

Introduction to Big Data

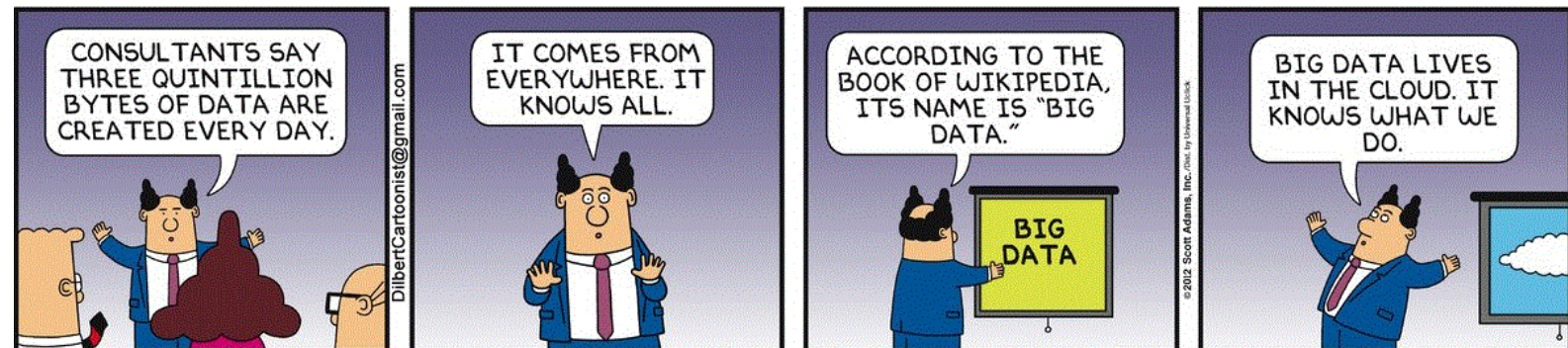
- CSCI316 -
**Big Data Mining Techniques and
Implementation**

What is Big Data?

Definition:

Big Data is used in the singular and refers to a collection of data so large and complex, it's impossible to process them with the usual databases and tools. Because of its size, *Big Data* can be hard to capture, store, search, share, analyze and visualize.

Even so the term Big Data is well known its meaning is often not well understood:



What is Big Data?

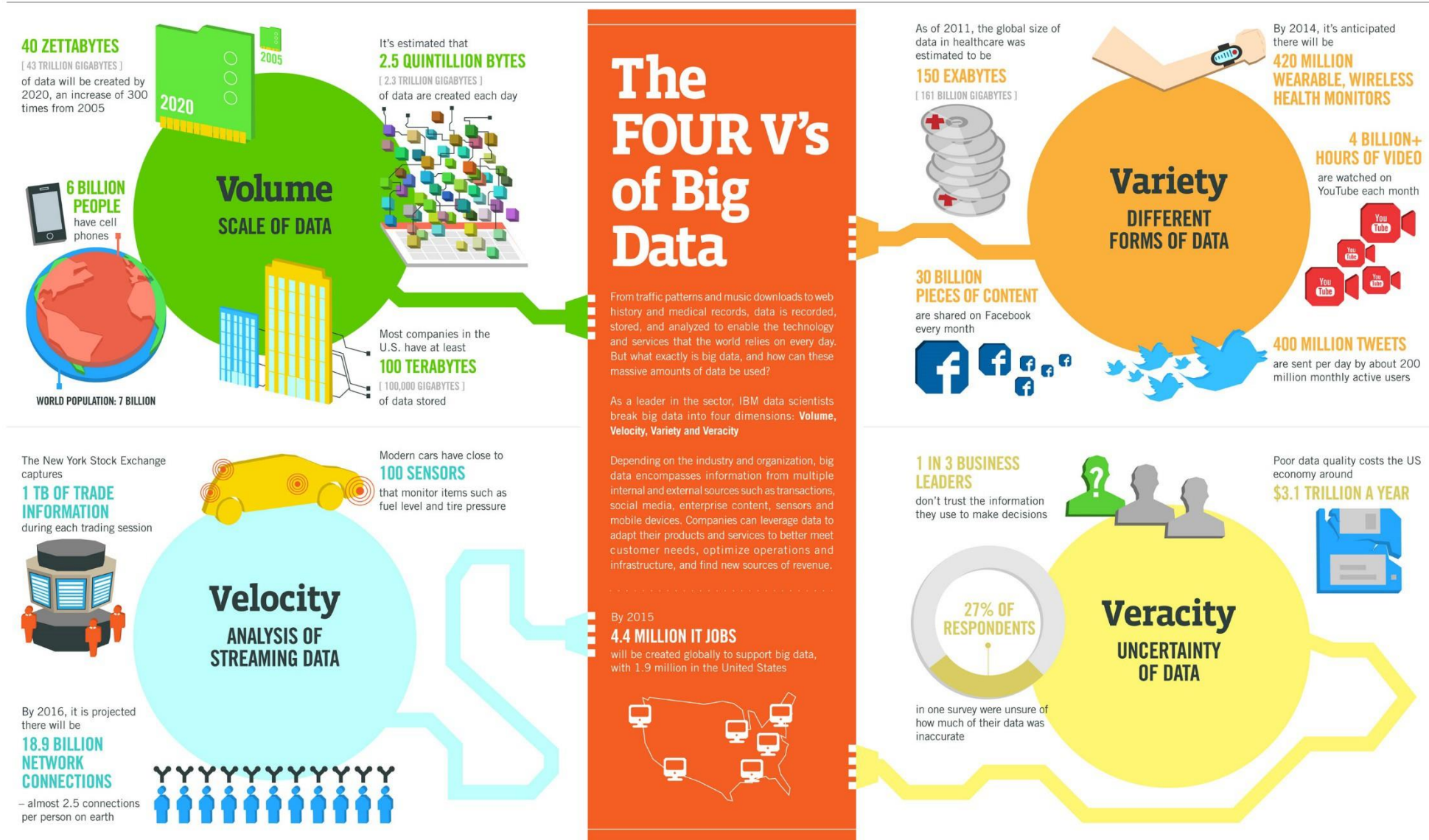
Big Data is also known under alternative terms such as

- Smart Data
- Predictive Analysis
- Data Science
- Massive Data

Big Data has general properties. These are:

- Velocity, Variety, Volume, Veracity
- famously known as the four Vs

What is Big Data? The 4 Vs.



The 4Vs

Volume:

- Concerns the sheer volume of data.
- A typical PC stores TBs of data but data is created at a much higher rate:
 - 3.8 Billion internet users per day (as of 2016)
 - Youtube: 400 hours of videos added every minute (as of 2016)
 - Facebook: 3 million posts per minute (as of 2016)
 - Google: 3,607,080 searches per minute
 - SMS: 15,220,700 texts per minute
 - Instagram: 46,740 pictures per minute.
- In 2016 an estimated 44 Billion GB (Exabyte) of data was created each day, predicted to grow to 463 billion GB by 2025.

The 4Vs

Variety:

- Data comes in many forms and can vary in:
 - Structure: structured (i.e. forms), semi-structured (i.e. newspaper article), unstructured (meta-data).
 - Media: Type of data (i.e. text, multimedia, audio, 3D, geo)
 - Semantic variety: Interpretation of values. (i.e. age=3 vs age=infant, income=55k vs income=above_average)
 - Availability variations: real time (i.e. sensory), intermitted (i.e. satellite data,) or stored (i.e. records).
- Increased data diversity
- Adds complexity

The 4Vs

Velocity:

- Refers to:
 - A. Speed by which data is generated, stored, and analysed.
 - How much data is generated per unit of time?
 - Speed by which results need to become available.
 - May require real time processing (i.e. streaming data).
 - B. Speed by which data changes over time
 - Domain changes, environmental changes, changes in user behaviour, changes in expectations.
 - May require regular update of models

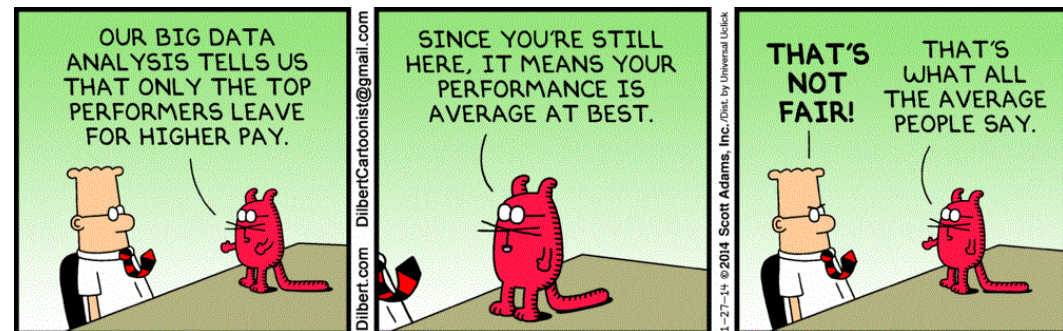
The 4Vs

Veracity:

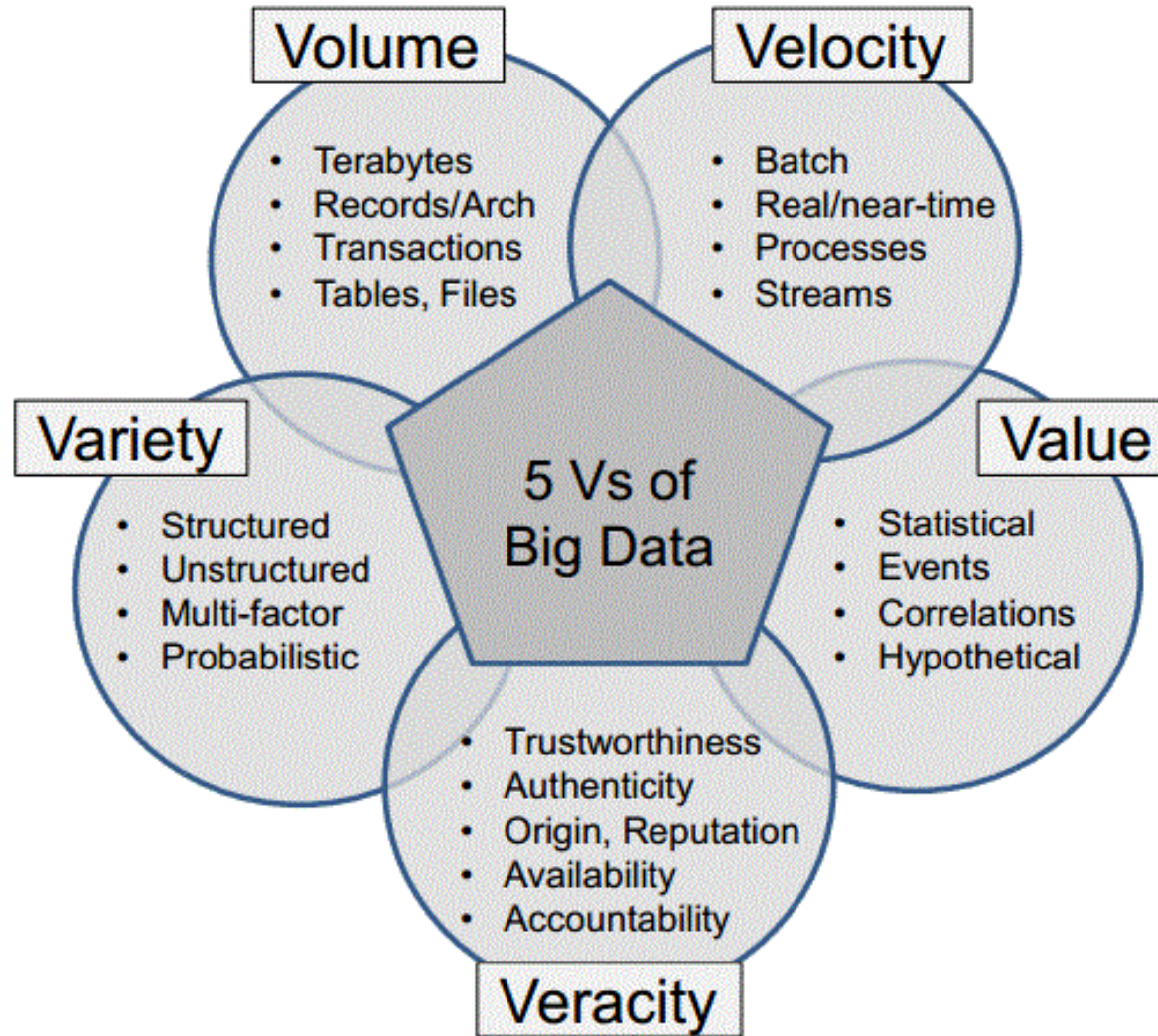
- Refers to data quality, data uncertainty, imprecise data types.
- Data validity:
 - Noise and accuracy of data.
 - Regulated vs unregulated
- Data volatility:
 - Is data collected in the past still valid today?
 - Are results from data collected today valid for future decision making applications?
- Big Data is only as good as the quality of the data (junk in = junk out)

The fifth V?

- The four Vs are said to be fundamental dimensions of Big Data.
- Although **Value** is at the heart of Big Data:
- Refers to the value of Big Data results (the new insights obtained):
 - Academic value: Domain understanding, method development,...
 - Statistical value: To get a better overview
 - Correlations: Discovery of links and relationships.
 - Business value: Buying and selling data, buying and selling results, decision support.
- Note that “value” is in the eye of the beholder:






4 or 5Vs?



From a practical perspective:

- Big data only makes sense when there is value associated with it.
- Volume, Velocity, Variety, and Veracity refer to property of data (the input) whereas Value refers to the envisaged results (the output).

How to collection data












Method	Effort	
Download	Low	
API (Application program interface)	Medium	
Scrape/Crawl	High	

Data You Can Just Download

- NYC Taxi data: Trip (11GB), Fare (7.7GB)
- StackOverflow (xml)
- Wikipedia (data dump)
- Atlanta crime data (csv)
- Soccer statistics
- ...

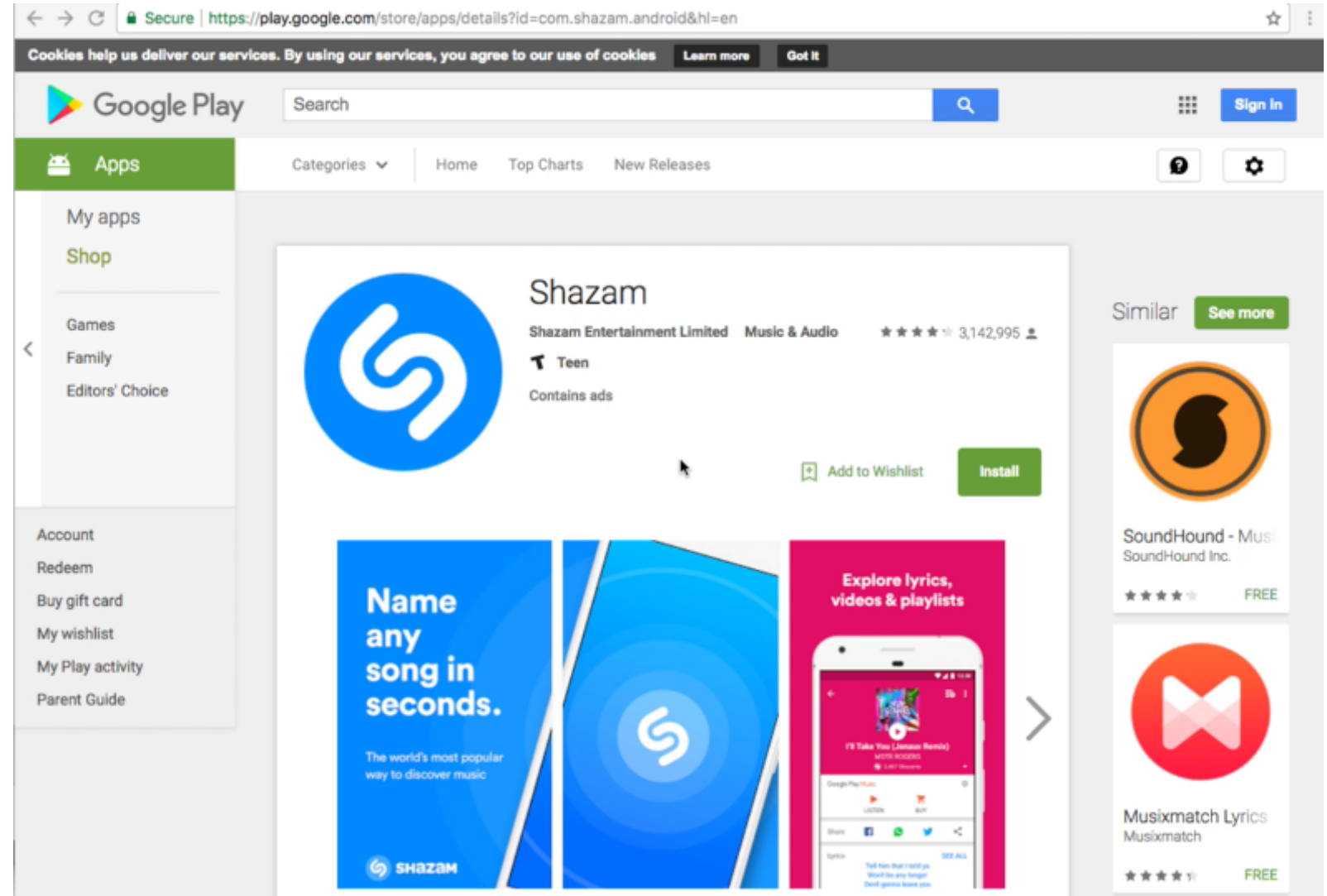
Collect Data via APIs

- Google Data API (e.g., Google Maps Directions API)
<https://developers.google.com/gdata/docs/directory>
- Twitter (small subset)
<https://dev.twitter.com/streaming/overview>
- data.gov
- Facebook (your friends only)

API	GData Status	See Also
 Google Analytics Data Export API	Replaced by Google Analytics Core Reporting API (starting at version 2.4).	Migration Guide: Moving from v2.3 APIs to v2.4 & v3.0
 Google Apps Provisioning API	Shut down. Replaced by the Admin SDK Directory API .	Current Google Apps APIs
 Google Base Data API	Not available since June 1, 2011. Replaced by the Content API for Shopping .	New Shopping APIs and Deprecation of the Base API
 Blogger Data API	Replaced by the latest Blogger API .	
 Google Book Search API	Shut down. Replaced by Google Books API Family .	Google books API searching by ISBN (on Stack Overflow)
 Google Calendar API v2	Shut down. Replaced by latest Google Calendar API .	
 Google Code Search Data API	Shut down in Jan 15, 2012. No replacement API.	A fall sweep (Google blog post)
 Google Contacts API	GData version is still live. Replaced by Google People API for read-only access.	Google Contacts API Google People API
 Google Documents List Data API	Shut down. Replaced by Google Drive API .	
 Google Finance Portfolio Data API	Shut down. No replacement API.	Spring cleaning for some of our APIs (Google blog post)
 Google Health Data API	The product was discontinued as of January 1, 2013. No replacement API.	An update on Google Health and Google PowerMeter

How to Scrape

- Google Play example
- Goal: Collect the network of Similar Apps



How to Scrape?

Goal: Write a **program/algorithm** to scrape Google Play to **collect a million-node network** of similar apps



Each **node** is an app

An **edge** connects two similar apps

Hint: start with some apps (e.g., Shazam), and go from there.

Icons from:

<https://play.google.com/store/apps/details?id=com.shazam.android&hl=en>

<https://play.google.com/store/apps/details?id=com.spotify.music>

<https://play.google.com/store/apps/details?id=com.soundcloud.android>

<https://play.google.com/store/apps/details?id=com.soundcloud.android>

How to Scrape?

Google Play example

Goal: collect the network of similar apps

<https://play.google.com/store/apps/details?id=com.shazam.android>



User browser's "developer Tools"
to extract data

<https://play.google.com/store/apps/details?id=com.spotify.music>

Data Unavailable via APIs

- Amazon (reviews, product info)
- ESPN
- eBay
- Google Play
- Google Scholar
- ...

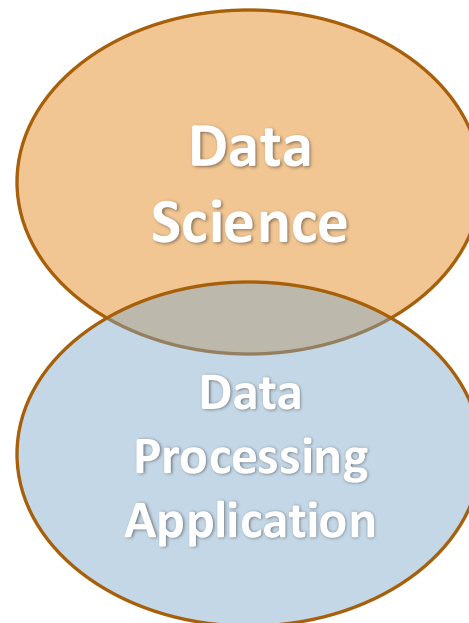
Data Scientist vs. Data Engineer

Data Scientist

- Analyse and model data
- Make prediction based on data
- Build data pipelines to fulfil certain tasks

Data Engineer

- Develop data processing applications
- Deploy the output of data scientists in production



Career Paths and Challenges

- ❖ What Does a Data Scientist Do?
- ❖ Skills Needed to Be a Data Scientist
- ❖ Where Do Data Scientists Work?
- ❖ Related Jobs in Data Science

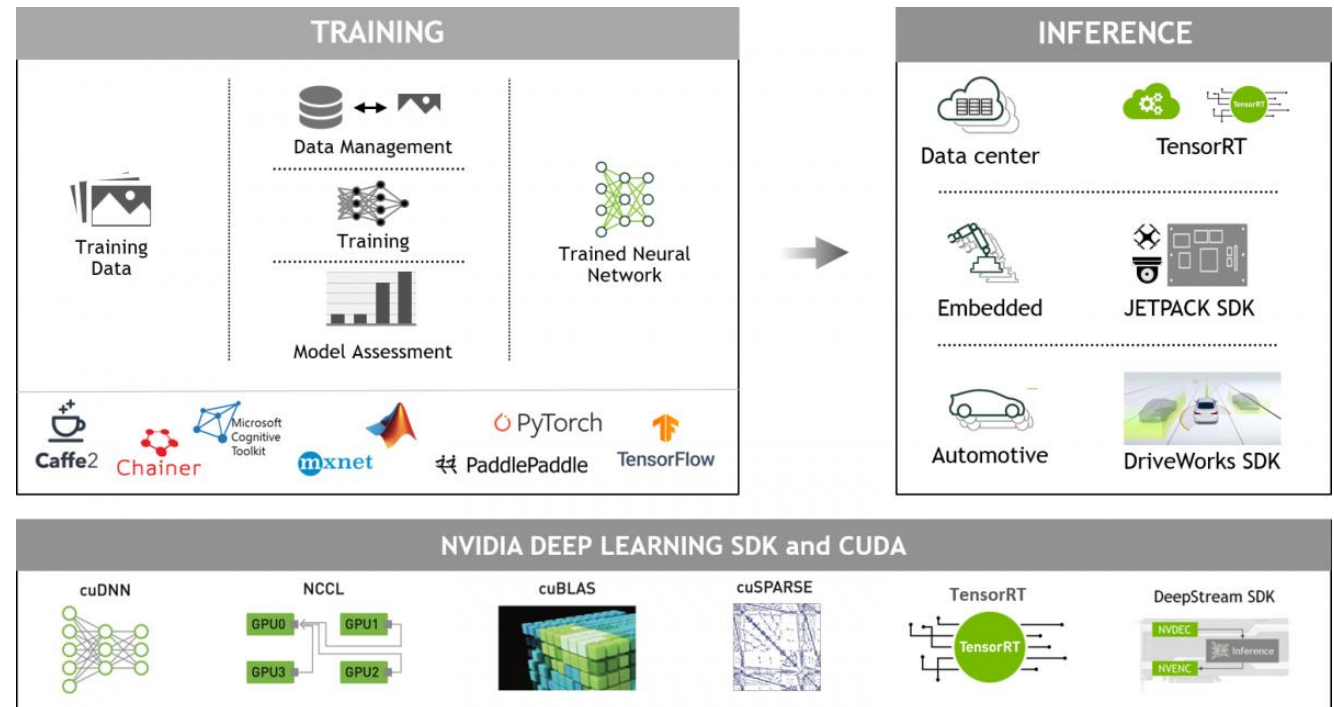
What Does a Data Scientist Do?

Data Collection

Data Preprocessing

Data Visualization

Data Analytics and Application



What Does a Data Scientist Do?

Data Collection

Data Preprocessing

Data Visualization

Data Analytics and Application



Collecting Data from Various Sources

Source: <https://innovativeadagency.com/blog/importance-data-collection/>

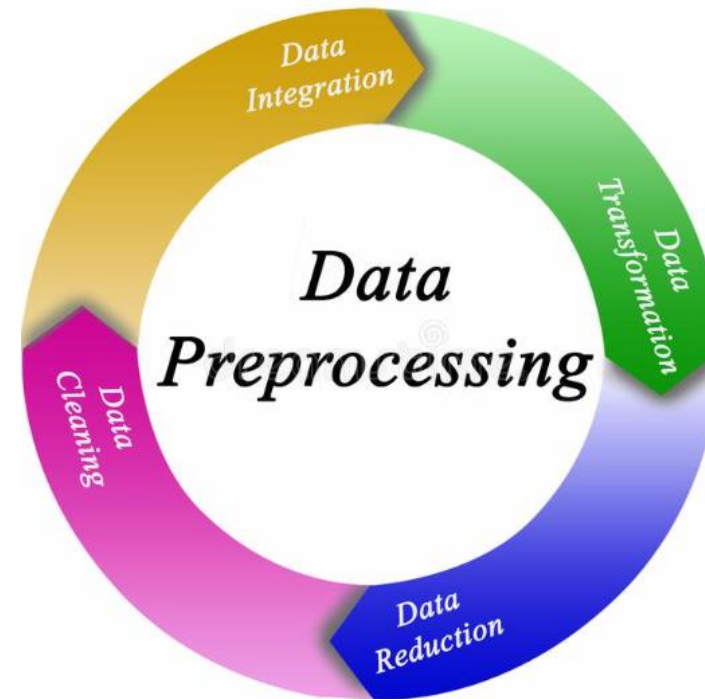
What Does a Data Scientist Do?

Data Collection

Data Preprocessing

Data Visualization

Data Analytics and Application



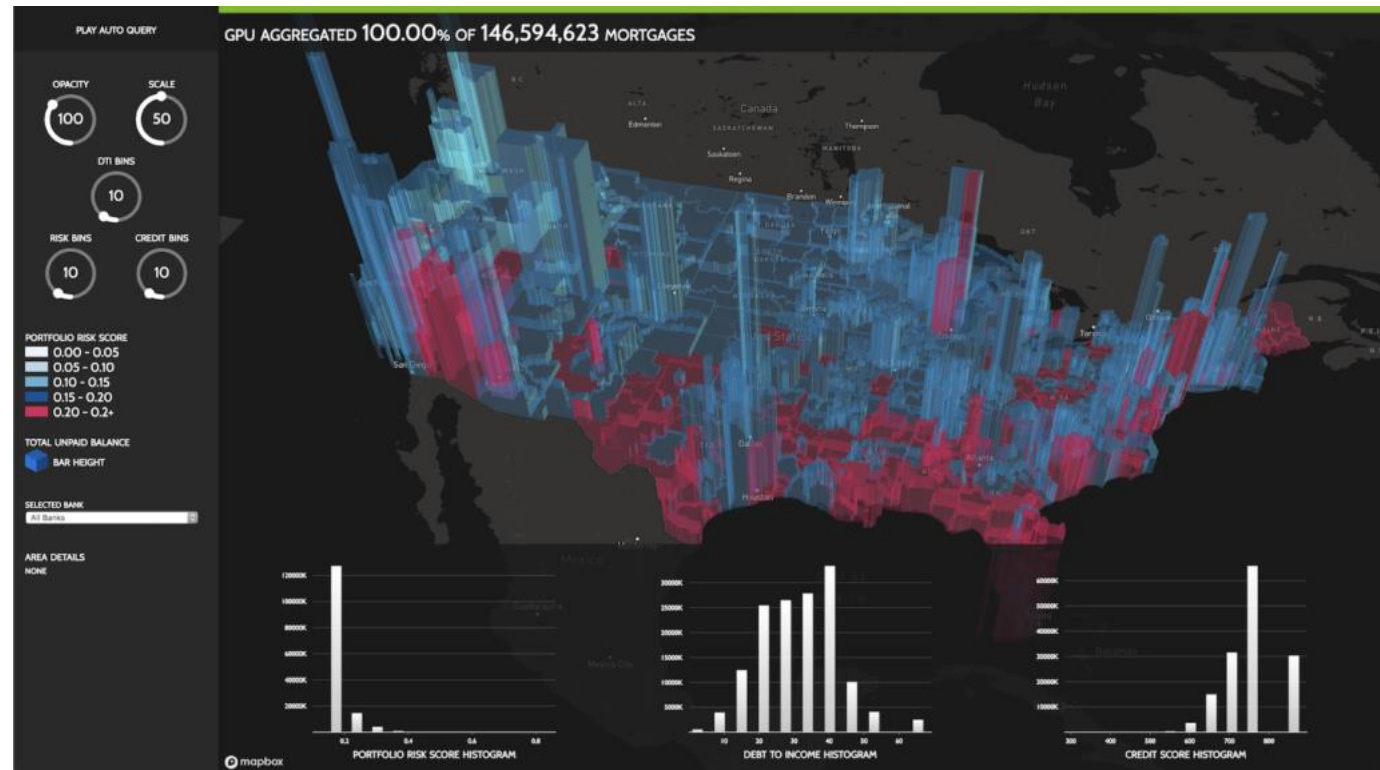
What Does a Data Scientist Do?

Data Collection

Data Preprocessing

Data Visualization

Data Analytics and Application



Source: <https://developer.nvidia.com/blog/gpu-accelerated-analytics-rapids/>

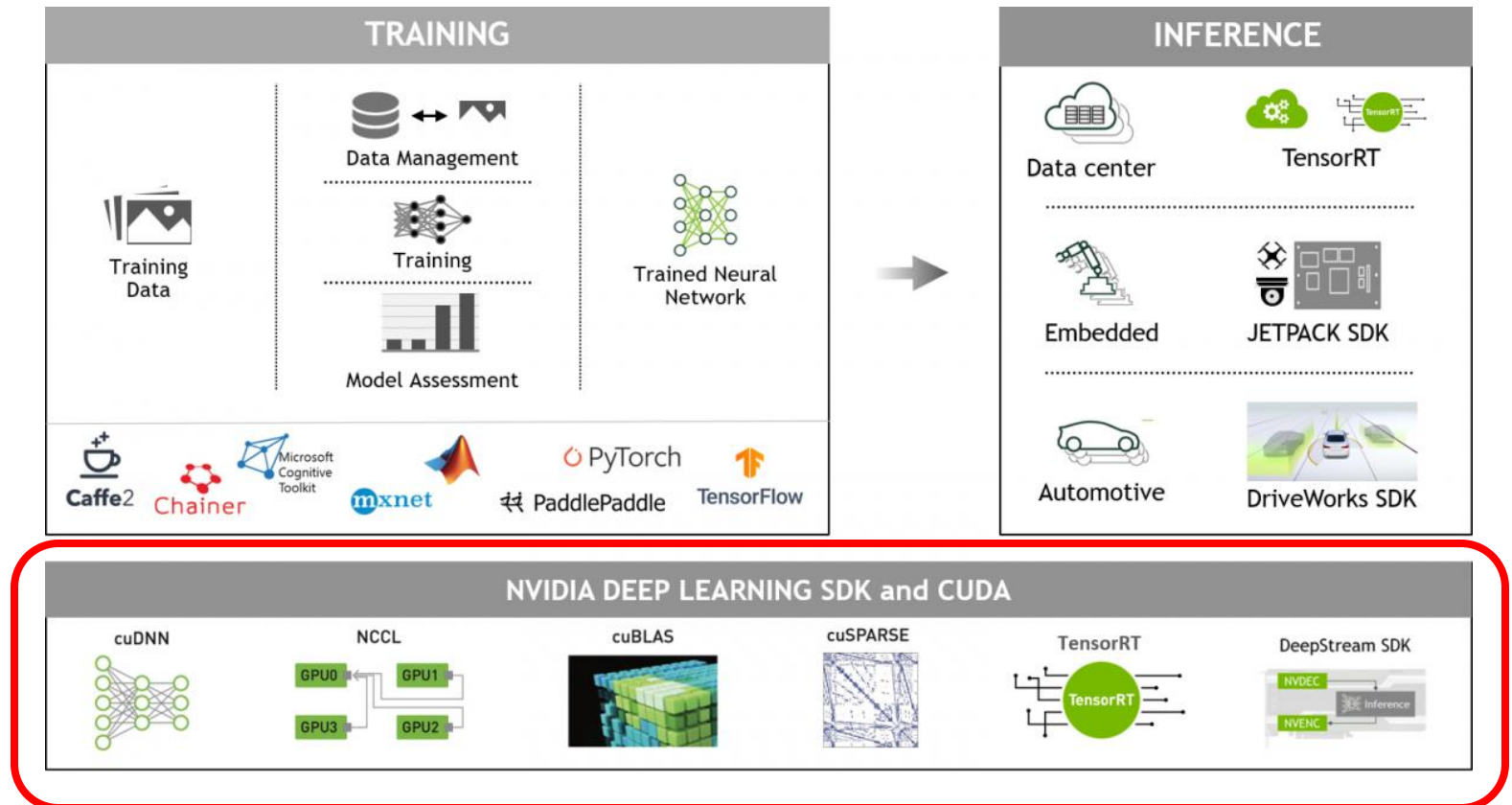
What Does a Data Scientist Do?

Data Collection

Data Preprocessing

Data Visualization

Data Analytics and Application



Skills Needed to Be a Data Scientist

Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency

Skills Needed to Be a Data Scientist

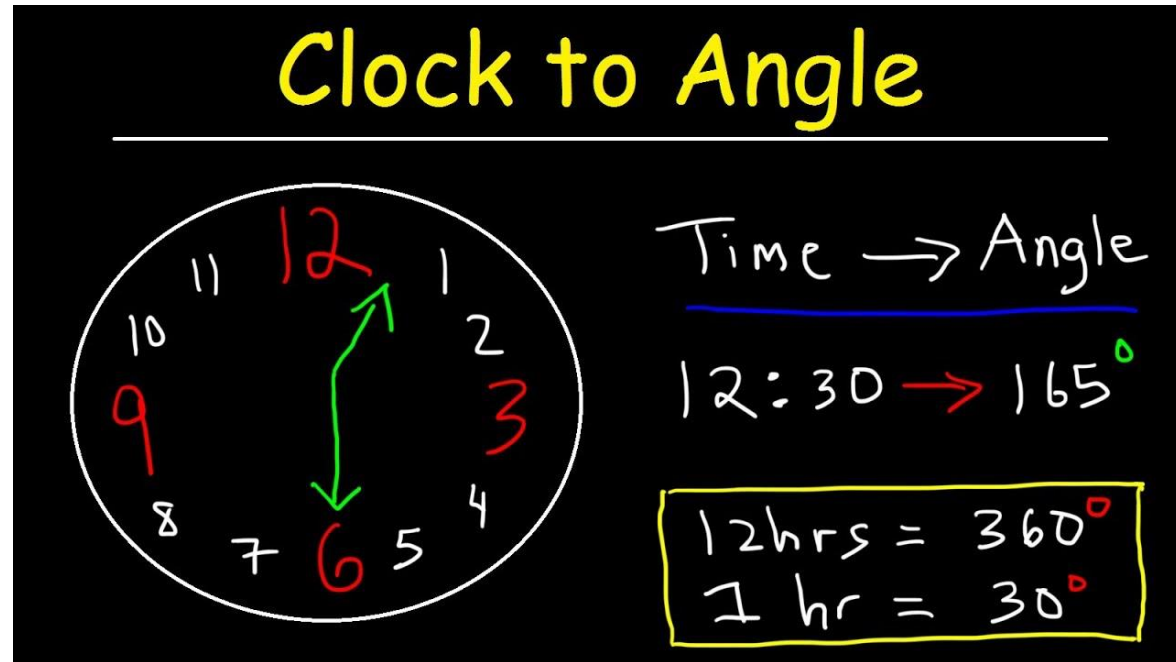
Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency



Source: <https://www.youtube.com/watch?v=LEHYr0XfSyl>

Skills Needed to Be a Data Scientist

Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency



Skills Needed to Be a Data Scientist

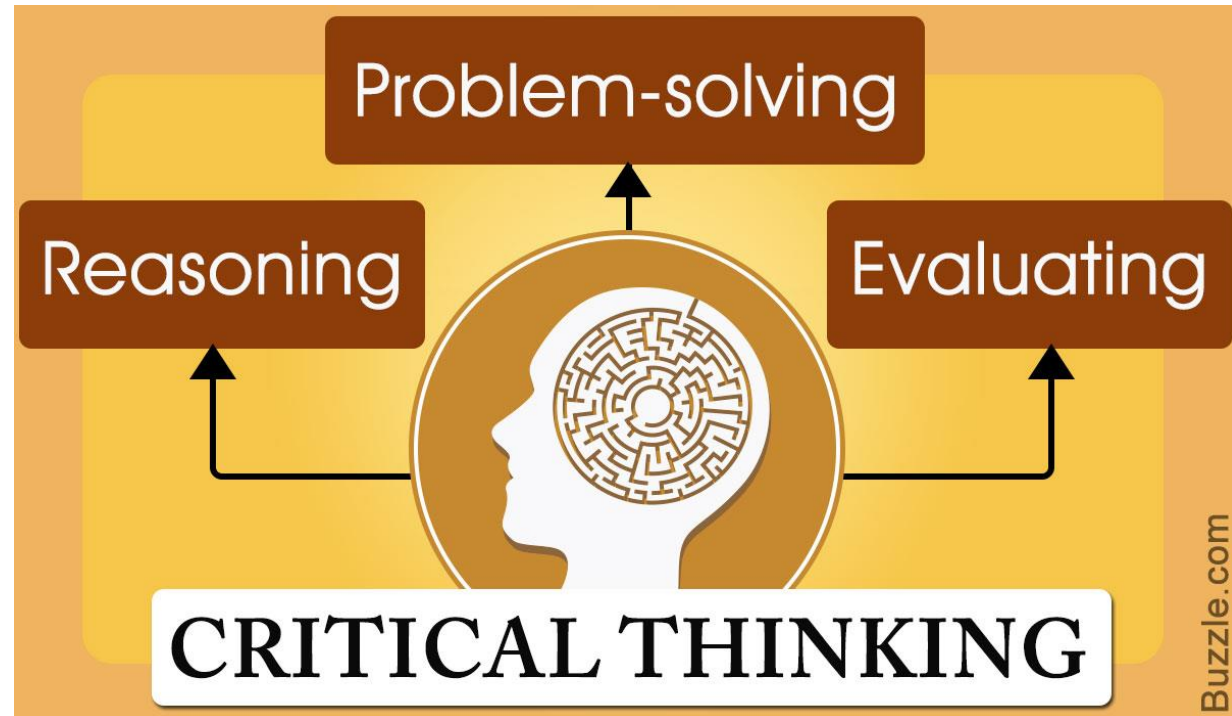
Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency



Skills Needed to Be a Data Scientist

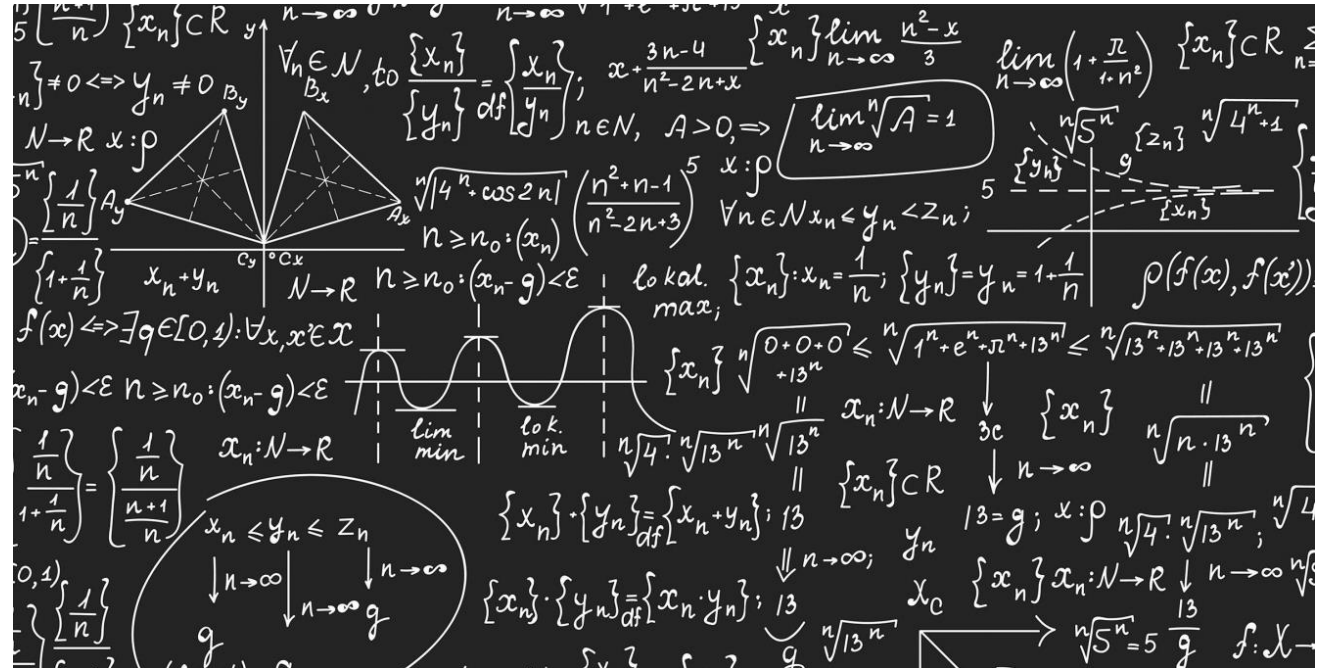
Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency



Skills Needed to Be a Data Scientist

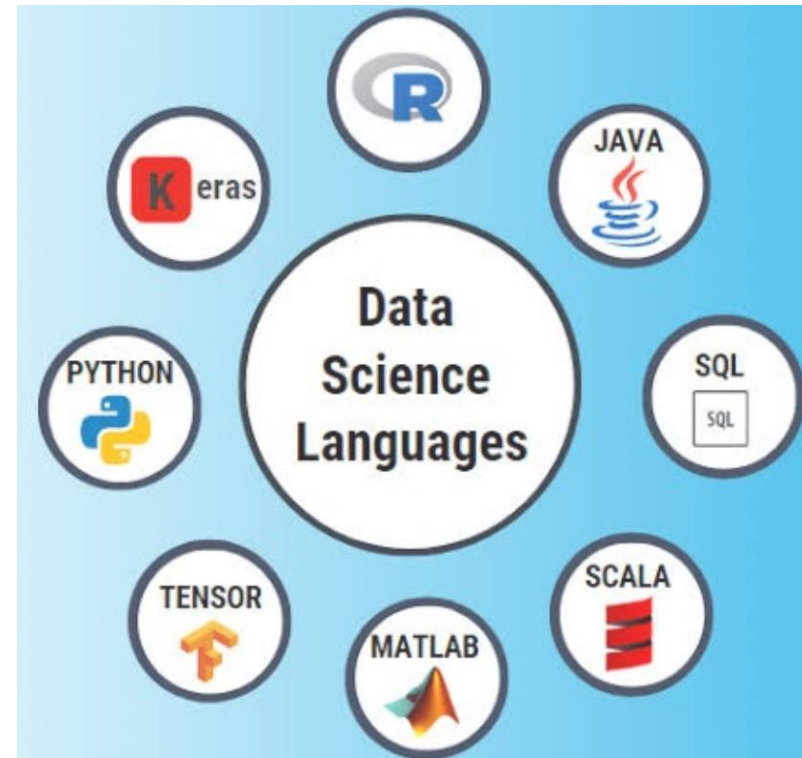
Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency

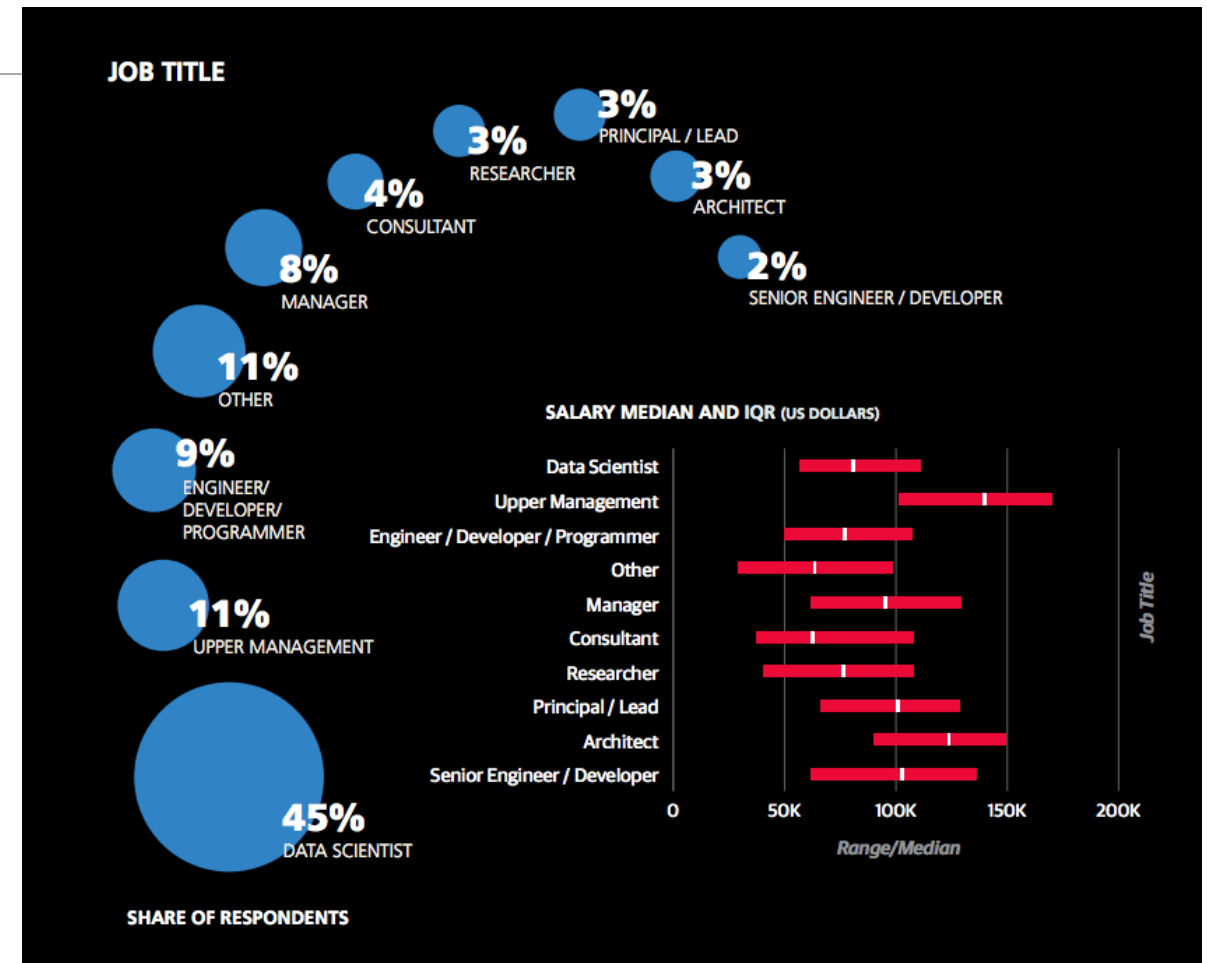


What Salary Can a Data Scientist Earn?

Data Scientist Salary by Job Title

According to an O'Reilly data science salary report, 45 percent of those surveyed said they hold the title of “data scientist.”

In general, the more a data science professional engages in managerial tasks, the higher the salary.



Source: <https://datasciencedegree.wisconsin.edu/data-science/data-scientist-salary/>

Where Do Data Scientists Work?

Academia

- Research and development
- Colleges and universities
- ...

Industry

- Software companies
- Car companies
- Delivery companies
- ...

Related Jobs in Data Science

Data analyst

Research scientist

Machine learning engineer

Big data engineer

...

Data Scientist Job Titles Include:

- Product analyst
- Data analyst
- Research scientist
- Quantitative analyst
- Machine learning engineer
- Data engineer
- Big data engineer
- Back-end engineer
- Natural language processing engineer
- Business analyst
- Statistician
- Economist
- Applied scientist
- Operations research scientist
- Research scientist
- Research engineer
- Machine learning scientist
- Product scientist
- Business intelligence analyst
- Natural Scientist