Data Science
Big Picture

Author: Prasad Siri
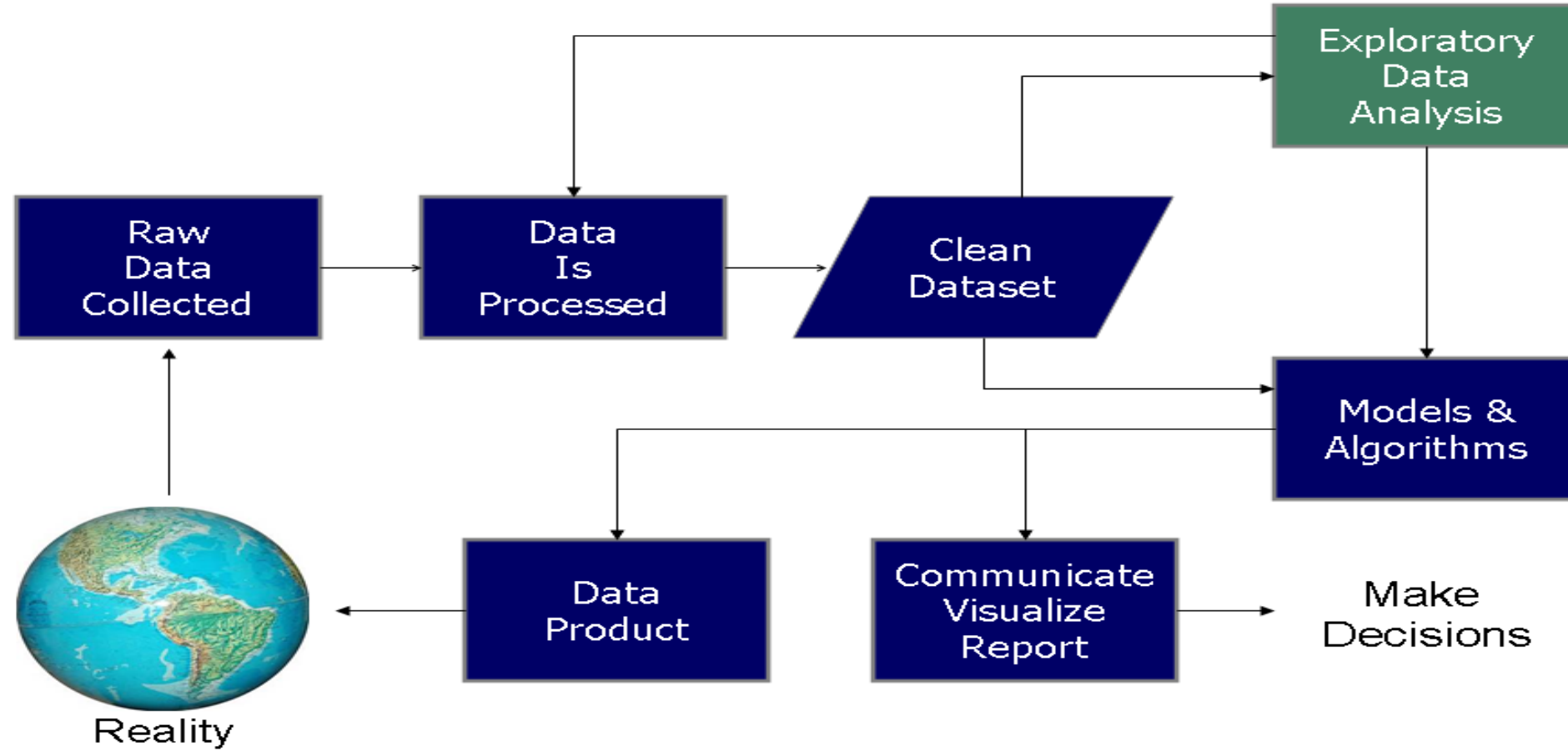
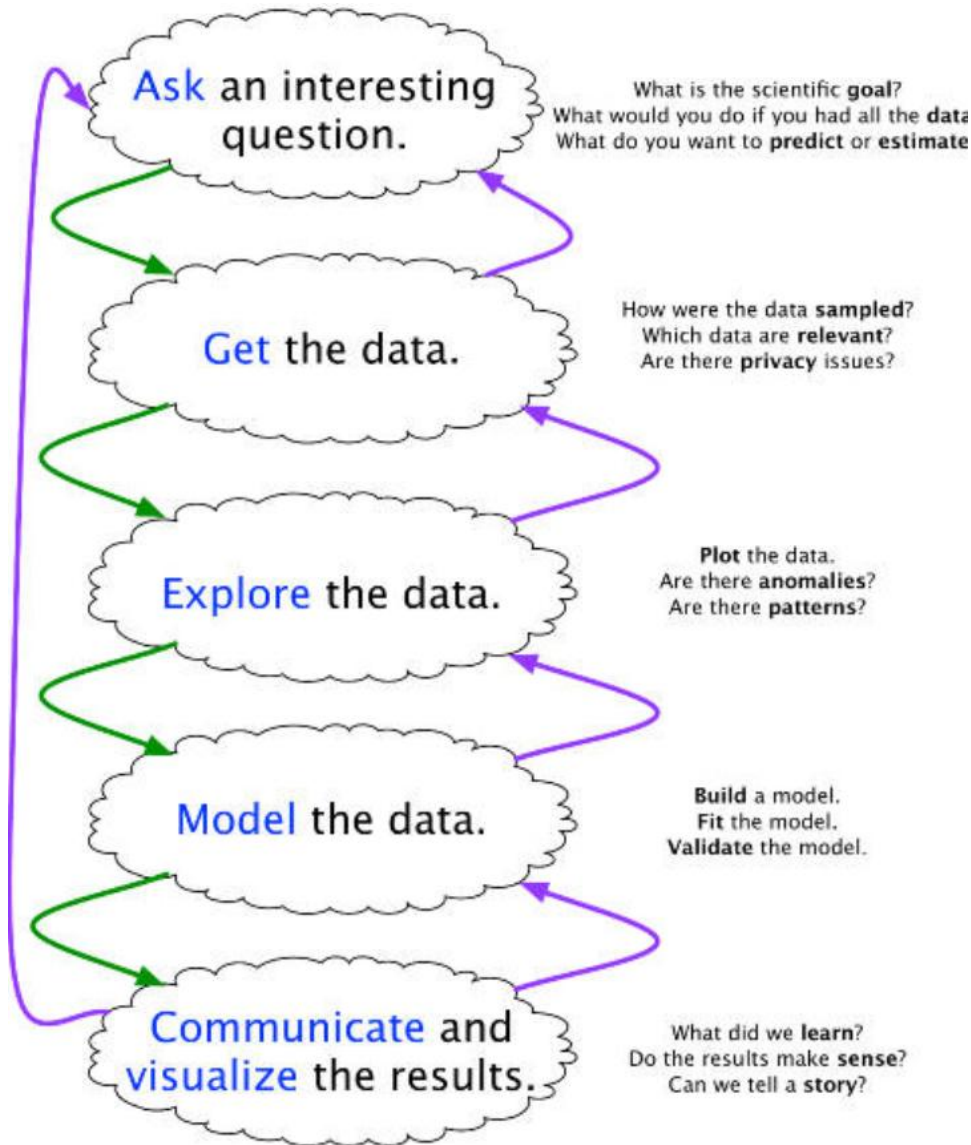Source: **Microsoft**

DATA that MATTERS

Definition

- **Data science** is an interdisciplinary field about processes and systems to extract knowledge or **insights** from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).

Process

Data Science Process

Process

Ask an interesting question.
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.
Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.
Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

Stage 1: Ask A Question

- Skills: science, domain expertise, curiosity
- Tools: your brain, talking to experts, experience

Stage 2: Get the Data

- Skills: web scraping, data cleaning, querying databases, CS stuff
- Tools: python, pandas

Stage 3: Explore the Data

- Skills: Get to know data, develop hypotheses, patterns? anomalies?
- Tools: matplotlib, numpy, scipy, pandas, mrjob

Stage 4: Model the Data

- Skills: regression, machine learning, validation, big data
- Tools: scikits learn, pandas, mrjob, mapreduce

Stage 5: Communicate the Data

- Skills: presentation, speaking, visuals, writing
- Tools: matplotlib, adobe illustrator, powerpoint/keynote

Colaberry

DATA that MATTERS

Infogrex

Prerequisite

- little Mathematics
- little Statistics
- little Programming

- More Commonsense

"…the six-foot tall man who drowned crossing the river that was six inches deep on average…."

What you enjoy about being a Data Scientist
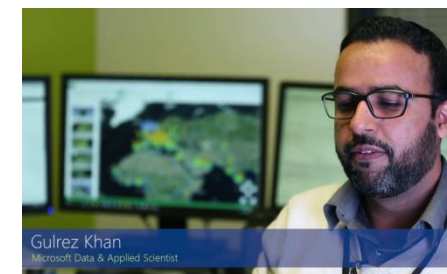
Gauher Shaheen
Microsoft Data Scientist

Danielle Dean
Microsoft Senior Data Scientist Lead

Marjan Baghaie
Microsoft Data Scientist

Robert Musson
Microsoft Principle Data Scientist, Skype

Gulrez Khan
Microsoft Data & Applied Scientist

olaberry

Infogrex

DATA that MATTERS

What you enjoy about being a Data Scientist

Danielle Dean
Microsoft Senior Data Scientist Lead

What you enjoy about being a Data Scientist

- Problem Solver & Value to the customer

- Playing with Data and Making Sense with data

- Investigation

- Solving Puzzle and New Knowledge

- Pattern

- Listen carefully … it will talk , User trend , not opinion

- Working with Different background

What Data Scientist Do

Robert Musson
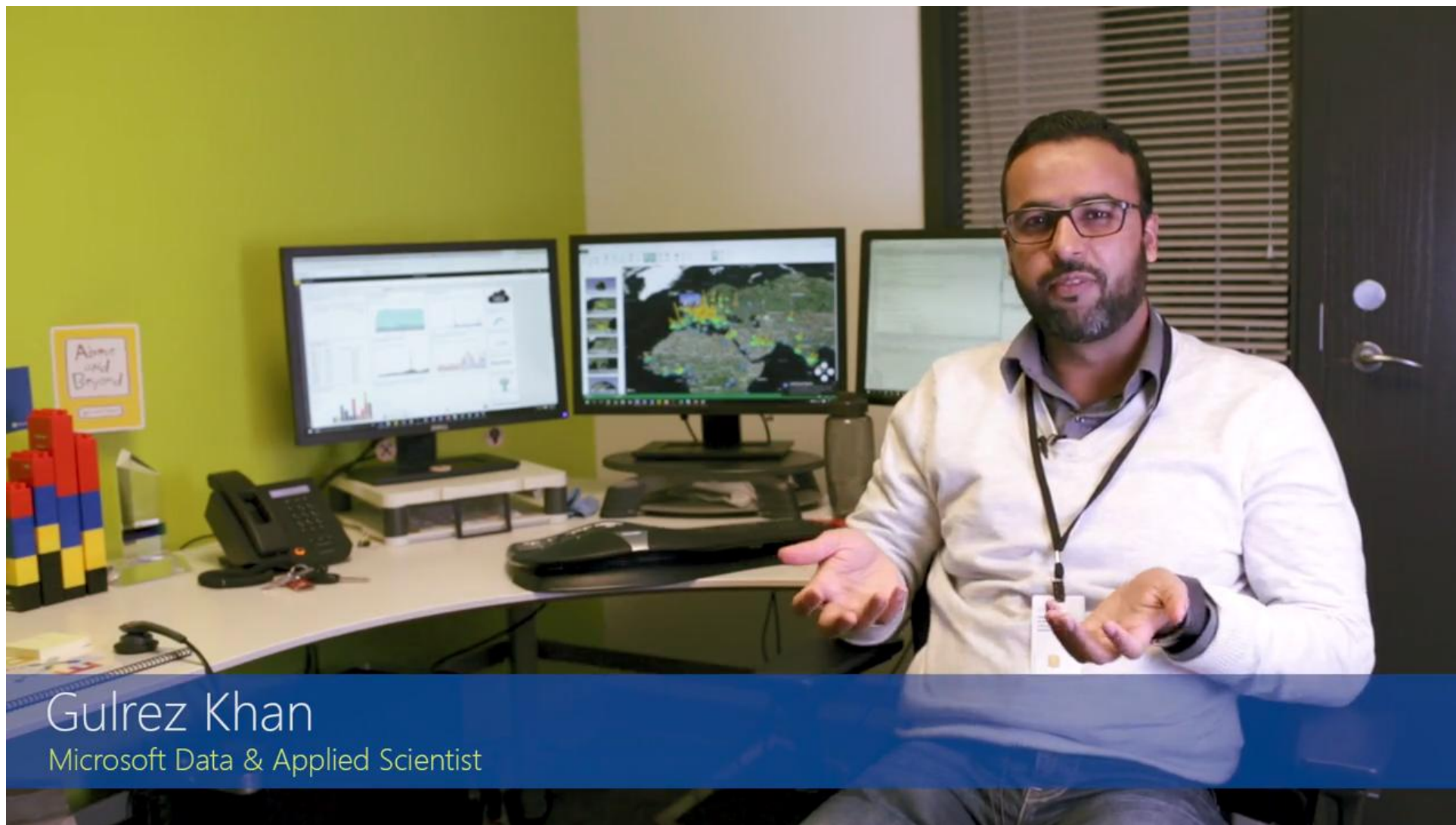Microsoft Principle Data Scientist, Skype

What Data Scientist Do

- Dealing With Data

- Raw Data to Transforming Data

- Cleaning Data , Good Enough to Trust worthy

- Understanding and Preparing Data , Multisource,

- Creating Algorithm to create insight

What Kind of Person Succeeds as Data Scientist

Gulrez Khan
Microsoft Data & Applied Scientist

What Kind of Person Succeeds as Data Scientist

- Little bit Crazy , like data

- Love the Data , Dream about data, Creativity

- Curious , Work with many individual

- Quick learner , ask question … the data

- Analyzing the data

laberry

Infogrex
DATA that MATTERS

What Skills Does a Data Scientist Need

Marjan Baghaie
Microsoft Data Scientist

What Skills Does a Data Scientist Need

- Statistics

- Mathematics

- Soft skills

- Programing language : R , Python

- Excel, Power BI, Tableau, Click and etc

- Graphs, Charts

What Advice Would You Give an Aspiring Data Scientist

Gauher Shaheen
Microsoft Data Scientist

What Advice Would You Give an Aspiring Data Scientist

- Don't wait , Explore the Data Sets …. Data.gov

- Learn from other people and solve some

- Seek Problem and share with community

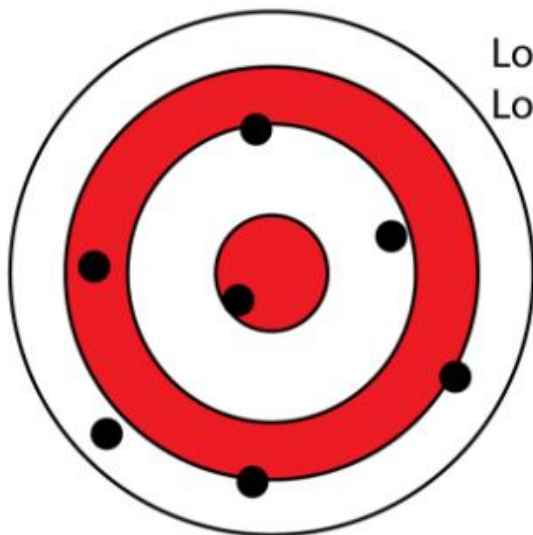- Methodology (There is no right answer from lucky guess)

Question First …

The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data - John Tukey
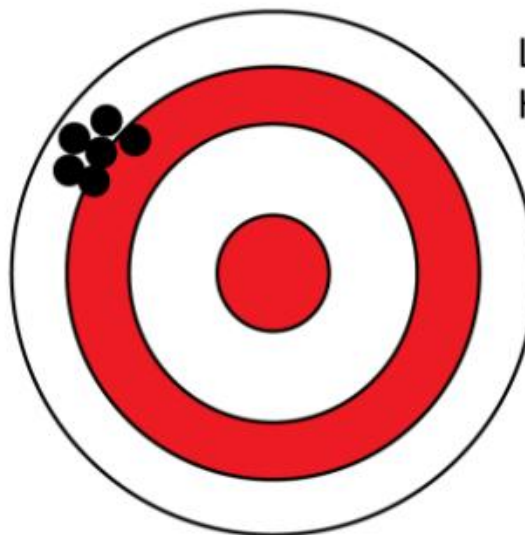
Be approximately right
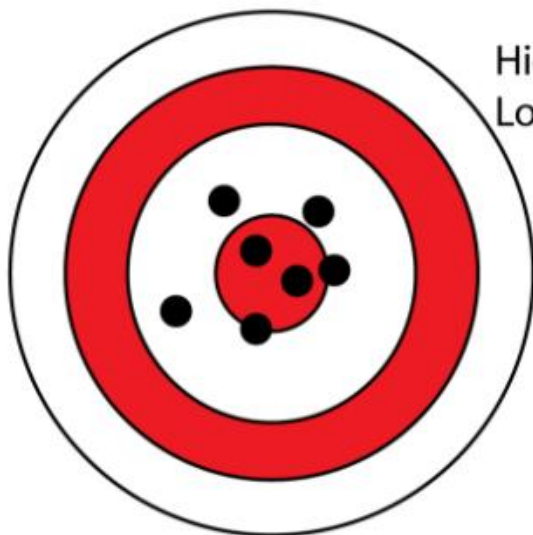rather than exactly wrong.
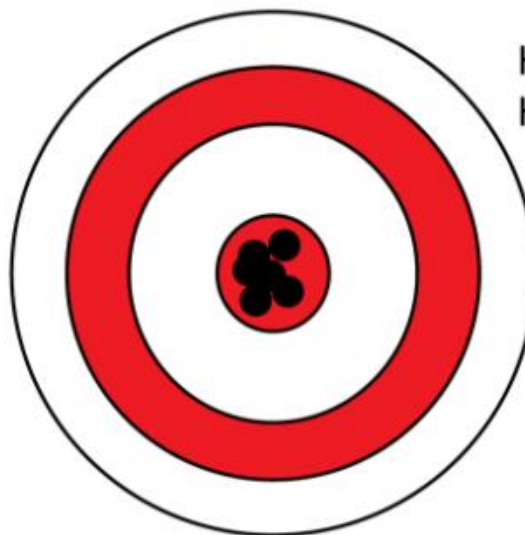
Precision vs Accuracy

Low accuracy
Low precision

Low accuracy
High precision

High accuracy
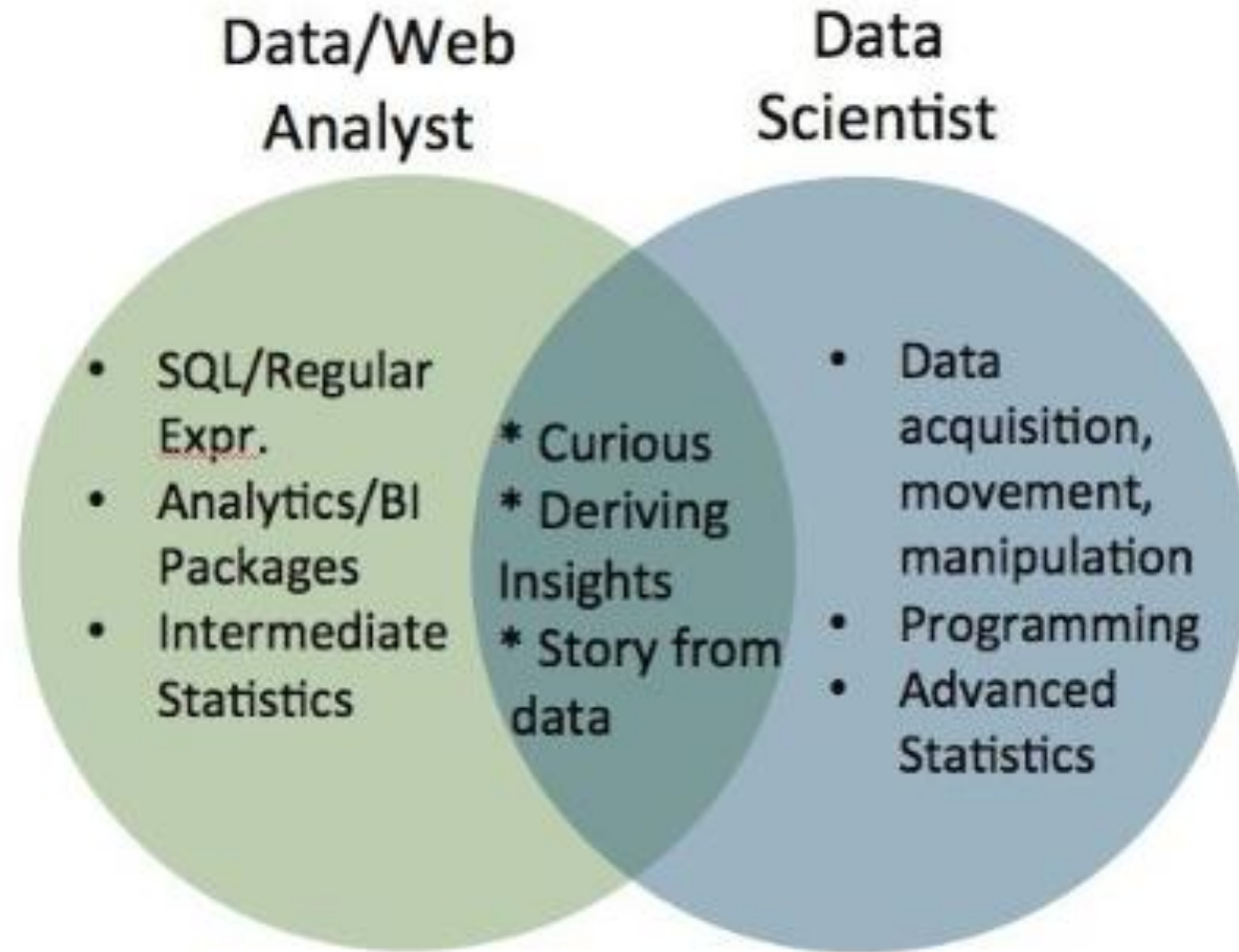Low precision

High accuracy
High precision

Data Scientist Vs Data Analyst

Data/Web Analyst
- SQL/Regular Expr.
- Analytics/BI Packages
- Intermediate Statistics

* Curious
* Deriving Insights
* Story from data

Data Scientist
- Data acquisition, movement, manipulation
- Programming
- Advanced Statistics

Colaberry

Infogrex
DATA that MATTERS

Data Science vs Big Data vs Data Analytics

WHAT ARE THEY?

Data Science is a field that comprises of everything that related to data cleansing, preparation, and analysis.

Big Data is something that can be used to analyze insights which can lead to better decision and strategic business moves.

Data Analytics Involves automating insights into a certain dataset as well as supposes the usage of queries and data aggregation procedures.

Job Opportunity

8 Data Science skills 1.5 million jobs

Opportunities for Data Scientists—one of today's hottest jobs—are rapidly growing in response to the exponential amounts of data being captured and analyzed. Companies hire Data Scientists to find insights and to solve meaningful business problems. Get the real-world knowledge and hands-on experience that can help you succeed in one of these new jobs.

Prove that you have what it takes in the Microsoft Professional Degree Program.

8 Skills

1. Programming Languages – R, Python, SAS
2. Database – SQL NOSQL
3. Mathematics – Linear Algebra, Multivariable Calculus
4. Statistics – Descriptive , Inferential
5. Data Munging – 80% of time spent on cleaning data
6. Data Visualization– SQL , NOSQL
7. Thinking like Data Scientist– Data driven problem solver
8. Strong Business Acumen – Domain Knowledge

Data Science vs Big Data vs Data Analytics

Few Application Areas

Internet Search

Target Advertisements

Recommender Systems

Image Recognition

Speech Recognition

Gaming

Price Comparison Websites

Airline Route Planning

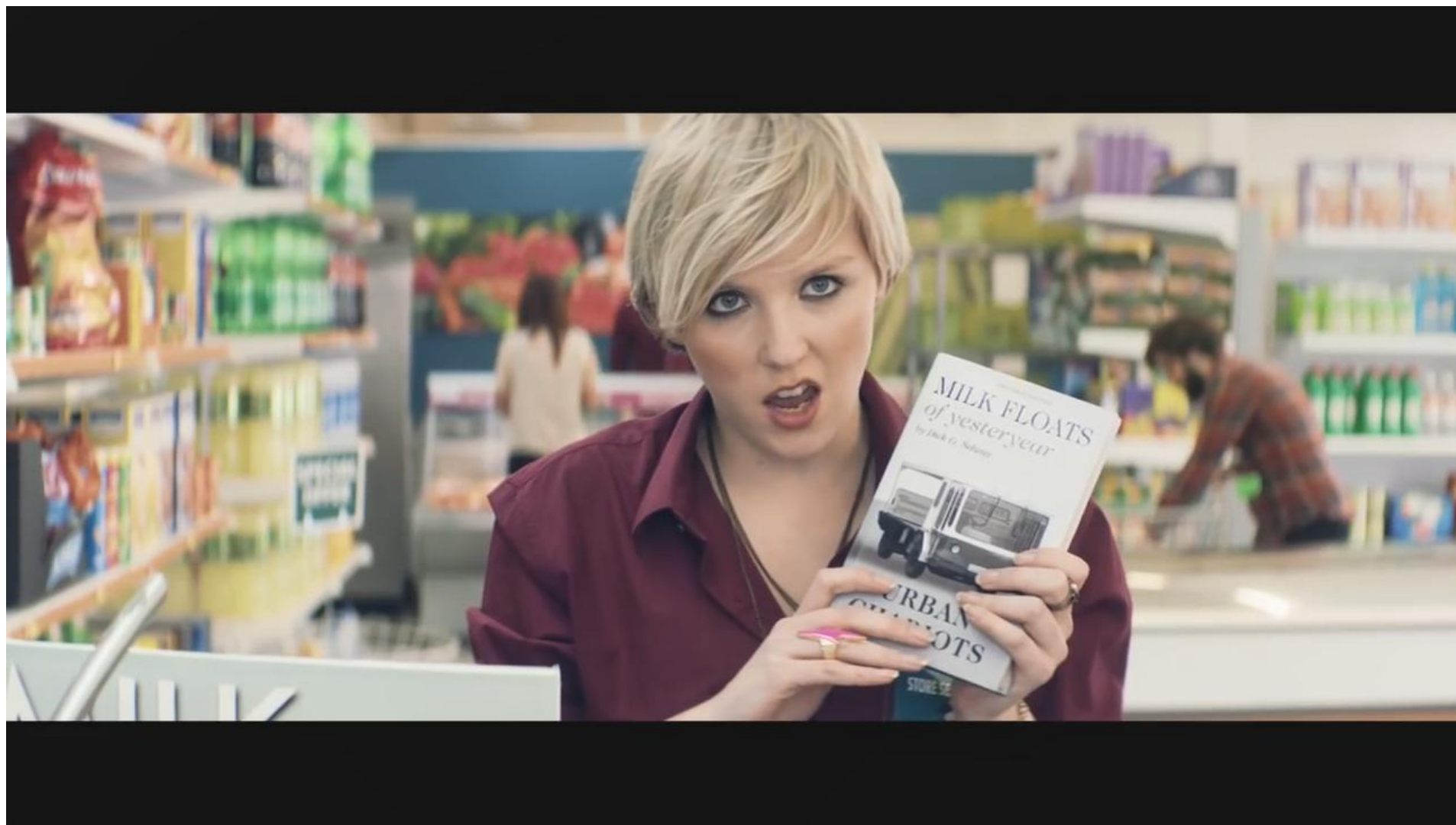Fraud and Risk Detection

Delivery Logistics

Self Driving Cars

Robots

Google Analytics In Real Life - Site Search

In Nutshell

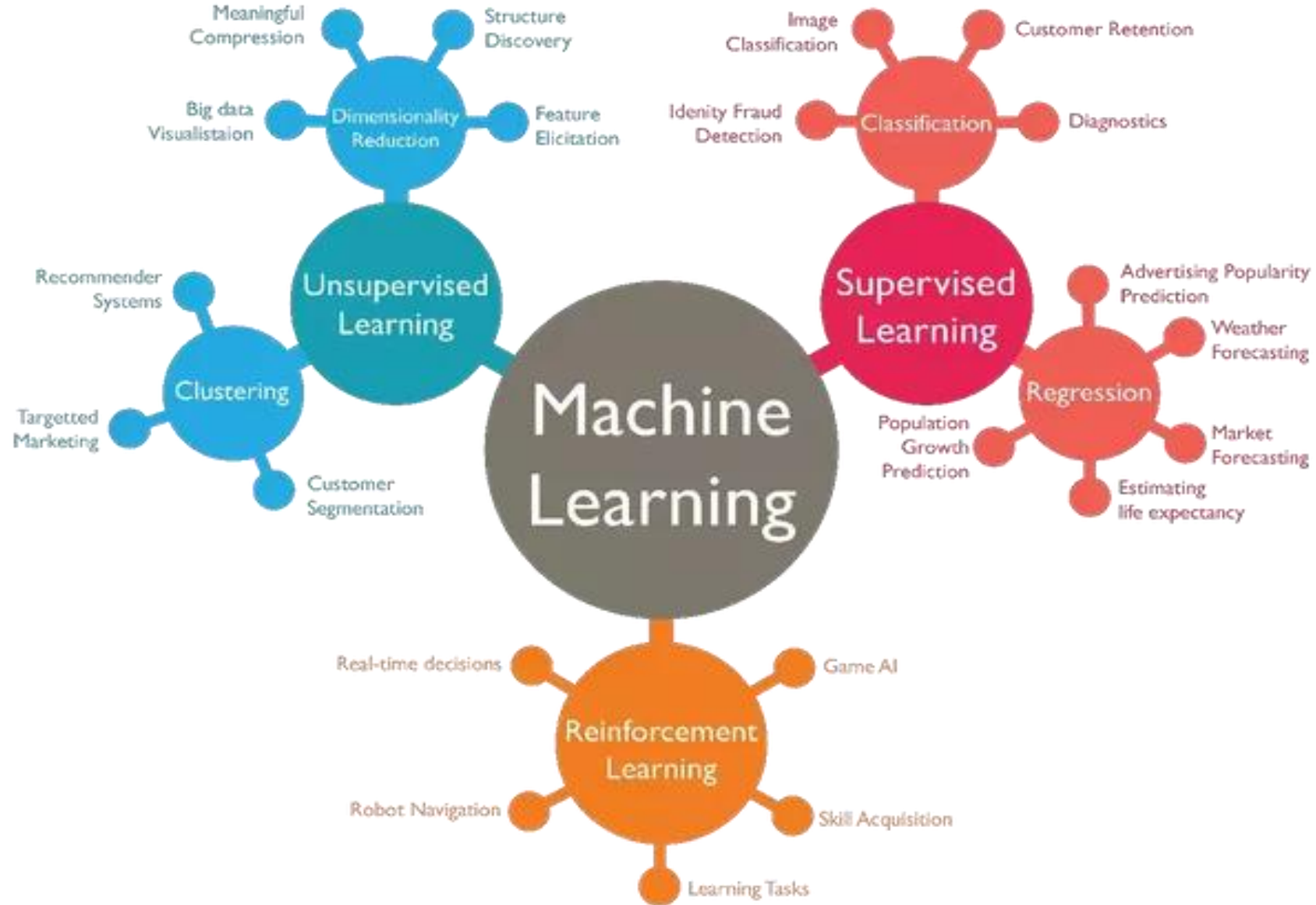If Study of Past and Present is **Data Analytics**

…. then study of Future is **Data Science**

IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Predictive Analytics ...

Patient: "Will I survive this risky operation?"

Surgeon: "Yes, I'm absolutely sure that you will survive the operation."

Patient: "How can you be so sure?"

Surgeon: "9 out of 10 patients die in this operation, and yesterday died my ninth patient."

Thank You

The more you know, the more you know you don't know.

Aristotle