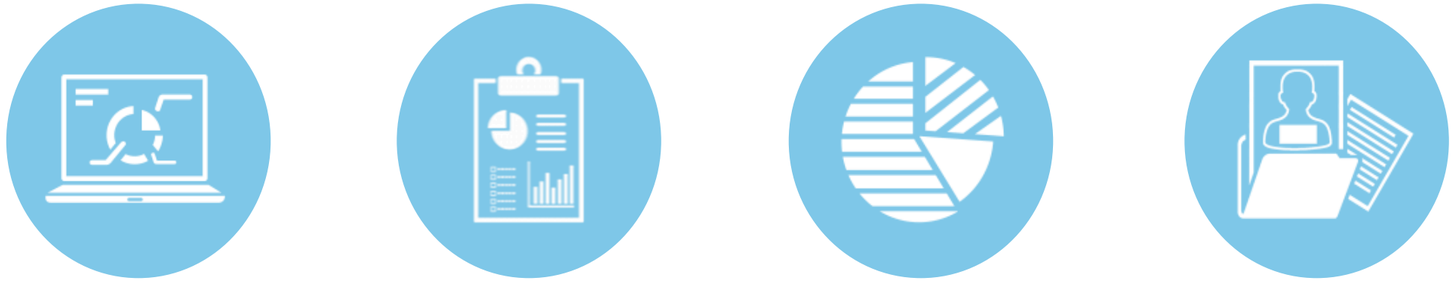
Data Science 101

An Introduction to Data Science



Prepared By: Niteen Kumar, Data Scientist

Objectives

After completing this lesson, you will be able to:

- Explain Data Science
- Discuss what does a Data Scientist do
- Discuss the applications of Data Science
- Understand how Data science and Big data play together
- Explain Data Science as a discipline and how it is shaping the world
- Discuss roles and responsibilities of a data scientist

My Bio

Name: Niteen Kumar, Data Scientist

Education:

- MS in Data Science, CUNY SPS
- Bachelor of Engineering – Electronics and Communication
- MIT – Big Data Specialization

Teaching and Advisory– Major Highlights:

- CUNY SPS – Instructor Led Data Science and Analytics Workshops
- Creator and Advisor of Data Science Course – Simplilearn.com – **8K+** online students
- Advisory and Education at Start Ups:
  - Chainhaus, NYC Start Up– An AI and Blockchain Company
  - Elphi – MIT Tech Start up – AI and Data Company

2003 — App Development

2005 — Data + App + Product Engineering

2010
- Solution Architecture
- Enterprise Architecture
- Data Engineering
- Micro Services

2019
- Data Science
- Advanced Analytics
- Data Services
- Product Development

What is Data Science

Some common definitions of Data Science are as follows:

A powerful new approach to make discoveries from data

An automated way to analyze enormous amounts of data and extract information

Data

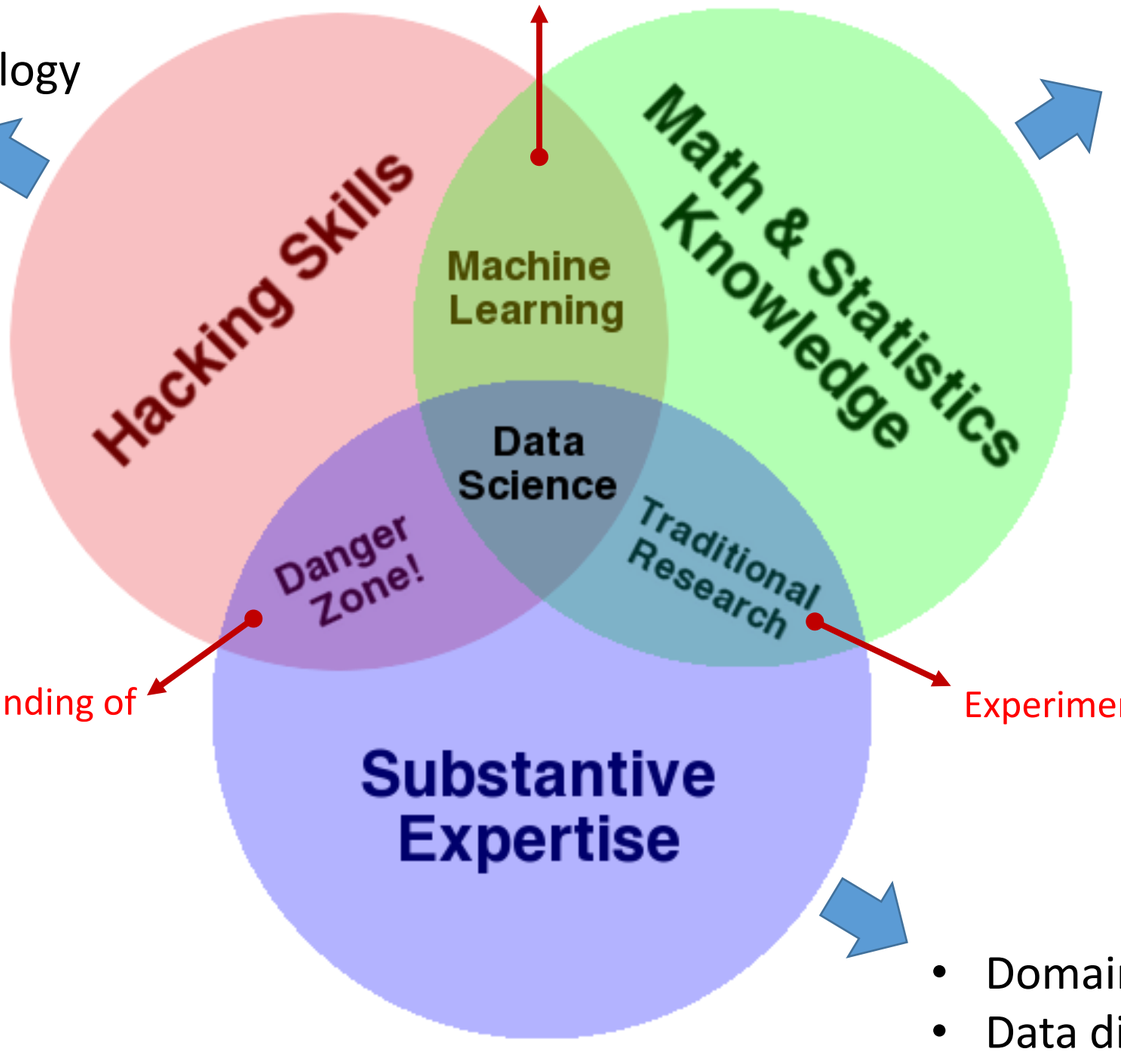
A new discipline that combines aspects of statistics, mathematics, programming, and visualization to turn data into information

What is Data Science

Build and train machine models

- Manipulate data using technology
- Learn and build new tools

Apply math and statistics methods



Data analysis without understanding of math & statistics methods

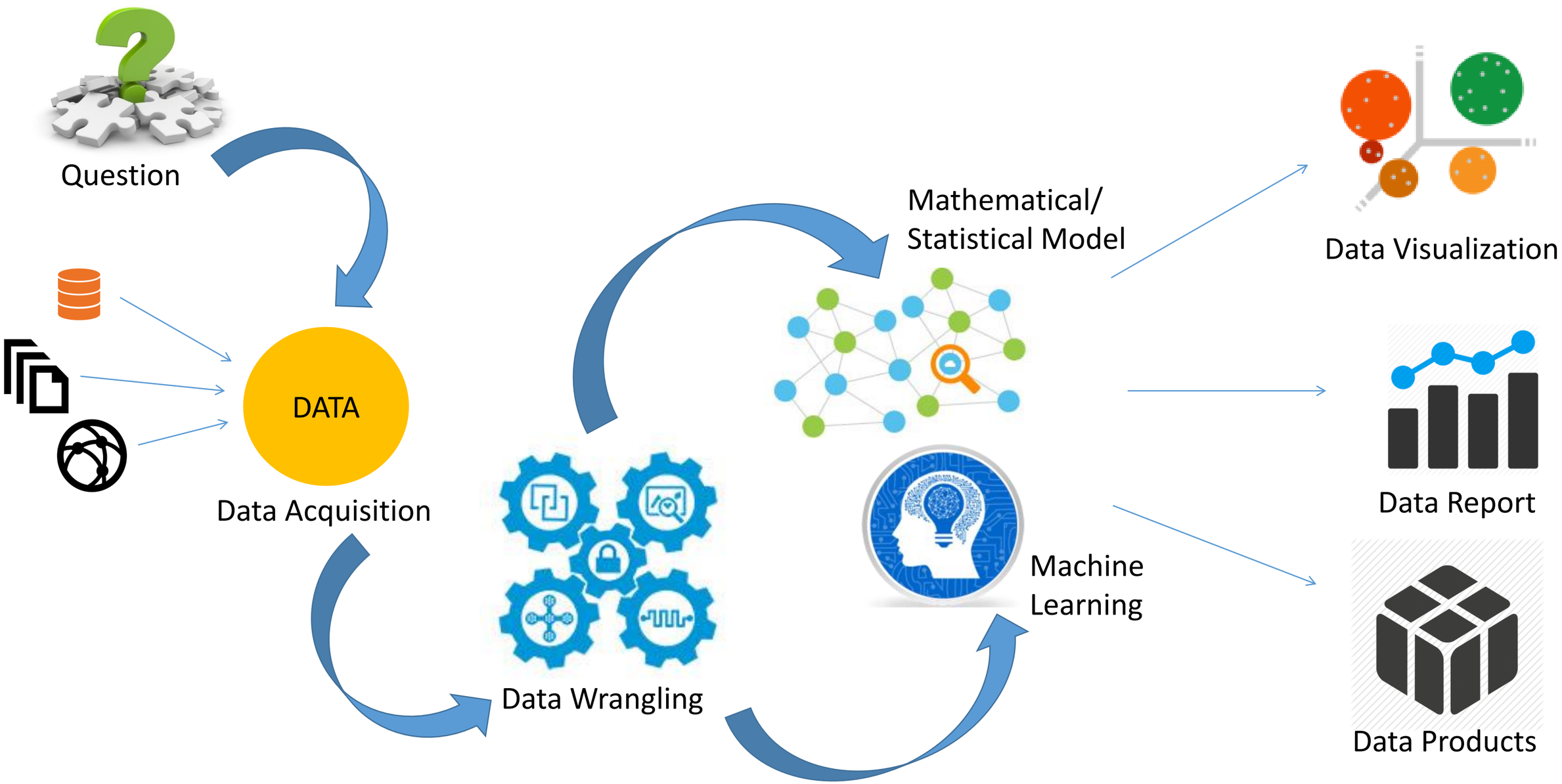
Experiment, hypothesis, control and predict

- Domain expertise and knowledge
- Data discovery

In GOD We Trust, All Others Must Bring Data

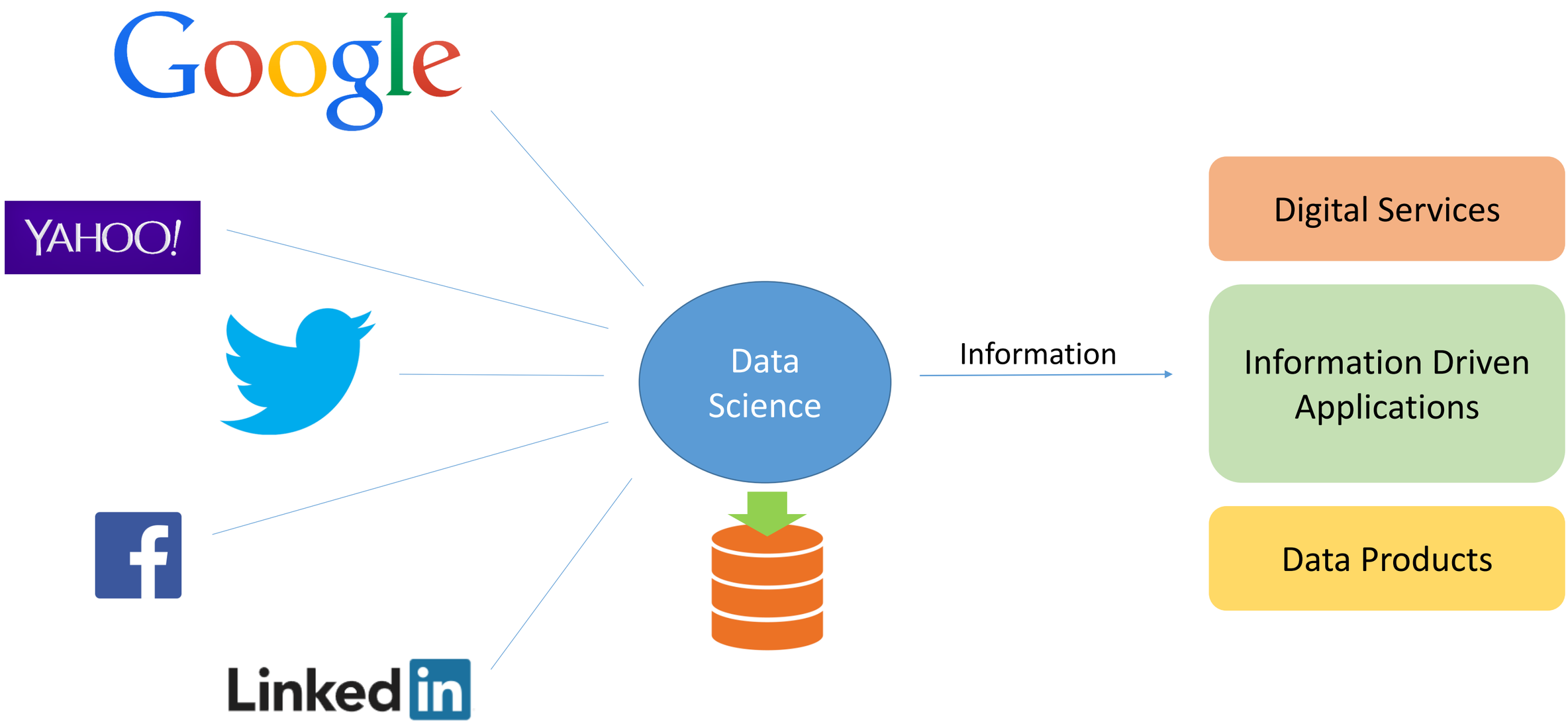
Dr. W. Edwards Deming

A Day in Data Scientist's Life



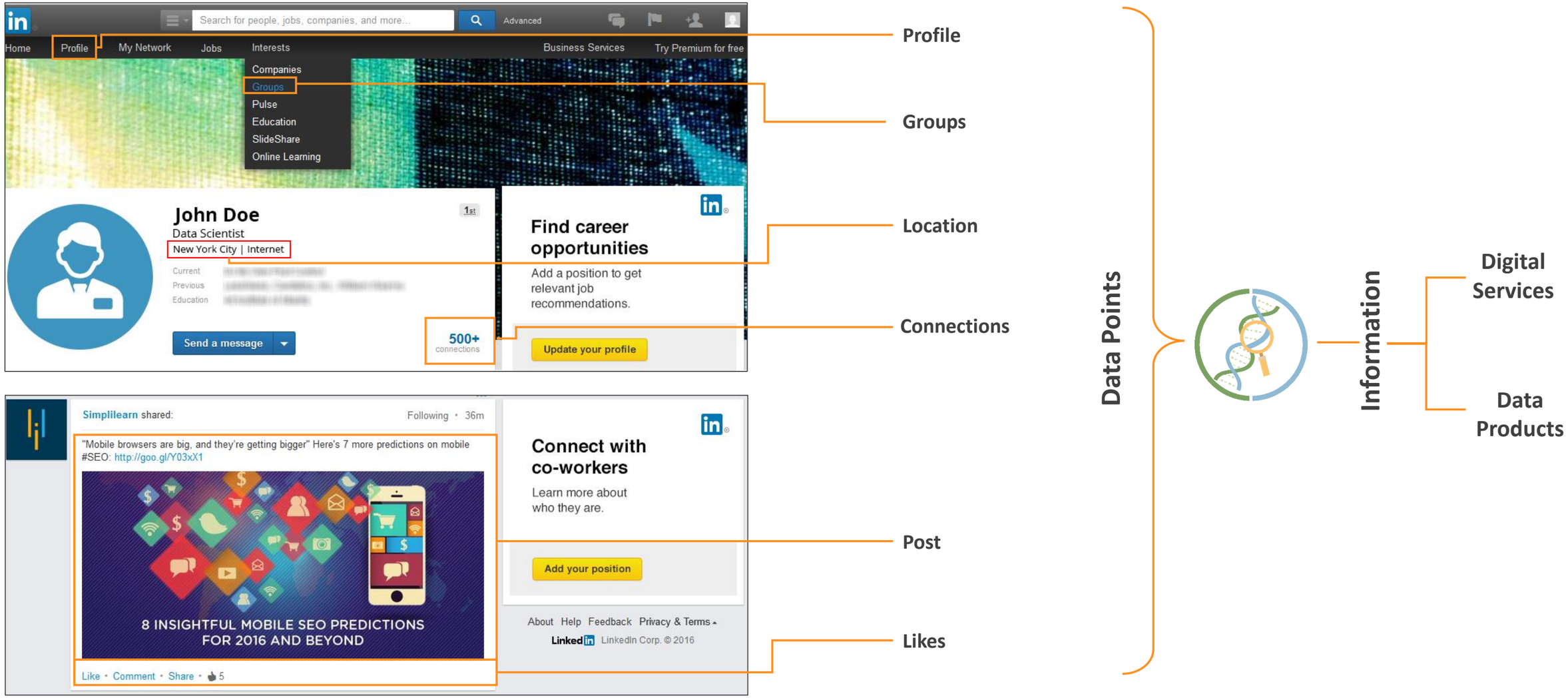
Who is using Data Science and How? Is it a reality already??

Data Science in Private Sector and Start Ups



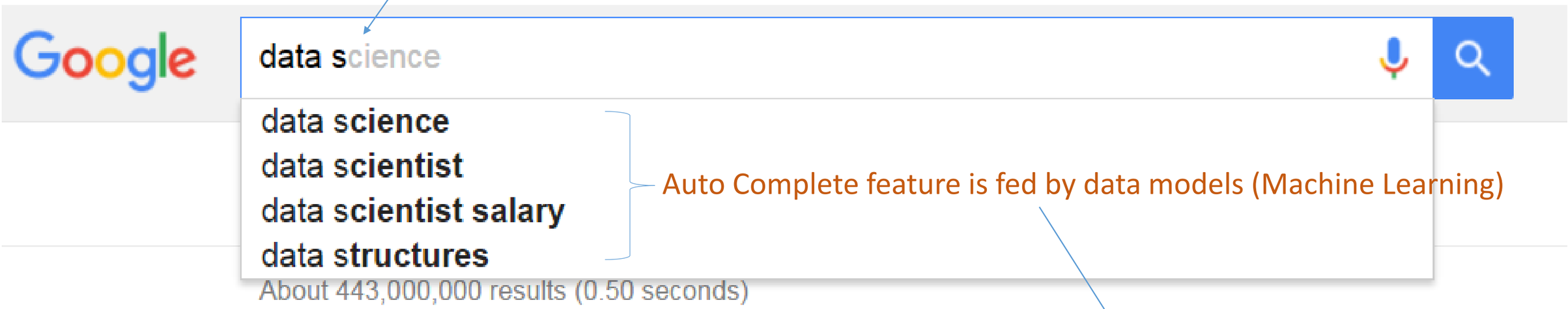
Data Science at LinkedIn

LinkedIn uses data points from its users to provide them with relevant digital services and data products.



Data Science at Google

Search keyword

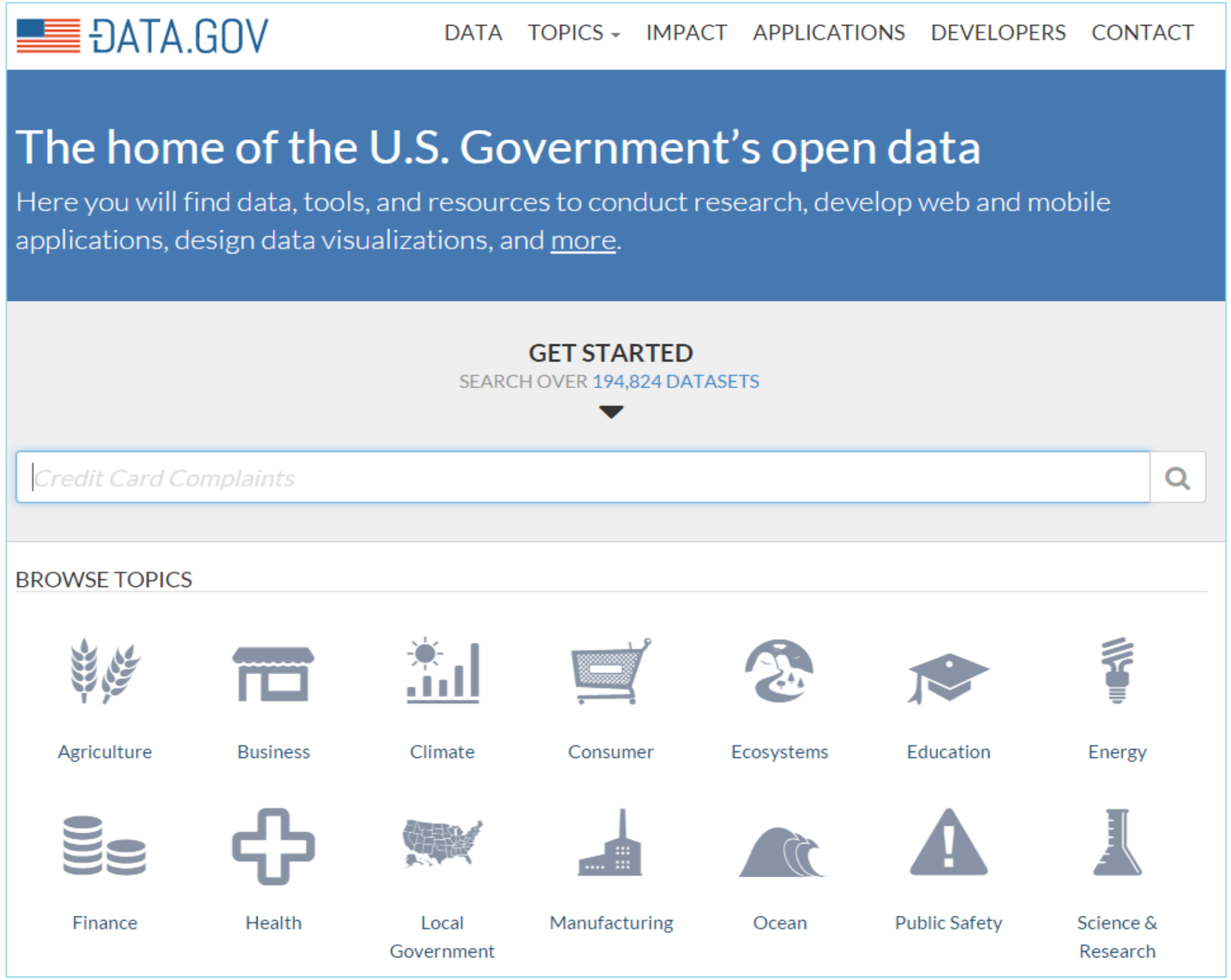


Auto Complete feature is fed by data models (Machine Learning)

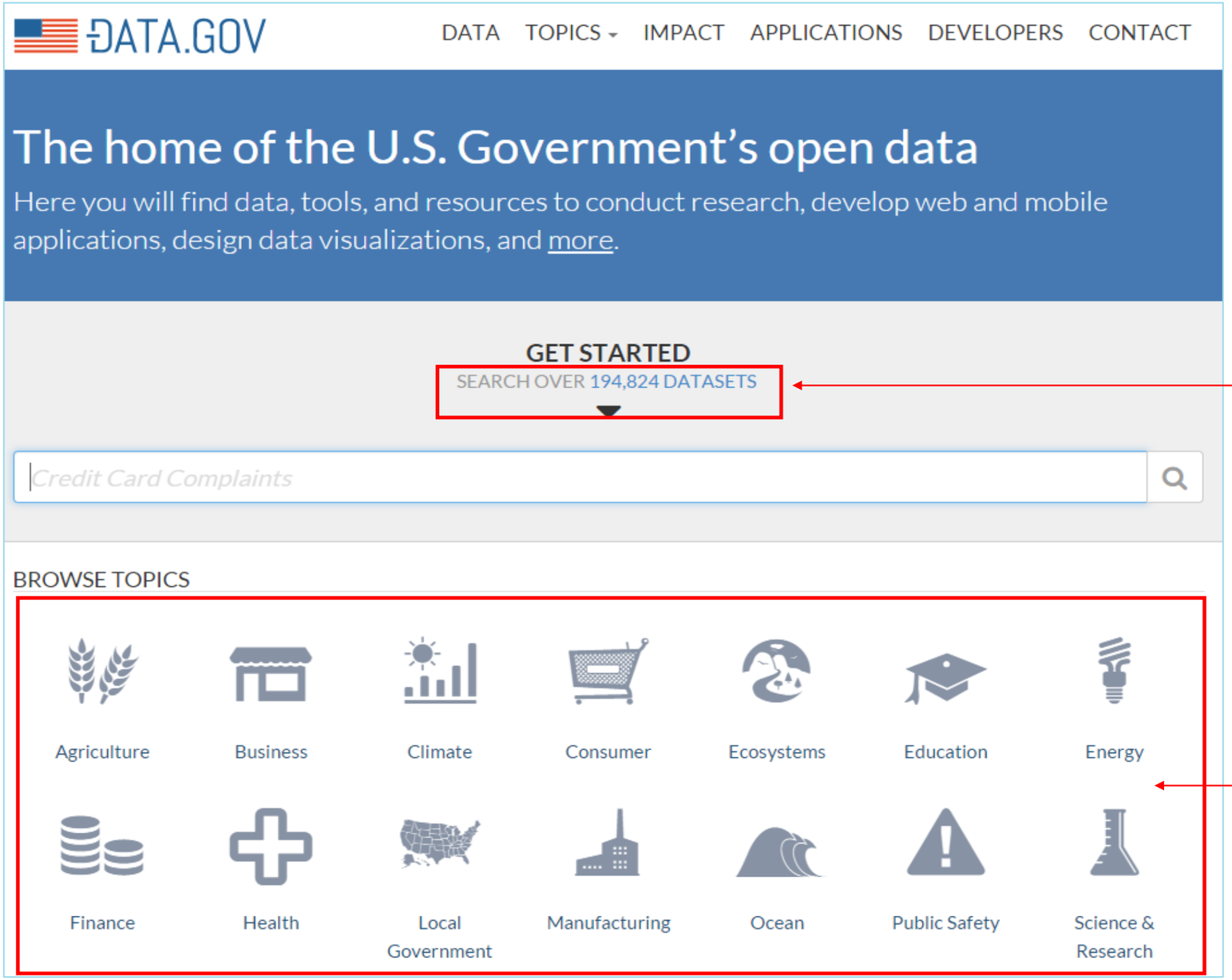
Fast and real time analytics is only made possible by modern and advanced infrastructure, tools and technologies

Influencing Factors
1.Query Volume – Unique and verifiable users 2.Geographical locations 3.Keyword /phrase mentions in the web 4.Some scrubbing for inappropriate content

Data.gov


DATA.GOV
DATA
TOPICS
IMPACT
APPLICATIONS
DEVELOPERS
CONTACT
The home of the U.S. Government's open data
Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.
GET STARTED
SEARCH OVER 194,824 DATASETS
Credit Card Complaints
BROWSE TOPICS
Agriculture
Business
Climate
Consumer
Ecosystems
Education
Energy
Finance
Health
Local Government
Manufacturing
Ocean
Public Safety
Science & Research

Data.gov



Large collection of datasets

Sectors/ Domains

Use Cases from Real World

Some use cases from real world..

Use case # 1 : Fitness and Lifestyle

Wearable device

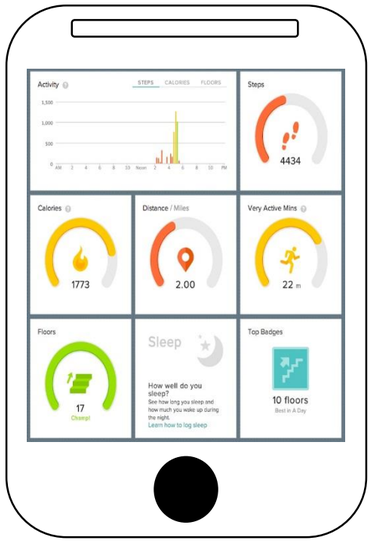
Biometric data transfer

IOT Gateway

Data transfer to Servers

Enterprise Infrastructure

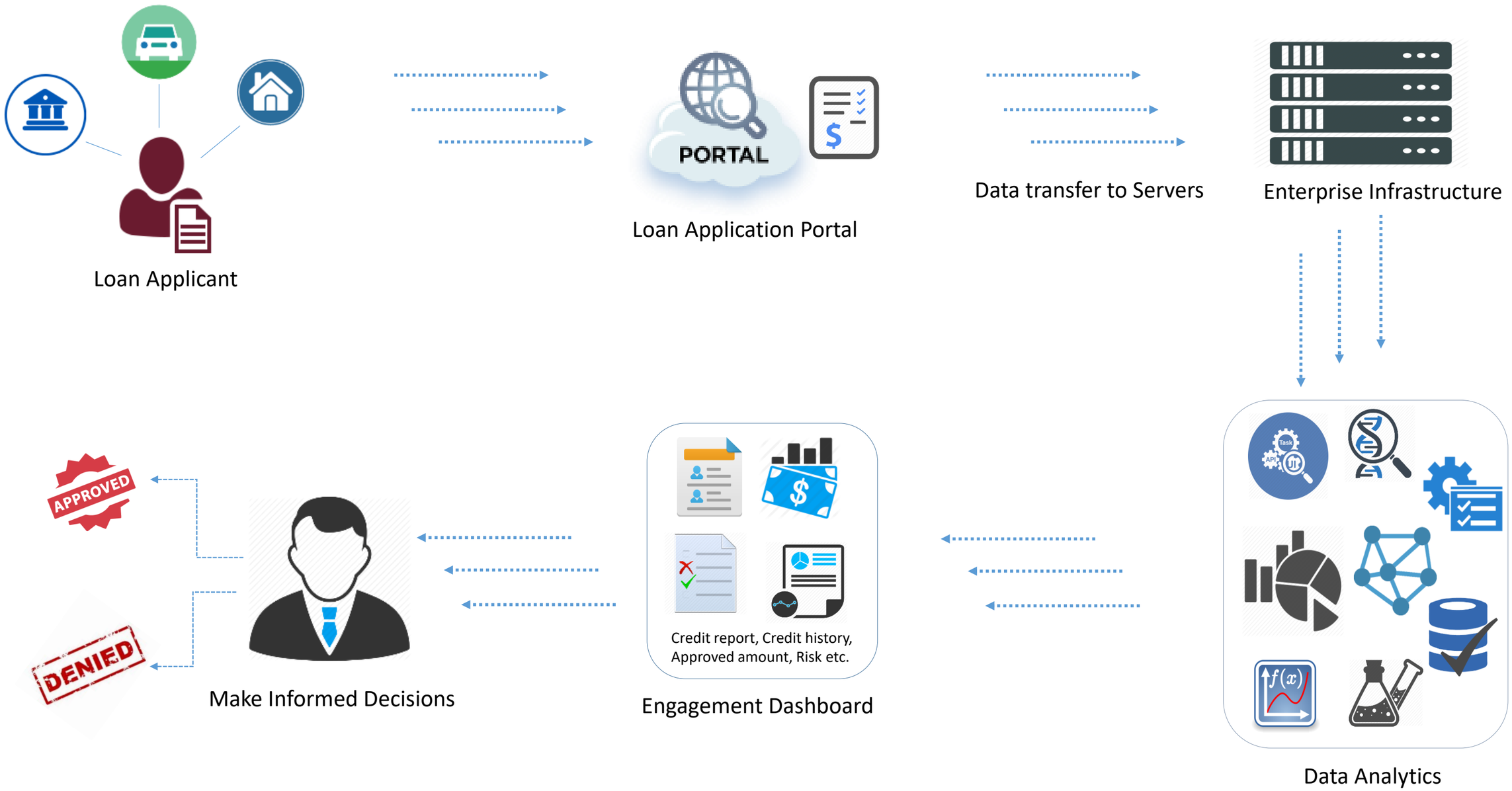
Data Analytics



Engagement Dashboard

Make Informed Decisions

Use case#2: Finance (Loan)



So, How BIG data and Data Science play together?

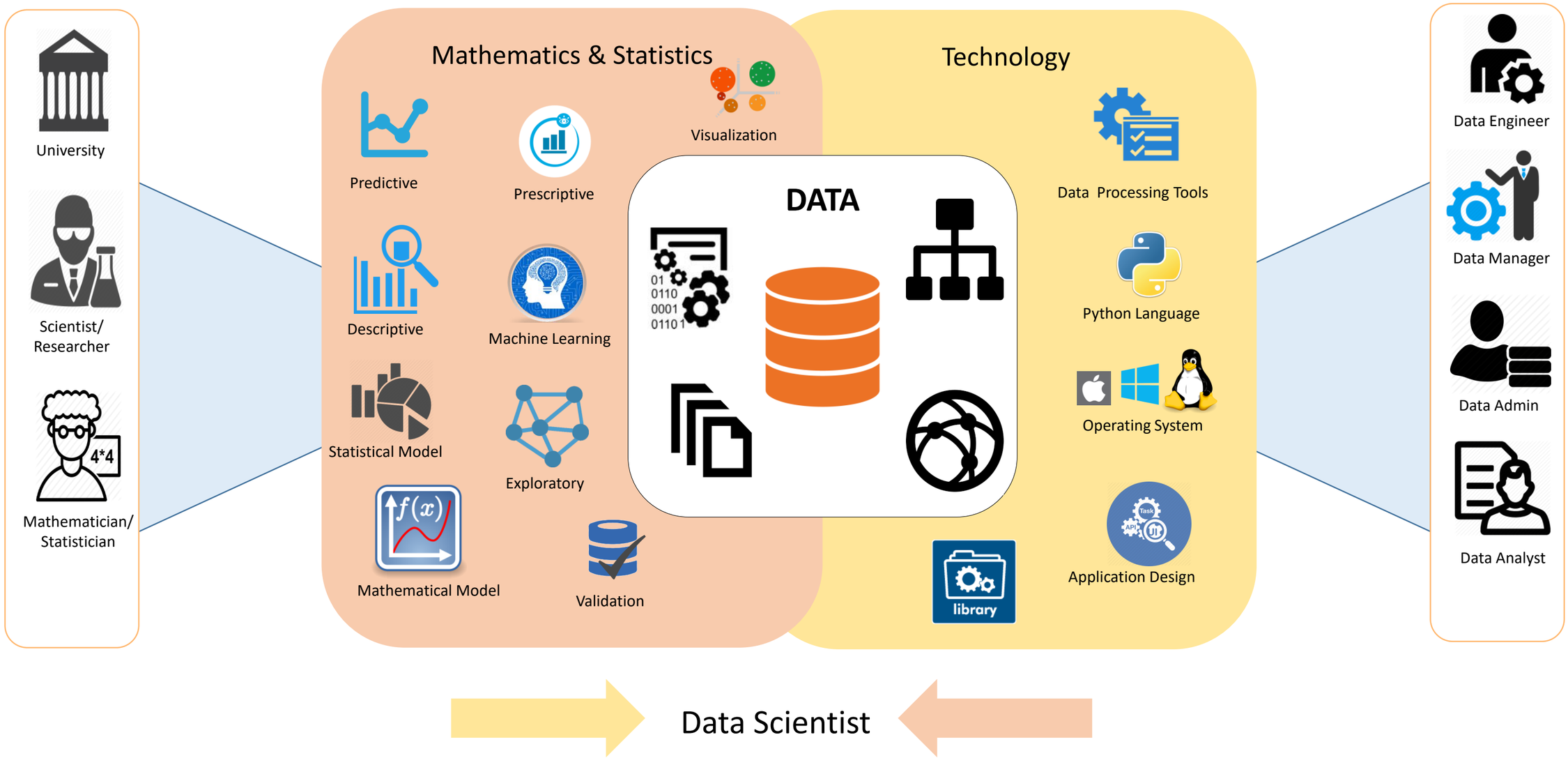
Big Data and Data Science

The Real Challenge

Data scientists face various challenges in real world:

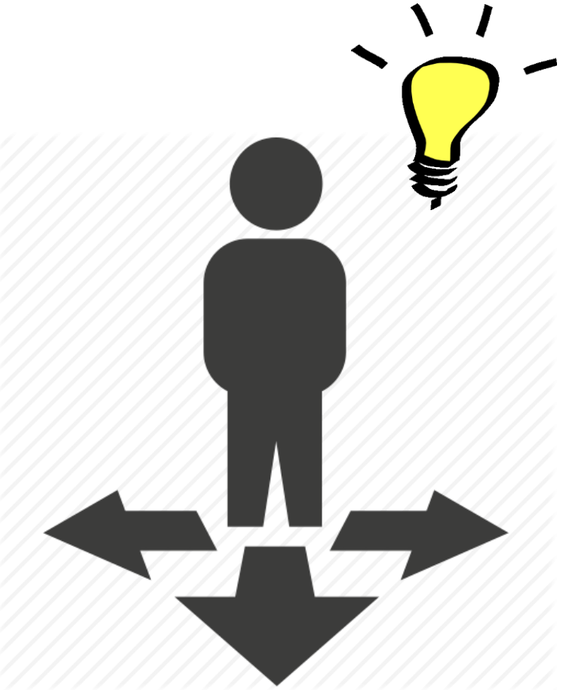
Volume	Data from various sources generating enormous amount of data
Velocity	Data is coming at tremendous speed and from all sort of devices, sensors and applications
Varity	Data in different formats: Structured, Semi Structures and Unstructured
Data Quality	Data are inconsistent, inaccurate, incomplete, not in desirable format, and anomaly present in dataset
Integration	Integrate with enterprise applications and systems
Unified Platform	An unified platform to ingest, process, analyze and visualize large datasets

Big Data and Data Science

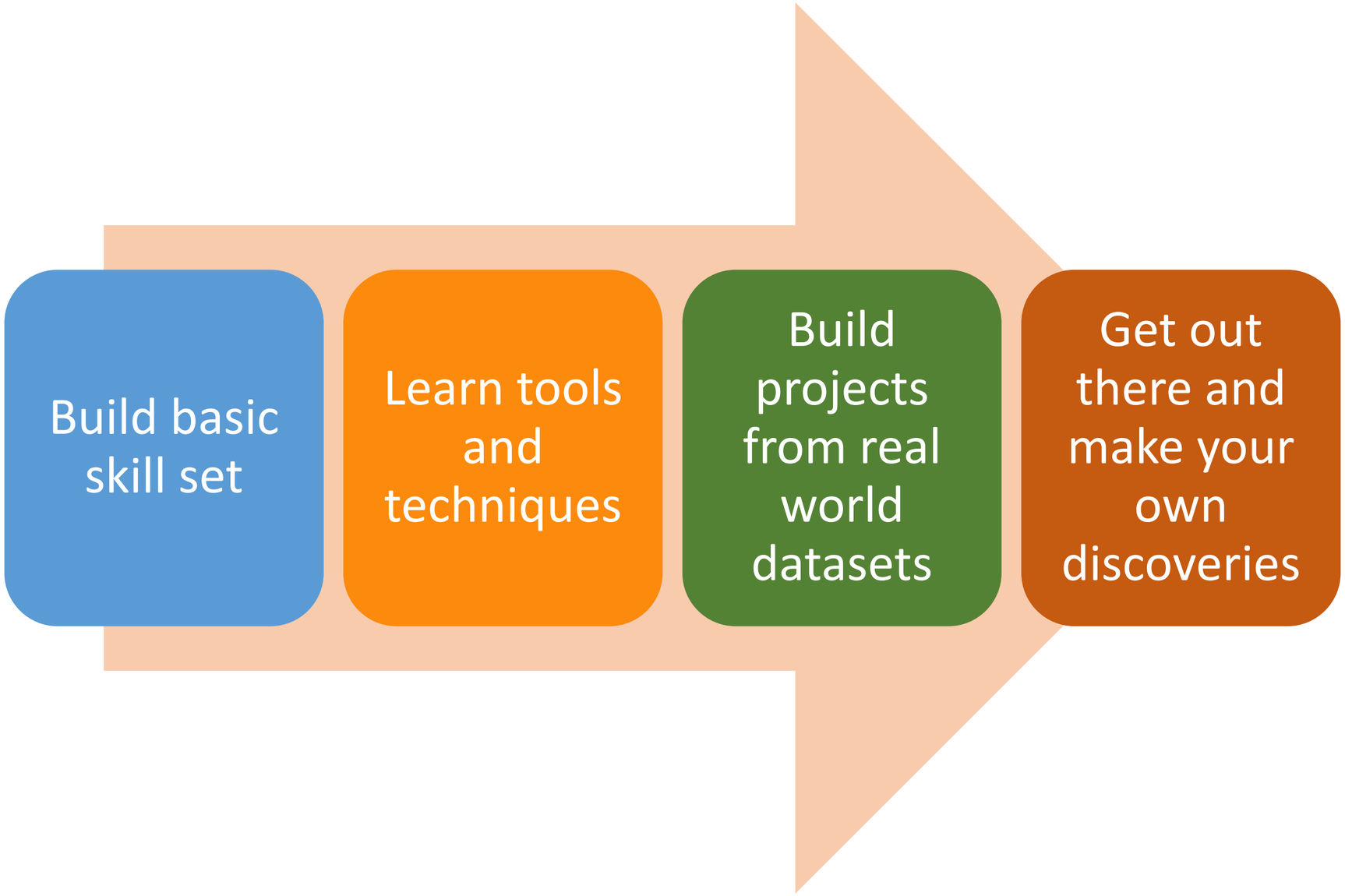


What’s Next..

OK. I got it. Great !!! But where do I go from here…



The Roadmap


Build basic skill set
Learn tools and techniques
Build projects from real world datasets
Get out there and make your own discoveries

Basic skills of a Data Scientist

Ability to ask right questions	Understand the structure of the data	Can interpret and wrangle data
Apply statistical and mathematical methods	Ability to visualize data and communicate	Function with different teams and groups

Summary

Let us summarize the topics covered in this lesson:

- Data Science is a discipline which combines aspects of statistics, mathematics, programming and substantive expertise
- Data scientists are solving bigger problems in public and private sectors
- There are lot datasets available for free to apply data science and turn them into data services and data products
- Data Analysts and Data Scientists are more in demand with the evolution of Big data and fast real time analytics
- Python is a powerful language and a preferred tool for Data scientist

Quiz

Quiz 1	The data science is a discipline which combines aspects of

a. Arts and traditional research

b. Hacking skills and substantive expertise

c. Mathematics & statistics knowledge

d. Traditional research and substantive expertise

The correct answer is **b and c**

Explanation: Data Science is a discipline which combines aspects of statistics, mathematics, programming and substantive expertise

Quiz 2	Danger zone, in Venn diagram, can do major damage to business because in this space data analysis:

a. Is done using hacking skills and substantive expertise only

b. Is done without mathematics and statistics knowledge

c. Is done only with substantive expertise

d. Is result of machine learning and traditional research only

The correct answer is **b**

Explanation: **Only** domain knowledge and technology are used to analyze data and without any knowledge of math & statistics methods. Hence it is inaccurate and incomplete and can damage business.

Quiz 3	A data scientist does the followings:

a. Ask the right question or business problem

b. data acquisition

c. Data wrangling and data visualization

d. All of the above

The correct answer is **d**

Explanation: A data scientist asks right questions to stakeholders, acquires data from various source and data points, performs data wrangling which is making it ready for data analysis and create reports and plots for data visualization

Quiz 4	Data science helps to create

a. Data products

b. Data services

c. Information driven application

d. All of the above

The correct answer is **d**

Explanation: Data scientists apply data science techniques to extract information from raw data and create: data products, data services and information driven applications

Quiz 5	Data science can be used in

a. Public sectors only

b. Private sectors only

c. Start ups only

d. All of the above

The correct answer is **d**

Explanation: There are more than 195K datasets (and growing)available to all for free. Data science can be used in public sectors, private sectors, start ups and everywhere else.

Quiz 6	The Big data includes _____.

a. Large volume of data

b. Volume and variety of data

c. Volume and velocity of data

d. Volume, variety and velocity of data

The correct answer is **d**

Explanation: Data becomes Big Data when its volume, velocity, or variety exceeds the capacity of the deployed IT systems to store, analyze, and process it.

Thank You