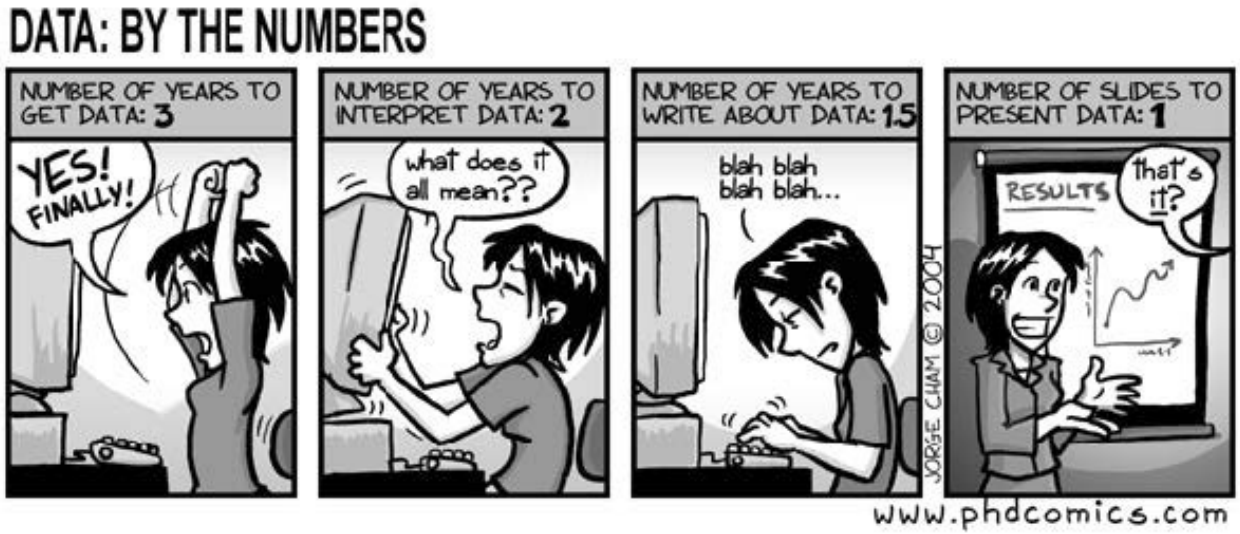
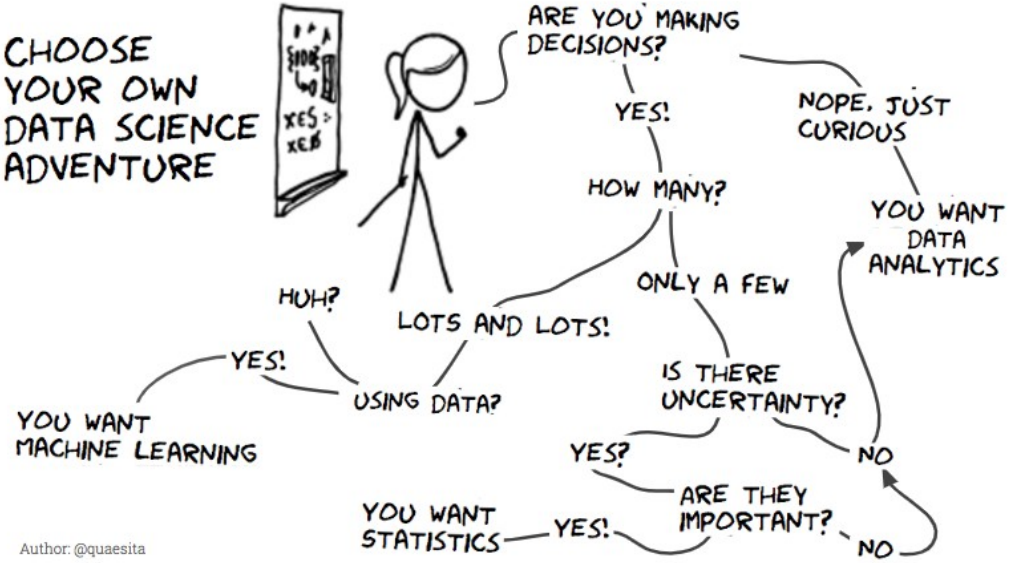


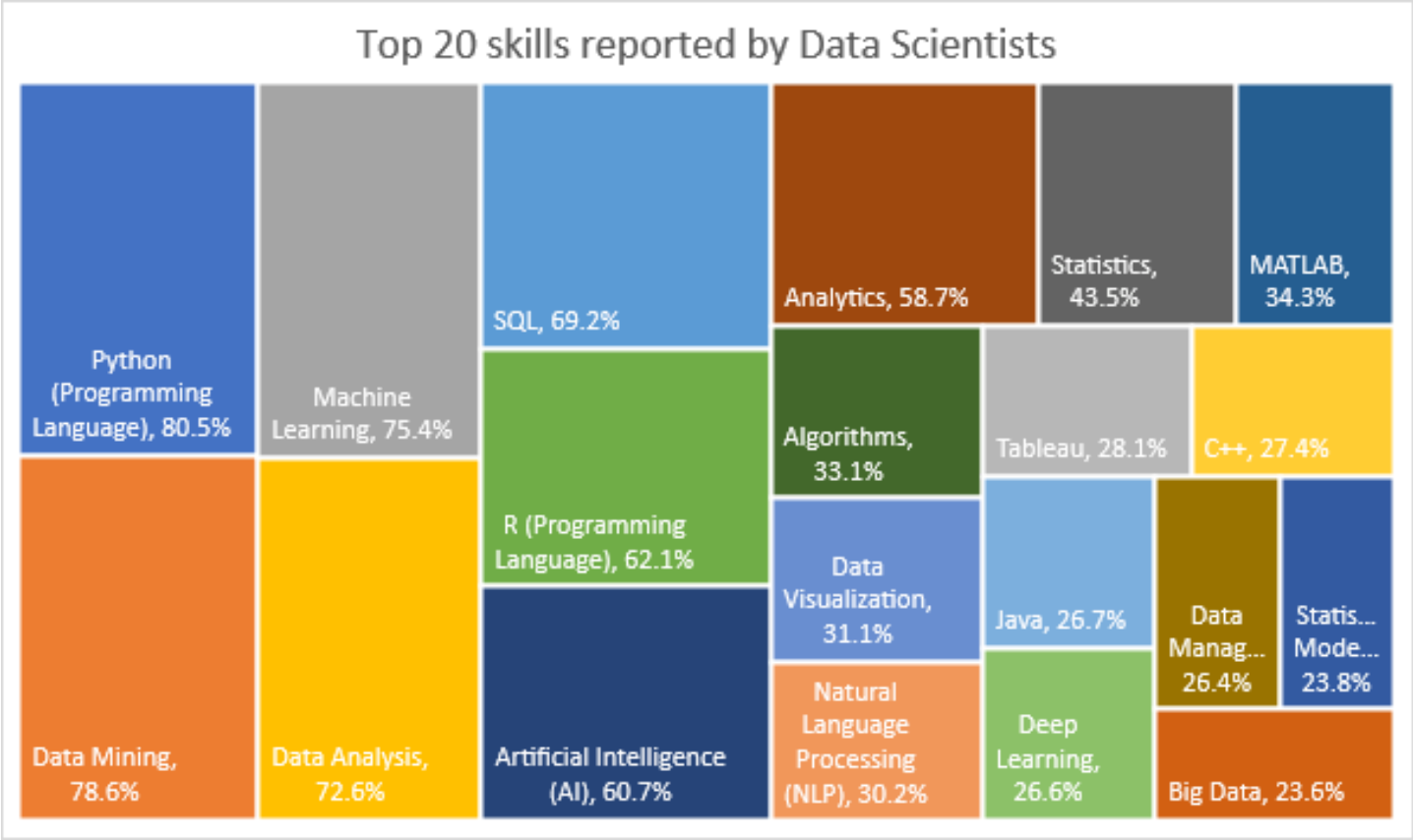
Data Science: intro, skills, works, explore, produce, interpret,…

...make sense for a particular problem, generating added value...

D. Sierra-Porta (2022)

Two pictures to understand our challenge!



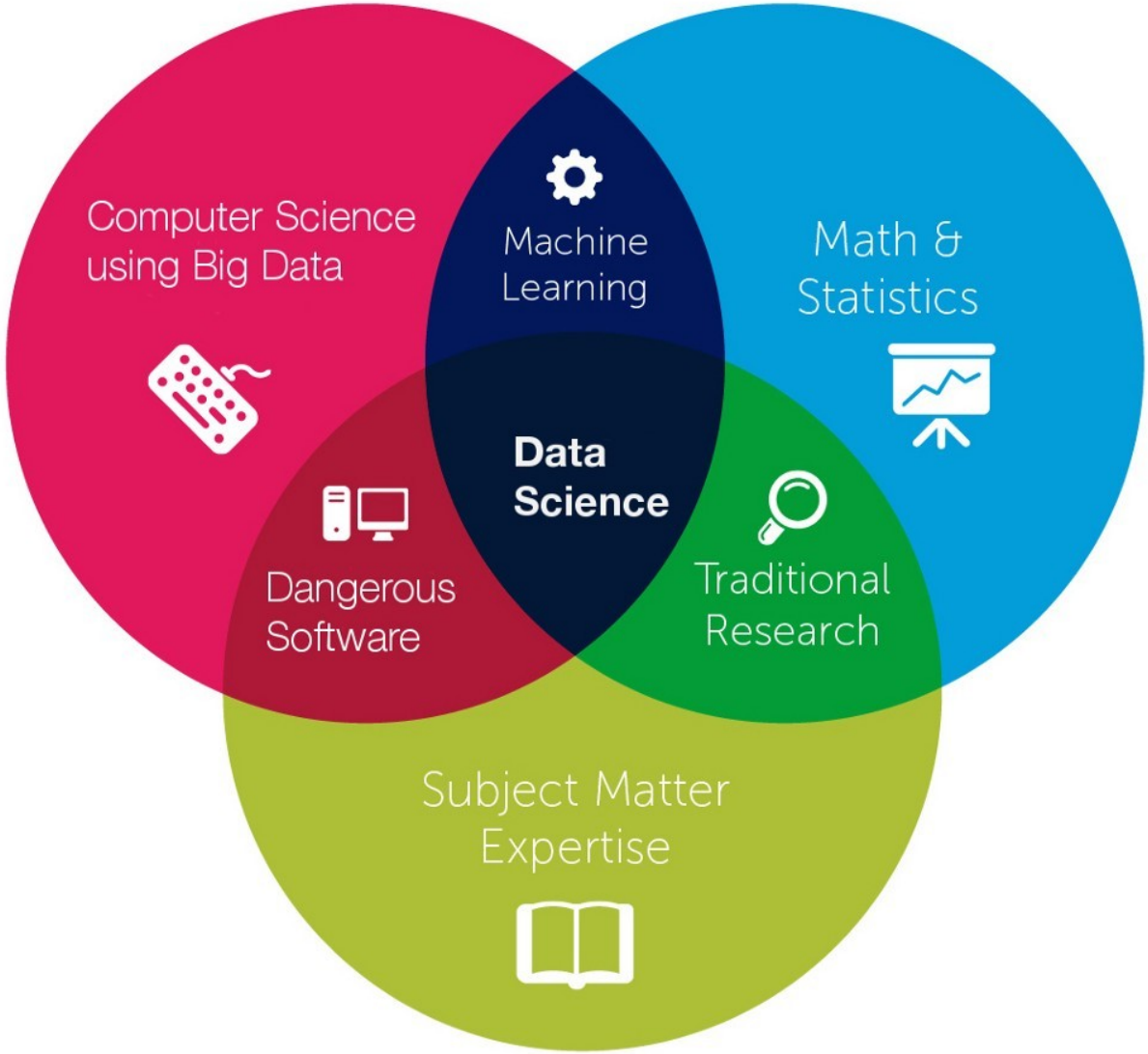


- Figure summarizes the **skills of 200 data scientists** from leading technology destinations such as the San Francisco Bay Area, New York City Metropolitan Area, Seattle, Dallas-Fort Worth Metroplex, Raleigh-Durham-Chapel Hill Area, Greater Chicago Area, Greater Boston, London Area, Bengaluru, and New Delhi.

- It is abundantly clear from Figure that the top skill that most data scientists have is Python, followed by data mining.

https://www.red-gate.com/simple-talk/development/data-science-development/how-to-become-data-scientist-data-driven-approach-careers-data/

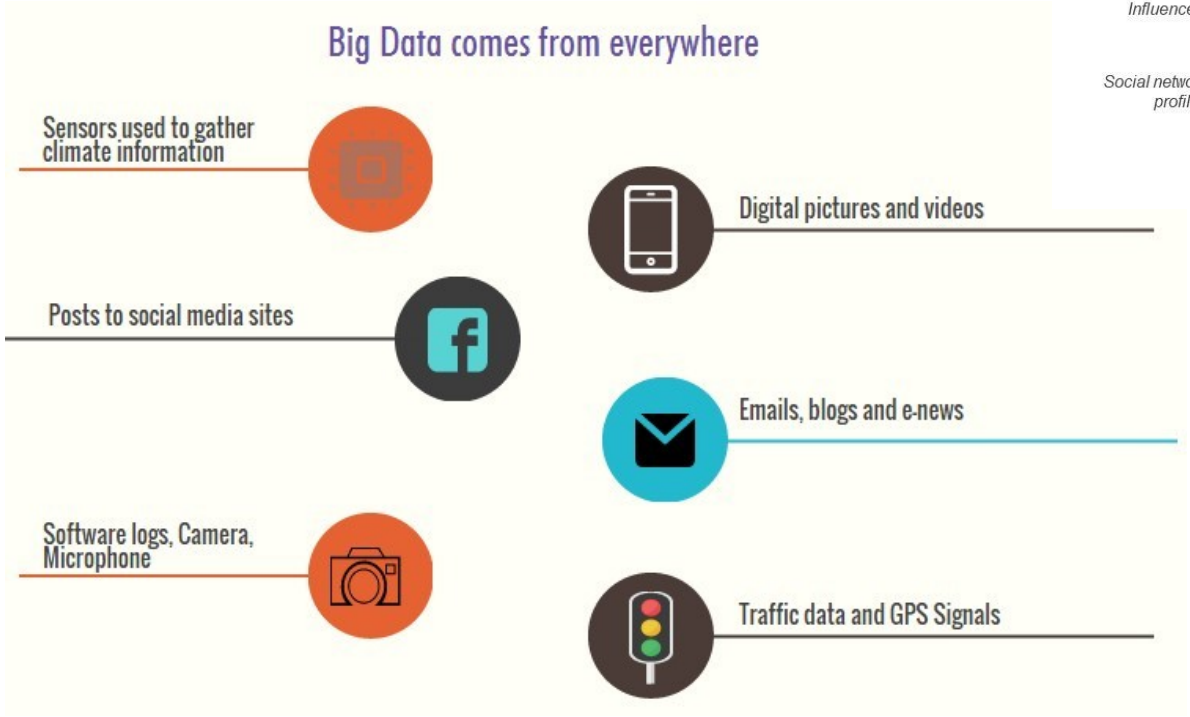
Data Science Venn Diagram

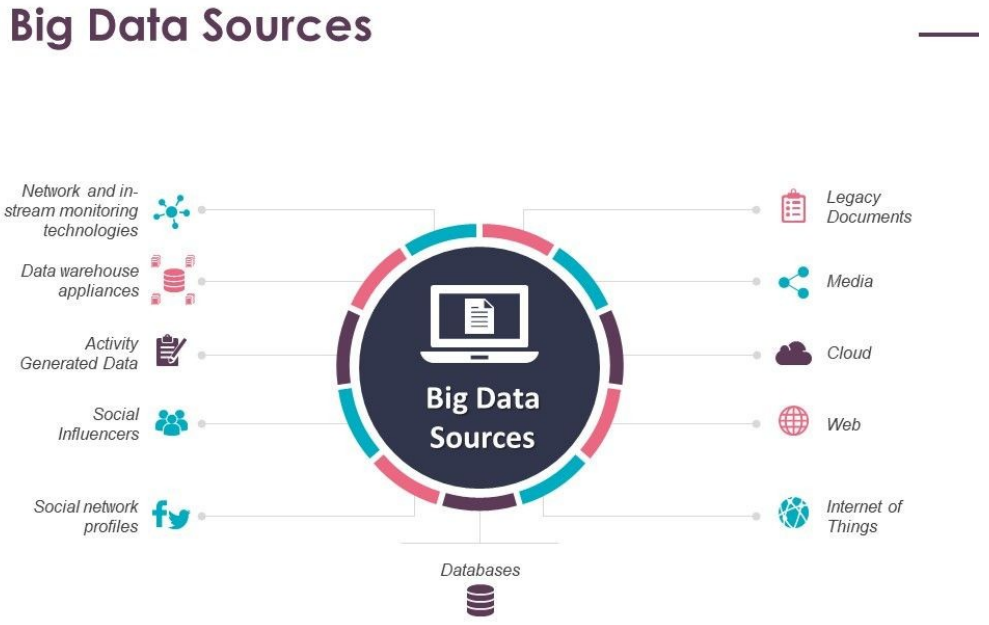


The **big** Picture

where does big data come from?

The bulk of big data generated comes from three primary sources: social data, machine data and transactional data.





The bulk of big data generated comes from three primary sources: **social data**, **machine data** and **transactional data**. In addition, companies need to make the distinction between data which is generated internally, that is to say it resides behind a company's firewall, and externally data generated which needs to be imported into a system.

Whether data is unstructured or structured is also an important factor. Unstructured data does not have a pre-defined data model and therefore requires more resources to make sense of it.

INTELLIGENCE BY VARIETY

WHERE TO FIND & ACCESS BIG DATA

KEY

SOME APIs
NO APIs

INTERNAL
EXTERNAL
BOTH

TERMINOLOGY

SOME APIs
Data that has a standard Web service

NO APIs
Data that has no standard Web service and requires alternative methods of integration

INTERNAL
Data that resides behind an organization's firewall

EXTERNAL
Data that resides outside of an organization's firewall

UNSTRUCTURED
Data that does not have a pre-defined data model or is not organized in a pre-defined manner

STRUCTURED
Data that resides in a fixed field within a record or file

VELOCITY
The rate at which data is generated and changed

VARIETY
The number of different data sources and types

VOLUME
The average quantity of data units per category

ARCHIVES
Archives of scanned documents, statements, insurance forms, medical record and customer correspondence, paper archives, and print stream files that contain original systems of record between organizations and their customers

DOCS
XLS, PDF, CSV, email, Word, PPT, HTML, HTML 5, plain text, XML, JSON, etc.

MEDIA
Images, videos, audio, Flash, live streams, podcasts, etc.

DATA STORAGE
SQL, NoSQL, Hadoop, doc repository, file systems, etc.

BUSINESS APPS
Project management, marketing automation, productivity, CRM, ERP content management systems, HR, storage, talent management, procurement, expense management, Google Docs, intranets, portals, etc.

PUBLIC WEB
Government, weather, competitive, traffic, regulatory, compliance, health care services, economic, census, public finance, stock, OSINT, the World Bank, SEC/Edgar, Wikipedia, IMDb, and other Web services

SOCIAL MEDIA
Twitter, LinkedIn, Facebook, Tumblr, Blog, SlideShare, YouTube, Google+, Instagram, Flickr, Pinterest, Vimeo, Wordpress, IM, RSS, Review, Chatter, Jive, Yammer, etc.

MACHINE LOG DATA
Event logs, server data, application logs, business process logs, audit logs, call detail records (CDRs), mobile location, mobile app usage, clickstream data, etc.

SENSOR DATA
Medical devices, smart electric meters, car sensors, road cameras, satellites, traffic recording devices, processors found within vehicles, video games, cable boxes or household appliances, assembly lines, office buildings, cell towers and jet engines, air conditioning units, refrigerators, trucks, farm machinery, etc.

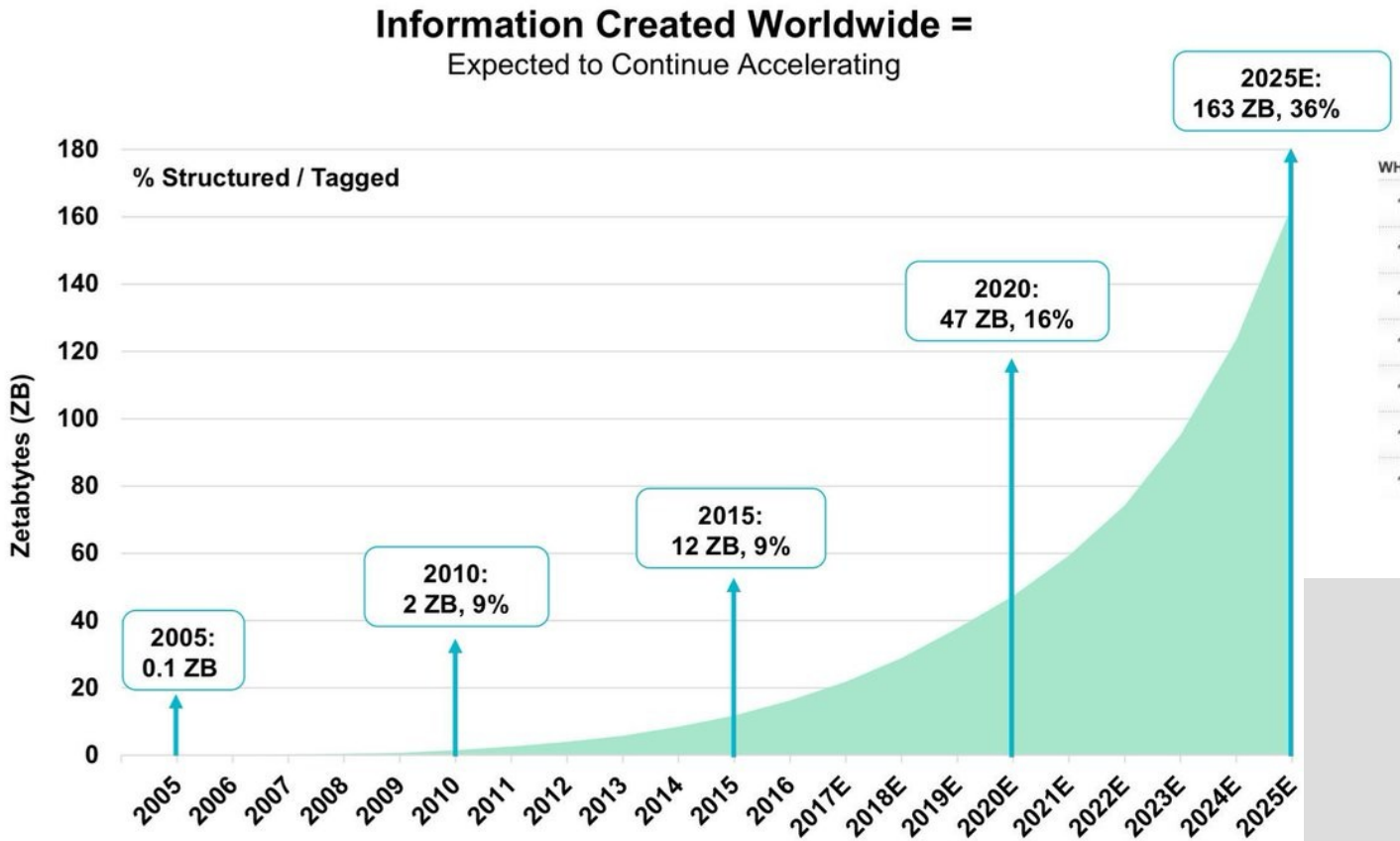
Data compiled by the domain experts at Kapow Software, a Kofax company, and is based on almost a decade of experience helping hundreds of large global enterprises and innovative start-ups across industries leverage critical data from disparate internal and external sources to meet business objectives.

kapow SOFTWARE
A Kofax Company

USING BIG DATA, ORGANIZATIONS CAN GENERATE ACTIONABLE INSIGHTS THAT ENABLE THEM TO DRIVE THEIR BUSINESS FORWARD. RAPID INTEGRATION OF THE EVER-EXPANDING POOL OF DATA SOURCES AND TYPES IS OPENING A WHOLE NEW WORLD OF POSSIBILITIES.

What amazing picture!

Data Volume Growth Continues @ Rapid Clip…
% Structured / Tagged (~10%) Rising Fast…



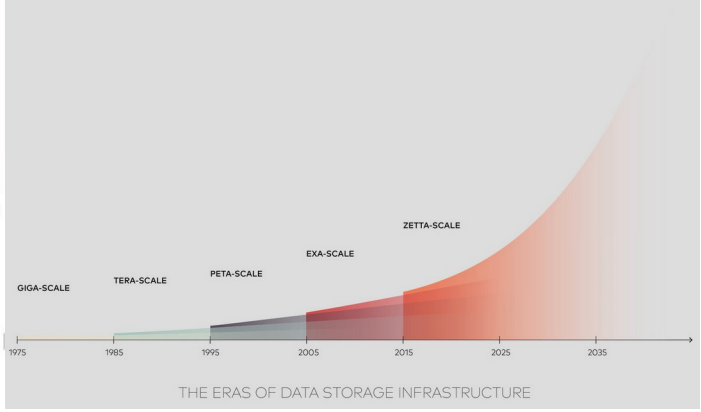
WHAT'S A ZETTABYTE?

1 kilobyte	1,000
1 megabyte	1,000,000
1 gigabyte	1,000,000,000
1 terabyte	1,000,000,000,000
1 petabyte	1,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000

KLEINER PERKINS

Source: IDC DataAge 2025 Study, sponsored by Seagate (3/17)
Note: 1 petabyte = 1MM gigabytes, 1 zeta byte = 1MM petabytes

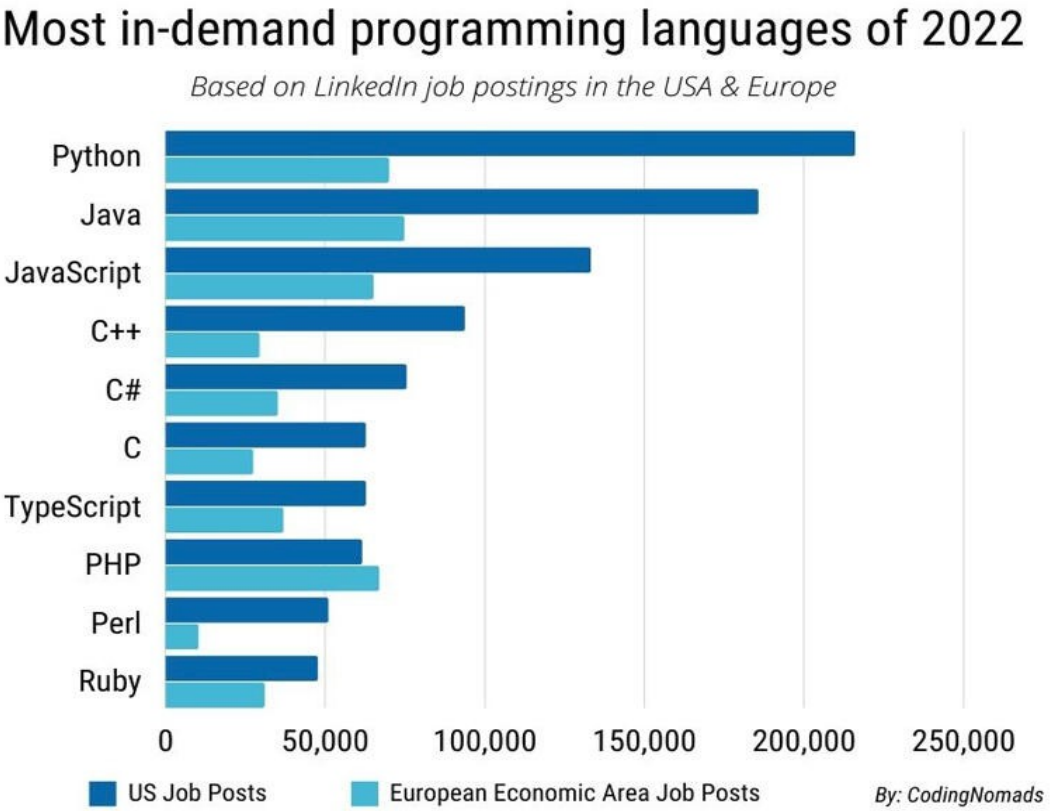
KP INTERNET TRENDS 2017

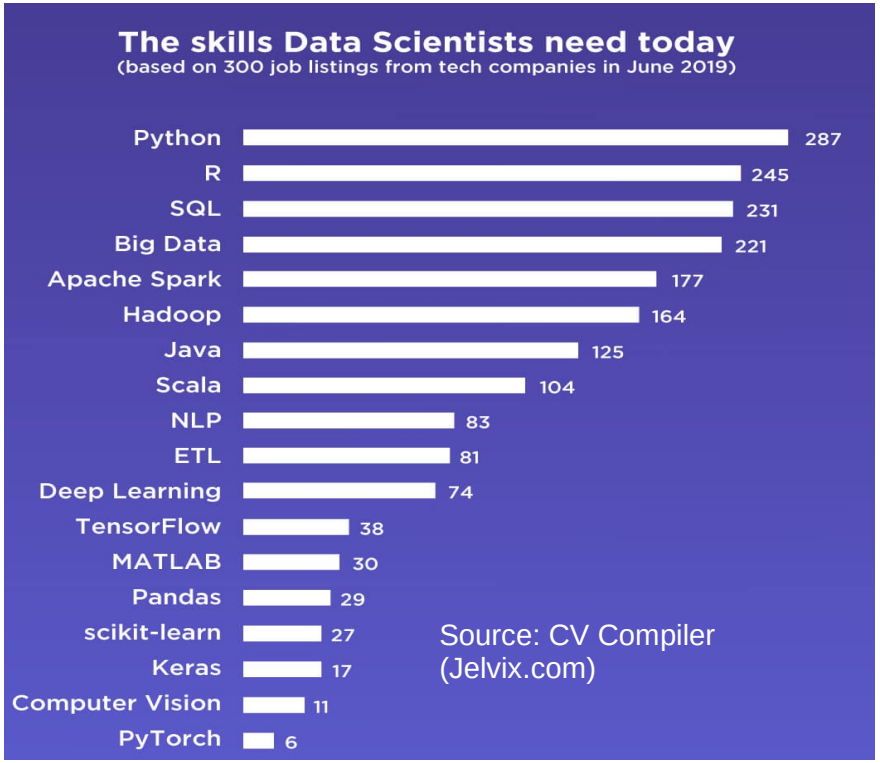


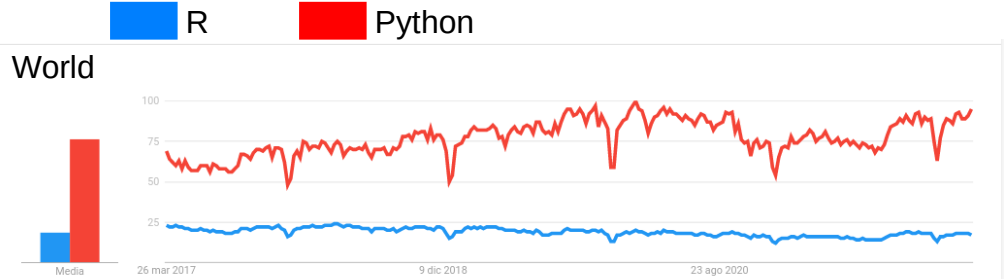
The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.







Source: https://trends.google.com/trends/explore?date=today%205-y&geo=CO&q=%2Fm%2F0212jm,%2Fm%2F05z1_

- What do I have consolidated?
- What do I need to learn and what do I still need to learn more about?

Data Scientist
Roadmap

Mathematics
- Linear Algebra
- Analytics Geometry
- Matrix
- Vector Calculus
- Optimization
- Regression
- Dimensionality Reduction
- Density Estimation
- Classification

Probability
- Discrete Distribution
  - *Binomial*
  - *Bernouli*
  - *Geometric etc*
- Continuos Distribution
  - *Uniform*
  - *Exponential*
  - *Gamma*
- Normal Distribution
- Introduction to Probability
- 1D Random Variable
- Function of One Random Variable
- Joint Probability Distribution

Statistics
- Introduction to Statistics
- Data Description
- Random Samples
- Sampling Distribution
- Parameter Estimation
- Hypotheses Testing
- ANOVA
- Reliability Engineering
- Stochastic Process
- Computer Simulation
- Design of Experiments
- Simple Linear Regression
- Correlation
- Multiple Regression
- Nonparametric Statistics
  - *Sign Test*
  - *The Wilcoxon Signed-Rank Test*
  - *The Wilcoxon Rank Sum Test*
  - *The Kruskal-Walls Test*
- Statistical Quality Control
- Basic of Graphs

Programming

Python
- Python Basics
  - *List*
  - *Set*
  - *Tuples*
  - *Dictionary*
  - *Function, etc.*
- NumPy
- Pandas
- Matplotlib/Seaborn, etc.

R
- R Basic
  - *Vector*
  - *List*
  - *Data Frame*
  - *Matrix*
  - *Array, etc*
- dplyr
- ggplot2
- Tidyr
- Shiny, etc.

DataBase
- SQL
- MongoDB

Other
- Data Structure
  - *Array, etc*
- Web Scraping
- Linux
- Git

Machine Learning

Introduction
- How Model Works
- Basic Data Exploration
- First ML Model
- Model Validation
- Underfitting & Overfitting
- Random Forests
- scikit-learn

Intermediate
- Handling Missing Values
- Handling Categorical Variables
- Pipelines
- Cross-Validation
- XGBoost
- Data Leakage

Deep Learning
- Artificial Neural Network
- Convolutional Neural Network
- Recurrent Neural Network
- Keras
- PyTorch
- TensorFlow
- A Single Neuron
- Deep Neural Network
- Stochastic Gradient Descent
- Overfitting and Underfitting
- Dropout Batch Normalization
- Binary Classification

Feature Engineering
- Baseline Model
- Categorical Encodings
- Feature Generation
- Feature Selection

Natural language Processing
- Text Classification
- Word Vectors

Data Visualization Tools
- Excel VBA
- Bi (Business Intelligence)
  - *Tableau*
  - *Power BI*
  - *Qlik View*
  - *Qlik Sense*

Deployment
- Microsoft Azure
- Heroku
- Google Cloud Platform
- Flask
- Django

Other Points
- Domain Knowledge
- Communication Skill
- Reinforcement Learning
- Case Studies
  - *Data Science at Netflix*
  - *Data Science at Flipkart*
  - *Project on Credit Card Fraud Detection*
  - *Project on Movie Recommendation , etc.*

Keep Practicing

PYPL PopularitY of Programming Language

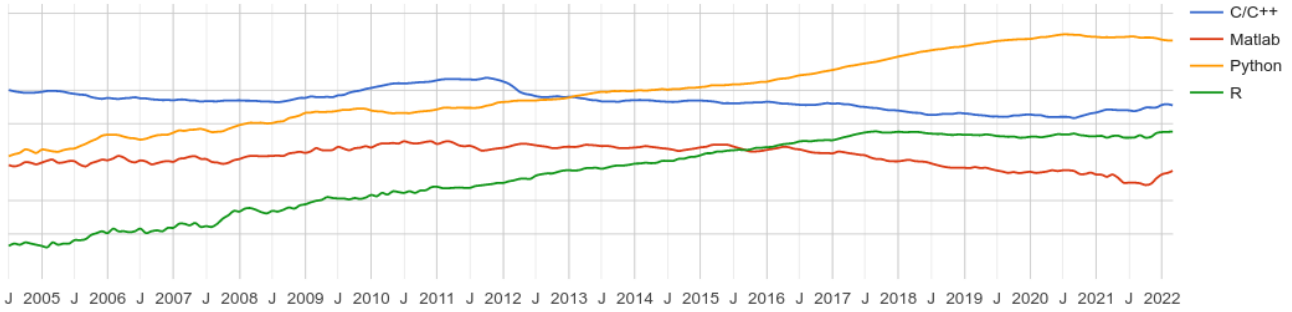
The PYPL PopularitY of Programming Language Index is created by analyzing how often language tutorials are searched on Google.

The more a language tutorial is searched, the more popular the language is assumed to be. It is a leading indicator. The raw data comes from Google Trends.

If you believe in collective wisdom, the PYPL Popularity of Programming Language index can help you decide which language to study, or which one to use in a new software project.

Worldwide, Mar 2022 compared to a year ago:

Rank	Change	Language	Share	Trend
1		Python	28.27 %	-2.0 %
2		Java	18.03 %	+0.8 %
3		JavaScript	8.86 %	+0.4 %
4		C#	7.51 %	+0.6 %
5		C/C++	7.32 %	+0.6 %
6		PHP	5.71 %	-0.4 %
7		R	4.23 %	+0.5 %
8		Objective-C	2.28 %	-1.2 %
9	↑	TypeScript	2.11 %	+0.3 %
10	↓	Swift	2.01 %	+0.2 %
11		Matlab	1.87 %	+0.2 %
12		Kotlin	1.57 %	-0.1 %



https://pypl.github.io/PYPL.html

Getting started!

Next...

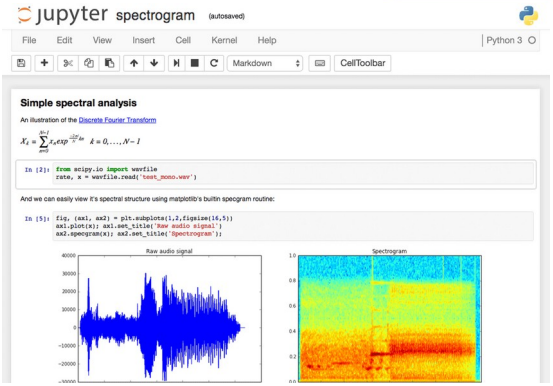
Creating environments and packages…..

- First: Install Acaconda (Windows)
  https://docs.anaconda.com/anaconda/install/windows/

- Next: Install dependences (Python)
  https://medium.com/@GalarnykMichael/install-python-anaconda-on-windows-2020-f8e188f9a63d

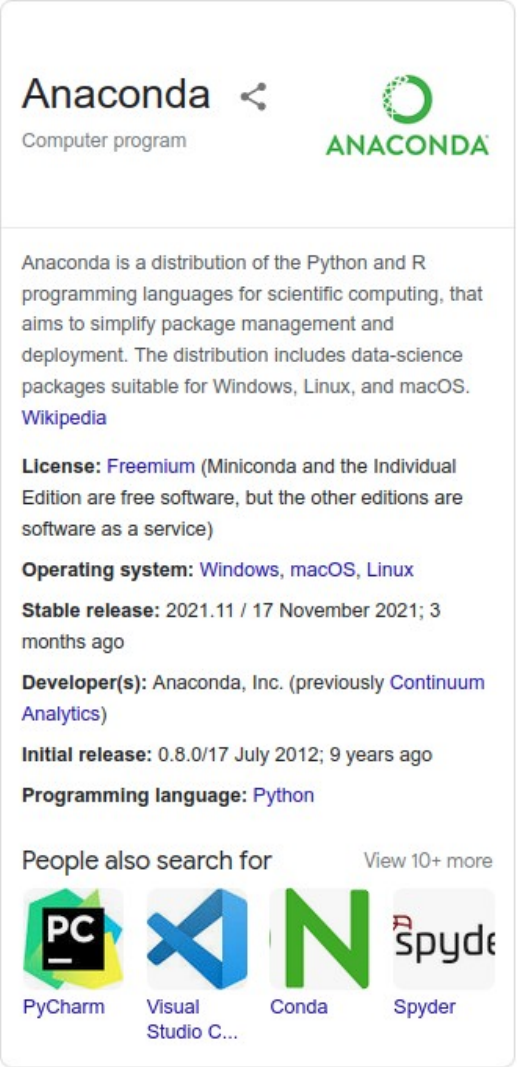
- Next: Install Jupyter Notebook
  https://www.geeksforgeeks.org/how-to-install-jupyter-notebook-in-windows/

- Next: Install another kernel
  https://datatofish.com/r-jupyter-notebook/



eoda

jupyter



We need **GitHub** too...

What is Git and what are its origins?

Have you ever had a document named something like report_final_draft_final(3).doc? If so, you've felt the pain of managing and sharing files. Keeping track of the changes to a file over time is difficult but important. Git is a Version Control System (VCS) – a tool that helps us to keep track of differences in a file or collection of files over time.

What is Git used for?

Git is used for managing the changes to a project over time. A project might be just a single file, a handful of files, or thousands of files. Those files can be anything from plain text to images or videos.

What is Github and what are its origins? How did its creation change the way people collaborate?

GitHub, developed in 2008, is a web application that hosts Git repositories. The team that started GitHub saw that Git could solve important problems for many teams – but Git itself is often difficult to use. GitHub adds a bunch of collaboration and exploration tools on top of Git to help you (and your team) be more productive.