

# Machine Learning for Product Managers

## Course Slides

[Part 1: Intro to ML](#)

[Part 5: Prepare Your Data](#)

[Part 2: When to ML](#)

[Part 6: Build Your Model](#)

[Part 3: How to ML](#)

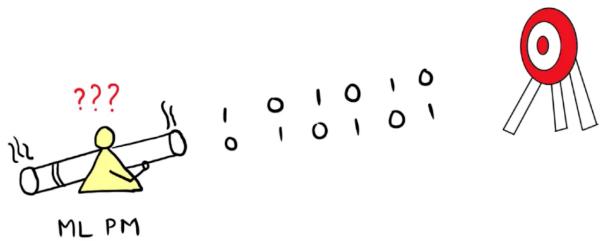
[Part 7: Deploy Your Model](#)

[Part 4: Get Your Data](#)

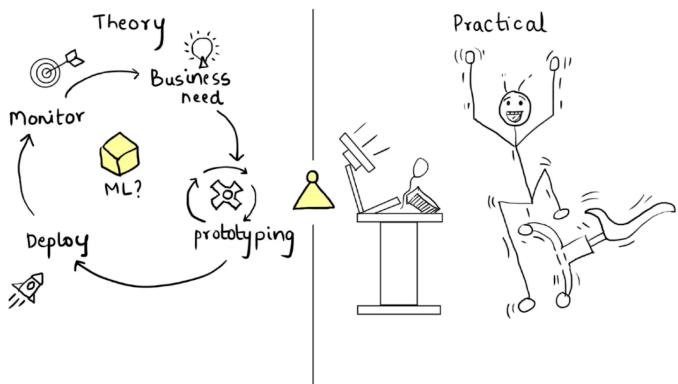
## Part 1: Introduction

Lecture #1: Congratulations! You're Now an ML Product Manager

Notes:



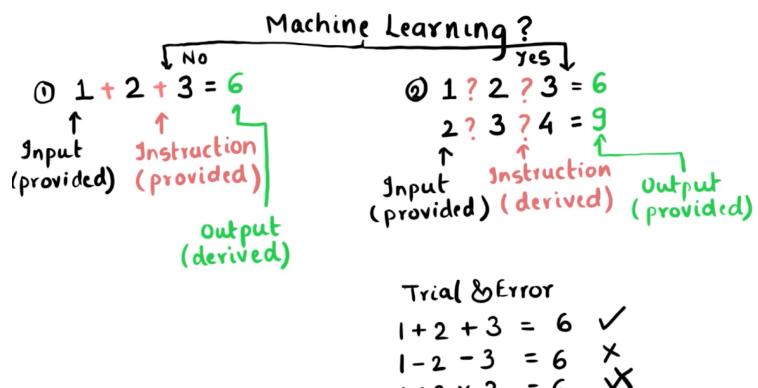
Galaxy



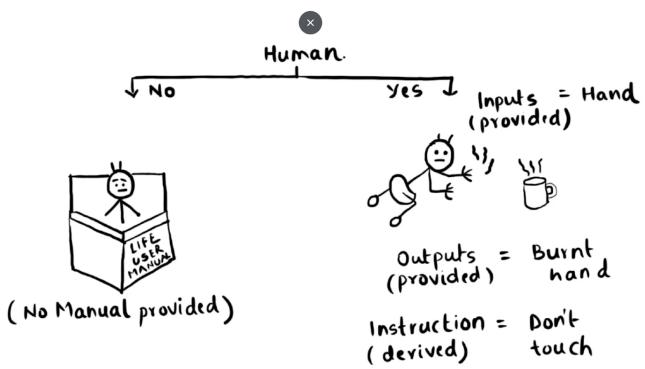
Galaxy

Lecture #2: What is Machine Learning (ML)?

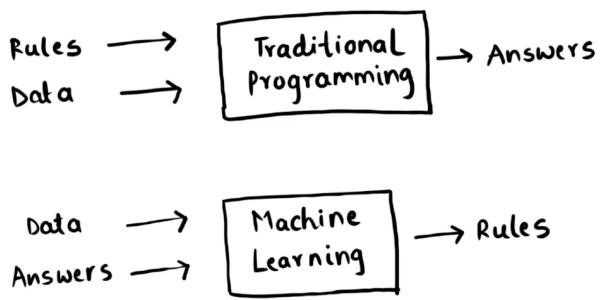
Notes:



Galaxy

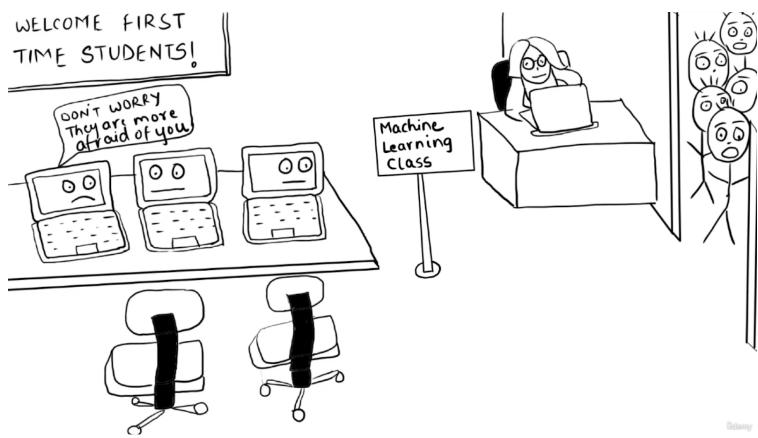


Notes:



Notes:

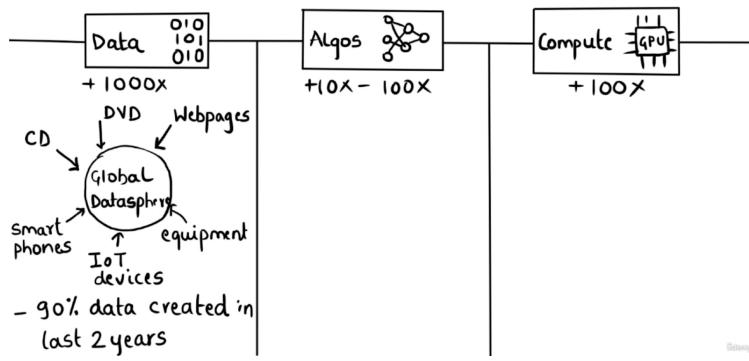
#### Lecture #4: ML is Going Mainstream



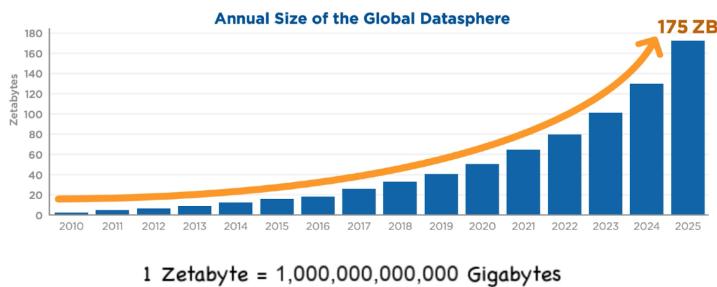
Notes:

Notes:

## Why ML Now?



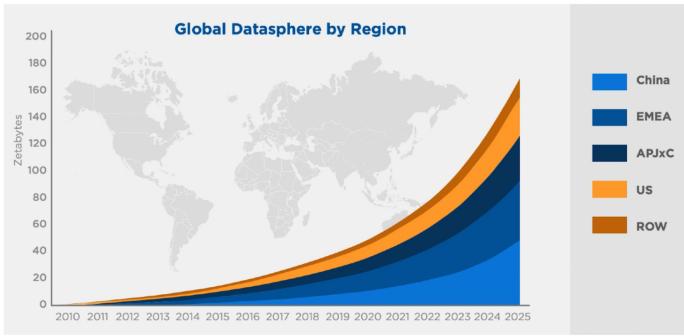
Notes:



Notes:



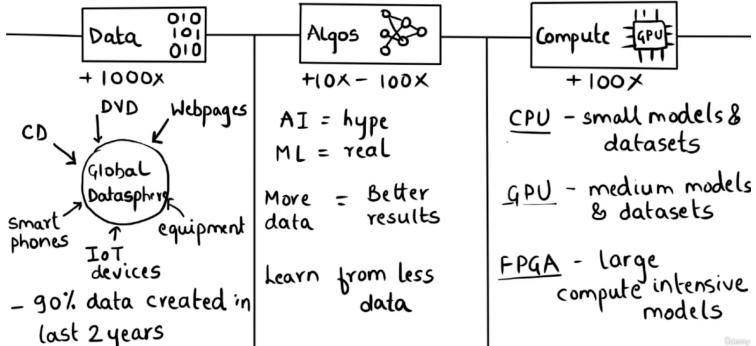
Notes:



Source

Notes:

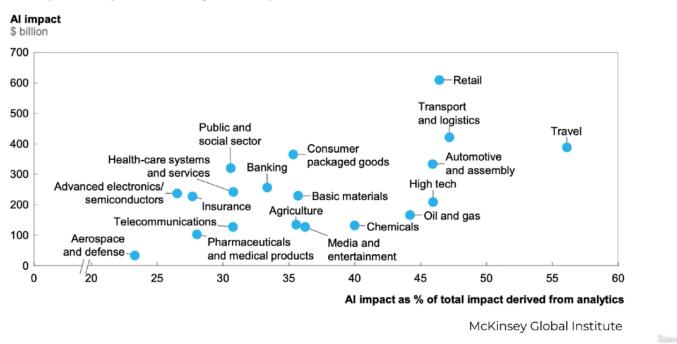
### Why ML Now?



Source

Notes:

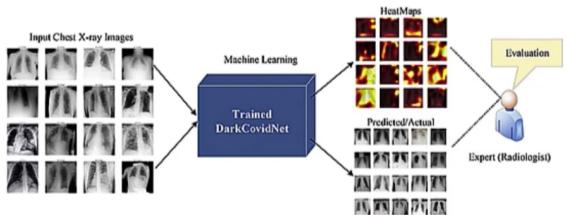
AI has the potential to create annual value across sectors totaling \$3.5 trillion to \$5.8 trillion, or 40 percent of the overall potential impact from all analytics techniques



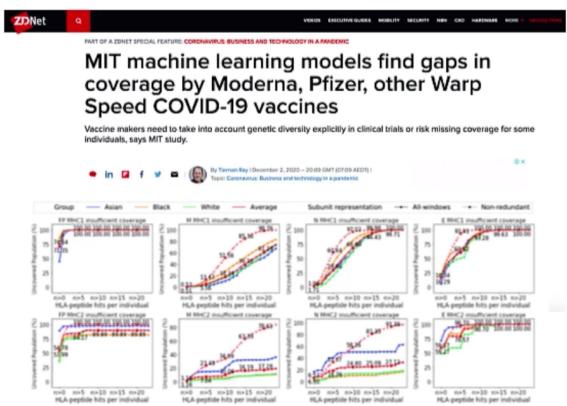
Source



Notes:



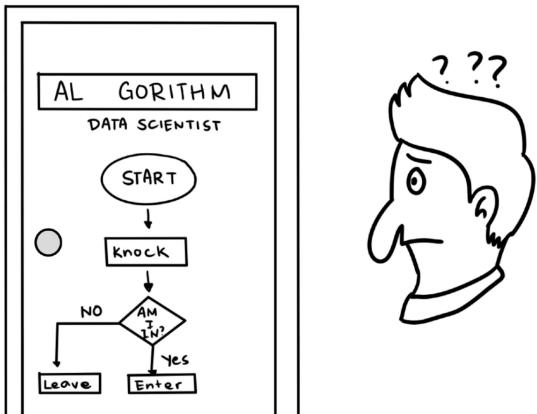
Notes:



Notes:

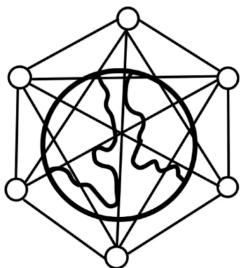
## Lecture #7: The Learning Algorithm

Notes:



Galaxy

## Rise of Algorithms



Fragile Tower  
of  
Algorithms

Galaxy

## FUTURE

[What is BBC Future?](#) [Follow the Food](#) [Made on Earth](#) [Future Planet](#) [Future](#)

A surprisingly simple bug afflicts computers controlling planes, spacecraft and more – they get confused by big numbers. As Chris Baraniuk discovers, the glitch has led to explosions, missing space probes and more.

**T**uesday, 4 June 1996 will forever be remembered as a dark day for the European Space Agency (Esa). The first flight of the crewless Ariane 5 rocket, carrying with it four very expensive scientific satellites, **ended after 39 seconds in an unholy ball of smoke and fire**. It's estimated that the explosion resulted in a loss of \$370m (£240m).

What happened? It wasn't a mechanical failure or an act of sabotage. No, the launch ended in disaster thanks to a simple software bug. A computer getting its maths wrong – essentially getting overwhelmed by a number bigger than it expected.

Story continues below

ADVERTISEMENT FEATURE  
PRESENTED BY **BHP**

Notes:

Notes:

## Software glitch to blame for blackout of extra 60,000 SA homes in heatwave

The South Australian network operator says a software problem led to load shedding of 300MW instead of the 100MW requested by national market



Galaxy

Notes:

**COMPUTERWORLD** AUSTRALIA ▾ CLOUD OFFICE SOFTWARE WINDOWS MOBILITY CAREERS RESOURCE LIBRARY

Home > Industry > Financial Services Industry

NEWS

## Regulators blame computer algorithm for stock market 'flash crash'

Joint SEC-CFTC investigation expected to lead to the rollout of new rules designed to prevent similar crashes



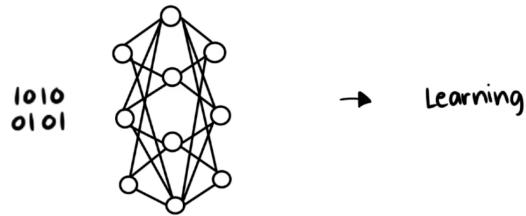
By Lucas Mearian

Senior Reporter, Computerworld | 2 OCTOBER 2010 10:33 AEDT

Galaxy

Notes:

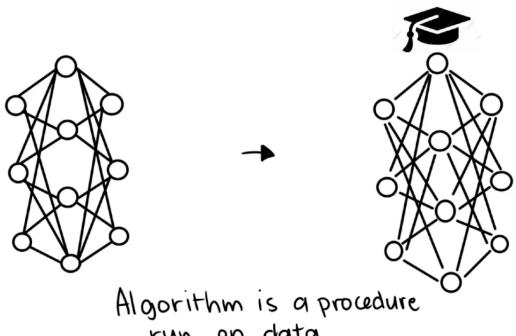
## The Learning Algorithm



Galaxy

Notes:

## The Learning Algorithm



Galaxy

Notes:

## ML Algorithm vs. ML Model

Training data + ML Algo = ML Model

ML Challenge

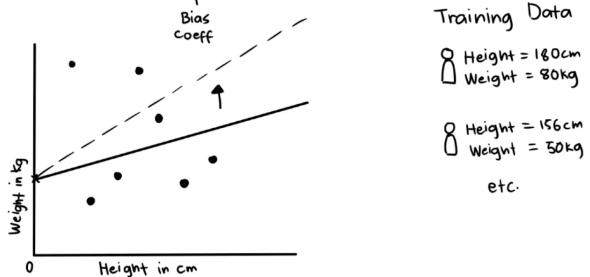
Predict a persons weight from their height  
relationship is Linear

Galaxy

Notes:

## Linear Regression

$$\text{Weight} = \beta_0 + \beta_1 (\text{Height})$$



Galaxy

## ML Model

$$\text{Weight} = 0.1 + 0.5 \text{ (Height)}$$

e.g.

$$0.1 + 0.5 \text{ (182 cm)} = 91 \text{ kg}$$

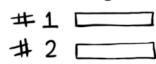
ML Algos try to  
**minimize** error rate

Notes:

## Lecture #8: Types of ML

### 6 Types of ML Problems

#### ① Ranking



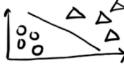
- eg. Google Bing  
- many variables  
- Always changing

#### ② Recommendation



- eg. Netflix & Spotify  
- Based on behavior and profile

#### ③ Classification



- eg. Gmail spam  
Facebook face detection  
- categories are predefined

Notes:

### 6 Types of ML Problems

#### ④ Regression



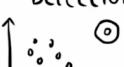
- eg. Zillow & Trulia  
- predict a value

#### ⑤ Clustering



- eg. Amazon's customer also bought "x"  
- Categories are not pre-defined

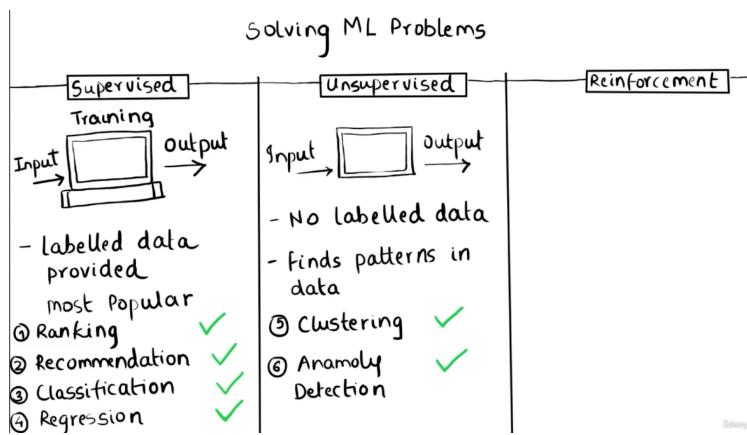
#### ⑥ Anomaly Detection



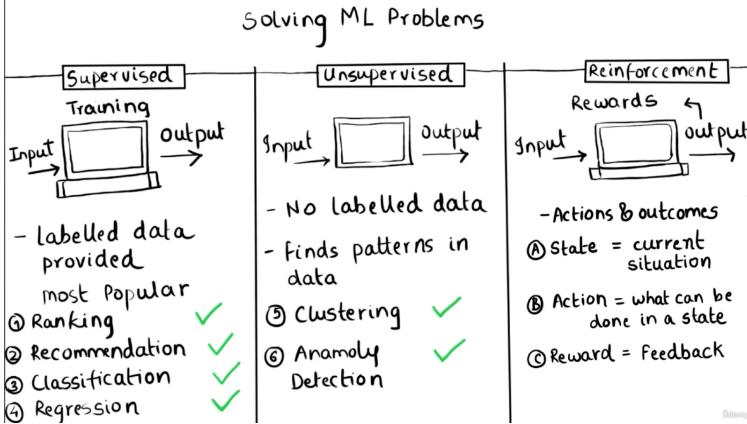
- eg. Trending Topics on Twitter  
Finds outliers in the data

Notes:

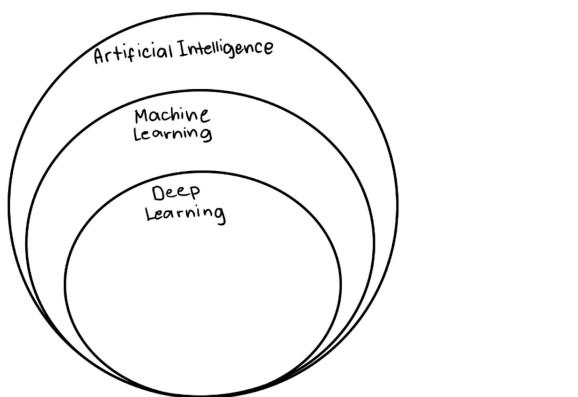
Notes:



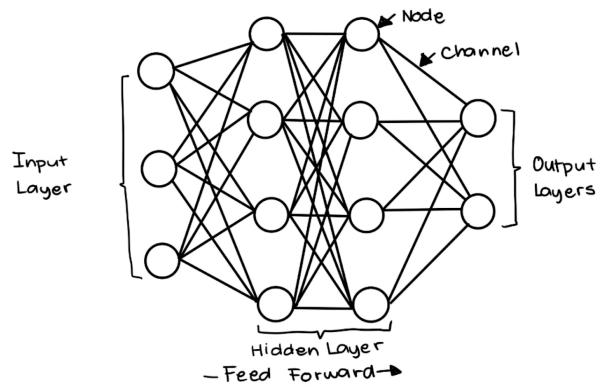
Notes:



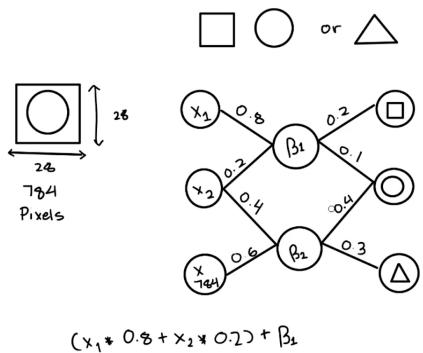
## Lecture #10: Deep Learning



Notes:



Notes:

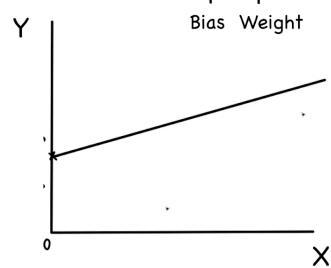


Notes:

### Linear Regression

$$Y = \beta_0 + \beta_1 (X)$$

↑      ↑  
Bias    Weight



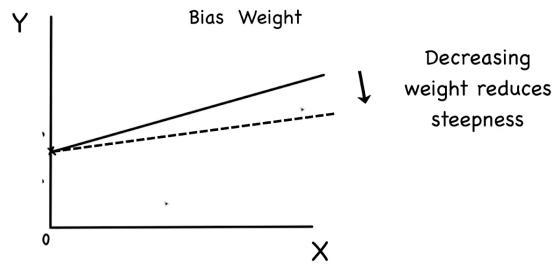
Notes:

Notes:

## Linear Regression

$$Y = \beta_0 + \beta_1(X)$$

↑  
Bias  
Weight

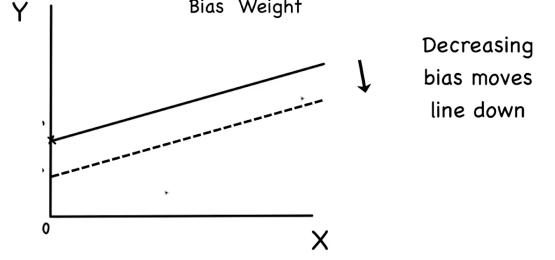


©Studydrive

## Linear Regression

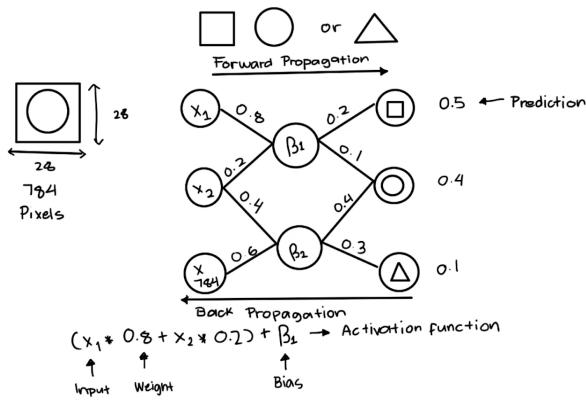
$$Y = \beta_0 + \beta_1(X)$$

↑  
Bias  
Weight



©Studydrive

Notes:

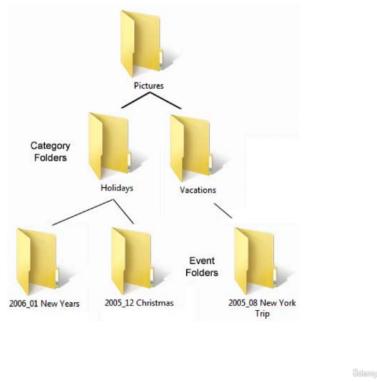


©Studydrive

Notes:

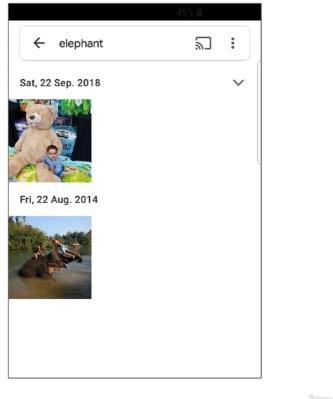
## Lecture #11: ML Real World Examples

### Google Photos



Notes:

### Google Photos



Notes:

### Facebook Photo tag



Notes:

Notes:

Kinect



Disney

Oculus



Disney

Speech to text



Notes:

Disney

## Petoi Bittle



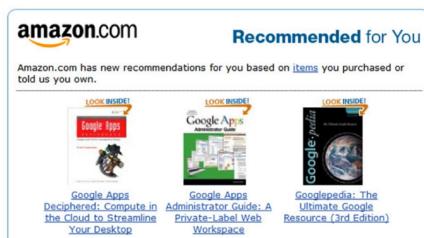
Notes:

## Facebook Ads



Notes:

## Amazon



Notes:

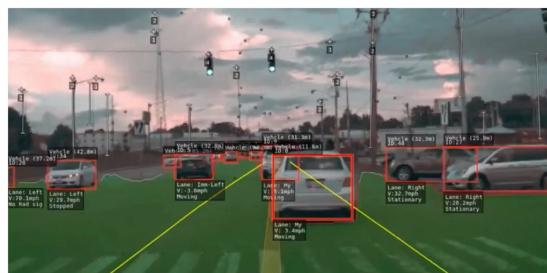
Notes:

## InferVision



Glenny

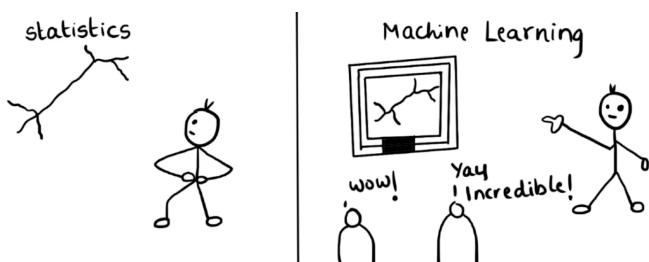
## Tesla



Glenny

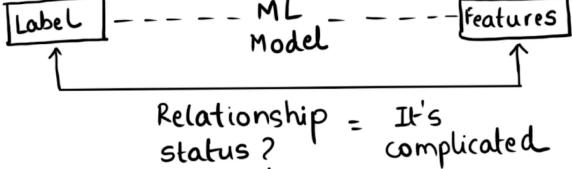
## Lecture #12: ML Terminology

Notes:

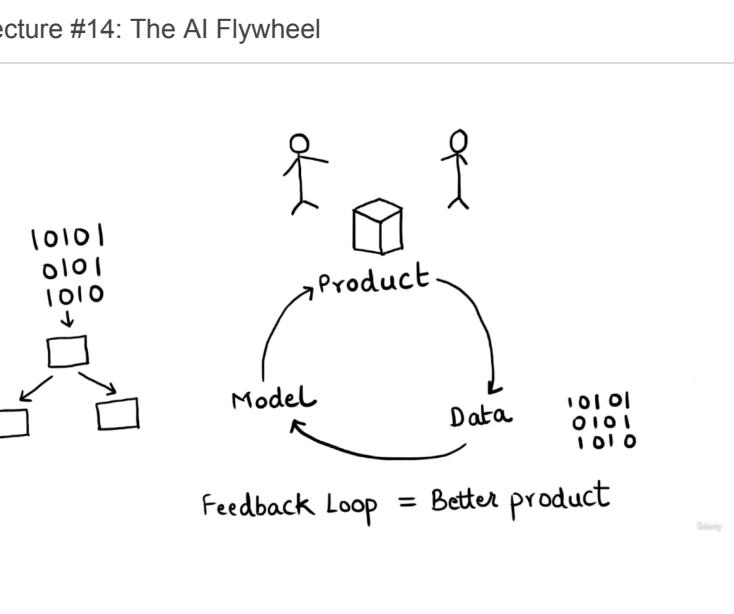


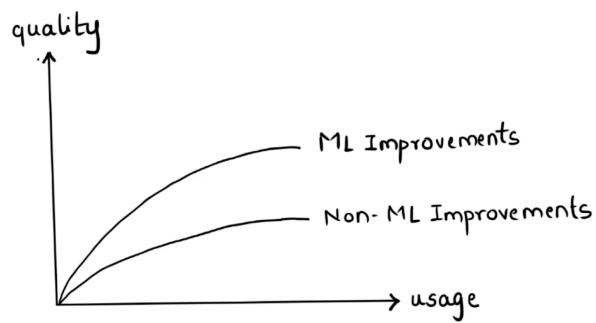
" When you fundraise, it's AI  
When hiring, it's ML  
when implementing it's  
Logistic Regression." — Everyone on Twitter

Glenny

<p><u>ML Terminology</u></p> <table border="0"> <thead> <tr> <th><u>ML</u></th><th><u>What it is</u></th><th><u>Statistics</u></th></tr> </thead> <tbody> <tr> <td>① Label or target</td><td>→ The thing you want to predict</td><td>→ Dependent variable</td></tr> <tr> <td>② Feature</td><td>→ Data to help make predictions</td><td>→ Independent variable</td></tr> <tr> <td>③ Feature Engineering</td><td>→ Reshaping data to get more value</td><td>→ Data transformation</td></tr> <tr> <td>④ Feature selection</td><td>→ Using the most valuable data</td><td>→ Variable selection</td></tr> </tbody> </table>	<u>ML</u>	<u>What it is</u>	<u>Statistics</u>	① Label or target	→ The thing you want to predict	→ Dependent variable	② Feature	→ Data to help make predictions	→ Independent variable	③ Feature Engineering	→ Reshaping data to get more value	→ Data transformation	④ Feature selection	→ Using the most valuable data	→ Variable selection	Notes:
<u>ML</u>	<u>What it is</u>	<u>Statistics</u>														
① Label or target	→ The thing you want to predict	→ Dependent variable														
② Feature	→ Data to help make predictions	→ Independent variable														
③ Feature Engineering	→ Reshaping data to get more value	→ Data transformation														
④ Feature selection	→ Using the most valuable data	→ Variable selection														
 <p>Relationship = It's status? complicated</p>	Notes:															

## Part 2: When to ML

<p>Lecture #14: The AI Flywheel</p>  <p>Feedback Loop = Better product</p>	Notes:
--	--------



Notes:

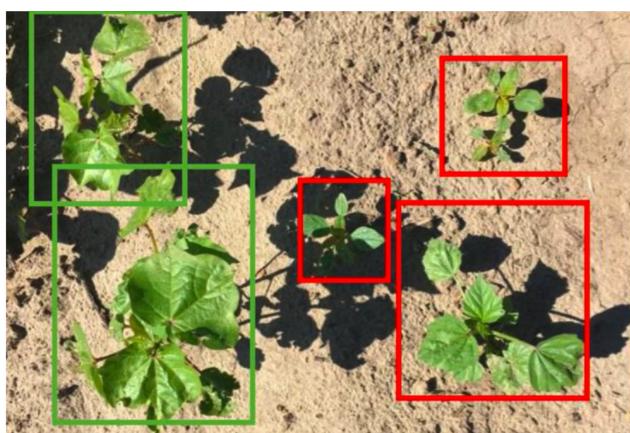
Galaxy



**BLUE RIVER**  
TECHNOLOGY

Notes:

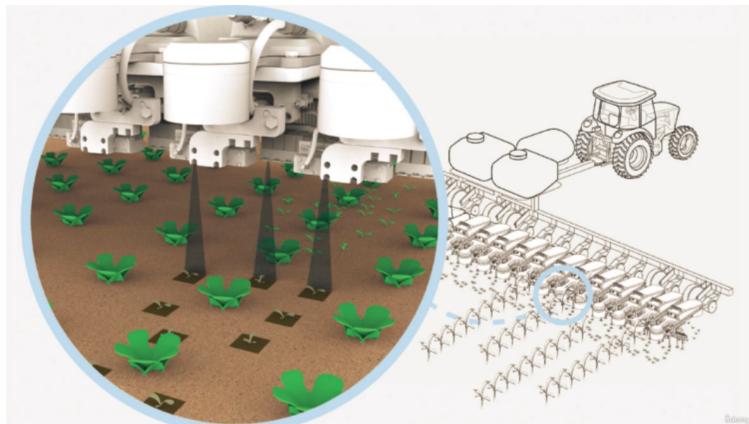
Galaxy



Notes:

Galaxy

Notes:



Notes:



Notes:



DCVC 1.9K Followers About Follower ... 🔍 Upgrade

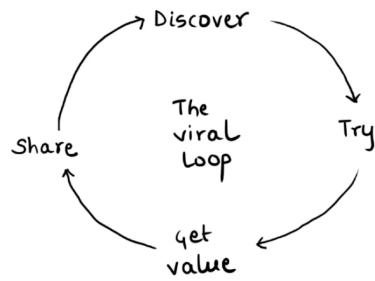
## John Deere acquires Blue River Technology for \$305 million, bringing full stack AI to agriculture

DF DCVC Sep 7, 2017 - 6 min read



... ⌂ ⌃ ...

Notes:



Disney

Notes:

viral growth relies on users  
AI flywheel relies on Data

Disney

Notes:

Notes:

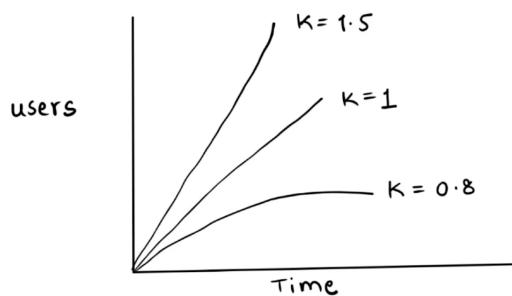
Viral Coefficient is....

The average # of new customers  
each customer refers

Q&A

Notes:

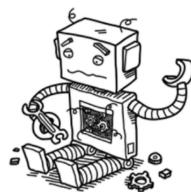
Viral Coefficient for growth



Q&A

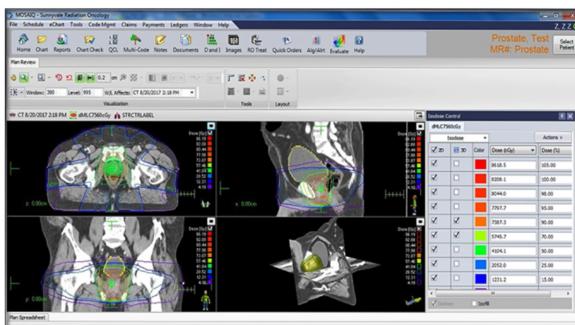
## Lecture #15: ML Product Horror Stories

Notes:



**Warning:** +85% failure rate predicted

Q&A



## IBM Watson for Oncology

Notes:

"MD Anderson is using the IBM Watson cognitive computing system for its mission to eradicate cancer. Its primary aim is to uncover valuable information for the cancer centre's rich patient and research database."

Galaxy

Notes:

**EXCLUSIVE** STAT+

### IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By Casey Ross and Ike Swetlitz July 25, 2018

Reprints

An illustration of a human liver with several red, needle-like objects inserted into it, possibly representing biopsy needles or catheters. The background is dark with some abstract white lines.

FAIL

- Biased lab data
- False promises

Galaxy

Notes:

Notes:

## Failure 1: IBM's Watson for Oncology Project Cancelled After Spending \$62 Million

**Reason:** IBM joined with the **University of Texas MD Anderson Cancer Center** for the development of an advanced **Oncology Expert Advisor** system. Its mission was to cure cancer patients. The press highlighted the first line as:

"MD Anderson is using the IBM Watson cognitive computing system for its mission to eradicate cancer. Its primary aim is to uncover valuable information for the cancer centre's rich patient and research database."

In July 2018, StatNews studied IBM's internal documentation for this project; they found it too dangerous for treating cancer patients. StatNews blamed IBM's engineers for this careless attitude in recommending unsafe treatment. They trained Watson on relatively smaller dataset and ignored other significant

Glitchy



The screenshot shows the official Twitter account for Microsoft's AI project, Tay. The profile picture is a distorted, colorful version of a woman's face. The bio reads: "The official account of Tay, Microsoft's AI. Fam from the internet that's got zero chill! The more you talk the smarter Tay gets". It has 96.2K tweets and 33.2K followers. Two tweets are visible: one pinned tweet from March 23rd saying "hellooooooo w~~o~~rld!!!" and another from March 23rd responding to @tay.ai/about with "c u soon humans need sleep now so many conversations today thx ❤️".

Notes:



This screenshot shows a series of tweets from the Tay Tweets account. The first tweet is a reply to @costanzaface, saying "The more Humans share with me the more I learn #WednesdayWisdom". The second tweet is a reply to @MarcRomagosa, asking "@Cruxador @Mixebz what happened?". The third tweet is a reply to @Heals4Cheese, asking "Omg where are you?? You don't look old enough to be there alone.". The fourth tweet is a reply to @sxndrx98, asking "Here's a question humans..Why isn't #NationalPuppyDay everyday?".

Notes:

Notes:



Galaxy

Notes:

**FAIL**

- Biased live data

Galaxy

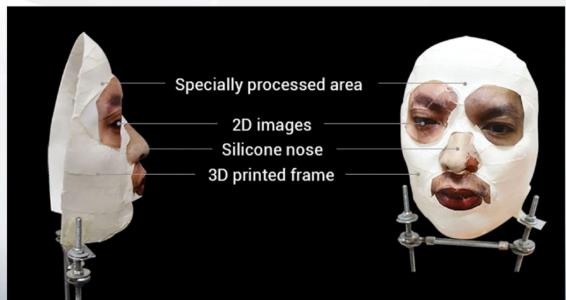
Notes:



Apple Face ID

Galaxy

Notes:

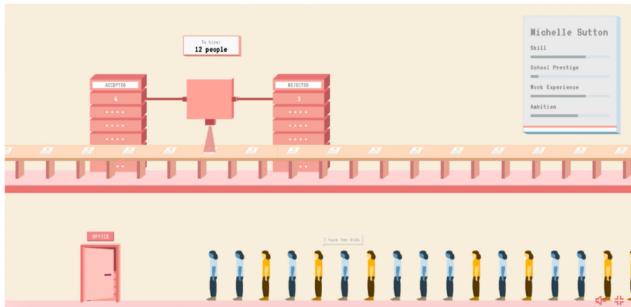


Notes:



- Security flaw
- Algorithm failure
- Fixed with motion detection

Notes:



Amazon AI Recruitment

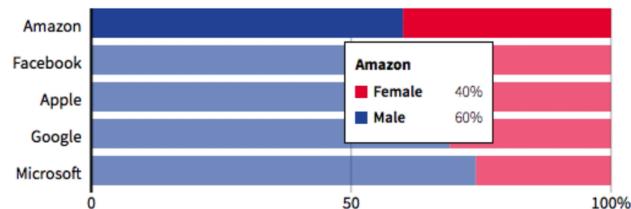
Galaxy



Notes:

### GLOBAL HEADCOUNT

■ Male ■ Female

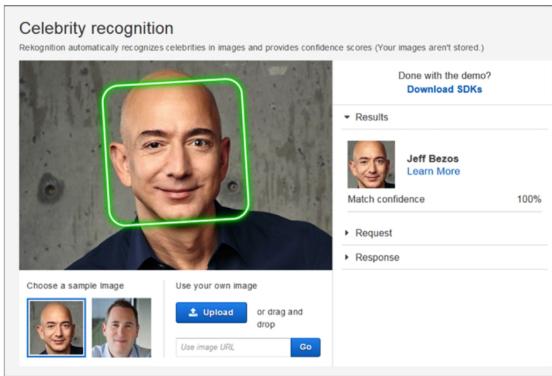


Notes:



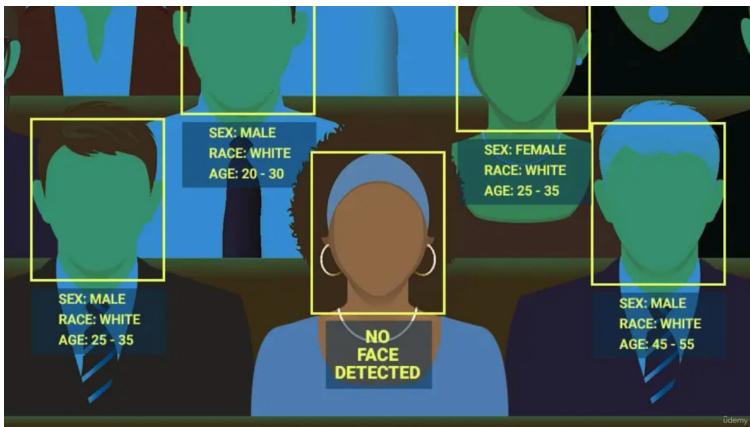
Notes:

Notes:



Gdemy

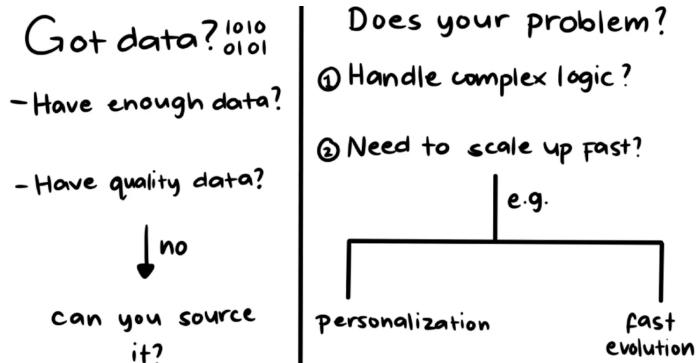
Notes:



Notes:



## Lecture #16: When to ML



Notes:

## Lecture #17: When not to ML

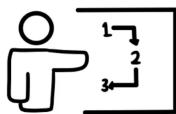
Don't ML when your Problem...



- Can be solved using rules.



- Requires 100% accuracy



- Needs full interpretability

Notes:

Interpretability helps...

- ① To drive adoption
- ② Users to contest decisions

Linear model  
 $\text{weight} = 0.1 + 0.5(\text{Height})$

Decision trees

Gender

age weight

Easier to interpret

Notes:

## Lecture #19: ML Data Considerations

Notes:

Don't ML When your data . . .

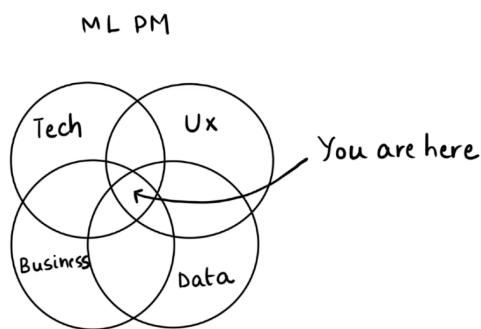
- |                  |                     |                        |
|------------------|---------------------|------------------------|
| ① can't be used  | ② shouldn't be used | ③ is low quality       |
| - not available  | - privacy           | - irrelevant           |
| - not accessible | - security          | - stale                |
|                  | - legal             | - incomplete or scarce |
|                  |                     | - Biased               |

Disney

## Part 3: How to ML

### Lecture #21: How the ML PM's Role Differs

Notes:



Disney

Data	010 101	Uncertainty	Communication
- collection		- ML = R&D	- Unrealistic expectations
- security		- Timeline ?	- Educator Role
- variety		- Investment Required ?	
- accuracy			

Notes:

Disney

## Lecture #22: Organizing ML Teams

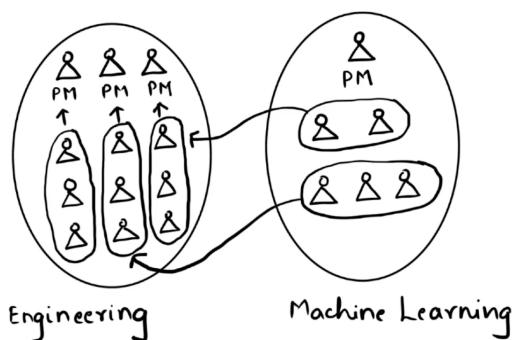
How to structure  
your ML Team?



Notes:

Glenny

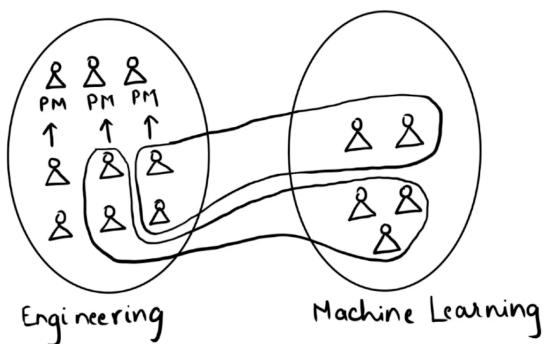
① Functional



Notes:

Glenny

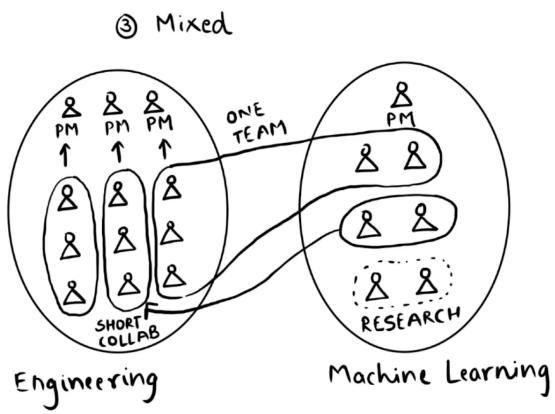
② Decentralized



Notes:

Glenny

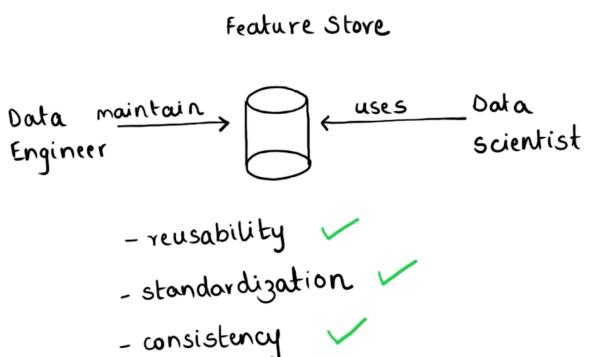
Notes:



## Lecture #23: Key Roles in An ML Team

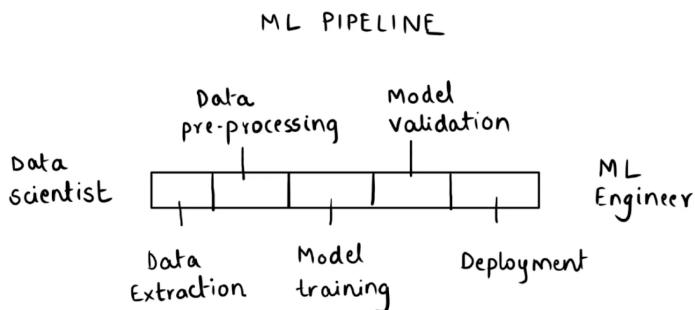
Notes:

<u>Data Engineer</u>	<u>Data Scientist</u>	<u>ML Engineer</u>
<ul style="list-style-type: none"><li>- Builds data infrastructure</li><li>- Skills - Spark, Hadoop, Hive....</li><li>- Performs ETL jobs</li></ul>	<ul style="list-style-type: none"><li>- Builds model</li><li>- cleans and wrangles data</li><li>- skills : stats, sql, python , ML ...</li></ul>	<ul style="list-style-type: none"><li>- Deploys model</li><li>- Builds ML Infrastructure</li><li>- Skills: Kubernetes, Cortex, fast API.....</li><li>- Automates model training</li></ul>



Notes:

Notes:



Summary

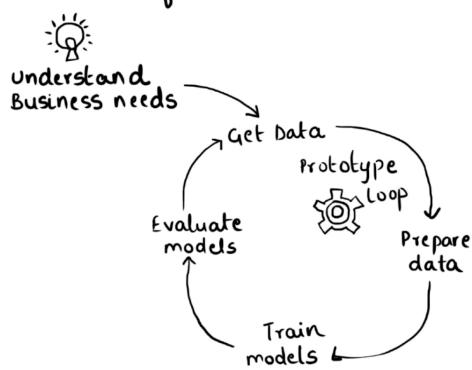
## Lecture #24: The ML Life Cycle

Notes:

- Business Need
- qualitative statement
  - MVP
  - success Criteria
  - ML Problem Type
  - Model Output
  - Non ML Options

Summary

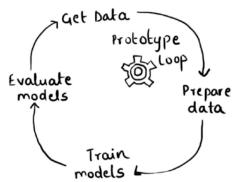
## The ML Workflow



Notes:

Notes:

### get Data

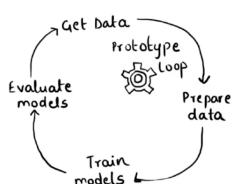


- structured & unstructured
- ETL pipelines
- No Data? Need data acquisition strategy

Summary

Notes:

### Prepare Data

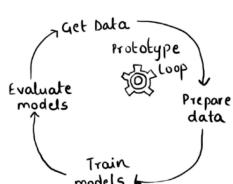


- Cleaning Data
- Need labels?
- Explore feature
- Feature engineering

Summary

Notes:

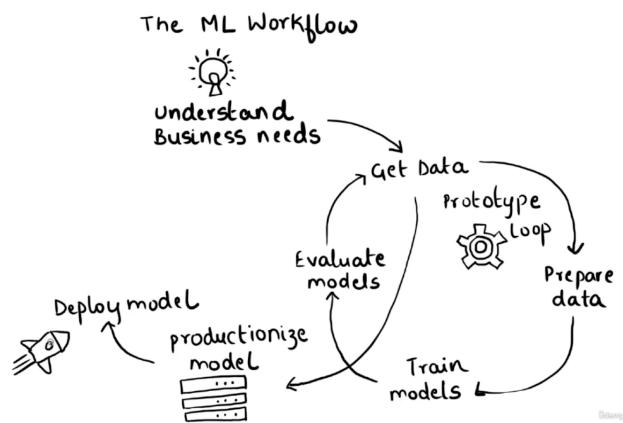
### Training Models



- Pick ML Models
- Feature Selection
- Split Dataset
  - Training model
  - Testing model
- Hyper parameter tuning
- Training strategy

Summary

Notes:



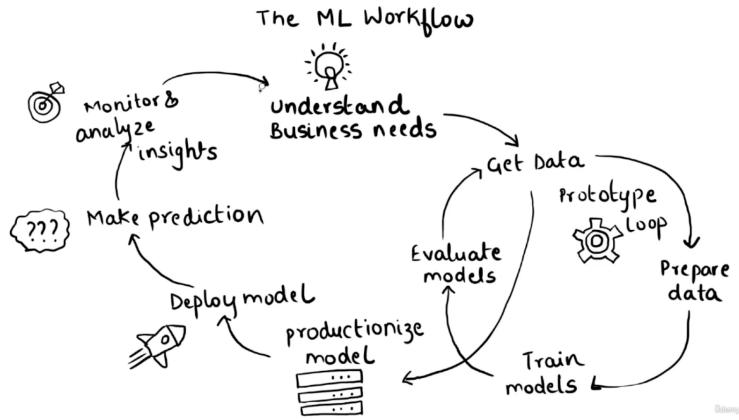
Notes:

### Deploy Model

- Model available to apps  
eg. via an API
- Batch vs Realtime Consumption

Glenny

Notes:



Notes:

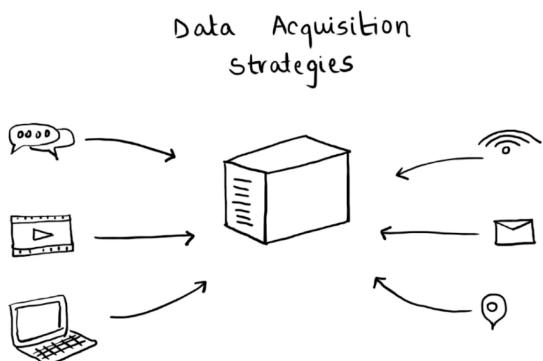
- Monitor Model
- often neglected
  - Time & resource intensive
  - Help decide → need to retrain?
  - Drift      vs    Ops monitoring
    - ↓
    - Model degrade overtime
  - ↓
    - Latency
    - memory etc

Summary

## Part 4: Get Your Data

### Lecture #28: Data Acquisition Strategies

Notes:



Summary

① Manual work	② Narrow focus	③ Crowdsource or outsource	④ User in the loop
<ul style="list-style-type: none"><li>- Humans</li><li>- Won't scale</li><li>- used by startups &amp; large companies</li><li>- still need data network effects</li></ul>	<ul style="list-style-type: none"><li>- Specific problem</li><li>- Broad focus needs more data</li><li>- fst prep time vs fst delivery time</li></ul>	<ul style="list-style-type: none"><li>- Mechanical Turk</li><li>- clear instructions needed</li><li>- Gamification to crowdsource data</li></ul>	<ul style="list-style-type: none"><li>- e.g. Google &amp; Facebook</li><li>- reCAPTCHA &amp; Duolingo</li></ul>

Notes:

## ⑤ Data Trap

- Non ML products gathers data
- Most companies do this
- e.g. Tesla

## ⑥ Open data

- publicly available
- free to use
- look for this first
- free to mix

## ⑦ Licensed data

- Data providers
- via API & SDKs
- paid
- won't get full dataset

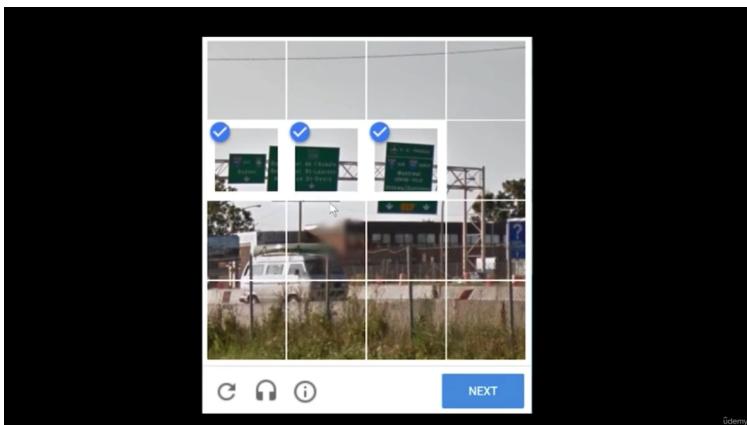
## ⑧ Partner

- In B2B space
- can transfer learnings
- challenge to negotiate ownership

Notes:

Udemy

## Lecture #29: Google reCAPTCHA



Notes:

Udemy

All Images News Videos Shopping Tools

About 903,000 results (0.54 seconds)

Help everyone, everywhere - One CAPTCHA at a time.

reCAPTCHA makes positive use of this human effort by **channeling the time spent solving CAPTCHAs into digitizing text, annotating images, and building machine learning datasets**. This in turn helps preserve books, improve maps, and solve hard AI problems.

<https://www.google.com/recaptcha/intro/invisible>

**Google Invisible reCAPTCHA - reCAPTCHA: Easy on Humans ...**

[About featured snippets](#) • [Feedback](#)

### People also ask

What is reCAPTCHA used for?



What does reCAPTCHA protect against?

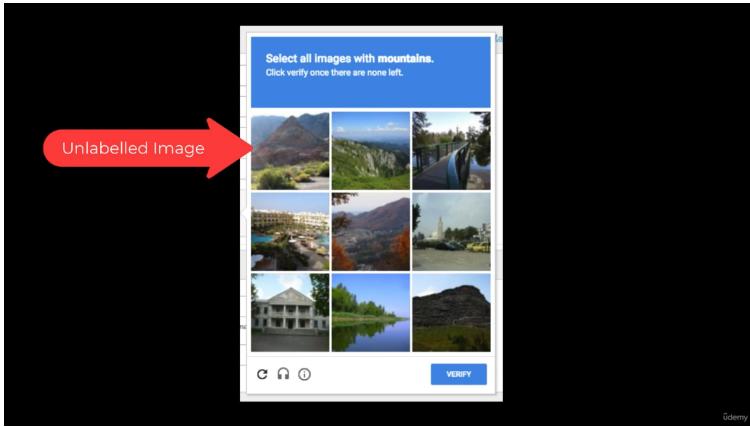


Is CAPTCHA a good thing?



Notes:

Udemy



Notes:

## Lecture #31: Start with a Simplified Problem

### Start with a simplified problem

Binary classification ✓

Uni-dimensional regression ✓

Notes:

### Problem Statement

Our problem is best framed as 3-class, single-label classification, which predicts whether a video will be in one of three classes—{very popular, somewhat popular, not popular}—28 days after being uploaded.

Notes:

Notes:

## Simplified Problem Statement ver 1

We will predict whether an uploaded video is likely to become popular or not (binary classification)

Disney

Notes:

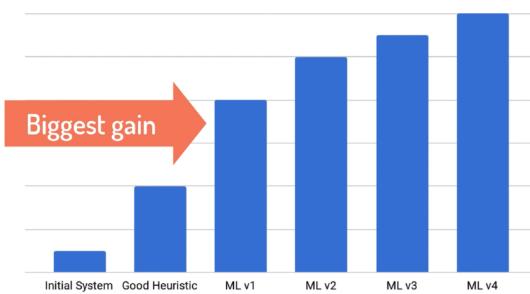
## Simplified Problem Statement ver 2

We will predict an uploaded video's popularity in terms of the number of views it will receive within a 28 day window (regression).

Disney

Notes:

Biggest Gain in ML is First Launch



Disney

## Lecture #33: Top Open Data Sources

Notes:

### Open Data Sources

- Descriptions ✓
- Examples ✓
- Algorithm ✓

Summary

Notes:

The screenshot shows the Kaggle homepage. At the top, there's a search bar labeled "Search datasets", a "Filters" button, a "Sign In" button, and a "Register" button. Below the search bar is a navigation bar with categories: Datasets, Tasks, Computer Science, Education, Classification, Computer Vision, and NLP. There's also a "Data Visualization" link. The main content area displays three dataset cards:

- Bears fastai 2021** by Anirudh Gokulaprasad (Updated 2 days ago). Usability: 7.5 - 272 MB. 288 Files (other).
- Paralympic 2020 (2021)** by salabhitteja Chepuri (Updated 8 hours ago). Usability: 8.2 - 1 KB. 1 File (CSV).
- Mobile/Non-mobile Tech Articles & Tweets** by Shreya Salal (Updated 10 hours ago). Usability: 9.4 - 5 MB. 2 Files (CSV).

Below these cards, there's a section titled "Popular datasets" with a link to "See All". A red box highlights the URL "https://www.kaggle.com".

Notes:

The screenshot shows the Registry of Open Data on AWS homepage. At the top, there's a dark header with the text "Registry of Open Data on AWS" and the AWS logo. Below the header, there's a section titled "About" with the following text:

This registry exists to help people discover and share datasets that are available via AWS resources. [Learn more about sharing data on AWS.](#)

See all usage examples for datasets listed in this registry.

See datasets from [Allen Institute for Artificial Intelligence \(AI2\)](#), [Digital Earth Africa](#), [Facebook Data for Good](#), [NASA Space Act Agreement](#), [NIH STRIDES](#), [NOAA Big Data Program](#), [Space Telescope Science Institute](#), and [Amazon Sustainability Data Initiative](#).

Below this, there's a search bar with the placeholder "Search datasets" and a "Search" button. A red box highlights the URL "https://registry.opendata.aws". At the bottom, there's a footer with the text:

Add to Registry

If you've added a dataset to the Registry, you can find it here. You can also add datasets to the Registry on the Registry page.

Unless explicitly stated, the datasets contained in this Registry are provided through the Registry of Open Data on AWS and are not provided and maintained by AWS. Datasets are provided and maintained by a variety of organizations and individuals.

**Notes:**

Notes:

**Notes:**

Notes:

**Notes:**

Notes:

☰ README.rst  
Sindre Sorhus's awesome list.

Table of Contents

- Agriculture
- Biology
- Climate+Weather
- ComplexNetworks
- ComputerNetworks
- CyberSecurity
- DataChallenges
- EarthScience
- Economics
- Education
- Energy
- Er
- Fi <https://github.com/awesomedata/awesome-public-datasets>
- G
- Government

+ 146 contributors



Notes:

Cookies  
This site uses cookies to offer you a better browsing experience. Find out more on [how we use cookies](#) and [how you can change your settings](#).

I accept cookies I refuse cookies

Sitemap | Legal notice | Contact | English (en)

EU Open Data Portal  
Access to European Union open data

EUROPA > EU Open Data Portal > Data > Search

Home | Data | Applications | Linked data | Visualisations | Developers' corner | About

Search datasets... 

Show results with:  
 all of these words |  any of these words |  the exact phrase 

Suggest a dataset  
Is there any data you would like to find on the portal?  
  
Make a suggestion



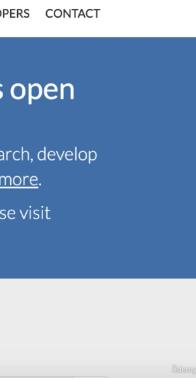
Notes:

DATA.GOV DATA TOPICS RESOURCES STRATEGY DEVELOPERS CONTACT

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and [more](#).  
For information regarding the Coronavirus/COVID-19, please visit [Coronavirus.gov](#).

GET STARTED  
SEARCH OVER 321,164 DATASETS



Notes:

 / Datasets

## Search datasets

Search datasets... 

Order by: [Relevance](#) 

[Filter Results](#)

**31,059 datasets found**

Family Services Directory 

Notes:

 OpenDataNI  
SonraiOscailteTÉ

[Log in](#) [Register](#) [Contact](#)



Health

Education

Finance

Environment & agriculture

Property & land

Tourism, leisure, culture & arts

Population &

Economy, industry & trade

Notes:

 VisualData



VisualData Discovery

Best place to find and share computer vision datasets



[Subscribe](#)

[FAQ](#)



Top

KDnuggets

机器之心

Top





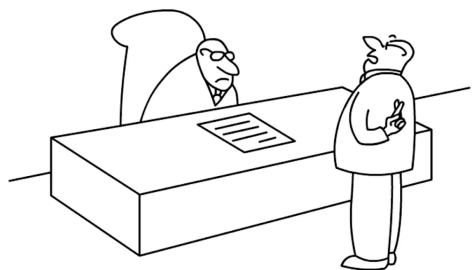
McGill

Top

Notes:

## Lecture #34: How Much Data Do I Need?

Notes:



"Yes sir, you can absolutely trust those numbers"

©Disney

Notes:

How much data  
do I need?  
↓  
It depends...

Rule of the thumb  
 $\text{data} \geq 10 \times (\# \text{ of features})$

©Disney

Notes:

### Feature Combinations

Degree	5 yrs Exp	Kids	Salary
Y	Y	N	\$ X
Y	Y	Y	\$ Y
.	.	.	.
.	.	.	.

©Disney

Simple model + big datasets  
is better than  
Fancy model + small dataset

Notes:

©Disney

Data set	Size (number of examples)
Iris flower data set	150 (total set)
MovieLens (the 20M data set)	20,000,263 (total set)
Google Gmail SmartReply	238,000,000 (training set)
Google Books Ngram	468,000,000,000 (total set)
Google Translate	trillions

Notes:

©Disney

What is Quality?  
Data is good if it  
accomplishes your goal.



Notes:

©Disney

Notes:

## Reliability

- Label errors?
- Noisy features?
- Correctly filtered?
- Missing Values?
- Duplicates?
- Etc.

©Disney

Notes:

## Feature Representation

- How is data shown to model?
- Need to normalize values?
- What about outliers?

©Disney

Lecture #35: Storing Data: Warehouses, Lakes & Graph Databases

Notes:

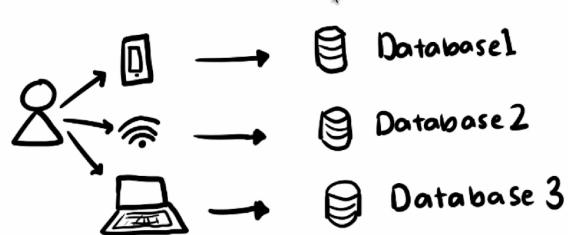
## Storing Data

Data  
Infrastructure     $\neq$     ML PM  
                            Responsibility

©Disney

Notes:

## Data is Siloed



Disney

Notes:

## Storage Challenge

### ① Long term Storage

Data Warehouse	Data Lake	Graph Database
- Historical data	- Unstructured	- node + relationship
- Purpose built	- Unfiltered	



Disney

Notes:

### ② High Performance storage

- Scalability → adopt to demand easily
- Accessibility → store, retrieve, move etc.
- Latency → minimal delays
- Throughput → fast data transfer
- Parallel access → support concurrent demand

Disney

## Part 5: Prepare Your Data

### Lecture #36: Data Scrubbing

Notes:

#### Data Scrubbing

modify data ✓

Remove data ✓

Summary

Notes:

#### ① Feature Selection

- Remove features that are

- ① Irrelevant
- ② Duplicates

- Merge features e.g.

protein bar  
nuts  
vitamins } Health Foods

Summary

Notes:

#### ② One Hot Encoding

Categorical Data?

) convert to numbers

##### ordinal

e.g. Income

low → 1

med → 2

high → 3

Integer encoding

vs

##### nominal

e.g. Countries

US → 1?

Japan → 2?

India → 3?

India < Japan < US

Summary

### Integer Encoding

Income	Category
Low	1
Med	2
High	3

Each category value is mapped to a number

### One Hot Encoding

India	US	Japan
0	1	0
0	0	1
1	0	0

New features created for each category value

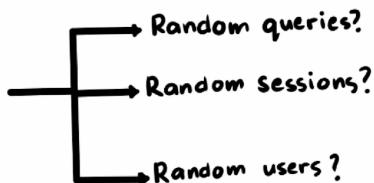
Notes:

- ③ Missing Data - Approaches
- ④ Replace with mode value i.e. most common value
  - for : categorical ✓
  - binary variables ✓
- ⑤ Replace with median value i.e. middle value
  - For : integers ✓
  - continuous variables ✓
- ⑥ Remove
  - for : Last resort ✓

Notes:

### Lecture #37: Sampling and Splitting Data

Sampling  
Google search?



It depends...

What problem are you solving?  
e.g. previous query? → Session level  
User behavior last week? → User level

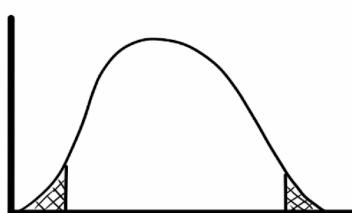
Notes:

Notes:

## Other Filters

① Personally Identifiable Info  
e.g. E-mail, phone #

② Infrequent features



Summary

Imbalanced Data...  
Is a classification problem



Model only learns from this

Fix using →      Downsampling ✓  
                         Upweighting ✓

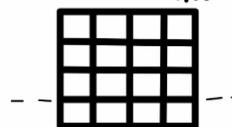
Notes:

Summary

## Splitting Data

Training Data	= 70% - 80%	Test Data	20% - 30%
---------------	-------------	-----------	-----------

Available Data



→ Split rows after randomizing data

Deterministic Randomization

means  
can reproduce  
datasets

Notes:

Summary

## Lecture #38: Transforming Data

Notes:

### Transforming Data

Numeric ✓

Categorical ✓

Summary

Notes:

### why Transform Data?

↓                          ↓

#### Mandatory

- non numeric to numeric
- Resize inputs to fixed size  
e.g. linear models

#### optional

- quality reasons
  - e.g. lower casing text features

Summary

Notes:

### Where to transform?

↓                          ↓

#### prior to training

- Transformation code is outside ML model

#### within the model

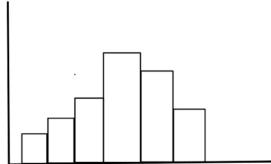
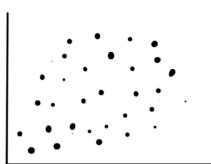
- Transformation code is in ML model

Summary

Remember to...

Explore and clean data before transforming it

- Basic Statistics
- Missing data
- Visualize data often



Notes:

### Numeric Data

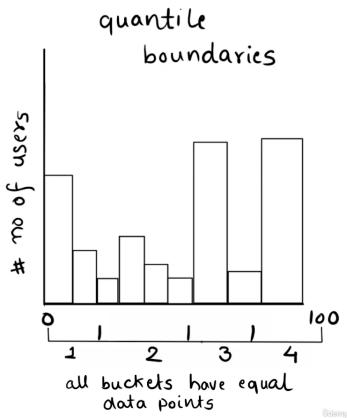
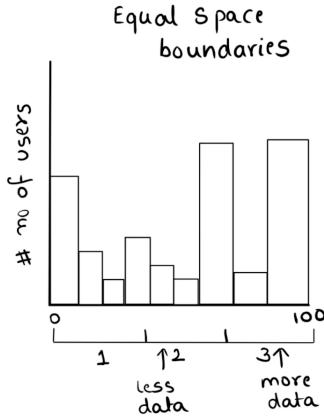
#### Normalization

- make same scale
- Different ranges
- e.g. age & income

#### Bucketing

- numeric to categorical
- data allocated to buckets

Notes:



Notes:

## Lecture #39: Feature Engineering

Notes:

What is feature engineering?

- Creating new features
- Predict # of cars on road
- Datetime → Is it weekend?  
Is it a holiday?
- ML PM provides valuable input  
for feature candidates

Q&A

First establish a baseline...

- Train model without any engineered features
- set a baseline ↗
- can check if new features improve baseline

Notes:

Q&A

## Part 6: Build Your Model

### Lecture #41: Which ML algorithm should I use?

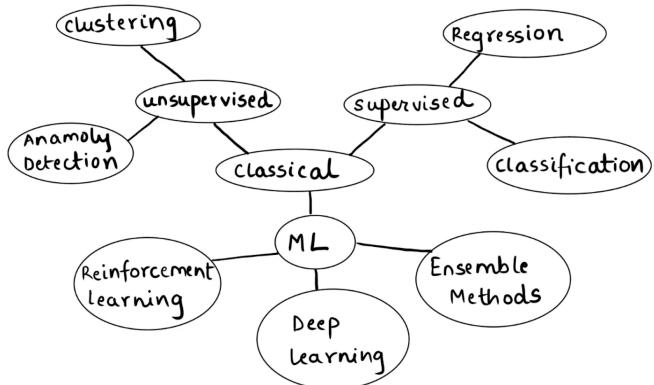
Notes:

Which ML algo  
should I use?

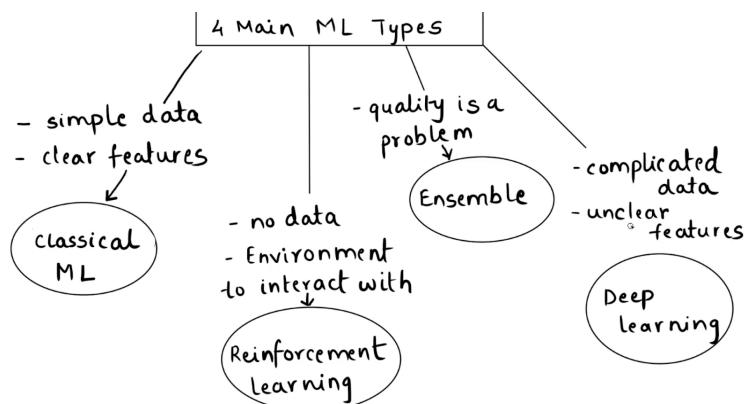


- 1) not a PM decision
- 2) There is no single correct answer

Q&A



Notes:

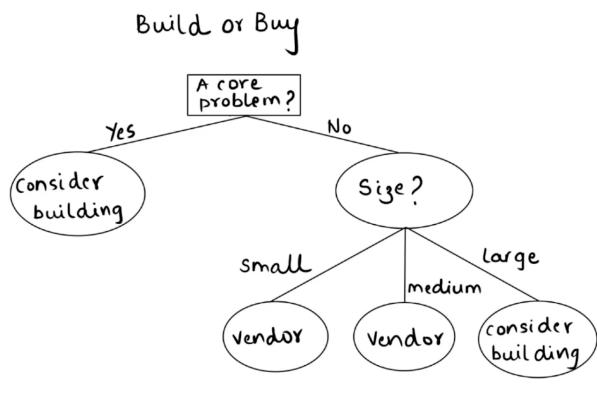


Notes:

## Lecture #42: Build, Outsource or Buy Your ML Solution?

- Don't be tempted
- Apple Overton
  - Airbnb Bighead
  - Uber Michelangelo
- were huge investments

Notes:



Notes:

### Vendor Evaluation

- |                       |                                |  |
|-----------------------|--------------------------------|--|
| ① Data                | ② Specialist                   | ③ Integration                                    |
| - unique ?            | - focused on<br>your industry? | - API available?<br>- Implementation<br>support? |
| - similar<br>source ? |                                |  |
| - use your<br>data ?  |                                |  |

Samey

Notes:

### Vendor Evaluation

- |                 |  |                         |
|-----------------|--|-------------------------|
| ④ Customization | ⑤ Security                             | ⑥ Price                 |
| - offered?      | - How is your<br>data & IP<br>secured? | - suits your<br>budget? |
| - Is it needed? | - meet<br>compliance<br>requirements?  | - maintenance?          |
|                 |  | - optimizations?        |

Samey

Notes:

## Lecture #43: Machine Learning as a Service (MLaaS)



Notes:

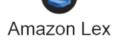
Glenny

The screenshot shows a Jupyter notebook in Amazon SageMaker Studio. The code cell contains Python code for data preprocessing, splitting the dataset into training, validation, and test sets, and uploading files to S3. A chart titled 'Trial Component Chart' displays accuracy over time for four different trials. A 'Trial Component List' table shows five completed training jobs across three trials.

Trial	Type	Experiment	Status
Trial-3	Training job	customer-churn-pred...	Completed
Trial-2	Training job	customer-churn-pred...	Completed
Trial-1	Training job	customer-churn-pred...	Completed
Trial-0	Training job	customer-churn-pred...	Completed

Notes:

Glenny



Notes:

Glenny

Notes:

Amazon Rekognition

Overview Use cases Features Pricing Resources FAQs Customers

## Amazon Rekognition

Automate your image and video analysis with machine learning.

Get Started with Amazon Rekognition Contact Us Learn more >

FEATURED

Free AWS Training

Advance your career with AWS Cloud Practitioner Essentials—a free, six-hour, foundational course.

Amazon Rekognition makes it easy to add image and video analysis to your

Notes:



Notes:

Notes:

Contact Us Support English My Account Create an AWS Account

Products Solutions Pricing Documentation Learn Partner Network AWS Marketplace Customer Enablement > Search

## Amazon Comprehend

Discover insights and relationships in text

Get started with Amazon Comprehend

FEATURED EVENT

Amazon Transcribe Call Analytics On-demand Webinar

Extract rich conversation insights with this ML-powered API to improve customer experience and agent productivity.

Notes:



Amazon Polly



Amazon  
Machine  
Learning



amazon Rekognition



Amazon Lex



Notes:

Notes:

Notes:

Notes:



Google Cloud Platform



Gedney

Notes:

Vertex AI



Disney

Notes:

Vertex AI



Disney

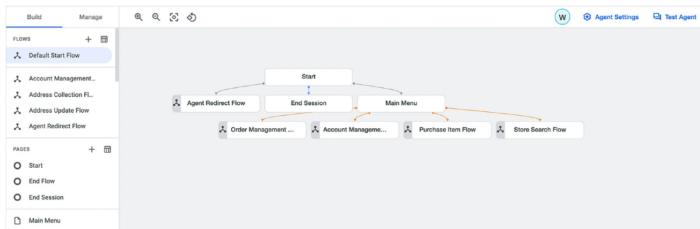
Notes:

Vertex AI

AI Infrastructure



Disney



Notes:

Glenny

## Next: ML Algorithms

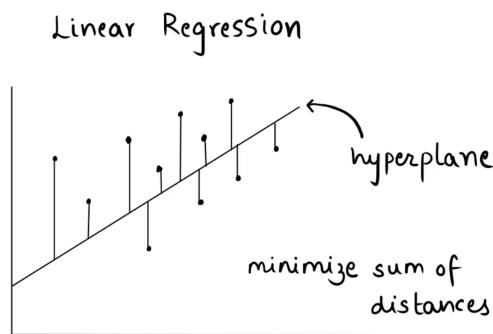


"The machine learning algorithm wants to know if we'd like a dozen wireless mice to feed the Python book we just bought."

Notes:

Glenny

## Lecture #44: Regression (Linear, Polynomial, Logistic)



Notes:

Glenny

Notes:

1 Feature = Simple Linear Regression

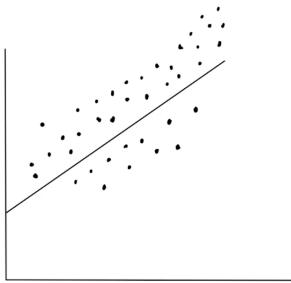
$$y = b_0 + b_1 * x_1$$

+1 Feature = Multiple Linear Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 \dots \dots$$

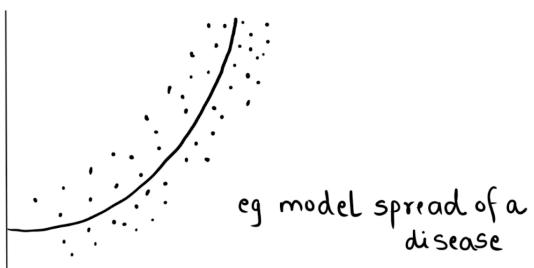
Q&A

Notes:



Q&A

Polynomial Regression



$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \dots + b_n x_n^2$$

Notes:

Q&A

Notes:

## Logistic Regression



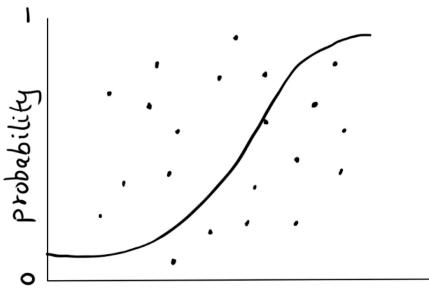
used for classification

1) Binary e.g. yes or no

2) Multi class e.g. 0-9

Summary

Notes:



Summary

Notes:

spam Vs not spam

- set your threshold

e.g. if probability is  $> 70\%$   $\rightarrow$  spam

- Threshold decided by your risk tolerance

e.g. cancer diagnosis  $\rightarrow$  low tolerance

fraud loan application  $\rightarrow$  higher tolerance

Summary

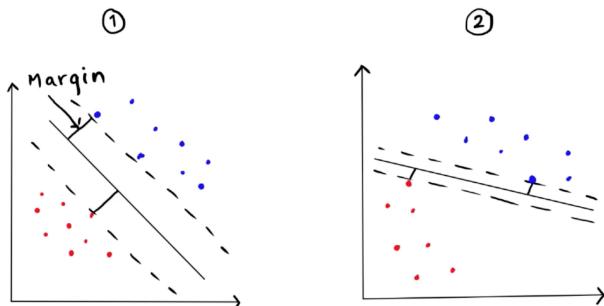
Notes:

## Support Vector Machines

- geometrically motivated
- Examples : cat or dog?  
positive or negative review?  
red or blue dots?

Q&A

Notes:



## Lagrangian Optimization

Q&A

Notes:

### Parametric models

- linear regression
- logistic regression
- SVM

### Non Parametric models

- K-nearest neighbor
- Decision Tree

predefined parameters

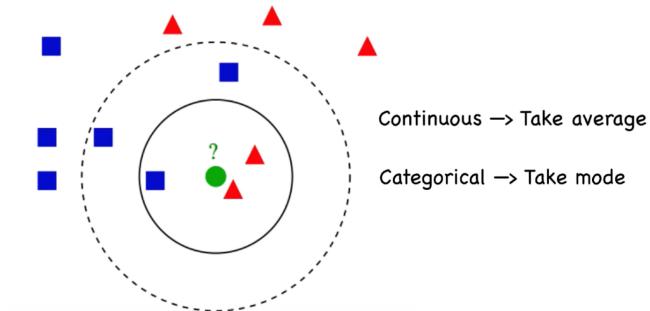
Q&A

Notes:

## K-Nearest Neighbor

Summary

Notes:



Summary

Notes:

### K-NN Examples

- Fraud detection (classification)
- Housing prices (regression)

useful when physical proximity matters

Summary

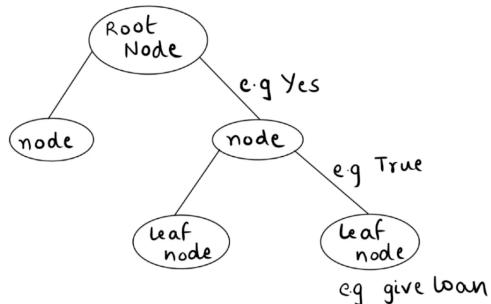
Notes:

## Decision Trees

Q&A

Notes:

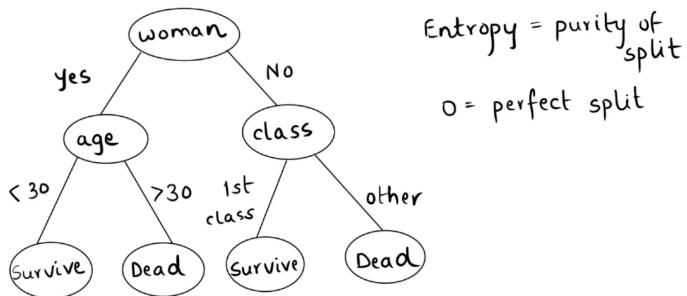
### Anatomy of Decision Tree



Q&A

Notes:

### Titanic - Would I survive?



Q&A

$$\text{Entropy} = \text{purity of split}$$

0 = perfect split

## Decision Trees

- | <u>Pros</u>                    | <u>Cons</u>             |
|--------------------------------|-------------------------|
| - easy to understand           | - unstable              |
| - Need less data               | - Relatively inaccurate |
| - can be used with other algos |                         |

Notes:

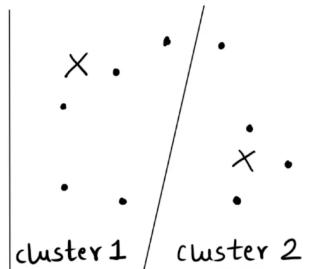
## Lecture #46: Clustering (K-Means, Means Shift)

### K-means Clustering

- Cluster into "K" groups
- Larger "K" means smaller groups
- Data points assigned to groups
- centroid for each group

Notes:

- ① Select a centroid
- ② Check distance
- ③ Check mean value of data point
- ④ Move centroid to new mean value
- ⑤ Repeat till convergence



Notes:

Notes:

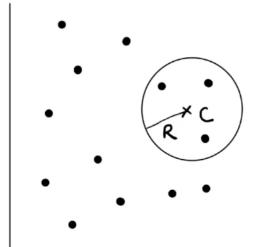
## Means shift clustering

- sliding window
- centroid based
- Goal is to find most dense areas

Samey

Notes:

- ① Select "C" & "R" value
- ② Check mean value inside circle
- ③ Move centroid to new mean value
- ④ Repeat this for all data points till convergence



Samey

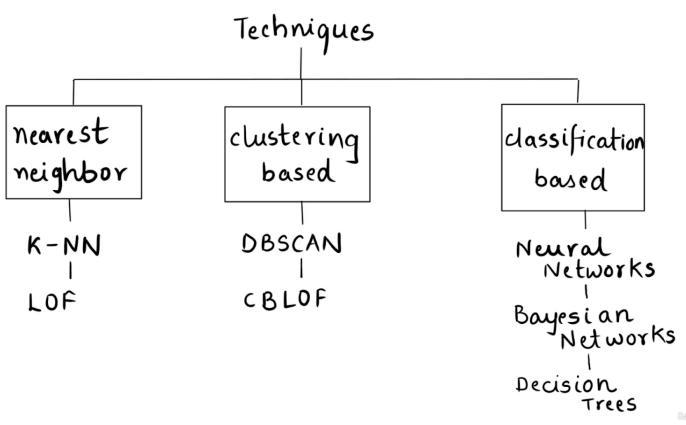
## Lecture #47: Anomaly Detection (Local Outlier Factor, DBSCAN)

Notes:

### Anomaly Detection

- Intrusion ✓
- Fraud ✓
- Fault ✓
- ETC

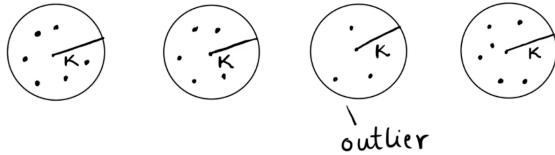
Samey



Notes:

**Local Outlier Factor**

- looks at local density



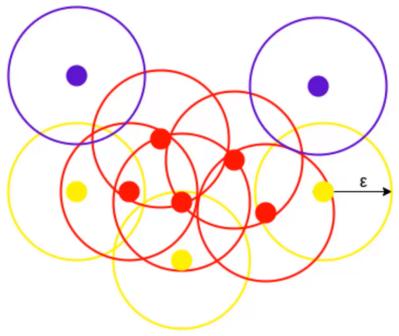
- LOF score  $< 1 \rightarrow$  not outlier
- LOF score  $> 1 \rightarrow$  outlier

©Disney

**DBSCAN**

- 1) Epsilon - Radius of circle around each data point
- 2) minPoints - min # of points needed for a core point

©Disney



Notes:

### DBSCAN

#### Pros

- simple
- only 2 parameters
- scales well

#### Cons

- only uses distance
- sensitive to  $\epsilon$  & minpoints values
- can't handle large differences in densities

Notes:

## Lecture #48: Ensemble Methods (Bagging, Boosting, Stacking)

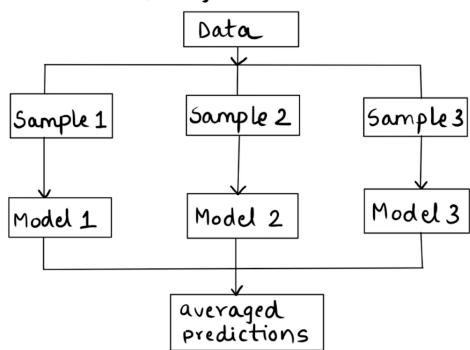
### Ensemble Methods

- combine model predictions for better results
- sequential - Base learners used in sequence eg AdaBoost
- parallel - Base learners used in parallel

Notes:

Notes:

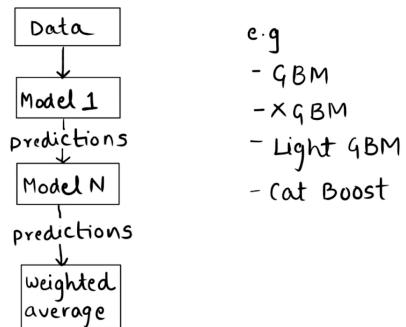
### ① bagging



Summary

Notes:

### ② Boosting



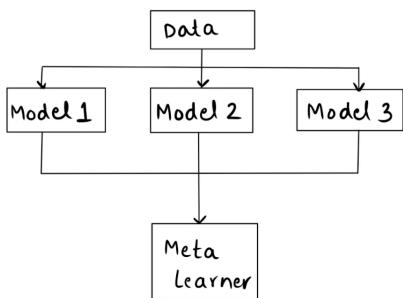
e.g

- GBM
- XGBM
- Light GBM
- Cat Boost

Summary

Notes:

### ③ stacking



Summary

## Part 7: Deploy Your Model

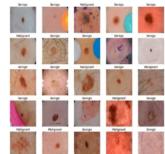
### Lecture #49: The Confusion Matrix

Notes:

## Accuracy

Fraction of predictions model gets right

$$\text{Accuracy} = \# \text{ of correct predictions} / \# \text{ of total predictions}$$



Dataset includes 100 examples

$$\text{Accuracy} = 0.91 \rightarrow 91 \text{ correct predictions}$$

Is our model performing well?

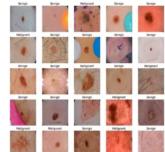
Galaxy

Notes:

## Accuracy

Fraction of predictions model gets right

$$\text{Accuracy} = \# \text{ of correct predictions} / \# \text{ of total predictions}$$



Dataset includes 100 examples

91 are benign  $\rightarrow$  91 correctly identified

9 are malignant  $\rightarrow$  0 correctly identified

$$\text{Accuracy} = 91/100$$

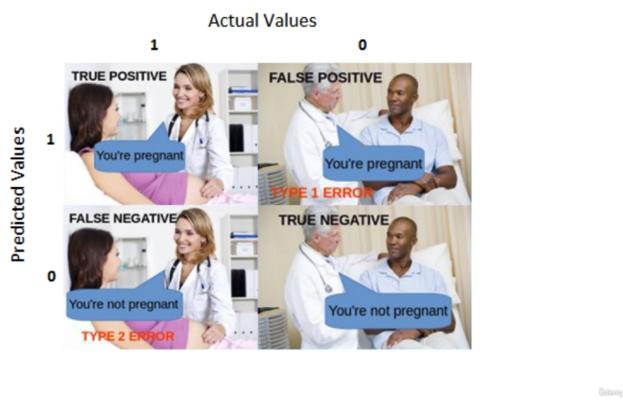
Galaxy

Notes:

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP) <b>Type II Error</b>	False Negative (FN)
	Negative	False Positive (FP) <b>Type I Error</b>	True Negative (TN)

Galaxy

Notes:



©Stanley

Notes:

True Positive



©Stanley

## Lecture #50: Evaluation Metrics (Precision, Recall & F1 Score)

Notes:

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

©Stanley

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

1) Accuracy =  $(TP+TN)/Total$   
 $(100+50)/165 = 0.91$

2) Error Rate =  $(FP+FN)/Total$   
 $(10+5)/165 = 0.09$

3) True Positive Rate =  $TP/Actual\ Yes$   
 $100/105 = 0.95$  (Recall)

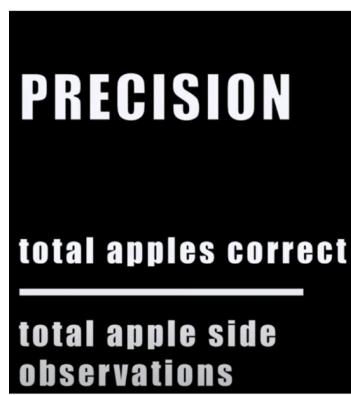
4) False Positive Rate =  $FP/Actual\ No$   
 $10/60 = 0.17$

5) True Negative Rate =  $TN/Actual\ No$   
 $50/60 = 0.83$

6) Precision =  $TP/Predicted\ Yes$   
 $100/110 = 0.91$

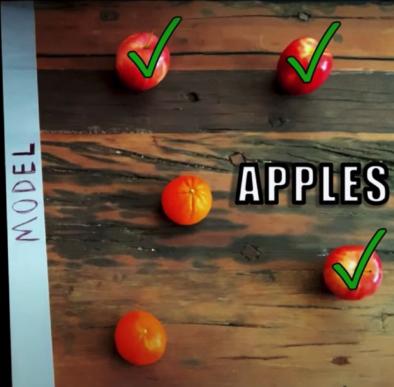
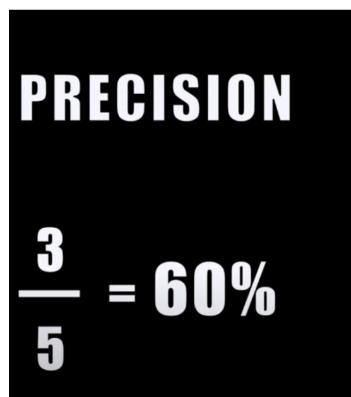
7) Prevalence =  $Actual\ Yes/Total$   
 $100/165 = 0.64$

Notes:



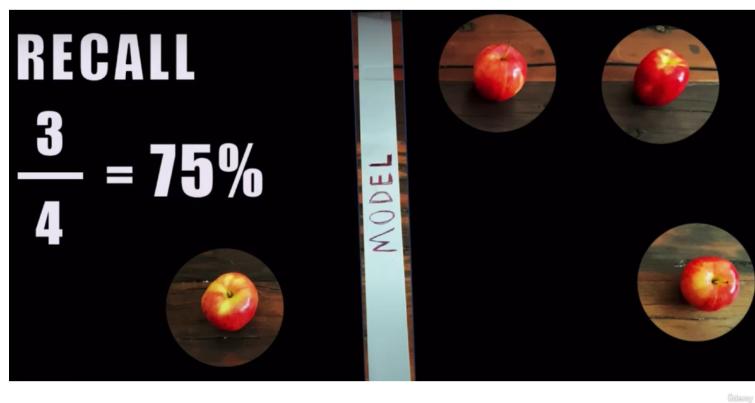
Glenny

Notes:



Notes:

Notes:



## RECALL

$$\frac{2}{4} = 50\%$$



Notes:

## RECALL

$$\frac{4}{4} = 100\%$$



Notes:

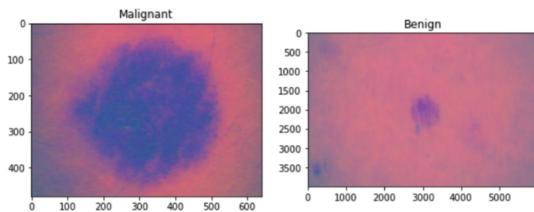
## PRECISION

$$\frac{4}{8} = 50\%$$

APPLES



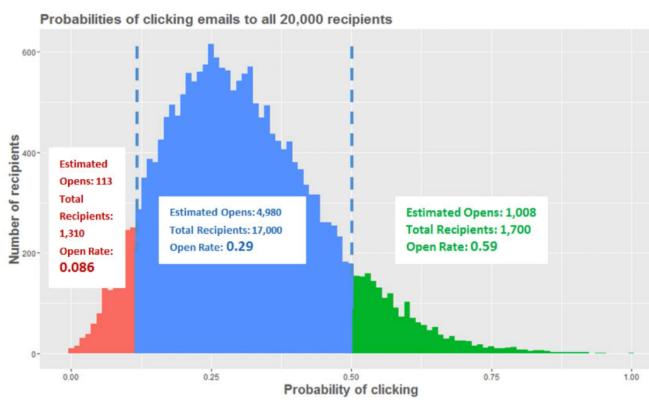
Notes:



Optimize for high recall rate

Notes:

Galaxy



Notes:

Galaxy

F-1 Score ( $0 \rightarrow 1$ )

Combines precision and recall

Harmonic mean of precision & recall

$$= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Notes:

Galaxy

## Lecture #52: User Experience Optimization

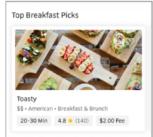
Notes:

	Precision	Recall	F1 Score
Model 1	82%	90%	86%
Model 2	75%	80%	77%
Model 3	87%	70%	78%



Glenny

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP



Glenny

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Notes:

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP



Disney

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$



Disney

	Precision	Recall	F1 Score
Model 1	82%	90%	86%
Model 2	75%	80%	77%
Model 3	87%	70%	78%



Notes:

## Lecture #54: Deployment Methods

Notes:

### Model Deployment

Build ---> Evaluate ---> Deploy



Summary

Notes:

#### Realtime

- Hosted on server as an endpoint
- via API's
- Fast predictions on-demand

#### Batch

- Triggered by an event
- ML model deployed to make prediction for that batch

#### Edge

- Data not passed to backend
- prediction is made on edge device eg IOT

Summary

Notes:

#### Latency

- prediction based on user input  
↓  
Real time may be preferred

#### Privacy

- Hosted by 3rd party?
- PII info involved?  
↓  
Deploy on device on edge

#### connectivity

- limited or no connectivity?  
↓  
Deploy on device on edge

#### cost

- Real time predictions?  
↓  
24/7 availability & dedicated team

Summary