Introduction to Big Data

- CSCI316 -

Big Data Mining Techniques and Implementation

Statement

I did not invent Big Data nor did anyone else. Big Data just happened. Coined by Roger Mougalas in 2005 and driven by technology and commercial ambitions it has become the new frontier for innovation and competition in a global marketplace.

M. Hagenbuchner

Disclaimer

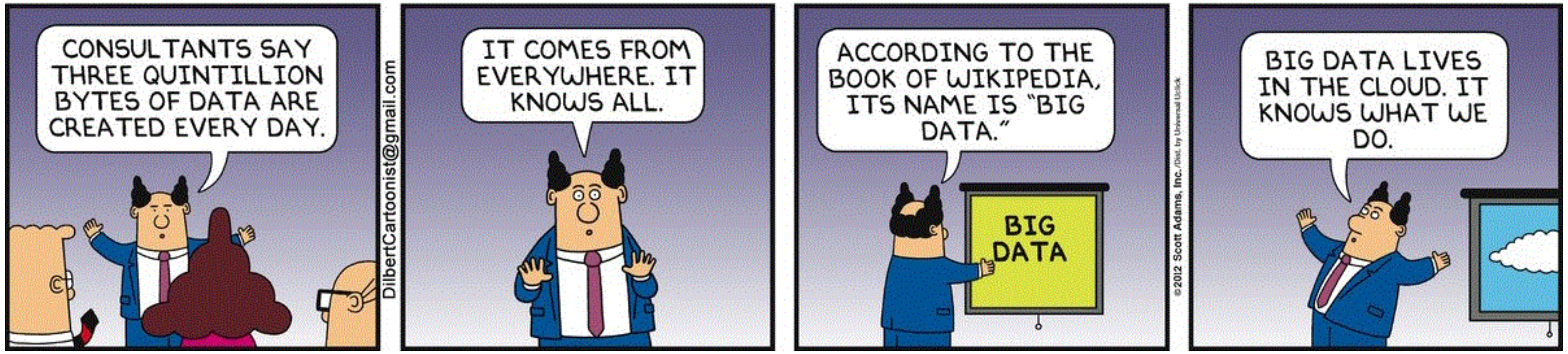
Some lecture notes contain artwork where appropriate to help with the explanation of Big Data concepts to undergraduate CSCI students. Some of the artwork may be protected by copyright and will be used in the lecture notes for purposes that differ from the originally intended purpose. The use of such artwork is hence covered by the fair use policy. Students are informed that **the slides must not be distributed or used beyond the purposes of these lectures**.

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:


CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
©2012 Scott Adams, Inc./Dist. by Universal Uclick
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.

What is Big Data?

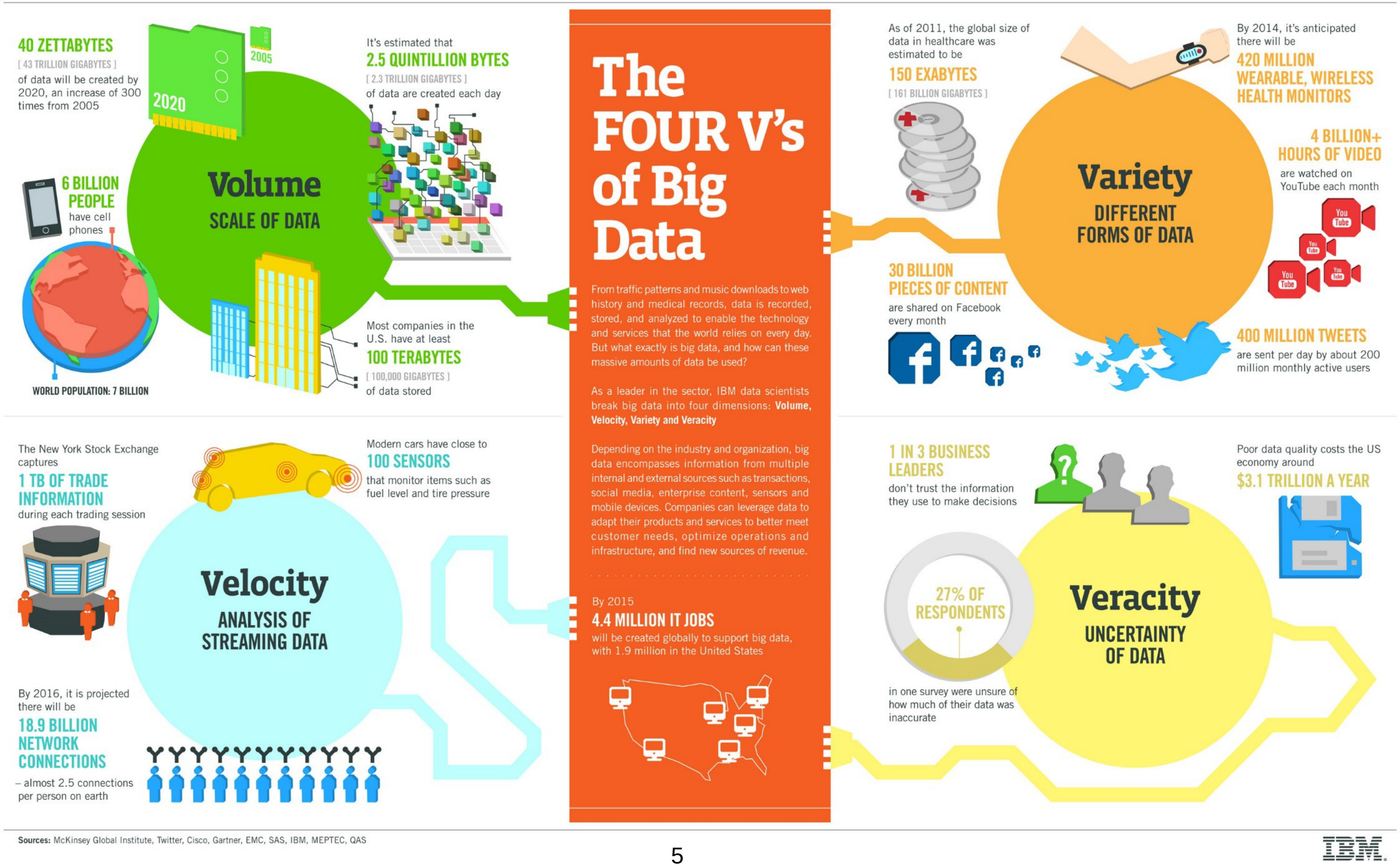
Big Data is also known under alternative terms such as

- 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 FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

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

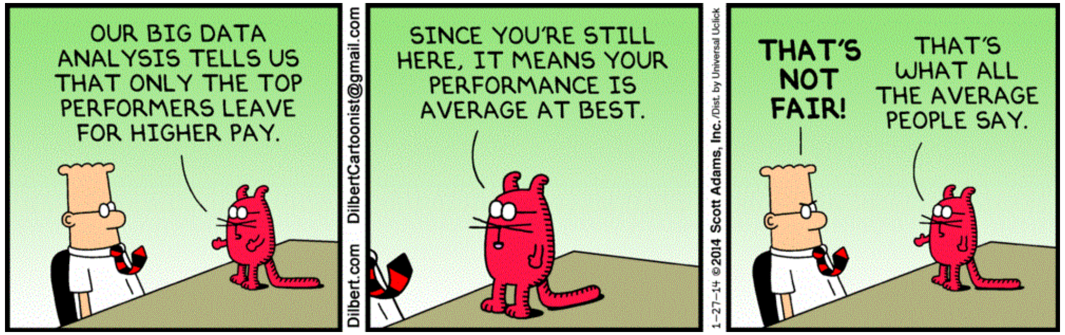
The 4Vs

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?

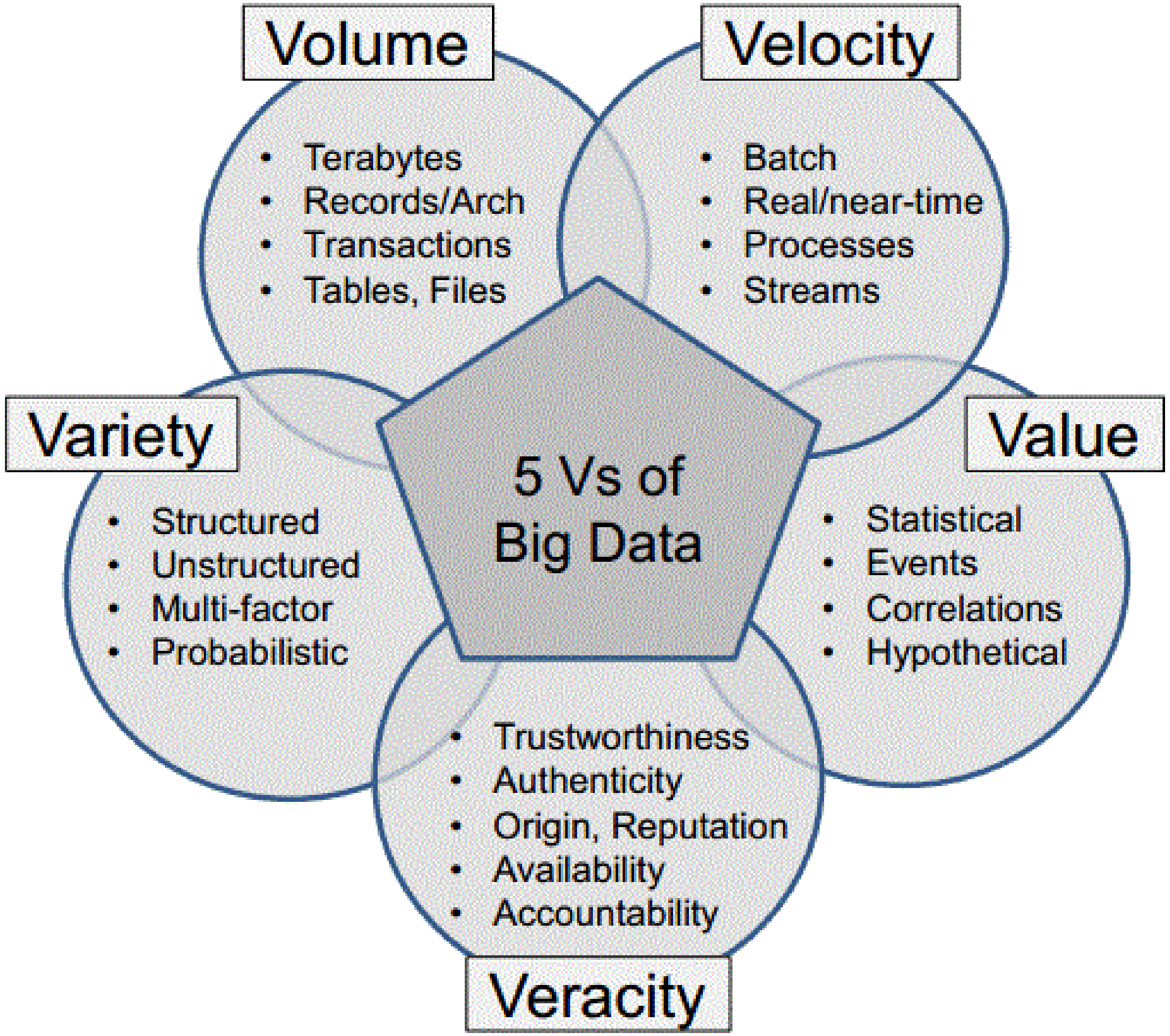
- The four Vs are said to be fundamental dimensions of Big Data.
- Although **Value** is at the heart of Big Data:
- Refers to the value of Big Data results (the new insights obtained):
  - Academic value: Domain understanding, method development,…
  - Statistical value: To get a better overview
  - 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:


OUR BIG DATA ANALYSIS TELLS US THAT ONLY THE TOP PERFORMERS LEAVE FOR HIGHER PAY.
SINCE YOU'RE STILL HERE, IT MEANS YOUR PERFORMANCE IS AVERAGE AT BEST.
THAT'S NOT FAIR!
THAT'S WHAT ALL THE AVERAGE PEOPLE SAY.

Dilbert.com DilbertCartoonist@gmail.com

1-27-14 ©2014 Scott Adams, Inc./Dist. by Universal Uclick

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).

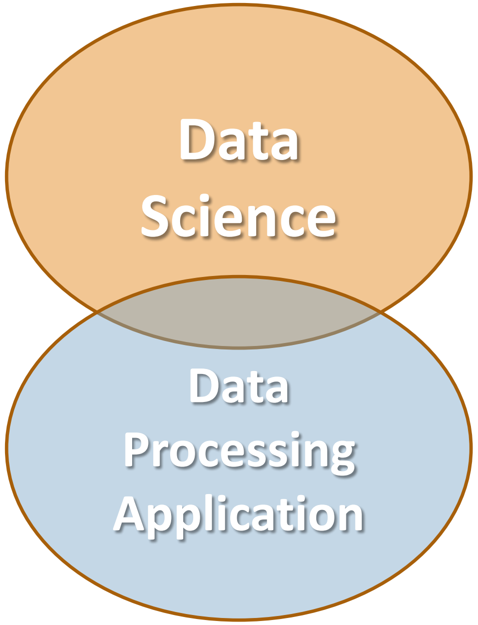
Data Scientist vs. Data Engineer

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
- Challenges in Data Science Careers

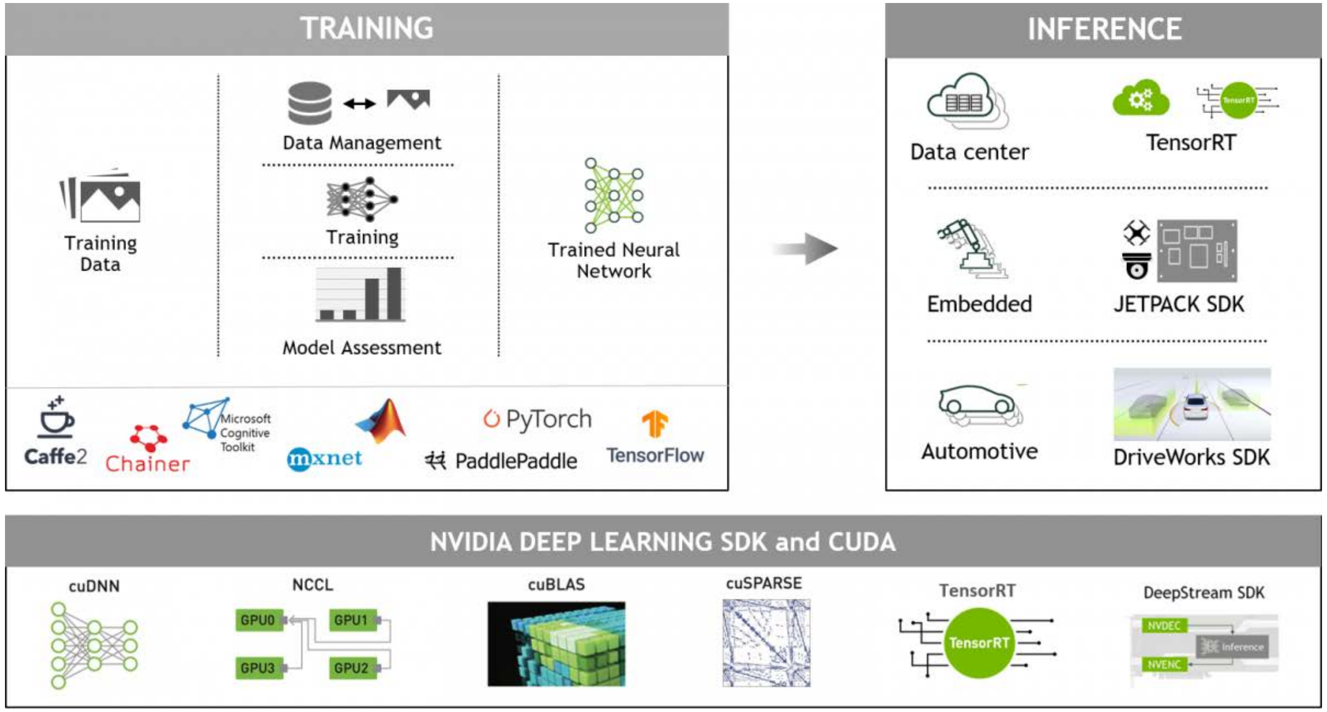
What Does a Data Scientist Do?

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/

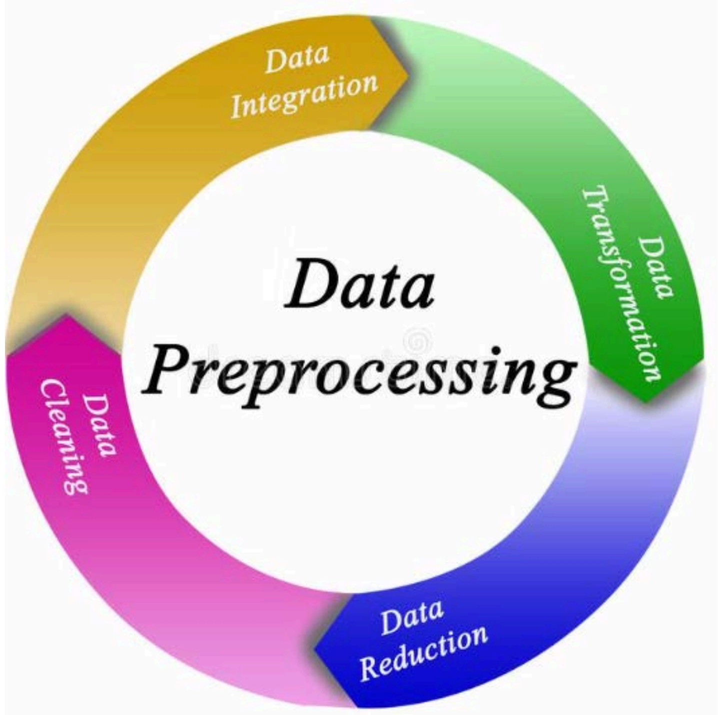
What Does a Data Scientist Do?

Data Collection

Data Preprocessing

Data Visualization

Data Analytics and Application



Source: https://www.dreamstime.com/illustration/data-preprocessing.html

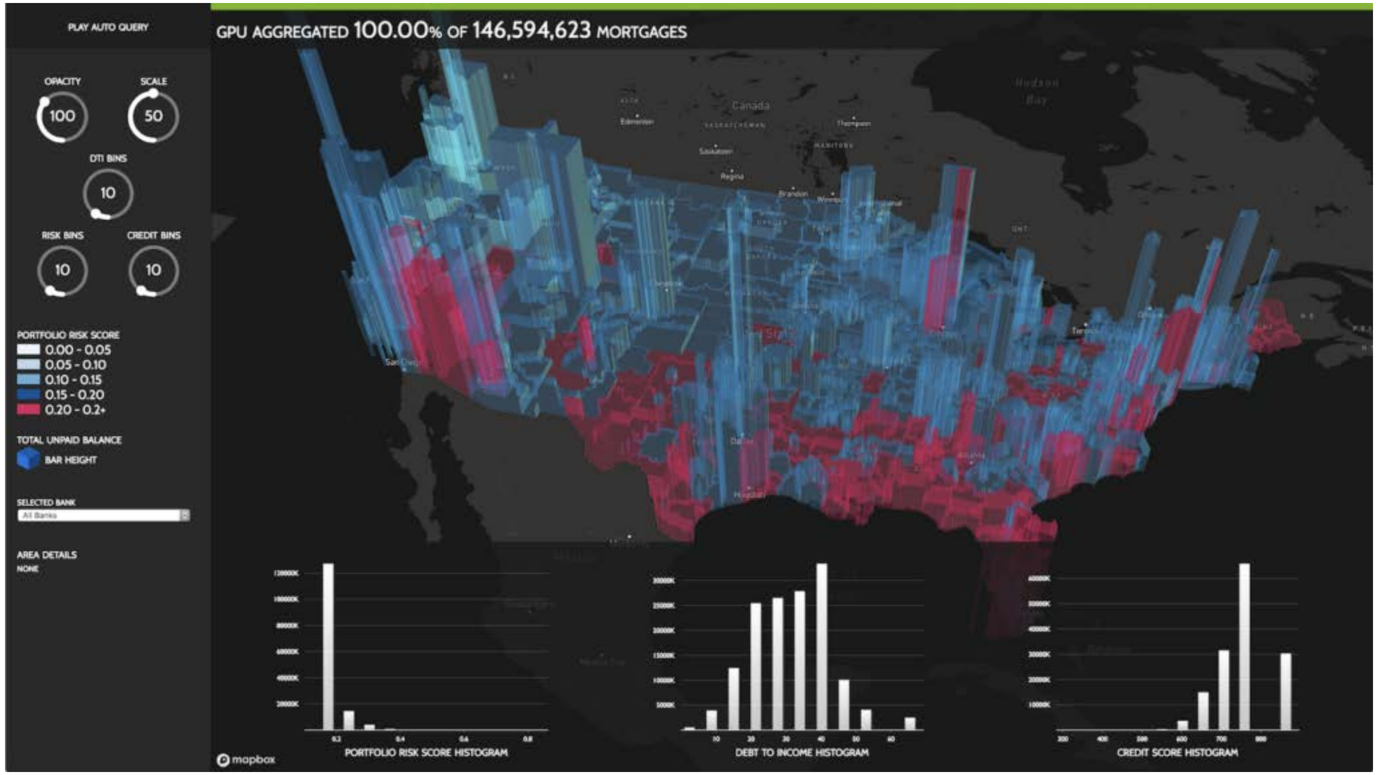
What Does a Data Scientist Do?

Data Collection

Data Preprocessing

Data Visualization

Data Analytics and Application



GPU AGGREGATED 100.00% OF 146,594,623 MORTGAGES

PLAY AUTO QUERY

OPACITY 100

SCALE 50

DTI BINS 10

RISK BINS 10

CREDIT BINS 10

PORTFOLIO RISK SCORE

- ⬜ 0.00 - 0.05
- ⬜ 0.05 - 0.10
- 🟦 0.10 - 0.15
- 🟦 0.15 - 0.20
- 🟥 0.20 - 0.2+

TOTAL UNPAID BALANCE

🟦 BAR HEIGHT

SELECTED BANK

All Banks

AREA DETAILS

NONE

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

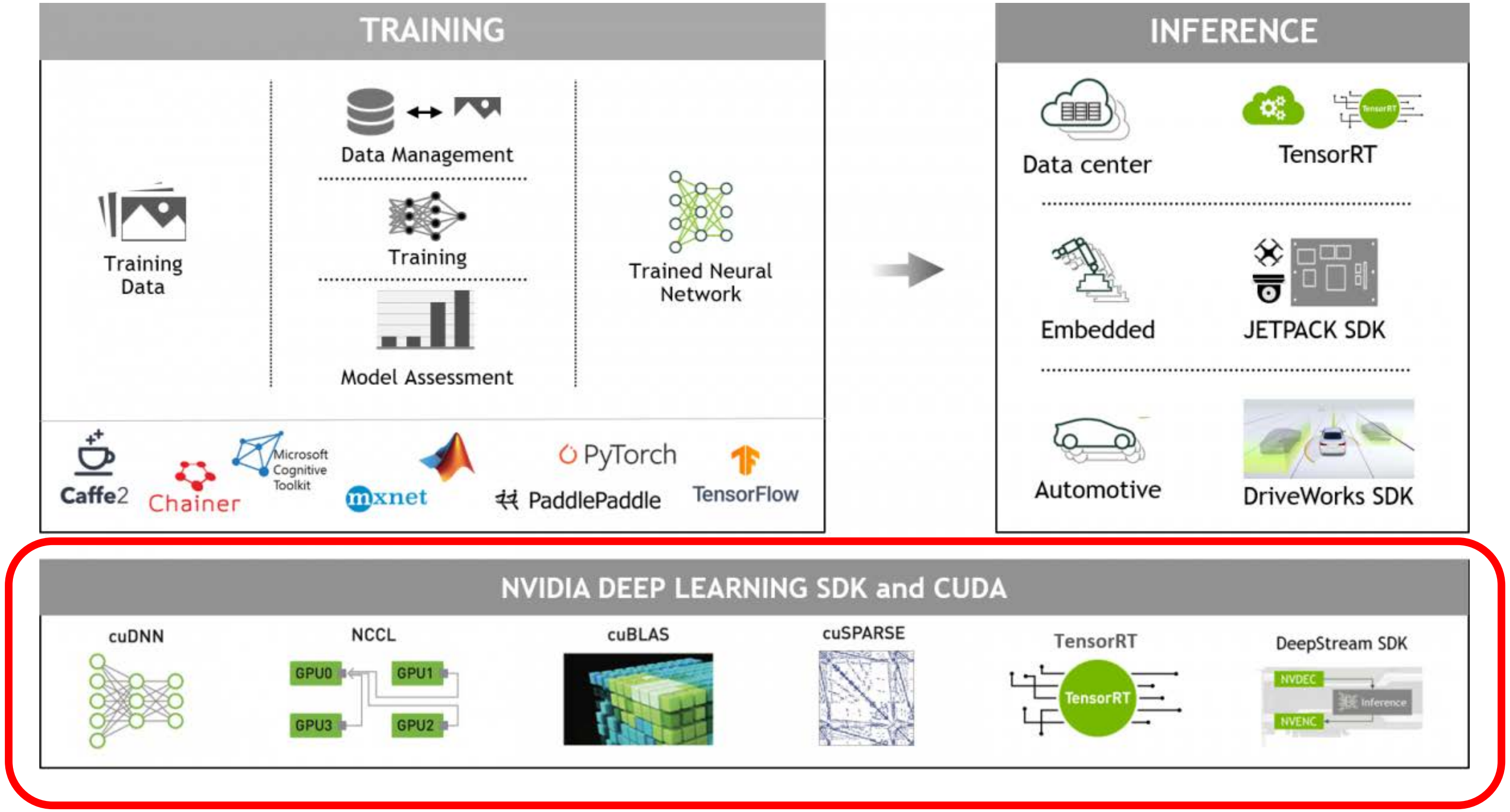
What Does a Data Scientist Do?

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

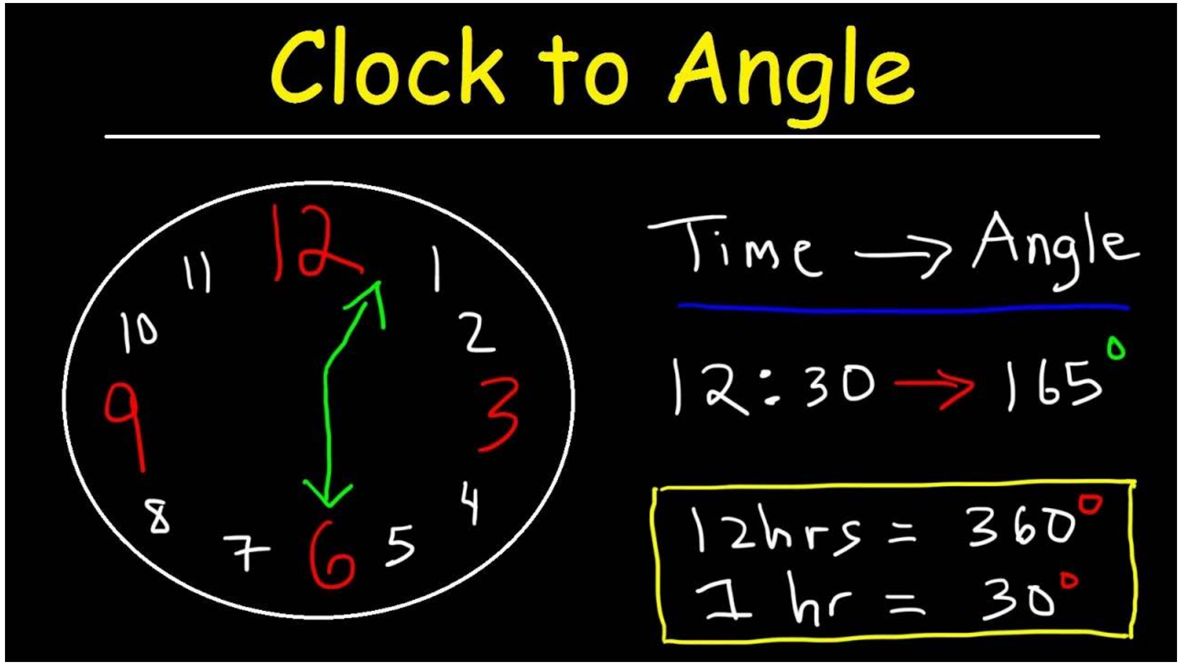
Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency



Source: https://www.youtube.com/watch?v=LEHYr0XfSyI

Skills Needed to Be a Data Scientist

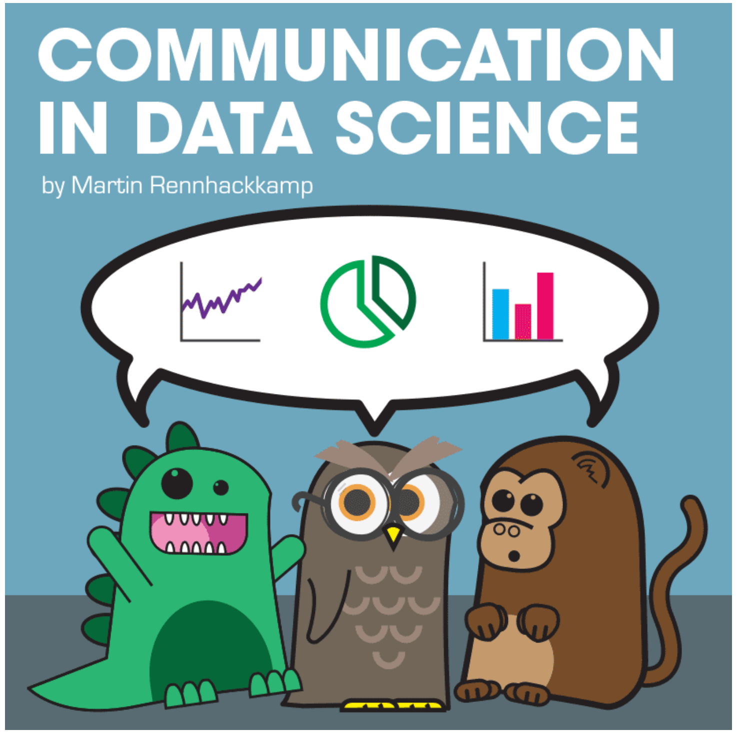
Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency



Skills Needed to Be a Data Scientist

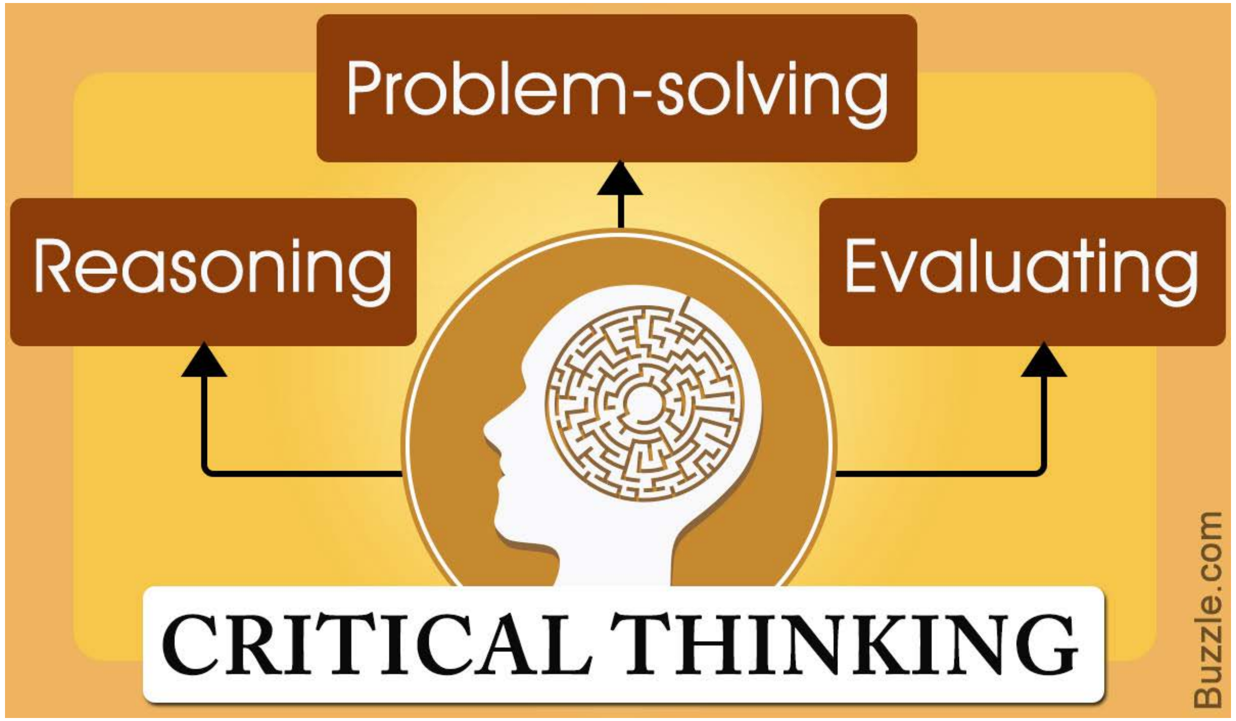
Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency



Source:https://kingdomecon.wordpress.com

Skills Needed to Be a Data Scientist

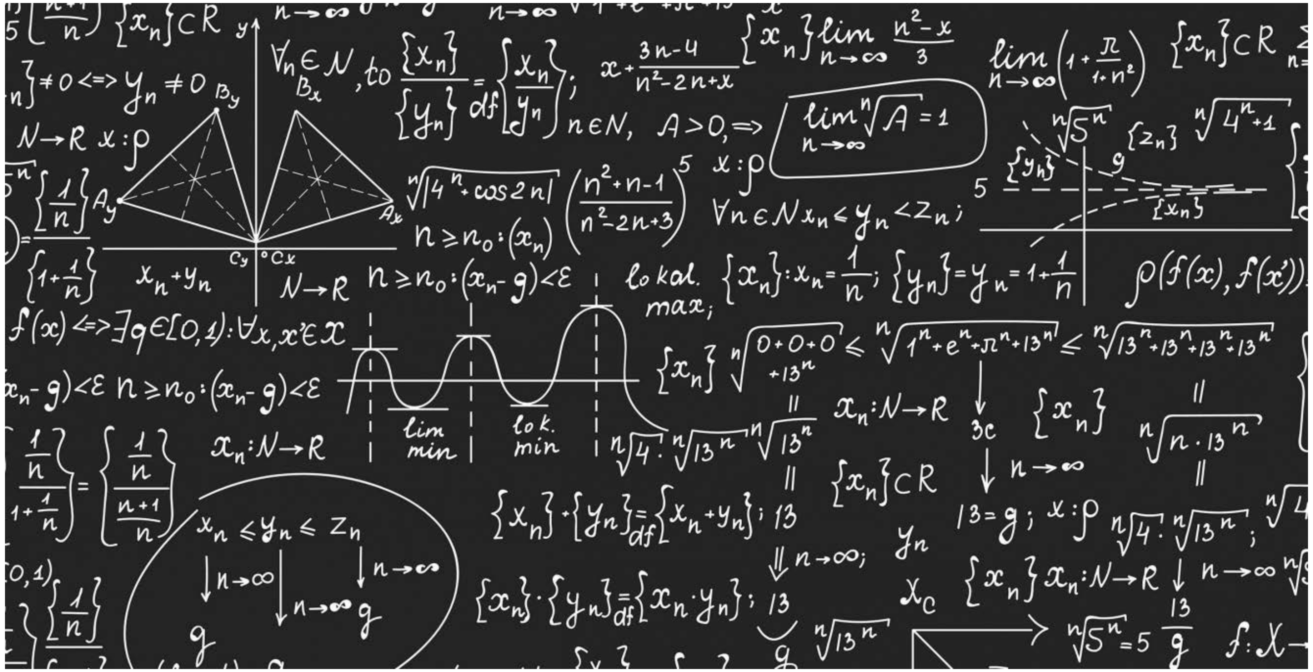
Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency



Skills Needed to Be a Data Scientist

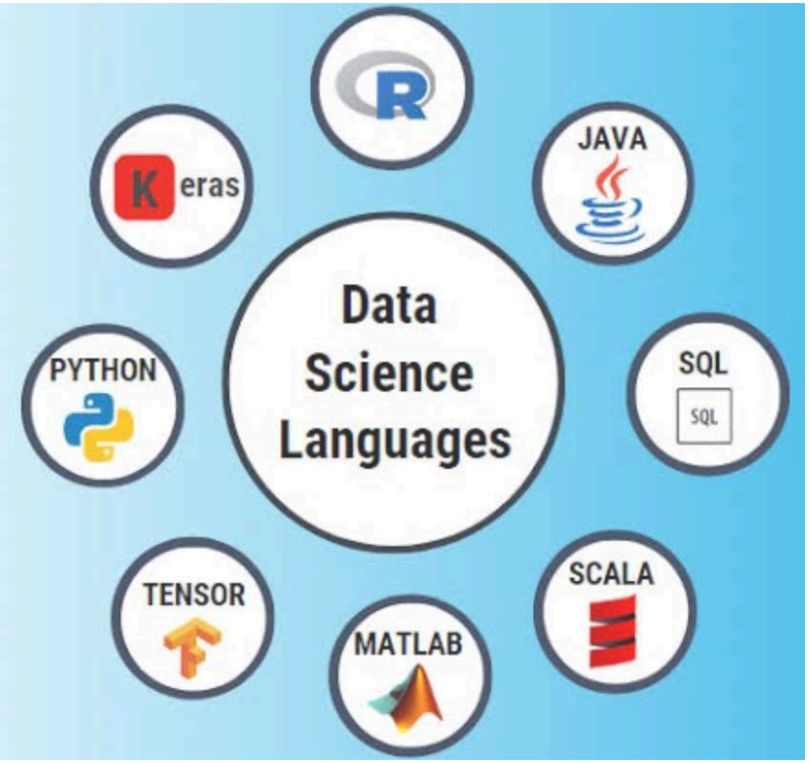
Analytical Skills

Communication Skills

Critical and Logical Thinking Skills

Math Skills

Computer Programming Competency



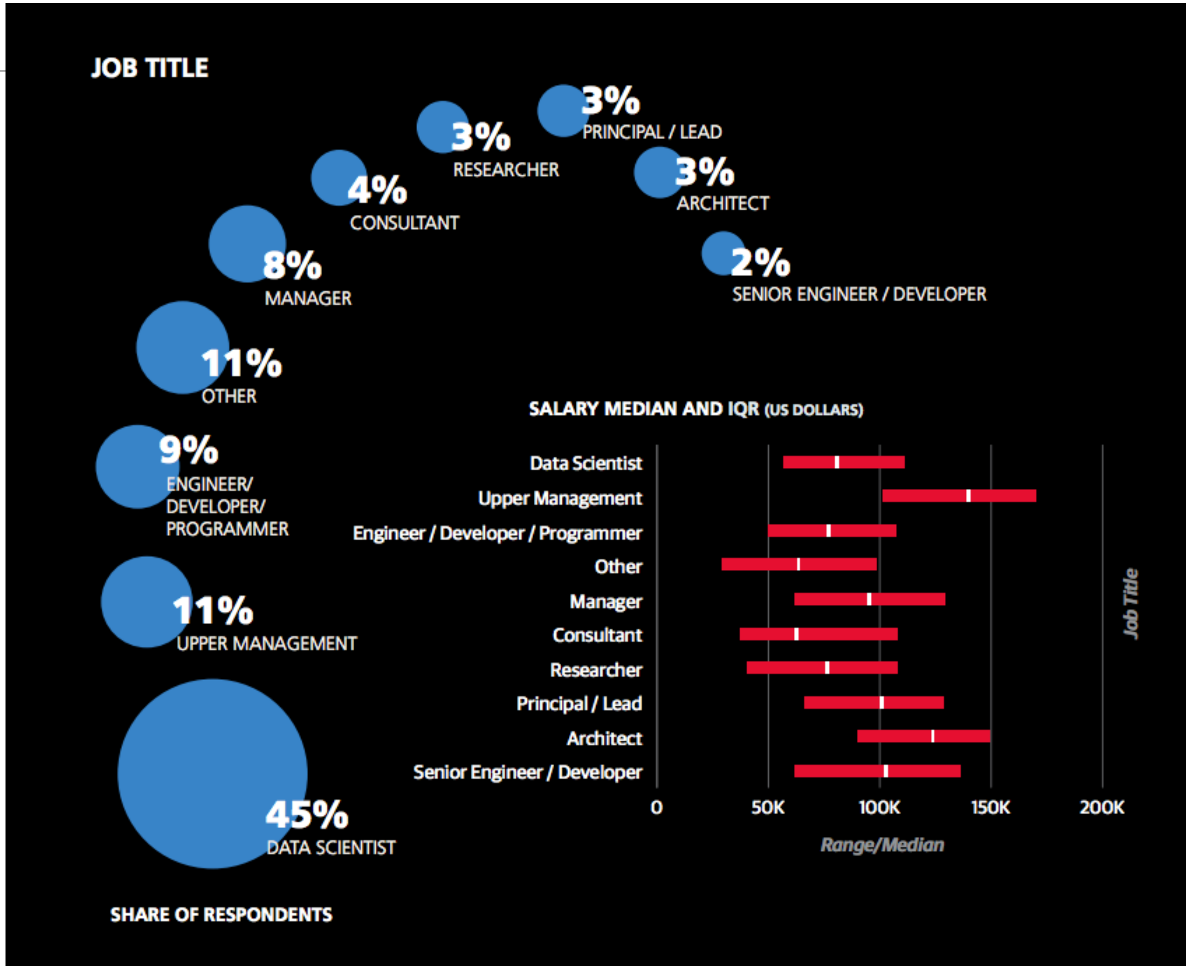
Source: https://www.henryharv[redacted]

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

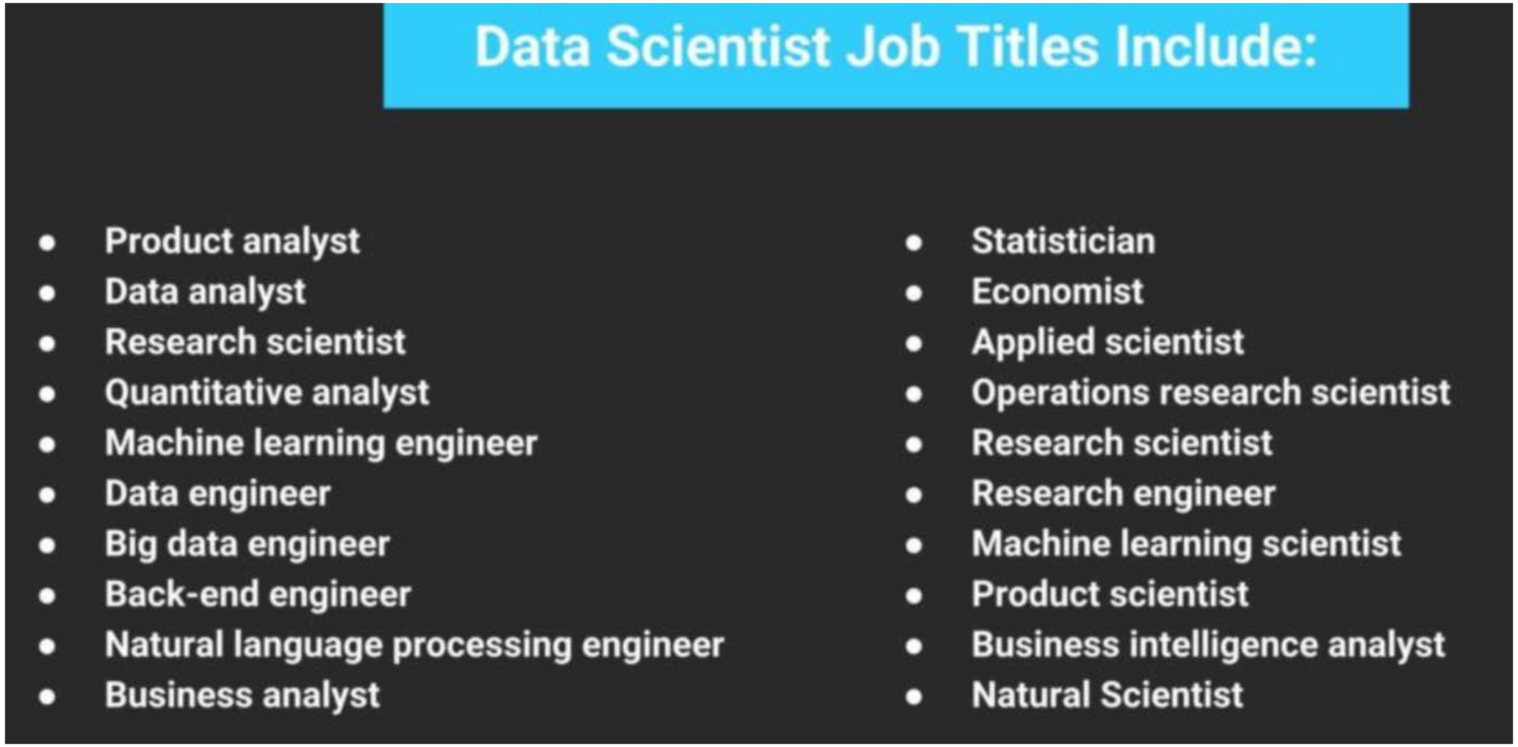
Data analyst

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

Source:https://www.google.com/search?q=d [redacted]
n&sa=X&ved=2ahUKEwif4PfehPLsAhXnAd8
c=zh7H_HOileXupM

Challenges in Data Science Careers

Insights not Used in Decision Making

Data Privacy, Veracity, Unavailability

Limitations of tools to scale/deploy

Wrong Questions Asked