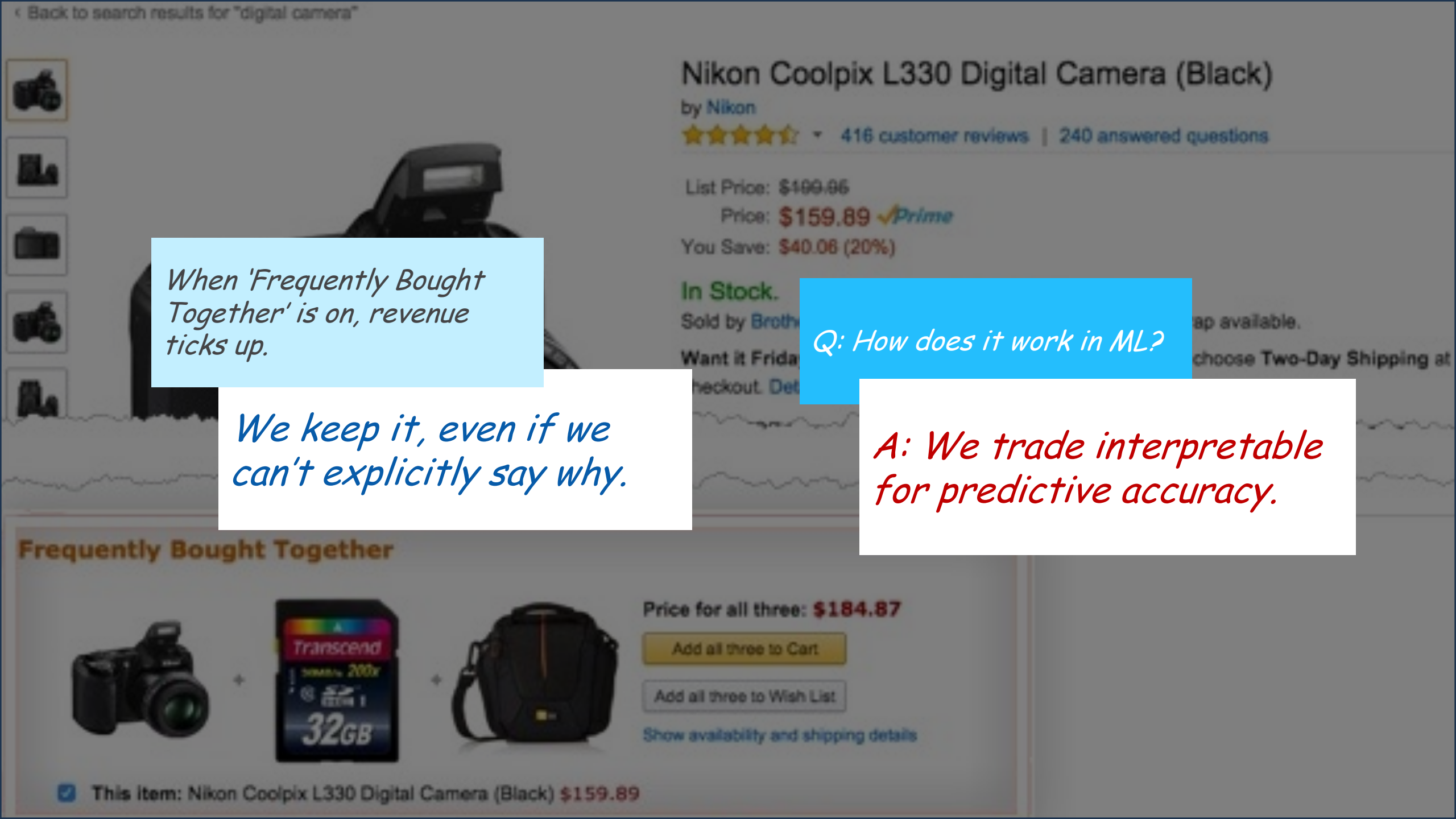
Machine Learning

An Introduction

Tarapong Sreenuch

8 February 2024

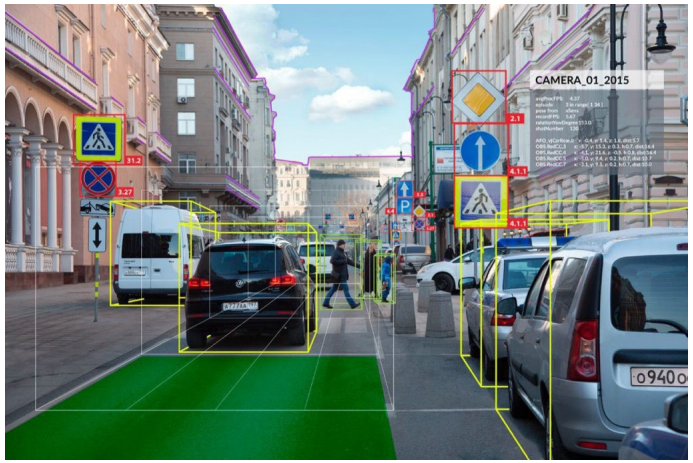
克明峻德，格物致知


‹ Back to search results for "digital camera"
Nikon Coolpix L330 Digital Camera (Black)
by Nikon
416 customer reviews | 240 answered questions
List Price: $199.96
Price: $159.89 Prime
You Save: $40.06 (20%)
In Stock.
When 'Frequently Bought Together' is on, revenue ticks up.
Q: How does it work in ML?
We keep it, even if we can't explicitly say why.
A: We trade interpretable for predictive accuracy.
Frequently Bought Together
Price for all three: $184.87
Add all three to Cart
Add all three to Wish List
Show availability and shipping details
This item: Nikon Coolpix L330 Digital Camera (Black) $159.89

Machine Learning

Machine Learning lowers cost of prediction.

The Economics of Artificial Intelligence, McKinsey



Cost of Prediction

Cost of Prediction

1950

2016

scikits learn machine learning in Python

amazon EMR

Microsoft Azure

TensorFlow

Azure Machine Learning

Amazon Machine Learning

Google's AutoML

SW Library

GPU

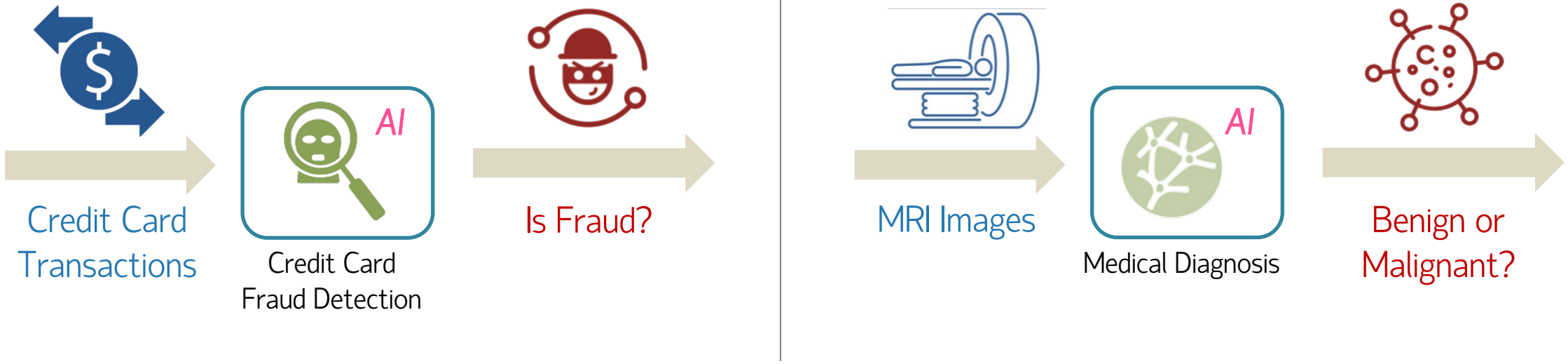
AI Algos

Cloud Computing

Technologies are being commoditised.

'Cheap Means Everywhere'

Prediction



PREDICTION *is the process of filling in missing information.*

Prediction takes information you have, often called '*data*', and uses it to generate information you don't have.

Cheap Means Everywhere

We're now starting to convert non-prediction tasks to be prediction tasks.

Self-Driving Car

What will a good human driver do?

Language Translation

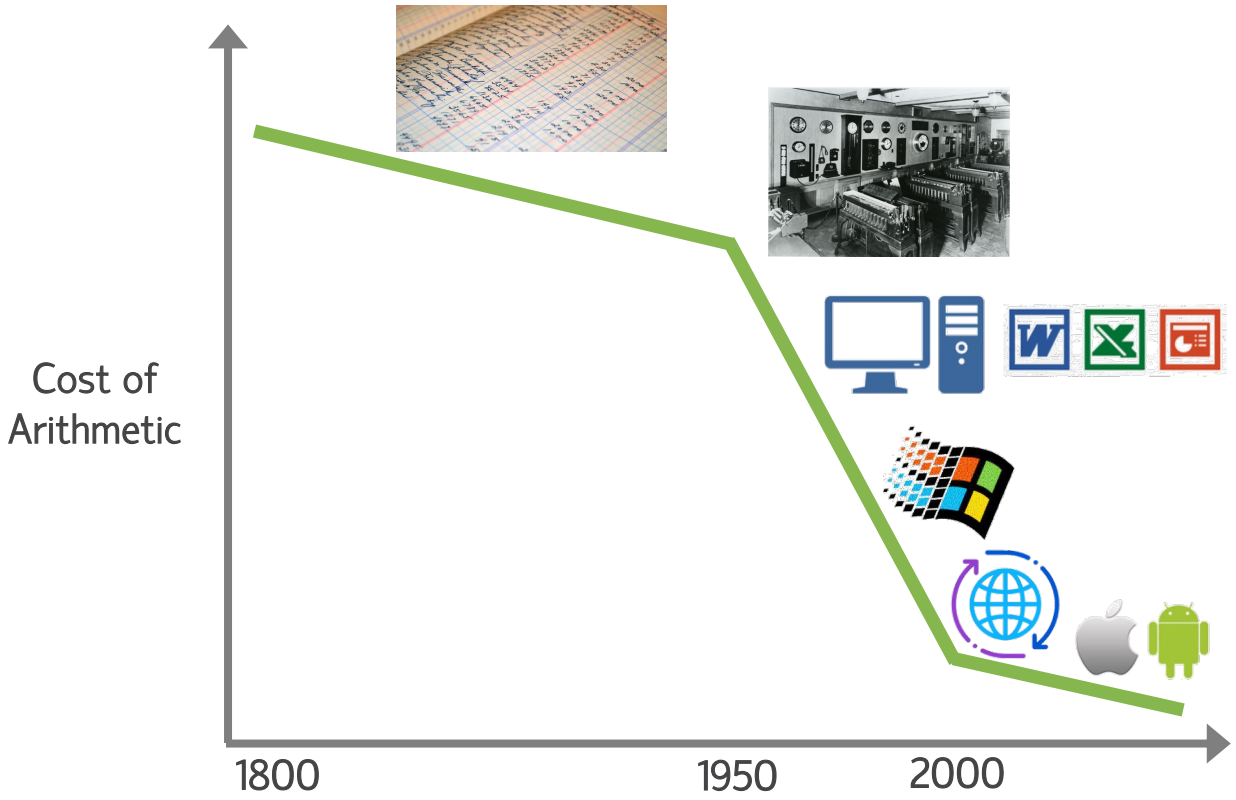
Used to be for experts who know the rule and exception of translation.

Computers

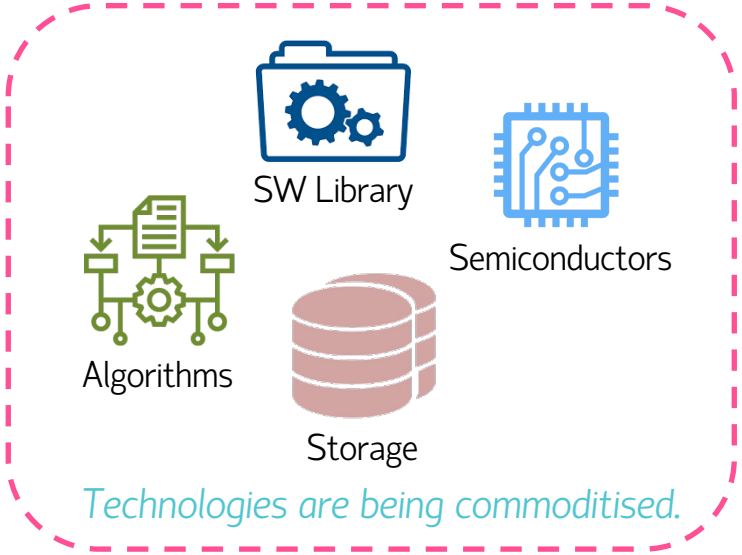
Computers do arithmetic and nothing more.

Tim Bresnahan, Stanford Economist

Cost of Arithmetic



'Cheap Means Everywhere'



Cheap Means Everywhere

We're now starting to convert non-arithmetic tasks to be arithmetic ones.

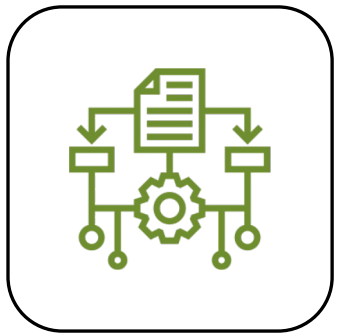
Play Music

Used to be Art & Humanity, not Machine

Digital Photography

Used to be a Chemistry Problem

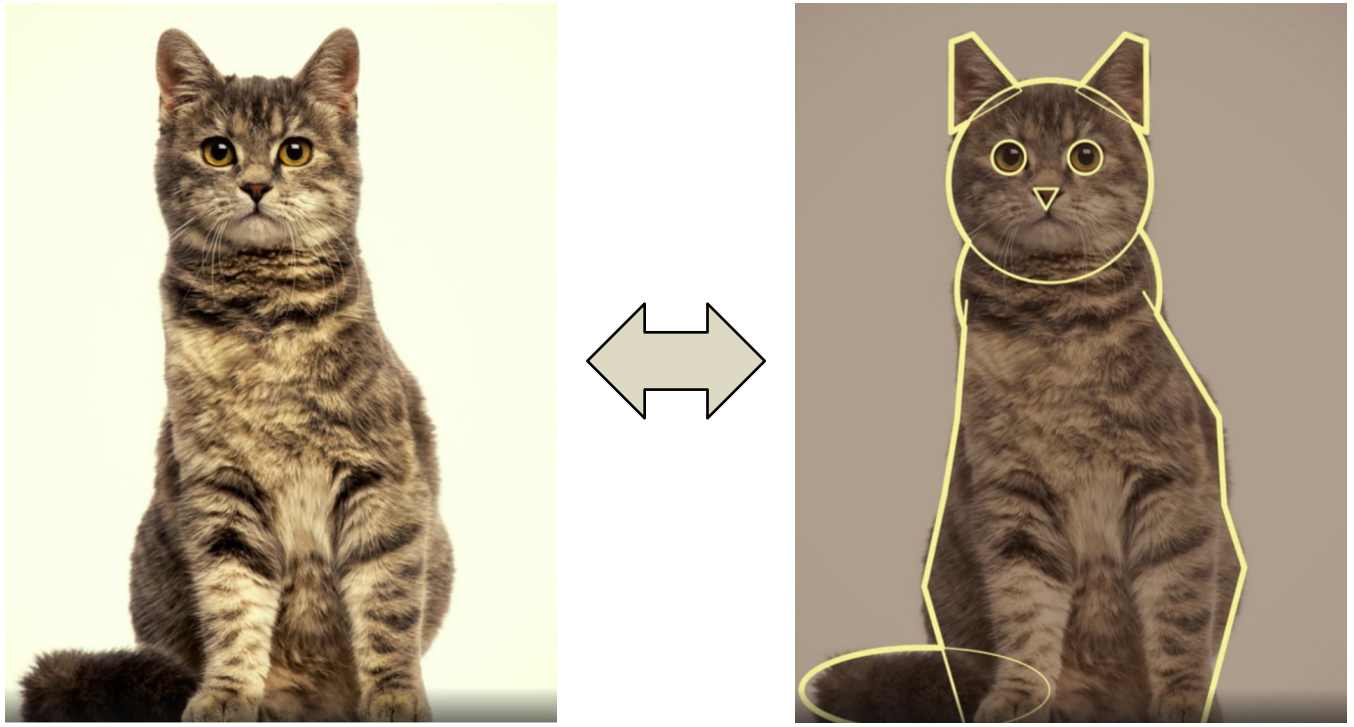
Prediction: Object Recognition



Object Recognition
Algorithm
(To be developed)

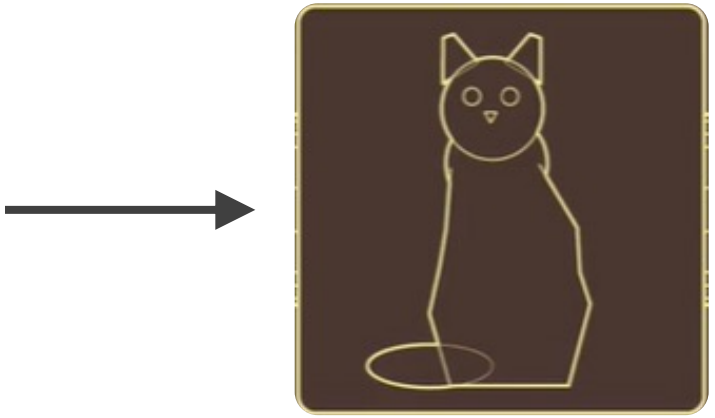
'Cat'

Algorithmic, Mathematical Model



- *'Model'* can be a geometric representation of an object type.
  - Cats, for an example, would be
    - 1 round ace
    - 2 pointed ears
    - 1 cylinder body
  - *This will be different for different object types.*
- *'Model'* is indeed an approximation of the real-world.

Geometric-Driven Object Recognition Model



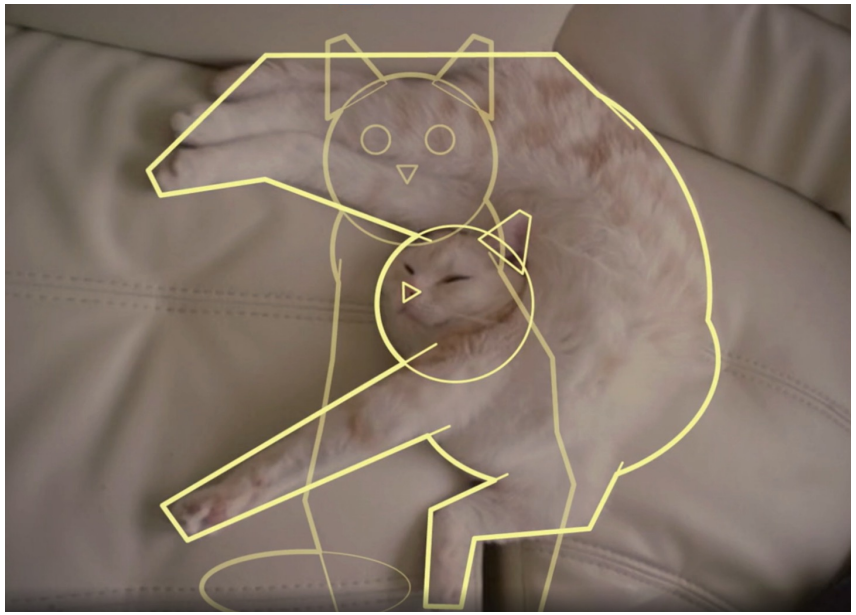
Algorithmic Model

'Cat'

The algorithm will try to correlate the object with the pre-defined geometric properties.

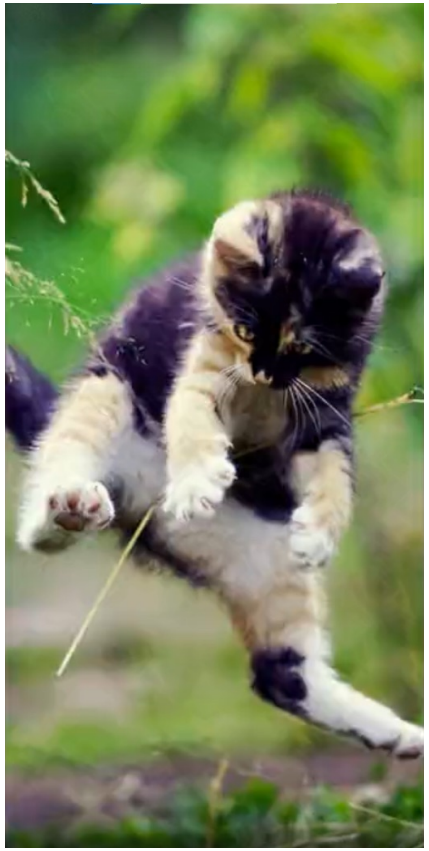
Handling Complexities

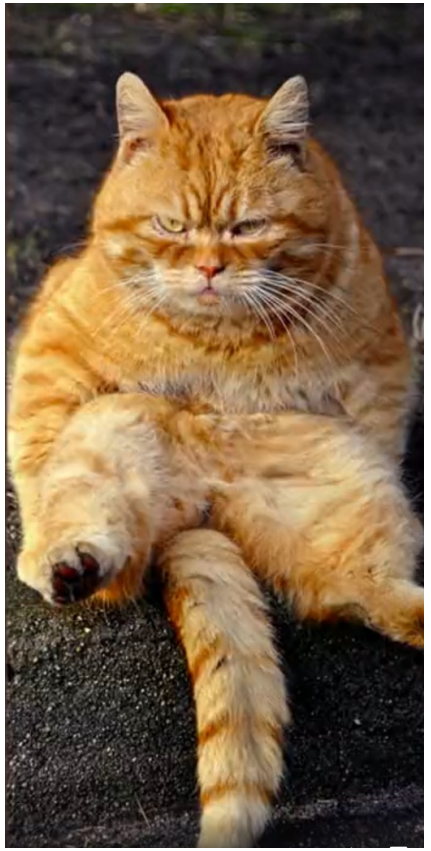
How about this New Scenario?



- Simply, we just need to code in more rules (or conditions) to hand a new scenario.
- We build in more possible geometric representation.
- *We are extending our code coverage.*

Infinite Number of Variations





This will mean infinite number of rules that we code them in.

How We Learn

Brain + Sensors

Data →
Answers →

Machine Learning

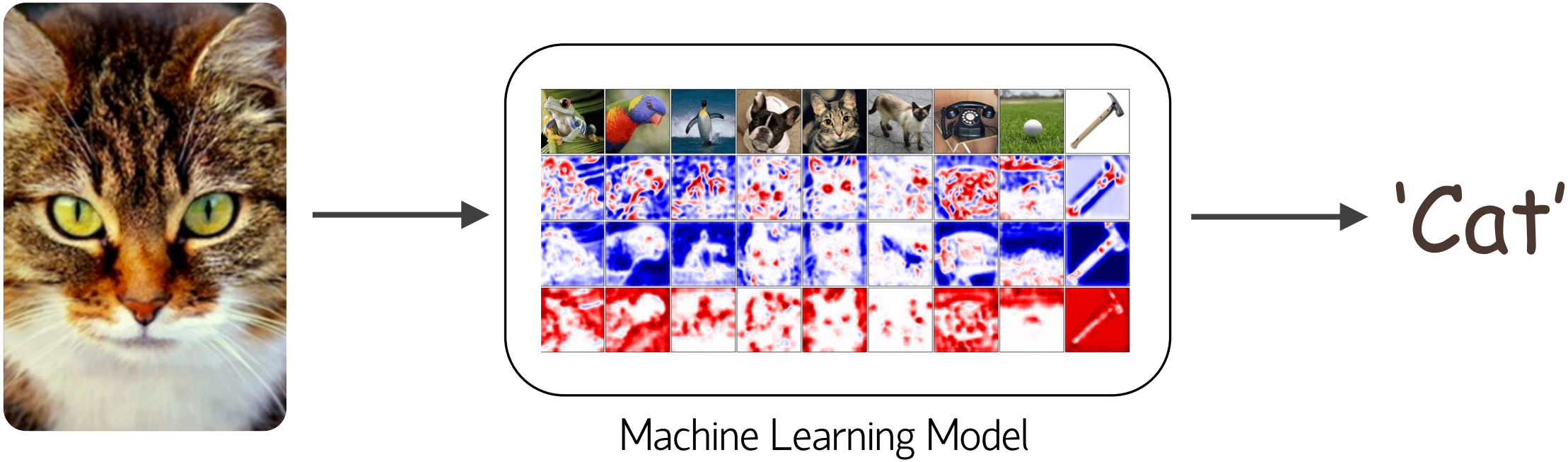
→ *"implicit"* Rules

... somehow we just can differentiate cats from dogs, but hard to explicitly describe ...

Why Now?

Commoditisation of computing and storage resources, and we now have tons of data available, e.g. ImageNet.

Machine Learning Model



With lots of data, we let the machine to learn those intricated features that are necessary to identify objects. We don't pre-define the model ourself.

Often, the model is so complex for us to understand, hence the implicitness.

Paradigm Shift

Many problems have transformed from algorithmic problems to prediction problems.

Algorithmic Model

"what are the features of a cat?"

It requires the articulation and hypothesis or a least of human intuition for model specification.

... first proving it works in theory ...

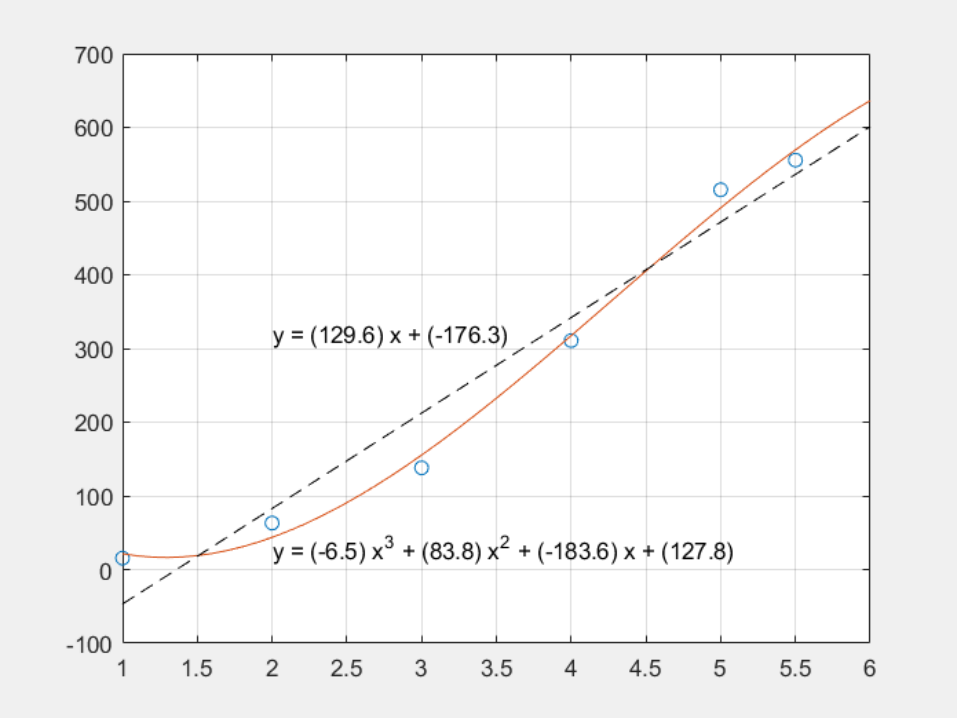
Machine Learning Model

"does this image with a missing label have the same features as that the cats that I have seen before?"

Machine Learning is more experimental, no prior assumption. It only needs to prove it works in practice.

... when the prediction miss, they often don't miss by much ...

Trade-Offs



We can fit any data with minimal errors if we keep increasing the complexity of our model.

We are good at comprehend and explain a simple model, e.g. linear (1st order), but not models with high complexities.

- In Science and Engineer, we feel uncomfortable if we cannot explain or understand our model.
- In Machine Learning, we essentially trade *understanding* for *predictive powers*. However, we do have applications, e.g. product recommendation, where *accurate predictions outweigh the need for throughout understanding of the model*.

Types of Machine Learning

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

… recall: our ML models're learning from data ….

Training Data: *'Machine Learning Jargon'*

- Training Data: existing data to learn from
- Training a model: when a model is being built from training data
  - It can take microseconds to weeks.

Training Data: Supervised Learning

Features

Target Variable

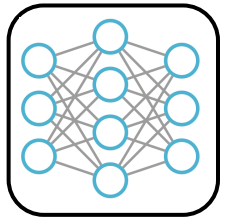
Observations (or Data Points)

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
55	M	221	5	True	typical angina	118	True
50	F	196	0	False	non-anginal pain	98	False
53	F	215	0	True	asymptomatic	110	True
62	M	245	3	False	typical angina	126	True
48	M	190	0	True	non-anginal pain	99	False
70	M	201	0	True	typical angina	105	False

Labels

After Training: Supervised Learning

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
65	F	208	2	False	typical angina	105	???



Machine Learning Model

Heart disease
False

Output (Prediction)

Supervised vs Unsupervised Learning

- Supervised Learning
  - Training data is "labelled".

- Unsupervised Learning
  - Training data only has features.
  - Useful for:
    - Anomaly detection
    - Clustering, e.g. dividing into groups

Target Variable

ood gar	Heart disease
8	True
3	False
0	True
6	True
)	False
5	False

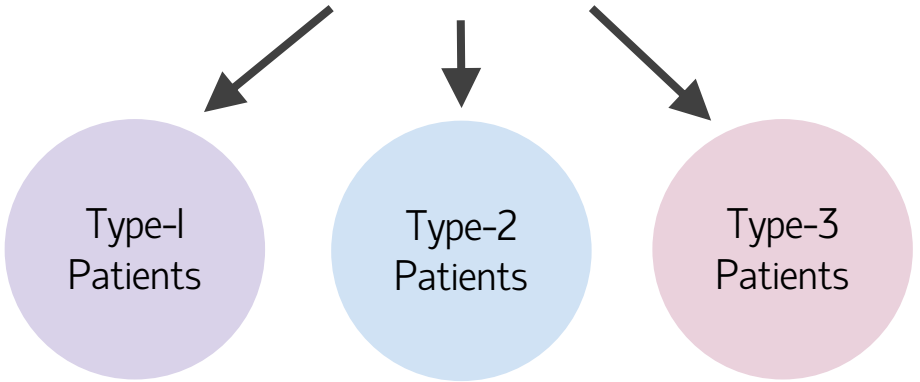
Labels

It's a bit like...

We start of with  *and end up like*  .

Training Data: Unsupervised Learning

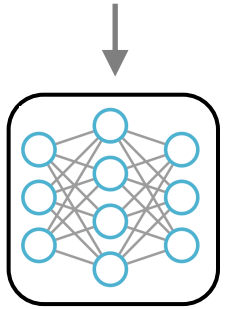
Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
55	M	221	5	True	typical angina	118	True
53	F	199	0	True	non-anginal pain	98	True
53	F	215	0	True	asymptomatic	110	True
62	M	245	3	False	typical angina	126	True
...	...	...	...	...	...	...	...



- Think about our 'T-Shirt' scenario.
- We group them based on their similarities.

After Training: Unsupervised Learning

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
65	F	208	2	False	typical angina	105	True



Machine Learning Model

Type-2 Patients

Output (Prediction)

Unsupervised Learning

- In reality, data doesn't always come with labels.
  - It requires manual labour to label, which is costly.
  - Labels are unknown.

- No labels: Model is unsupervised and finds its own patterns.

Machine Learning Lingo

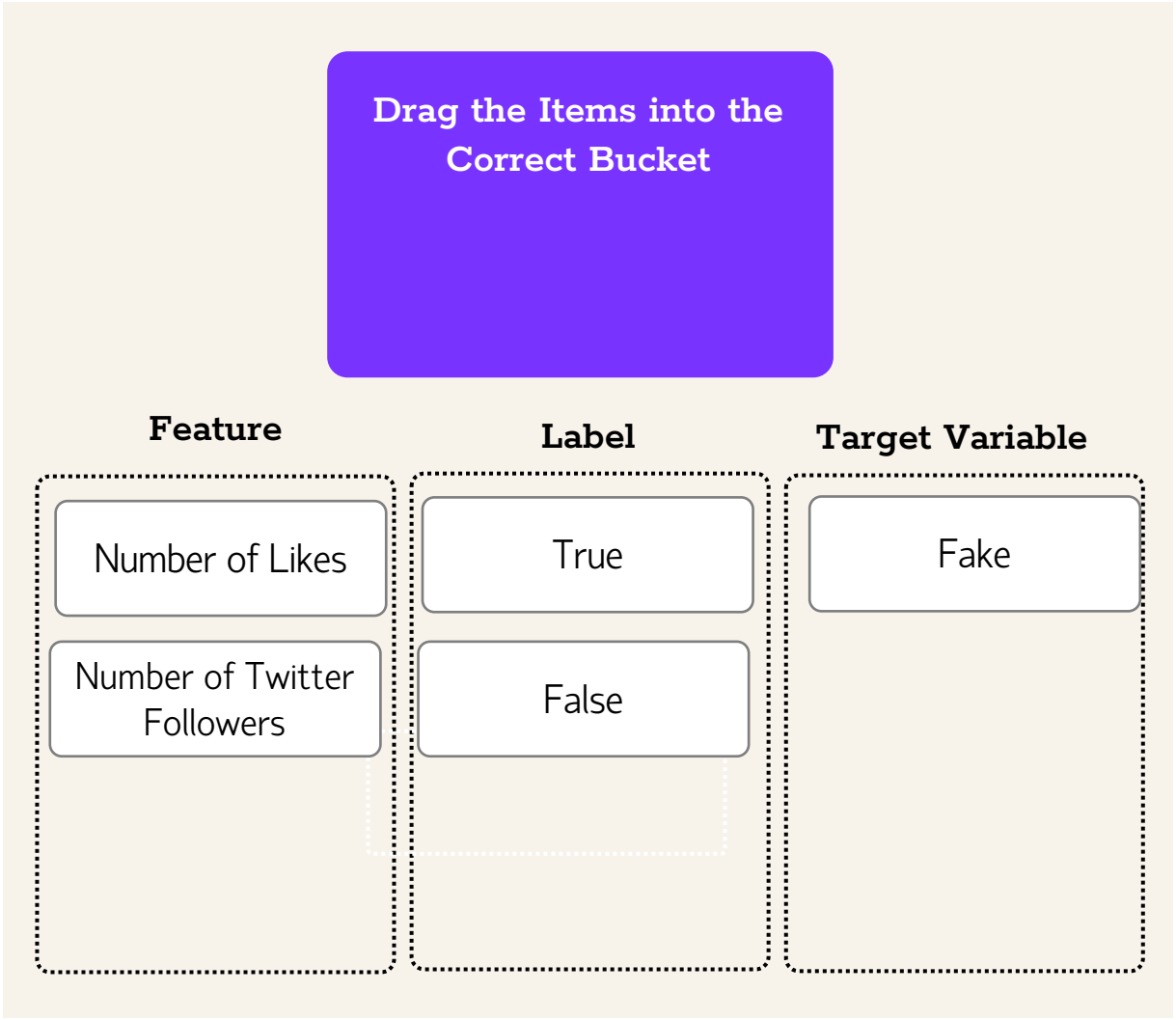
You overhear a group of data scientists discussing their latest machine learning project on predicting whether a tweet is fake or not. Twitter has provided them with a labeled dataset in hopes of improving their spam detection system.

Here are two observations from the dataset:

Tweet text	# of likes	# of Twitter followers	Fake
One-day sale on sunglasses -> check out this site!	0	2	True
Beautiful day for surfing 🏄	3	65	False

Instruction

- Based on this information, classify aspects of their training data as either a feature, label, or target variable.



Supervised Learning: Classification

Classification = *Assigning a* **Category**

- Will this customer **stop** his/her subscription?
  - *Yes, No*
- Is this mode **cancerous**?
  - *Yes, No*
- What **kind** of wine is that?
  - *Red, White. Rose*
- What **flower** is that?
  - *Rose, Tulip, Carnation, Lily*

Features *Target Variable*

Applicant ID	High school GPA	Test results	Accepted
0	3.5	2.4	False
1	4	2.2	False
2	4.2	4.3	True
3	4.8	2.9	False
...	...	...	...

Supervised Learning: Regression

Regression = *Assigning a* **Continuous** *Variable*

- How much will this stock be **worth**?
- What is this exoplanet's **mass**?
- How **tall** will this child be as an adult?

Features *Target Variable*

Reading ID	Humidity rate	Temperature in °C
0	0.89	7.388889
1	0.86	7.227778
2	0.89	9.377778
3	0.83	5.944444
...	...	...

Types of Supervised Learning

You know there are two flavors of supervised learning: classification and regression. Let's see if you can distinguish between these two types of problems.

Instruction

- Classify the problems on the right as classification or regression.

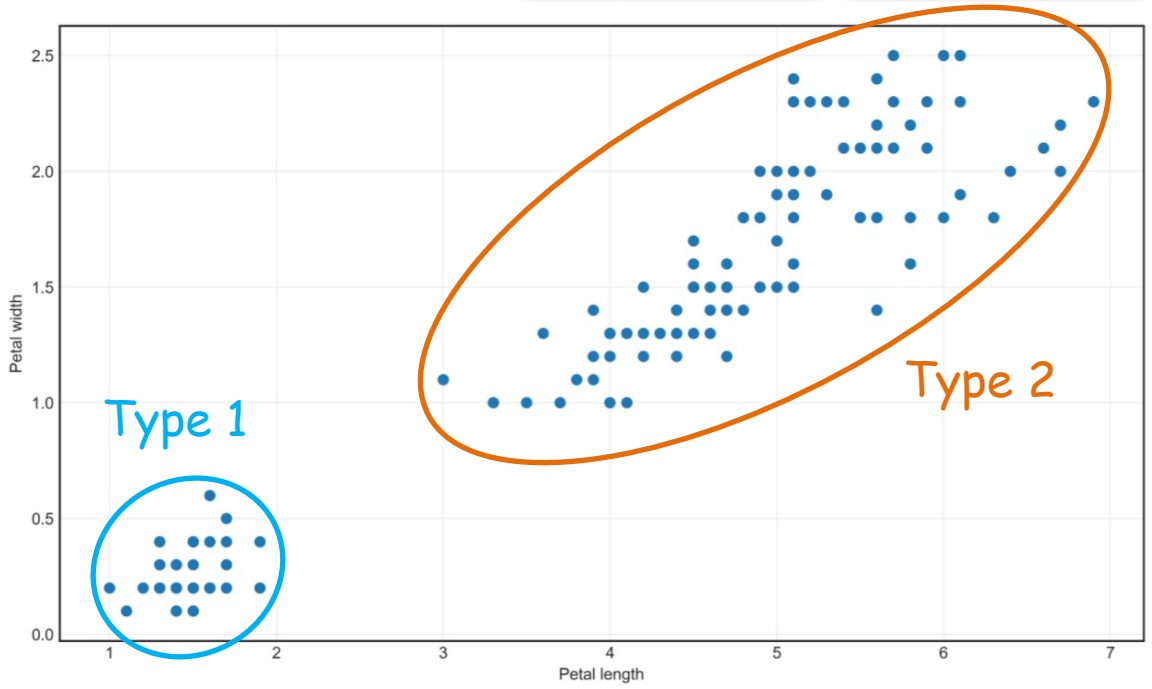
Drag the Items into the Correct Bucket

Classification	Regression
Based on chemical features (alcohol, pH, chlorides ...), predict whether a wine is red, white or rose.	Based on employee's attributes (seniority, income, department, distance from home...), predict how long until an employee looks for another job.
Based on space object attributes (discovery method, orbit, inclination, mass), predict whether this object is an exoplanet or not.	Based on chemical features (alcohol, pH, chlorides ...), predict the price of a wine.
Based on song features (length, key, loudness, tempo ...), predict a song's genre.	Based on people's attributes (level of education, area, job title, age ...), predict their income.

Unsupervised Learning: Re-Visit

Unsupervised Learning = No Target Column

- No guidance
- Looks at the whole dataset
- Tries to detect patterns



Types of Machine Learning

You now know about supervised and unsupervised learning. Let's see if you can distinguish them.

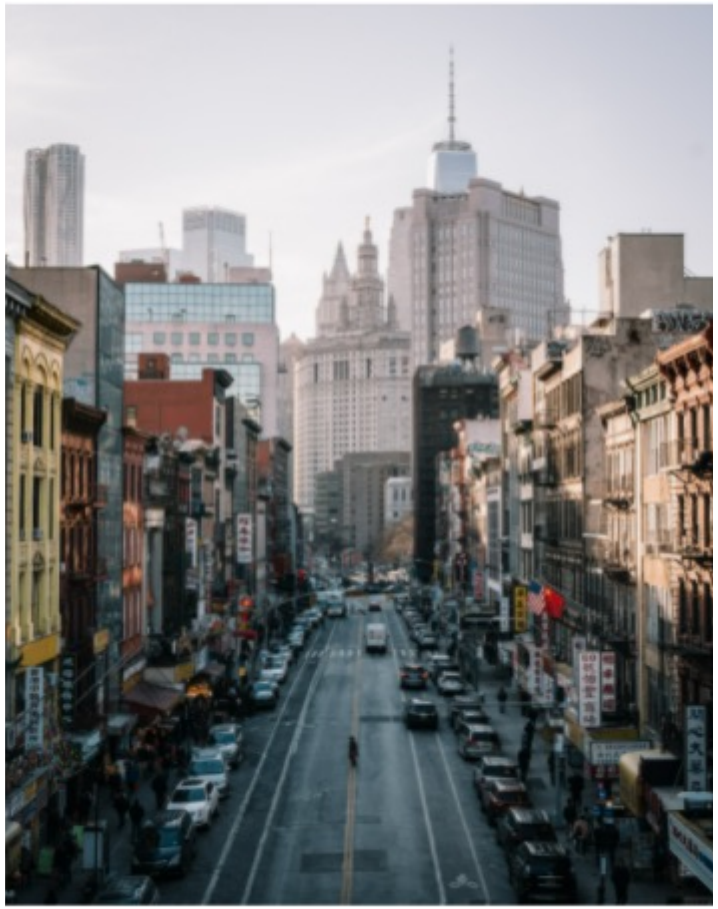
Instruction

- Some use cases are presented on the right. Decide if they require a supervised or unsupervised approach and drag them in the correct bucket.

Drag the Items into the Correct Bucket

Supervised	Unsupervised
Based on customer information (sign up date, ordering frequency, age, marital status), predict their yearly spending amount.	Based on customer's purchase history, find which items they are likely to be interested in next.
Based on email features (sender, topic, ratio of uppercased letter, proportion of 'money' term), predict if an email is spam or not.	Based on customer information (sign up date, ordering frequency, age, marital status, spending income), find segments of customers.
Based on stock information (open value. Highest value, lowest value, close value of each day), predict its future value.	Based on email information (sender, topic ...), find groups of emails by theme.

Our Scenario



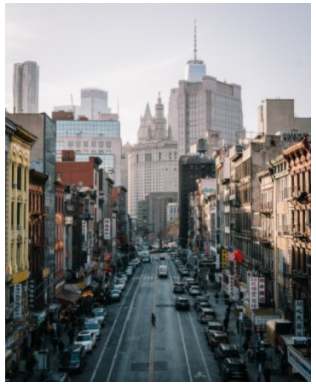
Our Dataset: NYC property sales from 2015–2019

Columns:

- Square feet
- Neighbourhood
- Year built
- Sale price
- Etc...

Our Target Variable: Sale Price

Machine Learning Workflow



Raw Data

1. Extract Features

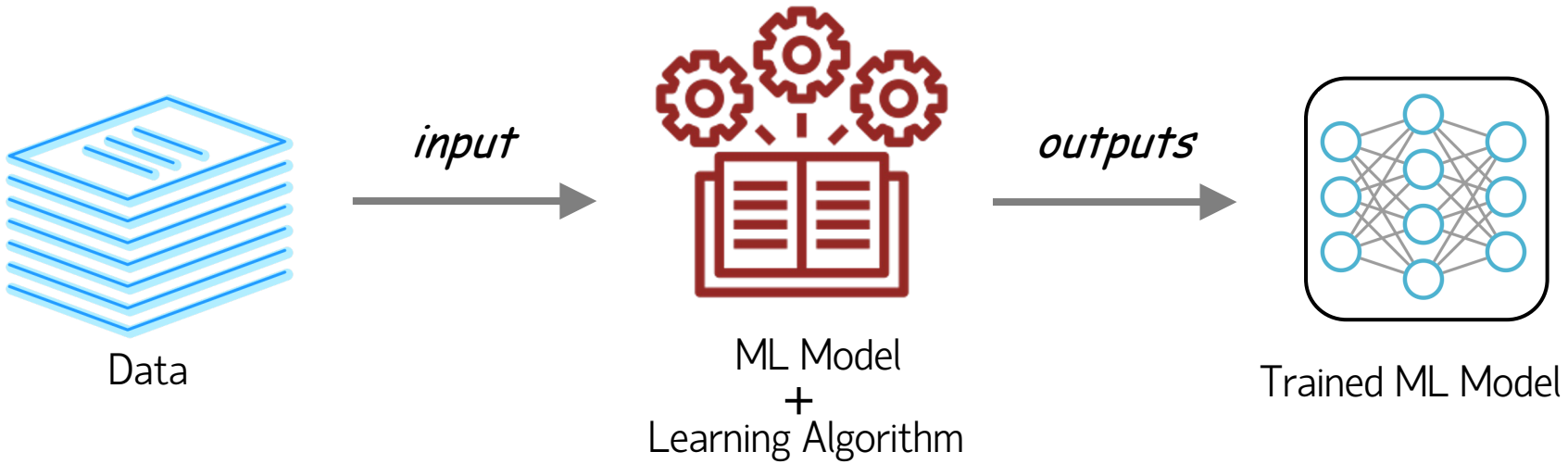
2. Split Dataset

Train Dataset

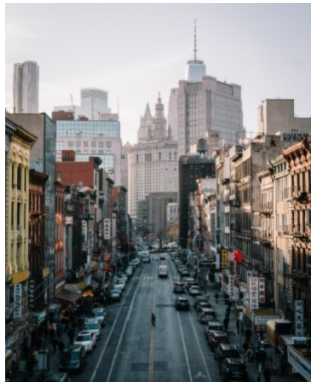
3. Train Model

Test Dataset

Step 3: Training Model



Machine Learning Workflow



Raw Data

1. Extract Features

2. Split Dataset

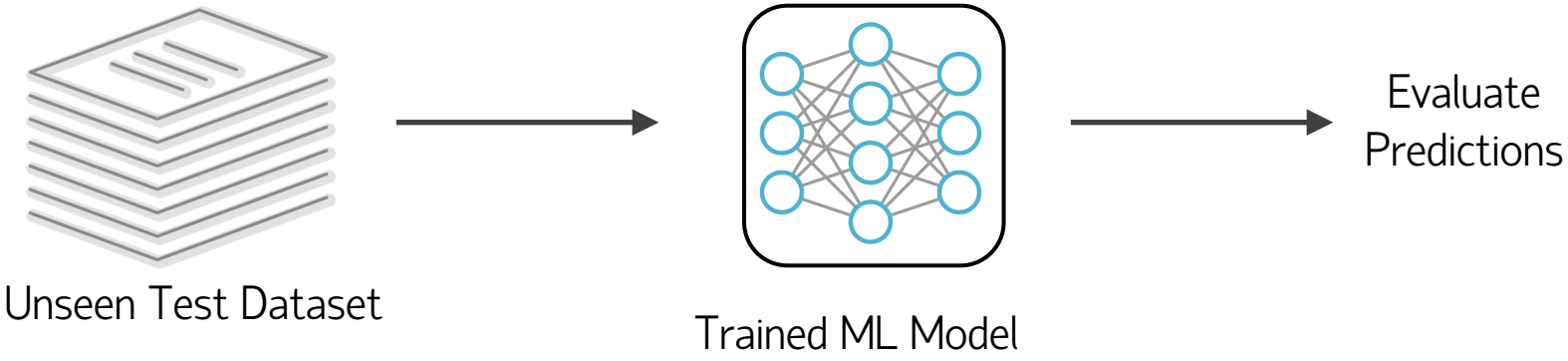
Train Dataset

3. Train Model

4. Evaluate

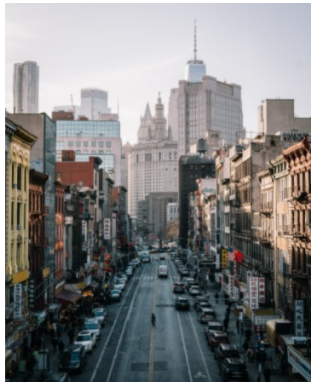
Test Dataset

Step 4: Evaluation



- Test Dataset: *'Unseen'*
- Many ways to evaluate:
  - What is the average error of the prediction?
  - What percent of apartments did the model accurately predict within a 10% margin?

Machine Learning Workflow



Raw Data

1. Extract Features

2. Split Dataset

Train Dataset

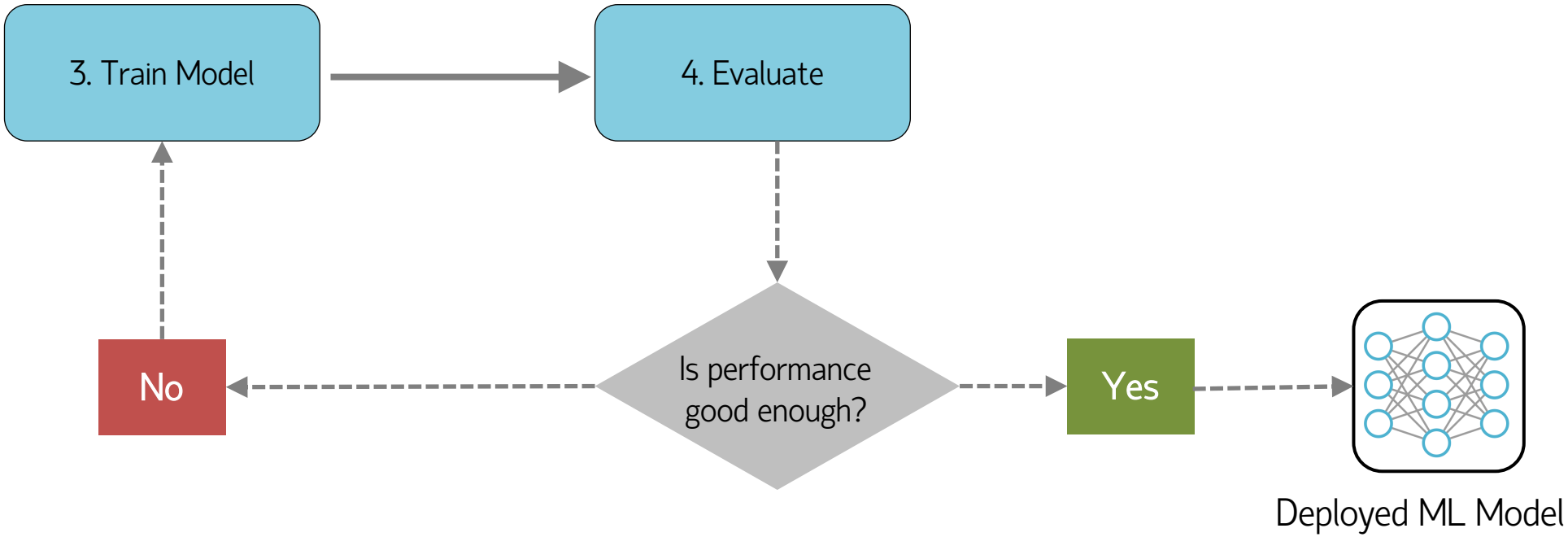
Test Dataset

3. Train Model

4. Evaluate

Deployed Machine Learning Model

Model Tuning



Summary of Steps

1. Extract Features
   - Choosing features and manipulating the dataset

2. Split Dataset
   - Train and test datasets

3. Train Model Dataset
   - Input train dataset into a machine learning model

4. Evaluate
   - If desired performance isn't reached: tune the model and repeat Step 3

Steps for Building a Model

Machine learning is integrated in many of the technologies we use everyday. For example, have you noticed that platforms will have personalized recommendations, whether it's another funny video on Youtube or a book by your favourite author on Amazon? These are "recommender systems" and they typically consist of a machine learning model trained on a user's browsing history.

Imagine the recommender system of your favorite online clothing store. They have data on all the clothes you've viewed and the clothes you ended up buying. This is enough to make a model to output personalized clothing recommendation for you. On the right are tasks to create this model, however, they are incorrectly ordered.

Instruction

- Correctly order the tasks.

Drag the Items into Order

Extract the features for each product in the shop, including brand, number of times viewed, cost, color and clothing type.

Split the dataset into 2/3 and 1/3 for the train and test dataset, respectively.

Train the model using the train dataset and a logistic regression model.

Evaluate the percentage of products in the test dataset that were accurately predicted as bought.

Summary

- Machine Learning lowers the cost of prediction. This accelerates an adoption of AI. We also observe ML being used in applications that not associated with prediction.

- Machine Learning is better in modelling intricated details of the real-world. ML models have high predictive power, but somewhat implicit.

- In Supervised Learning, models are created from the labelled data, learning from examples. On the contrary, Unsupervised Learning aims to self discover patterns hidden in the data.

- Regression assigns continuous variables, while Classification is categorical. Both ML models are trained and evaluated on train and unseen data, respectively.