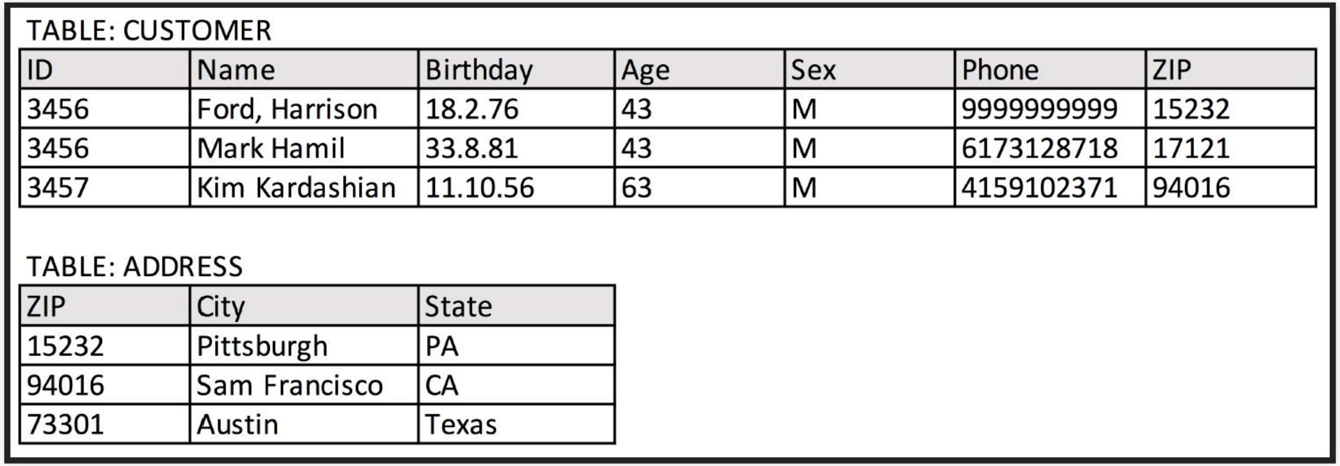


TABLE: CUSTOMER

ID	Name	Birthday	Age	Sex	Phone	ZIP
3456	Ford, Harrison	18.2.76	43	M	9999999999	15232
3456	Mark Hamil	33.8.81	43	M	6173128718	17121
3457	Kim Kardashian	11.10.56	63	M	4159102371	94016

TABLE: ADDRESS

ZIP	City	State
15232	Pittsburgh	PA
94016	Sam Francisco	CA
73301	Austin	Texas

Q. Can we (automatically) detect errors? Which errors are problem-dependent?

How do you avoid bad data?

Common strategies

- Enforce schema constraints
  - e.g., delete rows with missing data or use defaults
- Explore sources of errors (Data exploration)
  - e.g., debugging missing values, outliers
- Remove outliers
  - e.g., Testing for normal distribution, remove > 2σ
- Normalization
  - e.g., range [0, 1], power transform
- Fill in missing values

Validating the model

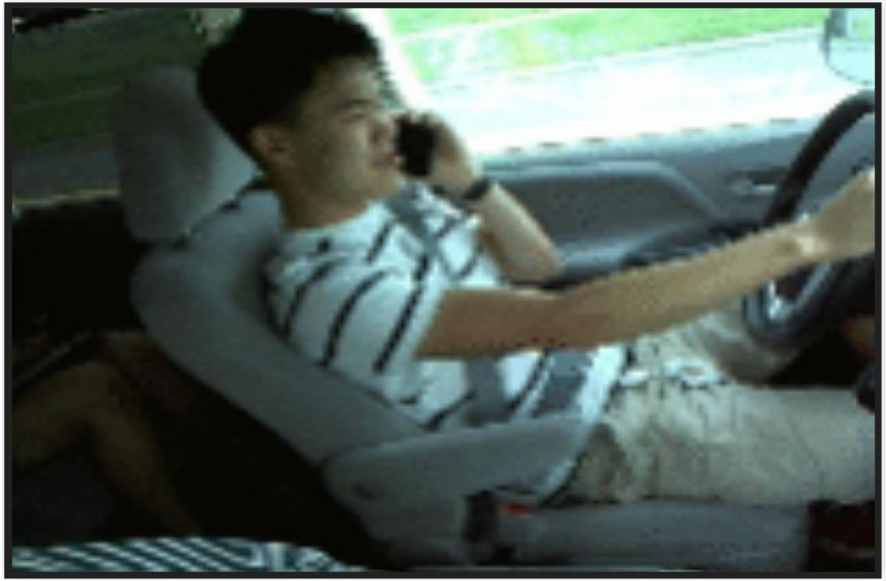
- Validation data should reflect usage data
- Be aware of data/concept drift? (face recognition during pandemic, new patterns in credit card fraud detection)

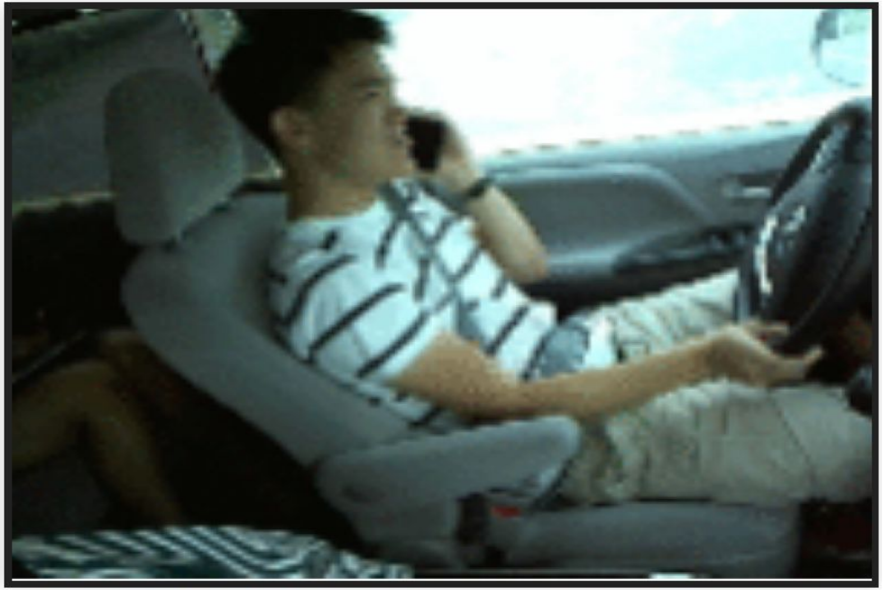
Training Data

Real Data

Independence of data

Kaggle competition on detecting distracted drivers





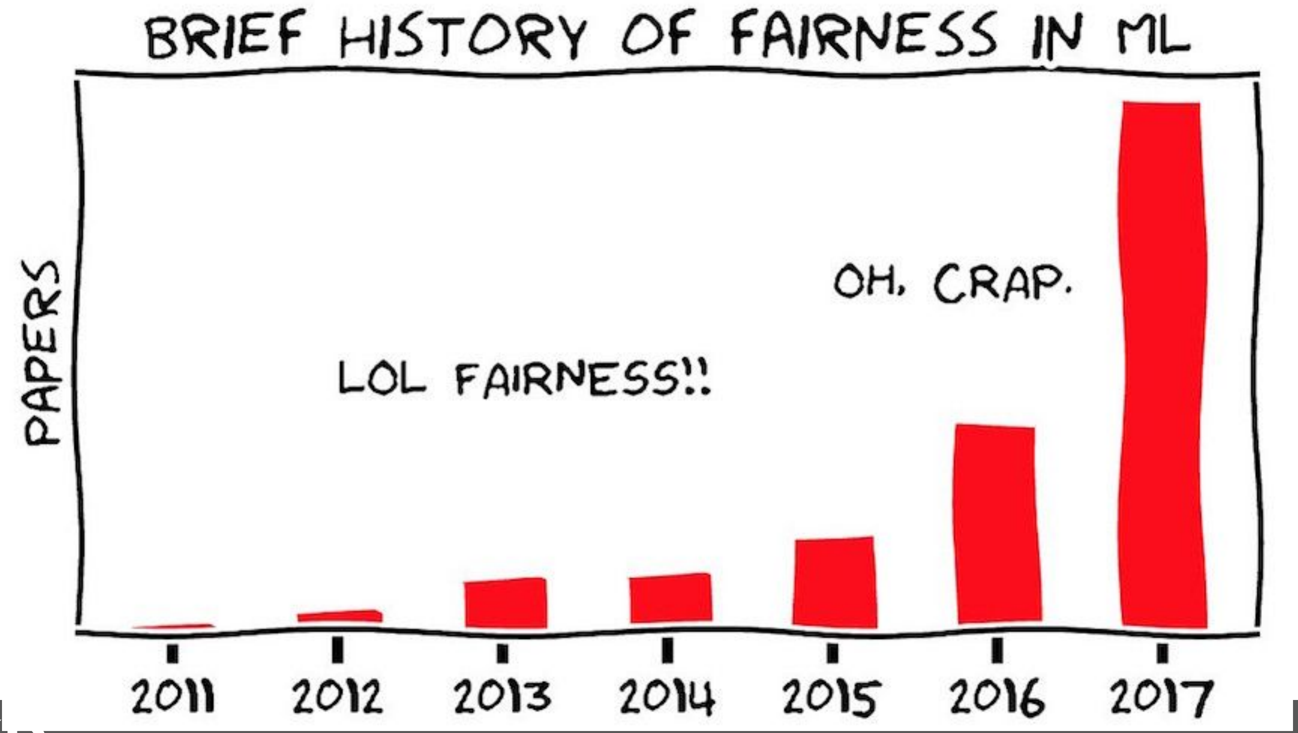
Fairness: What is fair?

Fairness discourse asks questions about how to treat people and whether treating different groups of people differently is ethical. If two groups of people are systematically treated differently, this is often considered unfair.

Philosophy: "what is fair is also what is morally right."

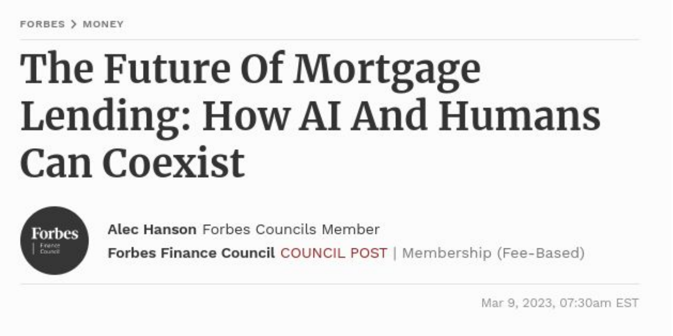
Law: "protect individuals and groups from discrimination or mistreatment with a focus on prohibiting behaviors, biases and basing decisions on certain protected factors or social group categories"

Fairness is still an actively studied & disputed concept!



Example - Mortgage Applications

- Home ownership is key path to build generational wealth
- Past decisions often discriminatory (redlining)
- Replace biased human decisions by objective and more accurate ML model
  - income, other debt, home value
  - past debt and payment behavior (credit score)
- Reduce operational costs and turn times within the mortgage process.



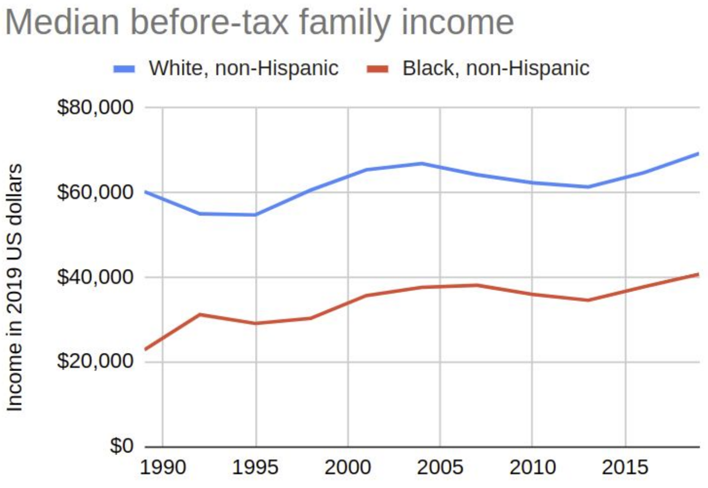
FORBES › MONEY

The Future Of Mortgage Lending: How AI And Humans Can Coexist

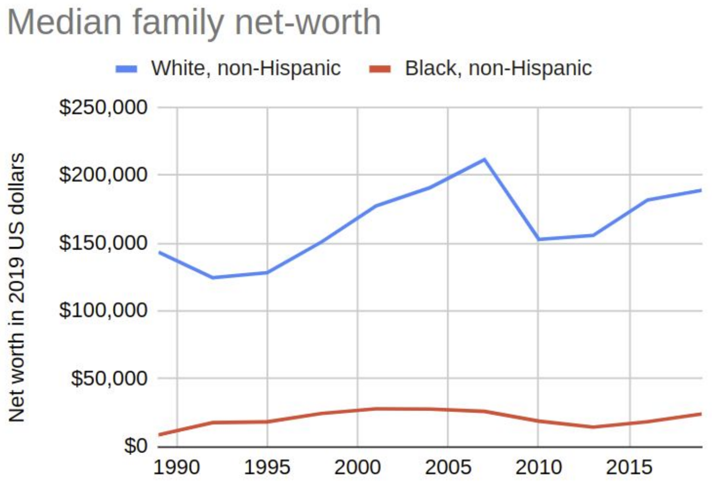
Alec Hanson Forbes Councils Member
Forbes Finance Council COUNCIL POST | Membership (Fee-Based)

Mar 9, 2023, 07:30am EST

Past bias, different starting positions



Median before-tax family income



Median family net-worth

Source: Federal Reserve’s Survey of Consumer Finances

Varieties of fairness

- Group unaware
- Demographic parity
- Equalized odds

Varieties of fairness

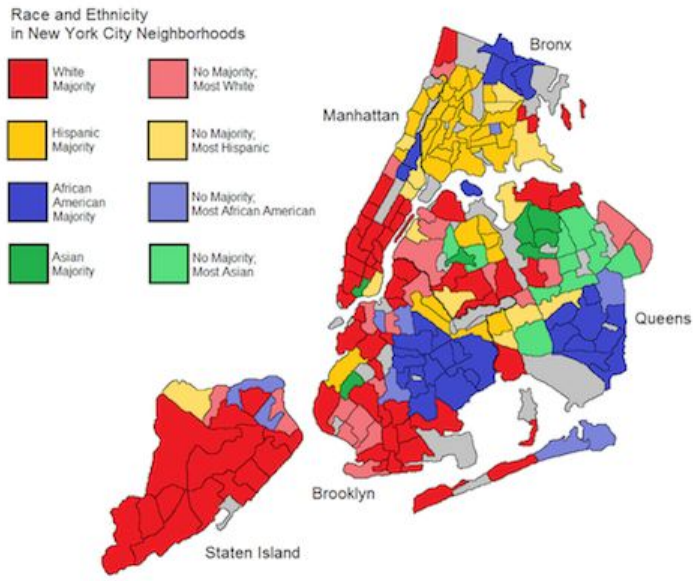
- **Group unaware (blindness)**
- Demographic parity
- Equalized Odds

Group Unaware

- Also called fairness through blindness or fairness through unawareness
- Ignore certain sensitive attributes when making a decision
- Example: Remove gender and race from mortgage model
- Easy to implement, but any limitations?

Group Unaware: Issues

- Proxies: Features correlate with protected attributes

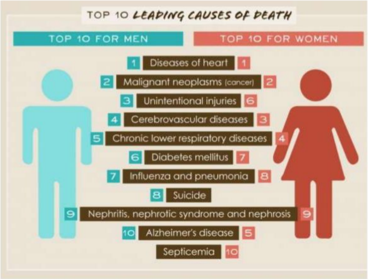


Group Unaware:
Is all discrimination is harmful?

Group Unaware:
Not all discrimination is harmful

FEDERAL TRADE COMMISSION

Mortgage discrimination is against the law.



TOP 10 LEADING CAUSES OF DEATH

TOP 10 FOR MEN		TOP 10 FOR WOMEN
1	Diseases of heart	1
2	Malignant neoplasms (cancer)	2
3	Unintentional injuries	6
4	Cerebrovascular diseases	3
5	Chronic lower respiratory diseases	4
6	Diabetes mellitus	7
7	Influenza and pneumonia	8
8	Suicide	
9	Nephritis, nephrotic syndrome and nephrosis	9
10	Alzheimer's disease	5
	Septicemia	10

- Loan lending: Gender and racial discrimination is illegal.
- Medical diagnosis: Gender/race-specific diagnosis may be desirable.
- Discrimination is a domain-specific concept!

Ensuring Group Unawareness

- How to train models that are fair wrt. group unawareness?
  - Simply remove features for protected attributes from training and inference data
  - If you can’t edit the model
    - Null/randomize protected attribute during inference
- How to test if models are fair wrt. group unawareness?
  - $\forall x. f(x[p \leftarrow 0]) = f(x[p \leftarrow 1])$
  - Test with any test data, e.g., purely random data or existing test data
  - Any single inconsistency shows that the protected attribute was used. Can also report percentage of inconsistencies.

Varieties of fairness

- Group unaware (blindness)
- **Demographic parity (independence)**
- Equalized odds

Demographic parity

Key idea: Compare outcomes across two groups

- Similar rates of accepted loans across racial/gender groups?
- Similar chance of being hired/promoted between gender groups?
- Similar rates of (predicted) recidivism across racial groups?
- Outcomes matter, not accuracy!

Varieties of fairness

- Group unaware (blindness)
- Demographic parity (independence)
- **Equalized odds (separation)**

Equalized odds

Key idea: Focus on accuracy (not outcomes) across two groups

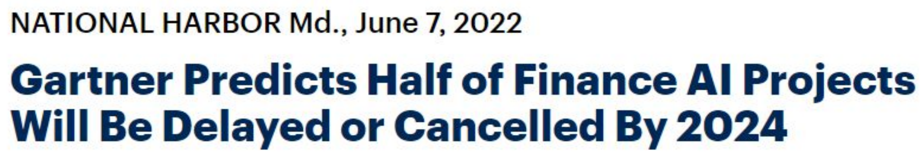
- Similar default rates on accepted loans across racial/gender groups?
- Similar rate of "bad hires" and "missed stars" between gender groups?
- Similar accuracy of predicted recidivism vs actual recidivism across racial groups?
- Accuracy matters, not outcomes!

Usually implemented by training different models

Outline

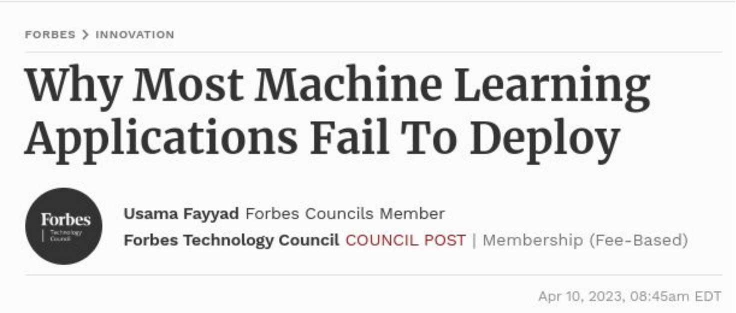
- Why ML/AI projects fail?
    - Data quality
    - Fairness issues
- **What’s wrong with the model-centric pipeline?**
- Are there any new challenges?
- What is ML Ops?

Why ML/AI projects fail? What’s wrong?



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

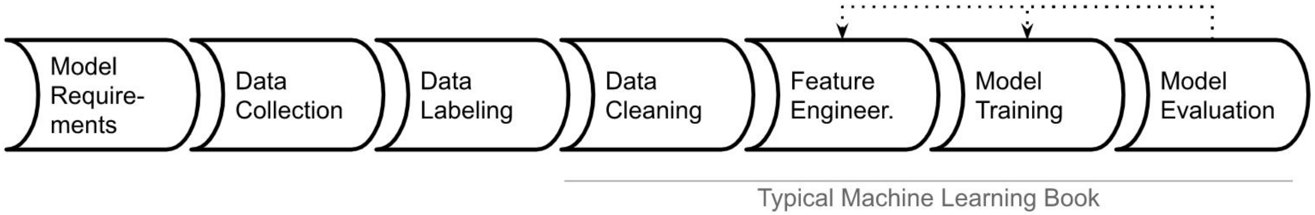
Usama Fayyad Forbes Councils Member

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)

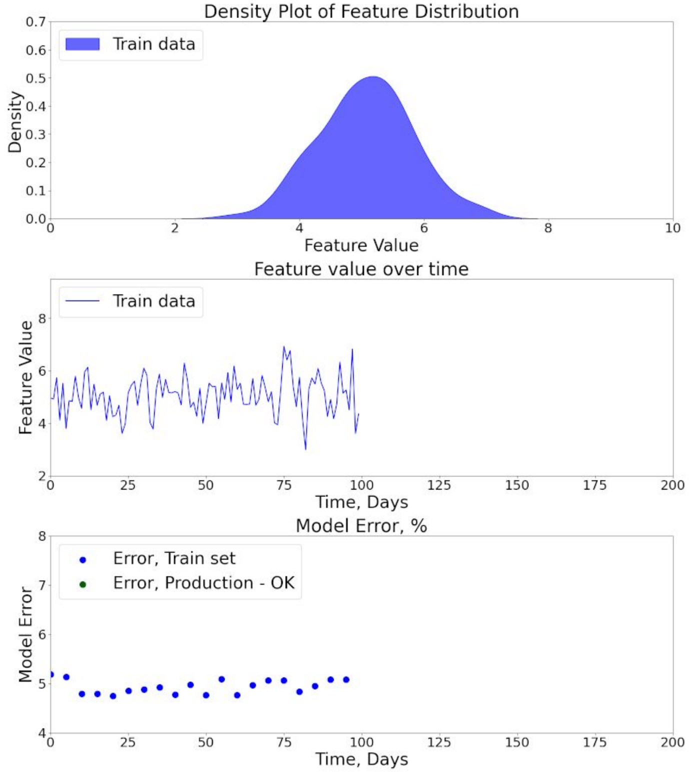


What’s wrong with the model-centric pipeline?

World is not static

- Concepts drift
  - ML estimates `f(x) = y`
  - What if the relationship between x & y changes over time?
- Data drift
  - Statistical properties of the input data change over time
  - Causes:
    - External factors (e.g., market trends, user behavior shifts).
    - Sensor recalibrations or environmental changes.
    - Changes in data collection methods or quality.
  - Impact:
    - Model performance degrades as the training data no longer accurately represents the real-world scenario.

Data drift - Monitoring



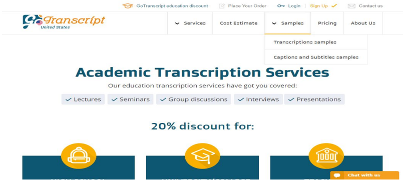
Activity

Pick one scenario based on where you are seating

- Transcription Services (front rows)
- Parking Sensor (middle rows)
- Surge Prediction (back rows)

Discuss in groups these questions:

- Identify a realistic reason why the model might start performing differently in your scenario.
- Discuss whether this change is due to **concept drift** or **data drift** and be ready to explain your reasoning.



GoTranscript education discount | Place Your Order | Login | Sign Up | Contact us

GoTranscript United States

Services | Cost Estimate | Samples | Pricing | About Us

Transcriptions samples

Captions and Subtitles samples

Academic Transcription Services

Our education transcription services have got you covered:

✓ Lectures ✓ Seminars ✓ Group discussions ✓ Interviews ✓ Presentations

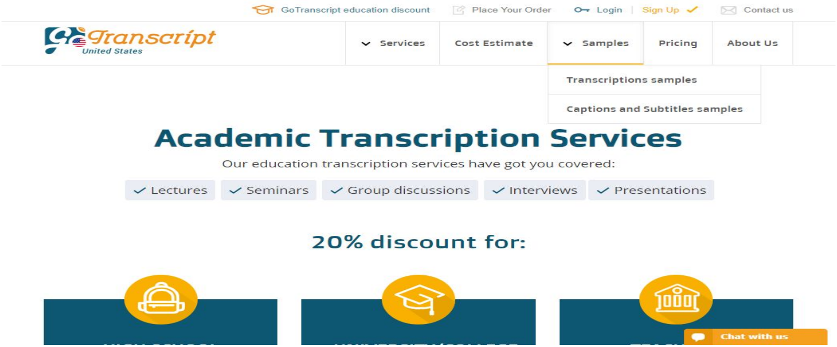
20% discount for:

Chat with us

Activity: Sample Answers

Reason for Change: New slang or terminology used in conversations.

Drift Type: **Concept Drift**, as the “definition” of typical language in transcription evolves.



Academic Transcription Services

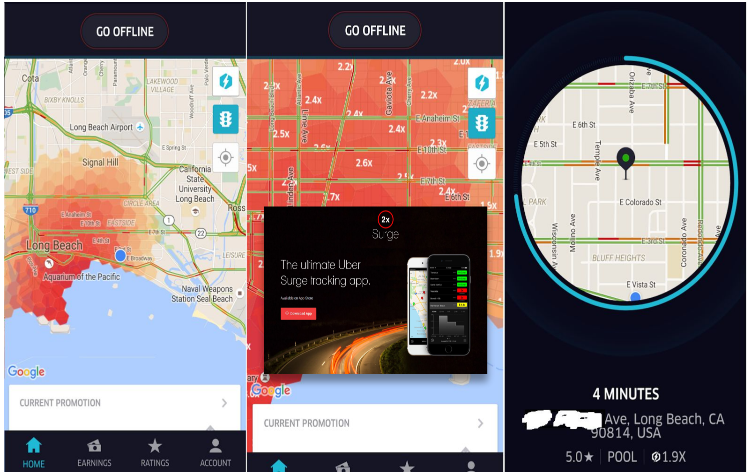
Our education transcription services have got you covered:

20% discount for:

Activity: Sample Answers

Reason for Change: New user behavior trends, like remote work reducing rush hour traffic.

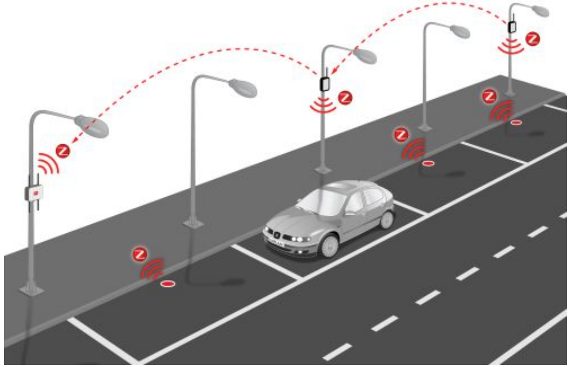
Drift Type: **Data Drift**, since the distribution of user activity data is changing without altering the fundamental concept of “surge prediction.”



Activity: Sample Answers

Reason for Change: Weather conditions, like snow, affecting sensor readings.

Drift Type: **Data Drift**, as the input data distribution (sensor readings) shifts due to environmental changes.



To be continued ...