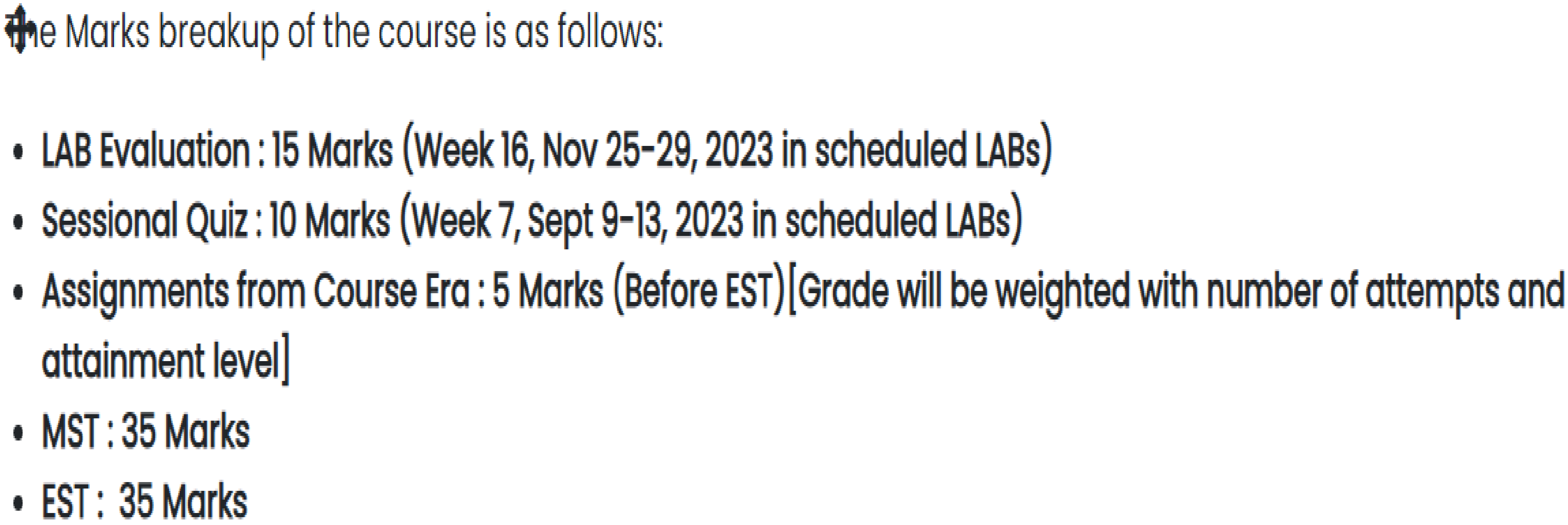
Foundation of Data Science

UCS548

Course Information

- Syllabus Copy
- Academic Calendar
- Attendance Requirement: **>75%**

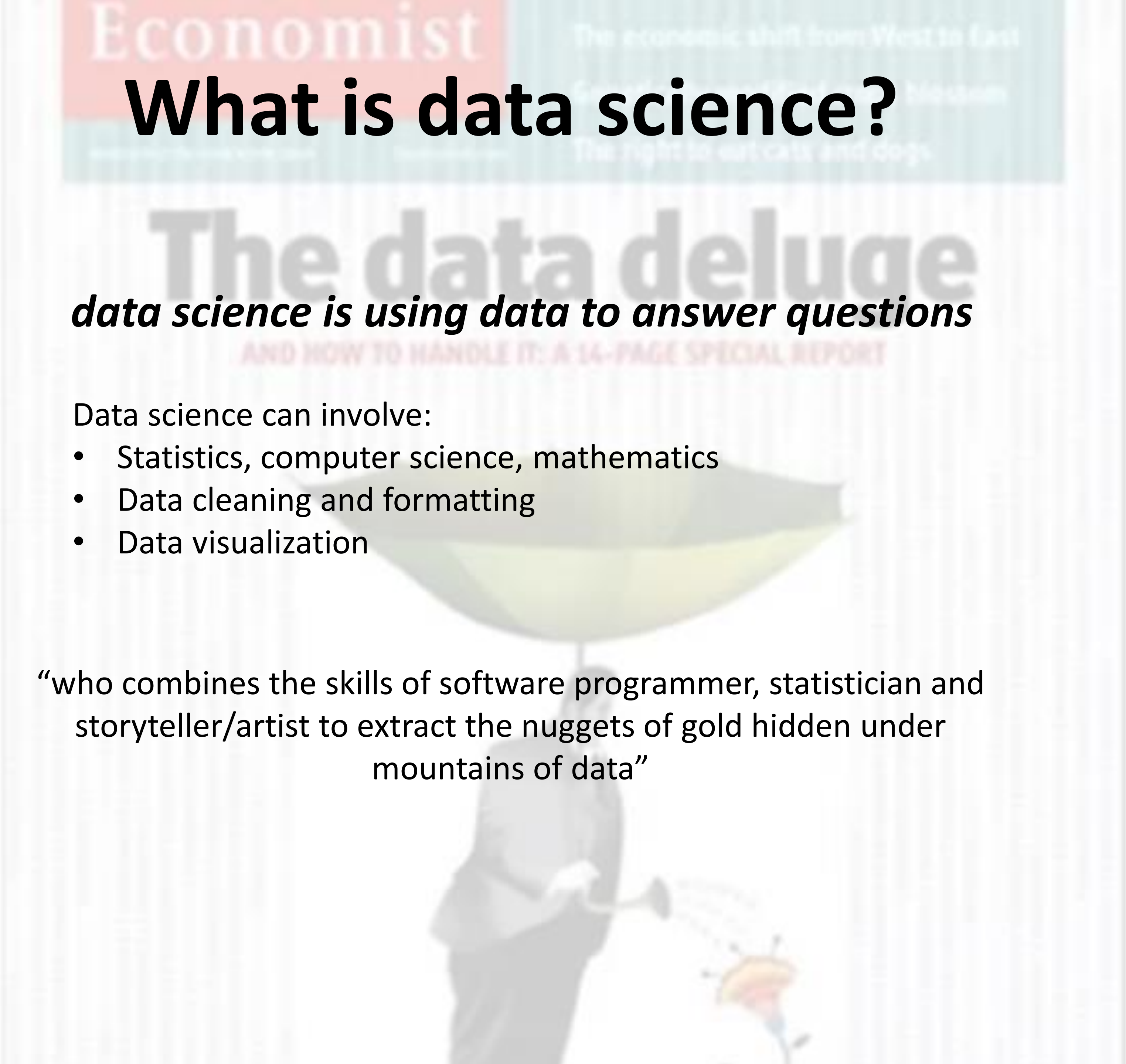
Tentative Evaluation Plana



The Marks breakup of the course is as follows:

- LAB Evaluation : 15 Marks (Week 16, Nov 25-29, 2023 in scheduled LABs)
- Sessional Quiz : 10 Marks (Week 7, Sept 9-13, 2023 in scheduled LABs)
- Assignments from Course Era : 5 Marks (Before EST)[Grade will be weighted with number of attempts and attainment level]
- MST : 35 Marks
- EST : 35 Marks

What is data science?


Economist
The economic shift from West to East
blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

data science is using data to answer questions

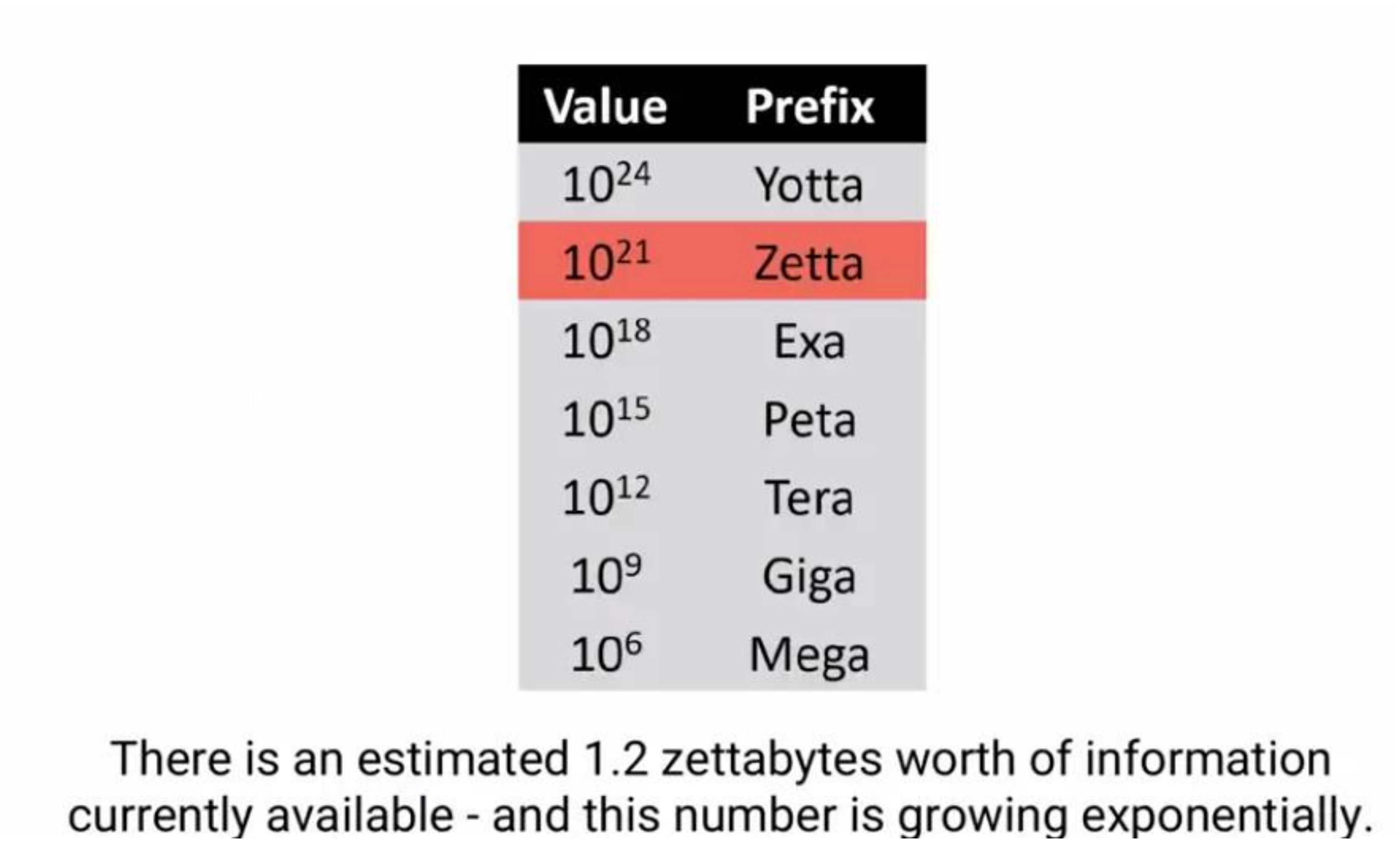
Data science can involve:

- Statistics, computer science, mathematics
- Data cleaning and formatting
- Data visualization

"who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data"

- **Data Science is an interdisciplinary field that focuses on extracting knowledge from data sets which are typically huge in amount.**
- **The field encompasses analysis, preparing data for analysis, and presenting findings to inform high-level decisions in an organization.**
- **As such, it incorporates skills from computer science, mathematics, statics, inform.**
- **Some of the techniques utilized in Data Science encompass machine learning, visualization, pattern recognition, probability modeling data, data engineering, signal processing, etc.**

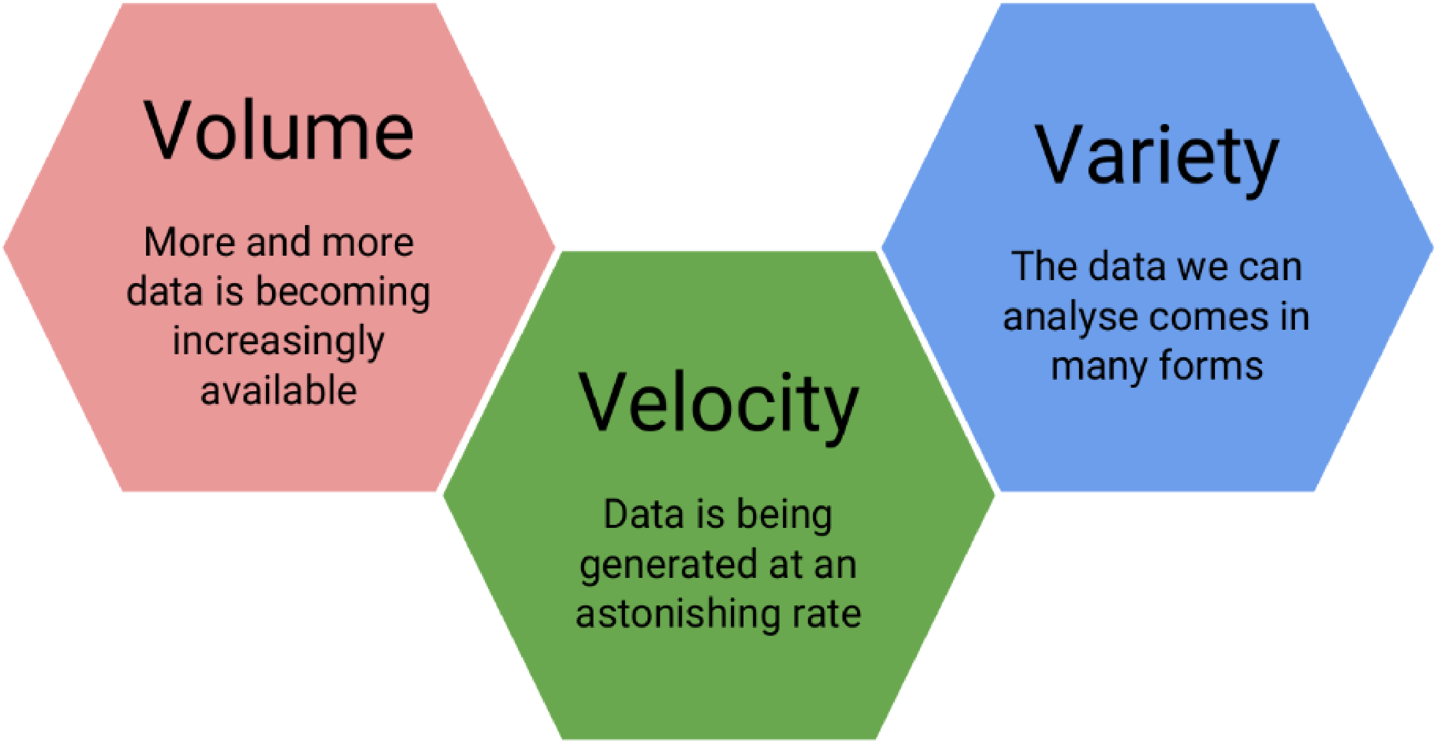
Why do we need data science?



Value	Prefix
10^{24}	Yotta
10^{21}	Zetta
10^{18}	Exa
10^{15}	Peta
10^{12}	Tera
10^{9}	Giga
10^{6}	Mega

There is an estimated 1.2 zettabytes worth of information currently available - and this number is growing exponentially.

What is big data?



Volume

More and more data is becoming increasingly available

Velocity

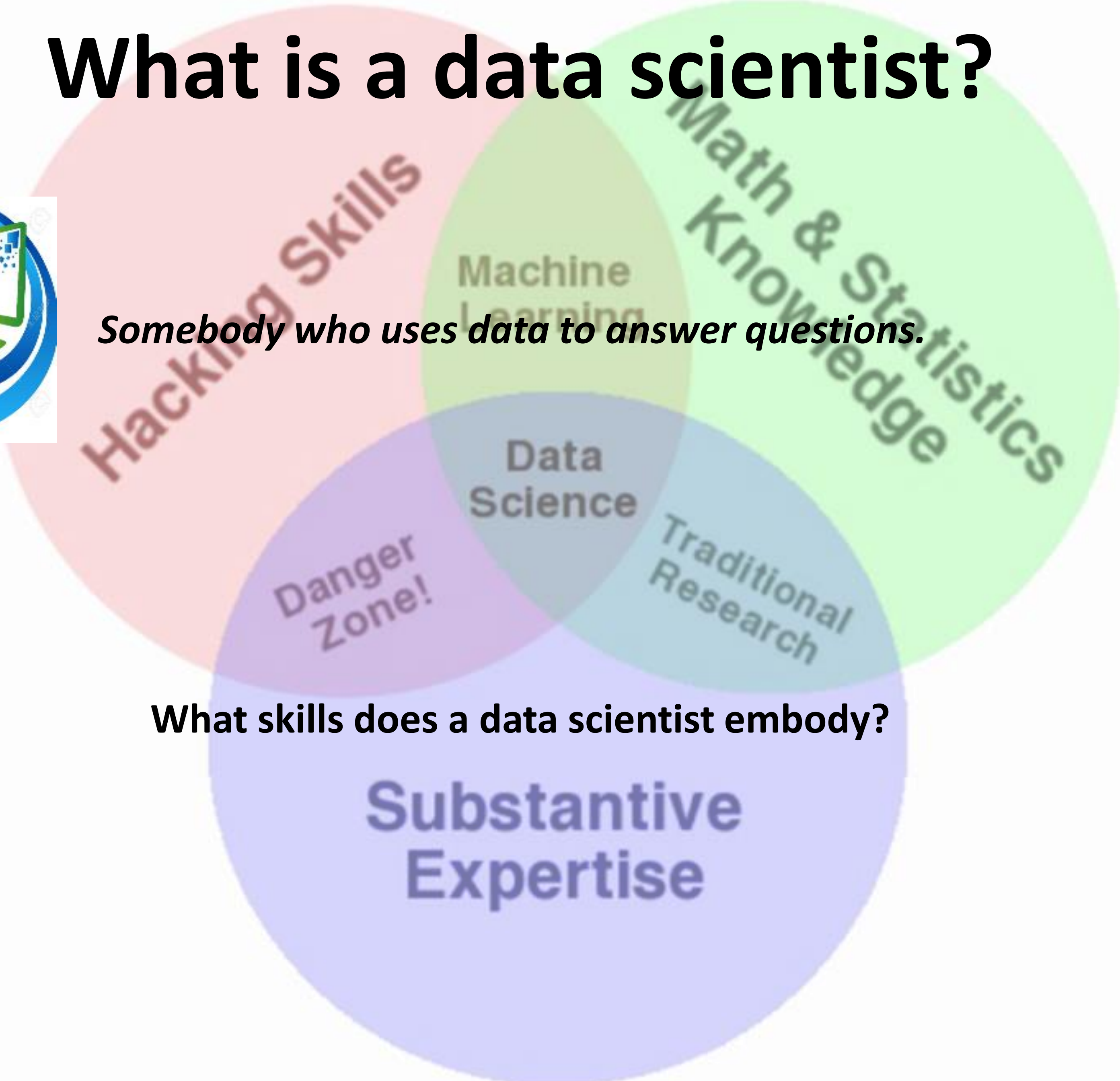
Data is being generated at an astonishing rate

Variety

The data we can analyse comes in many forms

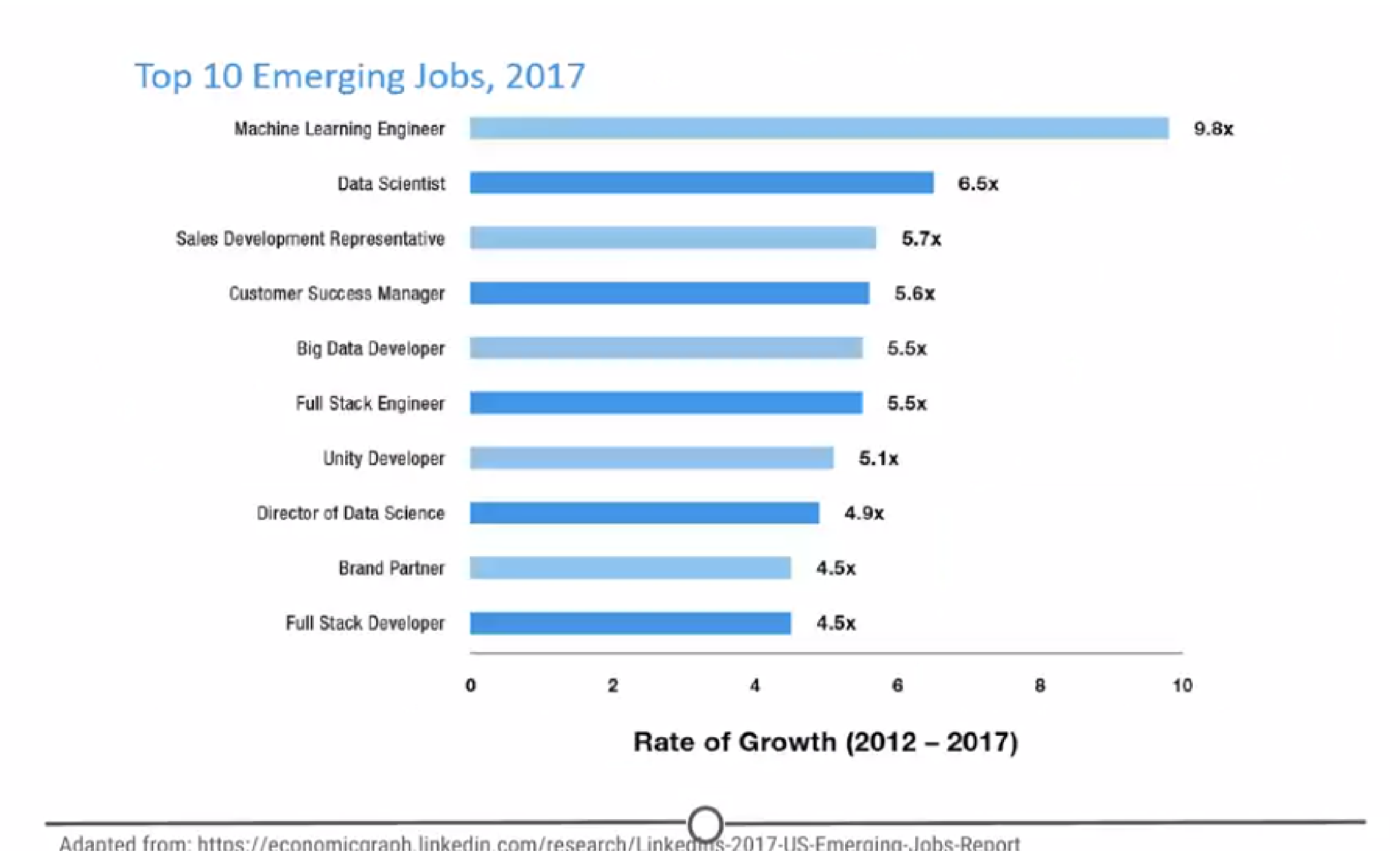
What is a data scientist?

Somebody who uses data to answer questions.



What skills does a data scientist embody?

Why do data science?

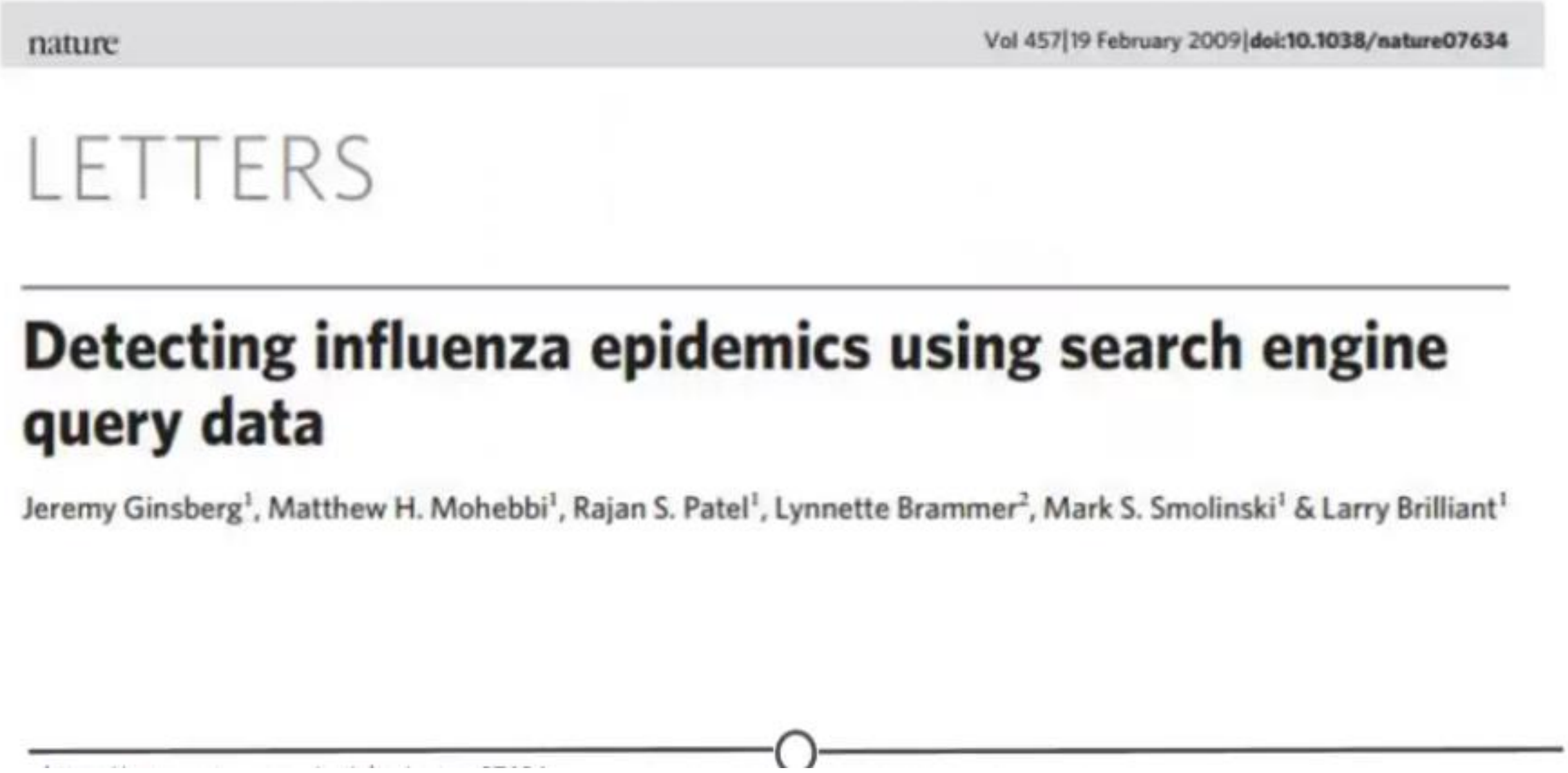


Adapted from: https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report

According to Glassdoor, in which they ranked the top 50 best jobs in America, Data Scientist is THE top job in the US in 2017, based on job satisfaction, salary, and demand.

Examples of data scientists

Data science in action!



nature

Vol 457|19 February 2009|**doi:10.1038/nature07634**

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Prerequisites for Data Science

The following are the three essential traits of Data Scientist:

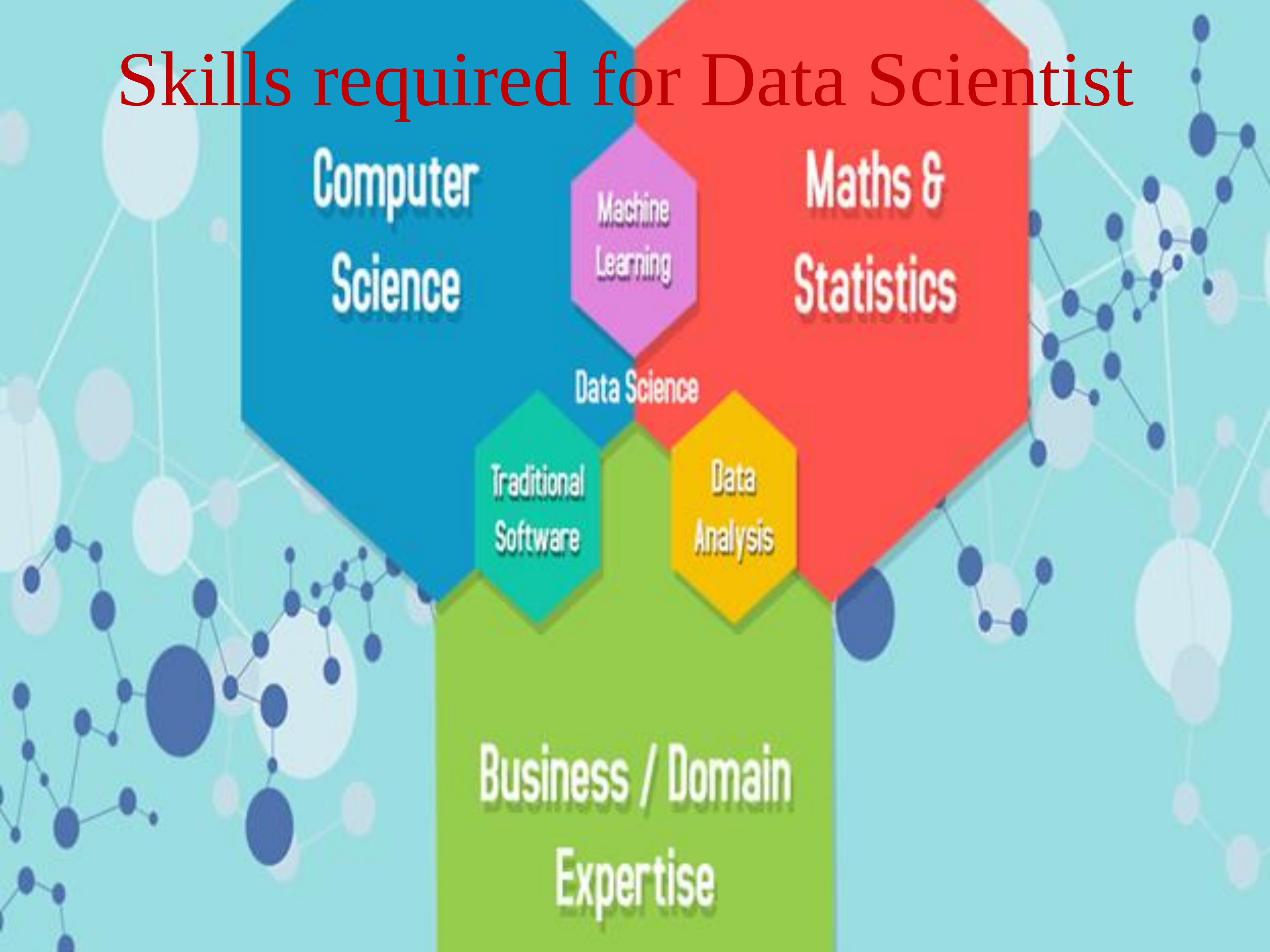
Curiosity

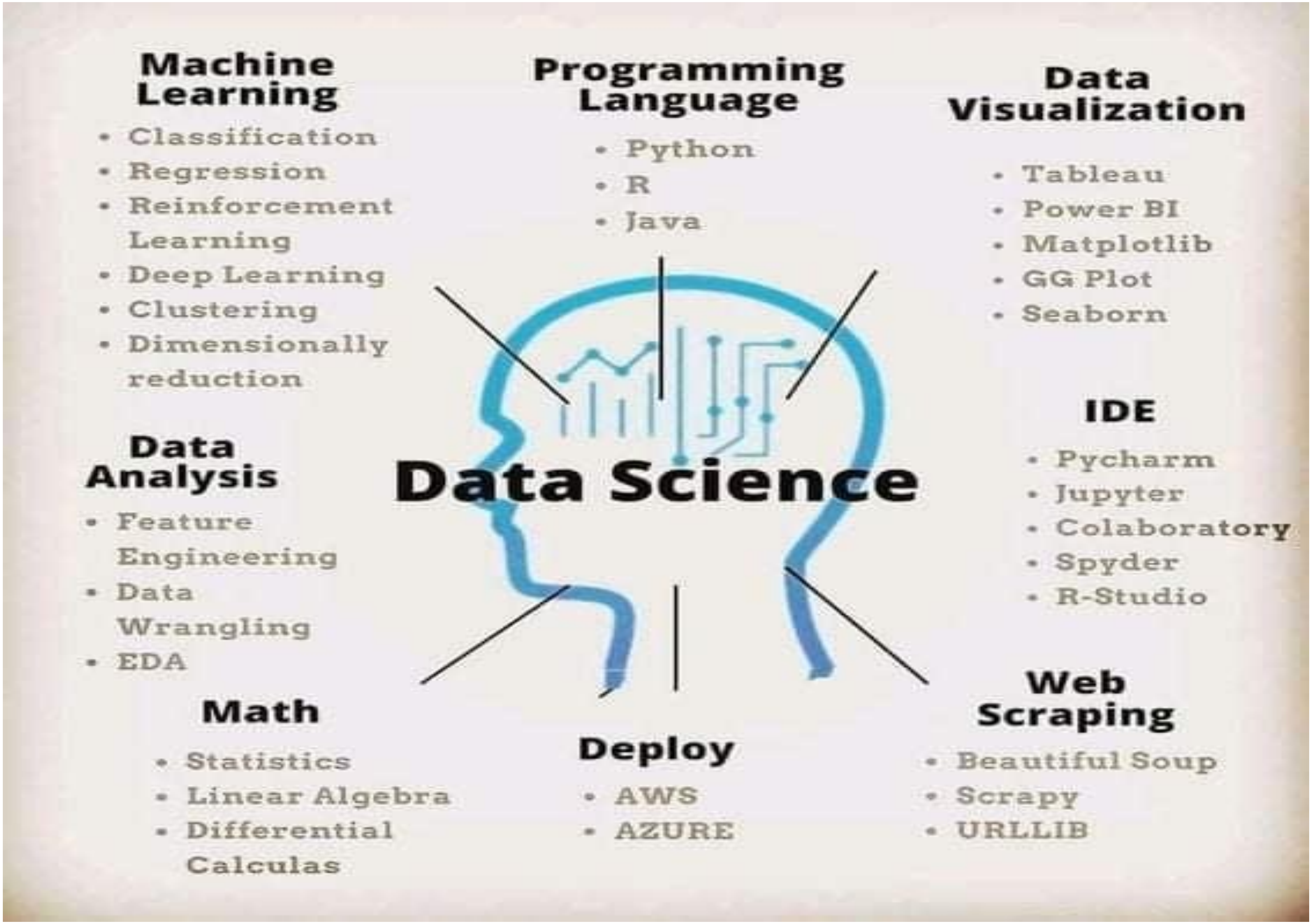
+

+

Curiosity: Only when you ask questions, you will have a better understanding of the business problem.
Common Sense: To identify new *ways to solve a business problems* and to detect priority problems.
Communication Skills: A Data Scientist needs to *communicate their findings* to business teams to act upon the insights

Skills required for Data Scientist





Data Science

Machine Learning

- Classification
- Regression
- Reinforcement Learning
- Deep Learning
- Clustering
- Dimensionally reduction

Programming Language

- Python
- R
- Java

Data Visualization

- Tableau
- Power BI
- Matplotlib
- GG Plot
- Seaborn

Data Analysis

- Feature Engineering
- Data Wrangling
- EDA

IDE

- Pycharm
- Jupyter
- Colaboratory
- Spyder
- R-Studio

Math

- Statistics
- Linear Algebra
- Differential Calculas

Deploy

- AWS
- AZURE

Web Scraping

- Beautiful Soup
- Scrapy
- URLLIB

Data Science Three skill tracks:

Engineering , Analysis, Modeling

Engineering

- **Involves in building the data pipeline infrastructure.**
- **It involves the software and the hardware used to store the data and perform data ETL (i.e., extract, transform, and load).**
- **Store and compute data on the cloud.**
- **The fundamental building block for automation is maintaining the data pipeline through modular, well-commented code and version control.**
- **Key task involved are:-**

Engineering

- **Key task involved are:-**

1. ***Data Environment*: Designing and setting up the entire environment to support data science workflow is the prerequisite for data science projects. It may include setting up storage in the cloud, Kafka platform, Hadoop and Spark cluster, etc**
2. ***Data Management*: Automated data collection, that includes parsing the logs (depending on the stage of the company and the type of industry you are in), web scraping, API queries, and interrogating data streams. Determine and construct data schema to support analytical and modeling needs. Use tools, processes, guidelines to ensure data is correct, standardized, and documented.**
3. ***Production*: Involves the whole pipeline from data access, preprocessing, modeling to final deployment. It is necessary to make the system work smoothly with all existing software stacks.**

Data Science Three skill tracks:

Analysis

- **Analysis turns raw information into insights in a fast and often exploratory way.**
- **In general, an analyst needs to have decent domain knowledge, do exploratory analysis efficiently, and present the results using storytelling.**
- **Key point includes are:-**

1. ***Domain Knowledge*: understanding of the organization or industry where you apply data science. You can’t make sense of data without context.**

2. ***Exploratory Analysis*: team look at as much data as possible so that the decision-makers can get a sense of what’s worth further pursuing. It often involves different ways to slice and aggregate data.**

3. ***Storytelling*: It is the art of telling people what the numbers signify. It usually requires data summarization, aggregation, and visualization. It is crucial to answering the following questions before you begin down the path of creating a data story.**
   - **Who is your audience?**
   - **What do you want your audience to know or do?**
   - **How can you use data to help make your point?**

Data Science Three skill tracks:

Modeling

- **A process that dives deeper into the data to discover the pattern we don’t readily see.**
- **A model only occupy a small part of a typical data scientist’s day-to-day time.**
- **Some of the models are :-**

1. ***Supervised Learning*: In supervised learning, each sample corresponds to a response measurement. There are two flavors of supervised learning: regression and classification.**
   - ✓ **In regression, the response is a real number, such as the total net sales in 2017 for a company or the yield of corn next year for a state.**
   - ✓ **The goal for regression is to approximate the response measurement as much as possible.**
   - ✓ **In classification, the response is a class label, such as a dichotomous response of yes/no.**
   - ✓ **The response can also have more than two categories, such as four segments of customers**

1. ***Un-supervised Learning*: In unsupervised learning, there is no response variable. Clustering approach is used for data analysis.**
2. ***Customized model development*: A data scientist may need to develop new models to accommodate the subtleties of the problem at hand. For example, people may use Bayesian models to include domain knowledge as the modeling process's prior distribution.**

Data Science Prerequisites

What type of problem you are solving?

***Description*:**

- The primary analytic problem is to summarize and explore a data set with descriptive statistics (mean, standard deviation, and so forth) and visualization methods.
- Data description is often used to check data, find the appropriate data preprocessing method, and demonstrate the model results.

***Comparison* :**

- The first common modeling problem is to compare different groups. Is A better in some way than B? Or more comparisons: Is there any difference among A, B, and C in a particular aspect?
- The commonly used statistical tests are chisquare test, t-test, and ANOVA. There are also methods using Bayesian methods.

Data Science Prerequisites

***Clustering* :**

- Please note that clustering is unsupervised learning; there are no response variables. The most common clustering algorithms include K-Means and Hierarchical Clustering.

***Classification* :**

- For classification problems, there are one or more label columns to define the ground truth of classes. We use other features of the training dataset as explanatory variables for model training. We can use the trained classifier to predict the labels of a new observation.
- The random forest algorithm is usually used as the baseline model to set model performance expectations.

Data Science Prerequisites

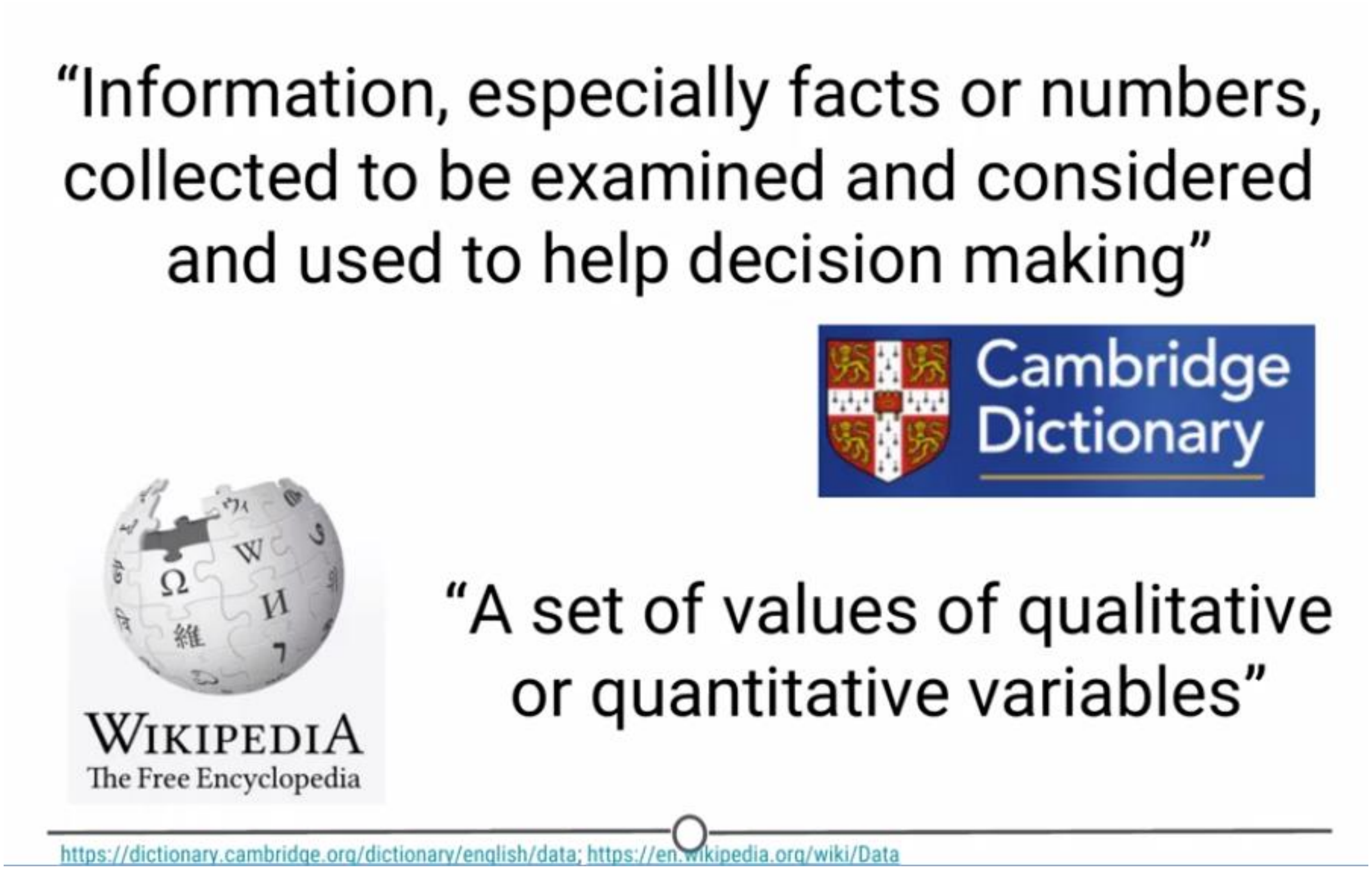
***Regression* :**

- Generally used for prediction and to answer the questions:-
- What will be the temperature tomorrow? What is the projected net income for the next season? How much inventory should we have?

***Optimization* :**

- It is an expansion of comparison problem and can solve problems such as:
- What is the best route to deliver the packages? What is the optimal advertisement strategy to promote a new product?.

What is data?



"Information, especially facts or numbers, collected to be examined and considered and used to help decision making"

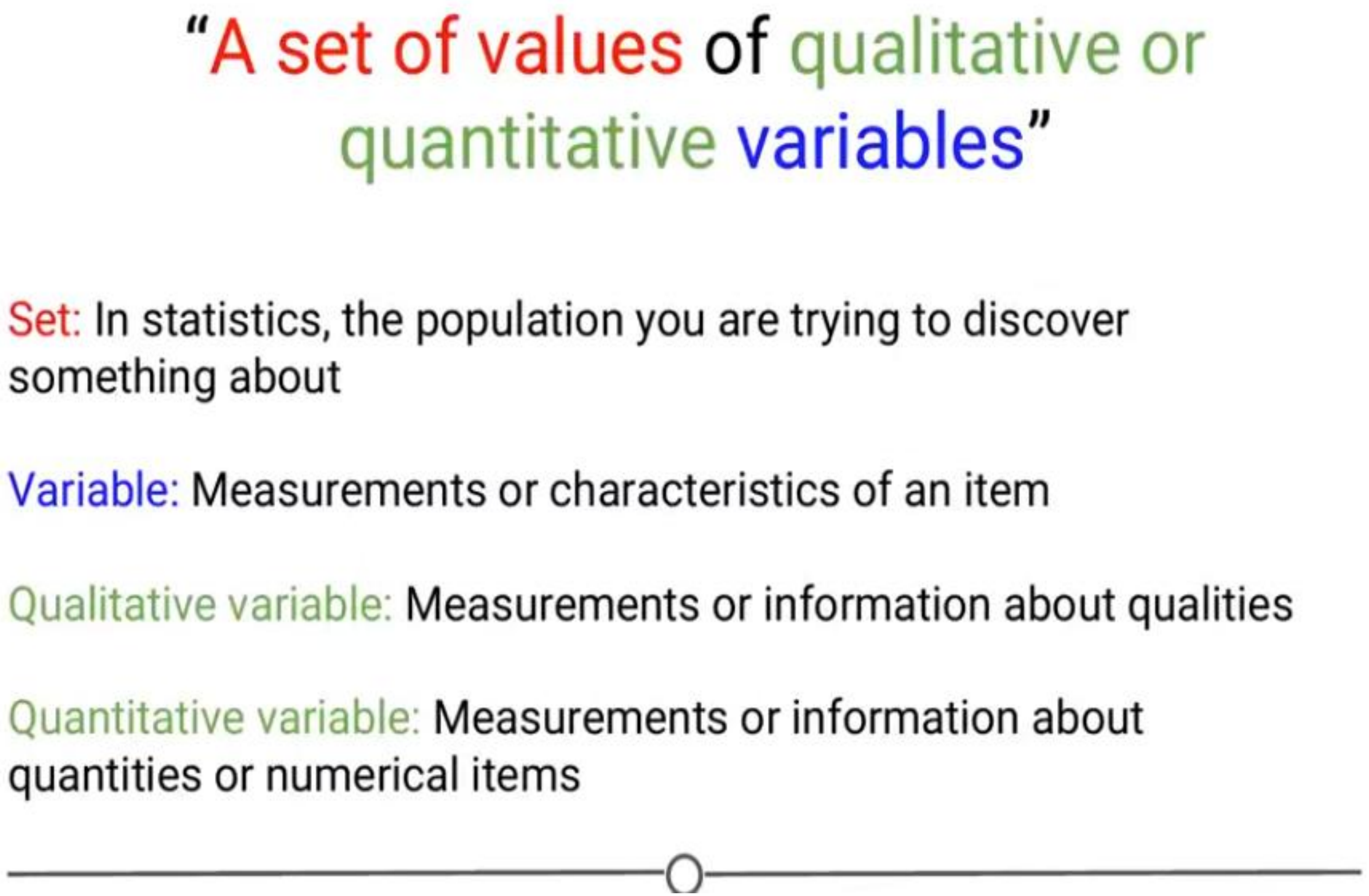
Cambridge Dictionary

WIKIPEDIA
The Free Encyclopedia

"A set of values of qualitative or quantitative variables"

https://dictionary.cambridge.org/dictionary/english/data; https://en.wikipedia.org/wiki/Data

What is data?



“A set of values of qualitative or quantitative variables”

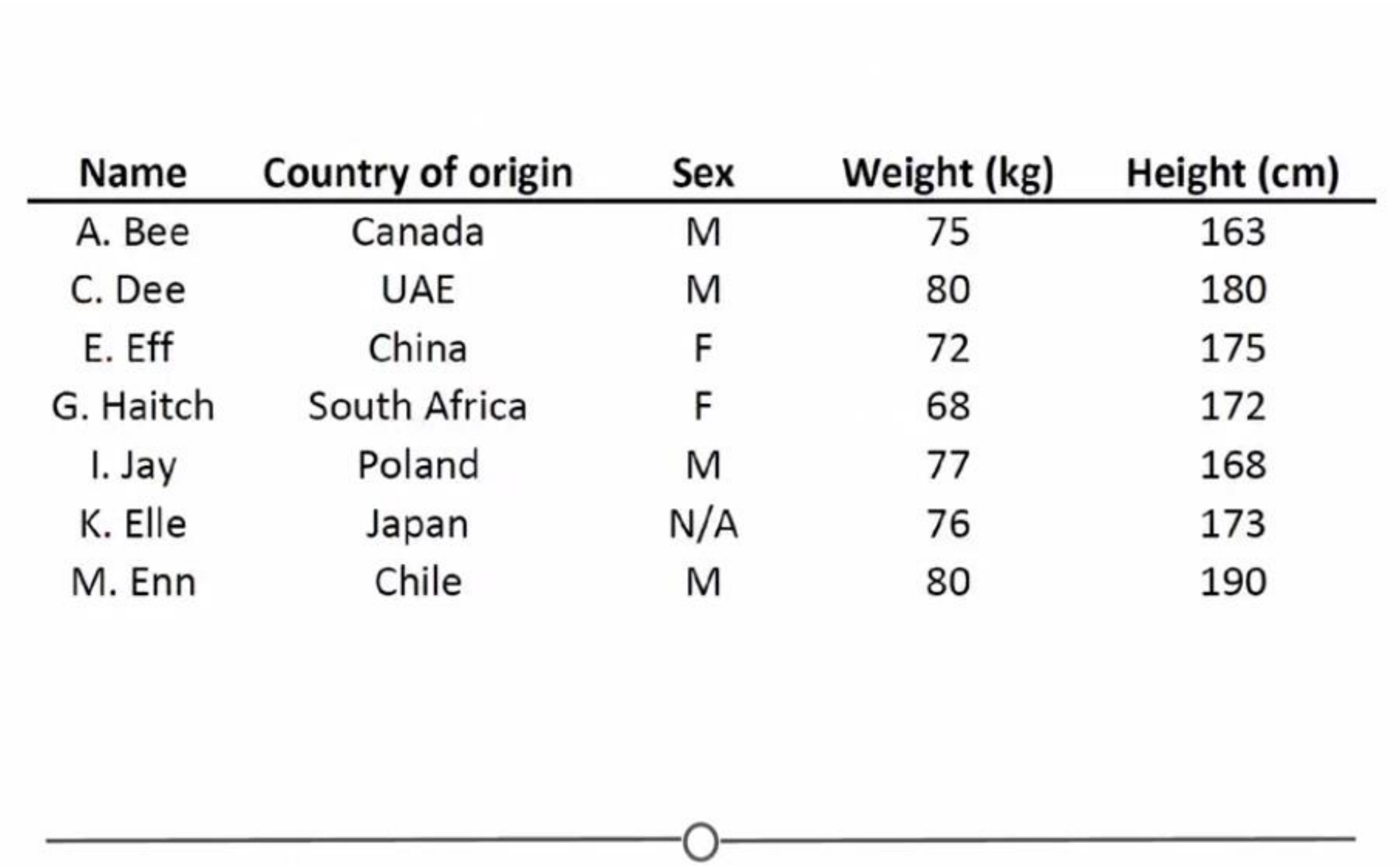
Set: In statistics, the population you are trying to discover something about

Variable: Measurements or characteristics of an item

Qualitative variable: Measurements or information about qualities

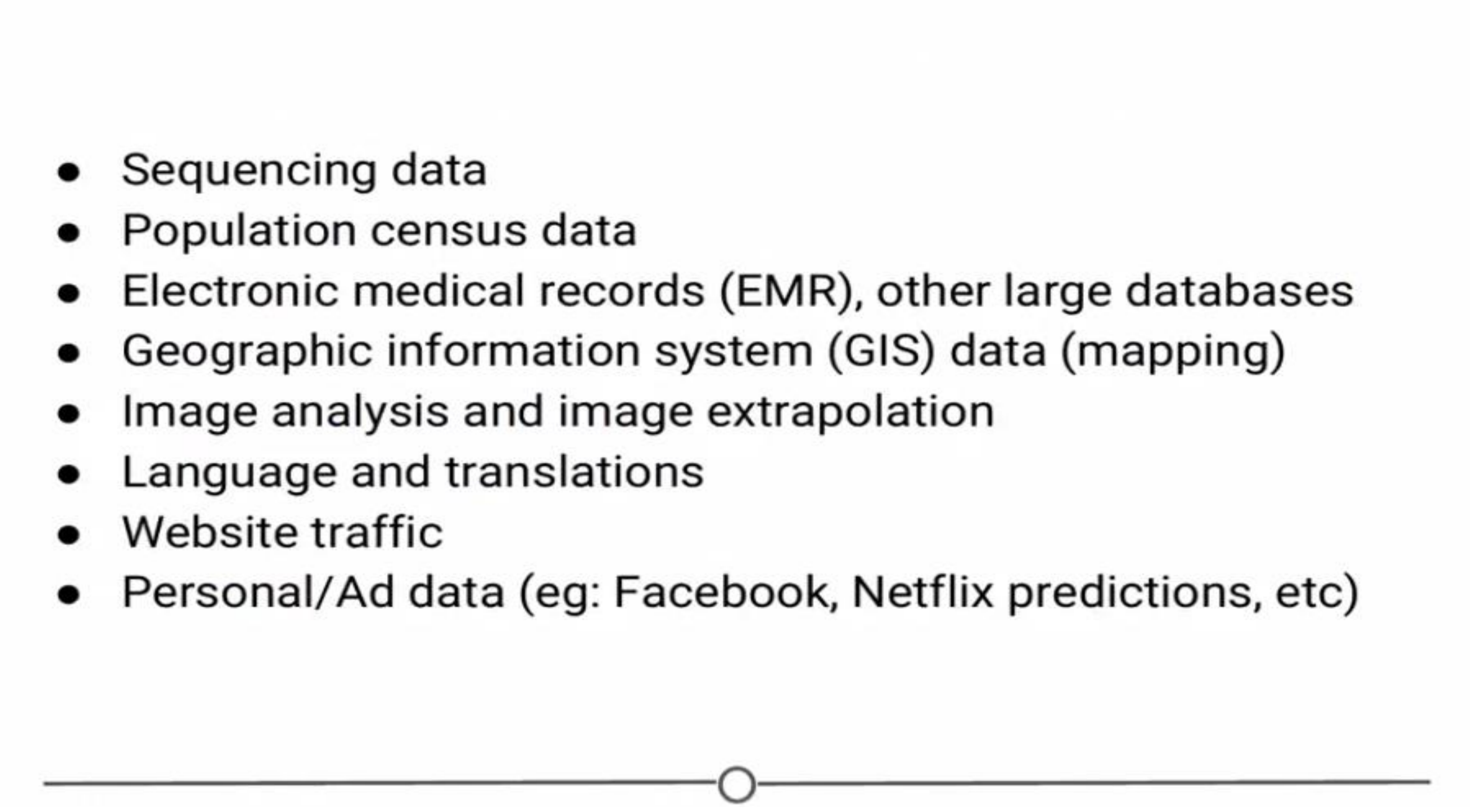
Quantitative variable: Measurements or information about quantities or numerical items

What can data look like? (rarely)



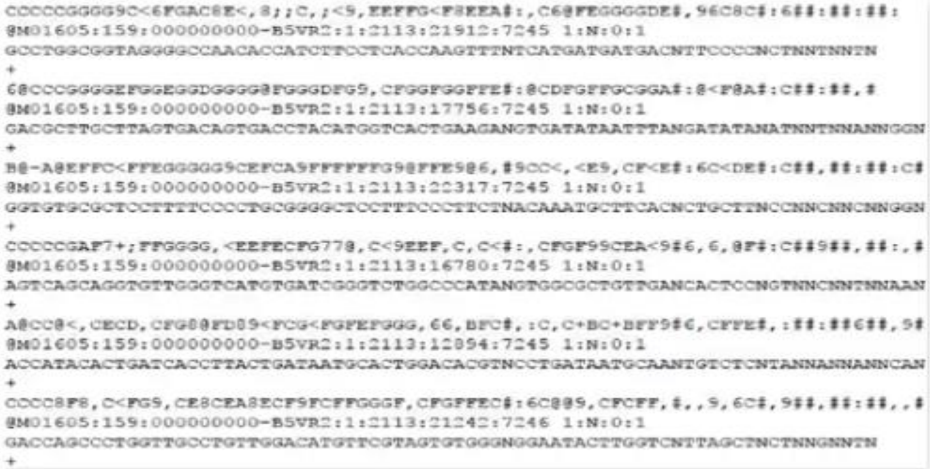
Name	Country of origin	Sex	Weight (kg)	Height (cm)
A. Bee	Canada	M	75	163
C. Dee	UAE	M	80	180
E. Eff	China	F	72	175
G. Haitch	South Africa	F	68	172
I. Jay	Poland	M	77	168
K. Elle	Japan	N/A	76	173
M. Enn	Chile	M	80	190

More common types of messy data

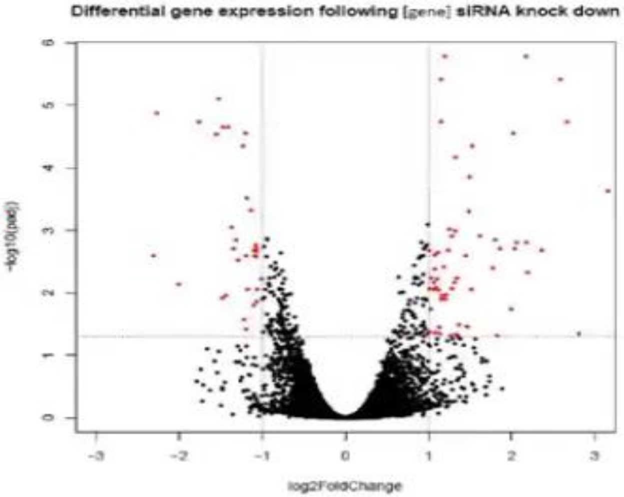


- Sequencing data
- Population census data
- Electronic medical records (EMR), other large databases
- Geographic information system (GIS) data (mapping)
- Image analysis and image extrapolation
- Language and translations
- Website traffic
- Personal/Ad data (eg: Facebook, Netflix predictions, etc)

Messy data: Sequencing

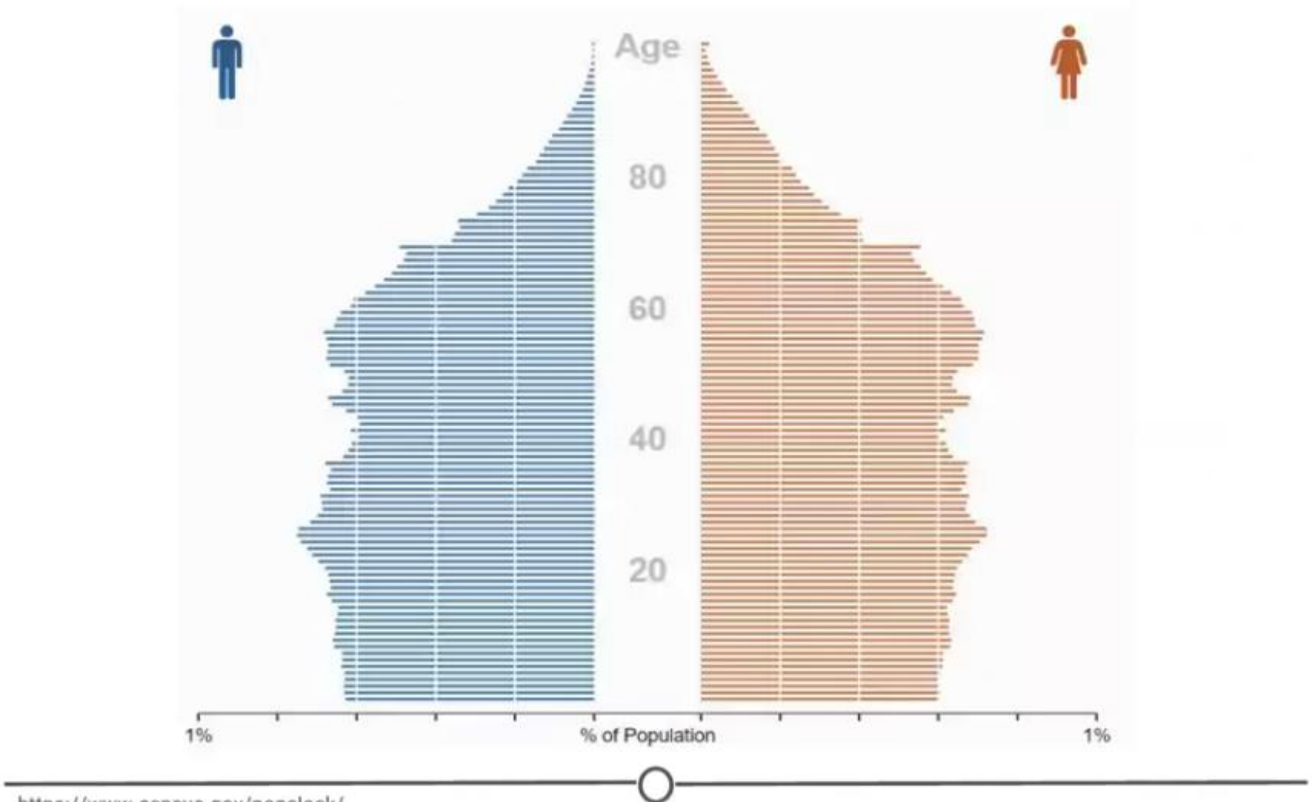


```
CCCCCGGGG9C<6FGAC8E<,8;;C,;<9,EEFFG<F8EEA#:,C6@FEGGGGDE#,96C8C#:6##:##:##:
@M01605:159:000000000-B5VR2:1:2113:21912:7245 1:N:0:1
GCCTGGCGGTAGGGGCCAACACCATCTTCCTCACCAAGTTTNTCATGATGATGACNTTCCCCNCTNNTNNTN
+
6@CCCGGGGEFGGEGGDGGGG@FGGGDFG9,CFGGFGGFFE#:@CDFGFFGCGGA#:@<F8A#:C##:##,#
@M01605:159:000000000-B5VR2:1:2113:17756:7245 1:N:0:1
GACGCTTGCTTAGTGACAGTGACCTACATGGTCACTGAAGANGTGATATAATTTANGATATANATNNTNNANNGGN
+
B@-A@EFFC<FFEGGGGG9CEFCA9FFFFFFG9@FFE9@6,#9CC<,<E9,CF<E#:6C<DE#:C##,##:##:C#
@M01605:159:000000000-B5VR2:1:2113:22317:7245 1:N:0:1
GGTGTGCGCTCCTTTTCCCCTGCGGGGCTCCTTTCCCTTCTNACAAATGCTTCACNCTGCTTNCCNNCNNCNNGGN
+
CCCCCGAF7+;FFGGGG,<EEFECFG778,C<9EEF,C,C<#:,CFGF99CEA<9#6,6,@F#:C##9##,##:,#
@M01605:159:000000000-B5VR2:1:2113:16780:7245 1:N:0:1
AGTCAGCAGGTGTTGGGTCATGTGATCGGGTCTGGCCCATANGTGGCGCTGTTGANCACTCCNGTNNCNNTNNAAN
+
A@CC@<,CECD,CFG8@FD89<FCG<FGFEFGGG,66,BFC#,:C,C+BC+BFF9#6,CFFE#,:##:##6##,9#
@M01605:159:000000000-B5VR2:1:2113:12894:7245 1:N:0:1
ACCATACACTGATCACCTTACTGATAATGCACTGGACACGTNCCTGATAATGCAANTGTCTCNTANNANNANNCAN
+
CCCC8F8,C<FG9,CE8CEA8ECF9FCFFGGGF,CFGFFEC#:6C@@9,CFCFF,#,,9,6C#,9##,##:##,,#
@M01605:159:000000000-B5VR2:1:2113:21242:7246 1:N:0:1
GACCAGCCCTGGTTGCCTGTTGGACATGTTCGTAGTGTGGGNGGAATACTTGGTCNTTAGCTNCTNNGNNTN
+
```



A volcano plot is produced at the end of a long process to wrangle the raw FASTQ data into interpretable expression data

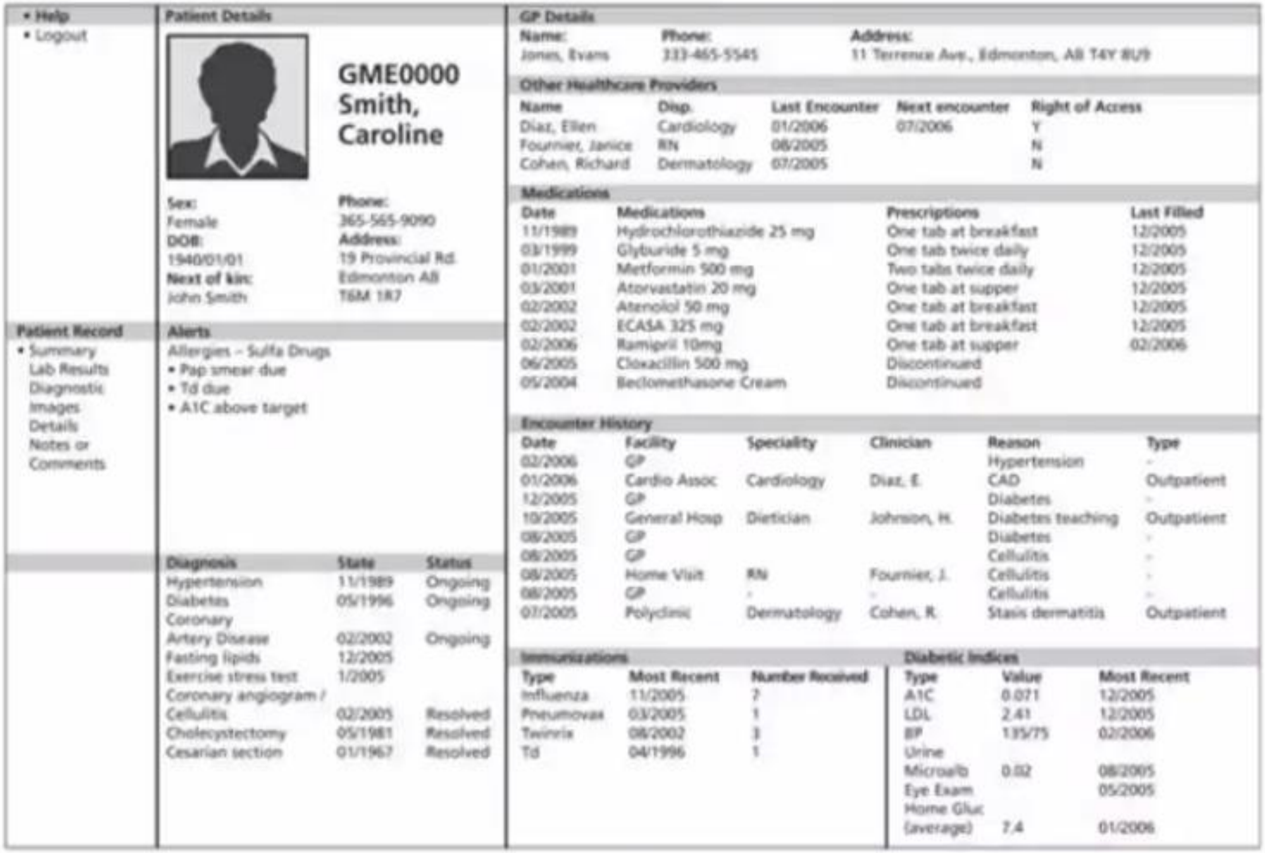
Messy data: Census information



https://www.census.gov/popclock/

A population pyramid plot

Messy data: Electronic medical records (EMR)



- Help
- Logout

Patient Details

GME0000
Smith, Caroline

Sex:	**Phone:**
Female	365-565-9090
DOB:	**Address:**
1940/01/01	19 Provincial Rd.
Next of kin:	Edmonton AB
John Smith	T6M 1R7

Patient Record

- Summary
  Lab Results
  Diagnostic Images
  Details
  Notes or Comments

Alerts

Allergies – Sulfa Drugs
- Pap smear due
- Td due
- A1C above target

Diagnosis	State	Status
Hypertension	11/1989	Ongoing
Diabetes	05/1996	Ongoing
Coronary Artery Disease	02/2002	Ongoing
Fasting lipids	12/2005	
Exercise stress test	1/2005	
Coronary angiogram /		
Cellulitis	02/2005	Resolved
Cholecystectomy	05/1981	Resolved
Cesarian section	01/1967	Resolved

GP Details

Name:	Phone:	Address:
Jones, Evans	333-465-5545	11 Terrence Ave., Edmonton, AB T4Y 8U9

Other Healthcare Providers

Name	Disp.	Last Encounter	Next encounter	Right of Access
Diaz, Ellen	Cardiology	01/2006	07/2006	Y
Fournier, Janice	RN	08/2005		N
Cohen, Richard	Dermatology	07/2005		N

Medications

Date	Medications	Prescriptions	Last Filled
11/1989	Hydrochlorothiazide 25 mg	One tab at breakfast	12/2005
03/1999	Glyburide 5 mg	One tab twice daily	12/2005
01/2001	Metformin 500 mg	Two tabs twice daily	12/2005
03/2001	Atorvastatin 20 mg	One tab at supper	12/2005
02/2002	Atenolol 50 mg	One tab at breakfast	12/2005
02/2002	ECASA 325 mg	One tab at breakfast	12/2005
02/2006	Ramipril 10mg	One tab at supper	02/2006
06/2005	Cloxacillin 500 mg	Discontinued	
05/2004	Beclomethasone Cream	Discontinued	

Encounter History

Date	Facility	Speciality	Clinician	Reason	Type
02/2006	GP			Hypertension	-
01/2006	Cardio Assoc	Cardiology	Diaz, E.	CAD	Outpatient
12/2005	GP			Diabetes	-
10/2005	General Hosp	Dietician	Johnson, H.	Diabetes teaching	Outpatient
08/2005	GP			Diabetes	-
08/2005	GP			Cellulitis	-
08/2005	Home Visit	RN	Fournier, J.	Cellulitis	-
08/2005	GP	-	-	Cellulitis	-
07/2005	Polyclinic	Dermatology	Cohen, R.	Stasis dermatitis	Outpatient

Immunizations

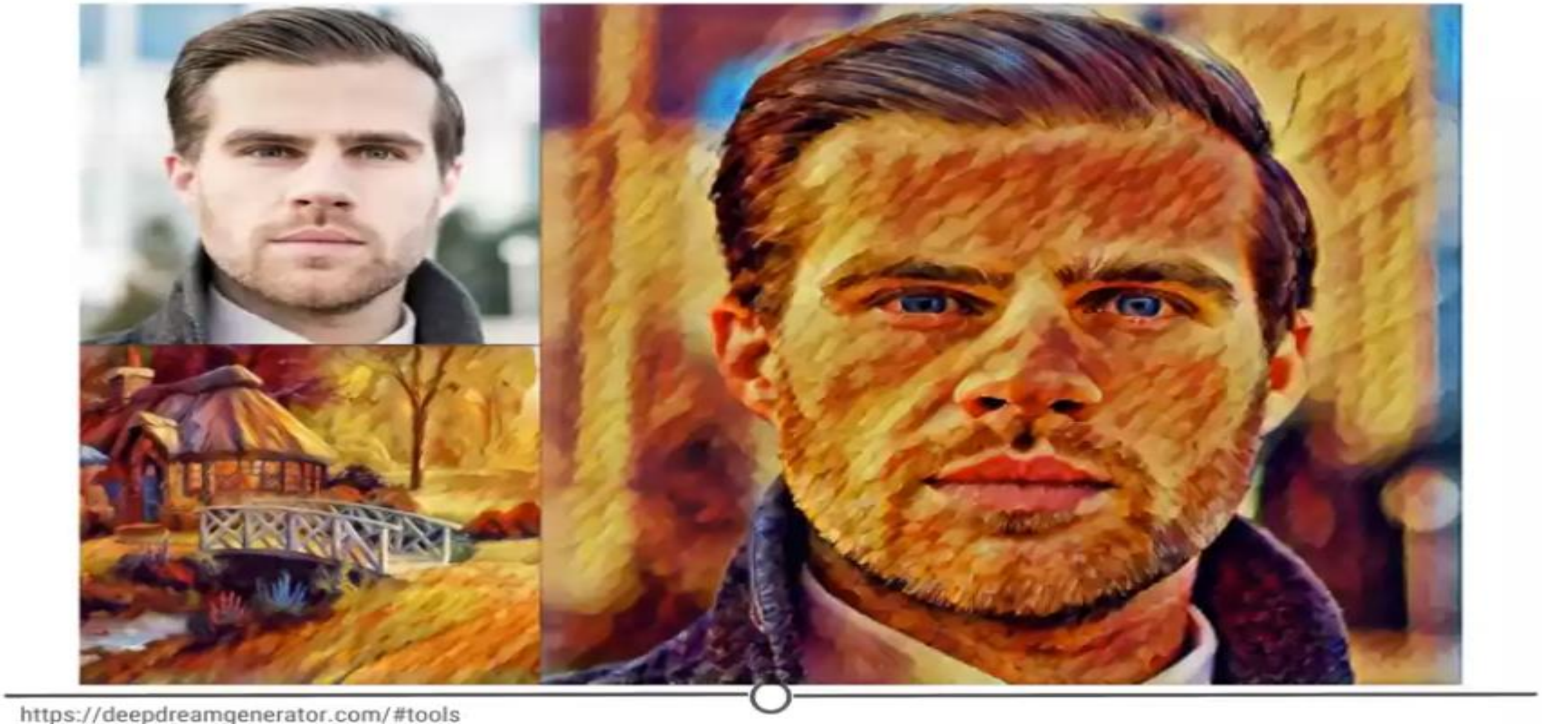
Type	Most Recent	Number Received
Influenza	11/2005	7
Pneumovax	03/2005	1
Twinrix	08/2002	3
Td	04/1996	1

Diabetic Indices

Type	Value	Most Recent
A1C	0.071	12/2005
LDL	2.41	12/2005
BP	135/75	02/2006
Urine Microalb	0.02	08/2005
Eye Exam		05/2005
Home Gluc (average)	7.4	01/2006

http://www.oaq-bvq.qc.ca/internet/English/parl_oaq_201004_07_e_33720.html

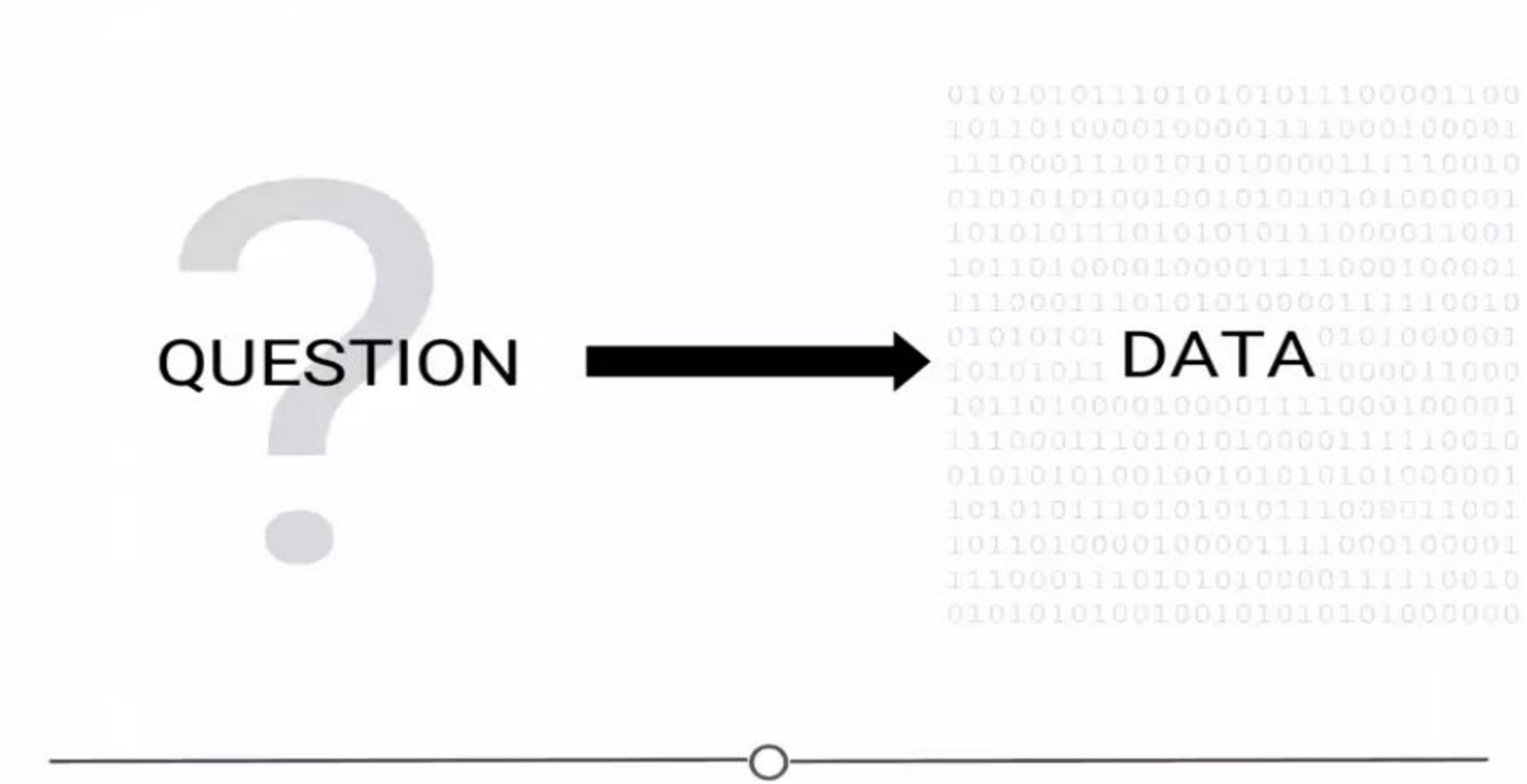
Messy data: Image analysis/extrapolation



https://deepdreamgenerator.com/#tools

The DeepDream software is trained on your image and a famous painting and your provided image is then rendered in the style of the famous painter

Data is of secondary importance



A good data scientist asks questions first and seeks out relevant data second.