Introduction to Data Science

Tools & Techniques for Data Science

Murk Marvi

Outline

- Course content
- Reference Books
- Course Assessment
- Data Science
- Business Value
- Needed Skills
- Contrast
- Tools

Course Content

- Introduction to Data Science
- Data Science Life cycle & Process for Building Data Products
- Introduction to Data
- Data pre-processing Stages
  - Aggregation, Sampling, Dimensionality Reduction, etc.
- Algebraic & Probabilistic View of Data
- Introduction to Python Libraries
- Relational Algebra & SQL

Course Content

- Scraping & Data Wrangling
- Basic Descriptive & Exploratory Data Analysis
- Introduction to Text Analysis
  - Stemming, Lemmatization, Bag of Words, TF-IDF
- Introduction to Prediction and Inference algorithms
  - Supervised, Unsupervised
- Bias-Variance Tradeoff
- Model Evaluation & Performance Metrics
  - Accuracy, Contingency Matrix, Precision-Recall,
  - F1 Score, etc

Reference Books

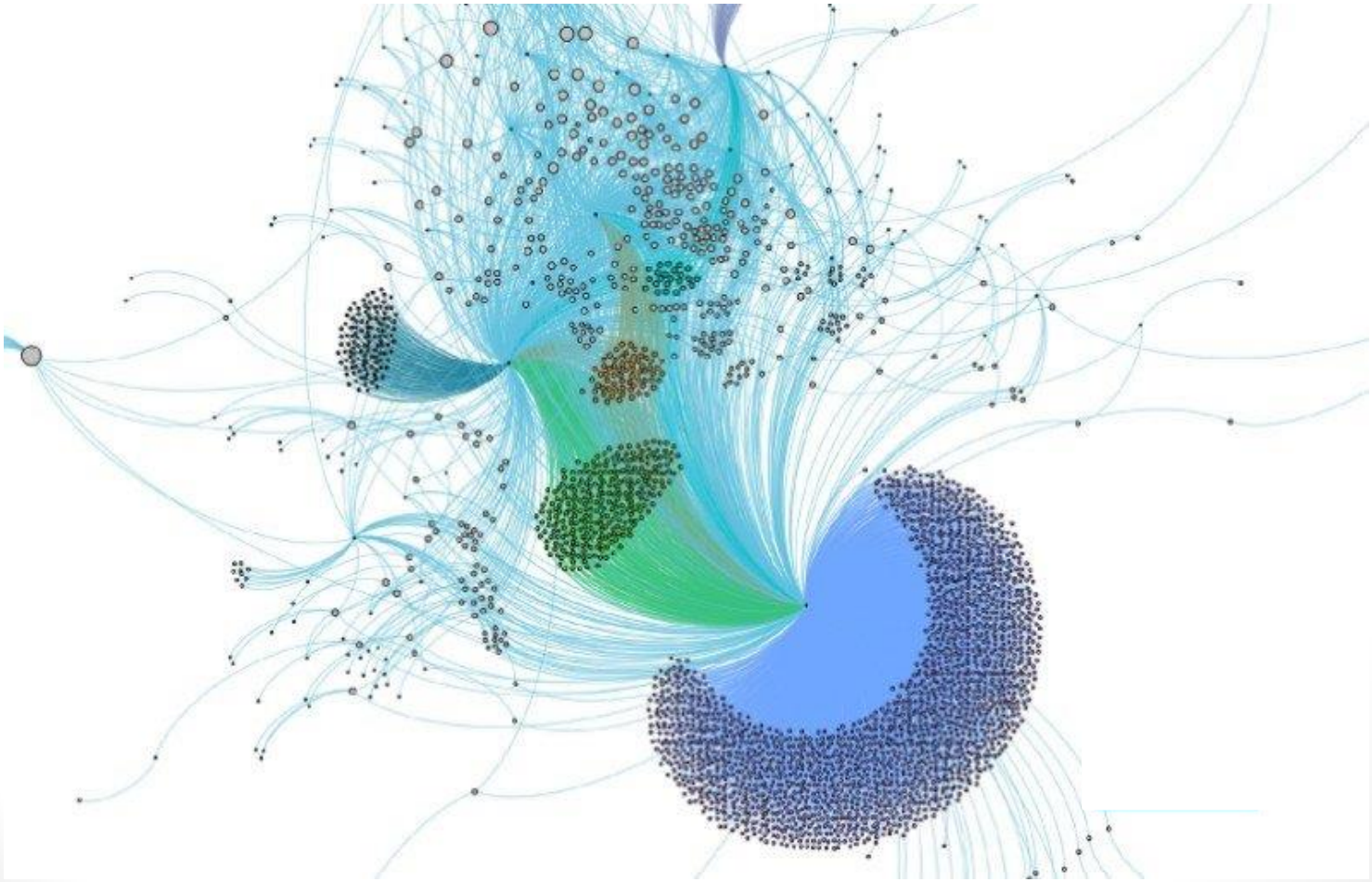
- Python for Data Analysis, 1st Edition, William McKinney
- An Introduction to Statistical Learning with Applications in R, 1st Edition, G. James, D. Witten, T. Hastie and R. Tibshirani
- Computational and Inferential Thinking: The Foundations of Data Science, 1st Edition, A. Adhikari and J. DeNero
- Data Mining and Analysis: Fundamental Concepts and Algorithms, 1st Edition, M. Zaki & W. Meira,
- Doing Data Science, 1st Edition, Cathy O'Neil and Rachel Schutt
- Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications, 1st Edition, Laura Igual.

Course Assessment

- Mid Exam – 15%
- Class performance + Assignment – 10%
- Project + Report + Presentation – 15%
- Final – 60%

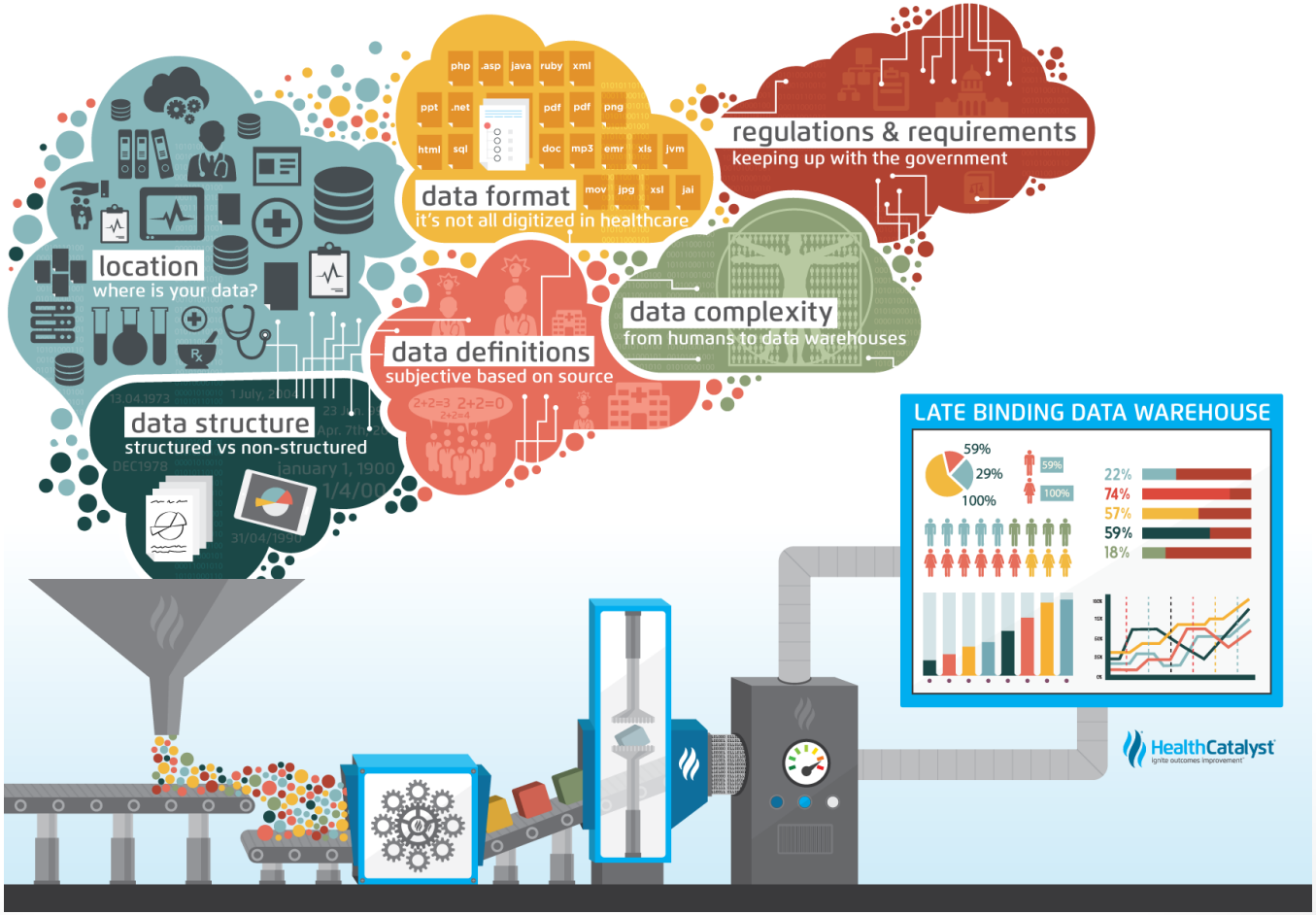
Data

"Data is the oil of the 21st century, and analytics is the combustion engine." – Peter Sondergaard, SVP, Garner Research



Data

"Hiding within those mounds of data is knowledge that could change the life of a patient or change the world."
– Atul Butte



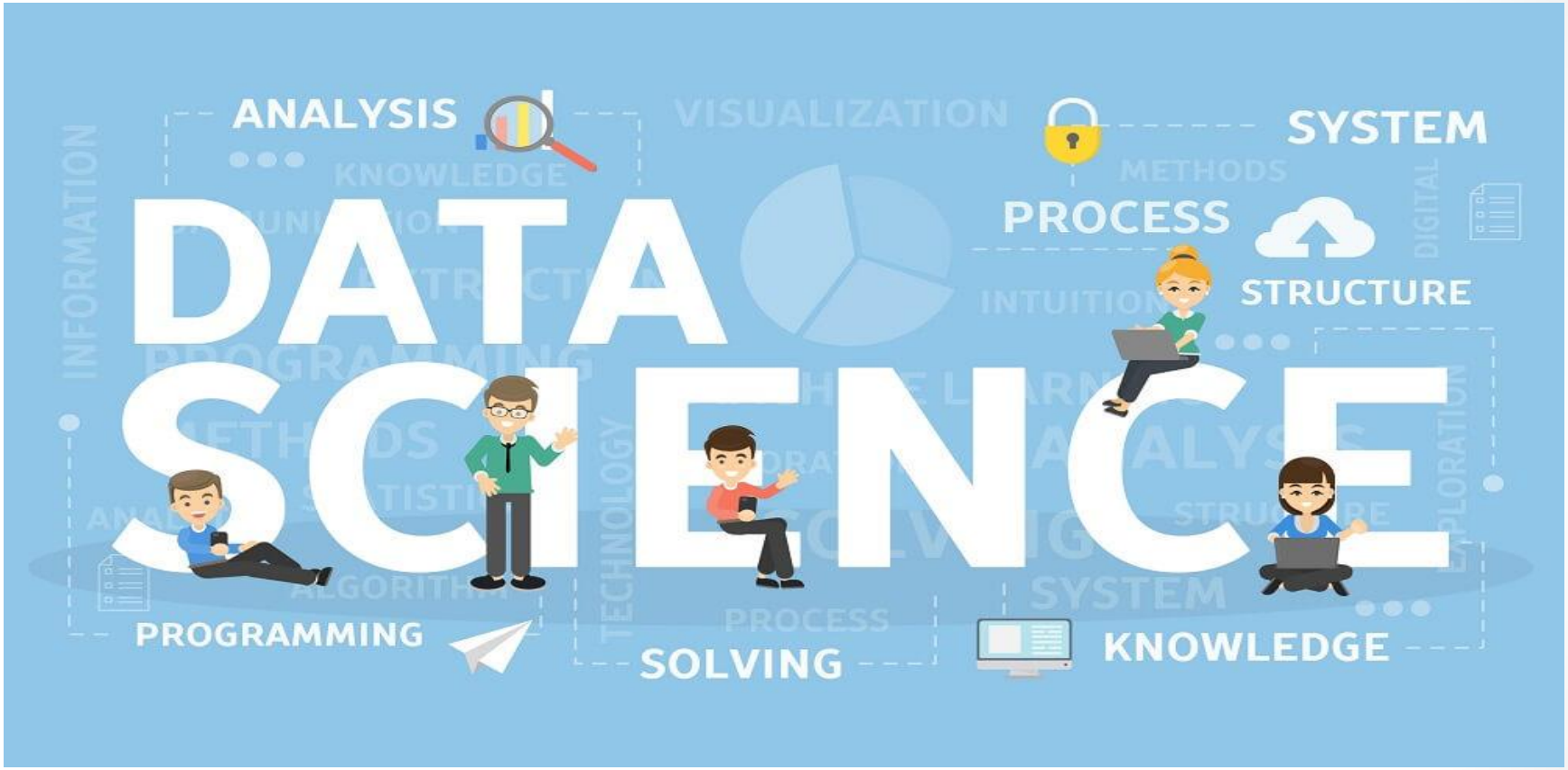
Data

"Data is all around us, and we humans exploit this data to gain experience and make intelligent decisions. Then, why not machines!"

A glass falling from a certain height, what is the probability that it will break?

Data

A coin flipped, what is the probability that a head will occur?


ANALYSIS
VISUALIZATION
SYSTEM
PROCESS
STRUCTURE
DATA
SCIENCE
PROGRAMMING
SOLVING
KNOWLEDGE

What Exactly the Data Science is?

- It is about extracting useful insights from the data in order to add business value or to solve complex problems.

What Exactly the Data Science is?

- It is about uncovering hidden information that may be useful to help companies make smarter choices for their business.

What Exactly the Data Science is?

- The end result of applying data science on a problem is a "**data product**".

Data Product

- The recommendation engine that Amazon uses suggests new items to its users, which is determined by their algorithms. Spotify recommends new music. Netflix recommends new movies.

amazon

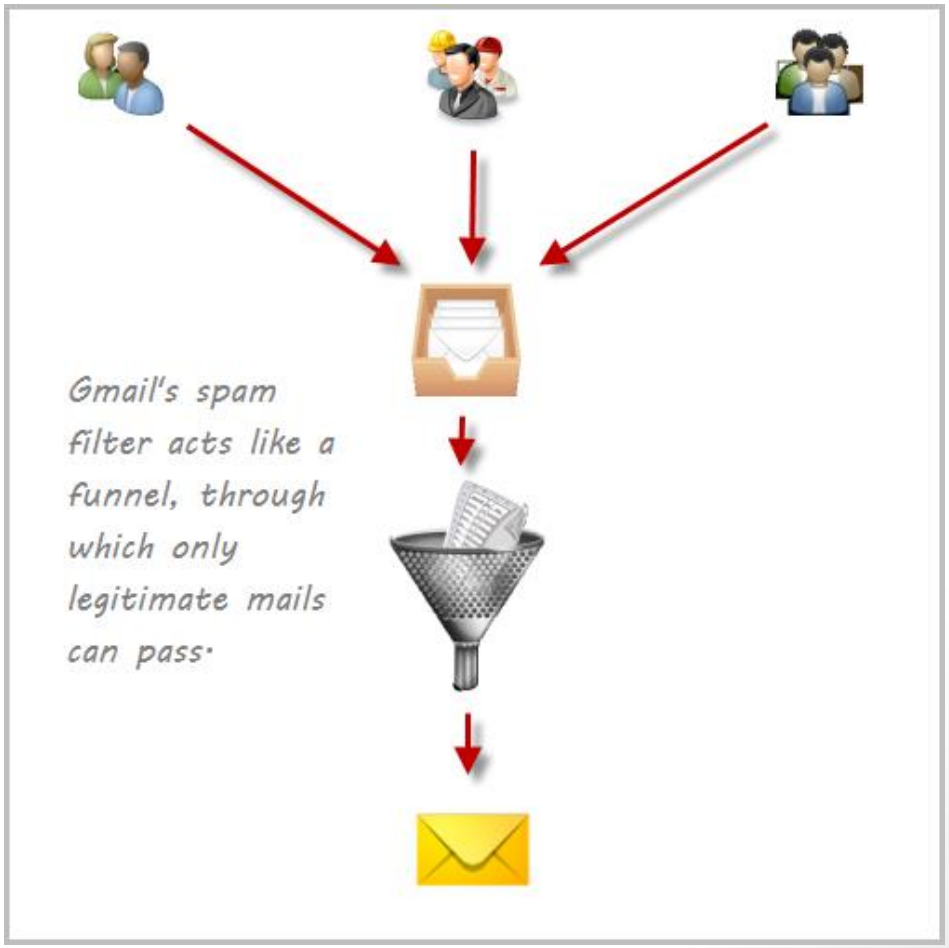
35% OF AMAZON'S REVENUE ARE GENERATED BY IT'S RECOMMENDATION ENGINE.

NETFLIX

75% OF USERS SELECT MOVIES BASED ON NETFLIX'S RECOMMENDATIONS.

Data Product

- The spam filter in Gmail is a data product. This is a behind the scenes algorithm that processes the incoming mail and decides whether or not it is junk.

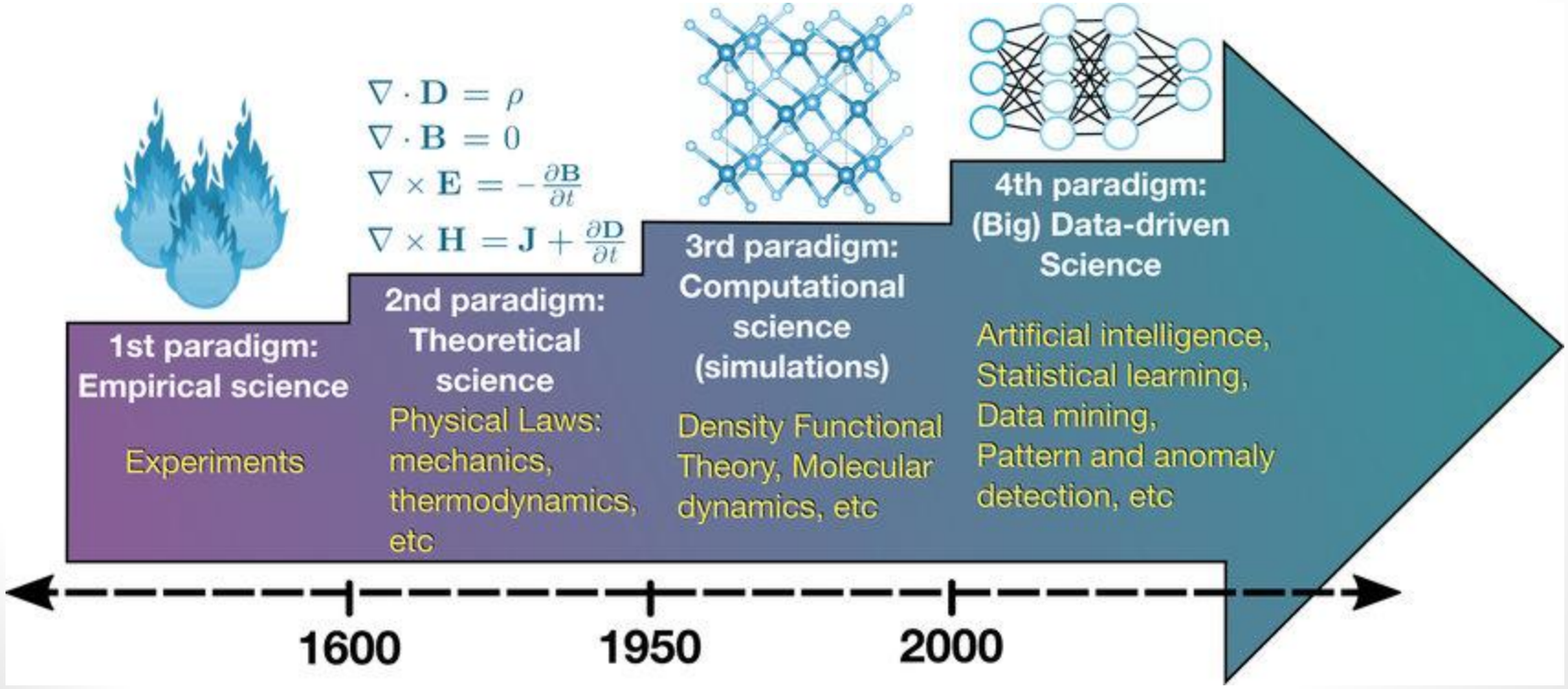


Data Product

- The computer vision that is used for self-driving cars is also a data product. Machine learning algorithms can recognize pedestrians, traffic lights, other cars, and so on.

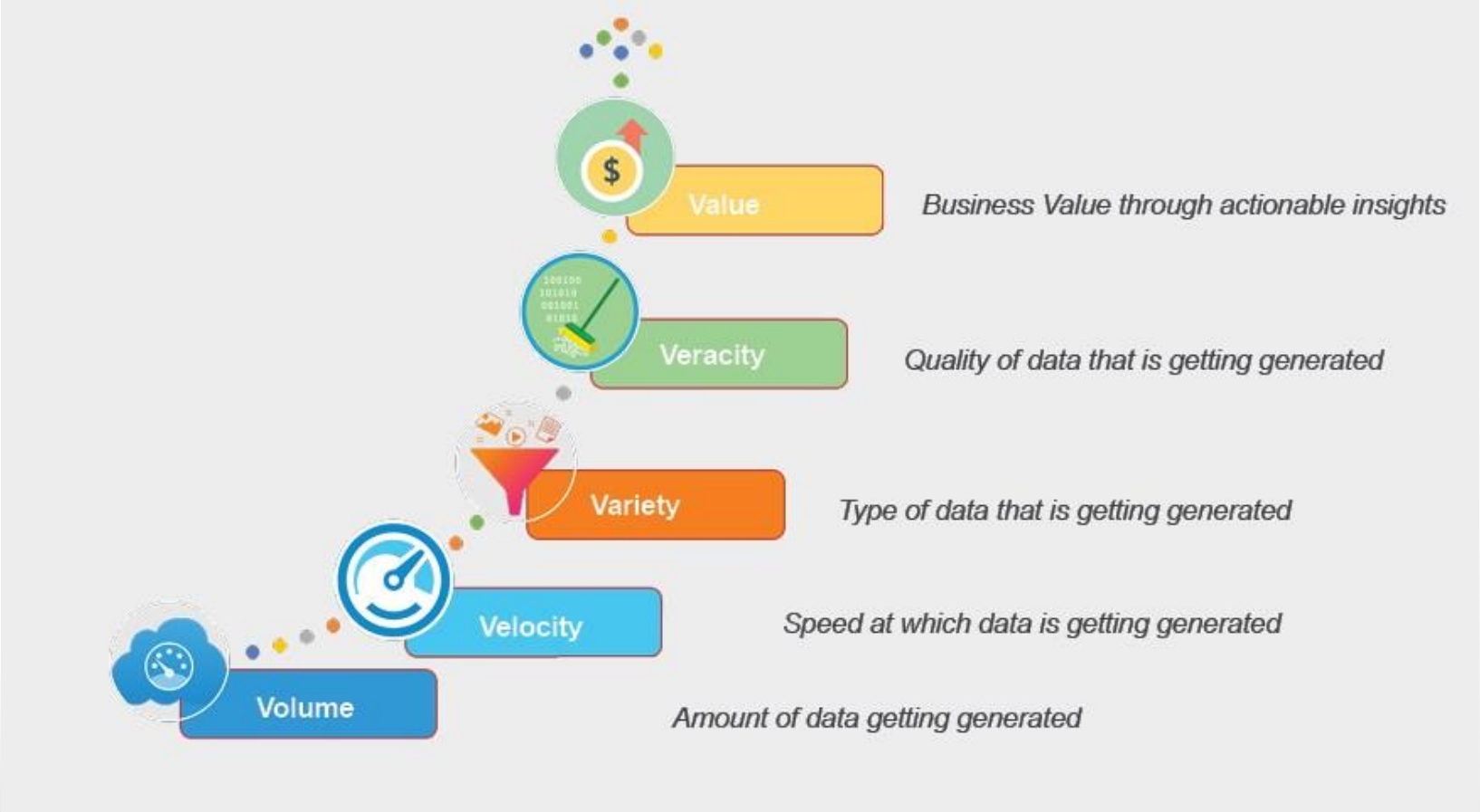
Data Science

- "Jim Gray, a Turing award winner, saw data science as a "**fourth paradigm**" of science: **computational, theoretical, empirical, and driven by data**.

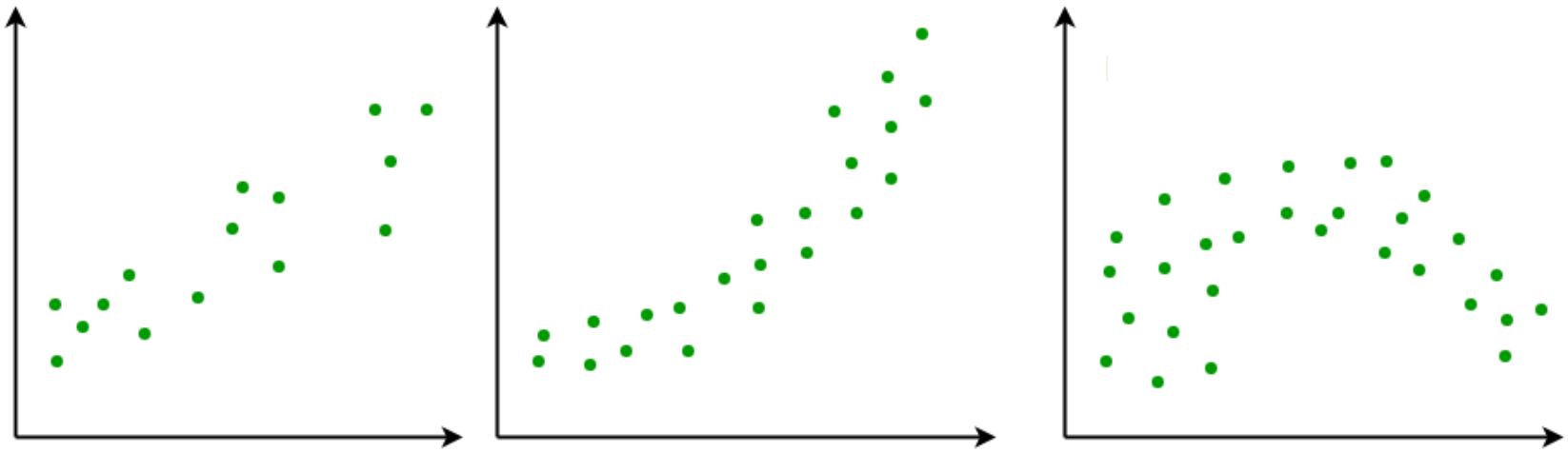


Data Science

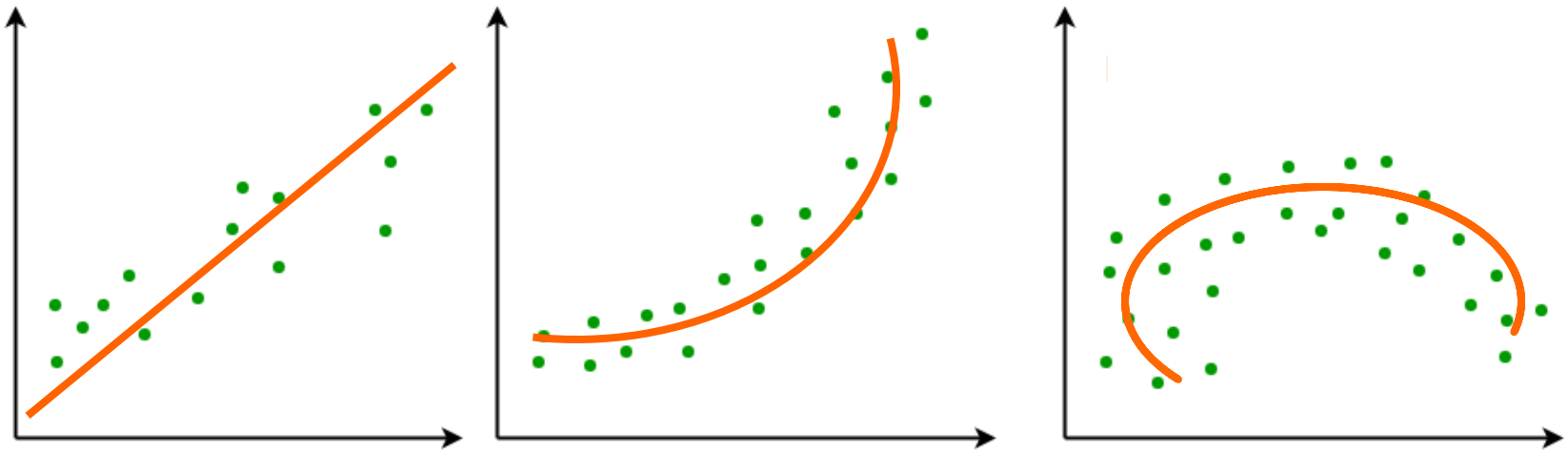
- He also asserted that all parts of science are changing due to the impact of data deluge and Information technology.



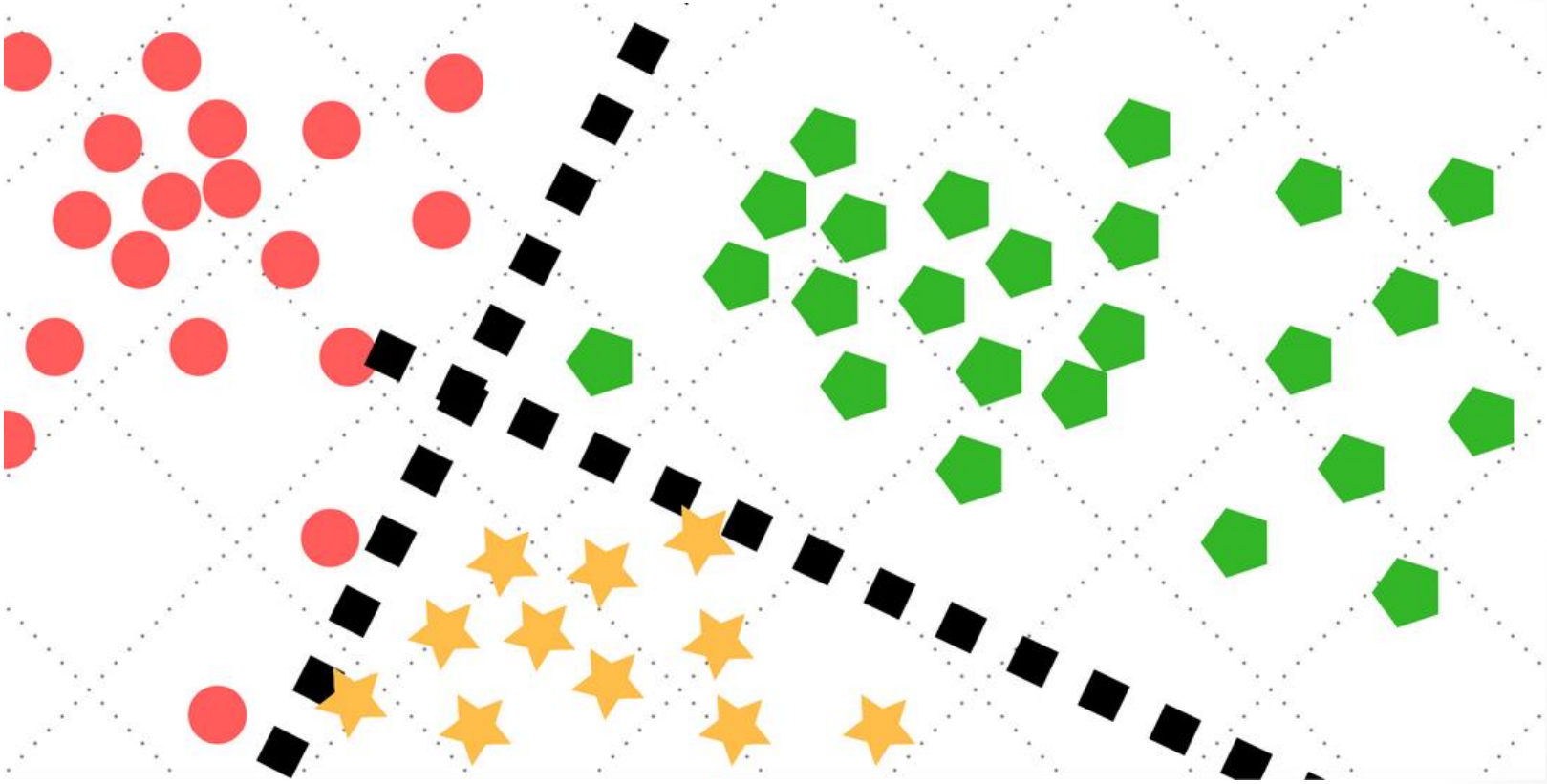
Interpret the data?



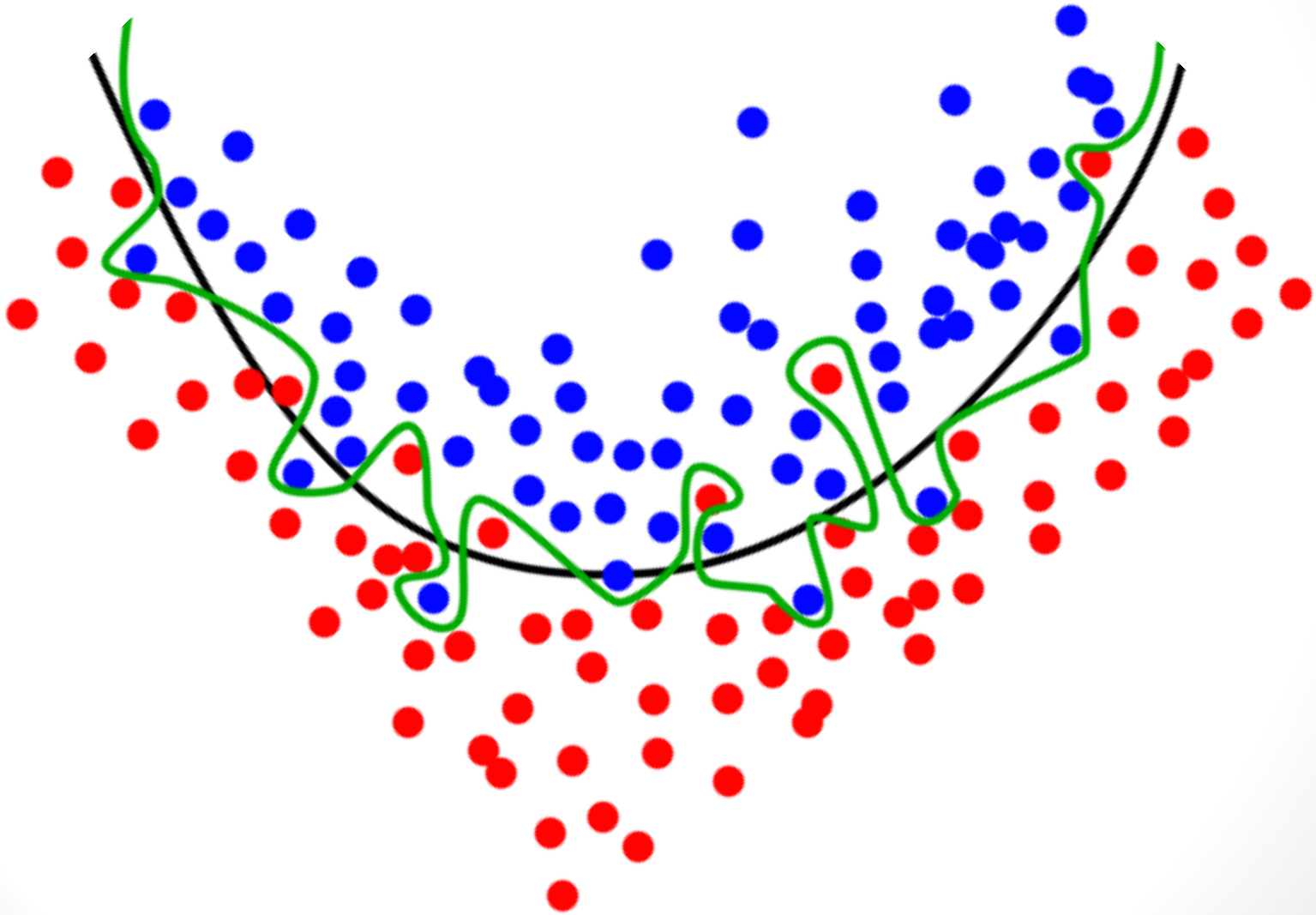
Interpret the data?



Interpret the data?



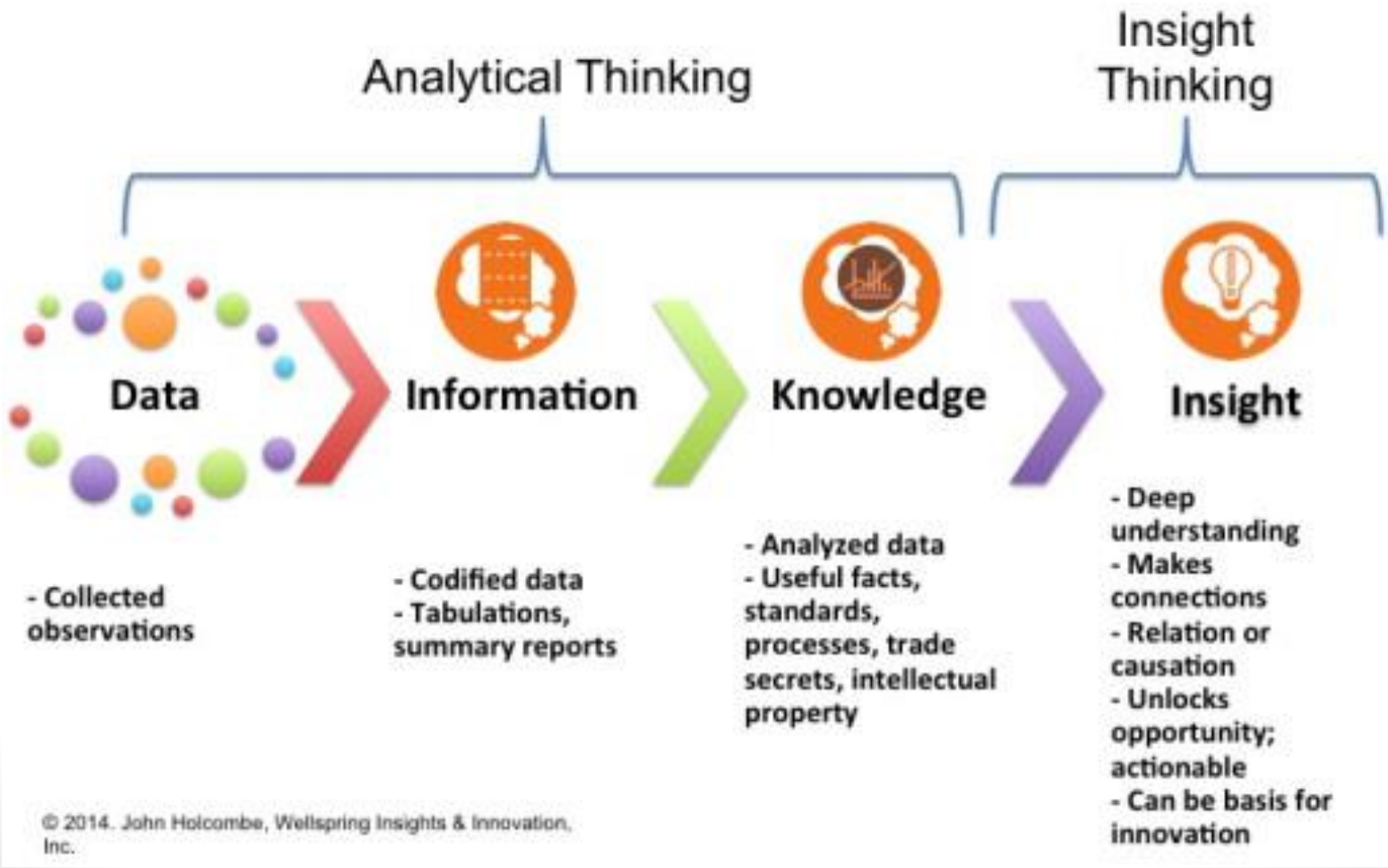
Interpret the data?



Data Product

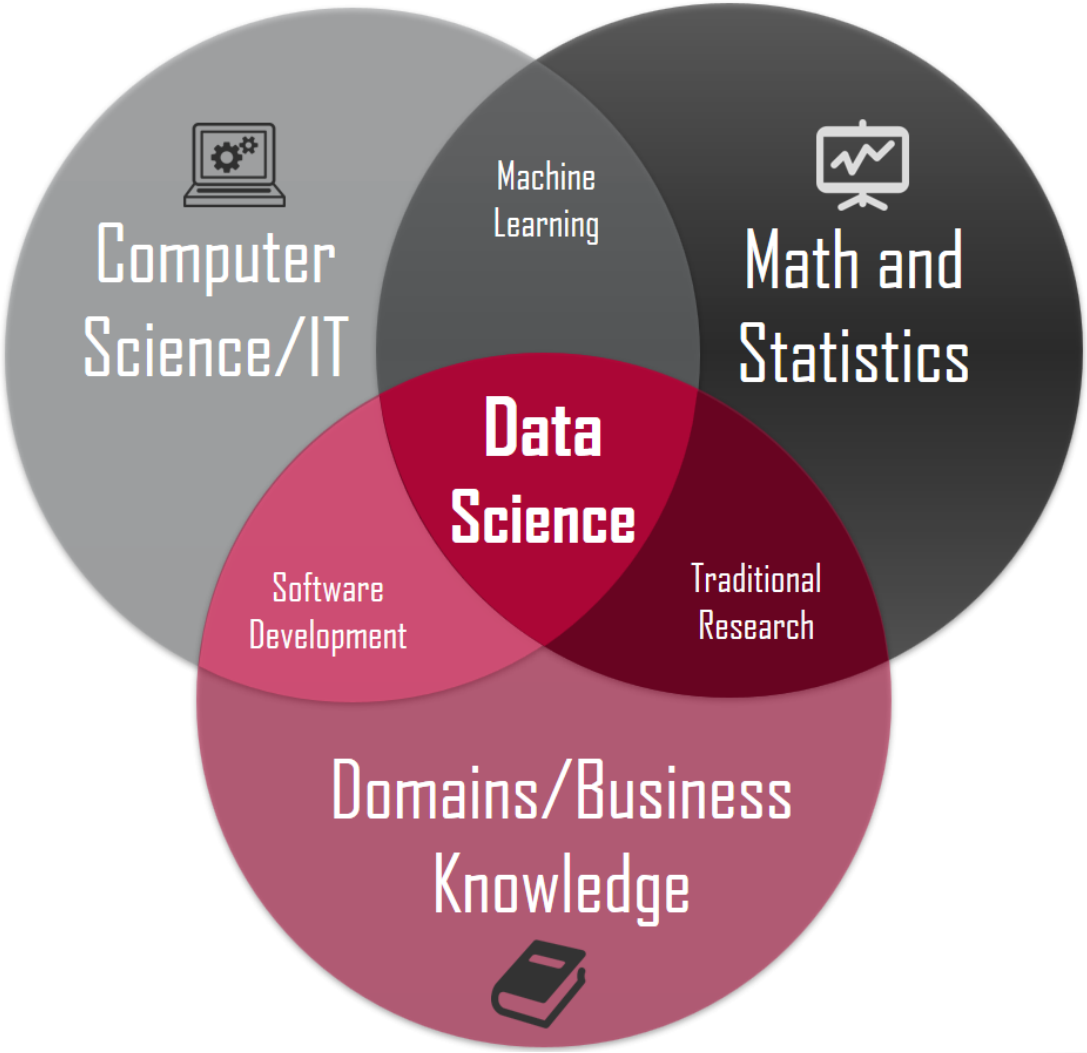
Data product work differently than data insight

- Data insights help to provide some advice to help a business executive make smarter decisions.

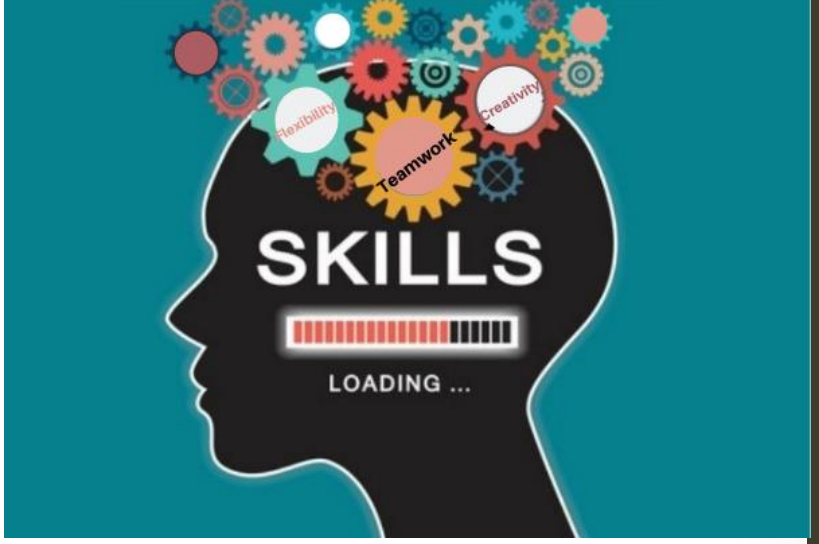


Needed Skills

What do you need?



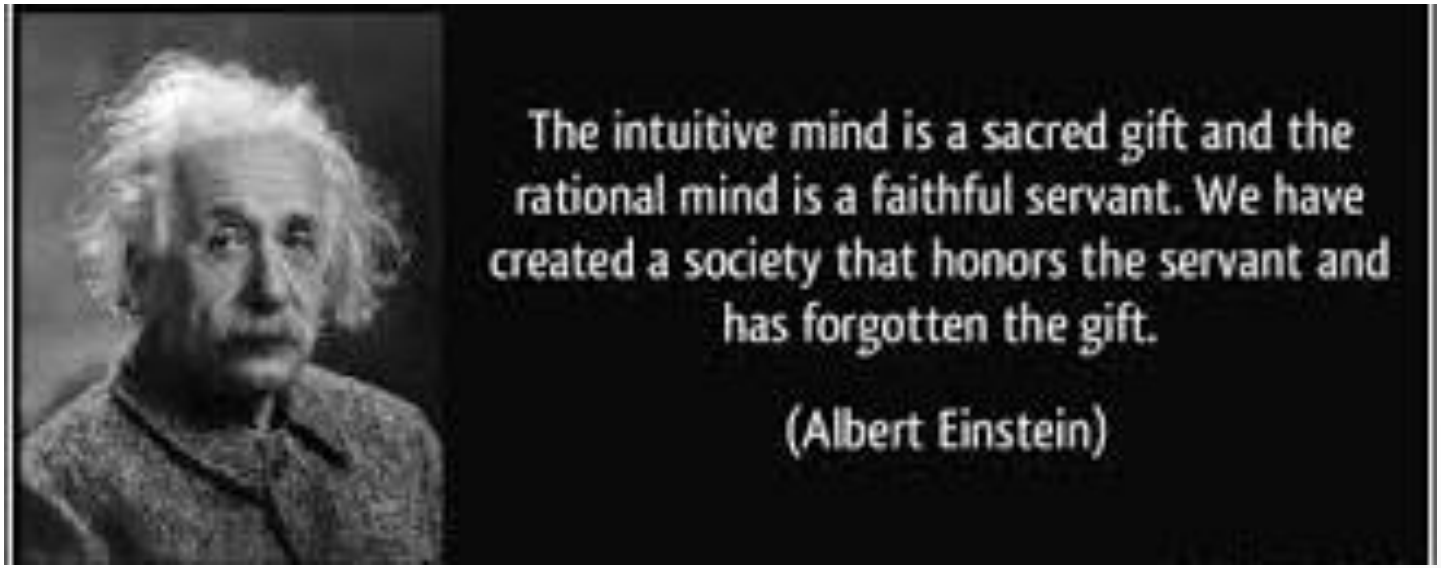
Technical Skills



- ✓ Skills in programming
- ✓ Skills in statistical analysis.
- ✓ Understanding of analysis tools.
- ✓ Adept at working data that is unstructured.
- ✓ Ability to data process and mine.
- ✓ Preferably a Master's or Ph.D. in engineering, statistics, or computer science.

Non-Technical Skills

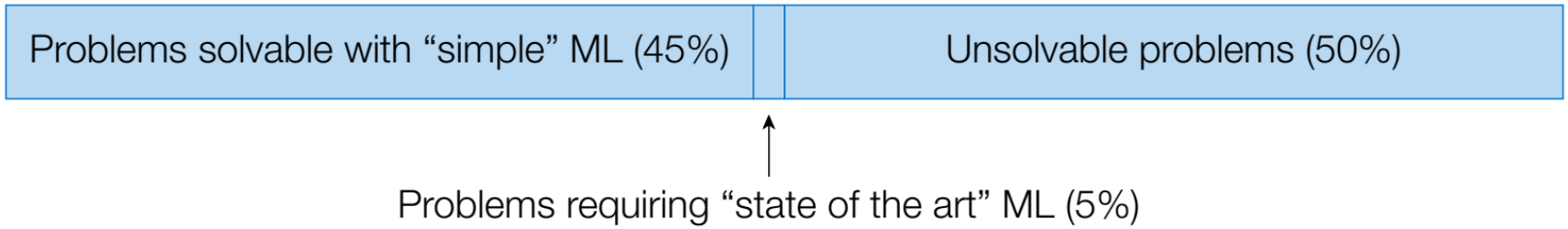
- ✓ Great data intuition.
- ✓ Strong communication skills.
- ✓ A strong business acumen.



Contrast

Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/ validate on a few, relatively clean, small datasets	Develop or use tools that can handle massive datasets
Publish a paper…	Take action…

Universe of machine learning problems

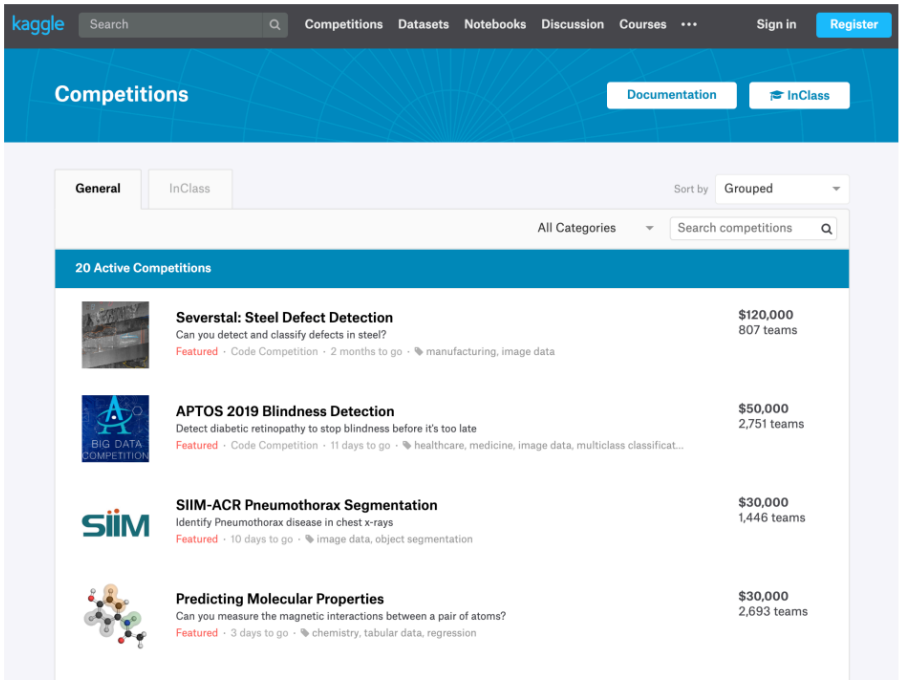


Contrast

Data Science is not Kaggle competitions either….

- Competitions like Kaggle ask for optimizing a metric on a fixed dataset.
- This may or may not ultimately solve the desired business/ scientific problem

Data science is an iterative process



Contrast

Big Data	Data Science
Organizations have to gather big data to help improve their efficiency, enhance competitiveness, and understand new markets.	It provides the mechanisms or tools to understand and use big data quickly.
There is no limit to how much valuable data that can be collected.	To use this data, the important information for business decisions has to be extracted.
People characterize big data by its volume, velocity, and variety, which is often referred to as the 3Vs.	It provides the techniques and methods to look at the data that is characterized by the 3Vs.

Developer Tools