INSTRUCTIONS

Version 1.02
Template

Dear author(s):

In this presentation template, you will find the formatting for all DevPulseCon presentations.

Please Google Share or email attach your presentation by 9am PT April 17th Wednesday to Eric (email address below). This will allow staff the opportunity to test run the presentation at the site.

Further details or questions, please email to:

- bizdev@codechix.org or eric.han.wg13@gmail.com

codechix Dev Pulse Con

codechix

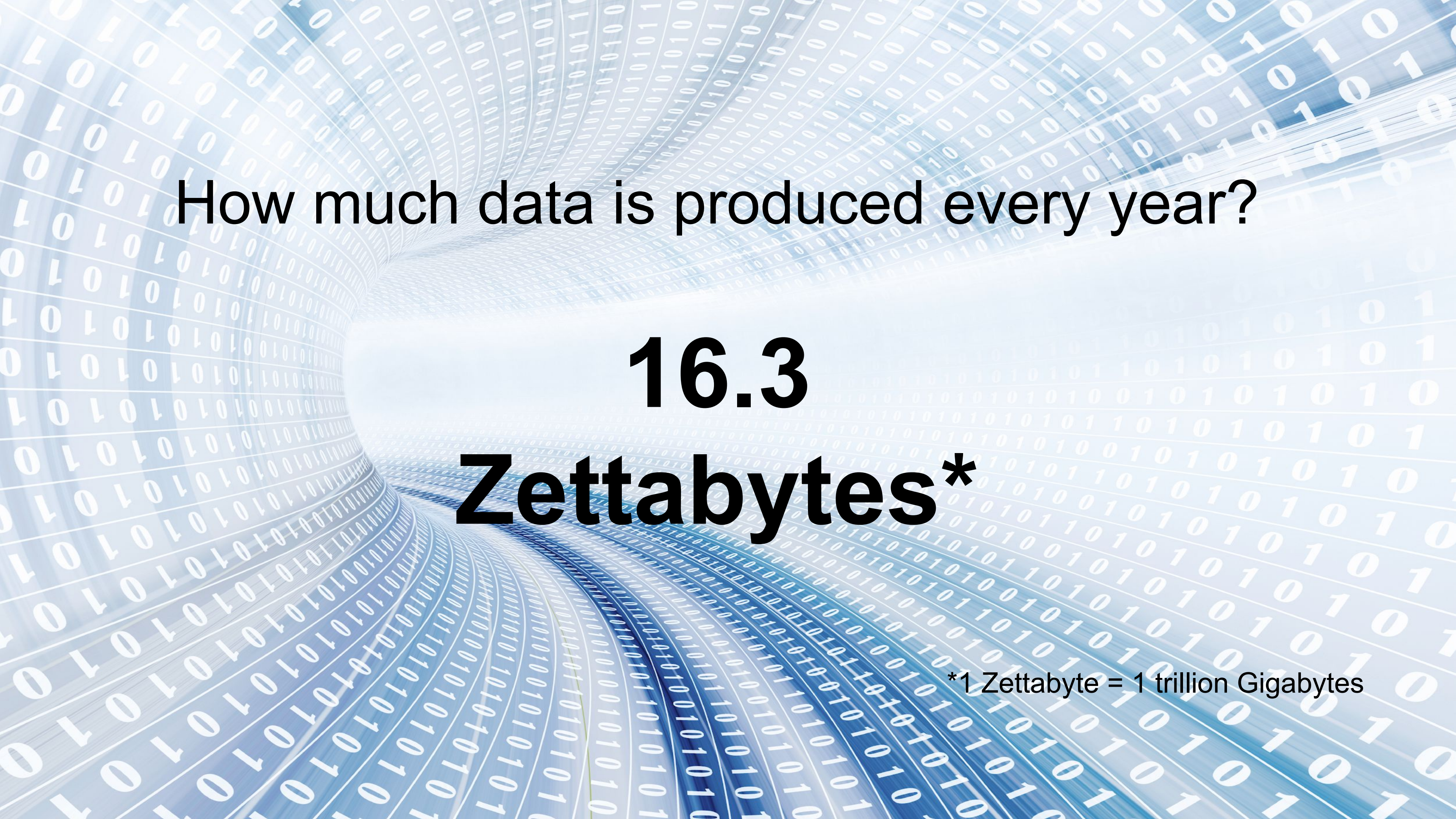
Dev Pulse Con

April 19-20, 2019
Computer History Museum
Mountain View, CA

Deciphering Data Science

Grishma Jena

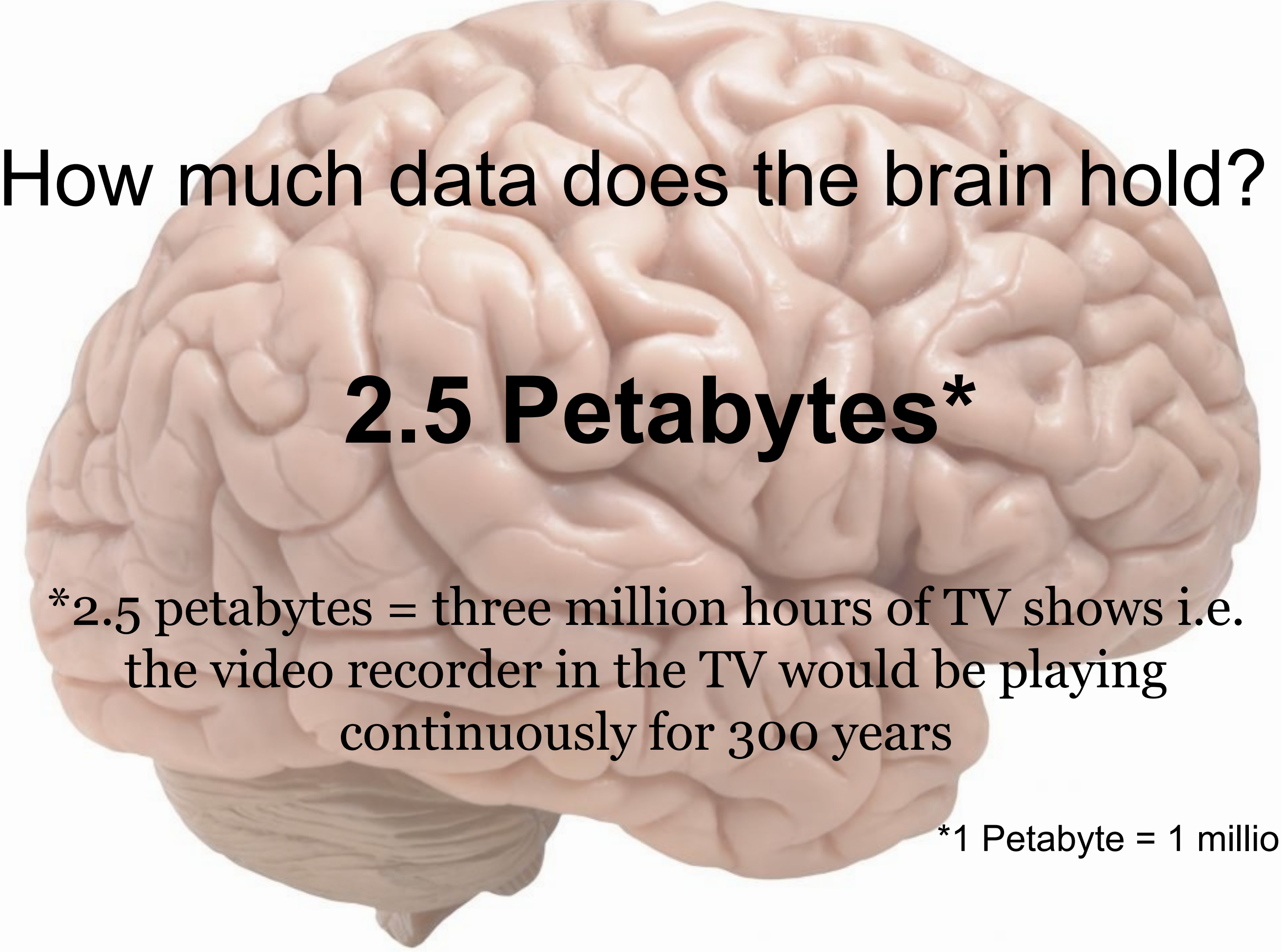
Cognitive Software Engineer
IBM Watson
Twitter: @DebateLover



How much data is produced every year?

16.3 Zettabytes*

*1 Zettabyte = 1 trillion Gigabytes



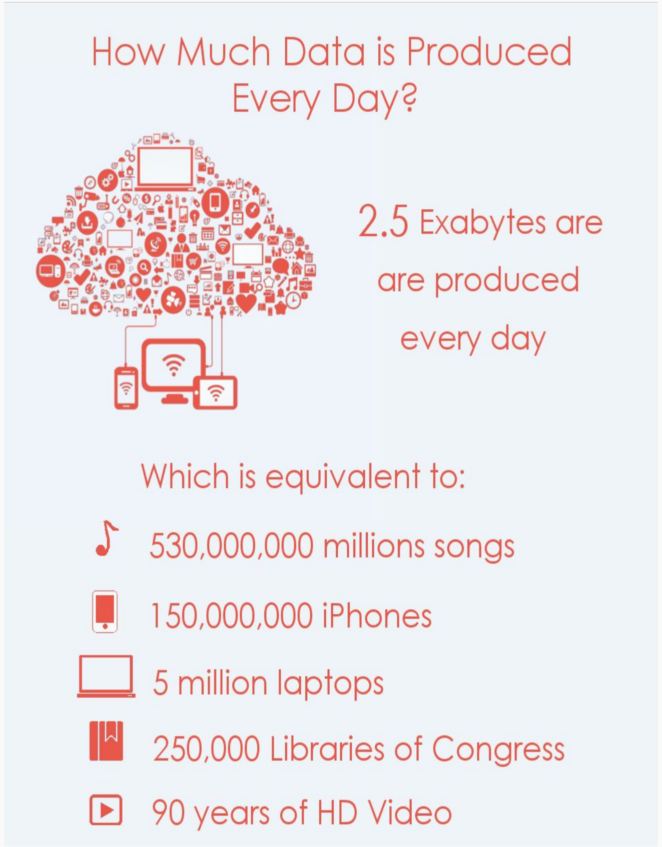
How much data does the brain hold?

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...



How Much Data is Produced Every Day?

2.5 Exabytes are are produced every day

Which is equivalent to:

- ♪ 530,000,000 millions songs
- 150,000,000 iPhones
- 5 million laptops
- 250,000 Libraries of Congress
- ▶ 90 years of HD Video

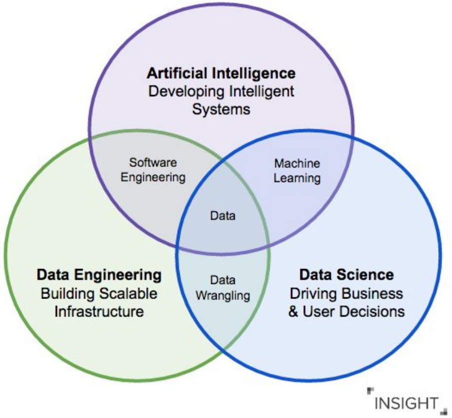
2020 estimates

1.7 megabytes per second

44 zettabytes

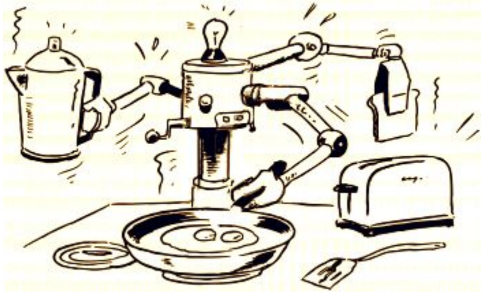
Buzzwords

- **Data** - any piece of information that can be stored and processed
- **Data science** - A 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



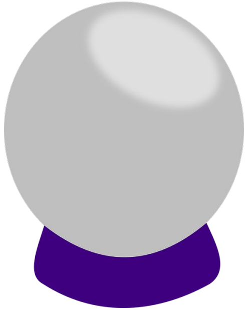
Source: Fast Forward Labs blog

Data Science capabilities



Enable automation of common tasks

Able to process lots of data quickly



Predicting and forecasting results

Making customized recommendations

Optimize procedures and actions

Find patterns not easily identified by humans

Data Science Pipeline

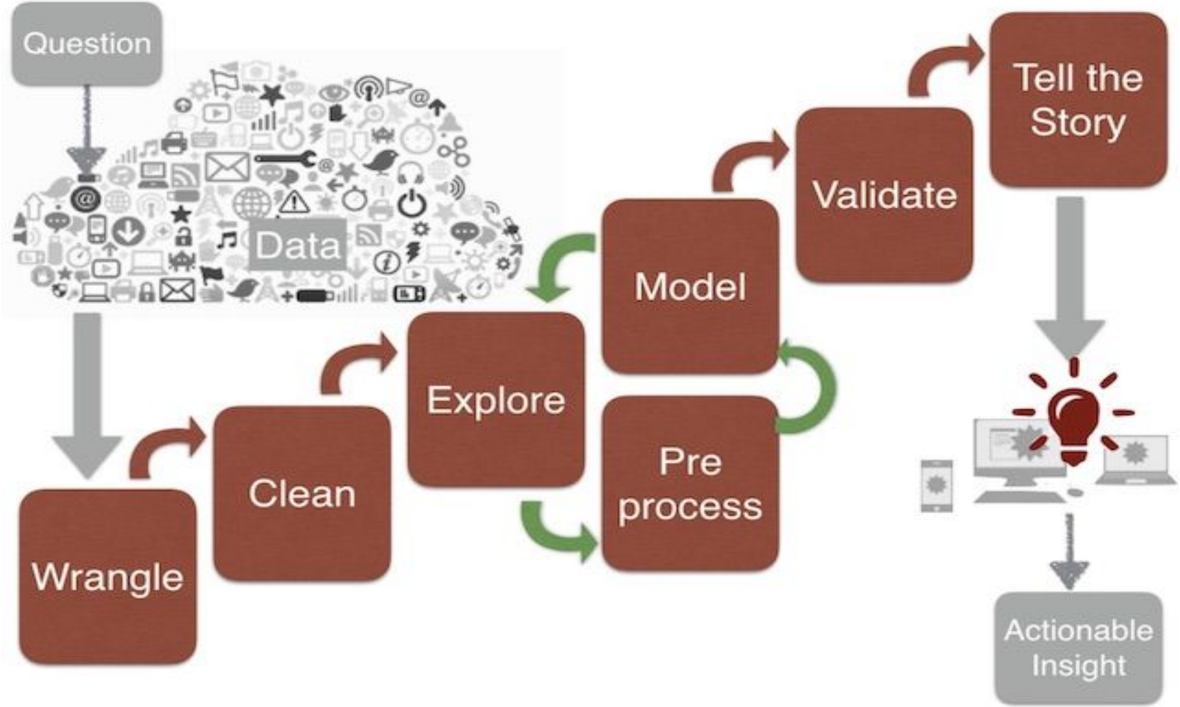


Image by Wolfram Research from uvm.edu

What question to answer?

Formulate a question the stakeholder is trying to answer

Created by b farias
from Noun Project

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

Created by Template
from Noun Project

How do we identify and classify spam emails?

Created by ProSymbols
from Noun Project

Is this a fraudulent credit card transaction?

Created by Gregor Cresnar
from Noun Project

How likely is it the user will buy our product?

Created by anbileru adaleru
from Noun Project

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

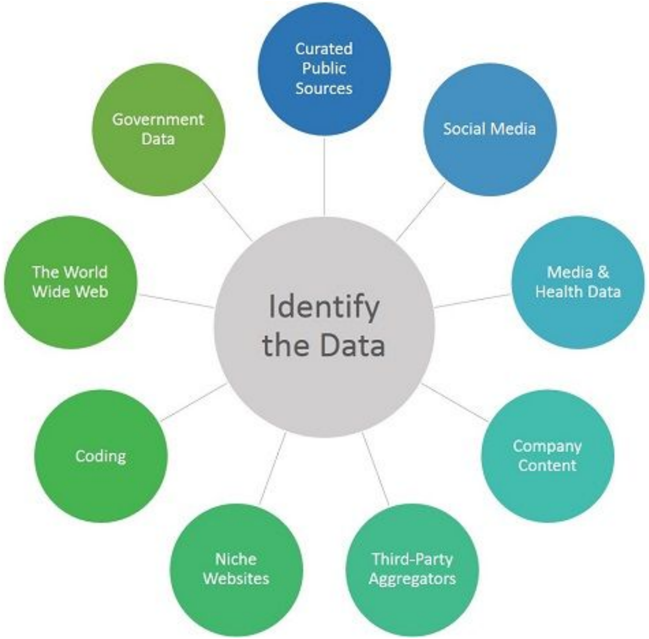
Data sources

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

Structured: Highly organized, tables with rows and columns like database, CSV

Unstructured: No structure present, like audio, video, documents

Semi-structured: Has some organizational properties like XML, JSON, web pages

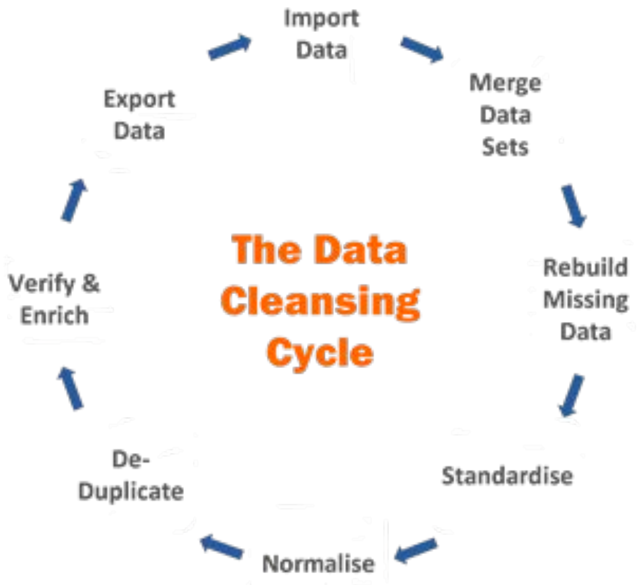


Source: Search technologies

Data wrangling

Data wrangling - gathering, selecting, transforming data to make useful

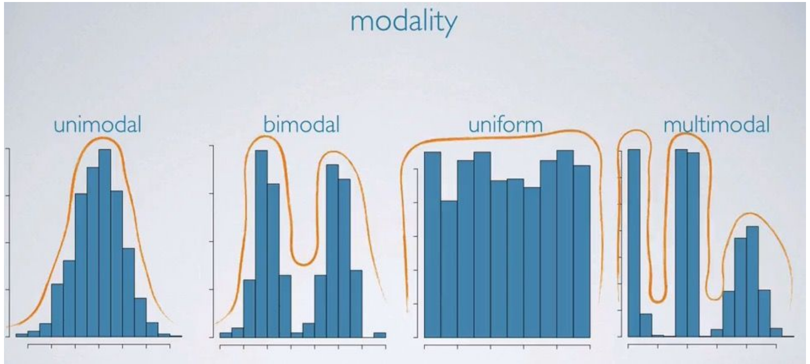
- Standardize - Gender == Sex, NY == New York
- Discard - missing, NAs, negative values
- Replace - with average, median, 0, etc.
- Interpolate values
- Convert categories - Spam = 1, Not spam = 0
- Scaling or normalization
- Deduplicate records



Source: Data cleansing group

Data exploration

- Exploratory data analysis (EDA)
- Initial investigation
- Extract important variables
- Summarize characteristics
- Uncover initial patterns, points of interest
- Form hypotheses about defined problem
- Visualize properties of data using graphs
  - Plot data using traces, histograms
  - Plot simple statistics like mean, standard deviation
  - Univariate, bivariate and multivariate visualizations



Source: Napitplu Jon

Model building

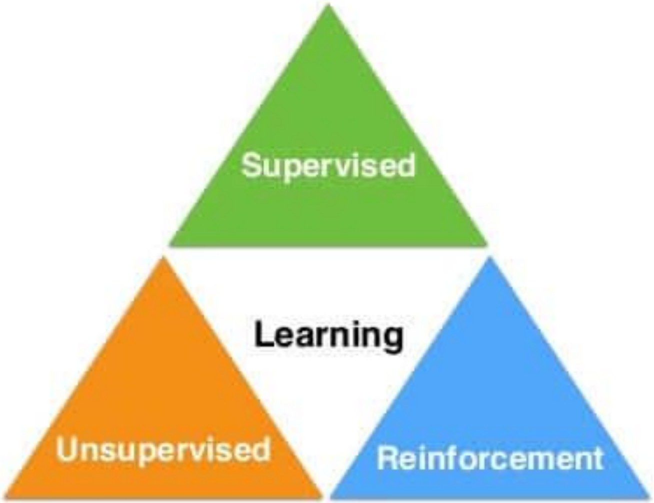
- **Feature engineering** - select important features and construct more meaningful ones, using domain knowledge
- Divide the data into training and test sets
- Create **Machine Learning** model
  - Choose supervised or unsupervised learning
  - Tune model parameters
  - Train the model
  - Monitor against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

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Source: XKCD comic

Machine learning approaches

- Labeled data
- Direct feedback
- Predict outcome/future

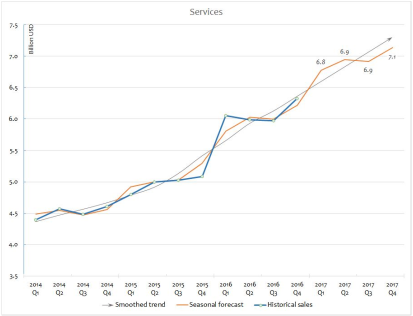


- No labels
- No feedback
- “Find hidden structure”

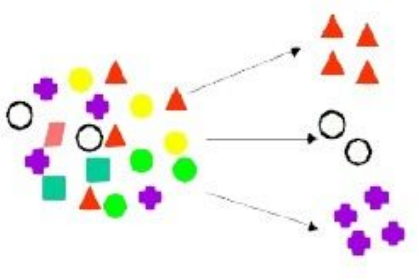
- Decision process
- Reward system
- Learn series of actions

Algorithms

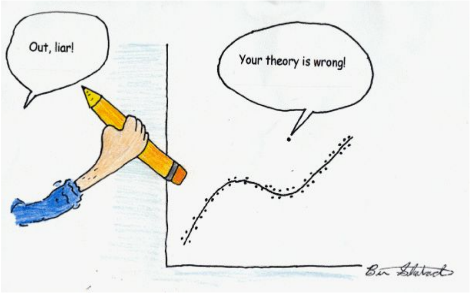
Classification: Cat or Dog?



Regression: how much or how many?



Clustering: how is this organized?

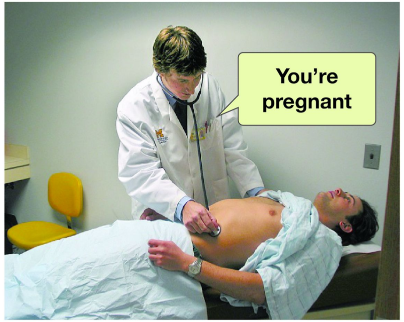


Anomaly detection: is this weird?

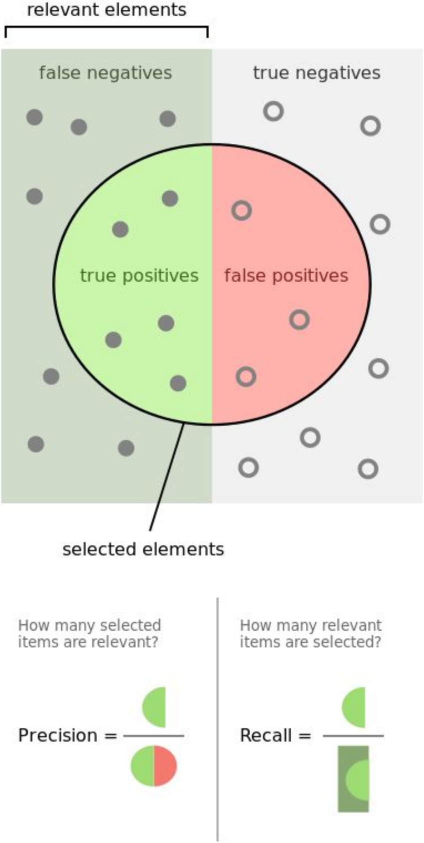
Model validation

- Measure model quality - how good is it?
- Use cross-validation for robustness
- Use metrics like accuracy, precision, recall, F1 score

Type I error
(false positive)



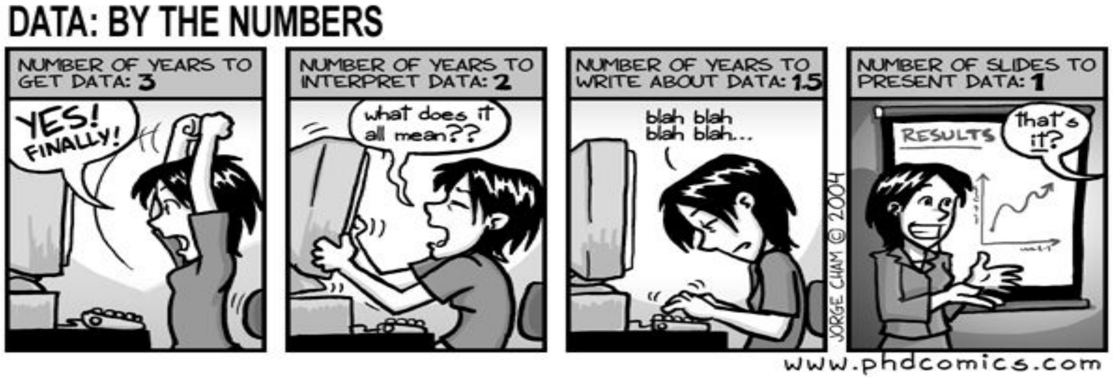
Type II error
(false negative)



Source: Wikipedia

Data visualization

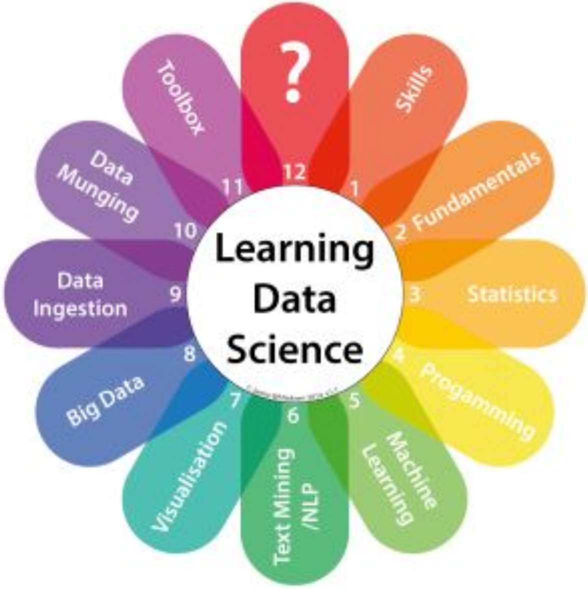
- Tell a story with data
- Communicate findings to key stakeholders
- Use plots and interactive visualizations
- Answer the original questions


DATA: BY THE NUMBERS
NUMBER OF YEARS TO GET DATA: 3
YES! FINALLY!
NUMBER OF YEARS TO INTERPRET DATA: 2
what does it all mean??
NUMBER OF YEARS TO WRITE ABOUT DATA: 1.5
blah blah blah blah...
NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?
JORGE CHAM © 2004
www.phdcomics.com

Source: PhD comics

Data science tools

- Languages: Python, SQL, R, Scala, Java, Matlab
- Cloud-based environments: IBM Watson Studio, Amazon ML, Google Cloud
- Tools: Jupyter, RStudio, Tableau
- Big Data: Spark, Hadoop, Hive, Cassandra, Elastic search, Kafka, Mesos
- Deep Learning: TensorFlow, PyTorch, Keras
- Visualization: Bokeh, Matplotlib, ggplot, plotly, D3, Fusion Charts, Ember, Networkx
- A lot of these are open-source
- A more comprehensive list here



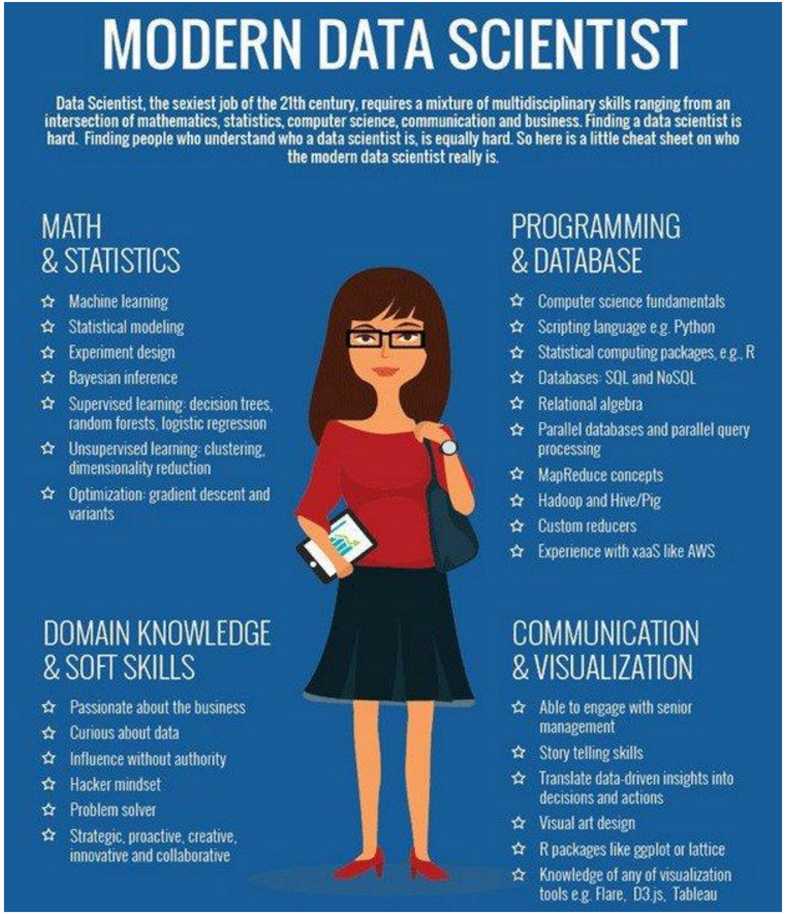
Source: Towards Data Science

Pick up an interesting dataset and play with it to discover something fascinating

Nothing like some hands-on experience :)

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



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

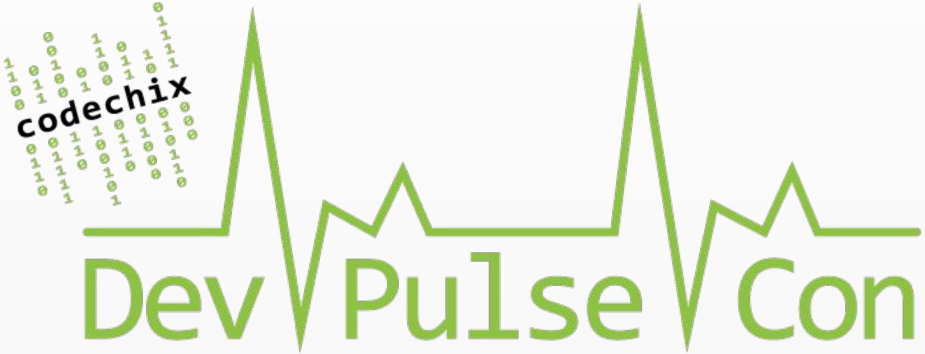
Source: Data driven

gjena.github.io

grishmajena

DebateLover

GJena



codechix
Dev Pulse Con