Becoming a Data Scientist

GGE 6505/GGE5405 Introduction to Big Data & Data Science
University of New Brunswick
Winter 2022

What is Data Science?

- More than 60 years since the term "data science" was coined.
  - 1962 → Data Analysis
  - 1974 → Data science (Alternative to CS)
  - 1985 → Data Science (Alternative to Statistic)
  - 1996 → Data Science Conference
  - 2002 → Data Science Journal

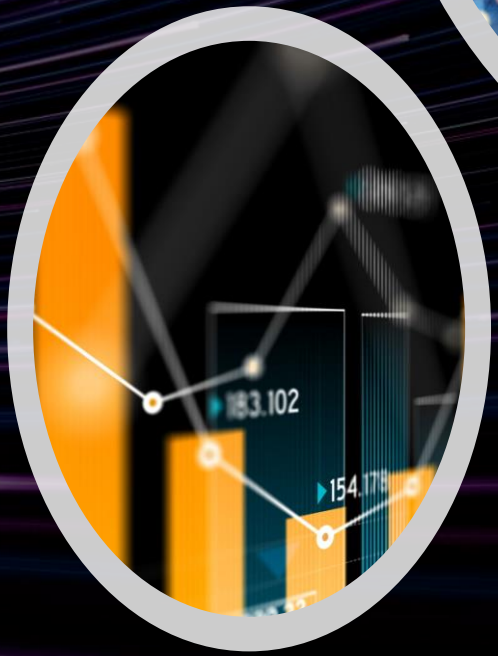
- The key word in Data Science is not data, it is SCIENCE.

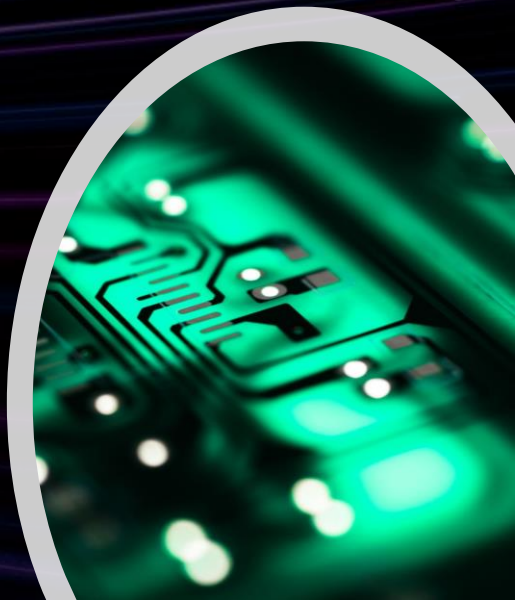
Science is the intellectual and practical activity encompassing the systematic study of the structure and behavior of the physical and natural world through observation and experiment. (Cambridge Dictionary)

What is Data Science?

Data science combines the scientific method, math and statistics, specialized programming, advanced analytics, AI, and even storytelling to uncover and explain the business insights buried in data.

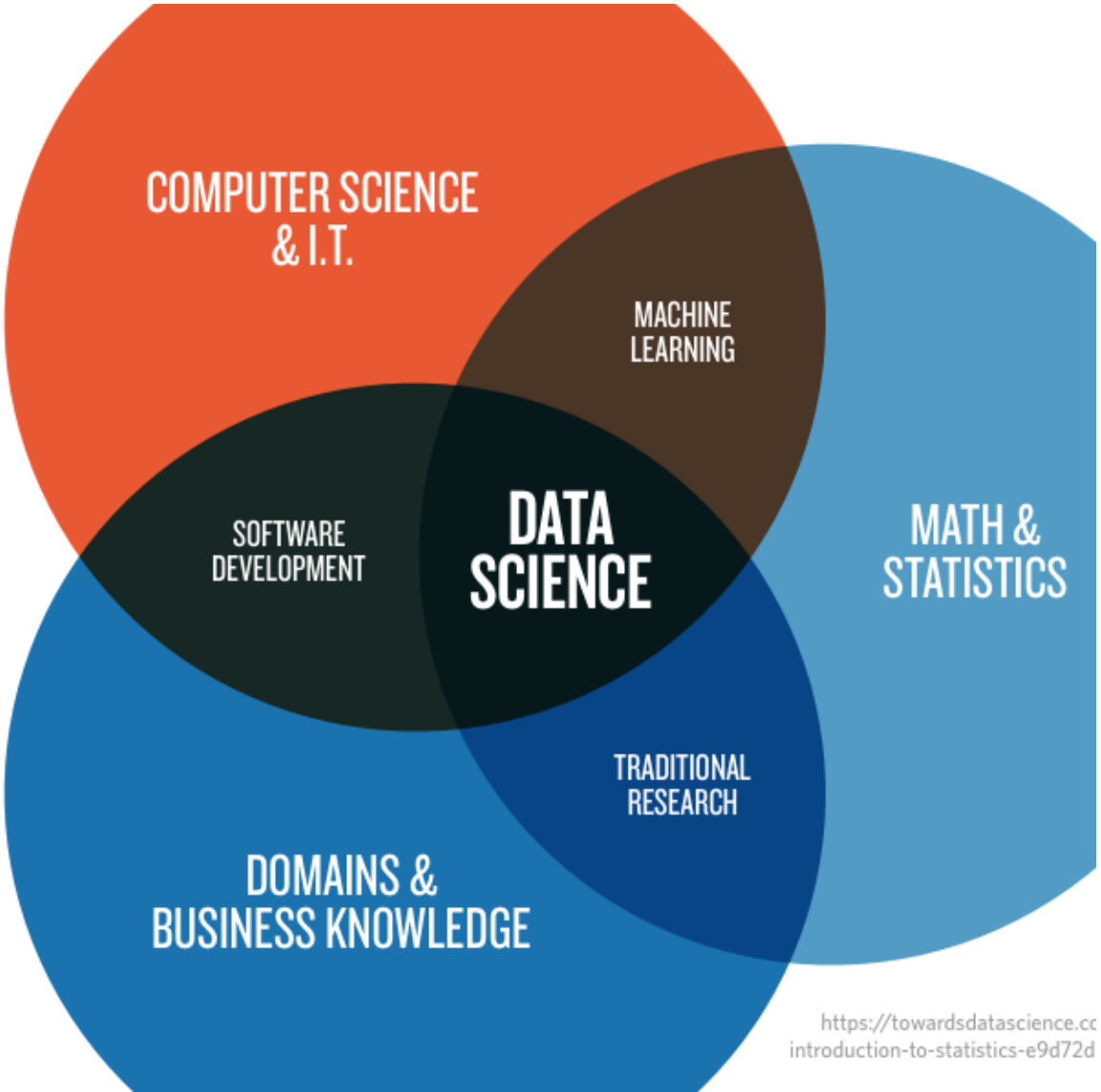
IBM





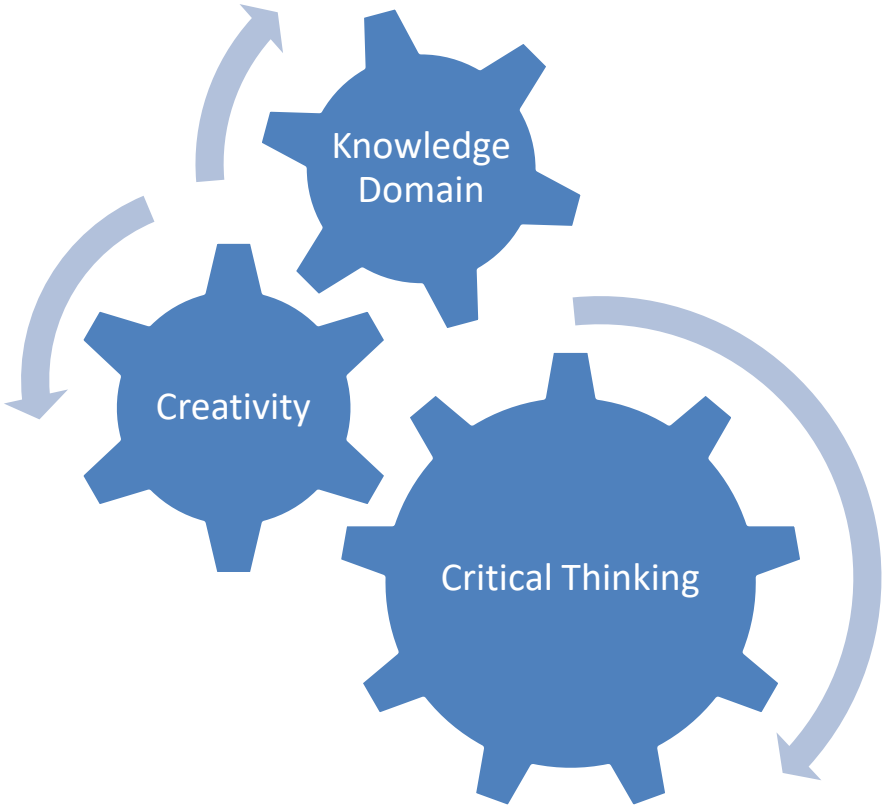
Why does a data scientist differ from a statistician, computer scientist and domain experts?

What is Data Science



https://towardsdatascience.cc
introduction-to-statistics-e9d72d

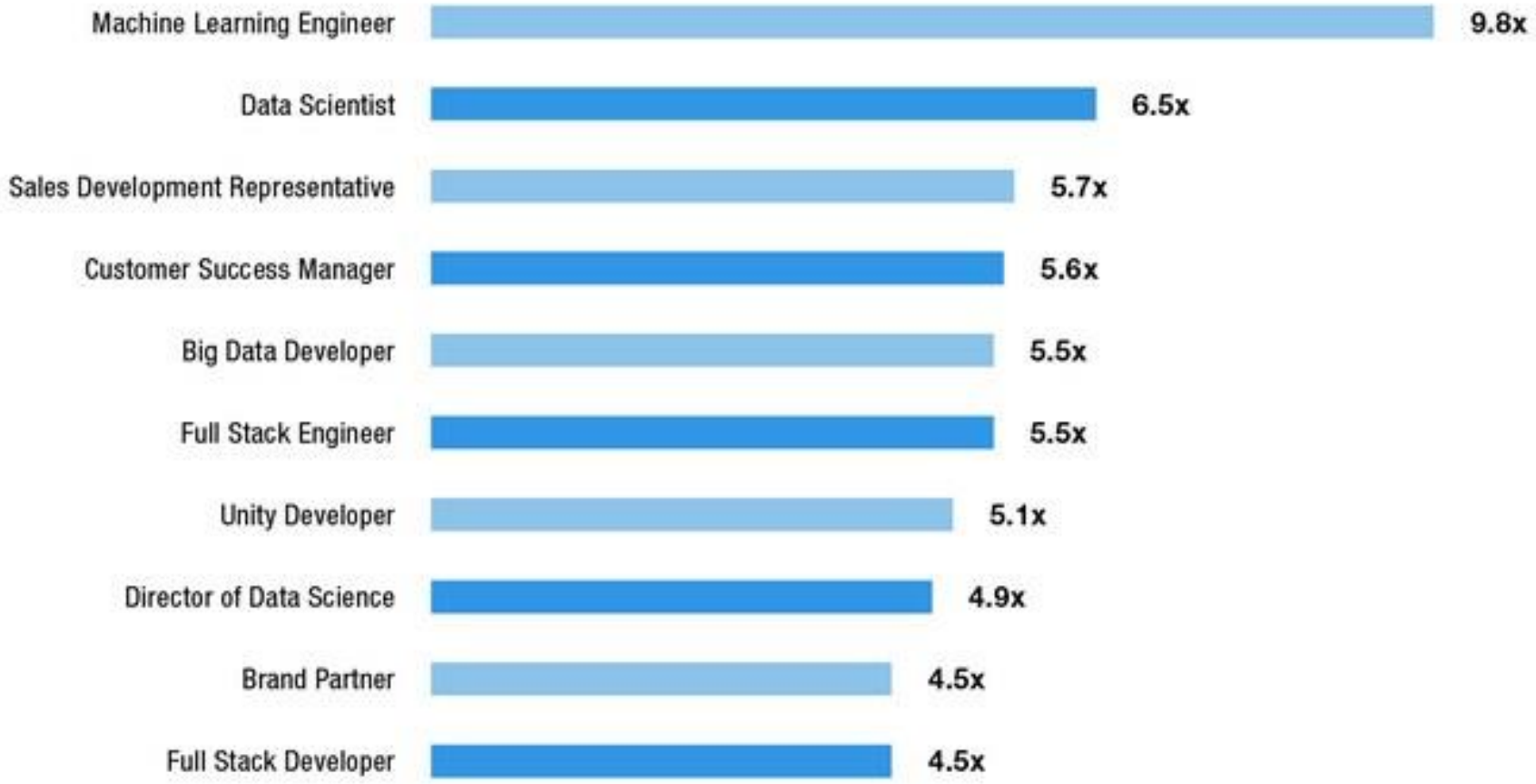
But for actually unlocking the data's secrets…



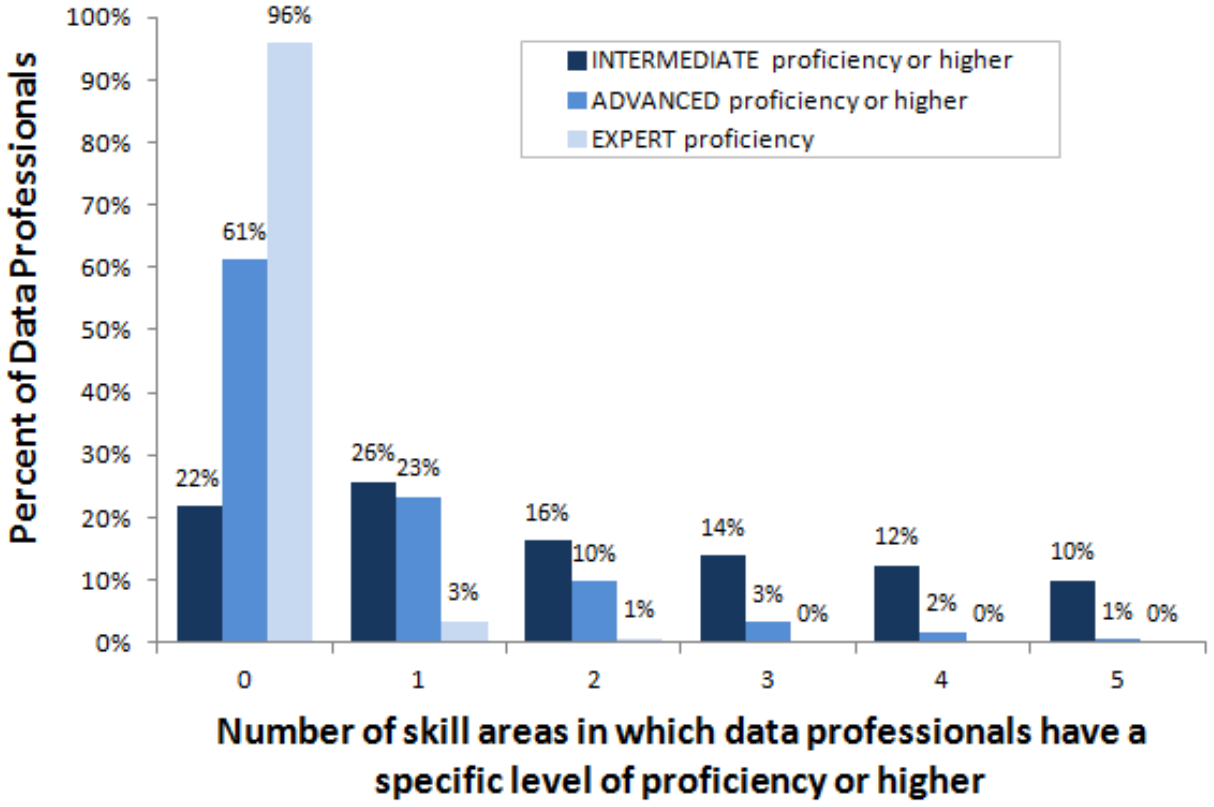
Why does data scientist remains a top job?

	Job Title	Median Base Salary	Job Satisfaction	Job Openings
#1	Front End Engineer	$105,240	3.9/5	13,122
#2	Java Developer	$83,589	3.9/5	16,136
#3	Data Scientist	$107,801	4.0/5	6,542
#4	Product Manager	$117,713	3.8/5	12,173
#5	DevOps Engineer	$107,310	3.9/5	6,603

Top Emerging Jobs



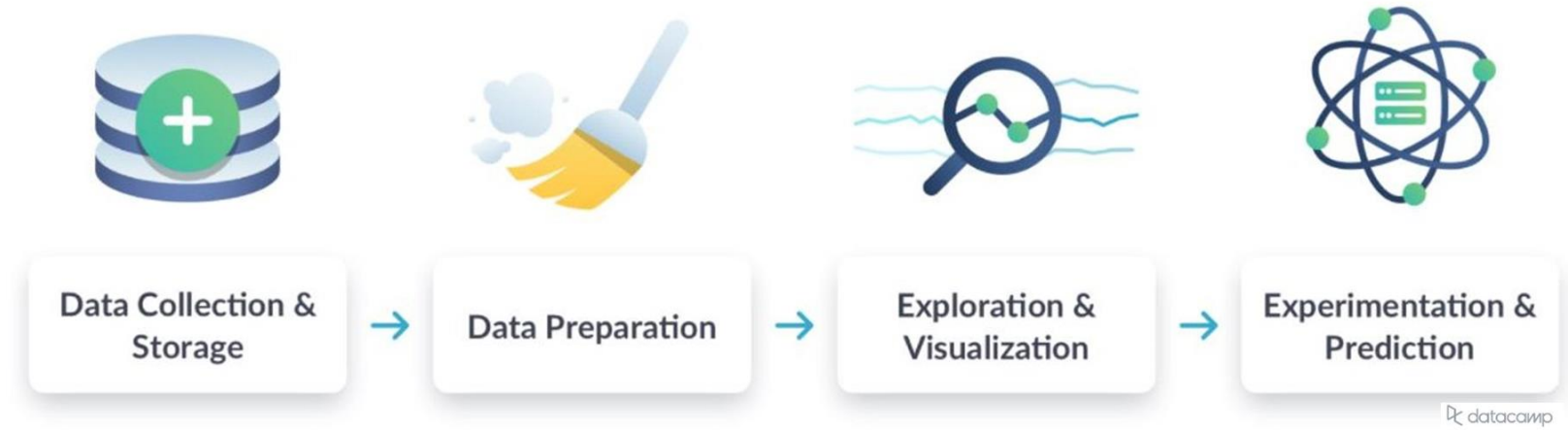
There are Only a Handful of Data Professionals who are Proficient in All Skill Areas



The five data skill areas were: Business, Technology, Math & Modeling, Programming and Statistics.


Salary Comparison By Experience
Data Scientist
Canada
124,000 CAD
+26%
157,000 CAD
+32%
206,000 CAD
+18%
243,000 CAD
+11%
269,000 CAD
+6%
286,000 CAD
< 2 Years
2 to 5
5 to 10
10 to 15
15 to 20
20+ Years
salaryexplorer.com

Data Science Workflow

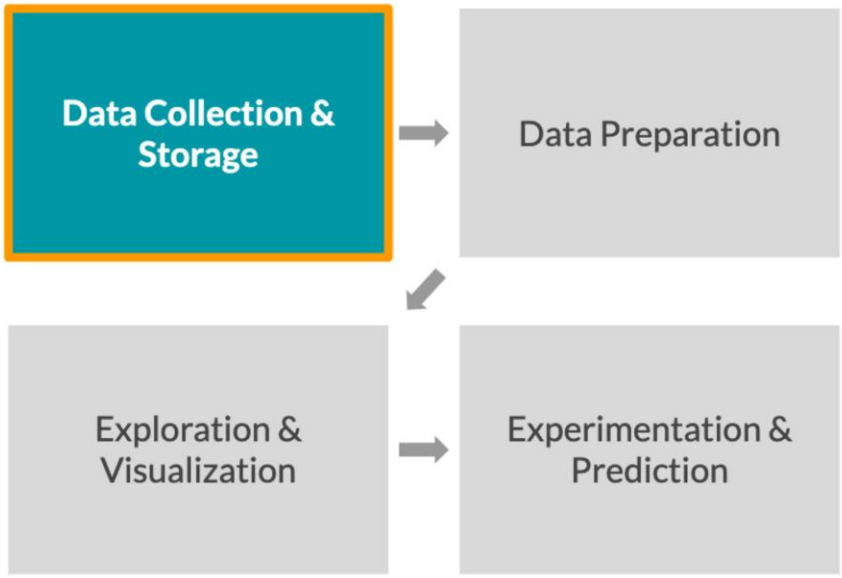


Roles in Data Science and related Fields

- Data Engineer
- Data Analyst
- Data Scientist
- Machine Learning Engineer

Data engineer

- Information architects
- Build data pipelines and storage solutions
- Maintain data access

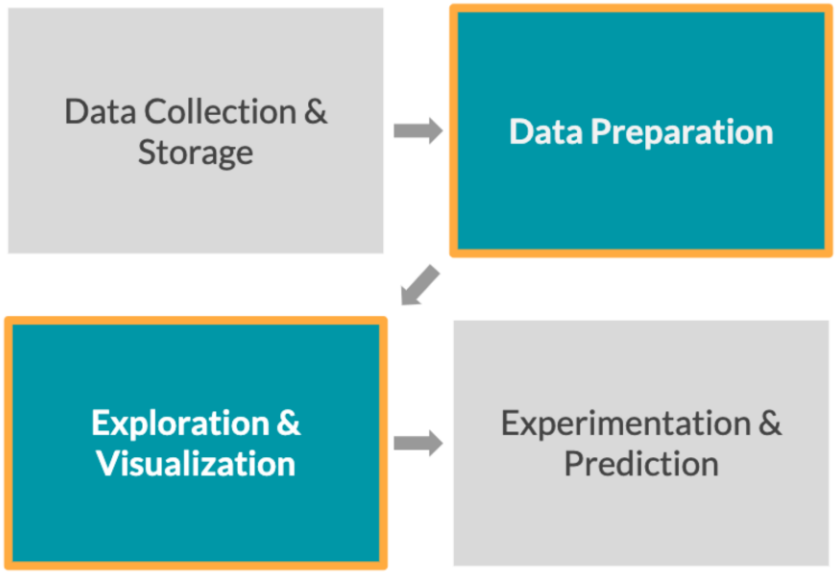


Data engineering tools

- **SQL**
  - To store and organize data
- **Java**, **Scala**, or **Python**
  - Programming languages to process data
- **Shell**
  - Command line to automate and run tasks
- **Cloud computing**
  - AWS, Azure, Google Cloud Platform

Data analyst

- Perform simpler analyses that describe data
- Create reports and dashboards to summarize data
- Clean data for analysis

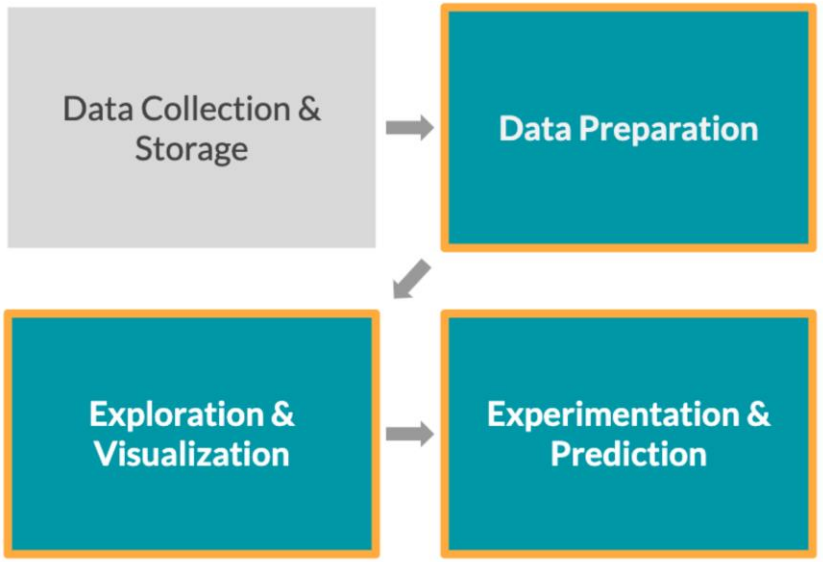


Data analyst tools

- **SQL**
  - Retrieve and aggregate data
- **Spreadsheets (Excel or Google Sheets)**
  - Simple analysis
- **BI tools (Tableau, Power BI, Looker)**
  - Dashboards and visualizations
- *May have:* Python or R
  - Clean and analyze data

Data scientist

- Versed in statistical methods
- Run experiments and analyses for insights
- Traditional machine learning

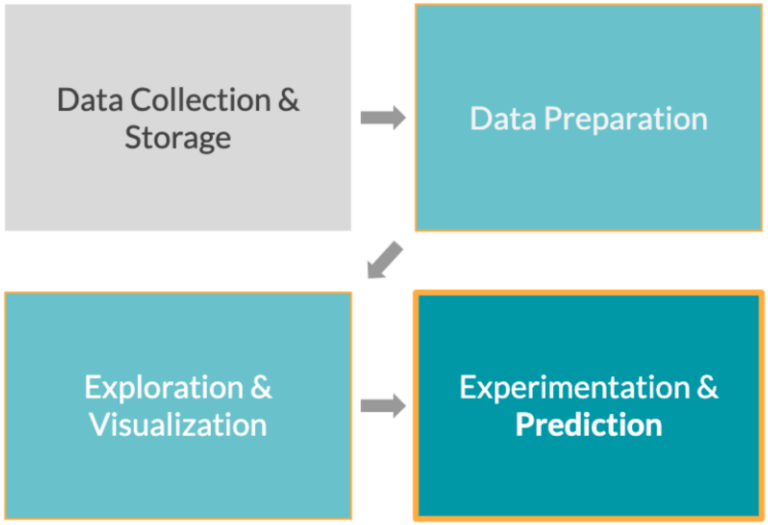


Data scientist tools

- **SQL**
  - Retrieve and aggregate data
- **Python and/or R**
  - Data science libraries, e.g., `pandas` (Python) and `tidyverse` (R)

Machine learning scientist

- Predictions and extrapolations
- Classification
- Deep learning
  - Image processing
  - Natural language processing



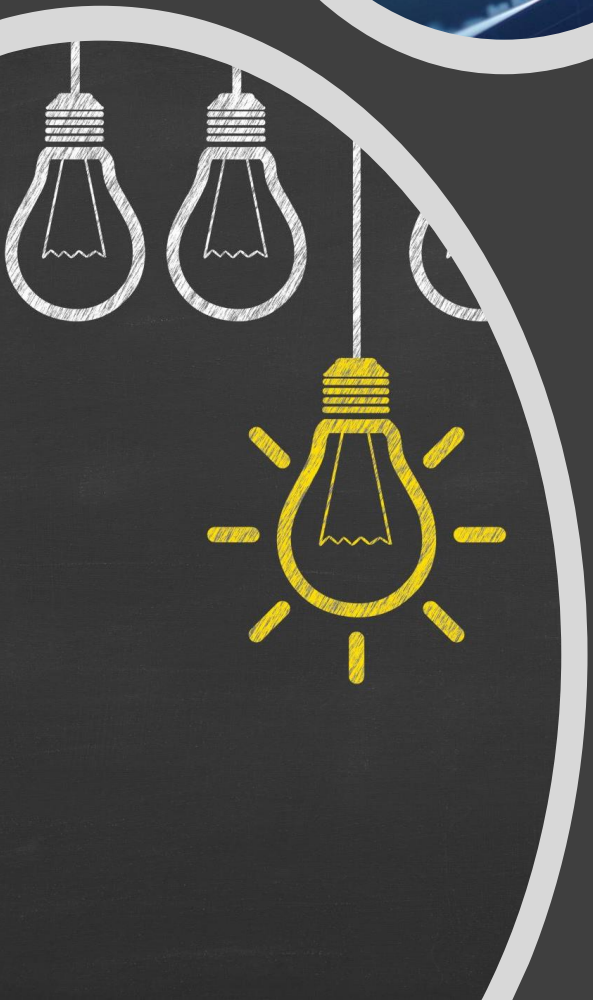
Machine learning tools

- **Python and/or R**
  - Machine learning libraries, e.g., TensorFlow or Spark

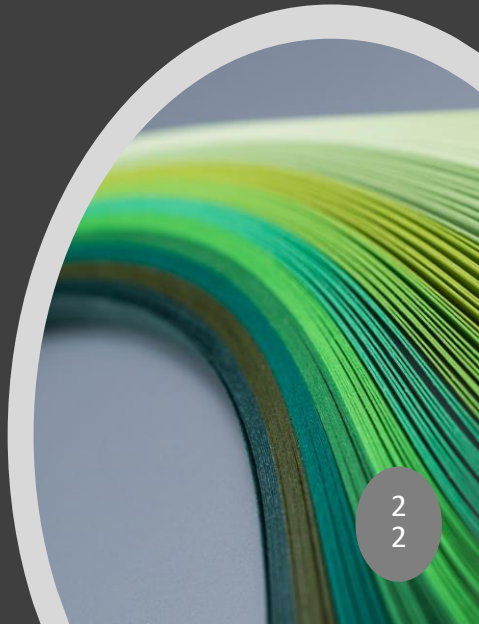
Summary of Data Related Jobs

Data Engineer	Data Analyst	Data Scientist	Machine Learning Scientist
Store and maintain data	Visualize and describe data	Gain insights from data	Predict with data
SQL + Java/Scala/Python	SQL + BI Tools + Spreadsheets	Python/R	Python/R

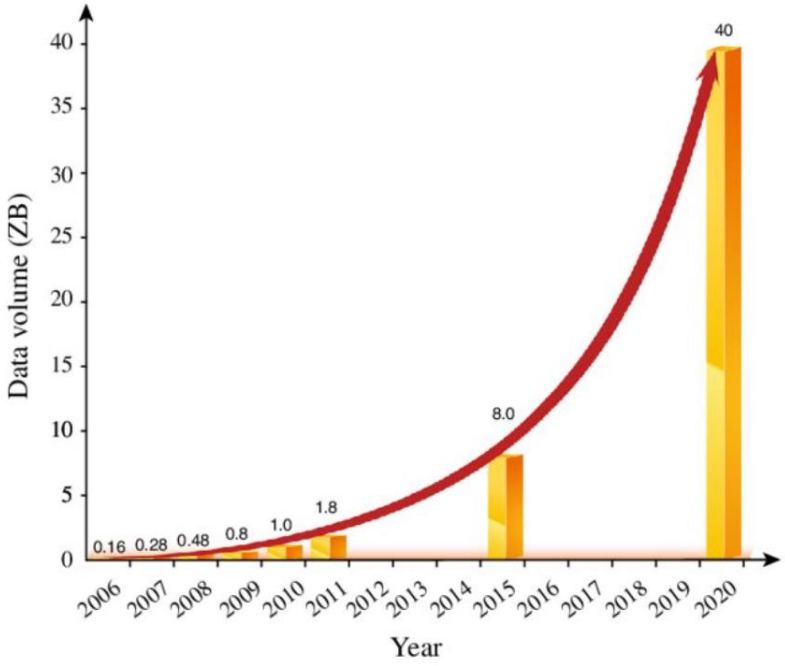
datacamp



BIG DATA



Global growth trend of data volume 2006-2020



This figure was uploaded by Dong Liang

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

if we're talking about how much data is created every day the current estimate stands at **1.145 trillion MB per day**.

What is big data?

The term was added to the Oxford English Dictionary in 2013

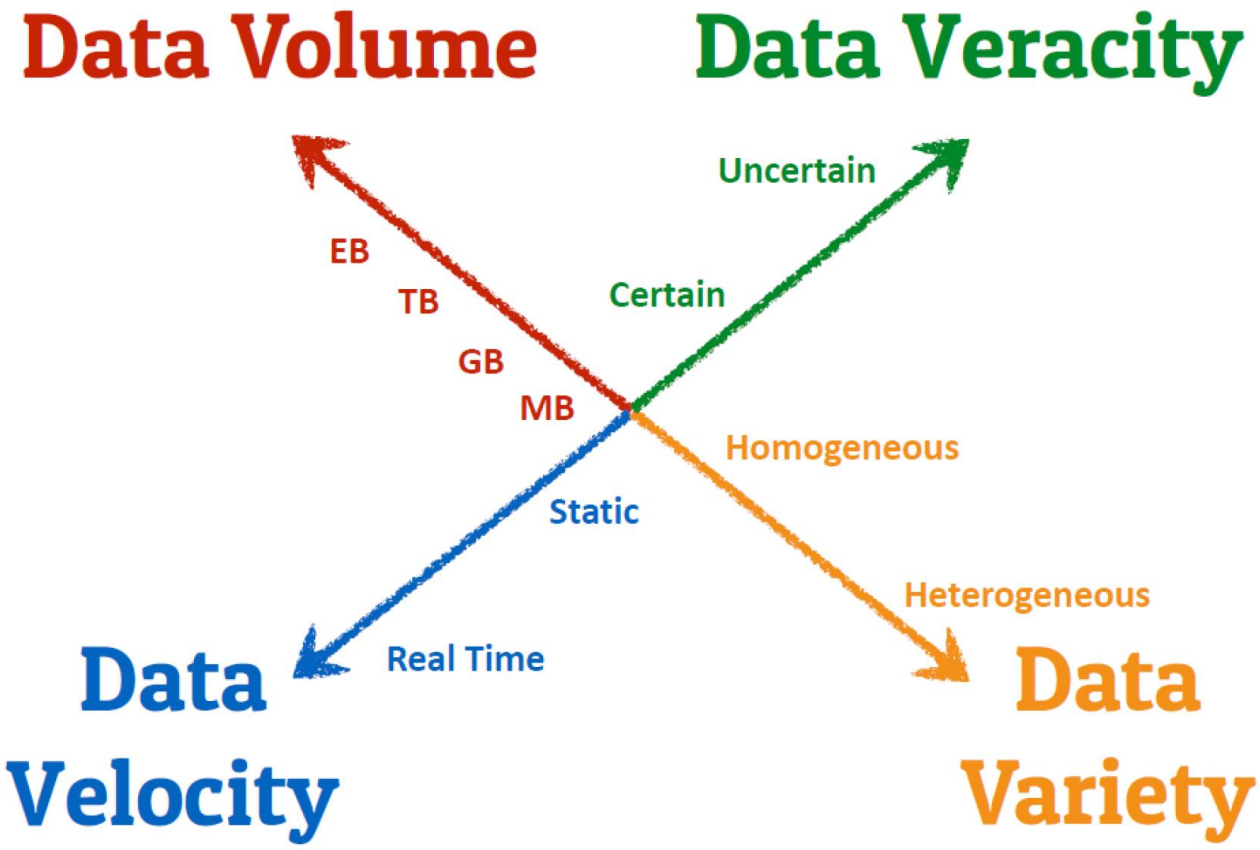
- "big data n. Computing (also with capital initials) data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data."

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

and appeared in Merriam-Webster's Collegiate Dictionary in 2014

- " big data: an accumulation of data that is too large and complex for processing by traditional tools."

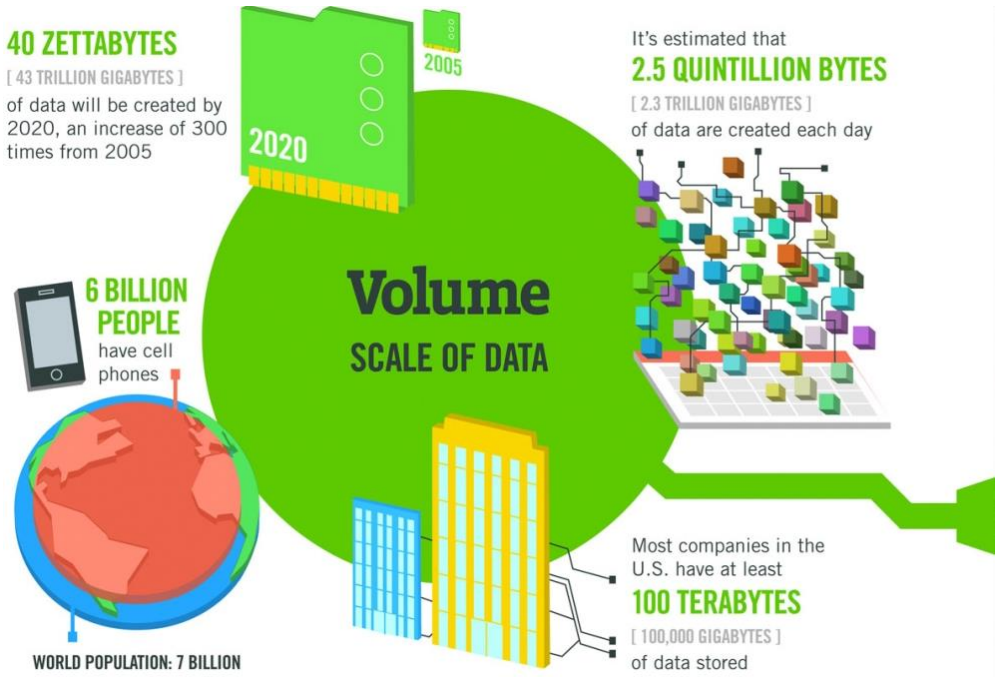
Characteristics of Big Data



Src: Marc Streit

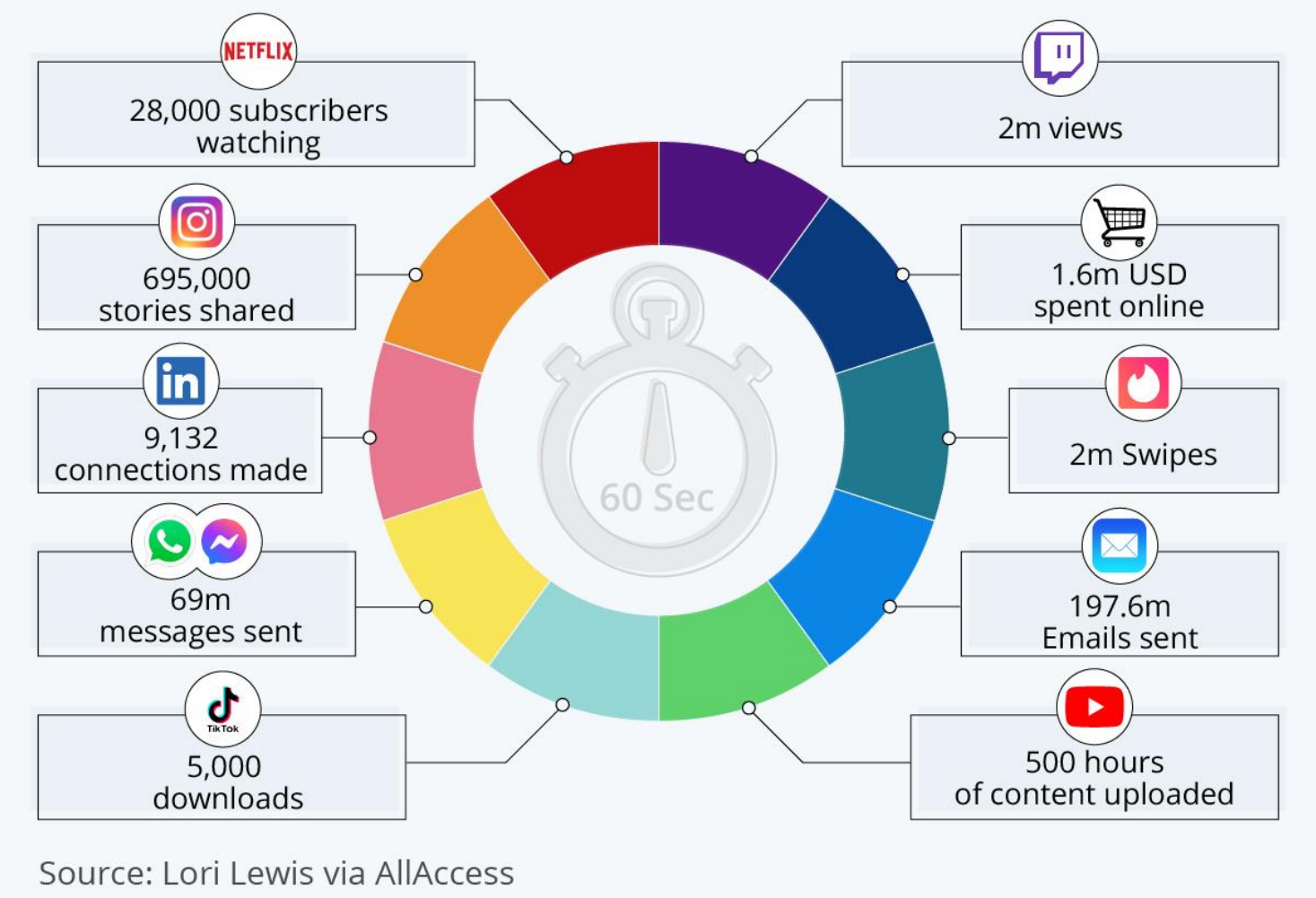
Data Volume

- Amount of Data Generated
- Saved in Records, Files, Tables
- In Kilobytes or Terabytes
- Data comes from many sources
  - Sensors
  - photos
  - Business transactions
  - Location-based data
  - Social media



A Minute on the Internet in 2021

- Estimated amount of data created on the internet in one minute!



Data Variety

- Multiple data Formats: Structure/ unstructured and semi-structure data
  - Online images and videos
  - Human generated text
  - Machine generated reading

- To extract knowledge all these types of data need to linked together (complexity)
- Integrate complex and multiple data types

Unstructured data

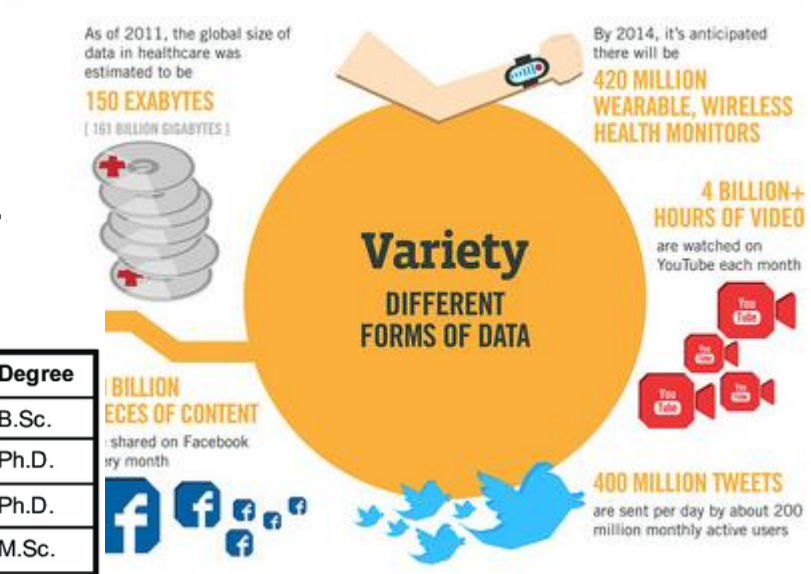
The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.



Structured data

Characteristics

Predefined data models
Easy to search
Text-based
Shows what is happening

Resides in

Relational databases
Data warehouses

Stored in rows and columns

Examples

Dates
Phone numbers
Social security numbers
Customer names
Transactional information

Unstructured data

Characteristics

No predefined data models
Difficult to search
Text, PDF, Images, Video
Shows the **why**

Resides in

Applications
Data warehouses and lakes

Stored in various forms

Examples

Documents
Emails and messages
Conversation transcripts
Image files
Open-ended survey answers

Semi-structured data

Characteristics

Loosely organized
Meta-level structure that can contain unstructured data
HTML, XML, JSON

Resides in

Relational databases
Tagged-text format

Stored in abstracts & figures

Examples

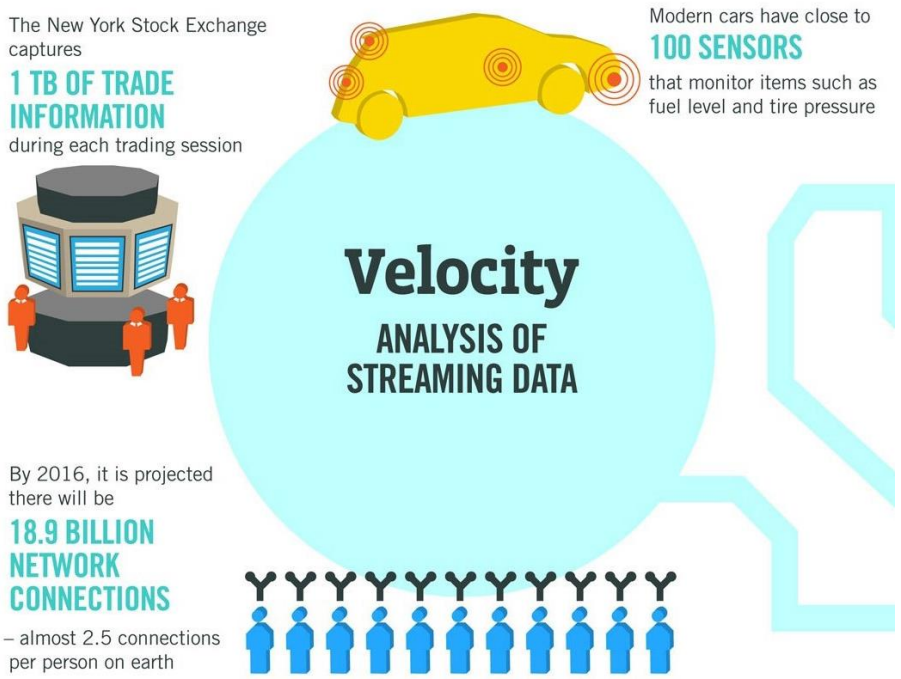
Server logs
Tweets organized by hashtags
Email sorting by folders (inbox; sent; draft)

Data Velocity

- In big data, Velocity demonstrate two things mainly,
  - Speed of growth of data
  - Speed of transmission of data

- Velocity refers to data generating, increasing and sharing at a particular speed through the resources

- Sources of fast data: business processes, machines, networks and human interaction with social media sites, mobile devices, etc.

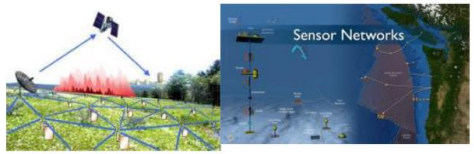


Data Velocity

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

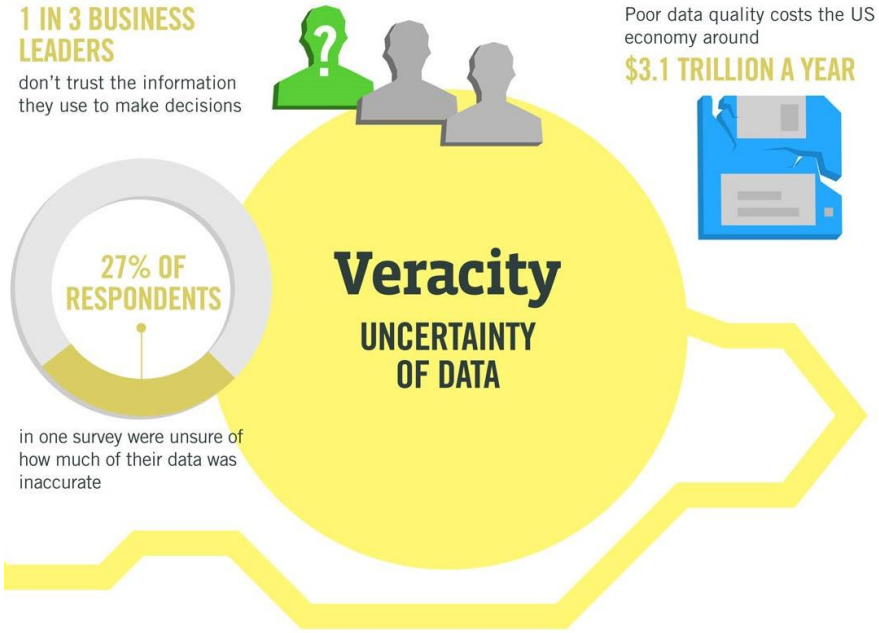
- Real-time and fast data
- Ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion
- Data Stream
  - Unbounded
  - Real-time
  - Scan the data once

Data Veracity

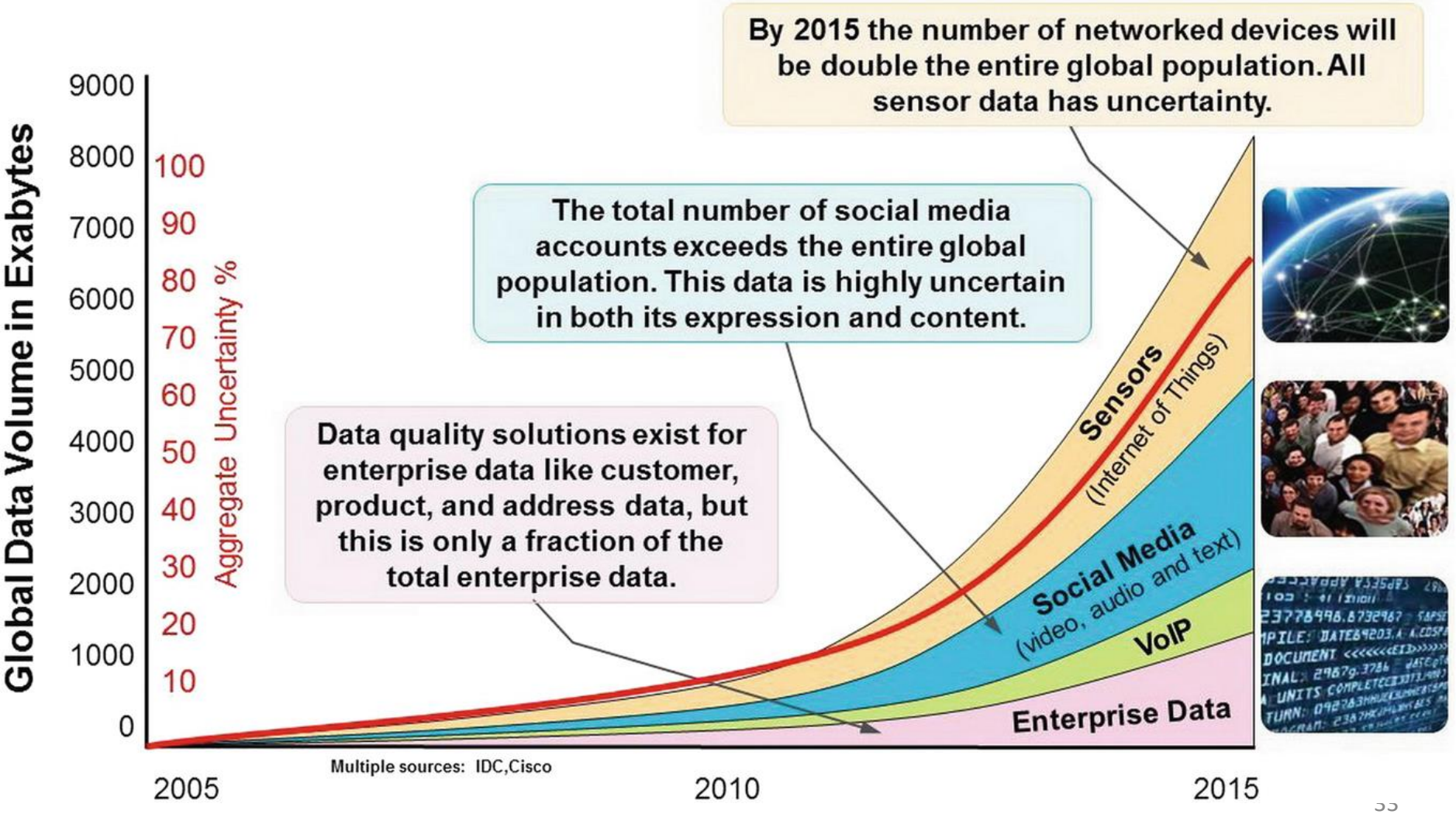
- It refers to the assurance of **quality/integrity/credibility/accuracy** of the data. Since the data is collected from multiple sources, we need to check the data for accuracy before using it for business insights

Uncertainty due to

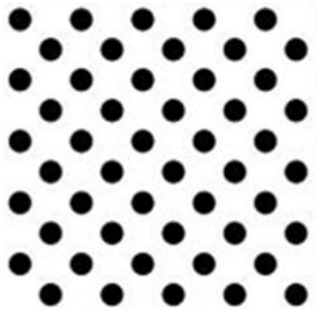
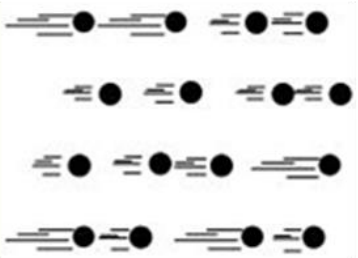
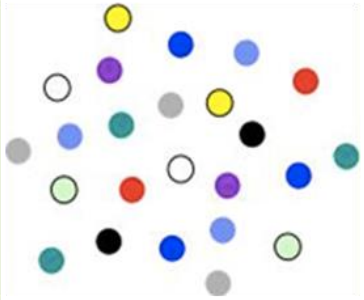
- Inconsistency, incompleteness, ambiguities and model approximations
- Since 2015, over 80% of all data is uncertain



Data Veracity



Summary: The 4 V's of Big Data

Volume	Velocity	Variety	Veracity
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,…	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Adapted by a post of M

"3D Data Management: Controlling Data Volume, Velocity and Variety", Douglas Laney, Gartner, 2001

"If You Think Big Data's Big Now, Just Wait", Ron Miller, TechCrunch, 2014

What can big data be used for?

1. Discover hidden patterns, trends and outliers

2. Improve decisions by enriching information for decision makers

3. Improve automated process (e.g. supply chain, business transactions)

Big Data Applications

Retail/Consumer

- Merchandizing and market basket analysis
- Campaign management and customer loyalty programs
- Supply-chain management and analytics
- Event- and behavior-based targeting
- Market and consumer segmentations

Finances & Frauds Services

- Compliance and regulatory reporting
- Risk analysis and management
- Fraud detection and security analytics
- Credit risk, scoring and analysis
- High speed arbitrage trading
- Trade surveillance
- Abnormal trading pattern analysis

Web and Digital media

- Large-scale clickstream analytics
- Ad targeting, analysis, forecasting and optimization
- Abuse and click-fraud prevention
- Social graph analysis and profile segmentation
- Campaign management and loyalty programs

Health & Life Sciences

- Clinical trials data analysis
- Disease pattern analysis
- Campaign and sales program optimization
- Patient care quality and program analysis
- Medical device and pharmacy supply-chain management
- Drug discovery and development analysis

Telecommunications

- Revenue assurance and price optimization
- Customer churn prevention
- Campaign management and customer loyalty
- Call detail record (CDR) analysis
- Network performance and optimization
- Mobile user location analysis

Ecommerce & customer service

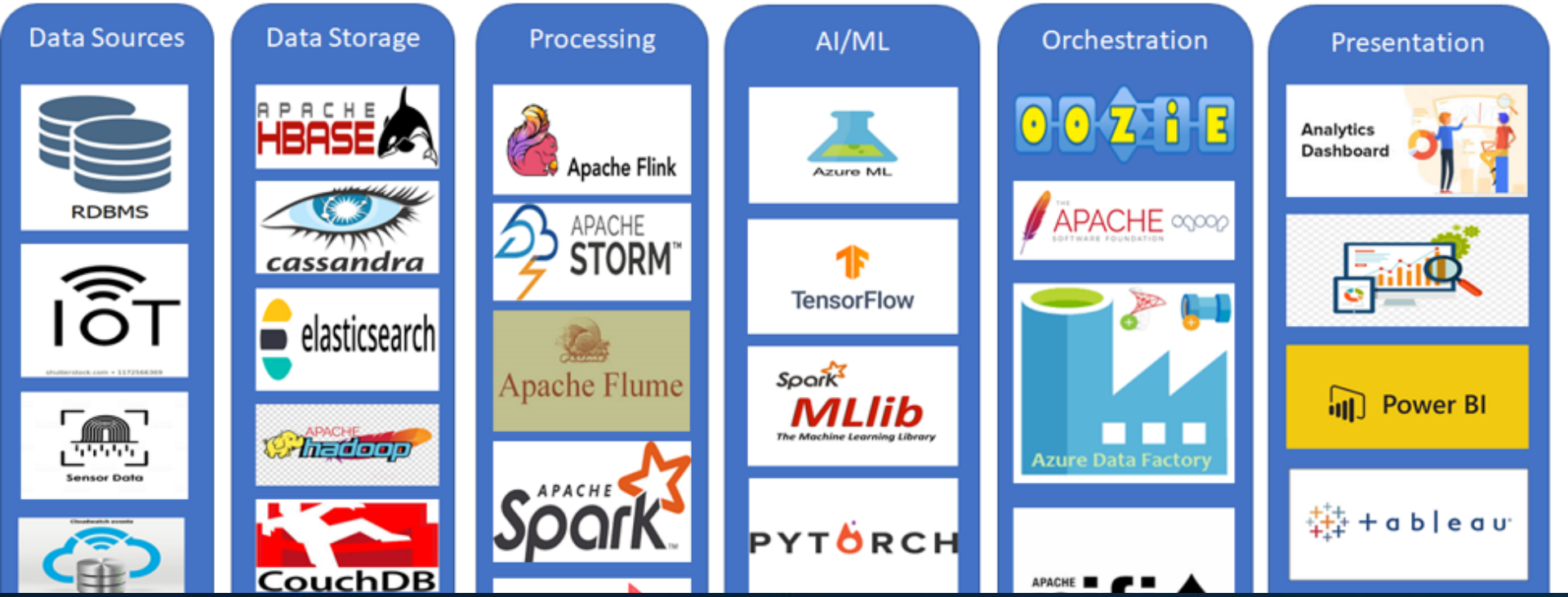
- Cross-channel analytics
- Event analytics
- Recommendation engines using predictive analytics
- Right offer at the right time
- Next best offer or next best action

Challenges with Big data

- Shortage of Skills
- Storage huge and exponentially growing dataset
- Processing data having complex structure
  - structure
  - unstructured
  - semi-structure

- Bringing huge amount of data to computation unit becomes a bottleneck

Component of Big Data Architecture



Big Data: Biomedicine


Spatial Data
Science

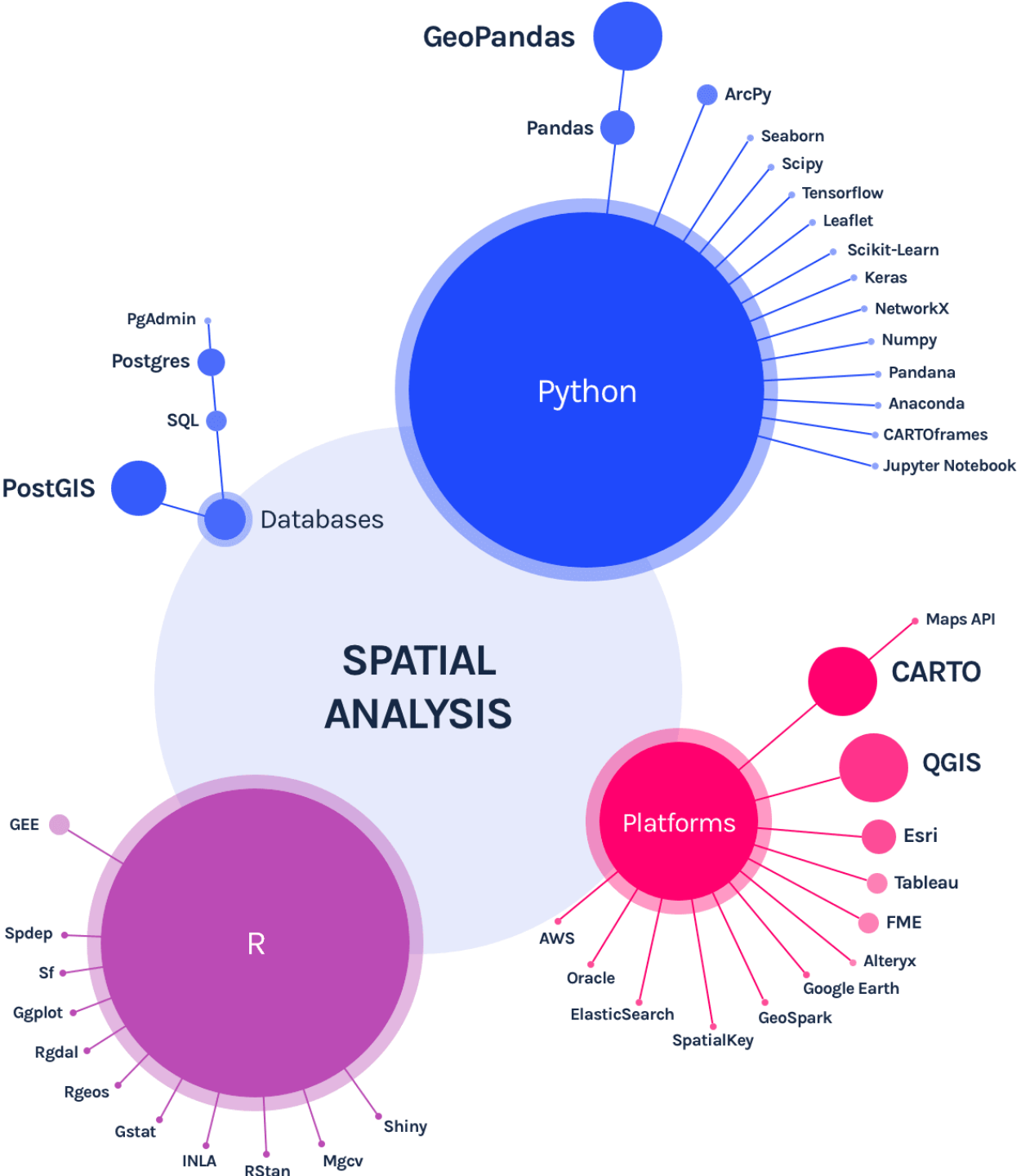
Spatial Data Scientist

Spatial data science can be viewed as a subset of generic “data science” that focuses on the special characteristics of spatial data, i.e., the importance of “where.” ‘

other. The former treats spatial information, such as the latitude and longitude of data points as simply an additional variable, but otherwise does not adjust analytical methods or software tools. In contrast, “true” spatial data science treats location, distance, and spatial interaction as core aspects of the data and employs specialized methods and software to store, retrieve, explore, analyze, visualize and learn from such data. In this sense, spatial data science relates to data science as spatial statistics to statistics, spatial databases to databases, and geocomputation to computation."

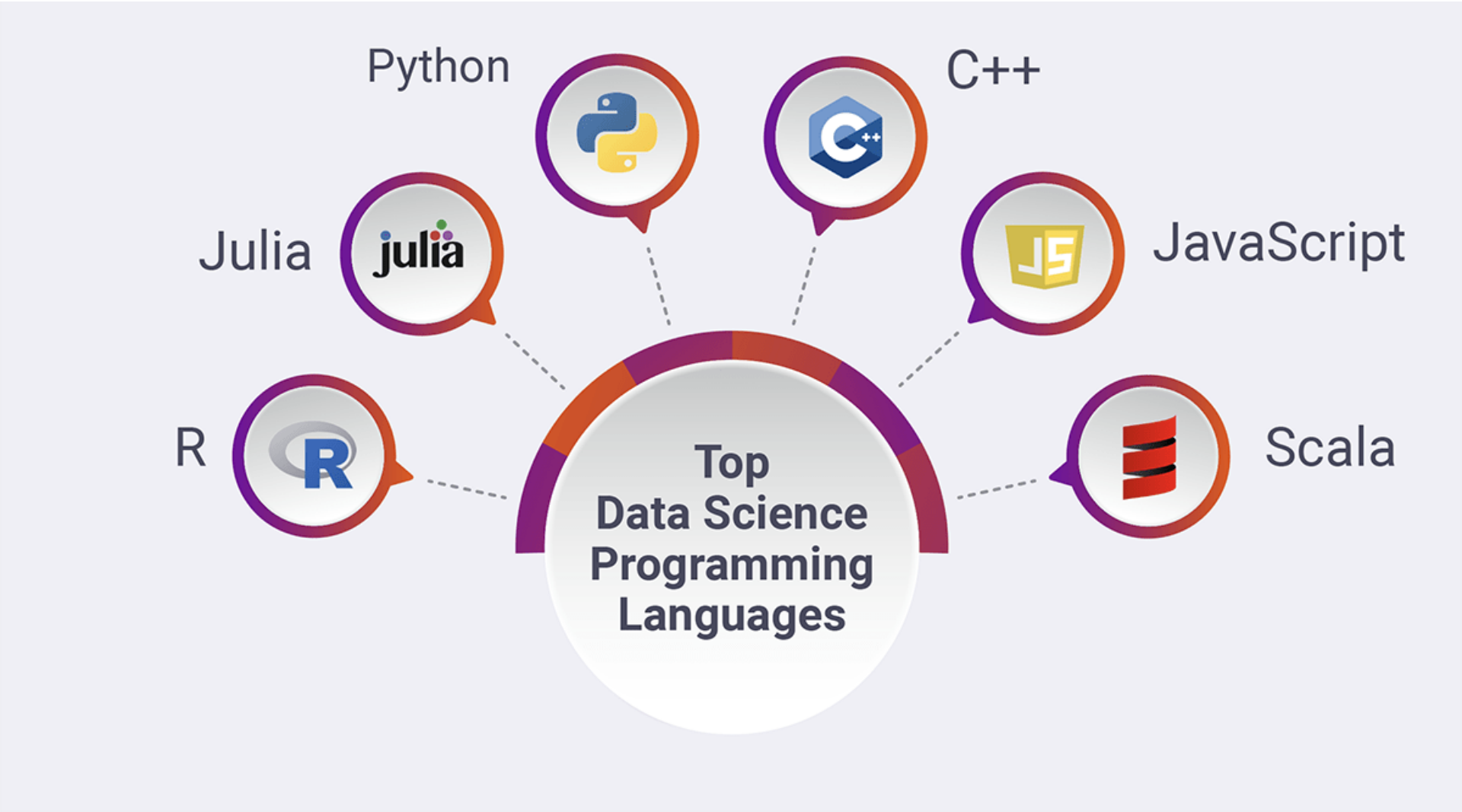
Professor Luc Anselin

father in the field of spatial data science


GeoPandas
Pandas
ArcPy
Seaborn
Scipy
Tensorflow
Leaflet
Scikit-Learn
Keras
NetworkX
Numpy
Pandana
Anaconda
CARTOframes
Jupyter Notebook
Python
PgAdmin
Postgres
SQL
PostGIS
Databases
SPATIAL ANALYSIS
Maps API
CARTO
QGIS
Esri
Tableau
FME
Alteryx
Google Earth
GeoSpark
SpatialKey
ElasticSearch
Oracle
AWS
Platforms
GEE
Spdep
Sf
Ggplot
Rgdal
Rgeos
Gstat
INLA
RStan
Mgcv
Shiny
R


What is Spatial Data Science?

Data Science Tools


Python
C++
Julia
JavaScript
R
Scala
Top
Data Science
Programming
Languages

Comparison Points	Java	Scala	Python	R
Performance	Faster	Faster (about 10x faster than Python)	Slower	Slower
Learning Curve	Easier than Java Tougher than Python	Steep learning curve than Java & Python	Easiest	Moderate
User Groups	Web/Hadoop programmers	Big Data Programmers	Beginners & Data Engineers	Data Scientists/ Statisticians
Usage	Web development and Hadoop Native	Spark Native	Data Engineering/ Machine Learning/ Data Visualization	Visualization/ Data Analysis/ Statistics use cases
Type of Language	Object-Oriented, General Purpose	Object-Oriented & Functional General Purpose	General Purpose	Specifically for Data Scientists. Needs conversion into Scala/Python before productizing
Concurrency	Support Concurrency	Support Concurrency	Does not Support Concurrency	NA
Ease of Use	Verbose	Lesser Verbose than Scala	Least Verbose	NA
Type Safety	Statically typed	Statically typed (except for Spark 2.0 Data frames)	Dynamically Typed	Dynamically Typed
Interpreted Language (REPL)	No	No	Yes	Yes
Maturated machine learning libraries availability/ Support	Limited	Limited	Excellent	Excellent
Visualization Libraries	Limited	Limited	Excellent	Excellent
Web Notebooks Support	Ijava Kernel in Jupyter Notebook	Apache Zeppelin Notebook Support	Jupyter Notebook Support	R Notebook

What makes Python great

It is a simple, open-source, general-purpose language and is very easy to learn. It has a rich set of libraries, utilities, ready-to-use features and support to a number of mature machine learning, big data processing, visualization libraries.

- **Advantages of Python:**

1. Easy to learn, easy debugging, fewer lines of code.
2. It is dynamically typed. i.e. can dynamically defined variable types. i.e. Python as a language is type-safe.
3. Python is platform agnostic and scalable

Python main Libraries

- Scientific Computing:
  - NumPy
  - SciPy
  - Pandas
  - Matplotlib
- Artificial Intelligent:
  - TensorFlow
  - Keras
  - SciKit-Learn
  - PyTorch
- Natural Language Processing (NLP):
  - Natural Language Toolkit (NLTK)

R Programming Language

- **R Language**
- R is the favourite language of statisticians. R is fondly called a language of statisticians. It's popular for research, plotting, and data analysis. Together with RStudio, it makes a killer statistic, plotting, and data analytics application.
- R is majorly used for building data models to be used for data analysis.
- **Advantages/Features of R:**

1. Strong statistical modeling and visualization capabilities.
2. Support for 'data science' related work.
3. It can be integrated with Apache Hadoop and Spark easily.

- **Drawbacks/Disadvantages of R:**

1. R is not a general-purpose language.
2. The code written in R cannot be directly deployed into production. It needs conversion into Java or Python.
3. Not as fast as Java / Scala.

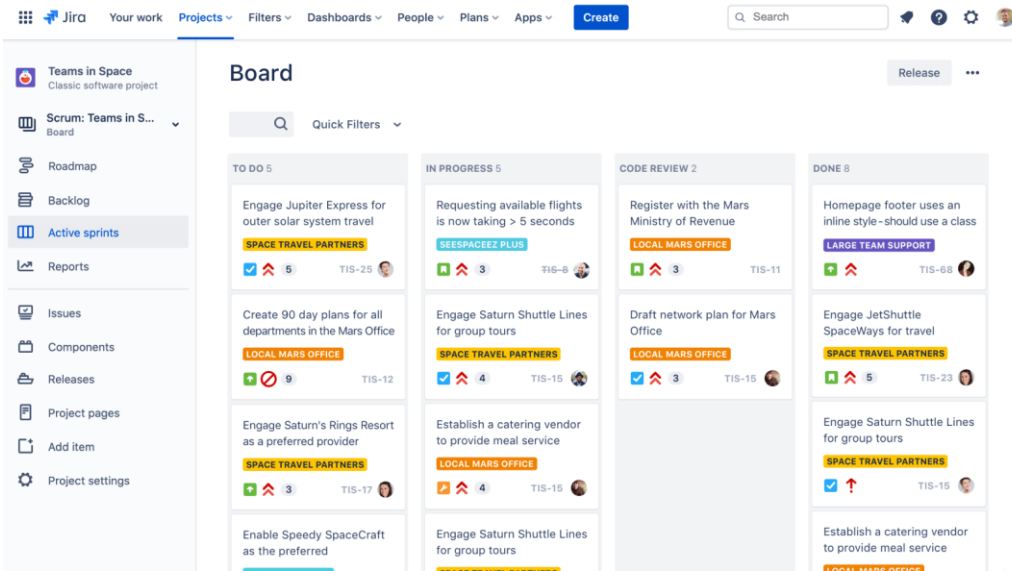
GitHub

- **Version control** is a system that records changes that are made to a file or a set of files over time.

- **GitHub** is an online interface for Git. Git is software used locally on your computer to record changes. GitHub is a host for your files and the records of the changes made. You can sort of think of it as being similar to DropBox - the files are on your computer, but they are also hosted online and are accessible from any computer. GitHub has the added benefit of interfacing with Git to keep track of all of your file versions and changes.

Jira

- Jira Software is an agile project management tool that supports any agile methodology
- Create tasks for yourself and members of your team to work on, complete with its details, due dates, and reminders. Utilize subtasks to breakdown larger items of work. Allow others to watch the task to track its progress and be notified when it's completed. Create sub-tasks within the parent task to break down the unit of work into digestible pieces for various members of the team. View all tasks on the board to easily visualize each's status.



Assignment 1

Due Date: Monday January 24, 9 a.m.

- Download Python on your machine (or Anaconda)
- Install Python
- Make Google Colab account: https://colab.research.google.com/
- Make GitHub account: GitHub
- Answer questions
- Upload your codes in the GitHub
- Submission: Individual