Machine Learning 101

QCon SF 2019

Grishma Jena
Data Scientist, IBM
@DebateLover

About me

- Cross-portfolio Data Scientist with IBM Data and AI in San Francisco
- Infusing data science in UX and Design
- Background in Machine Learning and Natural Language Processing
- Love to encourage women and youngsters in tech
- Speaker and mentor
  - Started with teaching Python at San Francisco Public Library
  - Mentor for non-profit AI4ALL for teenagers
  - Spoken at PyCon, OSCON and other conferences

gjena.github.io

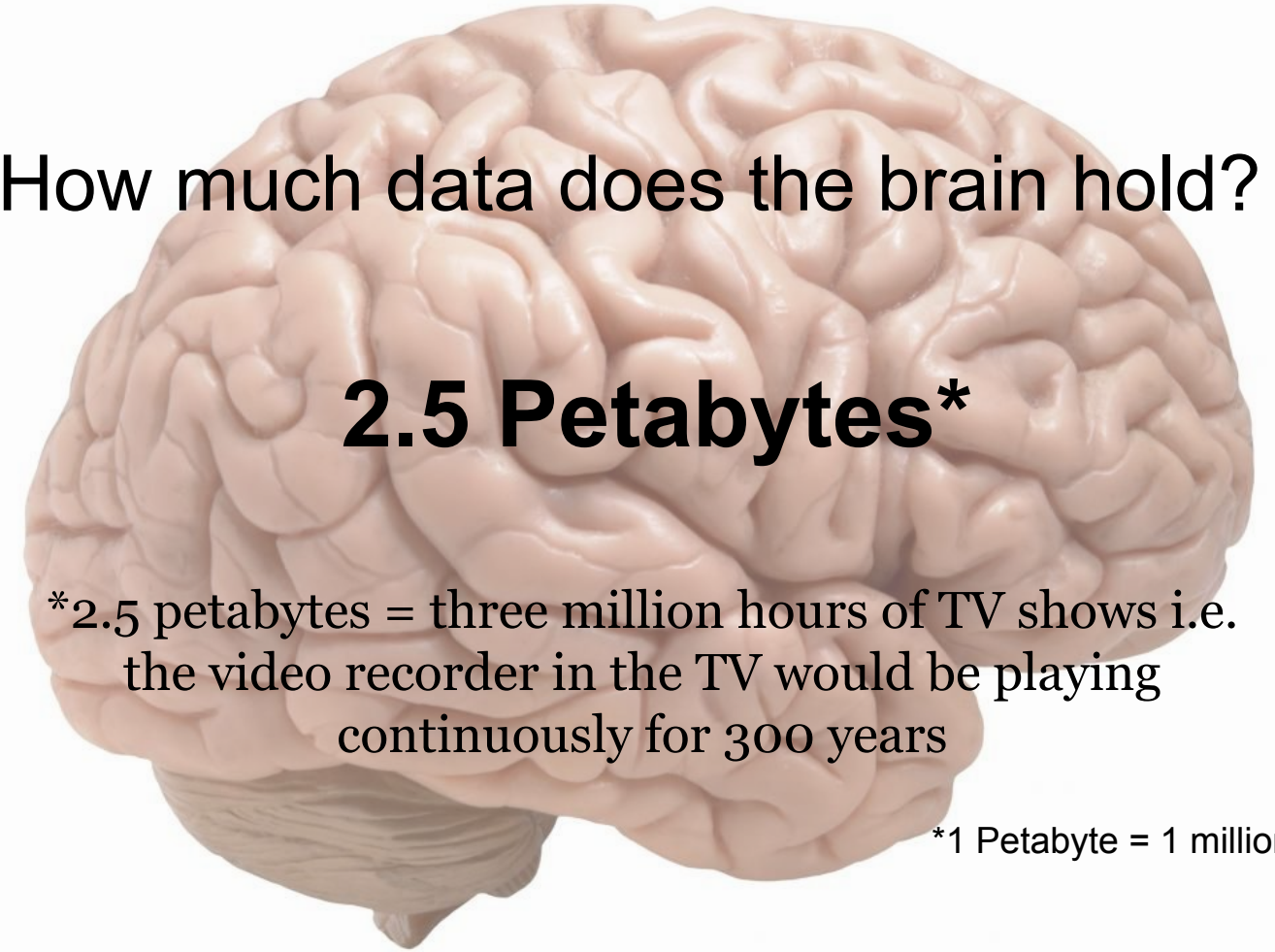
grishmajena

DebateLover

How much data is produced every year?

16.3 Zettabytes*

*1 Zettabyte = 1 trillion Gigabytes



How much data does the brain hold?

2.5 Petabytes*

*2.5 petabytes = three million hours of TV shows i.e. the video recorder in the TV would be playing continuously for 300 years

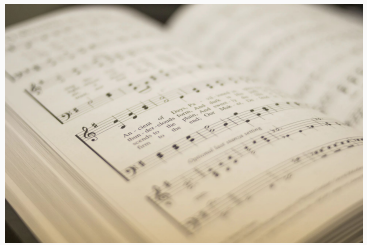
*1 Petabyte = 1 million Gigabytes

We generate more data than we realize...

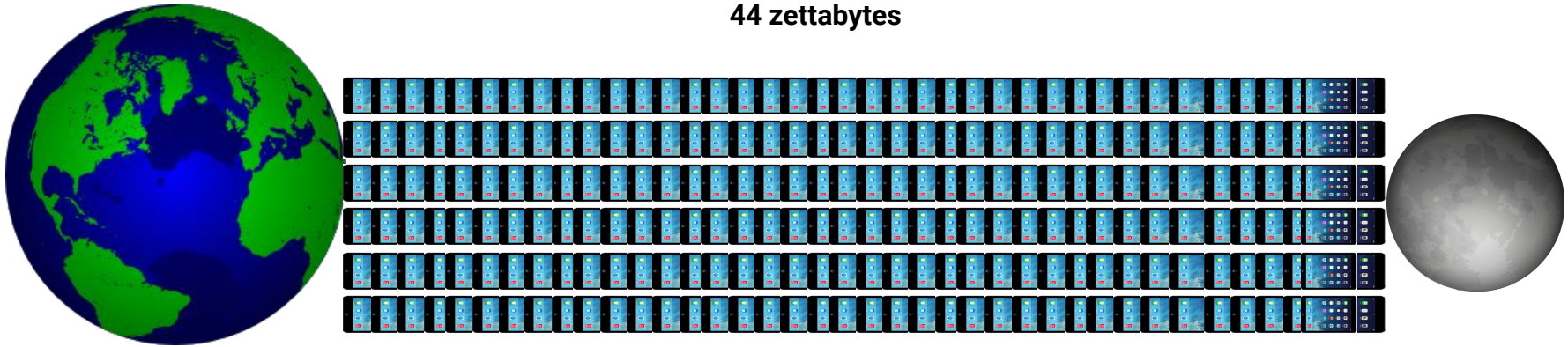
2.5
Exabytes
per day

5 million laptops

90 years HD video

150,000,000 iphones

530,000,000 million songs

44 zettabytes



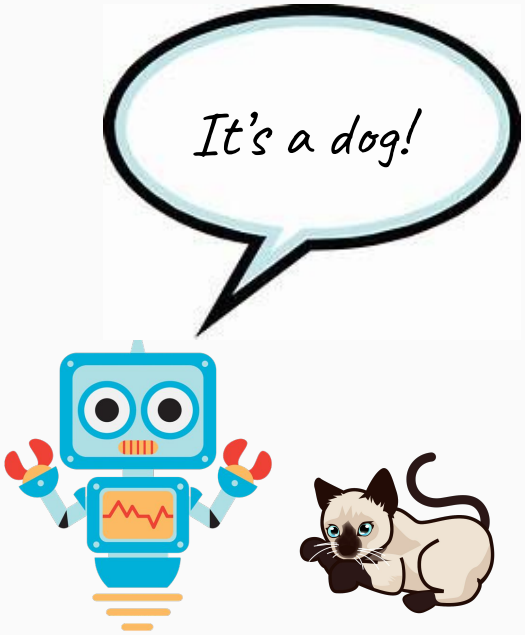
Digital Universe represented by the memory in a stack of iPad Air tablets

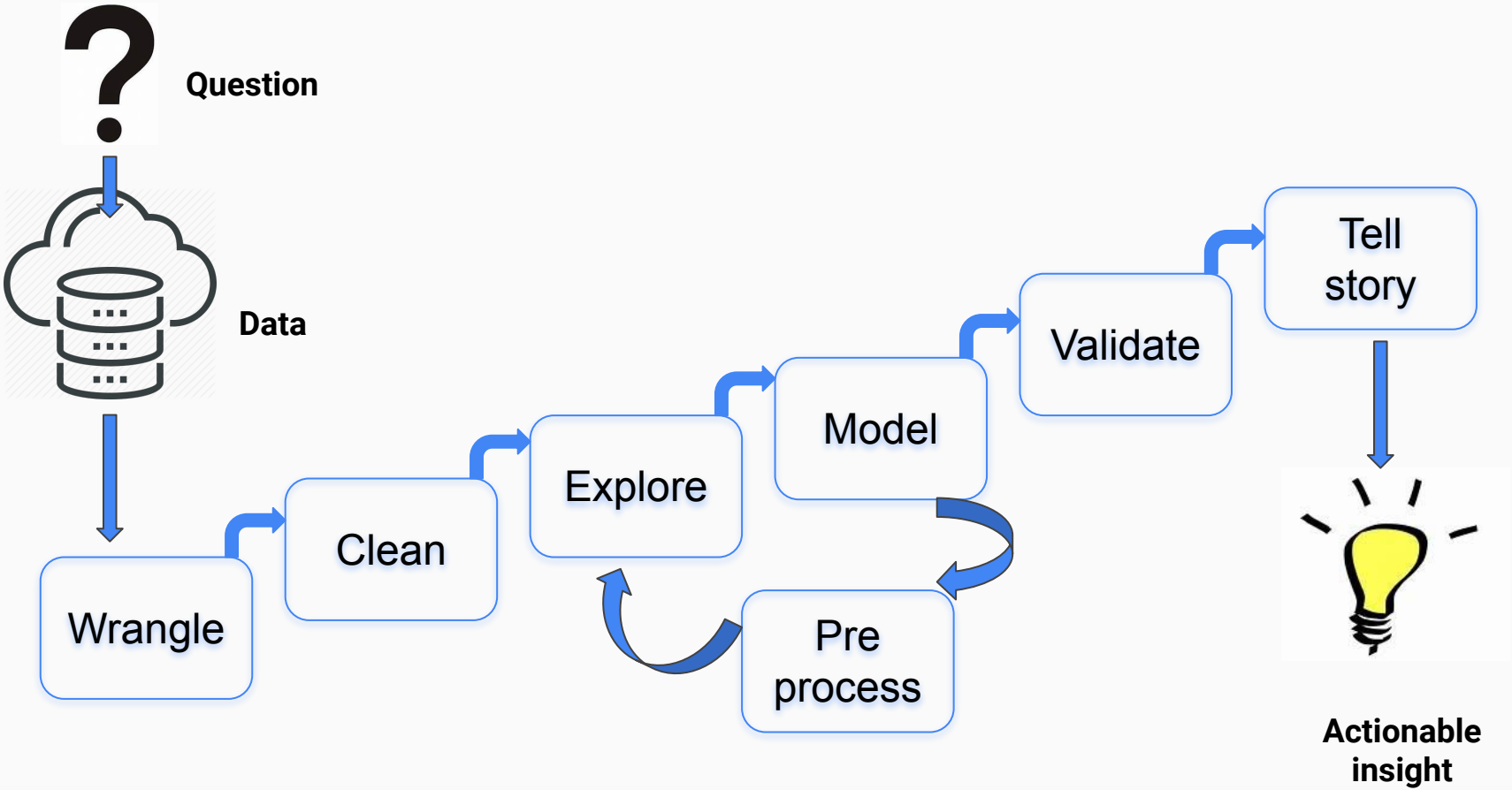
IPad Air
128 GB memory
0.29” thick

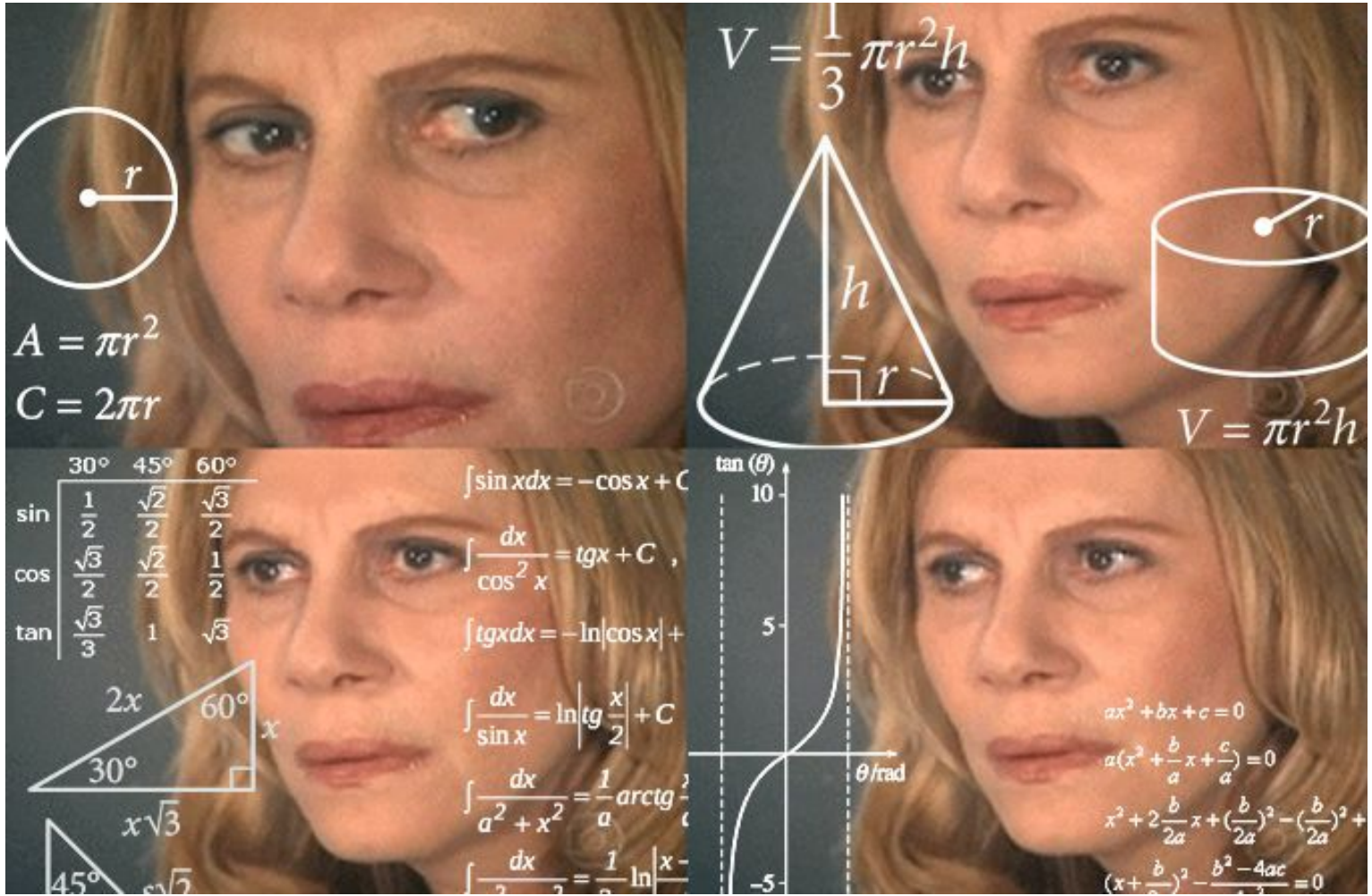
Source: EMC

Buzzwords

- **Data** - any piece of information that can be stored and processed
- **Data science** - Set of methods, processes, heuristics, and algorithms to extract insights from data
- **Big data** - extremely large amounts of data which traditional data processing systems fail to handle
- **Artificial Intelligence** - study of intelligent agents or developing intelligent systems
- **Machine Learning** - allow computer systems to learn from the data without explicitly programming




Question
Data
Wrangle
Clean
Explore
Model
Pre process
Validate
Tell story
Actionable insight


r
$A = \pi r^2$
$C = 2\pi r$
$V = \frac{1}{3}\pi r^2 h$
h
r
r
$V = \pi r^2 h$
30° 45° 60°
sin $\frac{1}{2}$ $\frac{\sqrt{2}}{2}$ $\frac{\sqrt{3}}{2}$
cos $\frac{\sqrt{3}}{2}$ $\frac{\sqrt{2}}{2}$ $\frac{1}{2}$
tan $\frac{\sqrt{3}}{3}$ 1 $\sqrt{3}$
2x
60°
x
30°
$x\sqrt{3}$
45°
$\int \sin x dx = -\cos x + C$
$\int \frac{dx}{\cos^2 x} = tgx + C$,
$\int tgx dx = -\ln|\cos x| +$
$\int \frac{dx}{\sin x} = \ln\left|tg\frac{x}{2}\right| + C$
$\int \frac{dx}{a^2 + x^2} = \frac{1}{a} arctg$
$\int \frac{dx}{} = \frac{1}{} \ln\left|x -\right.$
tan (θ)
10
5
θ/rad
-5
$ax^2 + bx + c = 0$
$a(x^2 + \frac{b}{a}x + \frac{c}{a}) = 0$
$x^2 + 2\frac{b}{2a}x + (\frac{b}{2a})^2 - (\frac{b}{2a})^2 +$
$(x + \frac{b}{}) - \frac{b^2 - 4ac}{} = 0$

What question to answer?

Formulate a question the stakeholder is trying to answer

Who are the next 1000 customers we will lose and why?

How do we identify and classify spam emails?

Is this a fraudulent credit card transaction?

How likely is it the user will buy our product?

How can we predict housing prices for the next few years?

DATA
DATA EVERYWHERE
memegenerator.net

Data sources

Data comes from variety of sources in different formats and is often messy.

Structured
Highly organized like CSV, database

Unstructured
No structure present like audio, video, documents

Semi-structured
Some organizational properties like XML, JSON, web pages

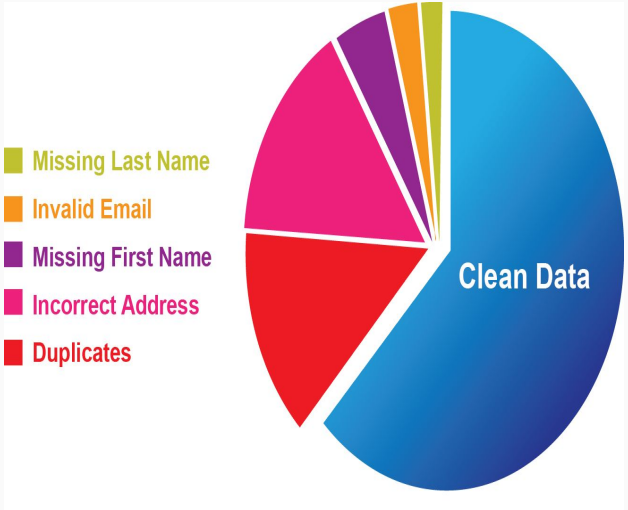
CLEANING

It's harder than it looks

Data wrangling

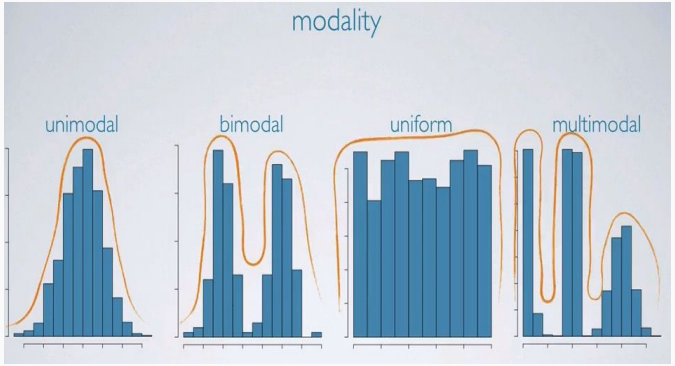
Data wrangling - gathering, selecting, transforming data for easy access and analysis

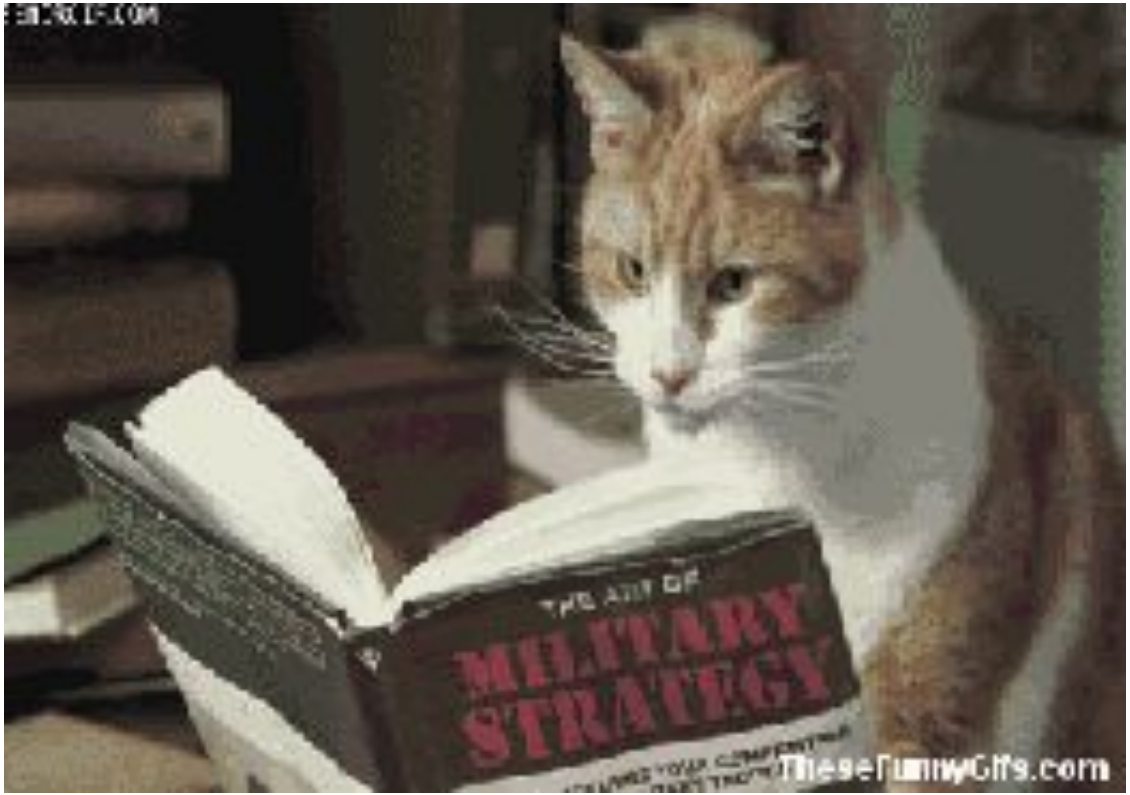
Standardize multiple sources NY=New York	Scaling or normalization	Discard or replace missing, NAs, negative values
Interpolate values	Convert categories to numbers Yes=1, No=0	Deduplicate records



Data exploration

- Initial investigation or Exploratory Data Analysis (EDA)
- Extract important variables
- Summarize characteristics
- Visualize properties of data using graphs
- Uncover initial patterns and points of interest
- Form hypotheses about defined problem




MILITARY
STRATEGY
TheseFunnyGifs.com

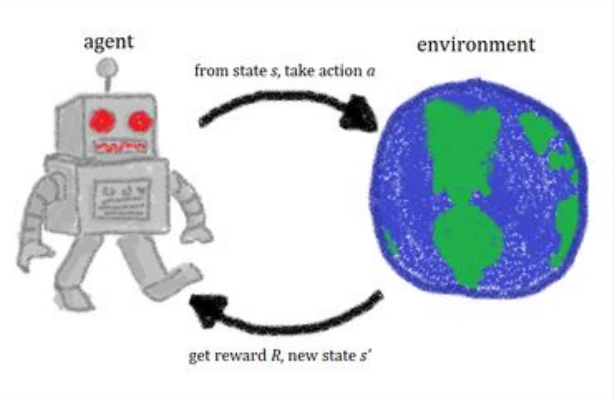
Model building

- **Feature engineering** - select important features and construct more meaningful ones, using domain knowledge
- Divide the data into training and test sets
- Create **Machine Learning** model
  - Choose supervised or unsupervised learning
  - Tune model parameters
  - Train the model
  - Monitor against **overfitting**
  - Evaluate model on unseen data i.e. test set
- Iterative process with different features
- Can have ensemble of models

Machine learning approaches

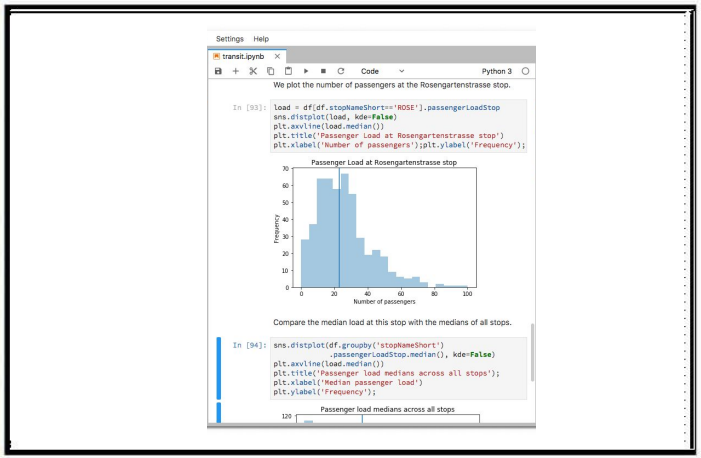
Supervised learning

Unsupervised learning



Reinforcement learning

Tool: Jupyter notebook



Algorithms : Classification

Algorithms: Regression

Algorithms: Clustering

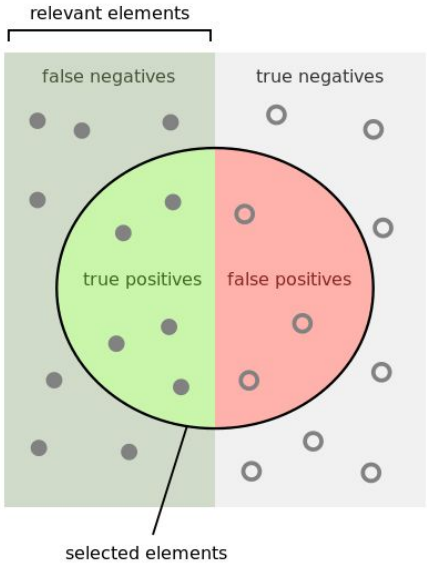
Algorithms: Anomaly detection

Reinforcement learning

Model validation

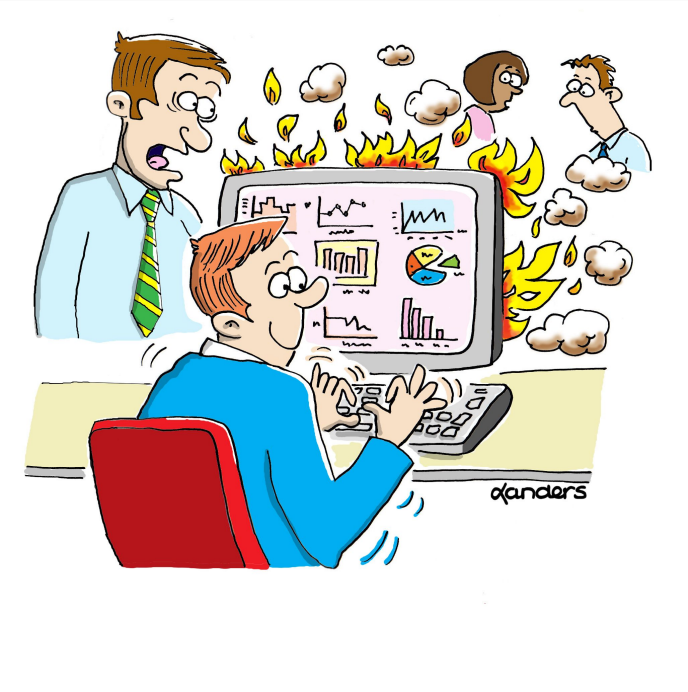
- Measure model quality - how good is it?
- Use cross-validation for robustness
- Use metrics like accuracy, precision, recall, F1 score, confusion matrix
- H_0 is the null hypothesis i.e. any observed difference in samples is due to chance or sampling error

	H_0 True	H_0 False
Reject H_0	Type I Error False positive	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error False negative



Data visualization and storytelling

- Tell a story with data
- Communicate findings to key stakeholders
- Use plots and interactive visualizations
- Answer the original questions
- Use powerful narratives for storytelling



DHH @dhh · Nov 7

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

1.6K 9.6K 29K

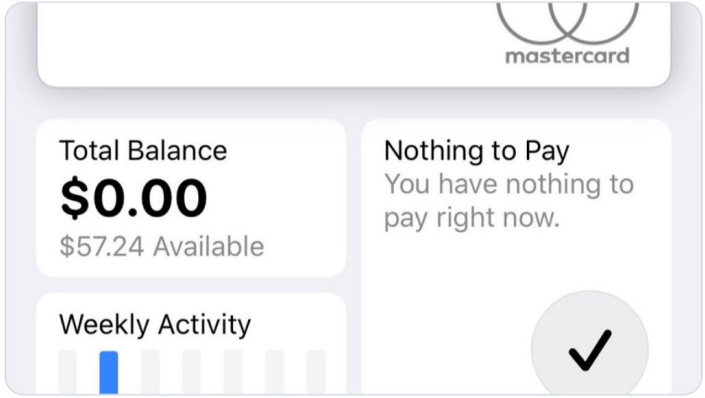
DHH @dhh · Nov 7

I'm surprised that they even let her apply for a card without the signed approval of her spouse? I mean, can you really trust women with a credit card these days??!

92 247 4.4K

DHH @dhh · Nov 7

It gets even worse. Even when she pays off her ridiculously low limit in full, the card won't approve any spending until the next billing period. Women apparently aren't good credit risks even when they pay off the fucking balance in advance and in full.



77 391 3.7K

DHH @dhh · Nov 8

She spoke to two Apple reps. Both very nice, courteous people representing an utterly broken and reprehensible system. The first person was like "I don't know why, but I swear we're not discriminating, IT'S JUST THE ALGORITHM". I shit you not. "IT'S JUST THE ALGORITHM!".

73 596 4.8K

DHH @dhh · Nov 8

Second rep went on about how she couldn't actually access the real reasoning (again IT'S JUST THE ALGORITHM is implied). But then revealed that my wife really should check her credit score, again implying that it was less than mine, and that's why. Even named TransUnion to check.

13 151 2.4K

DHH @dhh · Nov 8

So obviously we both furiously signup for the fucking $25/month credit-check bullshit shakedown that is TransUnion. Maybe someone stole my wife's identity? Even though we've verified there was nothing wrong previously. Guess what: HER CREDIT SCORE WAS HIGHER THAN MINE!!!

55 453 5.9K

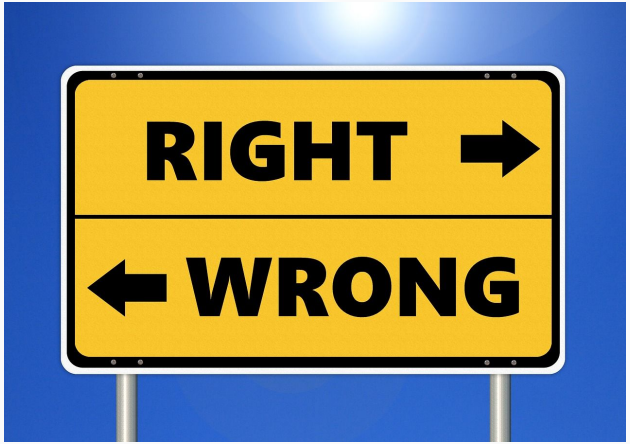
DHH @dhh

So let's recap here: Apple offers a credit card that bases its credit assessment on a black-box algorithm that 6 different reps across Apple and GS have no visibility into. Even several layers of management. An internal investigation. IT'S JUST THE ALGORITHM!

Ethics in Data Science

All involved in handling data should have an ethical discussion about the way the data is used. Checklist by Mike Loukides, Hilary Mason, DJ Patil:

- How can the tech be attacked or misused
- Fair and representative training data
- Study and understand possible sources of bias
- Diverse team - opinions, backgrounds, thoughts
- Clear, explicit user consent and data protection
- Ensure fairness over time, and for different groups
- Shut down in production if behaving badly and redress those harmed

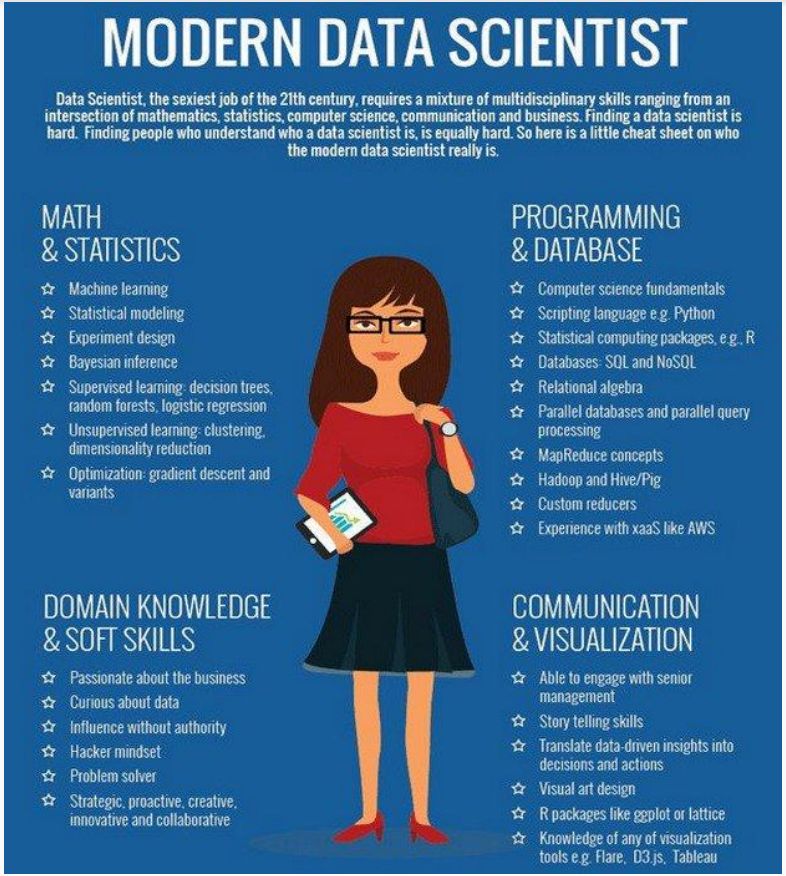


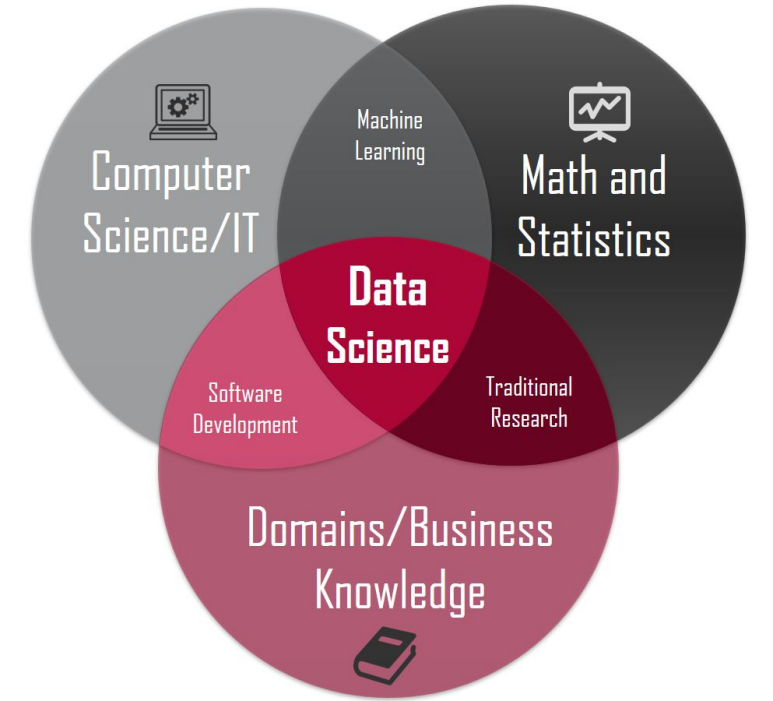
Recap

- What is Machine Learning?
- Data pipeline
  - Question
  - Data sources
  - Data cleaning
  - Data exploration
  - Model building
  - Model validation
  - Data visualization and storytelling
- Machine Learning approaches
  - Supervised (Classification, Regression)
  - Unsupervised (Clustering)
  - Reinforcement learning
- Ethics

Resources

- IBM’s Cognitive class
- Jupyter
- KD Nuggets
- Kaggle
- Towards Data Science
- Coursera
- Free Code Camp
- School of AI
- Seattle Data Guy’s Python resources
- Fast.ai
- Google ML crash course
- FiveThirtyEight





gjena.github.io

grishmajena

DebateLover