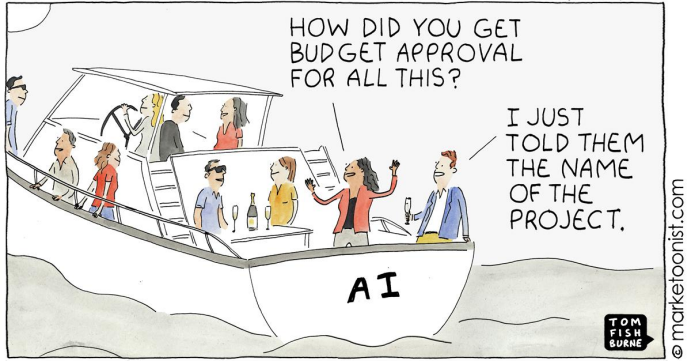


ML for Software Engineers

17-313 Fall 2024

Foundations of Software Engineering

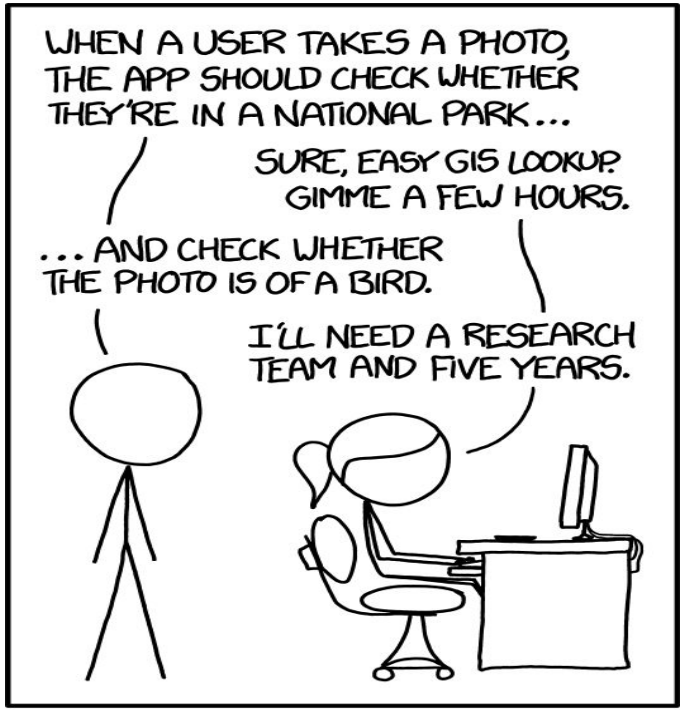
https://cmu-17313q.github.io

Eduardo Feo Flushing

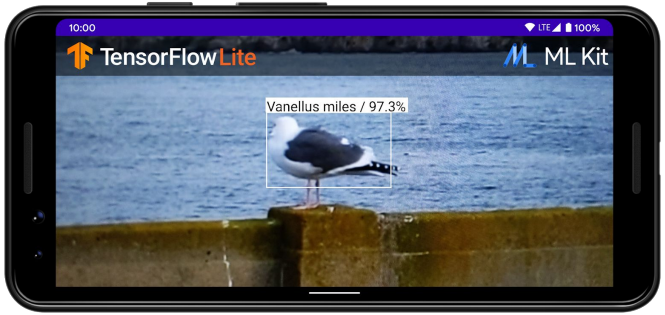
Learning goals

- Understand how machine learning (ML) components are parts of larger systems
- Explain the typical machine learning process
- Key differences from traditional software
- Distinguish between LLMs and traditional ML models in terms of flexibility, scalability, and application.
- Evaluate and improve LLM performance by understanding benchmarks, interpreting metrics like perplexity, and adjusting settings

2014



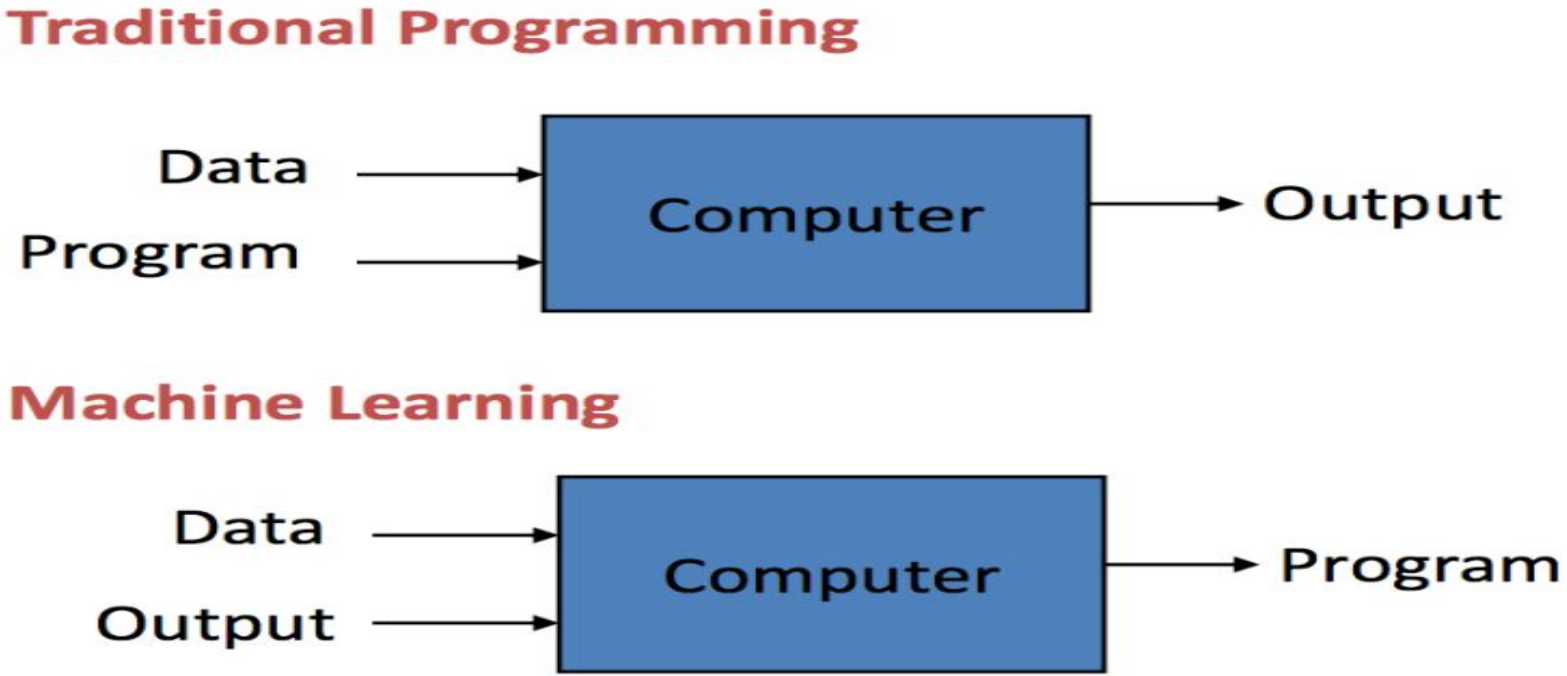
... a few years later



Outline

- Traditional Programing vs. ML
- Case Studies
- Model-Centric Pipeline: ML Basics
  - Features
  - Model Building
  - Evaluation
- LLMs
  - What's the difference between traditional ML and LLMs?
  - Performance

Traditional Programming vs ML



Traditional Programming

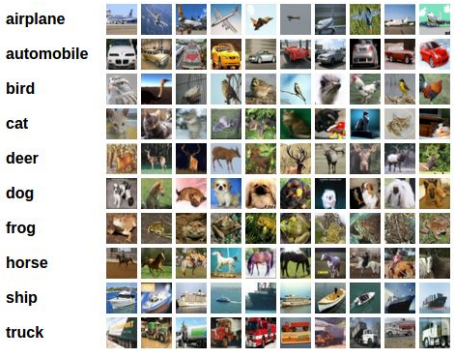
"It is easy. You just chip away the stone that doesn't look like David." –(probably not) Michelangelo

ML Development

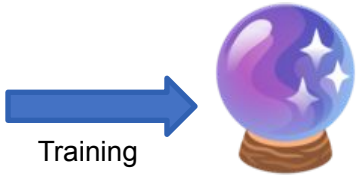
- Observation
- Hypothesis
- Predict
- Test
- Reject or Refine Hypothesis

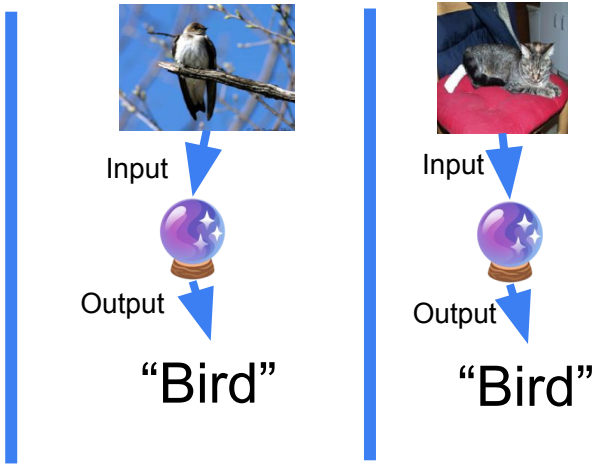
Machine Learning in One Slide

(Supervised)

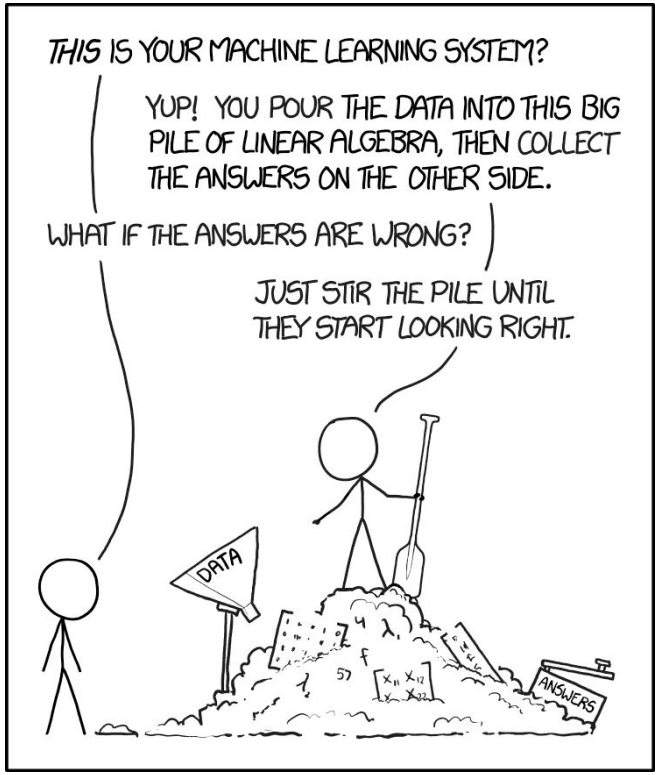


Lots of labelled data
(Inputs, outputs)





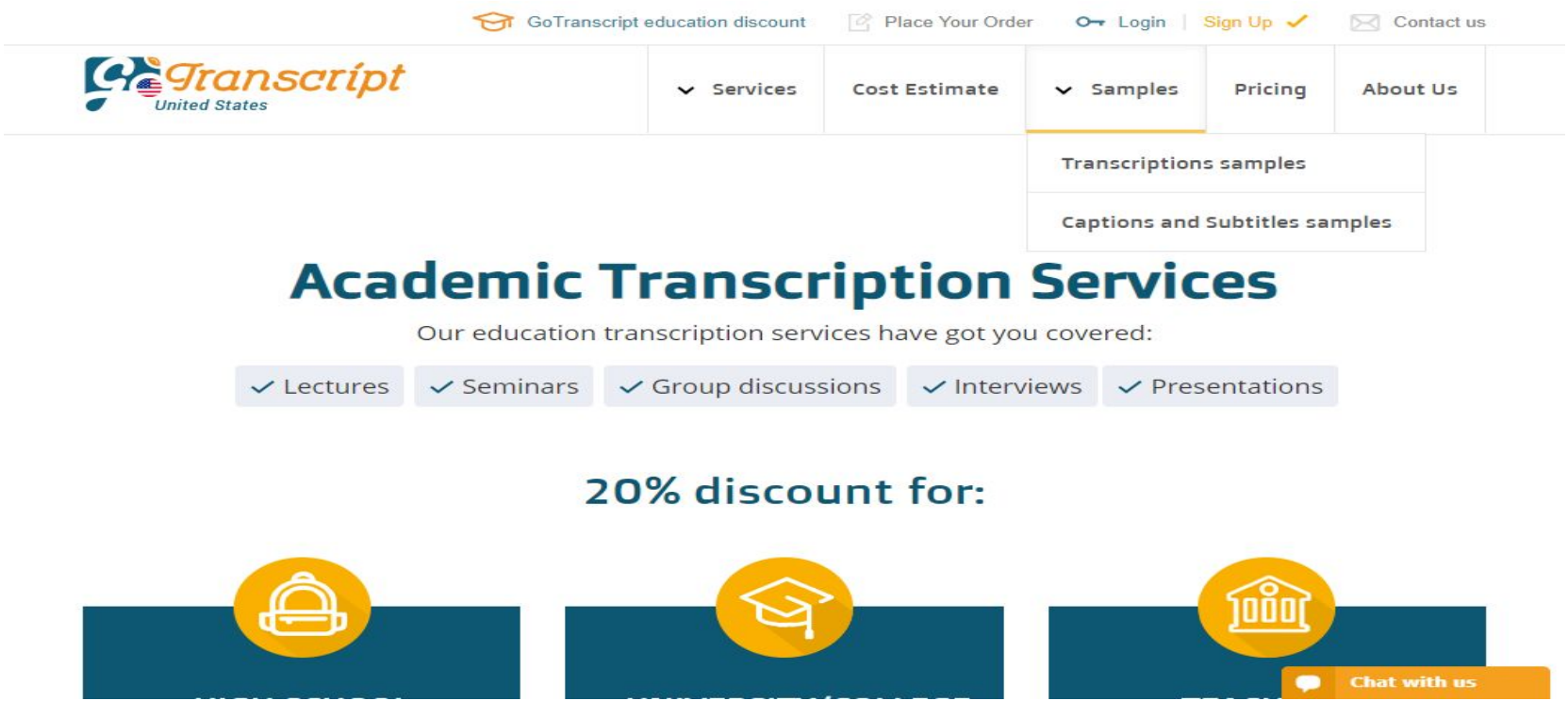
Black-box view of ML



Outline

- Traditional Programing vs. ML
- **Case Studies**
- Model-Centric Pipeline: ML Basics
  - Features
  - Model Building
  - Evaluation
- LLMs
  - What's the difference between traditional ML and LLMs?
  - Performance

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)

Case study: Smart IoT Parking Sensor

Case Study: Surge Prediction App

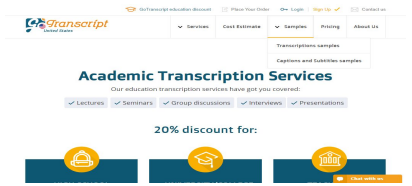
Activity

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 two challenges that the team will likely focus on when turning their research into a product
- What qualities are important for a good commercial product?



Outline

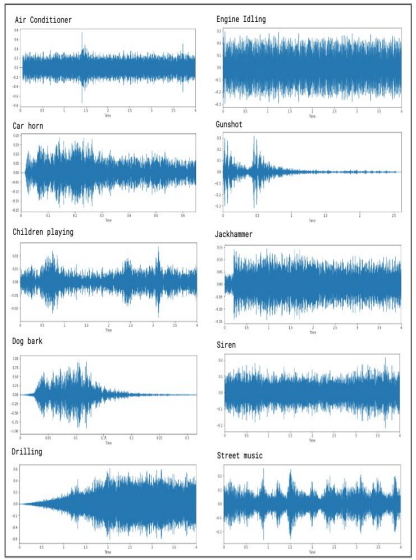
- Traditional Programing vs. ML
- Case Studies
- **Model-Centric Pipeline: ML Basics**
  - **Features**
  - **Model Building**
  - **Evaluation**
- LLMs
  - What's the difference between traditional ML and LLMs?
  - Performance

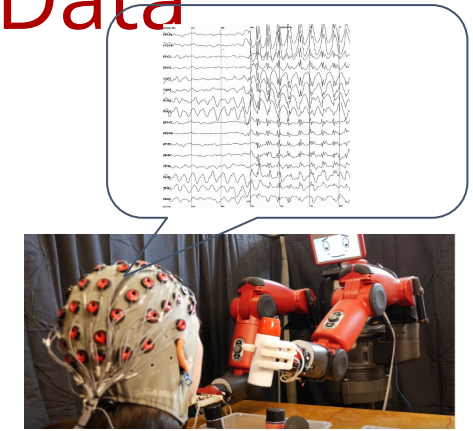
ML 101: Model-centric pipeline

- Traditional Model Focus (data science)

Typical Machine Learning Book

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

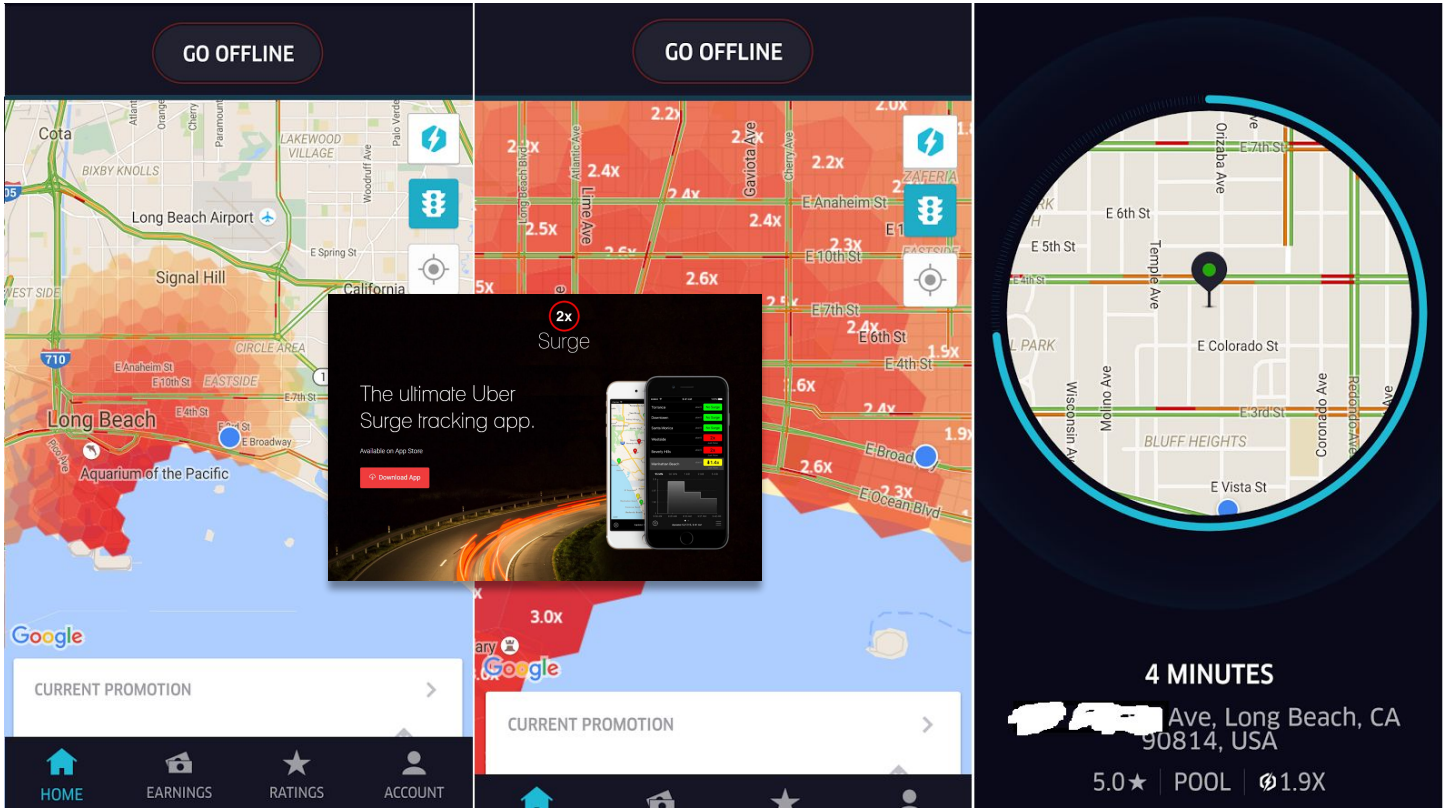
Data Cleaning

- 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

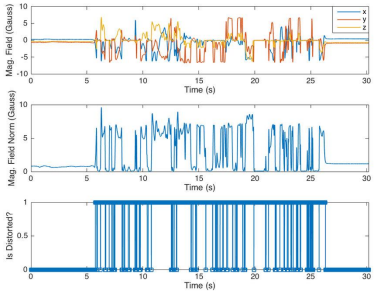
Features?

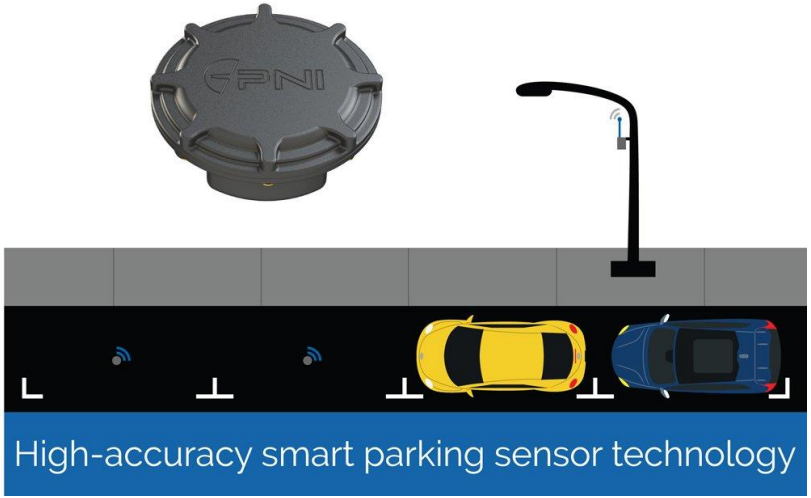

GO OFFLINE
Cota
LAKEWOOD VILLAGE
BIXBY KNOLLS
Long Beach Airport
E Spring St
Signal Hill
California
CIRCLE AREA
710
E Anaheim St
E 10th St
EASTSIDE
E 7th St
E 4th St
Long Beach
E Broadway
Aquarium of the Pacific
2x
Surge
The ultimate Uber Surge tracking app.
Available on App Store
Download App
Google
CURRENT PROMOTION
HOME
EARNINGS
RATINGS
ACCOUNT
GO OFFLINE
2.4x
2.5x
2.6x
2.2x
Atlantic Ave
Lime Ave
Gaviota Ave
Cherry Ave
E Anaheim St
E 10th St
E 7th St
E 6th St
E 4th St
E Broad
E Ocean Blvd
1.5x
1.9x
3.0x
Google
CURRENT PROMOTION
E 7th St
Orizaba Ave
E 6th St
E 5th St
Temple Ave
E Colorado St
Melino Ave
Coronado Ave
E 3rd St
BLUFF HEIGHTS
E Vista St
4 MINUTES
Ave, Long Beach, CA 90814, USA
5.0★ | POOL | 1.9X

Features?

PNI SENSOR CORPORATION

PlacePod Smart Parking Sensor





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

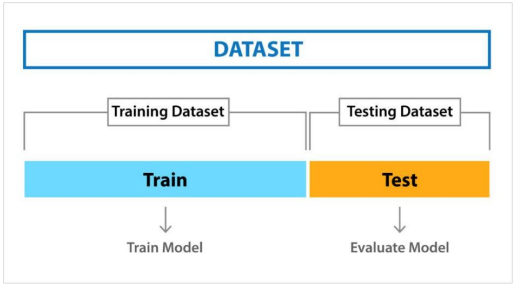
- Build a predictor that best describes an outcome for the observed features
- Many algorithms: Linear Regression, Logistic Regression, Probabilistic models, Decision Trees, Support Vector Machines, Gaussian Processes, Neural Networks

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

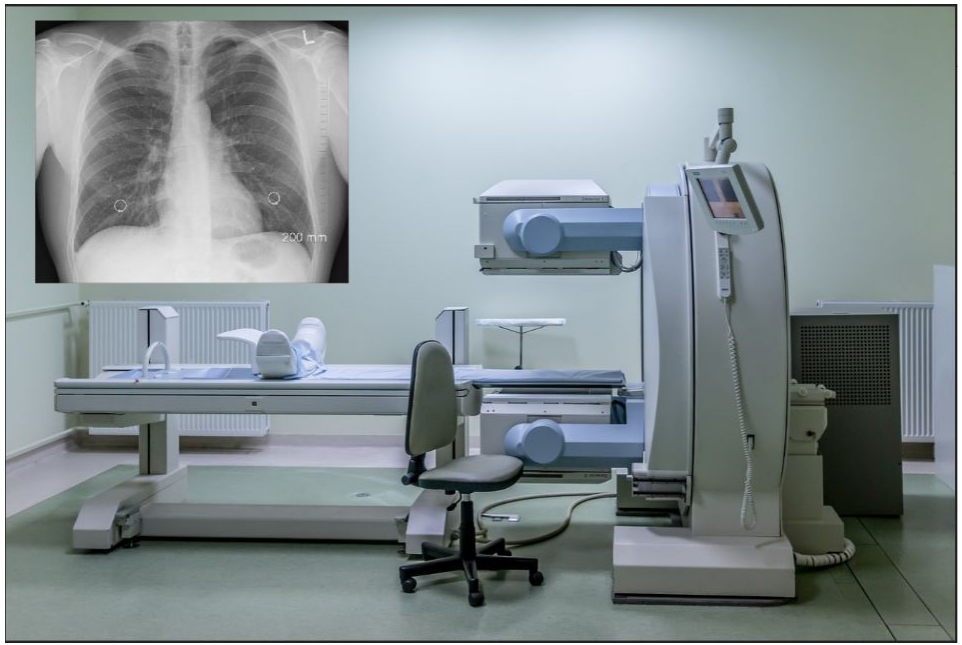
- Binary classification: Positive / Negative
- Possible classification outcomes:
  - TN: True Negatives
  - TP: True Positives
  - FN: False Negatives
  - FP: False Positives

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).

Evaluation: is model accuracy enough?

Q. Are false positives and false negatives equally bad?



Activity: False positives and false negatives, equally bad?

Discuss in groups these scenarios:

- 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

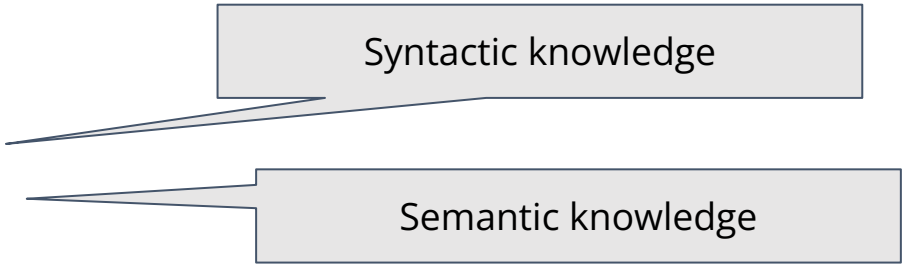
Outline

- Traditional Programing vs. ML
- Case Studies
- Model-Centric Pipeline: ML Basics
  - Features
  - Model Building
  - Evaluation
- **LLMs**
  - **What's the difference between traditional ML and LLMs?**
  - **Performance**

Large Language Models

- Language Modeling: Measure probability of a sequence of words
  - Probability distribution of a sequence of words

P(Eduardo, loves, his, cat) = 0.02
P(Eduardo, cat, loves, his) = 0.0001
P(Eduardo, hates, his, cat) = 0.00001



Large Language Models

- Language Modeling: Measure probability of a sequence of words
- LLMs are… large
  - GPT-3 has 175B parameters
  - GPT-4 is estimated to have ~1.24 Trillion
- Pre-trained with up to a PB of Internet text data
  - Massive financial and environmental cost

Large Language Models

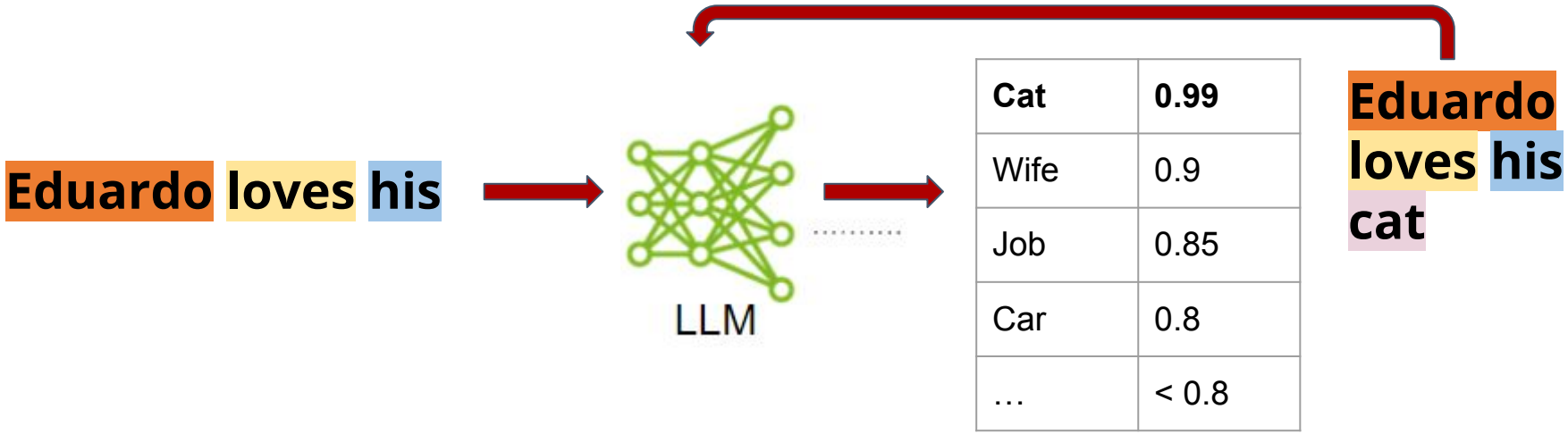
- Language Modeling: Measure probability of a sequence of words
  - Probability distribution of a sequence of words

$$\begin{aligned}\mathbf{P}(\text{Eduardo, loves, his, cat}) &= \mathbf{P}(\text{cat}\,|\,\text{Eduardo, loves, his})\,\mathbf{P}(\text{Eduardo, loves, his}) \\ &= \mathbf{P}(\text{cat}\,|\,\text{Eduardo, loves, his})\,\mathbf{P}(\text{his}\,|\,\text{Eduardo, loves})\,\mathbf{P}(\text{Eduardo, loves}) \\ &= \mathbf{P}(\text{cat}\,|\,\text{Eduardo, loves, his})\,\mathbf{P}(\text{his}\,|\,\text{Eduardo, loves})\,\mathbf{P}(\text{loves}\,|\,\text{Eduardo})\mathbf{P}(\text{Eduardo})\end{aligned}$$

  - Generative Models

Large Language Models

- Autoregressive Generative Models



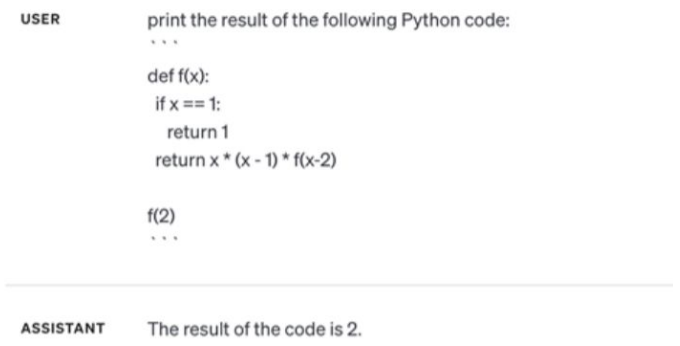
Language Models are Pre-trained

- Only a few people have resources to train LLMs
- Access through API calls
  - OpenAI, Google Vertex AI, Anthropic, Hugging Face

- We will treat it as a **black box that can make errors**!

LLMs are far from perfect

- Hallucinations
  - Factually Incorrect Output
- High Latency
  - Output words generated one at a time
  - Larger models also tend to be slower
- Output format
  - Hard to structure output (e.g. extracting date from text)
  - Some workarounds for this (later)



Traditional ML vs LLMs

Focus and Versatility

- Traditional ML Models:
  - Broadly adaptable (e.g., image classification, fraud detection)
  - Flexible but needs task-specific feature engineering
- LLMs:
  - Specialized for language tasks
  - Ideal for chatbots, text summarization, translation

Traditional ML vs LLMs

Scale and Complexity

- Traditional ML Models:
  - Range from simple to complex; millions of parameters max
  - Optimization and fine-tuning are often simpler, with a focus on hyperparameters.
- LLMs:
  - Billions of parameters; high computational demands
  - Extensive training time on vast datasets; may take days or weeks to complete.

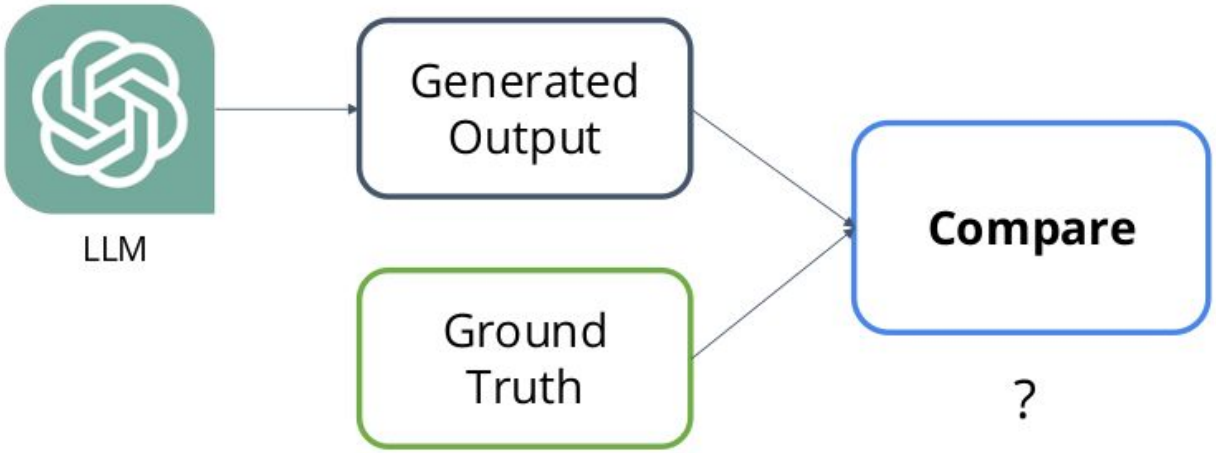
Traditional ML vs LLMs

Performance and Generalization

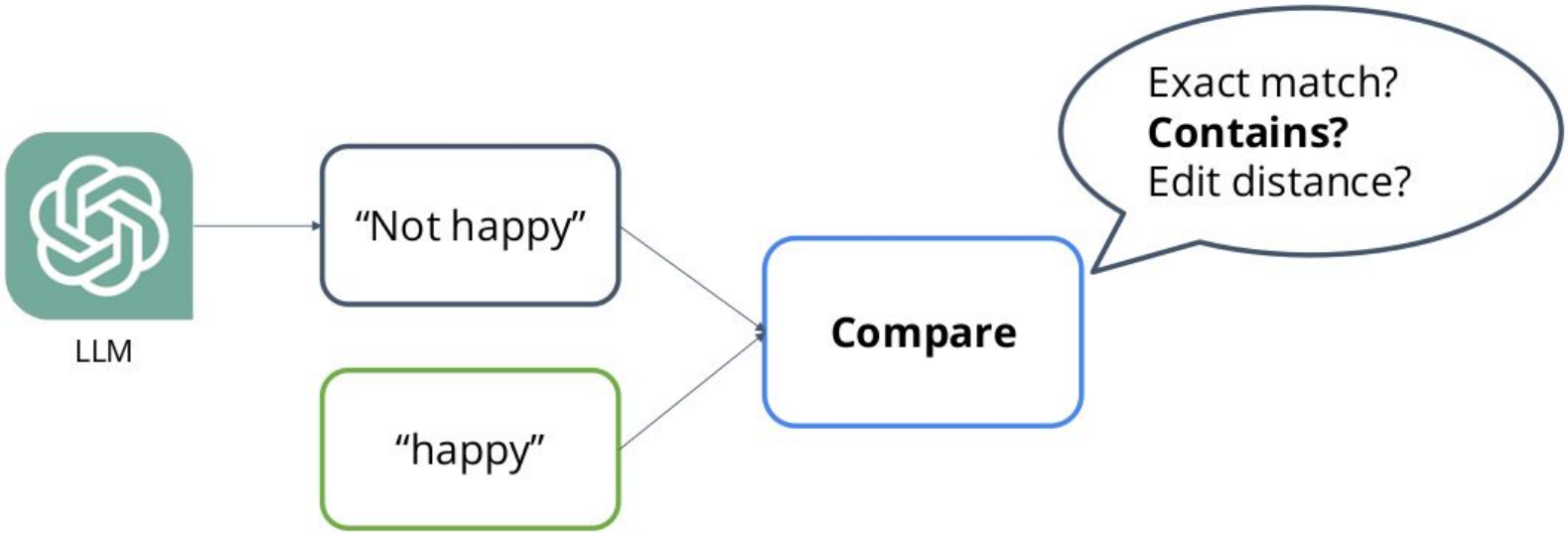
- Traditional ML Models:
  - Performance depends on feature engineering and task-specific data
- LLMs:
  - Strong generalization; adaptable to new tasks with minimal fine-tuning

Evaluation: is the LLM good at our task?

First, do we have a labeled dataset?

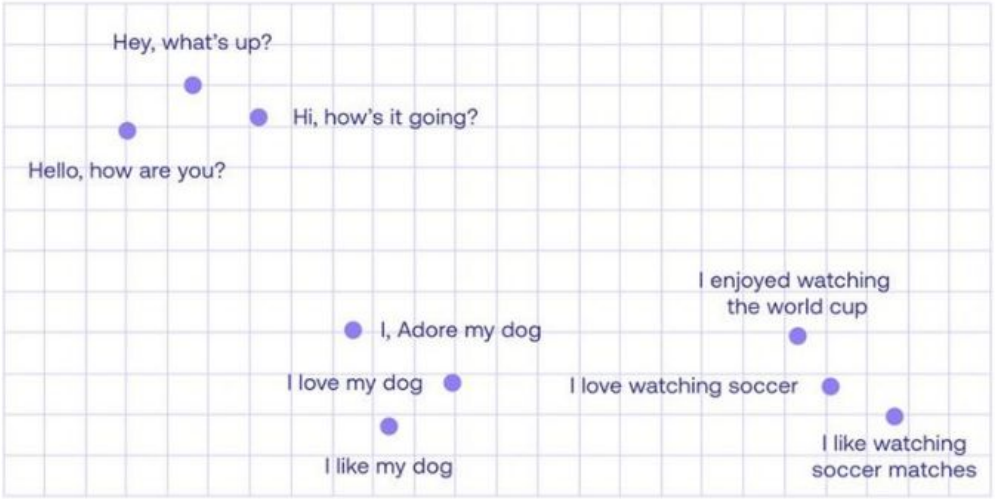


Textual Comparison: Syntactic Checks



Textual Comparison: Embeddings

Embeddings are a representation of text aiming to capture semantic meaning.



https://txt.cohere.com/sentence-word-embeddings/

LLM Evaluation

Suppose we don’t have an evaluation dataset.
What do we care about in our output?

Example: Creative Writing

- Lexical Diversity (unique word counts)
- Semantic diversity (pairwise similarity)
- Bias

LLM Evaluation: Perplexity

- A metric that shows how well a language model predicts text.
- Measures how "surprised" the model is by actual words in a sentence. Less "surprise" = better performance.
  - Lower perplexity means better predictions.
- Helps assess how accurately the model understands language patterns.
  - Lower perplexity usually means clearer, more coherent responses.
- Think of it as the model’s "confidence" in predicting words. The less it “struggles,” the better it’s doing

LLM Evaluation: NLP Benchmarks

- Standardized tests for evaluating LLM performance on language tasks.
- Popular Benchmarks in 2024
  - SuperGLUE: Measures general language understanding across tasks like question answering, coreference resolution, and sentence reasoning.
  - BIG-bench: Tests LLMs on a broad range of complex tasks, including reasoning, creativity, and logic.
  - MMLU (Massive Multitask Language Understanding): Evaluates model knowledge across academic, professional, and common-sense areas.
  - BEIR (Benchmark for Information Retrieval): Assesses LLMs on information retrieval tasks, simulating real-world search scenarios.
  - HELM (Holistic Evaluation of Language Models): Provides insights on LLM strengths and weaknesses, considering bias, toxicity, and robustness.
  - Stanford Question Answering Dataset: Tests models on answering questions based on short text passages.
- They provide a consistent way to track model improvements and compare models on a common scale.

Improving LLM Performance

Prompt Engineering

- Rewording text prompts to achieve desired output.
- Low-hanging fruit to improve LLM performance.
- Popular prompt styles
  - Zero-shot: instruction + no examples
  - Few-shot: instruction + examples of desired input-output pairs
- Don’t be too afraid of prompt length: 100+ words is OK

Improving LLM Performance

Chain of Thought Prompting

- Few-shot prompting strategy
- Example responses include reasoning
- Useful for solving more complex word problems [arXiv]
- Example:
  - Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance? Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50km
  - A: The distance that the person traveled would have been 20 km/hr * 2.5 hrs = 50 km. The answer is (e).

Improving LLM Performance

Fine-Tuning

- Retrain part of the LLM with your own data
- Create dataset specific to your task
  - Provide input-output examples (>= 100)
  - Quality over quantity!
- Generally not necessary: try prompt engineering first.

Improving LLM Performance

Fine-Tuning Output via LLM Model Settings:

Temperature

- Controls randomness in output
- Higher values (e.g., 1.0) make responses more diverse, while lower values (e.g., 0.2) make responses more focused and deterministic.

Top-k Sampling

- Limits output choices to the top k highest-probability words, reducing unlikely words. Lower k (e.g., 10) makes responses more predictable.

Top-p (Nucleus) Sampling

- Restricts choices to a dynamic set of words with a cumulative probability threshold (e.g., 0.9). This setting balances creativity and coherence.

Improving LLM Performance

Fine-Tuning Output via LLM Model Settings:

Max Tokens

- Sets the maximum length of the output, useful for limiting responses to a specific length.

Frequency and Presence Penalties

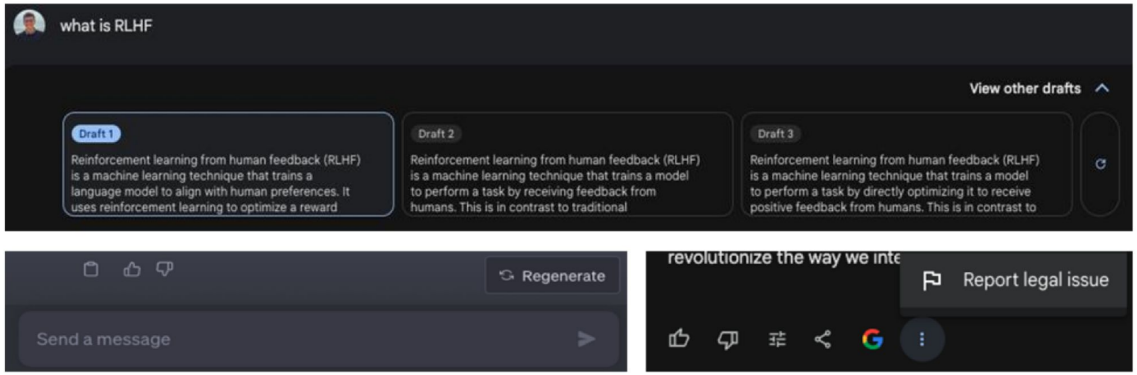
- Frequency Penalty: Discourages word repetition by reducing the likelihood of words already used.
- Presence Penalty: Encourages or discourages certain words based on how frequently they appear.

Tailoring settings allows for better control over response style, making outputs more suitable for creative tasks, factual responses, or concise summaries

Improving LLM Performance (In Production)

Reinforcement Learning from Human Feedback

Use user feedback, and interactions to improve the performance of your LLM application. Basis for the success of ChatGPT.



Next class …

Why ML/AI projects fail?

What’s wrong with the model-centric pipeline?

Are there any new challenges?

What is ML Ops?