What is Science in Data Science?

What is Science? Past Explanation

Science used to be about devising the best experiment to verify a specific hypothesis.

- Data acquisition was coupled to a specific hypothesis
- See Scientific Method

What is Science? Future (Now)

The abundance of data has led to a new paradigm

- >Data is ubiquitous
- >We need methods to sift through the data and extract meaning.
- –Existing data allows falsification and corroboration of many hypotheses

An Example: Genomics

"Omics" are the data-intense technologies in Bio-medicine

First Human Genome

- Time: **12 years** (1990 – 2002) to sequence first genome
- Cost: **3 Billion $**
- Use in Science: Too expensive to devise an experiment that requires whole-genome sequencing

An Example: Genomics Today

Time: 24 hours

Cost: $1000.

Use in Science: Most investigators do not even need to pay for sequencing since enough sequences already exist

- Expense and time are spent on sifting through the data.

What is Data in Data Science?

What is Data?

Data are observations that are put into context

Given that every observation has a context, an observation is a datum

- Not Data:
  - >1 (one)
  - >Chair
  - >Diabetes
- Data:
  - >I see a chair
  - >The patient has diabetes

Unstructured Data

Unstructured data does not exist.

- The essence of data is that they are structured.
- The context is what makes data.

What is meant by Unstructured Data?

Answer: poorly structured data that are hard to analyze

- Need to restructure the data through parsing, etc.
- A list of tweets is often used as an example of unstructured data.
  - > The tweets are organized into a list
  - > Any single tweet comes from one source at one time
  - > A tweet is a text with a length constraint.

Data Structure leads to Data Types

For example: a list of tweets

- The data type is “a list of tweets”
- A list inherits many characteristics of lists in general
- A tweet inherits many characteristics from the data type text

Typing makes Data

Typing is the **context** of the data

A list of tweets can be represented as a table

- The column header provides context
- The table structure states that all column values have comparable structures

Data Types in Physics

A data type is called a **unit**

- Data are well-typed and can be universally converted
  - > Meters per second, kilograms, joules, seconds

Data Types in Computer Science

Typing may demand structure

- Strong Typing

Structure and context determine typing

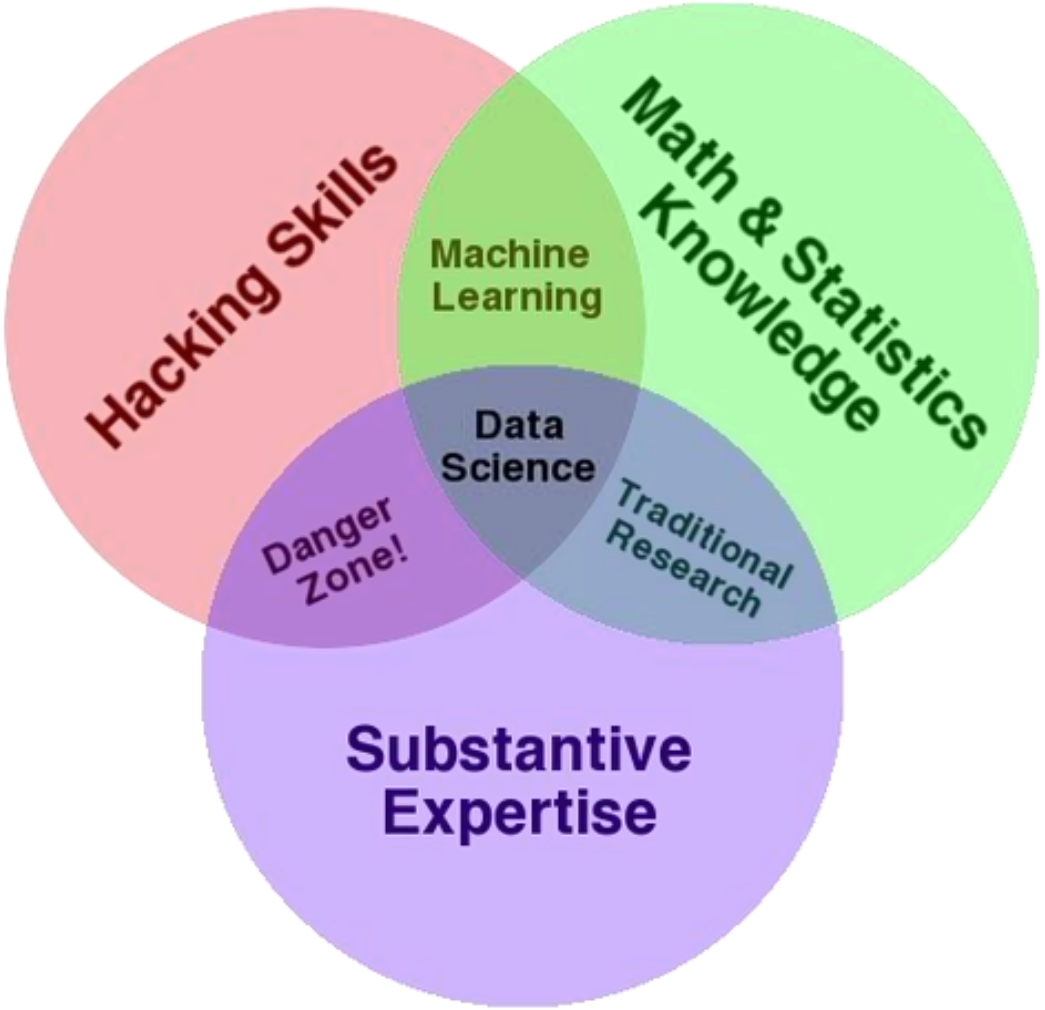
- Weak typing
  - > https://en.wikipedia.org/wiki/Strong_and_weak_typing

What is Data Science?

Data Science is made of two words: **Data** + **Science**

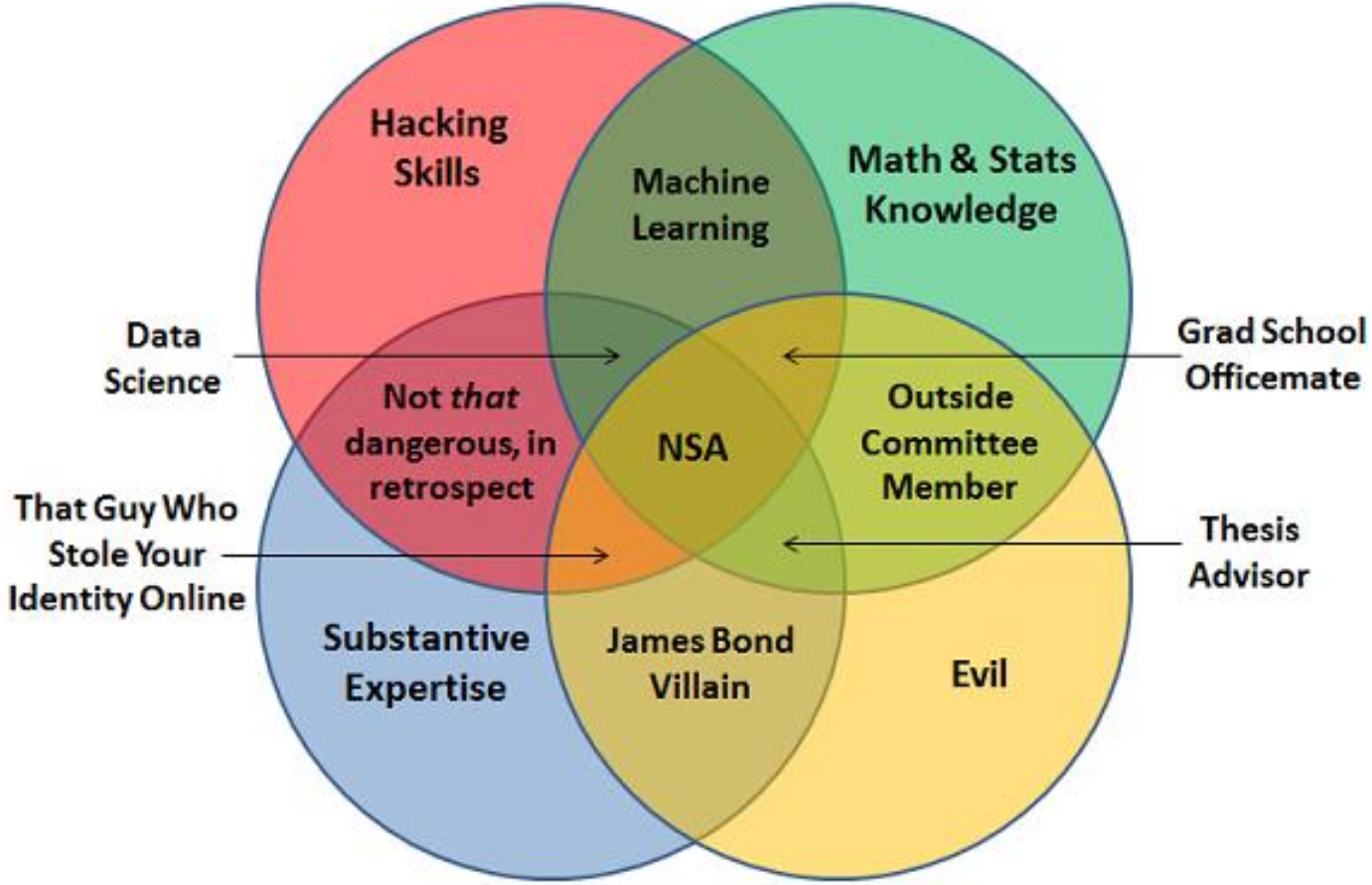
- **Data** and their structures are well explained in computer science
- The **Science** part of data science is explained by the scientific method.
- The synthesis of these two disciplines allows
  - > Data Visualization
  - > Data Extraction
  - > Data Processing / Transformation
  - > Hypothesis Verification or Falsification

What is Data Science?



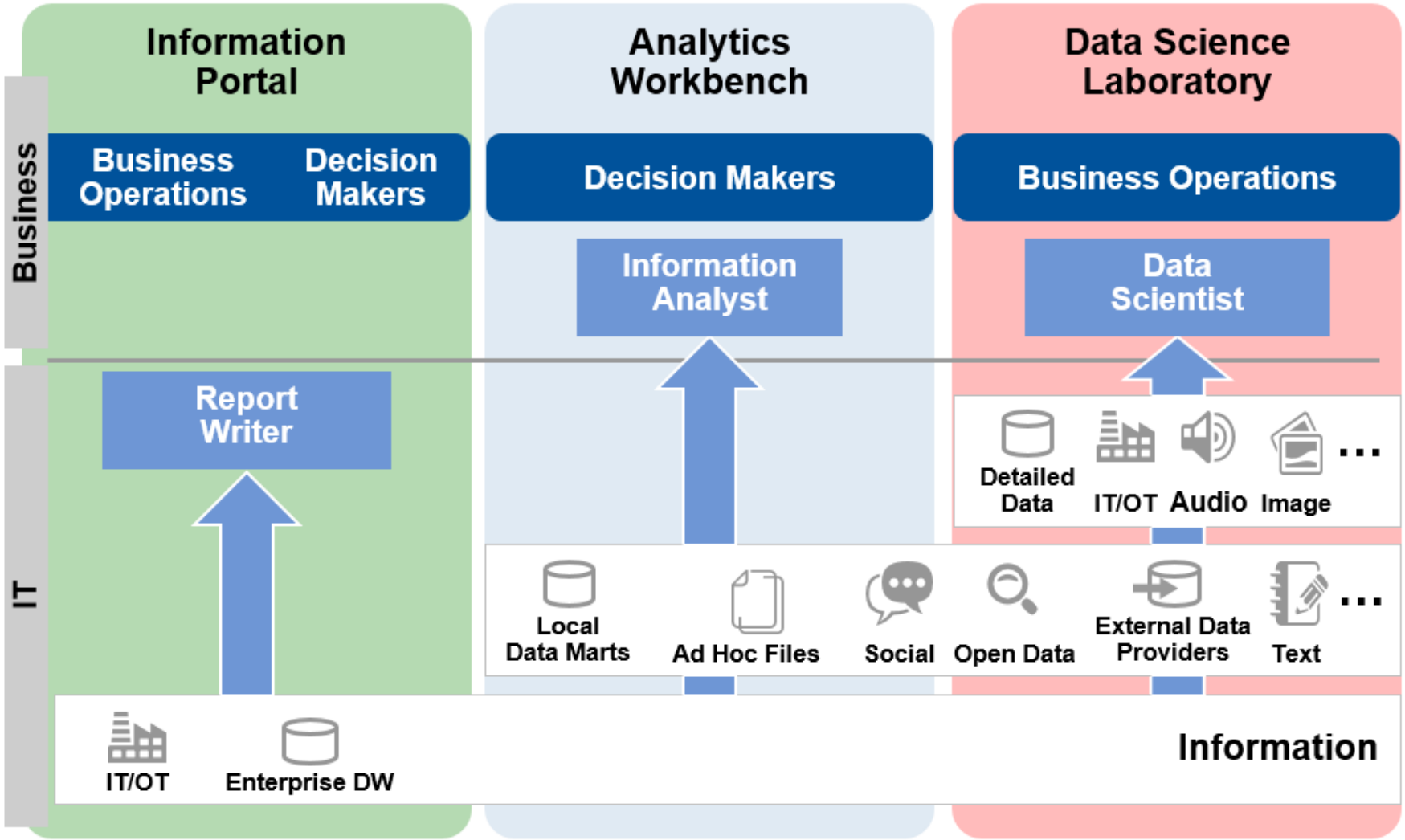
Drew Conway's Data Science Venn Diagram

What really is Data Science?



Drew Conway's Data Science Venn Diagram, Modified

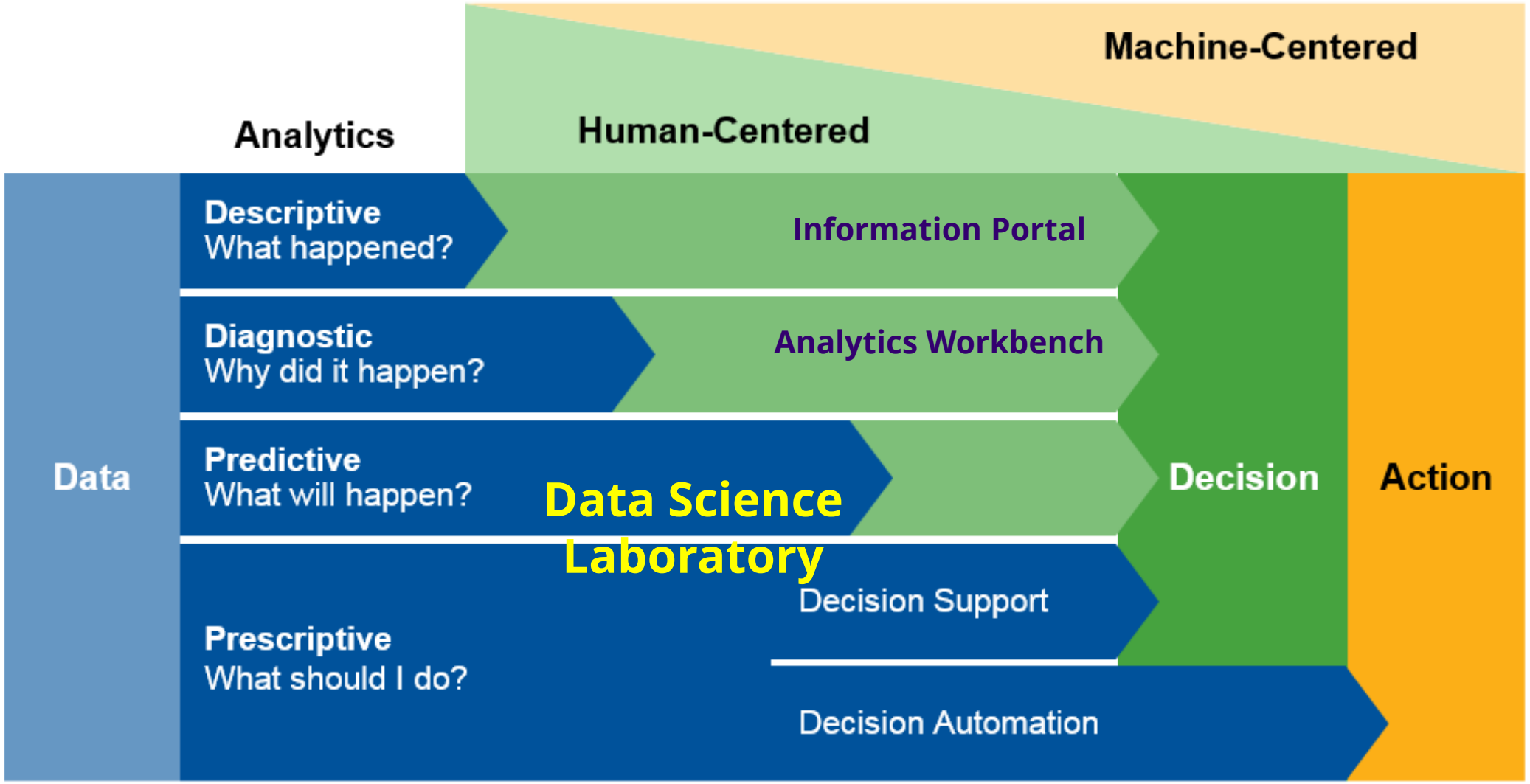
Where does Data Science fit?



DW = data warehouse; IT/OT = information technology/operational technology

Source: Gartner (October 2016)

Another View of Data Science



Source: Gartner (October 2016)

Summary

- Science is the set of methods to extract meaning from data.
- Data is observations in context.
- Data Science involves predictive, prescriptive, and machine-learning analytics efforts.

What is Data Science?

A look at Science and Data

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON