Foundations of Data Science

Master’s in Data Science

2022 / 2023

Course Information

About me

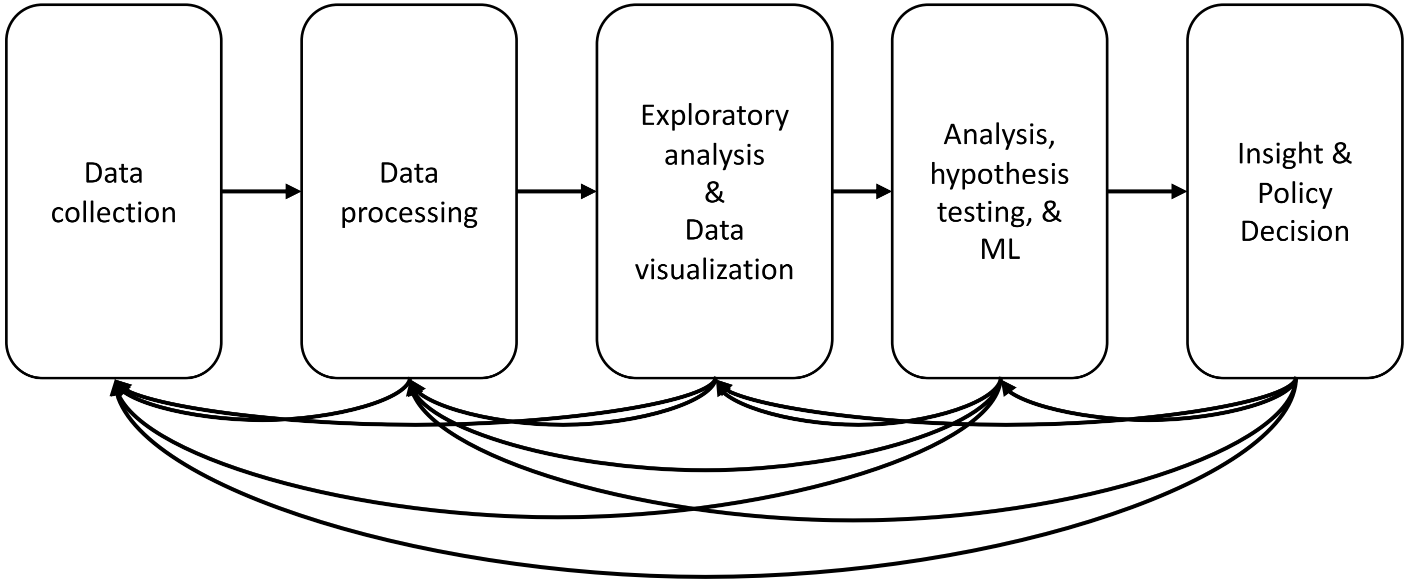
- Sérgio Matos
  aleixomatos@ua.pt

- Room: IEETA

- Research interests: (applied) NLP, ML/DL
  Data & Text Mining
  Information Retrieval & Extraction
  Clinical + Bio* applications

Objectives

- Data science pipeline, challenges, and application areas
  - Data collection, manipulation, transformations
  - Data representation, exploration and modelling
  - Visualisation
  - Reproducibility
  - Interpretability
  - Operationalization
  - Ethics and privacy


Data collection
Data processing
Exploratory analysis & Data visualization
Analysis, hypothesis testing, & ML
Insight & Policy Decision

The Data Lifecycle (from: CMSC641 Principles of Data Science)

Class methodology

- Mix of theory and practice
  - Understand the principles
    - Some aspects will be detailed in other UCs
  - Learn (more) by doing

- Practical exercises
  - Python
  - Python libraries: pandas, scikit-learn, …
  - Tools: conda, jupyter-lab, VS code, git

Grading methodology

Practical project:

- 3 intermediate submissions = 3*20%
- Final report and presentation = 40%
  - This is mandatory!

- Groups of two students (individual work can be accepted)
  - Must collaborate, not split the work
  - Follow-up during practical sessions
  - Grades may differ within a group

Books

- Data Science from Scratch, Joel Grus, 2nd Edition
- Python for Data Analysis, Wes McKinney, 3rd Edition
- Python Data Science Handbook, Jake VanderPlas

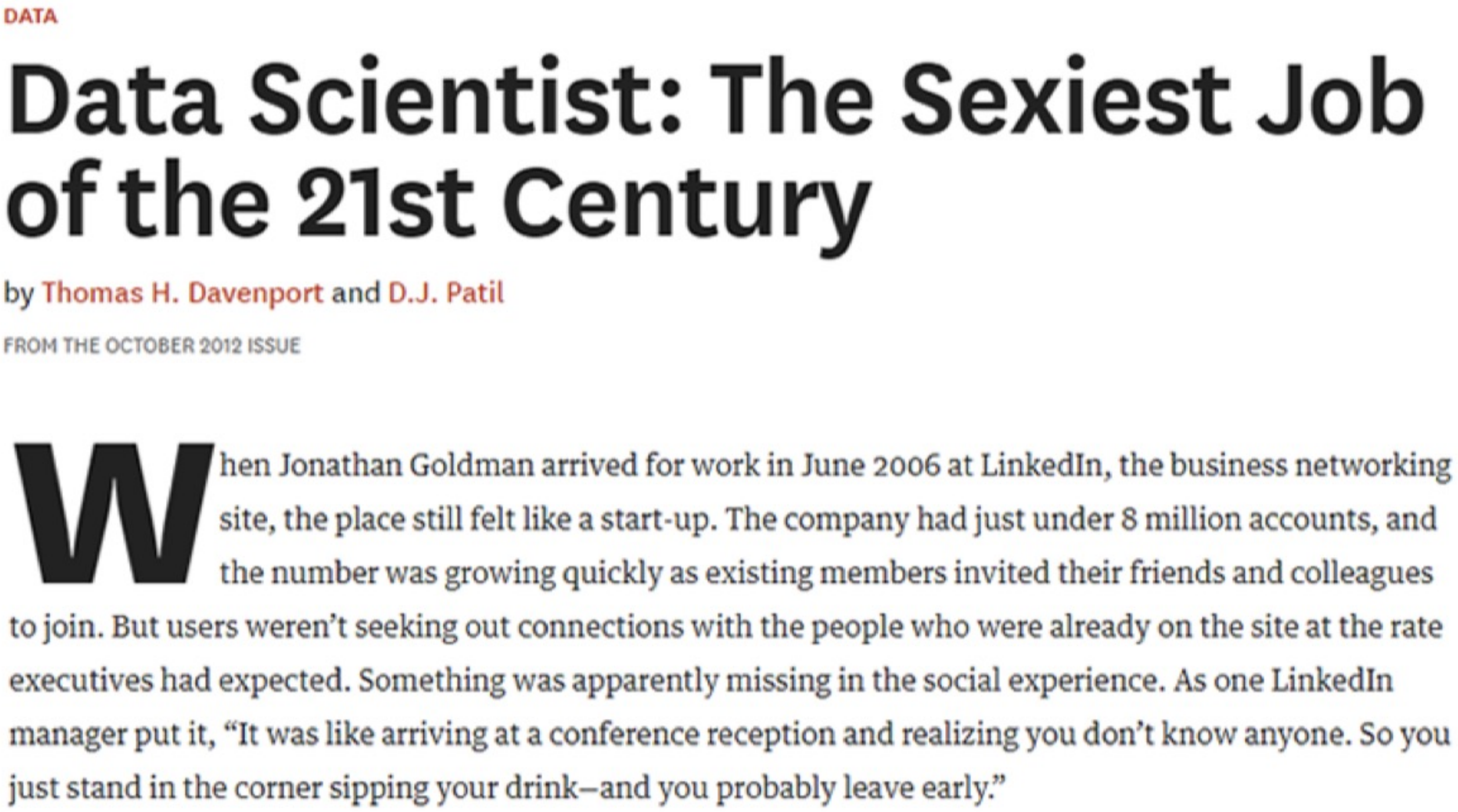
O’Reilly for Higher Education

INTRODUCTION

What is (the importance of) Data Science?

Harvard Business Review
OCTOBER 2012
The True Measures Of Success
10 Rules for Managing Global Innovation
What Ever Happened To Accountability?
GETTING CONTROL OF BIG DATA
How vast new streams of information are changing the art of management
PAGE 59

October 2012 Issue



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

"DATA IS THE NEW OIL."

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered **data to be an economic asset,** like oil.

From the beginning of recorded time until 2003, we created **5 exabytes** (5 billion gigabytes) **of data.**

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

Every hour, we create enough Internet traffic to fill **7 billion DVDs.** Side by side, that's that's seven times the height of Everest.

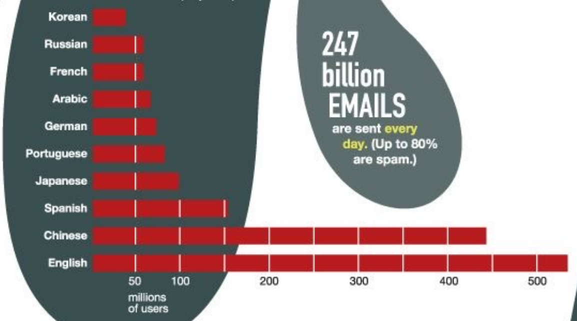
There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

As of August 2012, there were just over **4 million** articles in the English Wikipedia.

There are **133 million BLOGS** on the web.

English is the dominant language of the web. But by 2014 it will be **Chinese,** if its current rate of increase continues.

Top languages used on the web (May 2011):



80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)

247 billion EMAILS are sent every day. (Up to 80% are spam.)

10% of all photos ever taken were taken in 2011.

60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

50% of 5-year-old kids in the U.S. are given access to a smartphone.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders,** with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave **5 milliseconds** from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

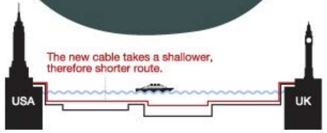
With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 milliseconds

The depth of the Atlantic Ocean varies.

The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.



Information Graphics by Nigel Holmes

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?

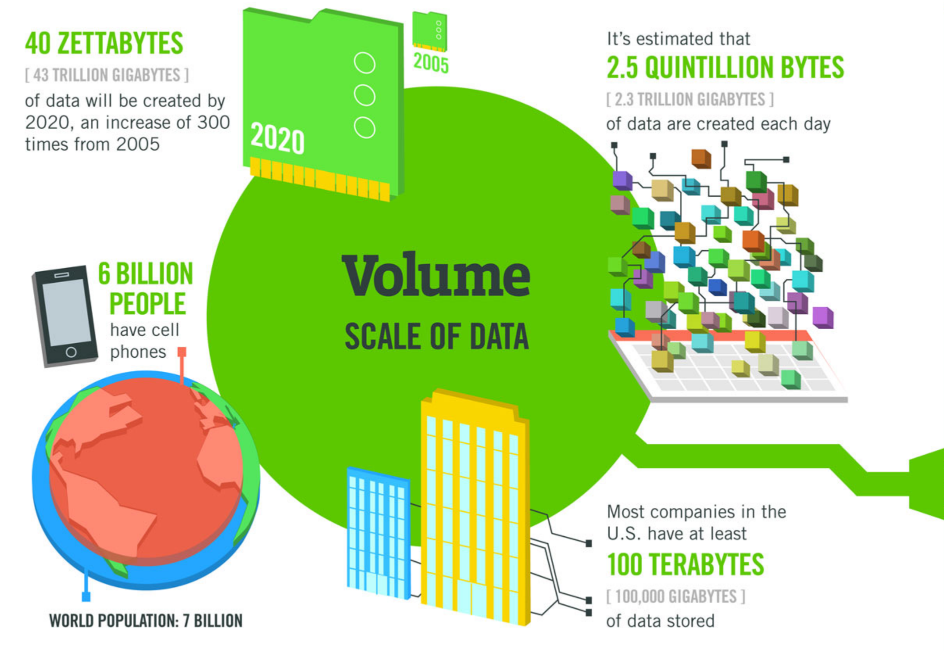
As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

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

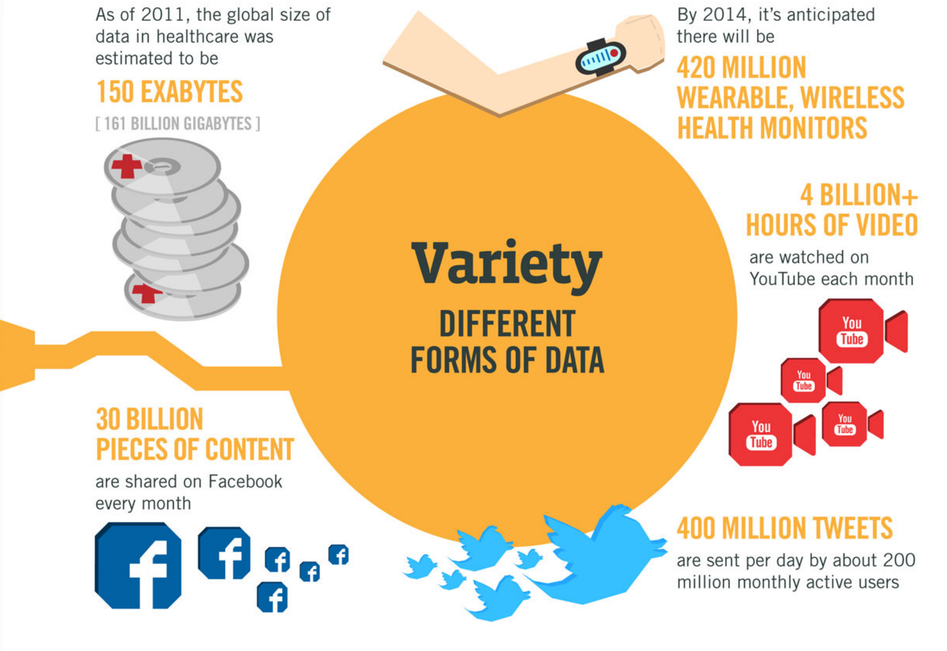
WORLD POPULATION: 7 BILLION

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

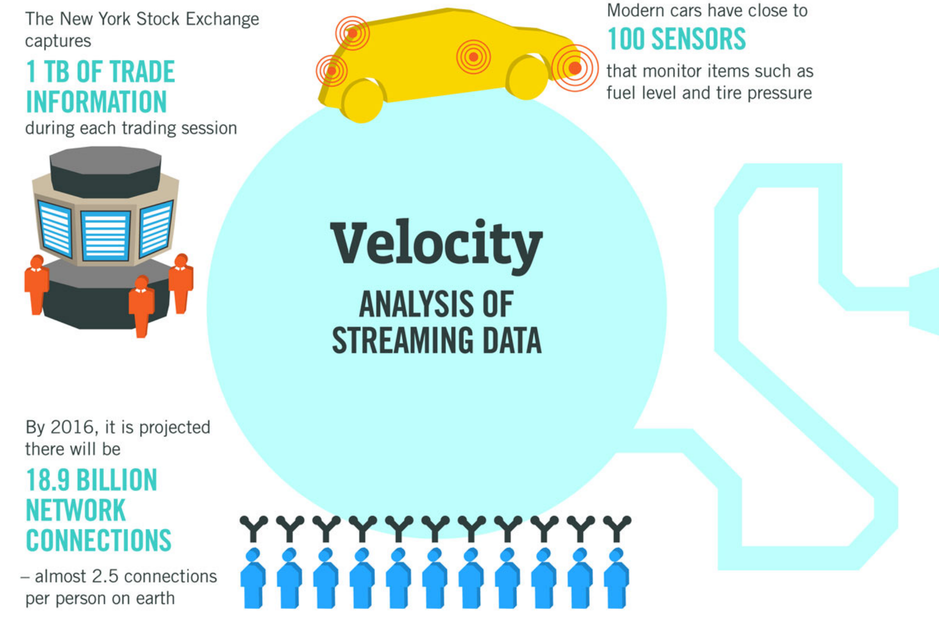
are watched on YouTube each month

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

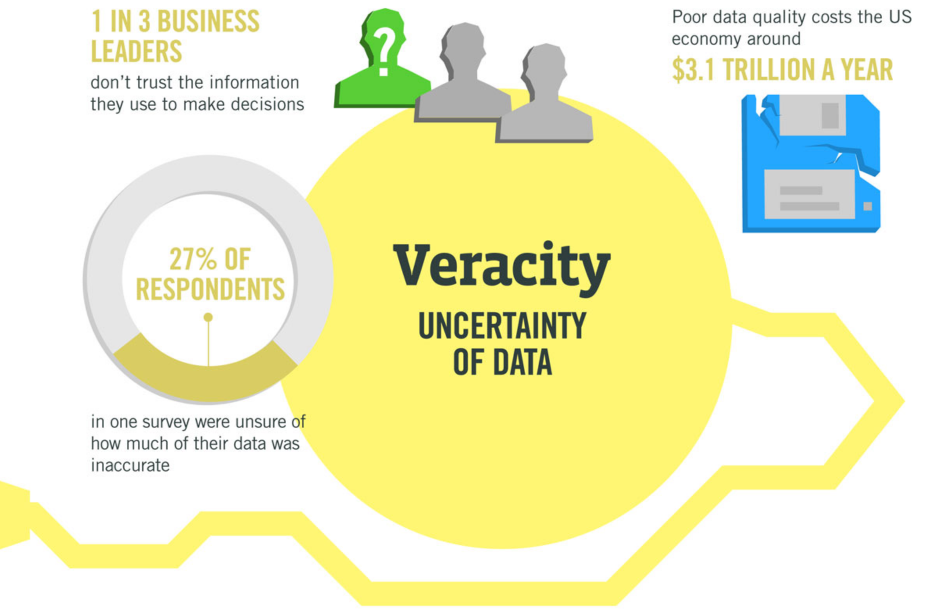
100 SENSORS

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

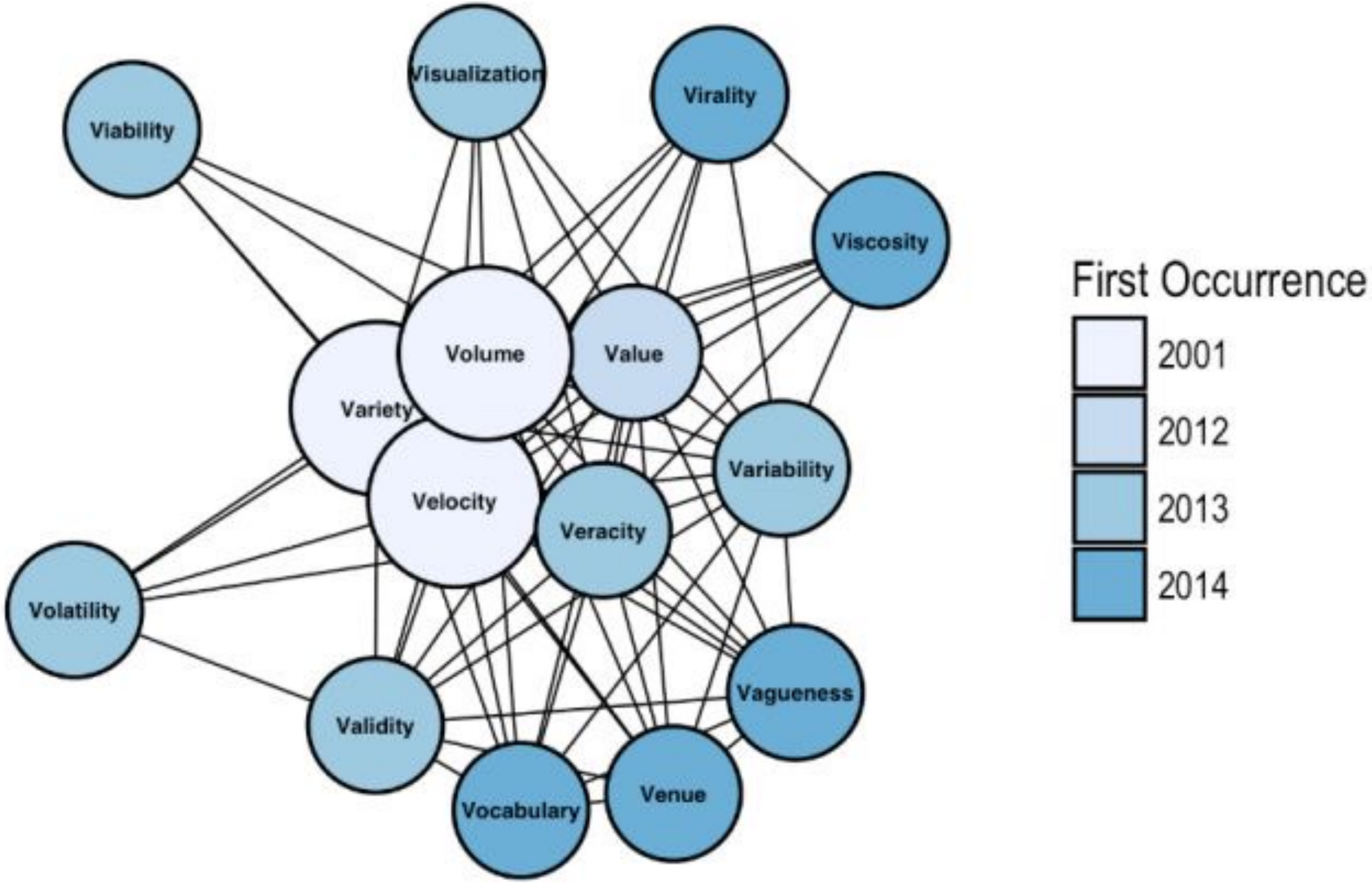
$3.1 TRILLION A YEAR

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



https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

Data

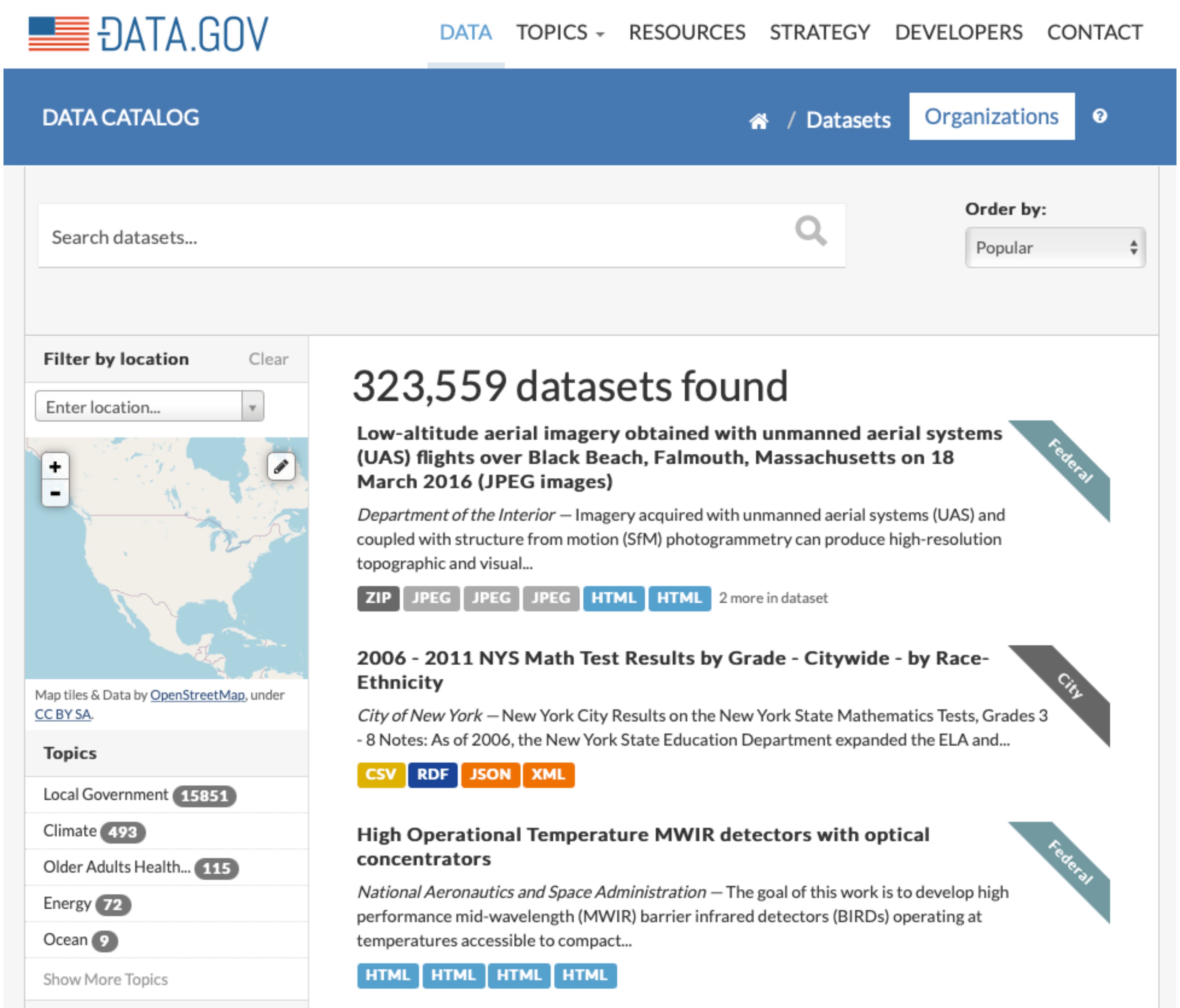
?

Data

Pieces of information,
in the form of numbers, text, measurements, observations, images,
that describe or represent an entity or environment

Data

- Structured data are organized according to a formal model that follows the domain and business logic, regularly using tables and relations in traditional data bases. Example: a company's DB
- Semi-structured data do not have a formal structure but contain markers that separate the elements and identify a hierarchy of the records and fields. Example: JSON, XML
- Unstructured data do not follow a formal model nor a pre-defined structure; may contain some internal structure, but it's usually inconsistent. Examples: text collections, videos, sensor data



Map tiles & Data by OpenStreetMap, under CC BY SA.

Topics

Local Government	15851
Climate	493
Older Adults Health...	115
Energy	72
Ocean	9

Show More Topics

323,559 datasets found

Low-altitude aerial imagery obtained with unmanned aerial systems (UAS) flights over Black Beach, Falmouth, Massachusetts on 18 March 2016 (JPEG images)

Federal

Department of the Interior — Imagery acquired with unmanned aerial systems (UAS) and coupled with structure from motion (SfM) photogrammetry can produce high-resolution topographic and visual...

ZIP JPEG JPEG JPEG HTML HTML 2 more in dataset

2006 - 2011 NYS Math Test Results by Grade - Citywide - by Race-Ethnicity

City

City of New York — New York City Results on the New York State Mathematics Tests, Grades 3 - 8 Notes: As of 2006, the New York State Education Department expanded the ELA and...

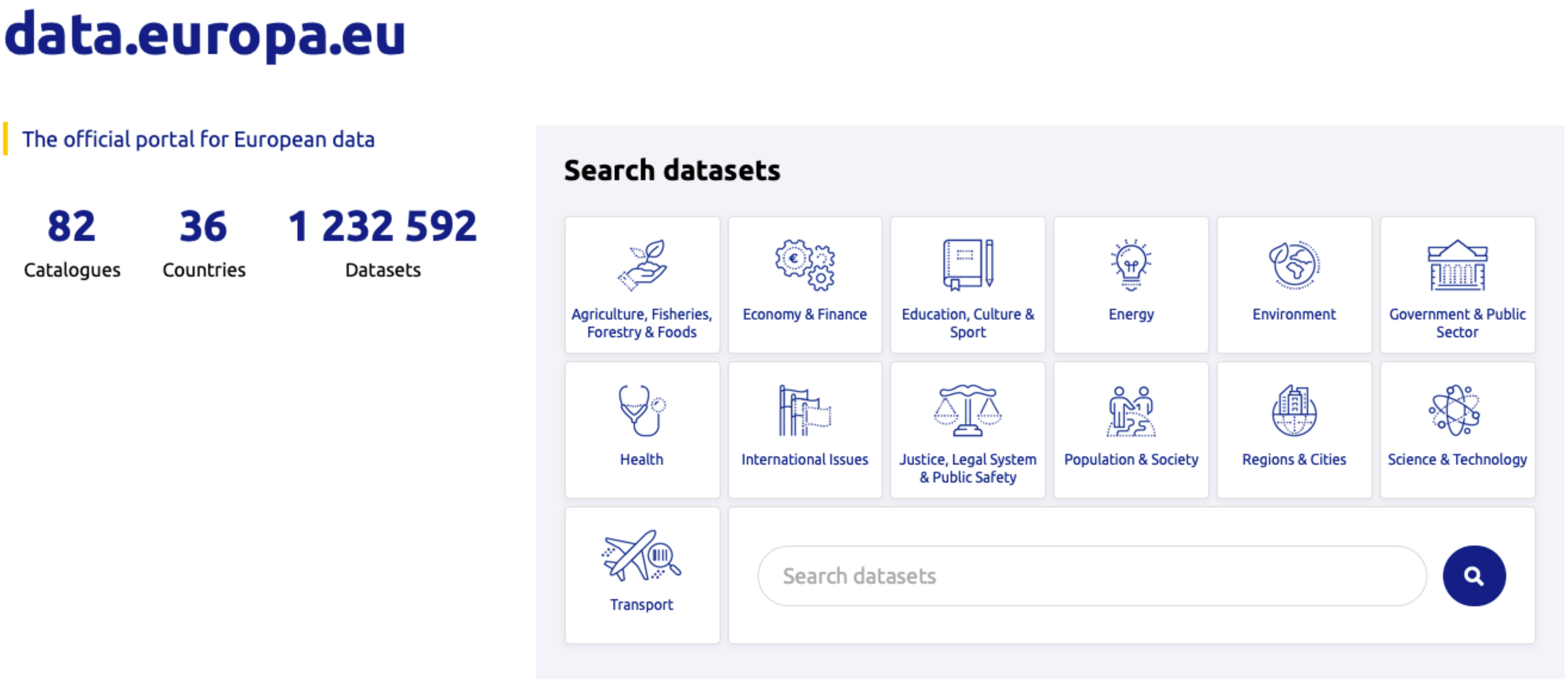
CSV RDF JSON XML

High Operational Temperature MWIR detectors with optical concentrators

Federal

National Aeronautics and Space Administration — The goal of this work is to develop high performance mid-wavelength (MWIR) barrier infrared detectors (BIRDs) operating at temperatures accessible to compact...

HTML HTML HTML HTML



data.europa.eu

The official portal for European data

82 Catalogues

36 Countries

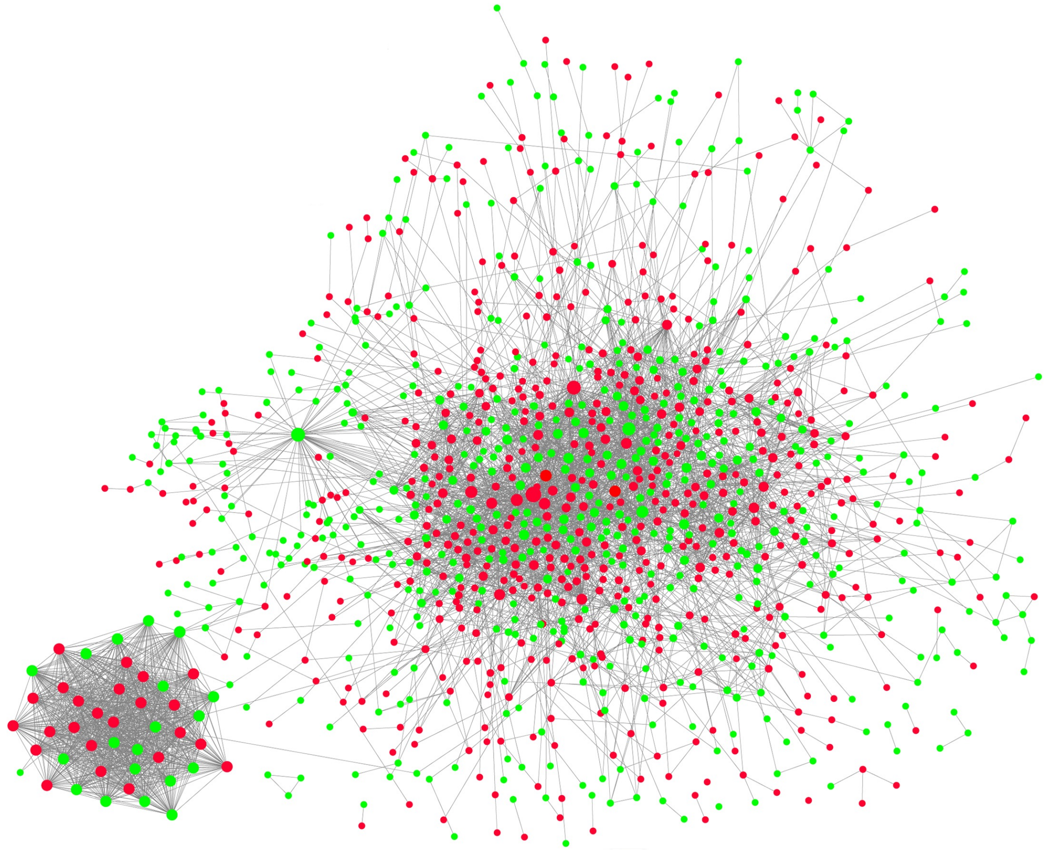
1 232 592 Datasets

Search datasets

Graph Data

- Lots of interesting data has a graph structure:
  - Social networks
  - Communication networks
  - Computer Networks
  - Road networks
  - Citations
  - Collaborations/Relationships
  - …

Some of these graphs can get quite large
(e.g., Facebook user graph; Protein interactic



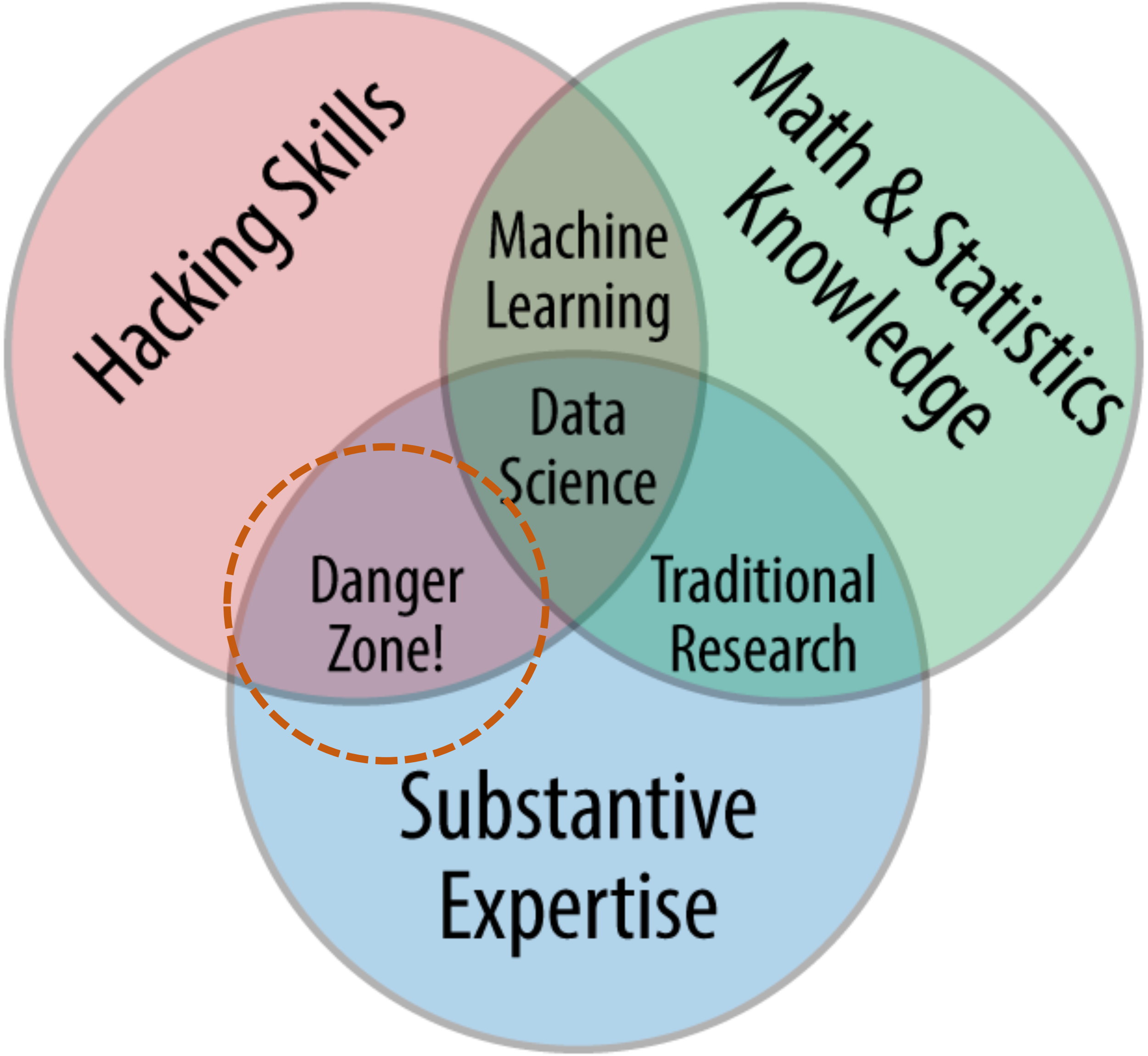
What is Data Science?

The future belongs to the companies and people that turn data into products

An O'Reilly Radar Report

By Mike Loukides

O'REILLY radar



Drew Conway’s Data Science Venn Diagram

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Applications

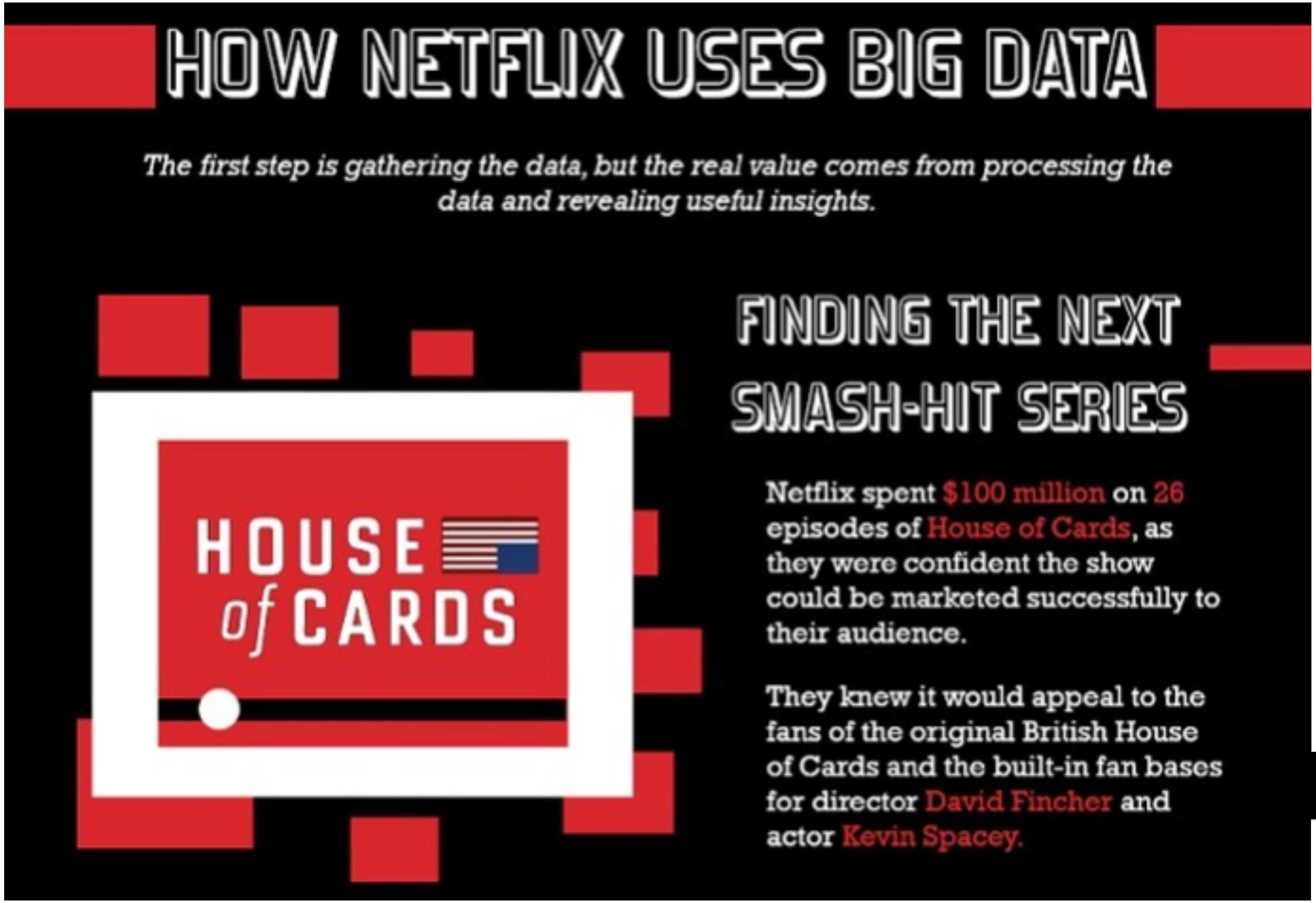
- Predictive maintenance
  - Predict equipment or component failures
  - Are systems/components in optimal operation
- Process optimization
  - Which parts of the process are causing delays?
  - What changes could improve production times?
- Medical research
  - Identification of biomarkers, therapeutic targets, new uses for existing pharmaceutical drugs

Applications

- Market analysis
    - Consumer profiling
    - Identifying buying patterns through time
    - Finding product associations
- Insurance fraud detection
    - Suspicious associations between road accidents, doctor/patients
    - Unnecessary lab exams

Success stories

- Shell
  - Predictive Maintenance: “leverage over a trillion data points through machine learning to detect anomalies…. predicting potential malfunctions of critical equipment, which allows preventive actions. ”
  - Inventory Optimisation: “apply AI to historical inventory data to optimise inventory stock levels.
- NHS
  - “a single integrated data, analysis and modelling platform … enabled … short-term forecasts, supply management, … , anticipate pressures and make best use of resources.”
- PepsiCo
  - “company’s clients provide reports that include their warehouse inventory and the POS inventory to the company, and this data is used to reconcile and forecast the production and shipment needs”



HOW NETFLIX USES BIG DATA

The first step is gathering the data, but the real value comes from processing the data and revealing useful insights.

FINDING THE NEXT SMASH-HIT SERIES

HOUSE of CARDS

Netflix spent $100 million on 26 episodes of House of Cards, as they were confident the show could be marketed successfully to their audience.

They knew it would appeal to the fans of the original British House of Cards and the built-in fan bases for director David Fincher and actor Kevin Spacey.

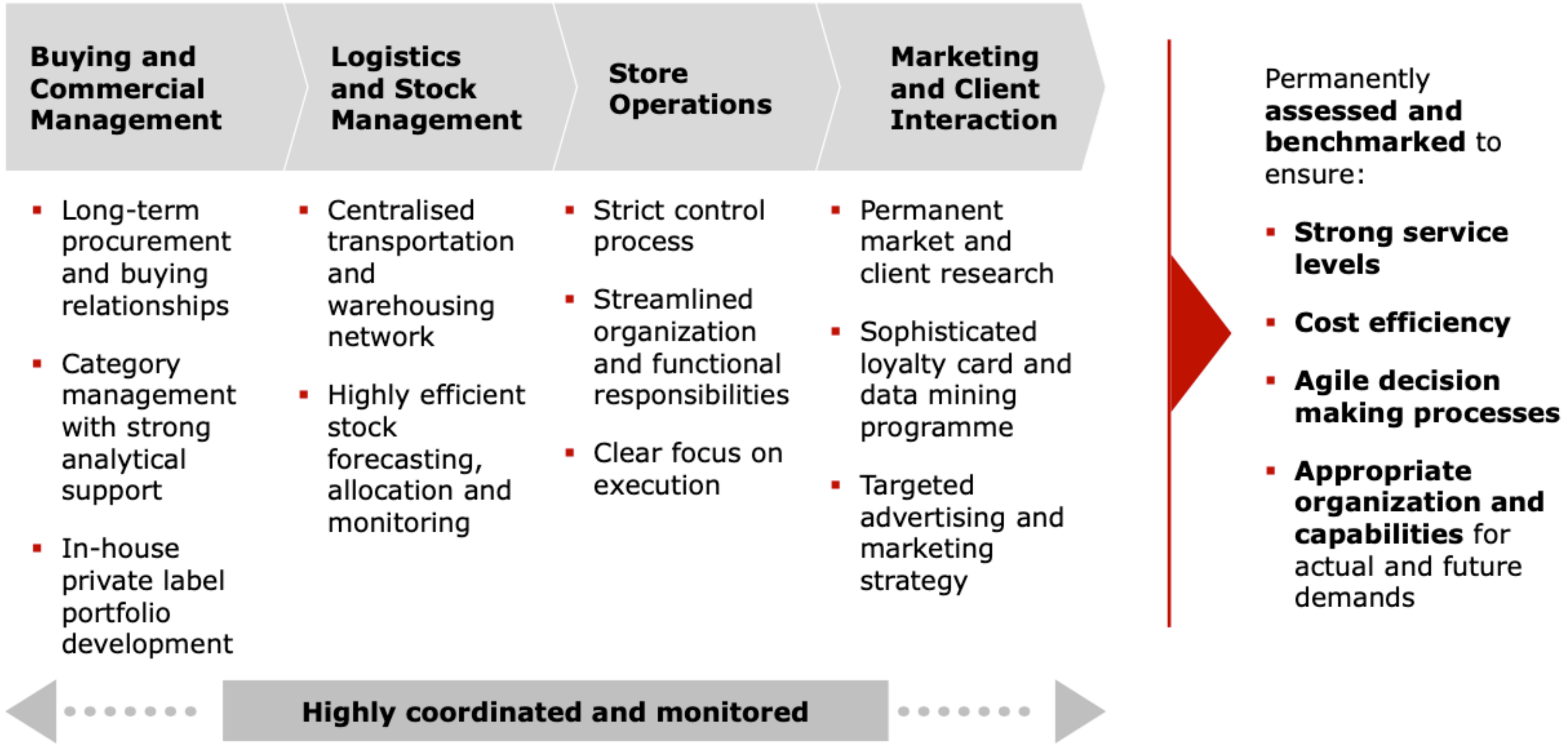
> “Big Data is not about the Data. It is about the Analytics.”

Gary King

Professor at Harvard University

Deep Culture of Efficiency and Innovation, with Best-in-Class Supply Chain Capabilities

Key strategic pillars of Sonae MC's supply chain



Buying and Commercial Management

- Long-term procurement and buying relationships
- Category management with strong analytical support
- In-house private label portfolio development

Logistics and Stock Management

- Centralised transportation and warehousing network
- Highly efficient stock forecasting, allocation and monitoring

Store Operations

- Strict control process
- Streamlined organization and functional responsibilities
- Clear focus on execution

Marketing and Client Interaction

- Permanent market and client research
- Sophisticated loyalty card and data mining programme
- Targeted advertising and marketing strategy

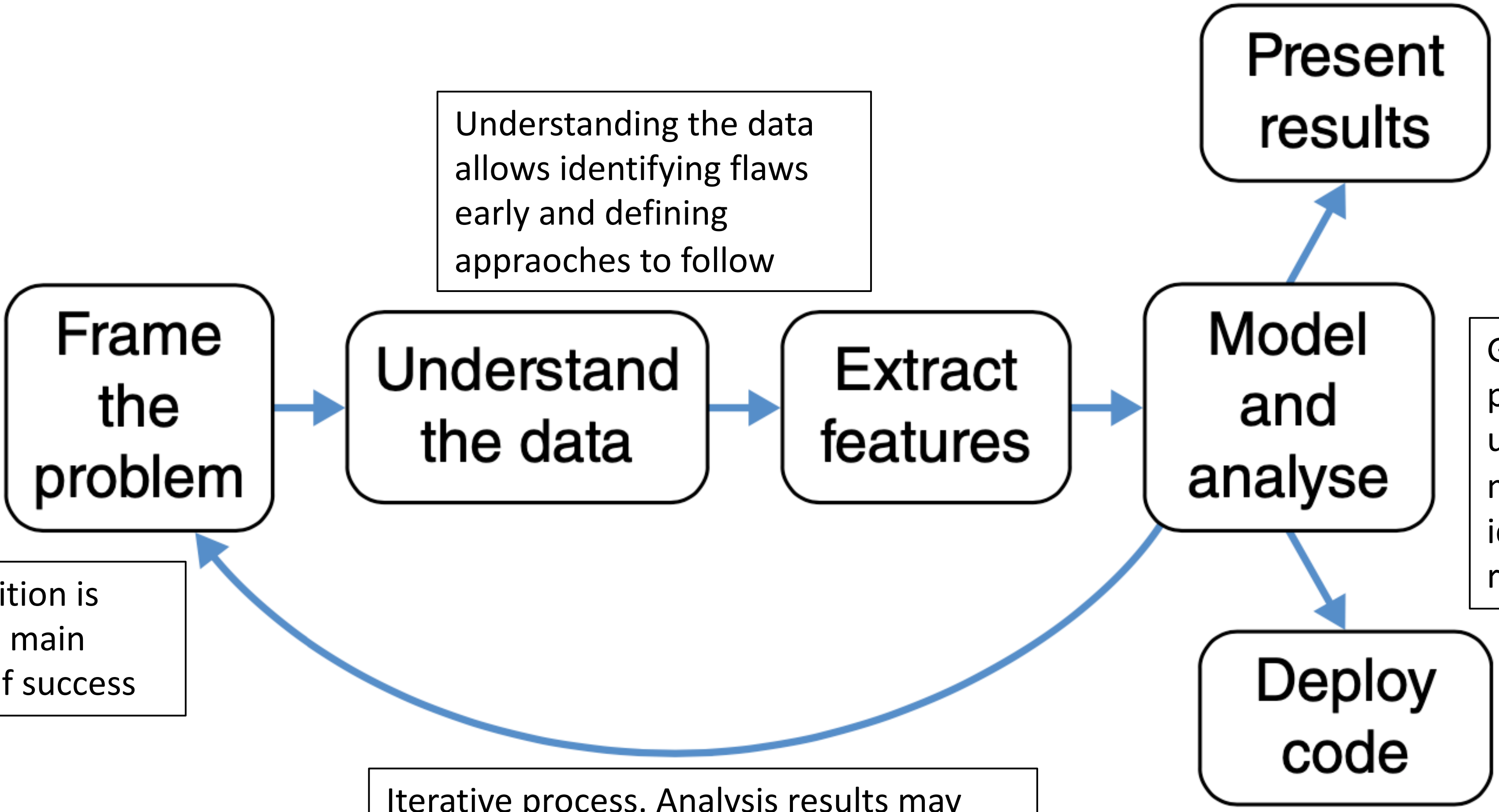
Highly coordinated and monitored

Permanently **assessed and benchmarked** to ensure:

- **Strong service levels**
- **Cost efficiency**
- **Agile decision making processes**
- **Appropriate organization and capabilities** for actual and future demands

Source: Company information

The data science roadmap



Understanding the data allows identifying flaws early and defining appraoches to follow

Getting quick preliminary results, using simplet methods, may help identifying flaws and redifining the problem

Problem definition is frequently the main determinant of success

Iterative process. Analysis results may reveal new ways of exploiting the data

Image: The Data Science Handbook, First Edition. Field Cady.