

data science

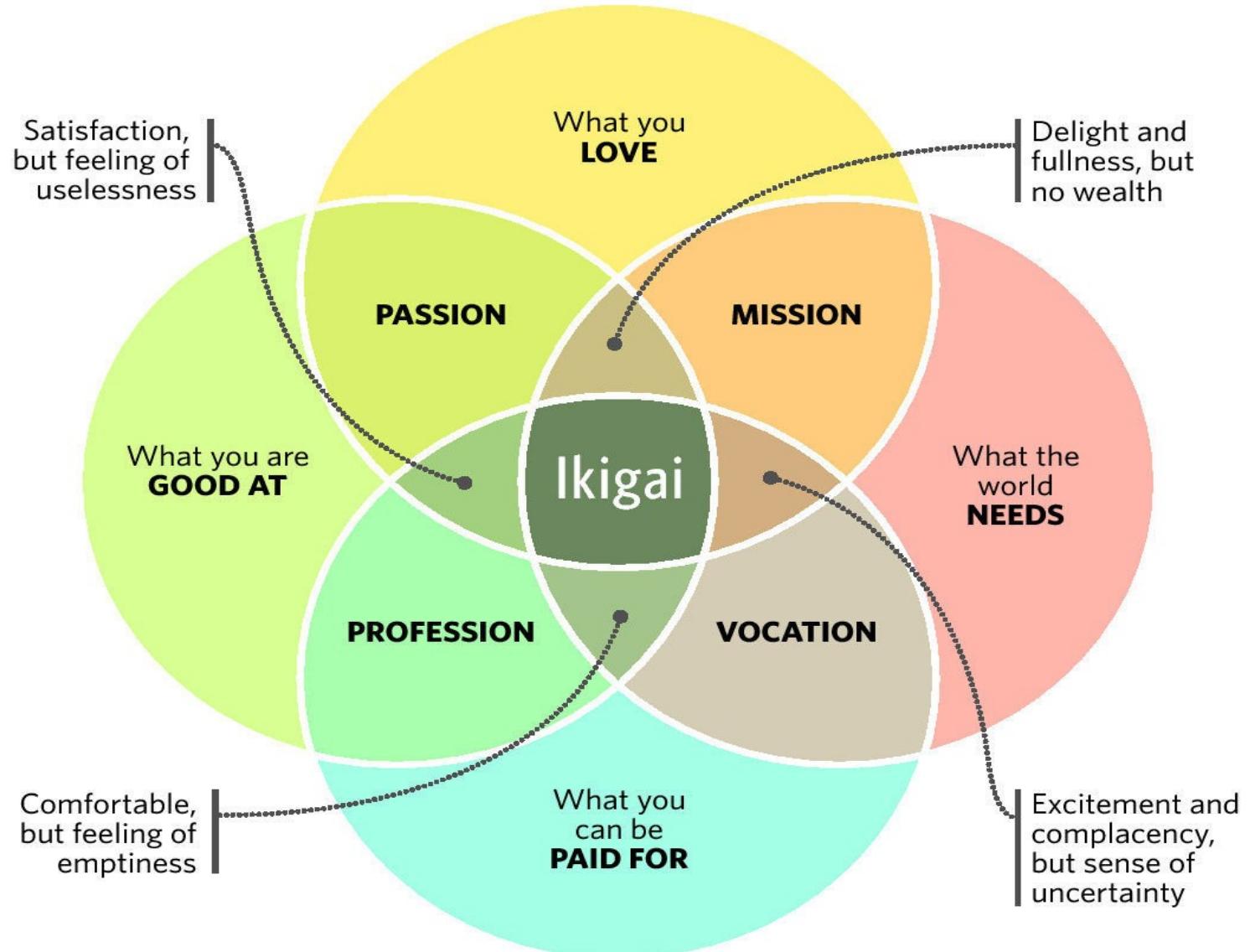
Alireza Vafaei Sadr



Feb-2020-TelAS

Ikigai

A JAPANESE CONCEPT MEANING "A REASON FOR BEING"





Empirical evidence

$$\Psi(x) = \frac{1}{\sqrt{R}}(A e^{i k x} + B e^{-i k x}) \quad x < 0$$

$$R = \sqrt{2mE/\hbar^2} \quad R_p = \frac{1}{2}Rg_{pp} + \Delta g_{pp} = \frac{8\pi G}{c^3}T_{pp}$$

$$H = \frac{P^2}{2m} + V(r) \quad P = -i\hbar\nabla$$

$$H|\psi(t)\rangle = i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle$$

$$A_{ij} = \frac{8\pi\hbar v^3}{c^3} B_{ij}$$

$$S_f = \langle f | S | i \rangle$$

$$S_B = \frac{k_B 4\pi G}{\hbar c} M^2$$

$$S = \frac{1}{2k} \int R \sqrt{-g} d^4x$$

$$S = \frac{c^3 k A}{4 \pi G}$$

$$E = mc^2 \quad E^2 = (pc)^2 + (mc^2)^2$$

$$r = \frac{\theta}{2\pi} + \frac{4\pi}{q^2}$$

$$I = \int e^{-\alpha x^2/2} dx = \sqrt{\frac{2\pi}{\alpha}}$$

$$E^2 = p^2 c^2 + m^2 c^4 = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \psi - \nabla^2 \psi + \frac{m^2 c^2}{\hbar^2} \psi = 0$$

$$P = \hbar k = \frac{\hbar v}{c} = \frac{\hbar}{\lambda}$$

$$S = \frac{1}{2} \int d^4x \left(R + \frac{R^2}{6M^2} \right)$$

$$\Omega_m = 10$$

Scientific theory



Computational science

BIG DATA

Data science

Data?!



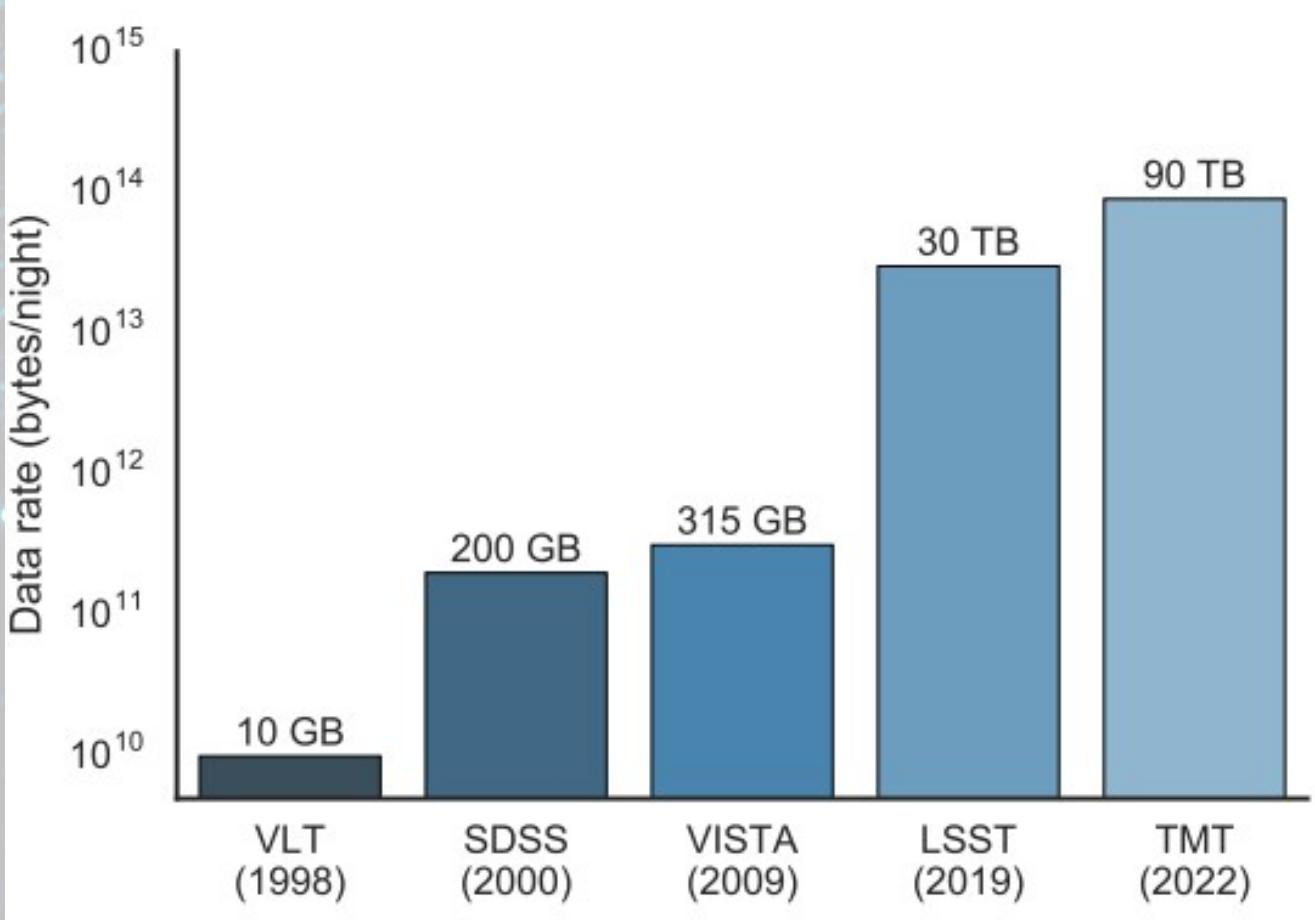
How BIG?

- New telescopes collect today 50 times the info they collected 5 years ago.
- Google process 24 PB per day = US library of congress X1000
- 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/>



Agriculture



Climate



Consumer



Ecosystems



Education



Energy



Finance



Health



Local
Government



Manufacturing



Maritime



Ocean



Public Safety



Science &
Research

<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://www.census.gov/data.html>

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

<https://www.yelp.com/dataset>

<https://www.google.com/finance>

<https://data.unicef.org/>

<https://wiki.dbpedia.org/>

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

<https://data.worldbank.org/>

<https://www.kaggle.com/datasets>

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

<https://lodum.de/>

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

<https://registry.opendata.aws/>

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

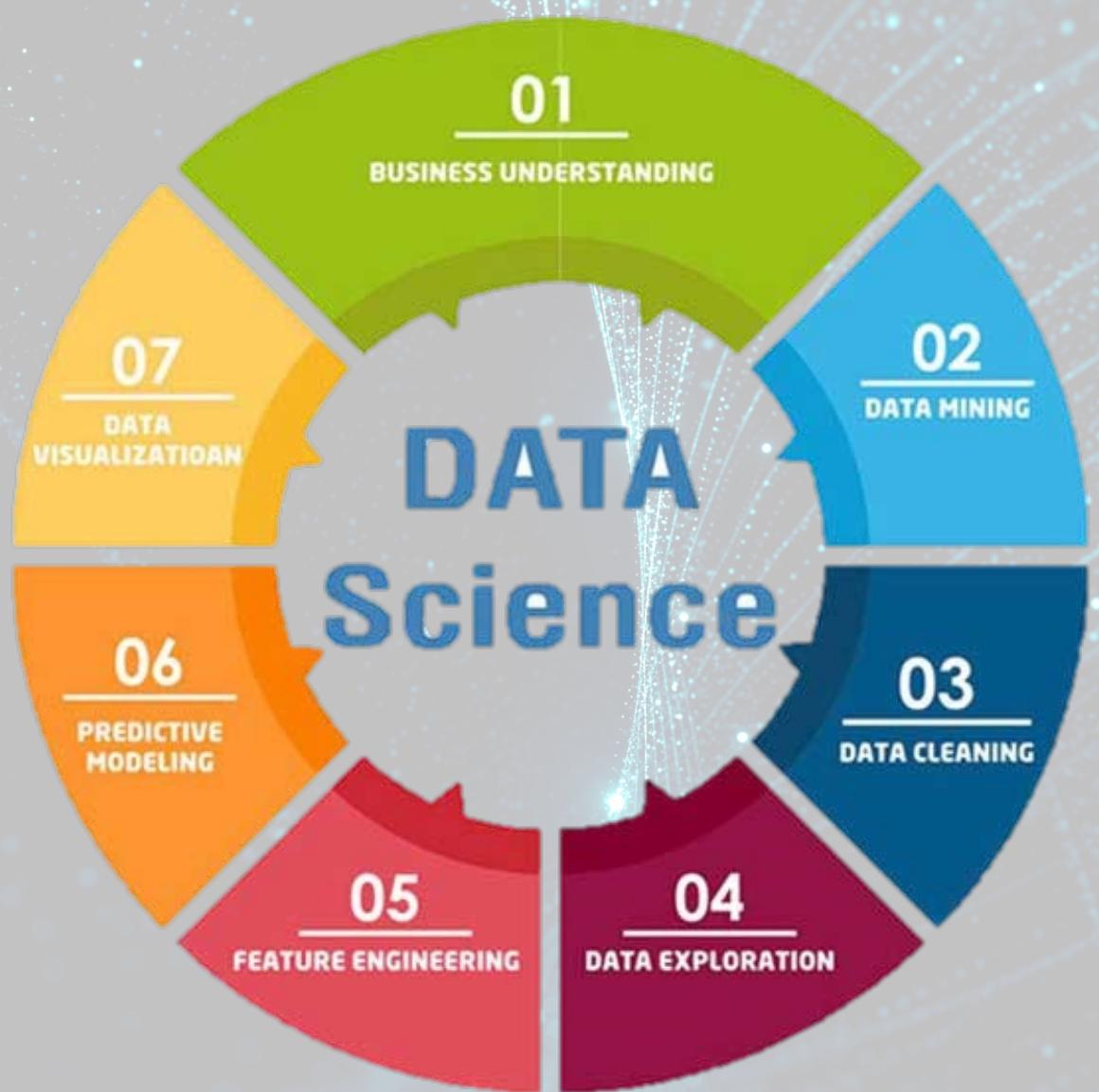
<https://data.fivethirtyeight.com/>

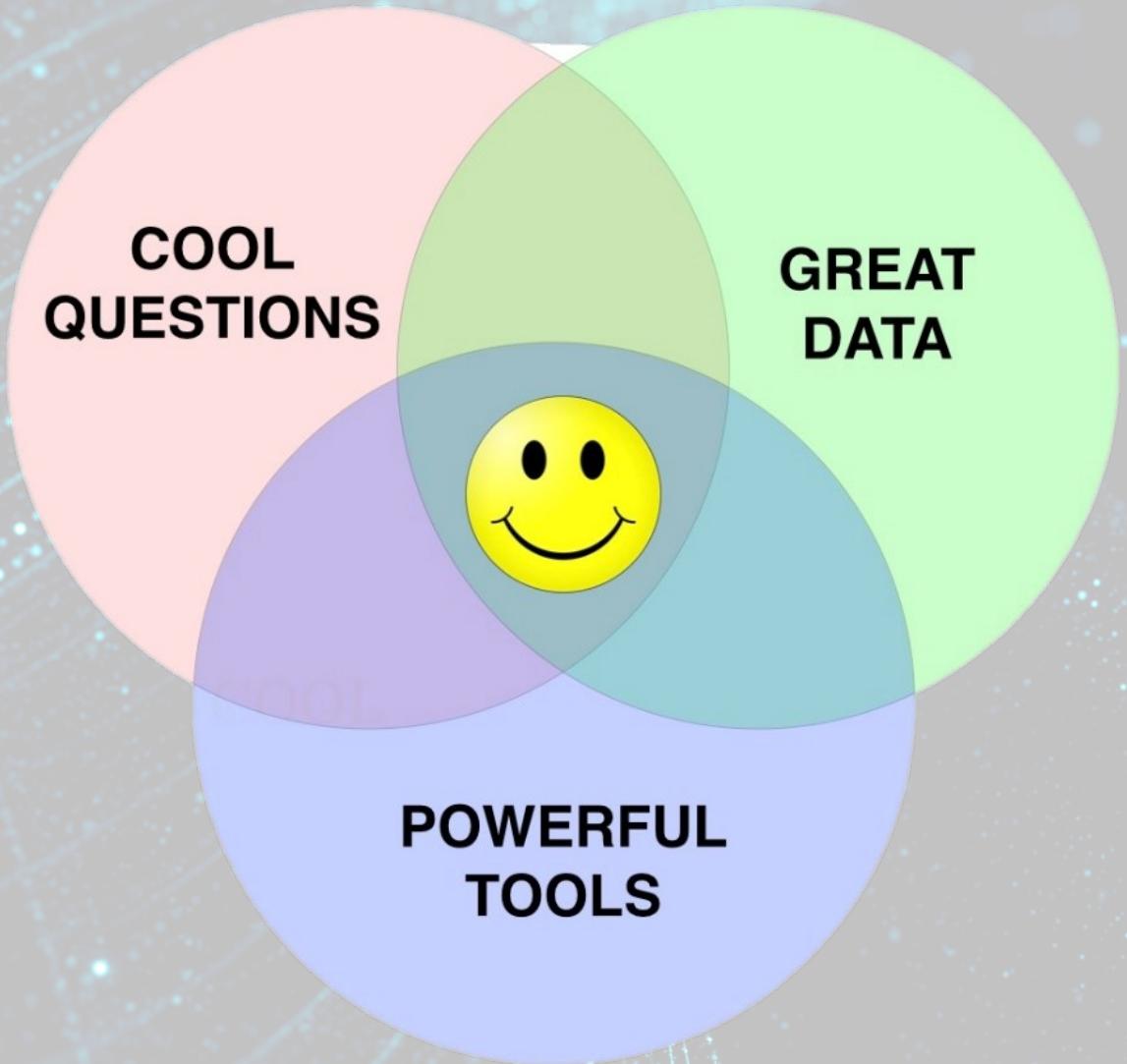


Science?!

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!?





Better half a loaf than no bread.



دريا را اگر نتوان کشيد هم به قدر تشنگي باید چشید

Creativity!



CAN: Creative Adversarial Networks

Generating “Art” by Learning About Styles and Deviating from Style Norms*

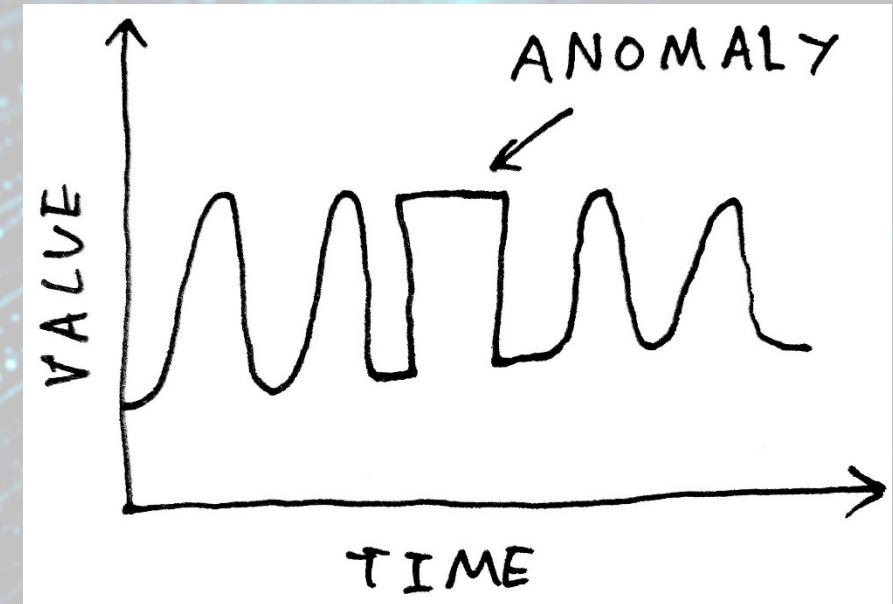
Ahmed Elgammal^{1†} Bingchen Liu¹ Mohamed Elhoseiny² Marian Mazzone³



Creativity!



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



Understanding!?



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 manner
- 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

Computer:

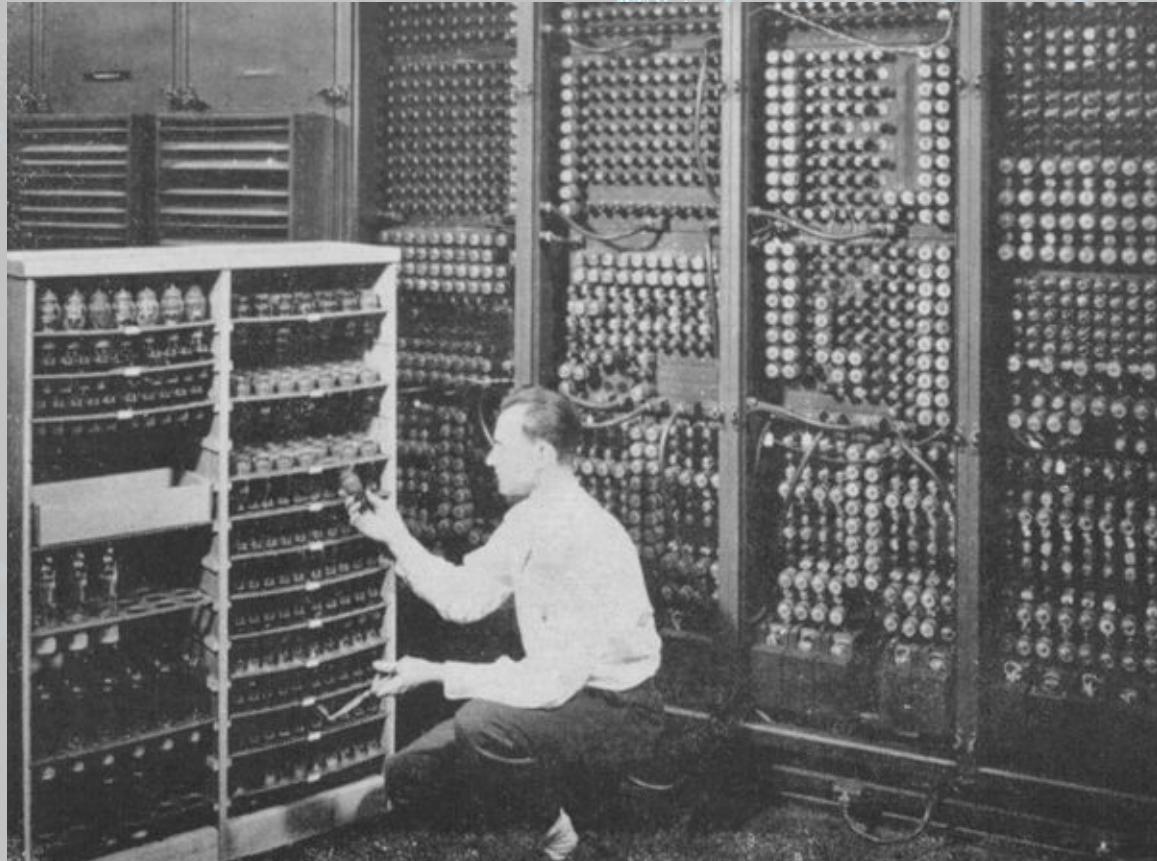


1613-1950

Computer:

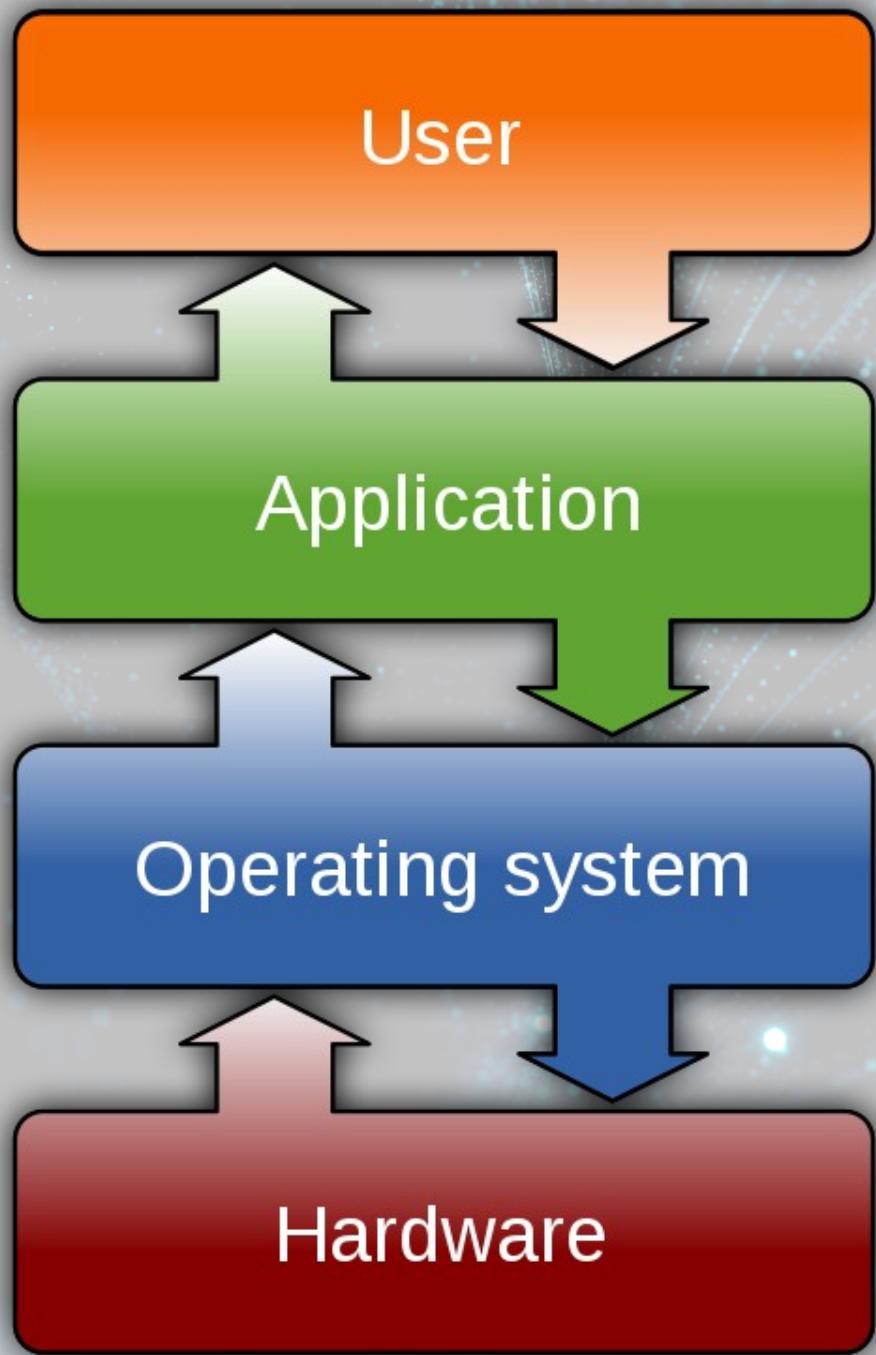
- Logical
- Arithmetic

Computer:



Computer:







**NOT SURE IF I AM A GOOD
PROGRAMMING**

**OR GOOD AT
GOOGLING**

LET'S DIVE!



Thanks!