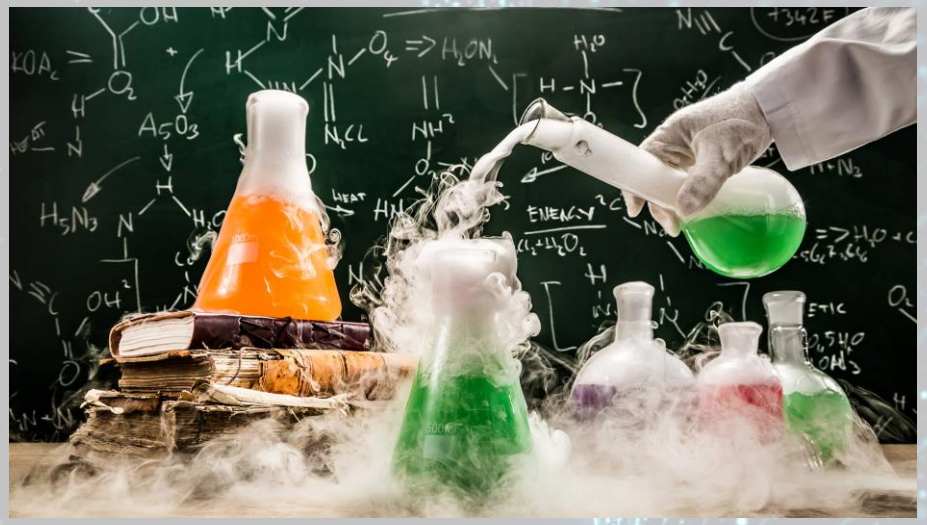
Data Science Workshop Series

WS1-SE1

What is data science?

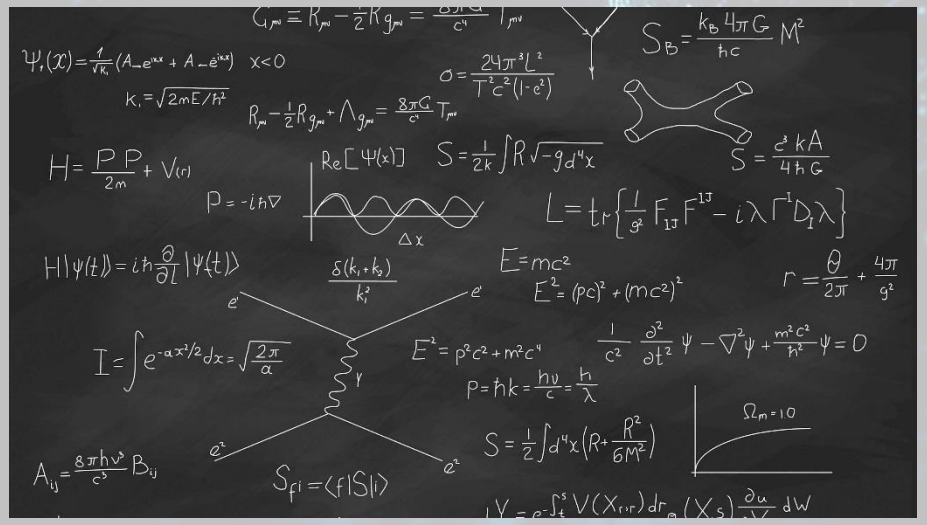
By: Alireza Vafaei Sadr

May-2019-IPM



Empirical evidence

Computational science



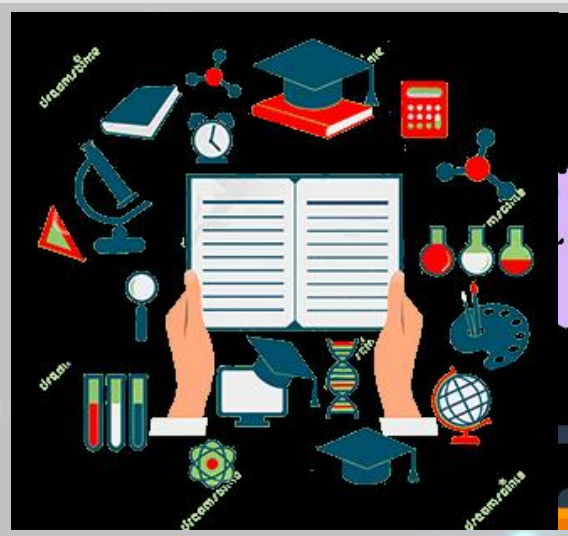
Scientific theory

BIG DATA

Data science

Data?!



https://www.smartdatacollective.com/big-data-20-free-big-data-sources-everyone-should-know/

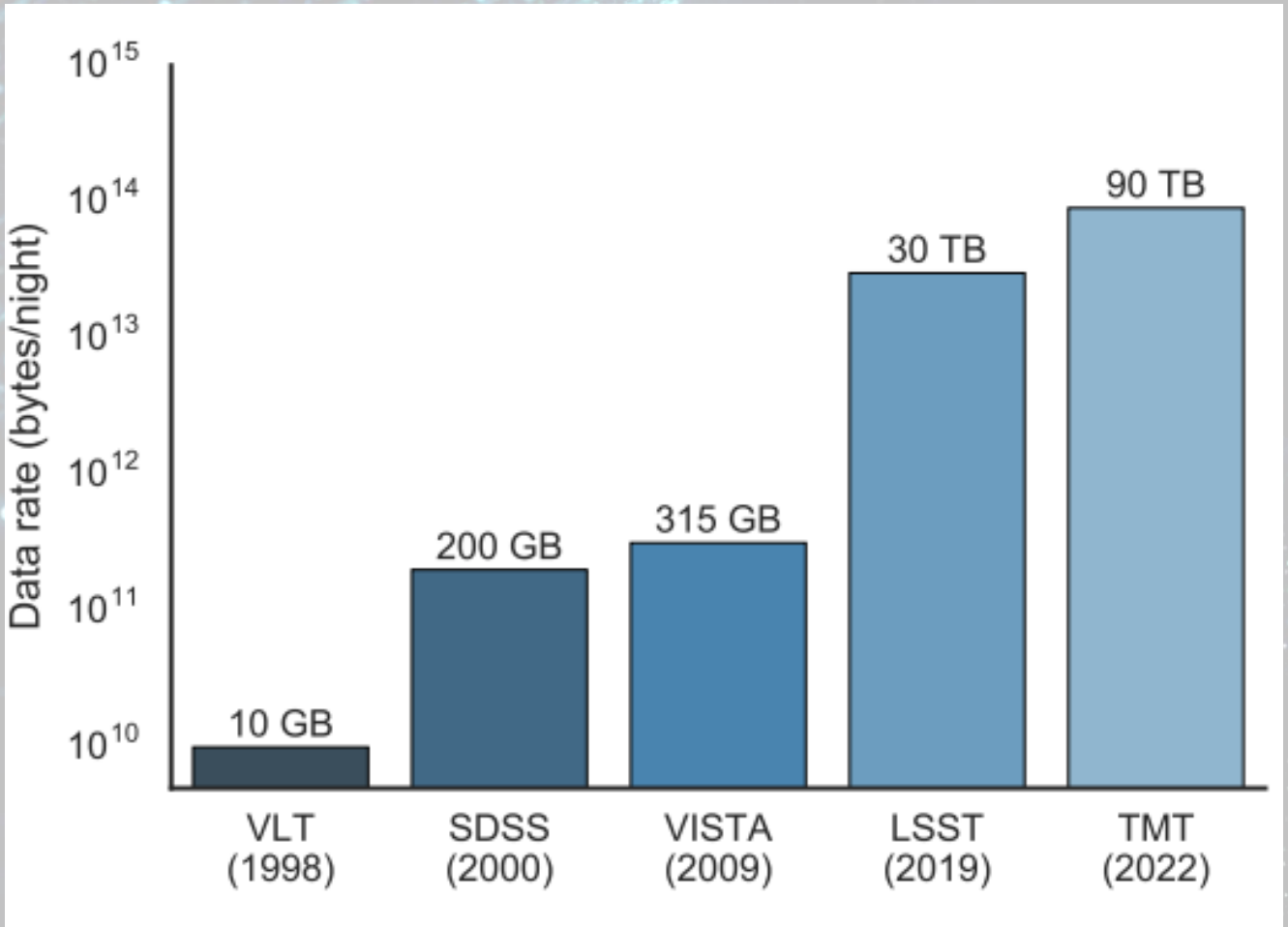
How BIG?

- New telescopes collect today 50 times the info they collected 5 years ago.
- Goggle process 24 PB per day = US library of congressX1000
- Facebook updates 10M photos per hour and 38B like per day.
- YouTube adds one hour of video every second
- …

An example in Physics!

Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy

Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, and Christian Igel



Do we have access to them?!

https://www.data.gov/

https://digital.nhs.uk/

https://healthdata.gov/

https://www.cia.gov/library/publications/the-world-factbook/

https://data.gov.uk/

http://data.europa.eu/euodp/en/data/

https://trends.google.com/trends/explore

https://www.google.com/finance

https://wiki.dbpedia.org/

https://aws.amazon.com/datasets/million-song-dataset/

https://data.worldbank.org/

https://www.who.int/gho/database/en/

https://www.google.com/publicdata/directory

https://registry.opendata.aws/

https://data.fivethirtyeight.com/

https://www.census.gov/data.html

https://www.yelp.com/dataset

https://data.unicef.org/

https://www.kaggle.com/datasets

https://lodum.de/

https://archive.ics.uci.edu/ml/index.php

https://www.xenonstack.com/blog/ingestion-processing-big-data-iot-stream/

Science?!

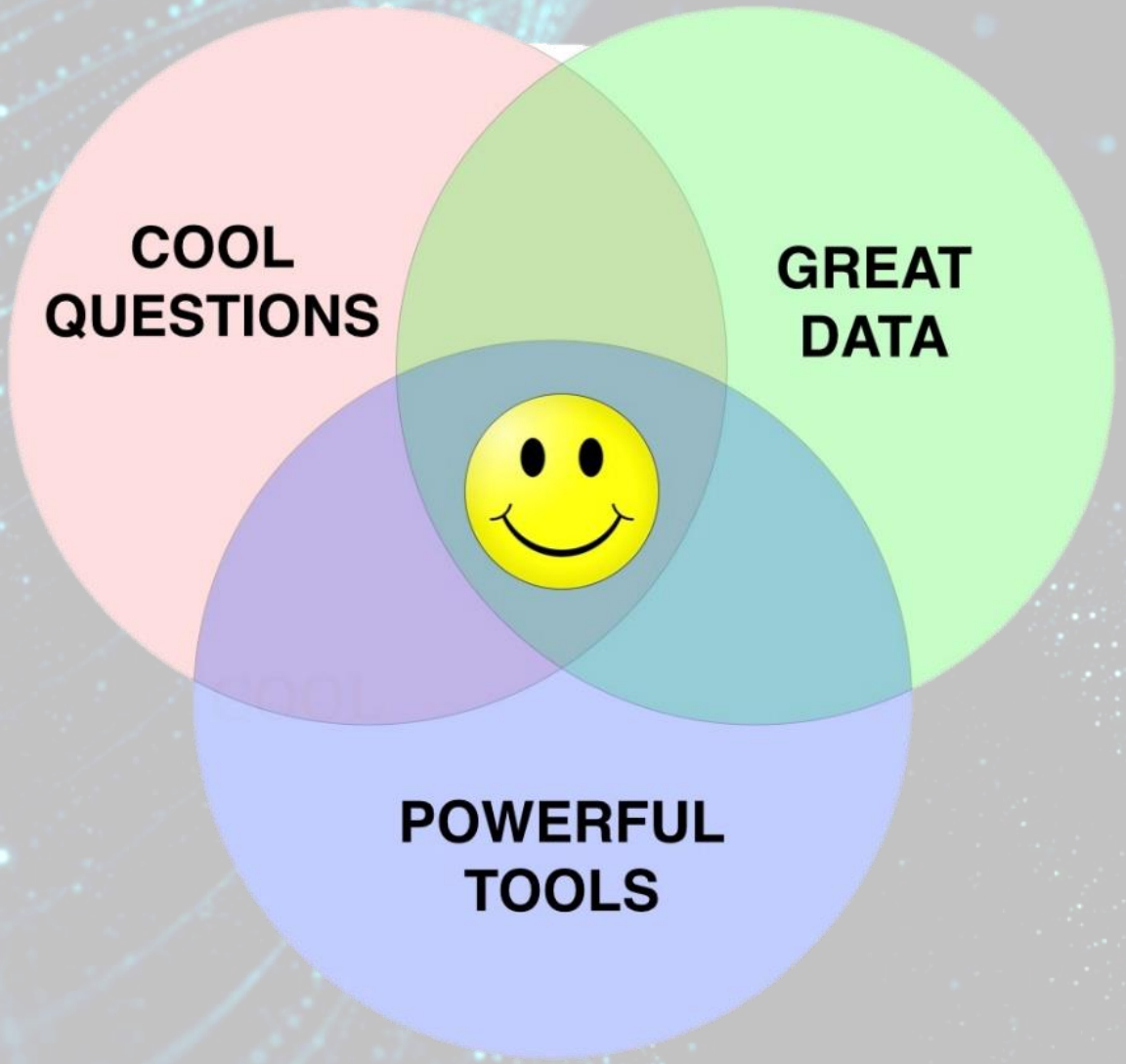
data research:

- Hypothesis-Driven:
  What kind of data do we need to help solve a problem?

- Data-Driven:
  What interesting problems can be solved with this data!?


DATA
Science
01
BUSINESS UNDERSTANDING
02
DATA MINING
03
DATA CLEANING
04
DATA EXPLORATION
05
FEATURE ENGINEERING
06
PREDICTIVE MODELING
07
DATA VISUALIZATIOAN


Data Science
Visualization
Statistical Modeling
Statistical Computing
Data Technology
Data Research
Data Consulting
Real-world Applications
Scientific Methods


COOL QUESTIONS
GREAT DATA
POWERFUL TOOLS

Better half a loaf than no bread.

آب دریا را اگر نتوان کشید هم به قدر تشنگی باید چشید

Creativity!

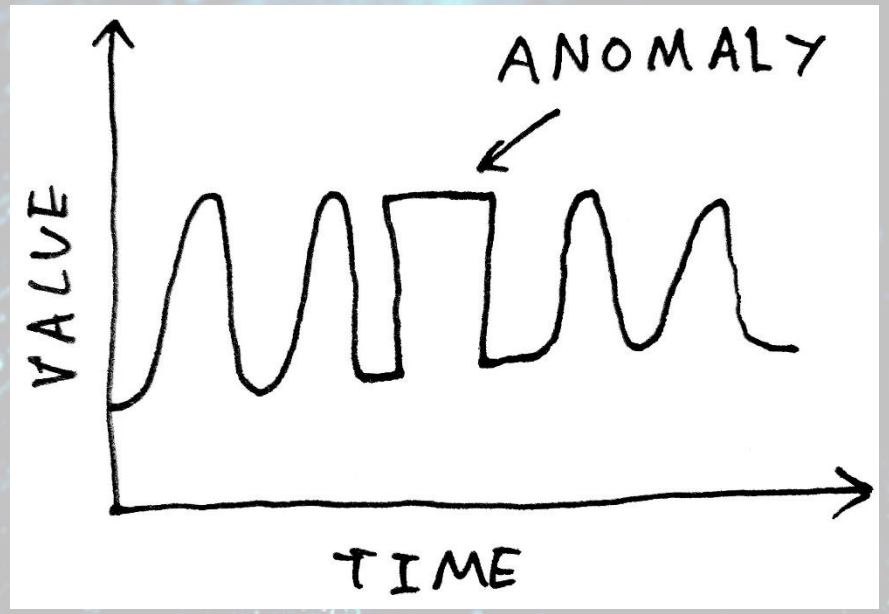
CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms*

Ahmed Elgammal[1†] Bingchen Liu[1] Mohamed Elhoseiny[2] Marian Mazzone[3]

Creativity!

https://towardsdatascience.com/do-gans-really-model-the-true-data-distribution-or-are-they-just-cleverly-fooling-us-d08df69f25eb

Discovery!



Discovery!

Table 1 Major discoveries made by the Hubble Space Telescope (*HST*). Of the *HST*'s "top ten" discoveries (as ranked by National Geographic magazine), only one was a key project used in the *HST* funding proposal (Lallo 2012). A further four projects were planned in advance by individual scientists but not listed as key projects in the *HST* proposal. Half the "top ten" *HST* discoveries were unplanned, including two of the three most cited discoveries, and including the only *HST* discovery (Dark Energy) to win a Nobel prize. This Table was previously published by Norris et al. (2015).

Project	Key Project?	Planned?	Nat Geo top ten?	Highly cited?	Nobel Prize?
Use cepheids to improve value of H_0	✓	✓	✓	✓	
UV spectroscopy of ig medium	✓	✓			
Medium-deep survey	✓	✓			
Image quasar host galaxies		✓	✓		
Measure SMBH masses		✓	✓		
Exoplanet atmospheres		✓	✓		
Planetary Nebulae		✓	✓		
Discover Dark Energy			✓	✓	✓
Comet Shoemaker-Levy			✓		
Deep fields (HDF, HDFS, GOODS, FF, etc)			✓	✓	
Proplyds in Orion			✓		
GRB Hosts			✓		

Walking! :D

Google “how google ai can play”

Understanding!?

Google “Audio-Visual Speech Separation”

I wanna be one of them!

Data: acquisition , structure, storage, cleaning, management …

Statistics: probability, error analysis, statistical significance …

Programming: OS, development (at least in one language) …

Machine learning: almost all of it!

Practice: (practice, experience, taste) real world examples!

You need to be passionate about data, your questions and
a lot of crazy things in programming!

You definitely need to

(Computer+”book”)

Be a ~~Book~~ Worm!

data acquisition:

- Data Sources: Companies/Proprietary Data, APIs, Government, Academic, Web Scraping/Crawling

Types of data

- Structured vs. Unstructured
- Quantitative vs. Categorical
- Discrete vs. Continuous
- Ordinal vs. Nominal

Structure and Formats:

- CSV, XML, SQL, JSON, H5
- Databases

Statistics:

- How events are alike?
- How much an event is probable?
- How one can compare different results?
- Correlation analysis
- Normalizations, compatibility
- Noise, errors and artifacts
- Data augmentation
- Data Imputation
- Outlier Detection

Statistics:

- Monte Carlo based techniques
- Distributions
- Modeling
  1. Parametric vs. Nonparametric
  2. Supervised vs. Unsupervised
  3. Blackbox vs. Descriptive (Prediction vs Inference)
  4. First-Principle vs. Data-Driven
  5. Deterministic vs. Stochastic
  6. Flat vs. Hierarchical
- Fitting

Model Evaluation

- Metrics:
  1. Accuracy
  2. Precision
  3. Recall
  4. Absolute Error
  5. MSE
- Methods:
  1. Cross Validation
  2. Bootstrapping

Feature engineering:

- Rounding
- Scaling
- Binning
- Interactions
- Transformation
- Dimensionality Reduction
- Encoding, Embedding

(machine learning) Models:

- Linear Regression
- Logistic Regression
- DistanceBased/Network algorithms
- Nearest Neighbor methods
- Clustering algorithms
- Naive Bayes
- Ensemble methods
- Random forests
- SVMs
- ANNs

Machine learning (concepts):

- Training/Validating/Testing
- Overfitting
- Bias/Variance
- Regularization
- Hyperparameters

Artificial Neural Networks

- Perceptron
- Activation Functions
- Optimizers
- Dropout
- Convolutions, Poolings
- Recurrents
- Regression, Classification, Detection, Segmentation …
- Transfer Learning
- Generative Adversarial Networks

How is one able to deal with this load of works?!?

One does not need to do that!

Data science does need a lot of collaboration, team work, discussion, …

Programming skills or how we can cook a data scientist?:

- Mentioned knowledge
- Computer and OS
- Programming concepts
- At least one hot programming language
- Good awareness and understanding of various packages
- Cooperative sprite
- A sufficient amount of confidence
- And, a massive amount of enthusiasm

Project management, collaboration and communication skills:

- GitHub
- Scrum
- Documentation
- Visualization

Fast Facts

Famous Data Scientist

Larry Page
CEO of Google

Job Opportunities

15,000%

increase in job postings for data scientists between 2011 & 2012.

Majors

- physics
- applied maths
- social sciences
- statistics
- analytics
- computer science
- marketing

$80K

average starting salary

$120K

average data science salary

$250K

data science team manager

$400K

highest paid data scientist

سپاس