

Data Science : An Introduction

Demystifying the “buzzword” data science

Fundamental Qs

- # 1 Why Data Science Matters?

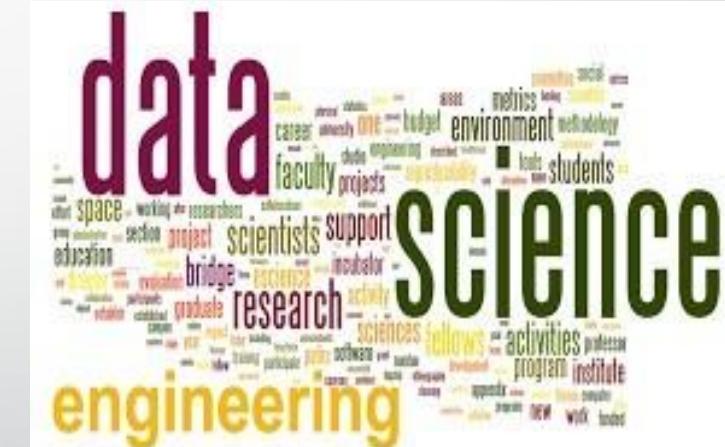
- ## 2 | What is Data Science?

- ## 3 Who does Data Science?

- ## 4 Where is Data Science Used?

- # 5 When is Data Science Applied?

- ## 6 How Data Science Works?



Learning Objectives

1. To describe **why** data science matters
2. To construct a definition for data science (the **what**)
3. To identify **who** data scientist are
4. To determine **where** and **when** data science is used / applied
5. To summarize **how** data science works



What ONE word that appears in your mind when you hear the term “Data Science”?

Pondering Questions



Q1 - Is Data Science new?

Q2 - Is Data Science the same as Statistics and Analysis?

Q3 - Do you need big data to do Data Science?

Q1 - Is Data Science New?

1962, JW Tukey, “*The Future of Data Analysis*”

“For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt...I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”

1997, C.F. Jeff Wu gave an inaugural lecture titled simply “[Statistics =Data Science?](#)”

2008. The term “data scientist” is often attributed to **Jeff Hammerbacher** (now, founder and Chief Scientist of Cloudera) and DJ Patil.

Brief History of Data Science

6th C BC - 1st C BC – **The Greeks!** Pyrrhonism, Skepticism & Empiricism...

1966 – Peter Naur @UoC Datalogy & Data Science

2001 – William S. Cleveland  "Data Science: An Action Plan" [@Purdue University](#)

2002 – Committee on Data for Science & Technology (CODATA)

2003 – Journal of Data Science

2009 – Jeff Hammerbacher @ Facebook What does a Data Scientist Do?

2010 – Drew Conway @NYU The Data Science Venn Diagram

2010 – Hillary Mason & Chris Wiggins @Dataists "

2010 – Mike Loukidis @O'Reilly "What is Data Science?"

2011 – DJ Patil @LinkedIn data scientist vs. data analyst

International Statistical Review (2001), 69, 1, 21–26, Printed in Mexico
© International Statistical Institute

Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland

Statistics Research, Bell Laboratories, 600 Mountain Avenue, Murray Hill NJ07974, USA
E-mail: wsc@research.bell-labs.com

Summary

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department, and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

Key words: Future; Applications; Computing; Methods; Models; Theory.

The Science of Datalogy

EDITOR:

This is to advocate that the following new **words**, denoting various aspects of our subject, be considered for general adoption (the stress is shown by an accent):

datalogy, the science of the nature and use of data,

datum&tics, that part of datalogy which deals with the processing of data by automatic means,

datamaton, an automatic device for processing data.

In this terminology much of what is now referred to as "data processing" would be datamatics. In many cases this will be a gain in clarity because the new word includes the important aspect of data representations, while the old one does not. Datalogy might be a suitable replacement for "computer science."

The objection that possibly one of these words has already been used as a proper name of some activity may be answered partly by saying that of course the subject of datamatics is written with a lower case d, partly by remembering that the word "electronics" is used doubly in this way without inconvenience.

What also speaks for these words is that they will transfer gracefully into many other languages. We have been using them extensively in my local environment for the last few months and have found them a great help.

Finally I wish to mention that datamatics and datamaton (Danish: datamatik and datamat) are due to Paul Lindgreen and Per Brinch Hansen, while datalogy (Danish: datalogi) is my own invention.

PETER NAUR
A/S Regnecentralen
Falkoner Alle 1
Copenhagen F, Denmark

The science of datalogy

Full Text:  [PDF](#)

Author: [Peter Naur A/S Regnecentralen, Copenhagen, Denmark](#)

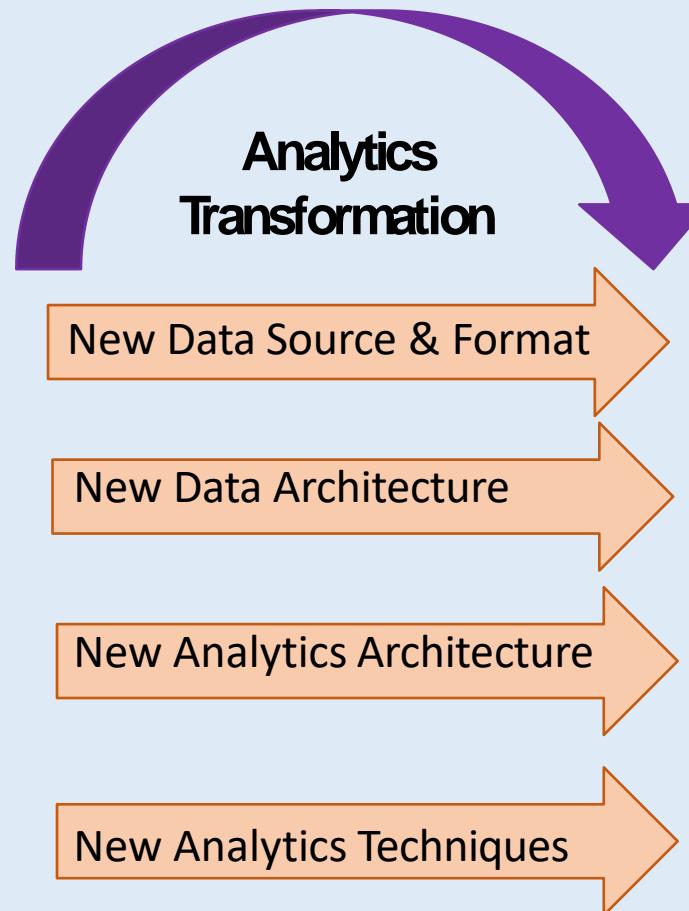
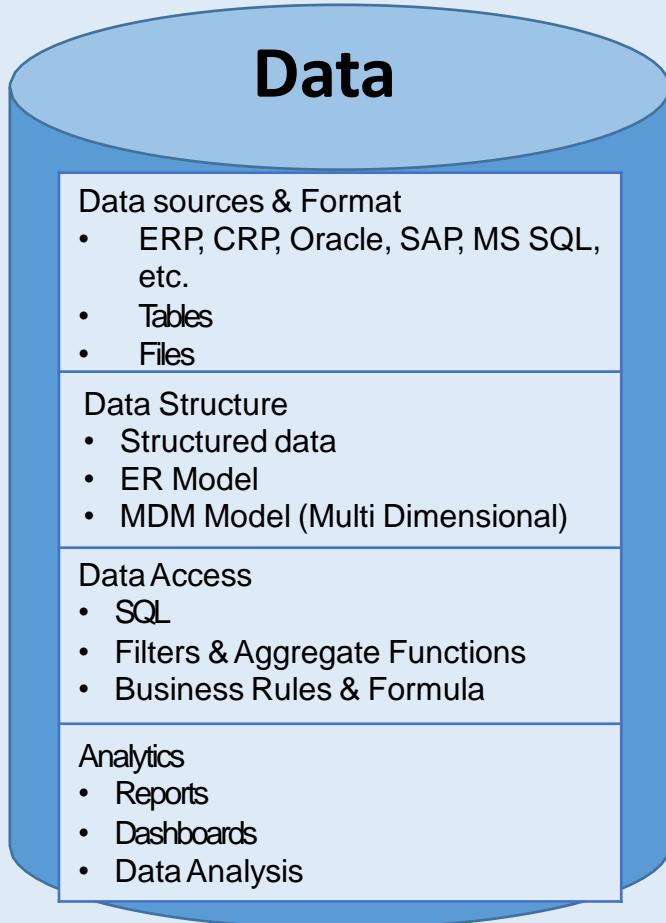
Published in:



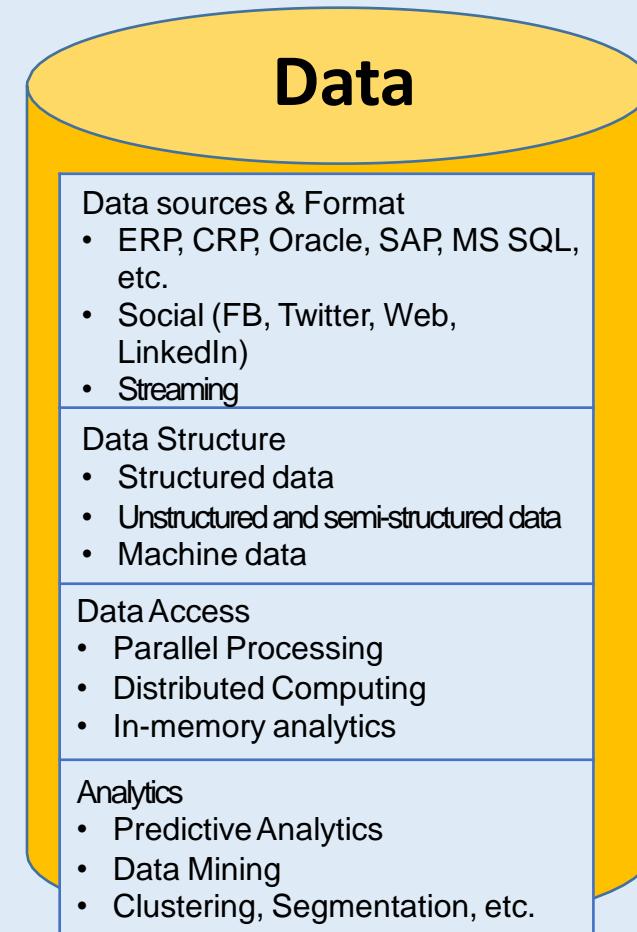
- Magazine
[Communications of the ACM CACM Homepage archive](#)
Volume 9 Issue 7, July 1966
Page 485
[ACM New York, NY, USA](#)
[table of contents](#) doi:>[10.1145/365719.366510](https://doi.org/10.1145/365719.366510)

Keyword: Transformation

Traditional Approach to Data & Analytics



Modern Approach to Data & Analytics = Data Science



Modernizing existing reporting, data management, data warehousing, analytics and BI solutions

New thinking, new ideas, innovation
→ new ways of solving problems



In 2008, **Dr DJ Patil** and **Jeff Hammerbacher**, heads of analytics and data at LinkedIn and Facebook respectively, coined the term '**data science**' to describe the emerging field of study that focused on teasing out the hidden value in the data that was being collected from **touchpoints** all over the retail and business sectors.

Top data scientist D J Patil's Tips to Build a Career in Data

<https://www.youtube.com/watch?v=UuAJMzpoq5E>

Q2 - Is it the same as Statistics or Analysis?

"I think data-scientist is a sexed up term for a statistician.... Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn't berate the term statistician."

Nate Silver, Statistician

Analyzing data is something people have been doing with statistics and related methods for a while.

It's all about the difference between explaining and predicting.

Data analysis has been generally used as a way of explaining some phenomenon by extracting interesting patterns from individual data sets with well-formulated queries.

Data science, on the other hand, aims to discover and extract actionable knowledge from the data, that is, knowledge that can be used to make decisions and predictions, not just to explain what's going on.

- **Statistical analysis** is used in order to gain an understanding of a larger population by analyzing the information of a sample. Statistical analysis allows inferences to be drawn about target markets, consumer cohorts and the general population by expanding findings appropriately to predict the behavior and characteristics of the many based on the few.
- **Data analysis** is the process of inspecting, presenting and reporting data in a way that is useful to non-technical people. Because data is next to useless if it can't be understood by the decision-makers who need to use it, data analysts act as translators between the numbers and figures and the people who need to know about them.

Statistics is part of Data Science.

Analysis is part of Data Science.

Q3 - Do you need Big Data to do Data Science?

BY VASANT DHAR

Data Science and Prediction

USE OF THE term “data science” is increasingly common, as is “big data.” But what does it mean? Is there something unique about it? What skills do “data scientists” need to be productive in a world deluged by data? What are the implications for scientific inquiry?

including our confidence in the inference. Why then do we need a new term like data science when we have had statistics for centuries? The fact that we now have huge amounts of data should not in and of itself justify the need for a new term.

The short answer is data science is different from statistics and other existing disciplines in several important ways. To start, the raw material, the “data”

The differences between small, medium, and Big data (Michael E Driscoll)

| class | size | manage with | how it fits | examples |
|---------------|----------|---------------------------------|------------------------------|----------------------------|
| small | < 10 GB | Excel, R | fits in one machine's memory | thousands of sales figures |
| medium | 10GB-1TB | indexed files, monolithic DB | fits on one machine's disk | millions of web pages |
| Big | > 1TB | Hadoop, distributed DBs | stored across many machines | billions of web clicks |

Data Scenario

Society is becoming **increasingly reliant on data** and the **tools and methods** to acquire and analyze it. Important sources of data that are only starting to be explored come from:

- social media,
- full-text scientific literature,
- video material,
- click and interaction patterns,
- financial transactions,
- customer behavior,
- sensors and scientific instrumentation.

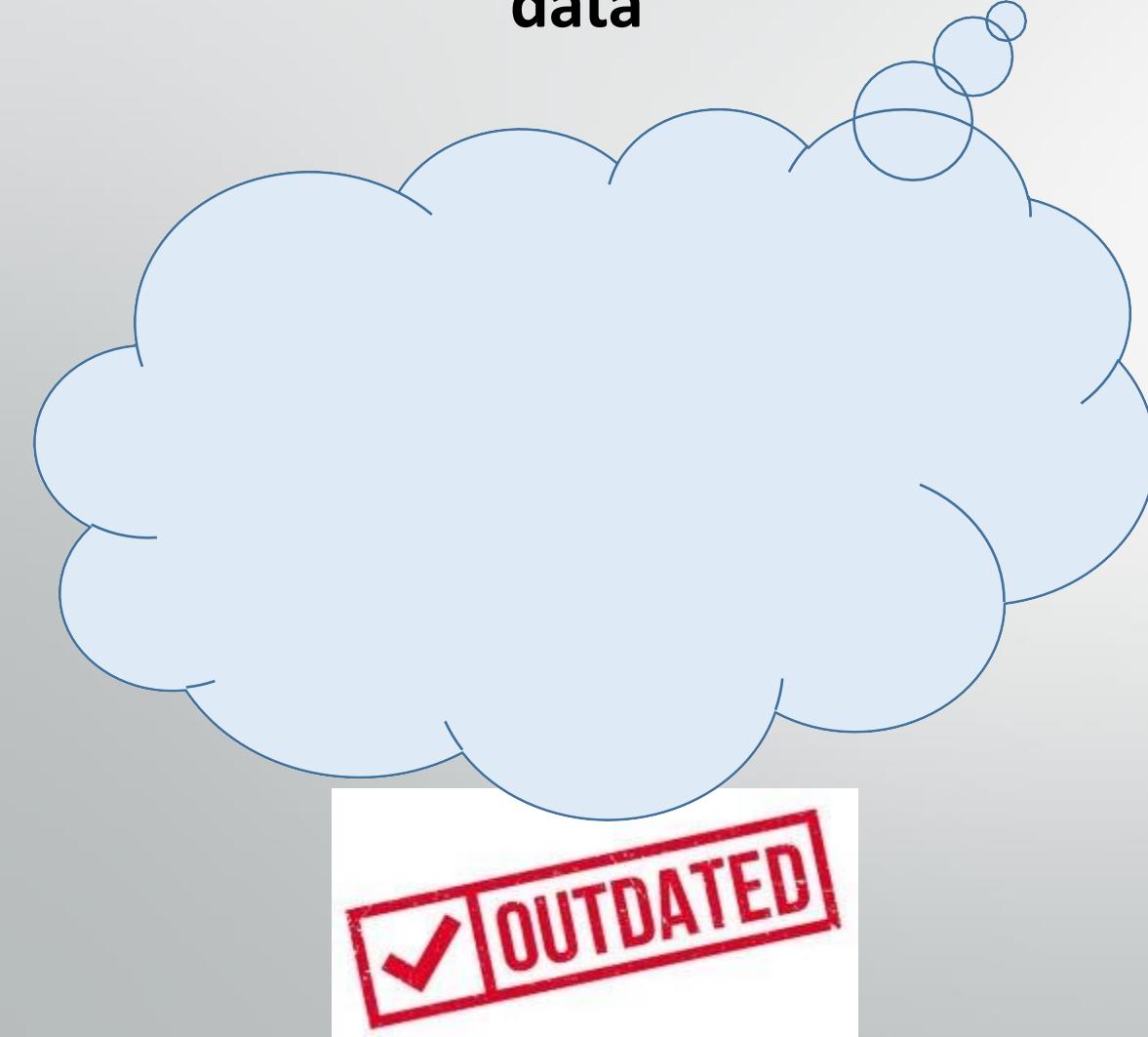


We can only fully utilize data if we have the methodologies to **store, process and transform** it into valuable and **accessible** information by **analysis and modelling**.

Being able to **fully utilize data** will, for example, improve healthcare quality, government efficiency, military effectiveness, steer innovation in business, and forge new scientific discoveries.

Data are Everywhere

- When people think of data



Data is much more **ubiquitous**.

It is the **by-product** of any and every action, pervading every part of our lives.

Wherever we go, whatever we buy, whatever interests we have, this data is all being collected and remodeled into trends that help advertisers and marketers push their products to the right people.

'Data are not taken for museum purposes; they are taken as a basis for doing something. If nothing is to be done with the data, then there is no use in collecting any. The ultimate purpose of taking data is to provide a basis for action or a recommendation for action.'

W. Edwards Deming, 1942

Companies Collect a Lot of Data, But How Much Do They Actually Use?

By Priceonomics Data Studio · 3,200 views · More stats

For all the talk of how data is the new oil and the most valuable resource of any enterprise, there is a deep dark secret companies are reluctant to share — most of the data collected by businesses simply goes unused.

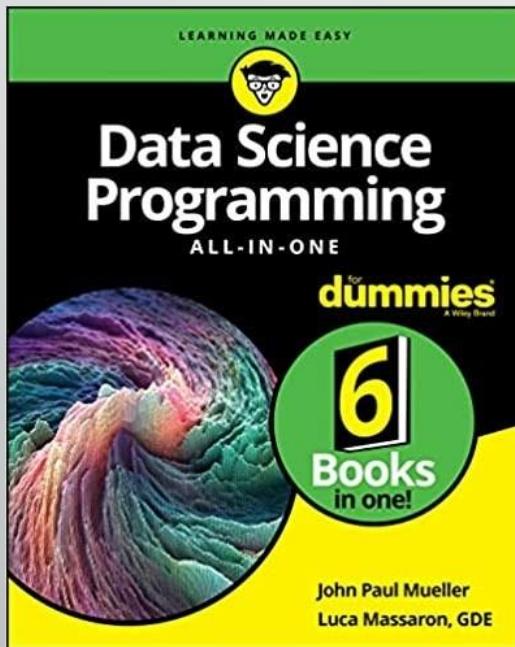
This unknown and unused data, known as **dark data** comprises more than half the data collected by companies. Given that some estimates indicate that 7.5 septillion (7,700,000,000,000,000,000) gigabytes of data are generated every single day, not using most of it is a considerable issue.

Gartner defines **dark data** as:

“The information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing).

<https://priceonomics.com/companies-collect-a-lot-of-data-but-how-much-do/>

Making life more interesting in other ways

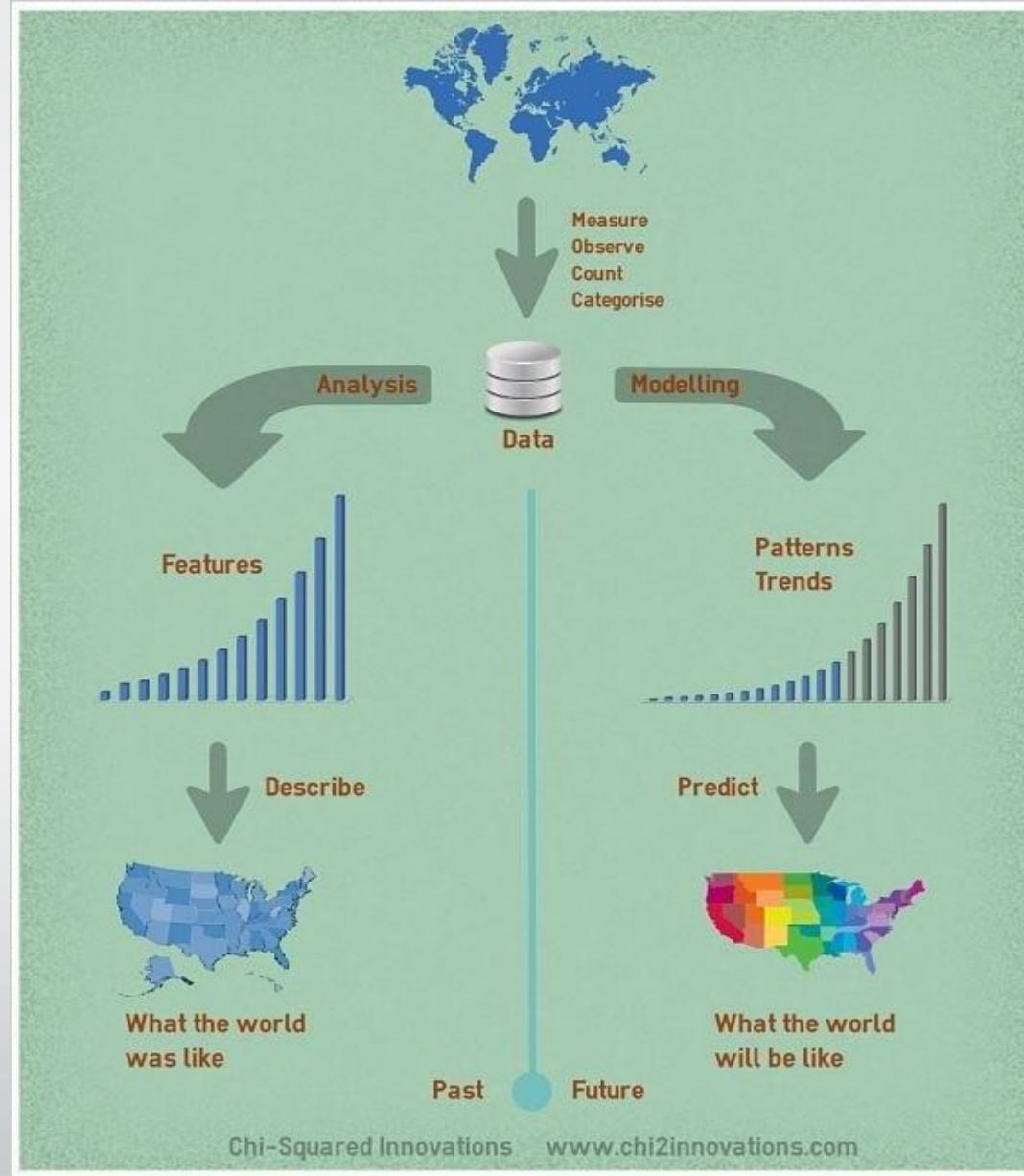


Data is part of your life. You really can't perform too many activities anymore that don't have data attached to them in some way. For example, consider **gardening**. You might think that digging in the earth, planting seeds, watering, and harvesting fruit has nothing to do with data, yet the seeds you use likely rely on research conducted as the result of gathering data. The tools you use to dig are now ergonomically designed based on human research studies. The weather reports you use to determine whether to water or not rely on data. The clothes you wear, the shoes you employ to work safely, and even the manner in which you work are all influenced by data. Now, consider that gardening is a relatively nontechnical task that people have performed for thousands of years, and you get a good feel for just how much data affects your daily life.

What do we do with the data?

We typically collect data to answer one of 2 questions:

- What is the world like?
- What is the world going to be like?



The Groceries Dataset

| A | B | C | D | E | F | G | H | I | J | K | L |
|------|-------------------|--------------------|---------------|---------------|---------------|-------------|--------------------------|-----------|-----------|------------------------|---|
| 1545 | pastry | long life bakery p | shopping bags | | | | | | | | |
| 1546 | sausage | pip fruit | whipped/so | detergent | dental care | | | | | | |
| 1547 | processed cheese | red/blush wine | newspapers | | | | | | | | |
| 1548 | napkins | | | | | | | | | | |
| 1549 | meat | whole milk | whipped/so | oil | pickled ve | long life b | napkins | | | | |
| 1550 | pastry | | | | | | | | | | |
| 1551 | sausage | root vegetables | other vegeta | whole milk | dessert | canned fr | soda | snack pro | waffles | shopping bags | |
| 1552 | pip fruit | root vegetables | whole milk | ice cream | bottled be | detergent | | | | | |
| 1553 | frozen vegetables | | | | | | | | | | |
| 1554 | citrus fruit | tropical fruit | coffee | | | | | | | | |
| 1555 | other vegetables | yogurt | whipped/so | soda | bottled be | salty snack | | | | | |
| 1556 | frankfurter | frozen dessert | white wine | chewing gum | shopping bags | | | | | | |
| 1557 | sausage | | | | | | | | | | |
| 1558 | beef | other vegetables | yogurt | beverages | rolls/buns | flour | margarine | cat food | bottled w | fruit/vegetable juice | |
| 1559 | hamburger meat | other vegetables | whipped/so | pastry | napkins | | | | | | |
| 1560 | yogurt | margarine | sugar | bottled water | bottled be | red/blush | long life bakery product | | | | |
| 1561 | beverages | | | | | | | | | | |
| 1562 | whole milk | | | | | | | | | | |
| 1563 | rolls/buns | | | | | | | | | | |
| 1564 | tropical fruit | bottled water | chewing gum | | | | | | | | |
| 1565 | citrus fruit | herbs | butter | domestic eggs | oil | bottled w | bottled beer | | | | |
| 1566 | whole milk | yogurt | whipped/so | soft cheese | candy | | | | | | |
| 1567 | citrus fruit | tropical fruit | root vegetab | yogurt | roll produ | margarine | sugar | bottled w | napkins | house keeping products | |

Groceries Data Set contains: a collection of receipts with each line representing 1 receipt and the items purchased. Each line is called a **transaction** and each column in a row represents an **item**.

Market Basket Use-Case

Domain of the Dataset:

Products and Retail. However, the application of the algorithm is not limited to only Products and Retail. The technique can be applied wherever we want to discover the co-occurrence relationship amongst various activities.

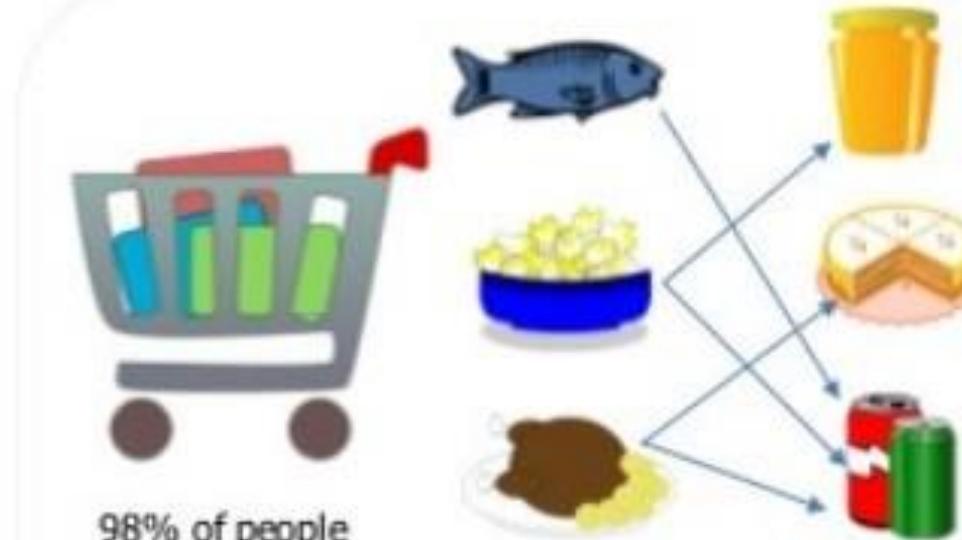
Problem Statement:

Market Basket Analysis.

A retail outlet wants understand the purchase behavior of a buyer. This information will enable the retailer to understand the buyer's needs.

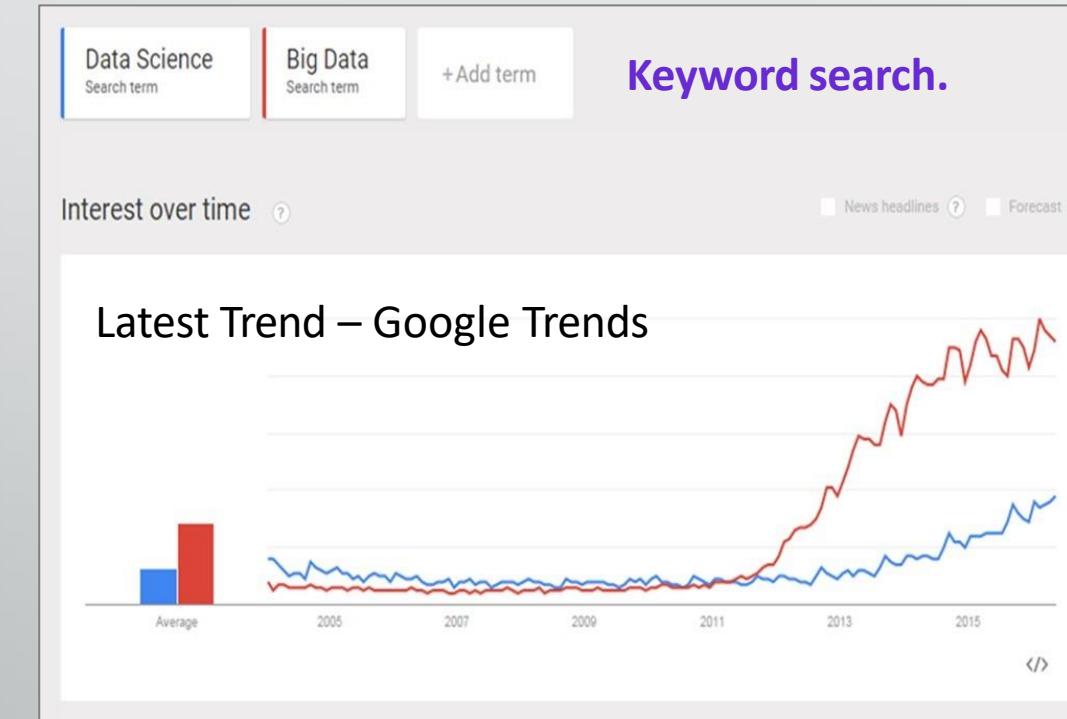
The analysis might tell a retailer that customers often purchase shampoo and conditioner together, so putting both items on promotion at the same time would create a significant increase in profit, while a promotion involving just one of the items would likely drive sales of the other.

Market Basket Analysis



98% of people who purchased items A and B also purchased item C

WHY Data Science?



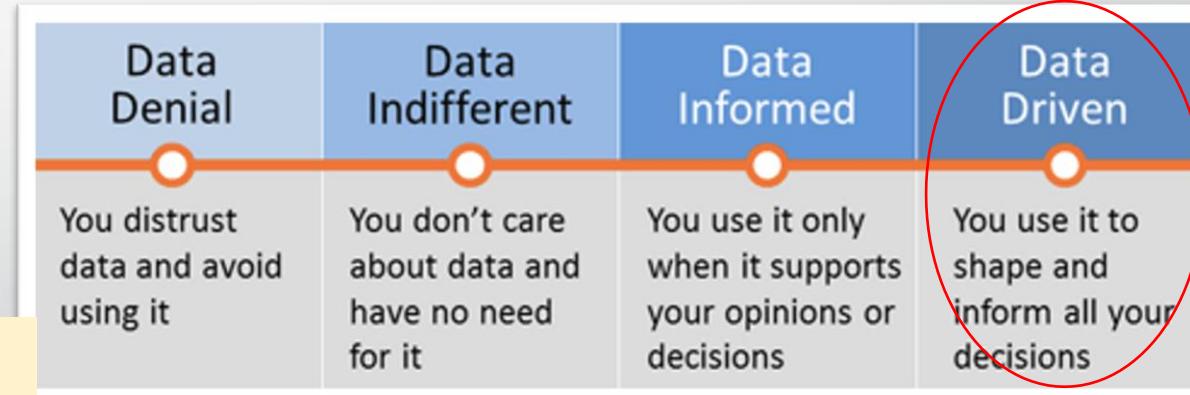
NEED: the future is increasingly complex and *difficult to predict*.

NEED: we don't have enough *qualified experts*, and experts often get it wrong.

RAW MATERIALS: we are generating *huge amounts of data* at an increasing rate.

ENABLER: new *hardware and software tools* are emerging.

THEREFORE: Data science is *inevitable!* We don't have a choice.



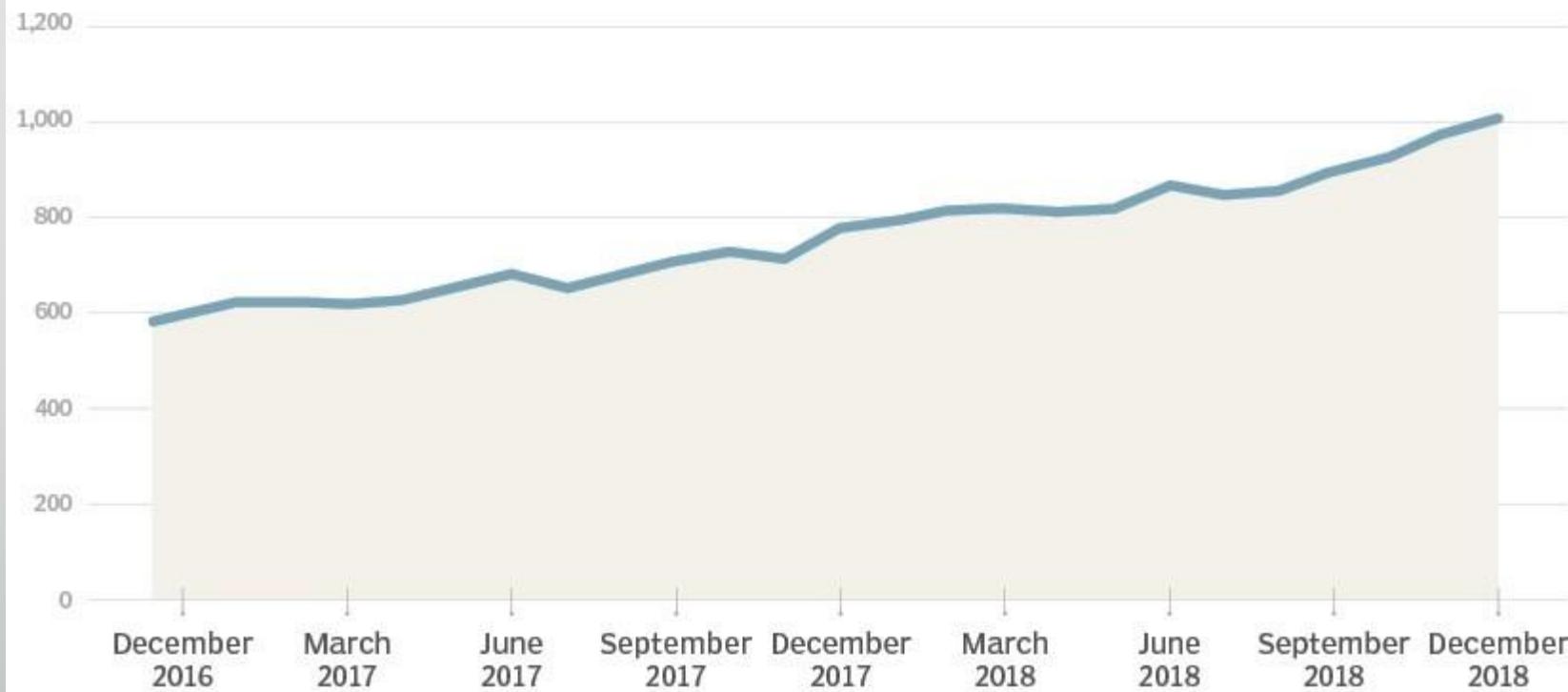
Data-driven decision making

↑ Accuracy

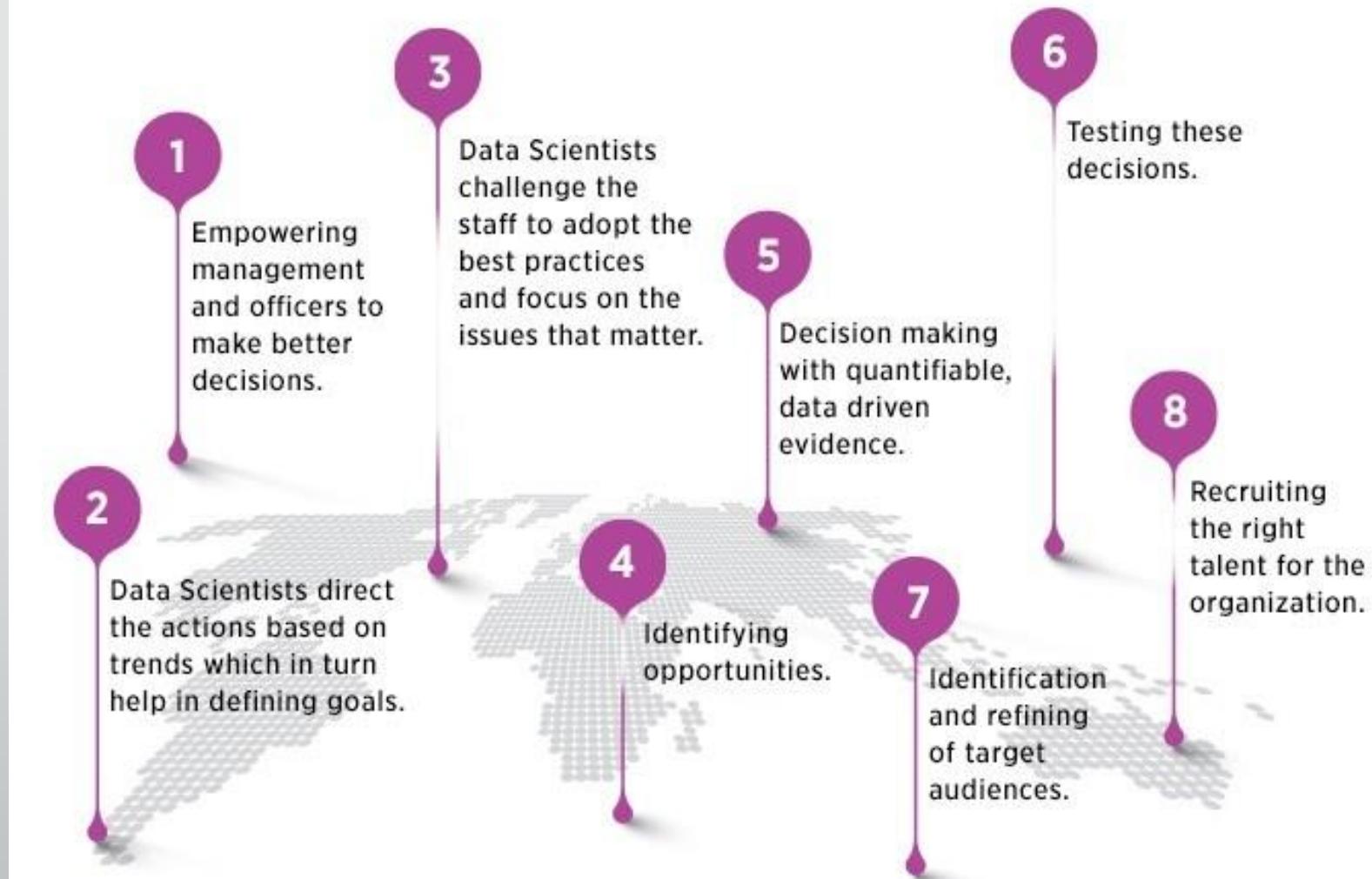
↑ Efficiency

Data scientists are in high demand

Data scientist job postings, per 1 million postings on Indeed



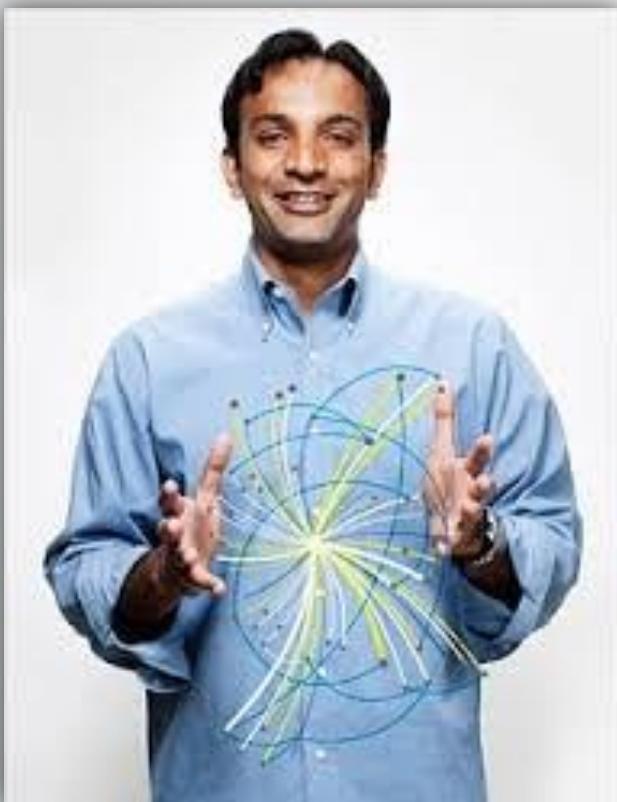
The growth in data scientist job postings on Indeed, from December 2016 to December 2018



Source: <https://www.simplilearn.com/why-and-how-data-science-matters-to-business-article>

Video

"Data Science: Where are We Going?" - Dr. DJ Patil (12:59 minutes)
https://www.youtube.com/watch?v=3_1reLdh5xw



USA first Chief Data Scientist

Former U.S. Chief Data Scientist

A screenshot of a LinkedIn profile card. It features a logo consisting of three overlapping colored shapes (orange, blue, and purple) on the left. To the right of the logo, the text reads: "Head Of Technology" followed by "Devoted Health". Below that, it says "Sep 2017 – Present · 1 yr 1 mo" and "San Francisco Bay Area & Boston".

Head Of Technology
Devoted Health
Sep 2017 – Present · 1 yr 1 mo
San Francisco Bay Area & Boston

Data – The Raw Materials

DATA - Facts and statistics collected together for reference or analysis. Watch the video on
“What is Data?” <https://www.youtube.com/watch?v=EMHP-q4GEDc>

“Data –Information – Knowledge” <https://www.youtube.com/watch?v=QsP5WGv0aQc>

SCIENCE - A systematic study through observation and experiment.

What is Science? <https://www.youtube.com/watch?v=hDQ8ggroeE4>

How does do science? | Figuring out what's true <https://www.youtube.com/watch?v=3MRHcYtZjFY>

DATA SCIENCE - The **scientific** exploration of **data** to extract meaning or insight, and the construction of software to utilize such insight in a business context.



Transform data into valuable insights
Transform data into data products
Transform data into interesting stories

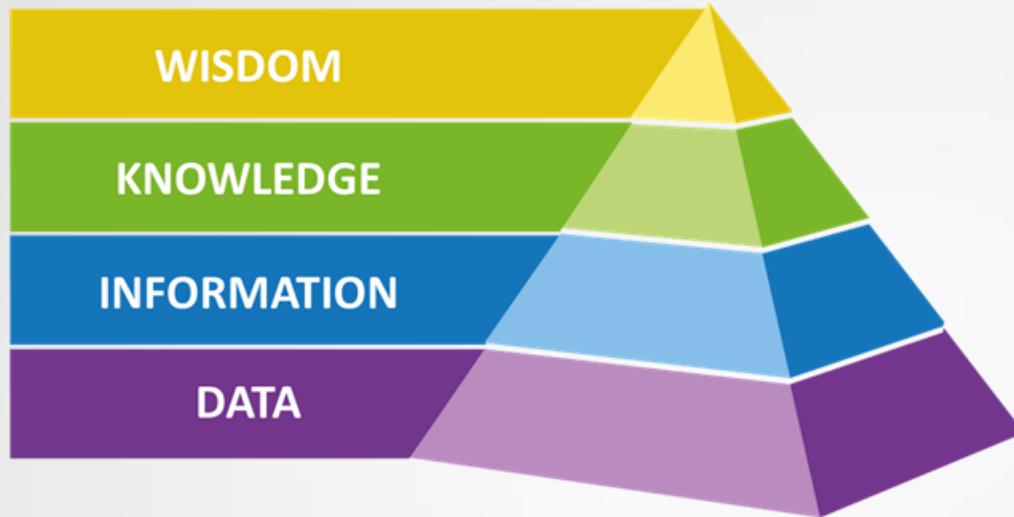


Table 2-1. Levels and definitions of data, information, knowledge, and wisdom (Ackoff, 1989).

| Concept | Definition |
|----------------|--|
| Data | Unprocessed facts and figures without interpretation or analysis |
| Information | Data given meaning that benefits the user |
| Knowledge | Combination of information, experience, and insight that benefits the user |
| Wisdom | Extrapolative and non-deterministic extension of knowledge |

WHAT is Data Science?

"We have lots of data – now what?"
(How can we unlock valuable insight from our data?)

The discipline of drawing useful conclusions from data

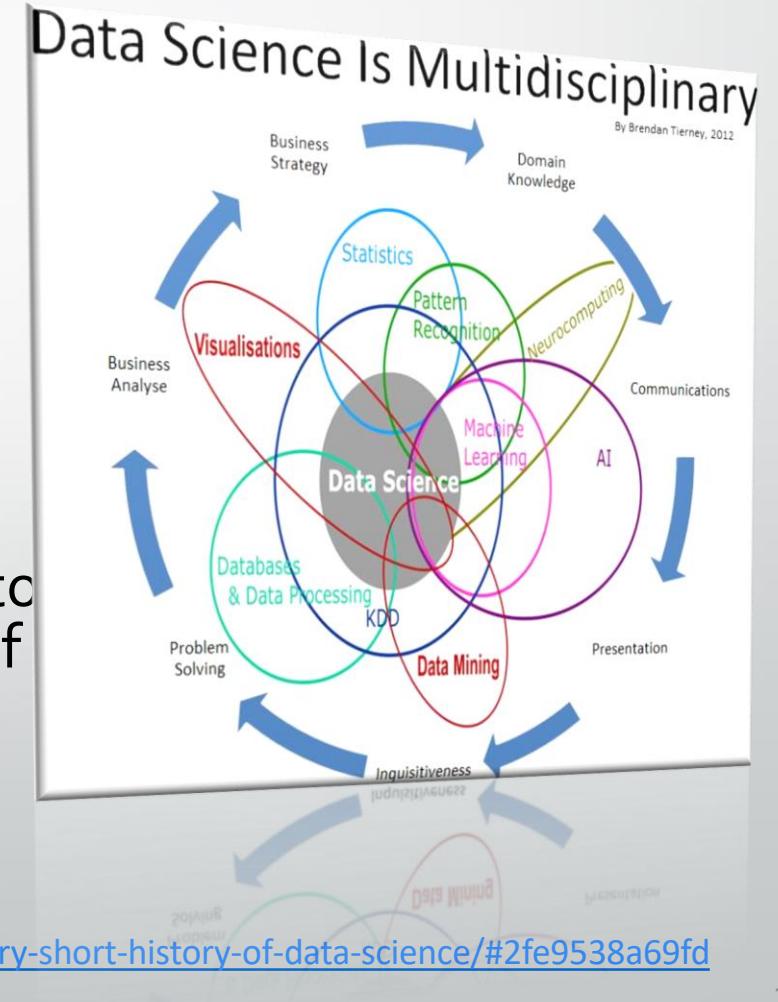
- using computation.
- It is geared toward helping individuals and organizations make better decisions from stored, consumed and managed data.
 - Data science is a broad field that refers to the collective processes, theories, concepts, tools and methods used to enable the review, analysis and extraction of information from raw data.

Data science is formerly known as **datalogy**.

Code of conduct for Data Science -
<http://www.datascienceassn.org/code-of-conduct.html>

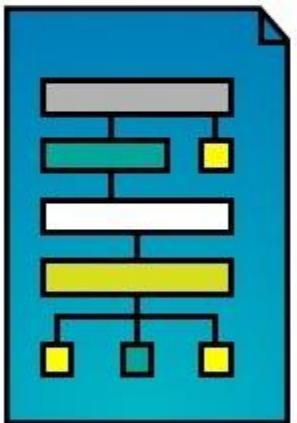
A Very Short History Of Data Science -

<http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#2fe9538a69fd>

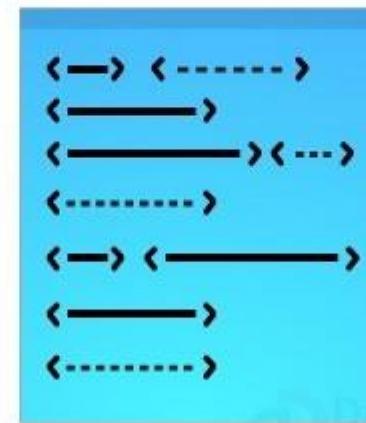


Purpose of Data Science

Extracting Data



Generating Insights from Data



Data Analysis & Processing

The Science of Data Science

The science to extract hidden values from any data by applying scientific, statistical, mathematical and computing techniques on it.

- Forming and testing hypothesis.
- Data science workflow / pipeline.
- Running data experiments to validate how certain we are of our interpretations of the results.

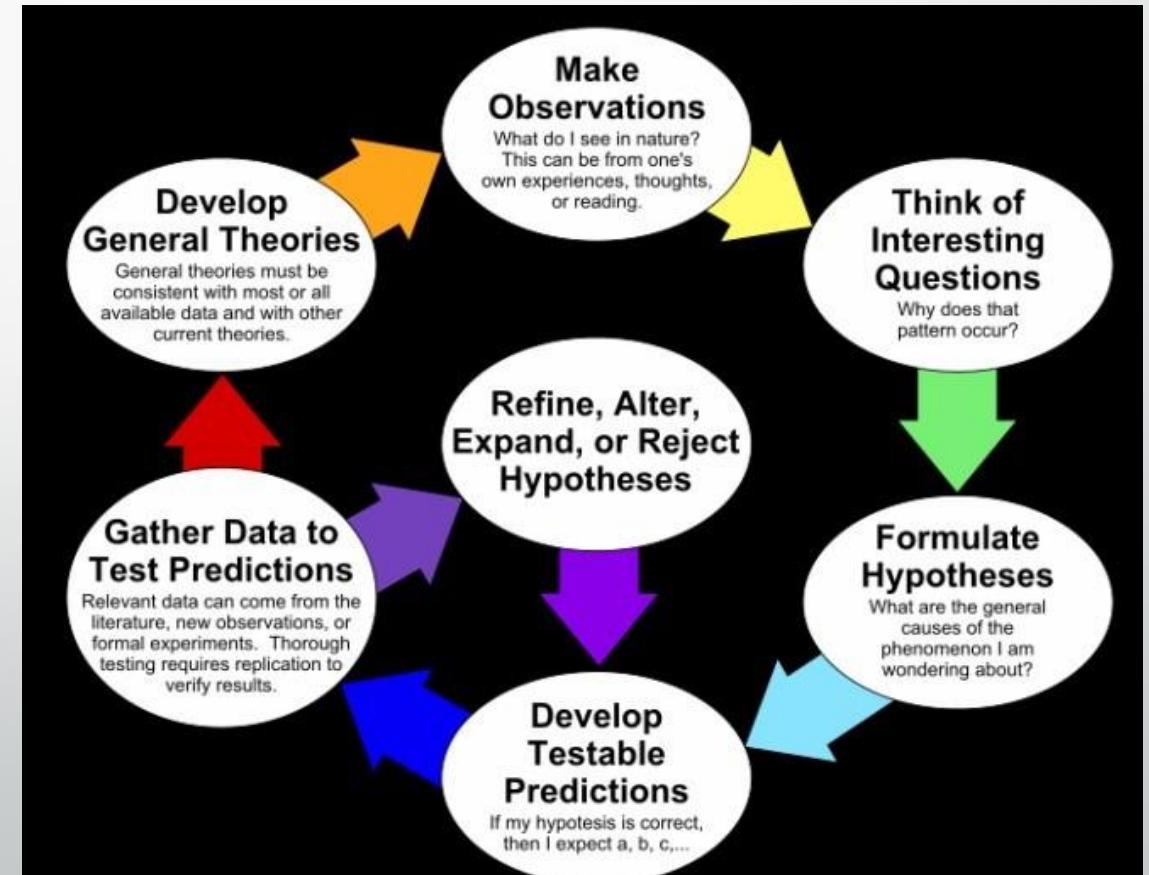
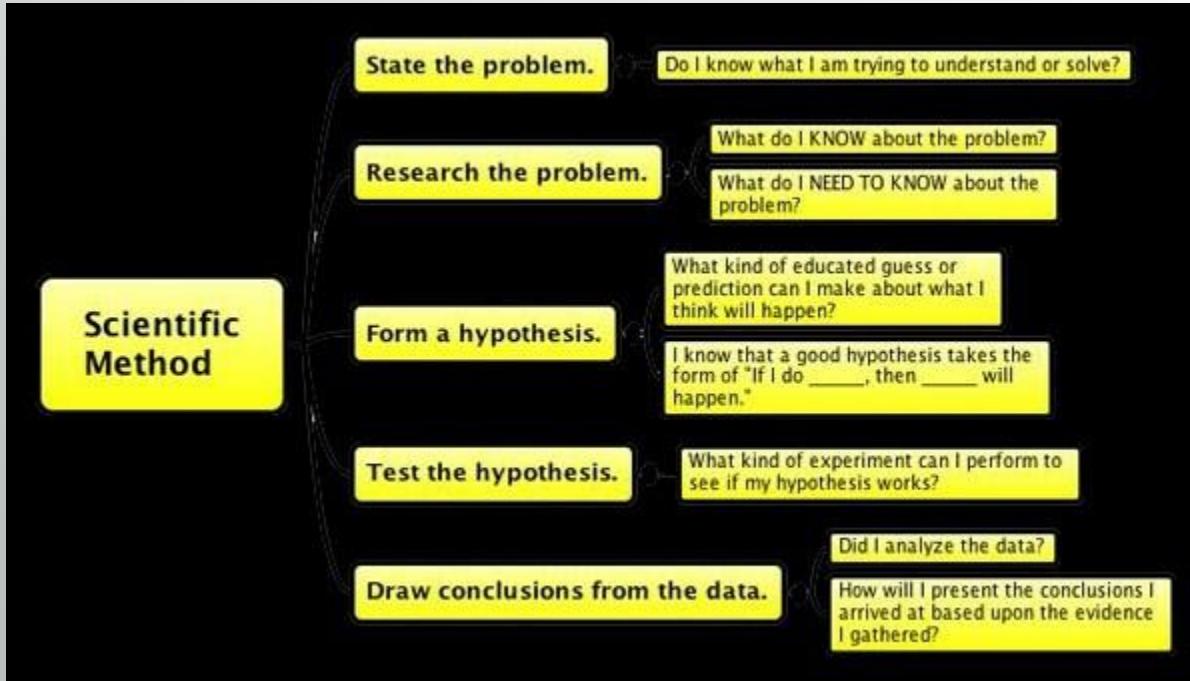
$$\begin{aligned} & \int d^3k e^{i\vec{k}\cdot\vec{x}} \frac{-G}{(2\pi)^{\frac{3}{2}}} \frac{1}{(k^2 + \mu^2)} \\ & \frac{-G}{(2\pi)^3} \int d^3k \frac{e^{i\vec{k}\cdot\vec{x}}}{(k^2 + \mu^2)} \\ & = \frac{-2\pi G}{(2\pi)^3} \int k^2 dk \int_{-1}^1 e^{ikr \cos \theta} \\ & = \frac{-G}{(2\pi)^2} \int \frac{k^2}{(k^2 + \mu^2)} \int_{-1}^1 e^{ikr \cos \theta} \\ & = \frac{-G}{(2\pi)^2} \int \frac{k^2}{(k^2 + \mu^2)} \left[\frac{1}{ikr} e^{ikr} - \frac{1}{ikr} e^{-ikr} \right] \\ & \frac{-G}{(2\pi)^2 ir} \int \frac{k}{(k^2 + \mu^2)} (e^{ikr} - e^{-ikr}) \\ & \approx \int_0^\infty \frac{k}{(k + i\mu)(k - i\mu)} (e^{ikr} - e^{-ikr}) \end{aligned}$$

The Art of Data Science

- Discovering what we don't know from data.
- Obtaining predictive, actionable insight from data.
- Creating Data Products that have business impact now.
- Communicating relevant business stories from data.
- Building confidence in decisions that drive business value.
- Feature Engineering
- Exploratory Data Analysis

The Science of Data Science

The **scientific method**: is a set of steps taken to ensure that conclusions are reached sensibly, experiments designed carefully, data is interpreted in accordance with the results of tests, and that procedures can be verified independently.



Core Aspects of Effective Data Analysis

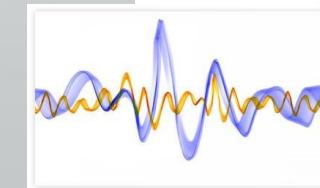
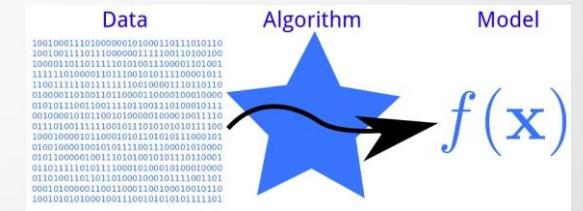
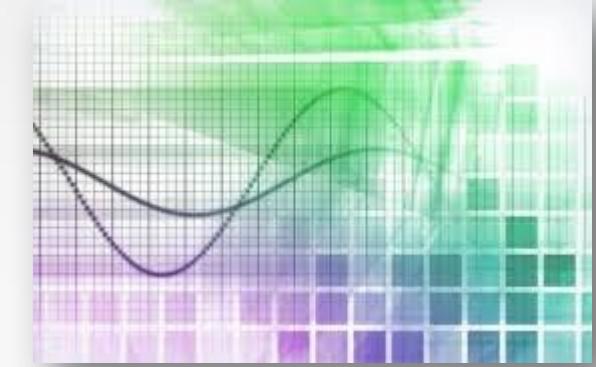
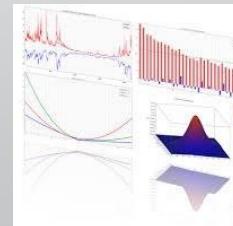
In order of difficulty: *Descriptive → Exploratory → Inferential → Predictive → Causal → Mechanistic*

- **Descriptive analysis** - describe set of data, interpret what you see (census, Google Ngram).
- **Exploratory analysis** - discovering connections (correlation does not = causation).
- **Inferential analysis** - use data conclusions from smaller population for the broader group.
- **Predictive analysis** - use data on one object to predict values for another (if X predicts Y, does not = X cause Y).
- **Causal analysis** - how does changing one variable affect another, using randomized studies, Strong assumptions, golden standard for statistical analysis.
- **Mechanistic analysis** - understand exact changes in variables in other variables, modeled by empirical equations (engineering/physics).

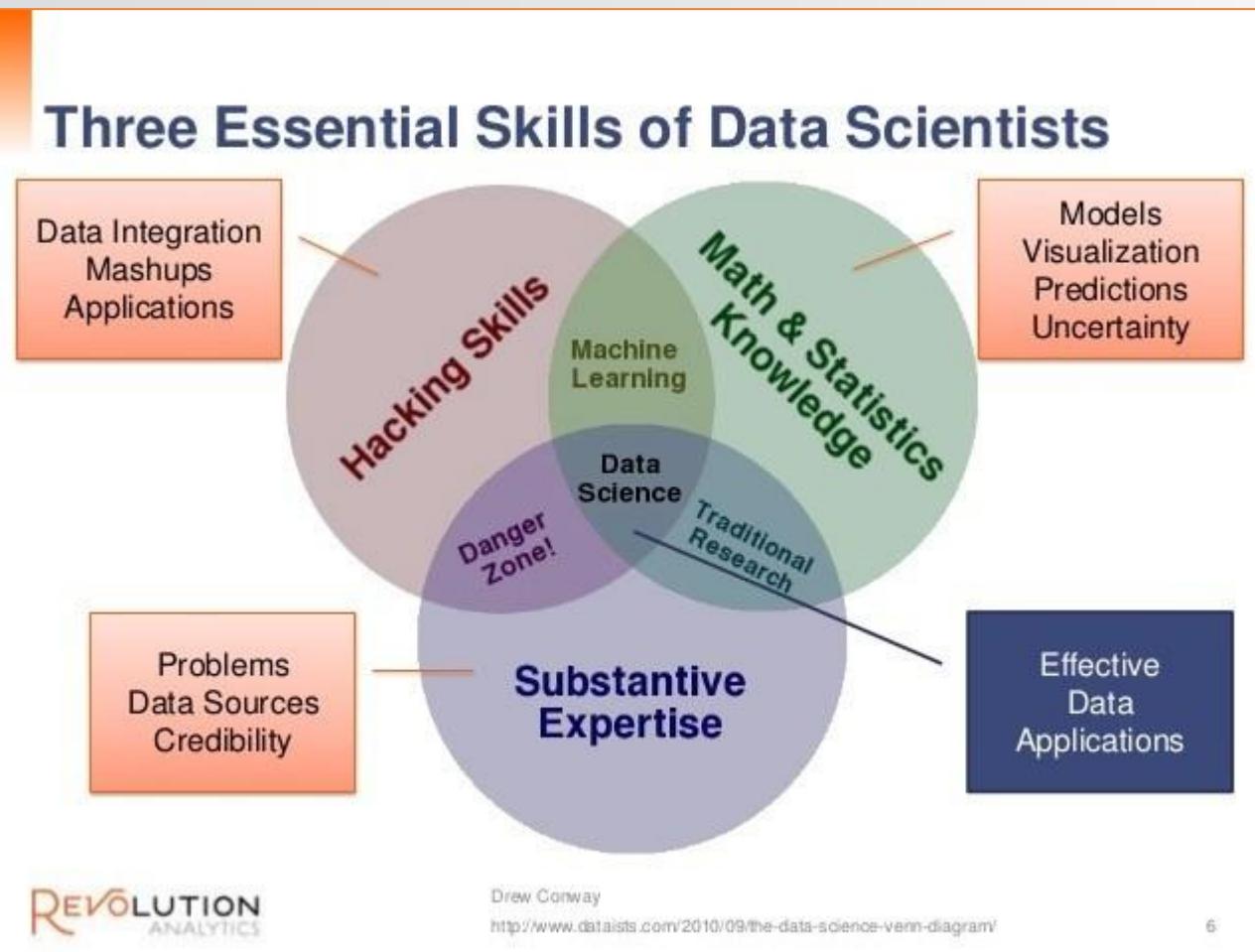
Source: Jeffery Leek <https://github.com/jtleek/dataanalysis/blob/master/week1/007typesOfQuestions/index.md>

Name of Data Analysis by Data Type

- ✓ **Biostatistics** for medical data.
- ✓ **Data Science** for data from web analytics.
- ✓ **Machine Learning** for data in computer science/computer vision.
- ✓ **Natural Language Processing** for data from texts.
- ✓ **Signal Processing** for data from electrical signals.
- ✓ **Business Analytics** for data on customers.
- ✓ **Econometrics** for economic data.



WHO is Involved?



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants



PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing packages, e.g., R
- ★ Databases: SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Curriculum via Metromap

Becoming a Data Scientist

Overall plan progressively
into the following areas:

1. Fundamentals
2. Statistics
3. Programming
4. Machine Learning
5. Text Mining / Natural Language Processing
6. Data Visualization
7. Big Data
8. Data Ingestion
9. Data Munging
10. Toolbox



Source: Swami Chandrasekaran

<http://nirvacana.com/thoughts/becoming-a-data-scientist/>

WHERE is Data Science Used?



Healthcare

- Predict diagnosis
- Prioritize screenings
- Reduce re-admittance rates



Financial services

- Fraud Detection/prevention
- Predict underwriting risk
- New account risk screens



Retail

- Product recommendation
- Inventory management
- Price optimization



Public Sector

- Analyze public sentiment
- Optimize resource allocation
- Law enforcement & security

Domain



Telco/mobile

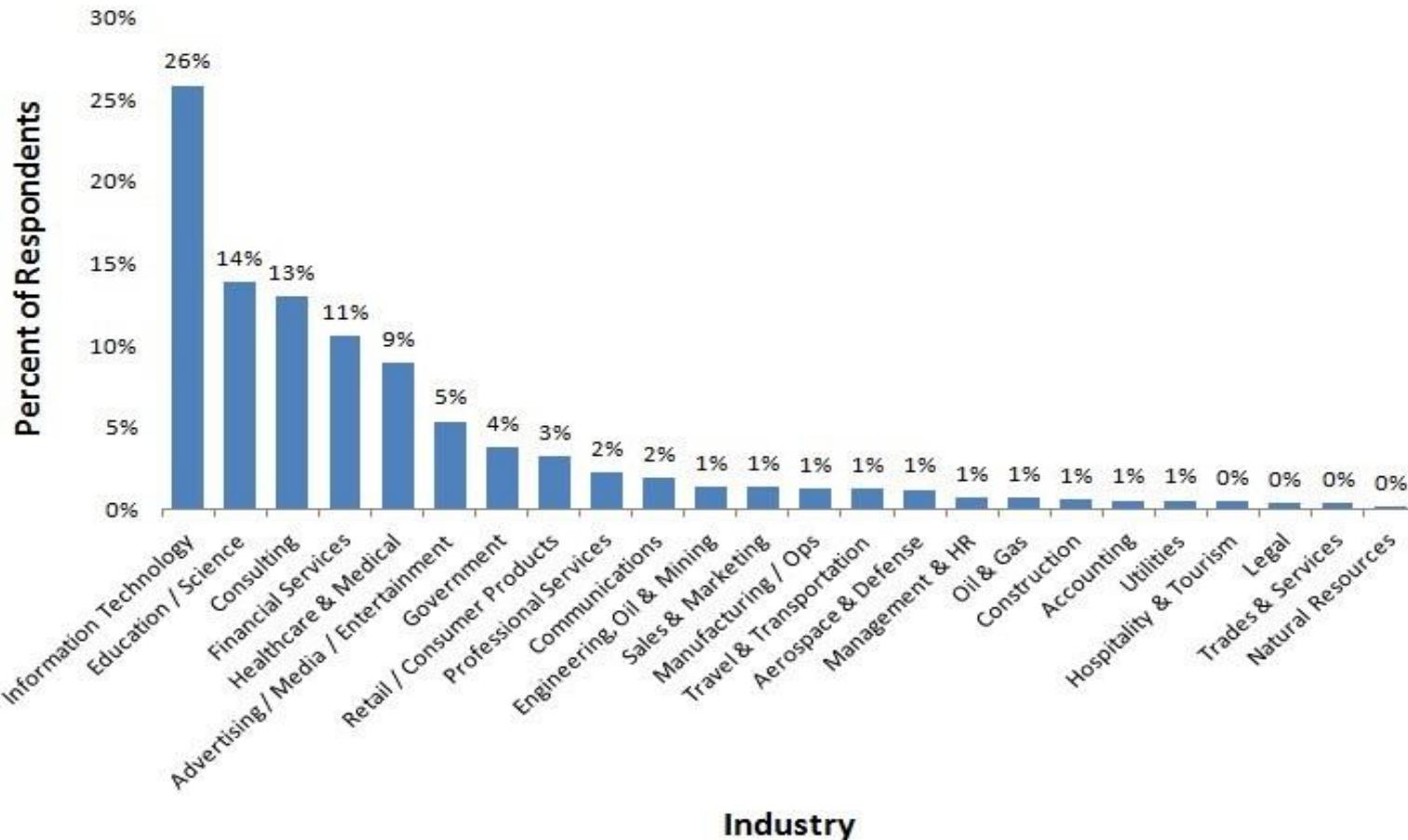
- Predict customer churn
- Predict equipment failure
- Customer behavior analysis



Oil & Gas

- Predictive maintenance
- Seismic data management
- Predict well production levels

Data Scientists Work in Many Industries



Data are based on over 1000 data professionals' responses to AnalyticsWeek and Business Over Broadway Data Science Survey. Due to small sample sizes ($N < 20$), 14 industries were not included in the chart.

WHEN is Data Science Applied?

Some examples of the outcome of data science i.e. the **data products**:

- ✓ Friend Recommendations on Facebook
- ✓ Music Recommendations on Spotify
- ✓ Product Recommendations on Amazon
- ✓ Dynamic Learning and Customized Assessments at Knewton Academy
- ✓ Trading Algorithms, Models and Credit Ratings in Finance.
- ✓ New government policies based on data.
- ✓ Predicting Flu Trends in Health
- ✓ Targeted Advertising

Applications



Data Jiu-Jitsu

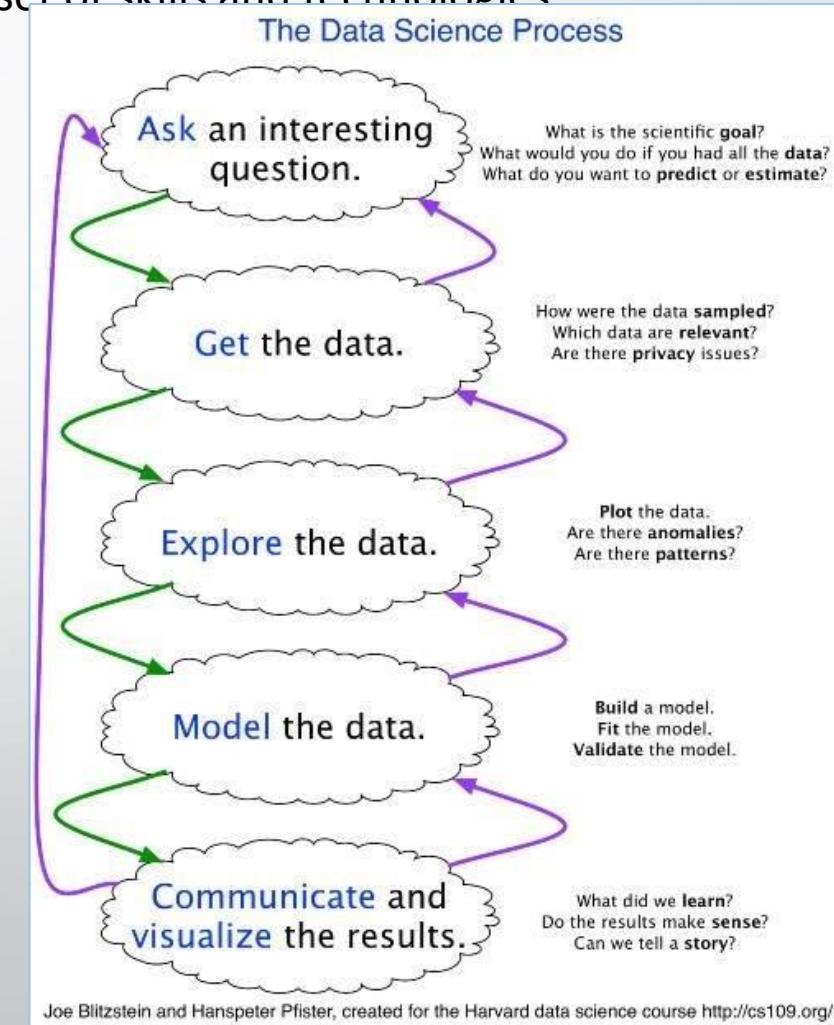
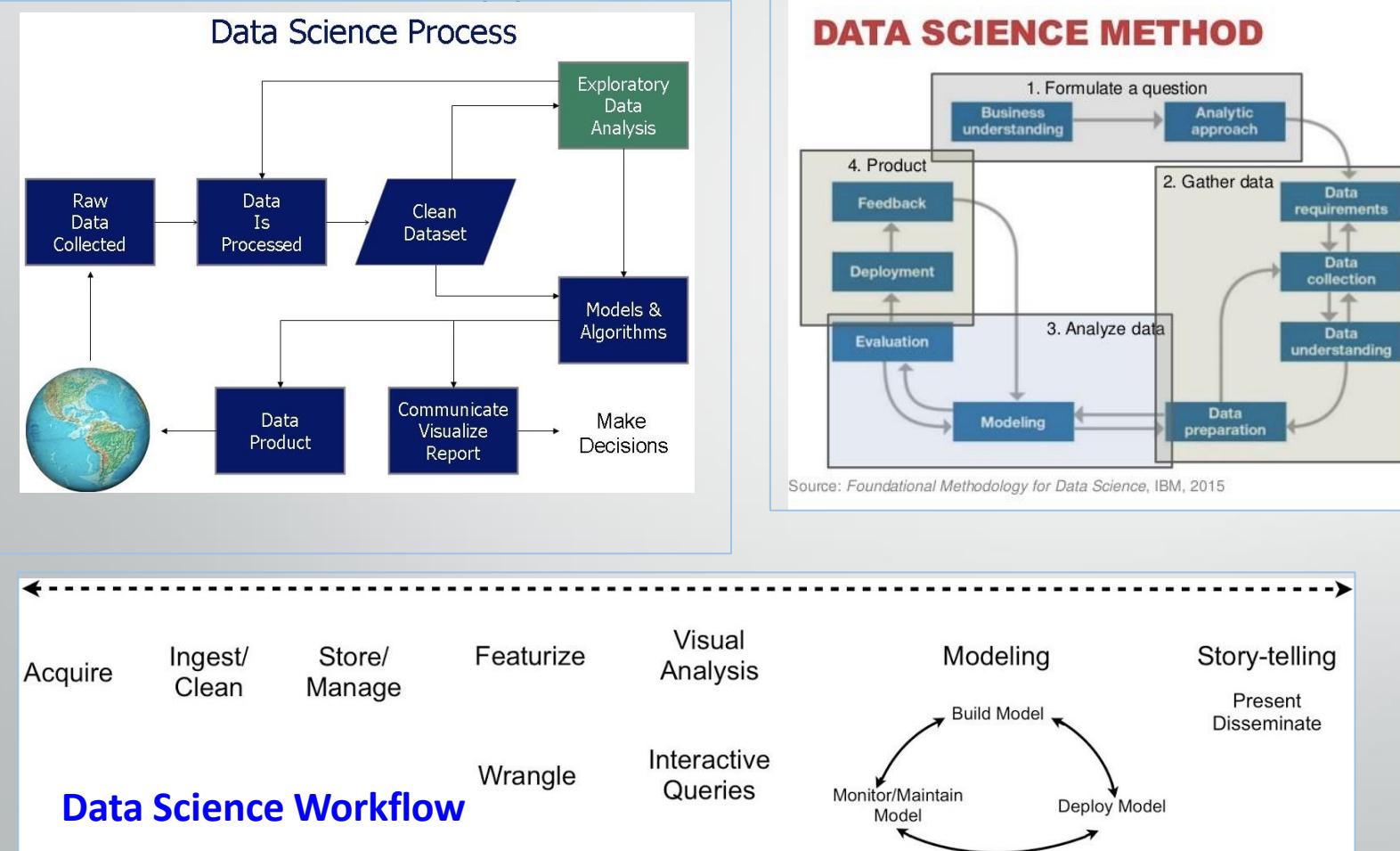


Data Jiu-Jitsu: ability to turn big data into data products that generate immediate business value
(DJ Patil @LinkedIn)

Amazon recommendations,
marketing campaigns, Uber, Siri,
price comparison sites, gaming
and image recognition, are all
powered, to varying degrees, by
data science.

HOW Data Science Works?

Data science is a multi-step process and each step in this process requires a diverse set of skills and technologies



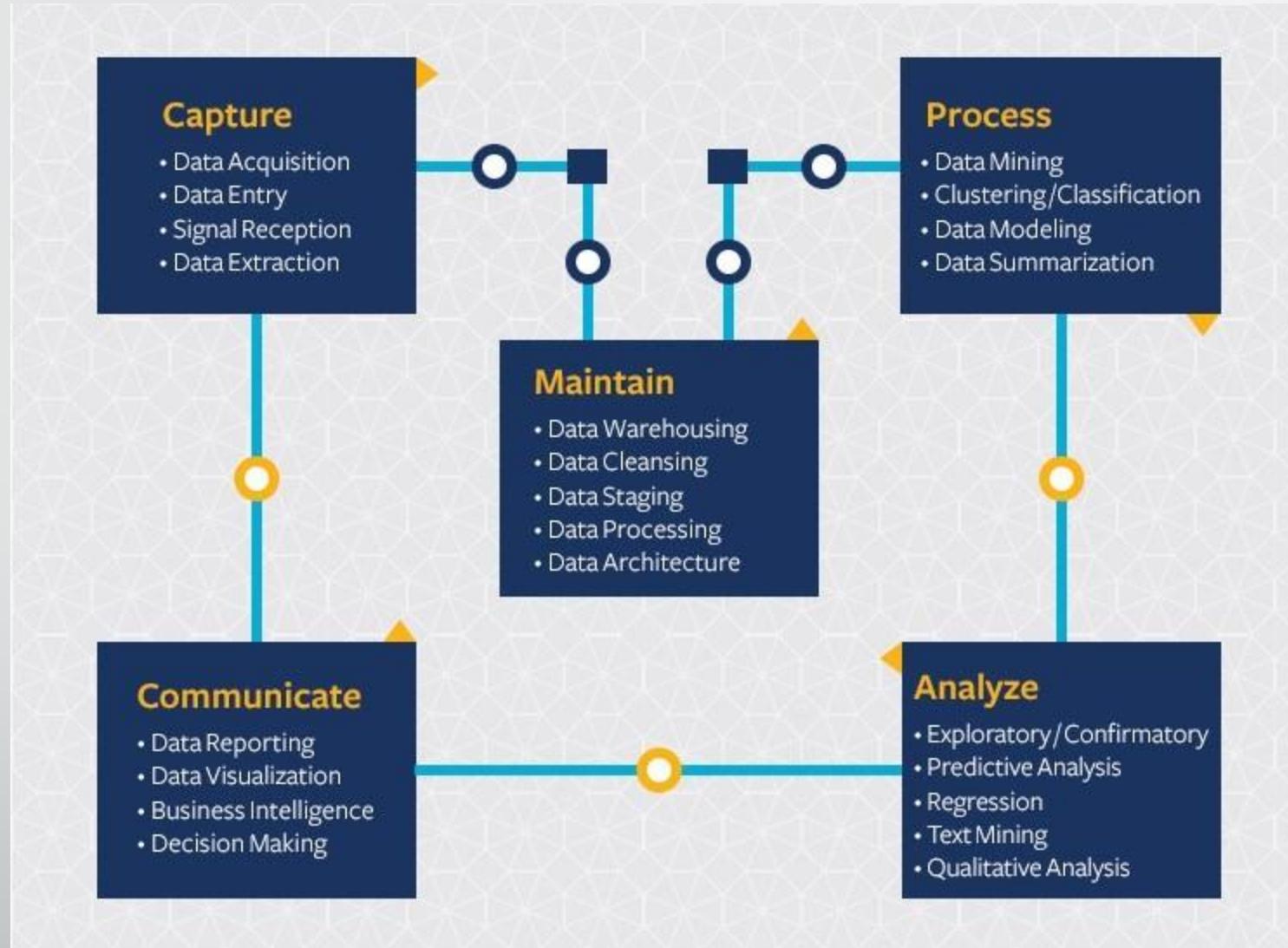
See also: <https://www.edureka.co/blog/what-is-data-science/>

Data Science

Methodology

- **Problem Formulation** – First, identify the problem to be solved. This step is easily overlooked. However, many dollars and hours have been spent solving the wrong problems.
- **Obtain The Data** – Next, collect new data and/or gather the data that already exists. In almost all cases, this data will need to be transformed and cleansed. It is important to note that this stage does not always involve big data or a data lake.
- **Analysis** – This is the part of the process where insight is to be extracted from the data. Commonly, this step will involve creating and optimizing statistical/machine learning models for prediction, but that is not always necessary. Sometimes, the analysis only contains graphs, charts, and basic descriptions of the data.
- **Data Product** – The end goal of data science is a data product. The insight from the Analysis phase needs to be conveyed to an end user. The data product might be as simple as a slideshow; more commonly it is a website dashboard, a message, an alert, or a recommendation.

The Data Science Life Cycle

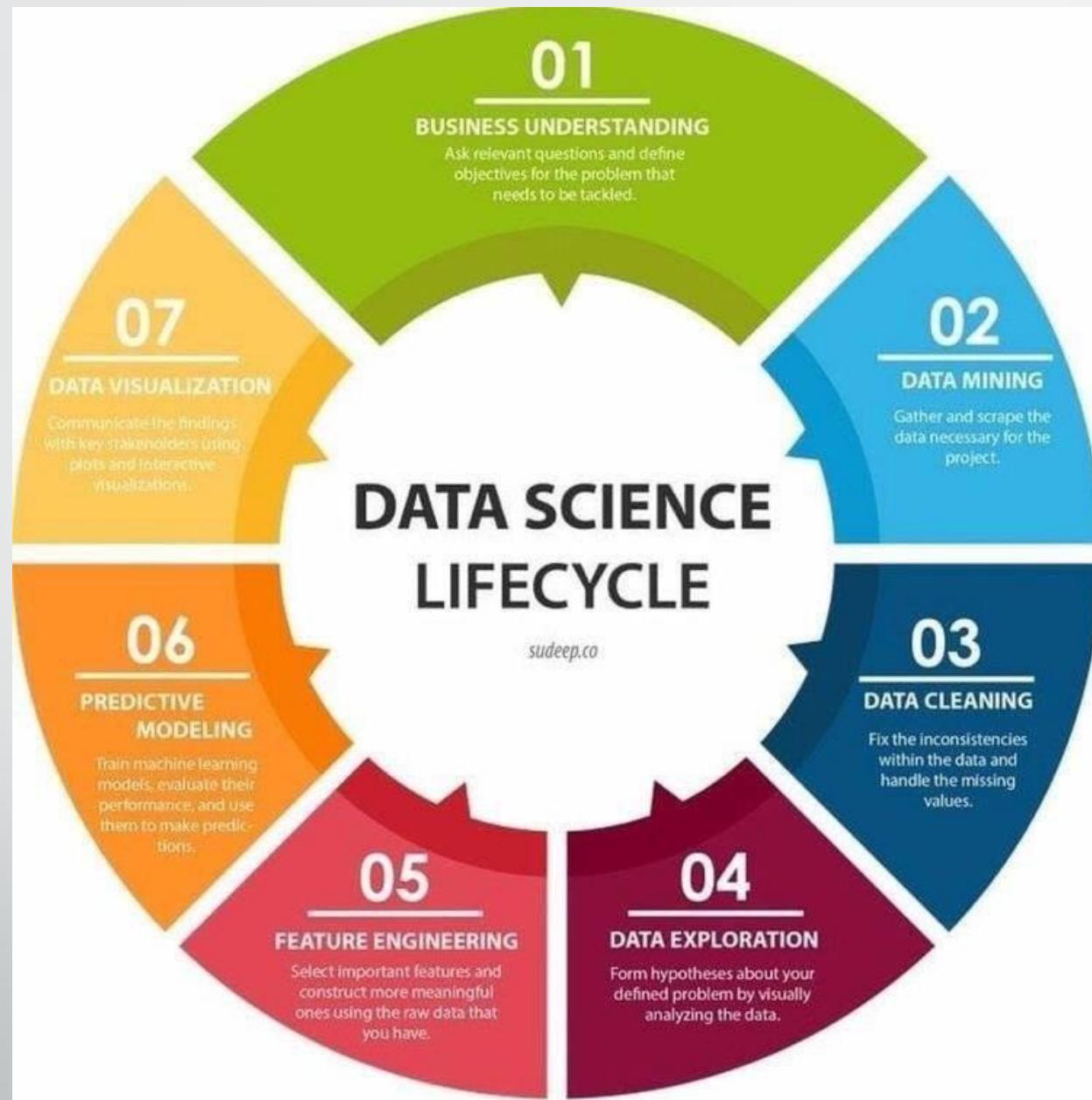


<https://datascience.berkeley.edu/about/what-is-data-science/>

OSEM model



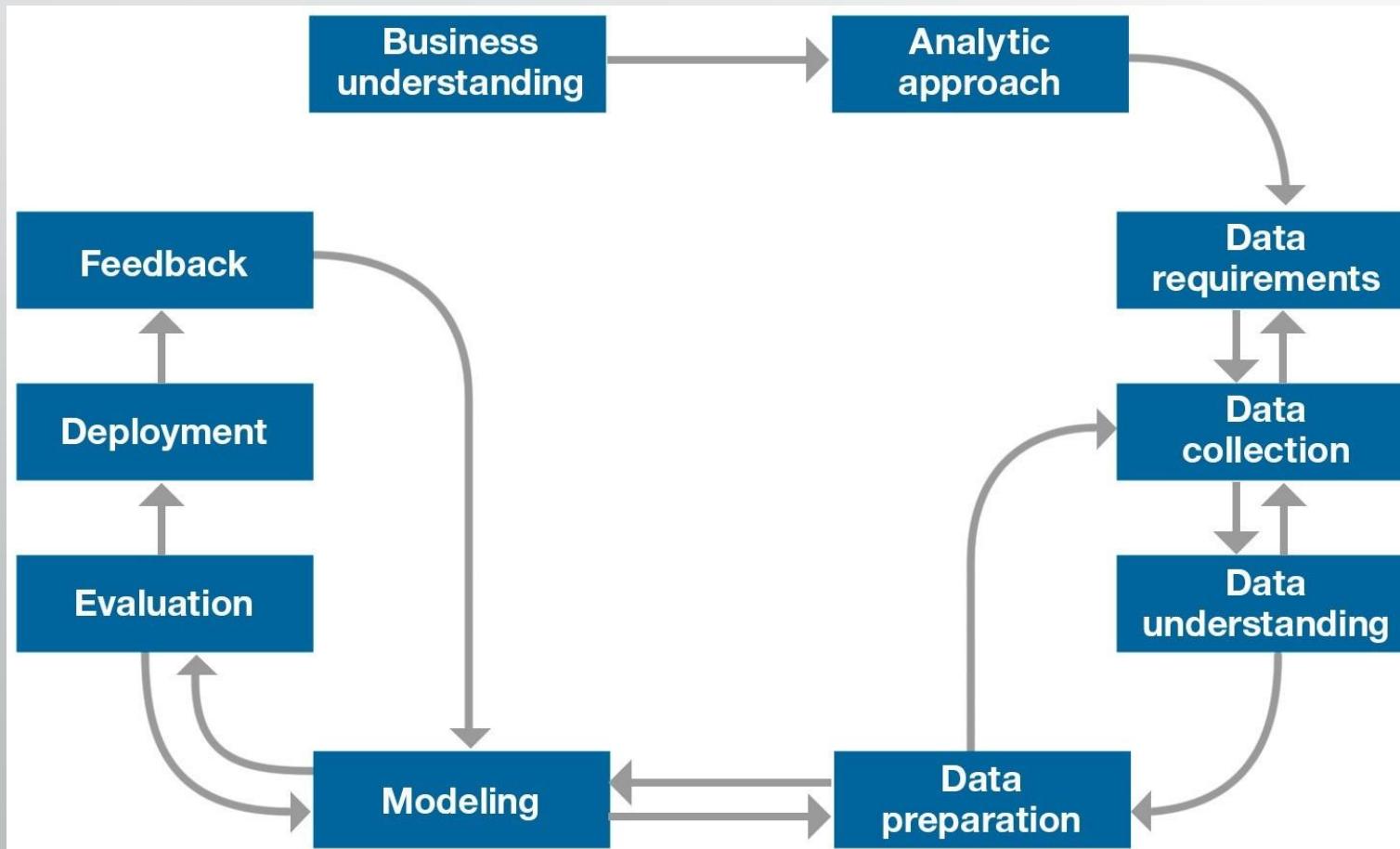
Source: [A Taxonomy of Data Science](#)



Data science process

- 1: Setting the research goal 
- 2: Retrieving data 
- 3: Data preparation 
- 4: Data exploration 
- 5: Data modeling 
- 6: Presentation and automation 

Foundational Methodology for Data Science



- A **methodology** is a general strategy that guides the processes and activities within a given domain.
- Methodology does not depend on particular technologies or tools, nor is it a set of techniques or recipes.
- Rather, a methodology provides the data scientist with a **framework** for how to proceed with whatever methods, processes and heuristics will be used to obtain answers or results.

CRISP-DM - Cross Industry Standard Process for Data Mining

Business Understanding



Don't ignore domain knowledge.
Do consult a subject matter expert

- Every project begins with **business understanding**.

- Clearly define project objectives and requirements from the business perspective... key to a successful solution
 - Business sponsors most critical in this stage
 - Define problem and solution requirements
 - Business sponsors involved throughout the project
 - Provide domain expertise
 - Review intermediate findings
 - Ensure that the work generates the intended solution

Don't Start with the Data!
Do Start with a Good Question.



- With a clear definition of the business problem, we define the **analytic approach** to solving the problem.
 - Express problem in context of statistical and machine learning techniques
 - Identify suitable technique(s)
 - Examples
 - *Classification* to predict response to a promotion ("yes" or "no")
 - *Clustering and Associations* for customer segmentation and market basket analysis

- The chosen analytic approach determines the **data requirements**.

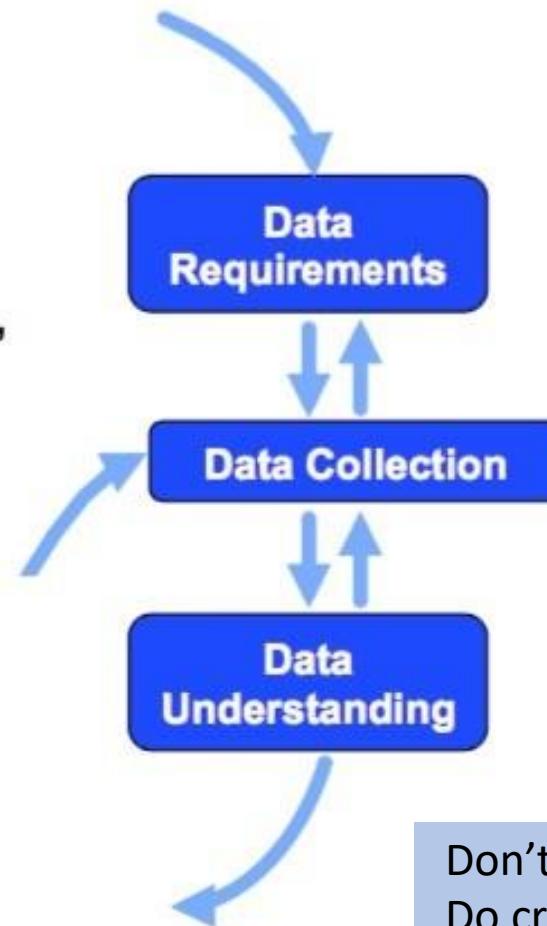
- Content, formats, representations

- Initial **data collection** is performed.

- Available data resources (structured, unstructured, semi-structured) relevant to the problem domain
 - Decide whether to obtain less-accessible data elements
 - Revise data requirements or collect more data, if needed

- Then **data understanding** is gained.

- Descriptive statistics and visualization
 - Content, quality, initial insights about data
 - Additional data collection to fill gaps, if needed



Don't brag about the size of your data.
Do collect relevant data.

Don't publish a table of numbers
Do create informative charts

Don't use just your own data
Do enhance your analysis with open data

- **Data preparation** encompasses all activities to construct the data set.
 - Data cleaning
 - Missing or invalid values
 - Eliminating duplicate rows
 - Formatting properly
 - Combining multiple data sources
 - Transforming data
 - Feature engineering
 - Text analysis
- Accelerate data preparation by automating common steps

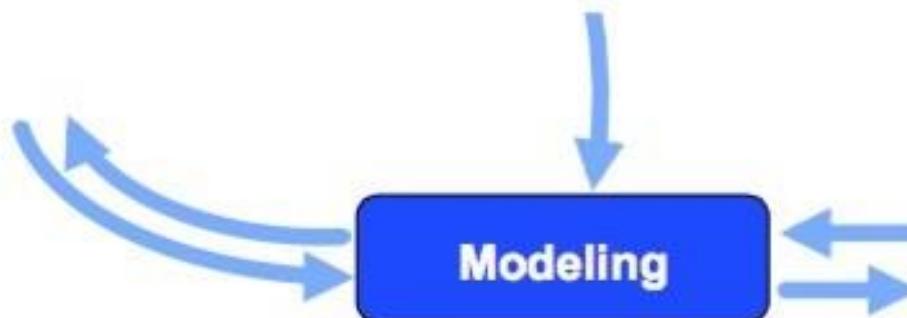


- **Modeling** focuses on developing models.

- Predictive or descriptive models
- According to the previously-defined analytic approach
- Training set for predictive modeling

- Highly iterative process

- Intermediate insights → refinements in data preparation & model specification
- Multiple algorithms & parameters to find best model for a given technique



- Model **evaluation** is performed during model development and before model deployment.

- Understand the model's quality
 - Ensure that it properly addresses the business problem

- Diagnostic measures

- Suitable to the modeling technique used
 - Testing set
 - Refine model as needed

- Statistical significance tests

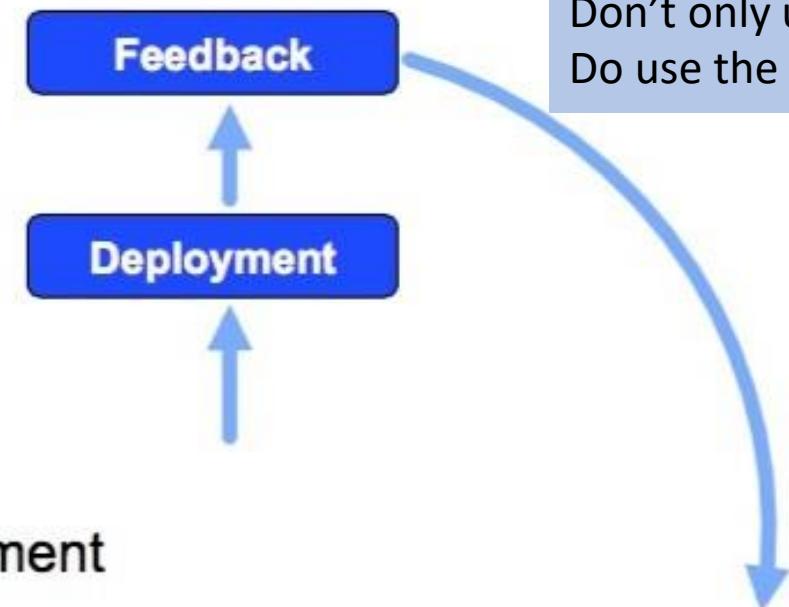


Don't always build your own tools
Do use lots of open-source tools

- Once finalized, the model is **deployed** into a production environment.

- May be in a limited / test environment until model is proven
- Involves additional groups, skills, and technologies
 - Solution owner
 - Marketing
 - Application developers
 - IT administration

Don't think one person can do it all
Do build a well-rounded team.



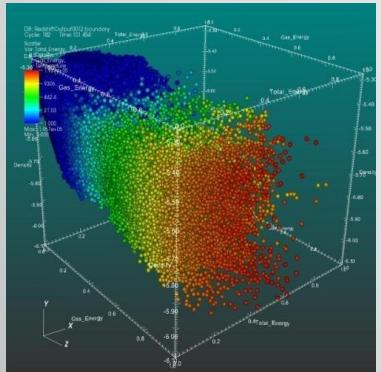
- **Feedback** to assess model performance

- Gathering and analysis of feedback for assessment of the model's performance and impact
- Iterative process for model refinement and redeployment
- Accelerate through automated processes

Don't only use one tool.
Do use the best tool for the job

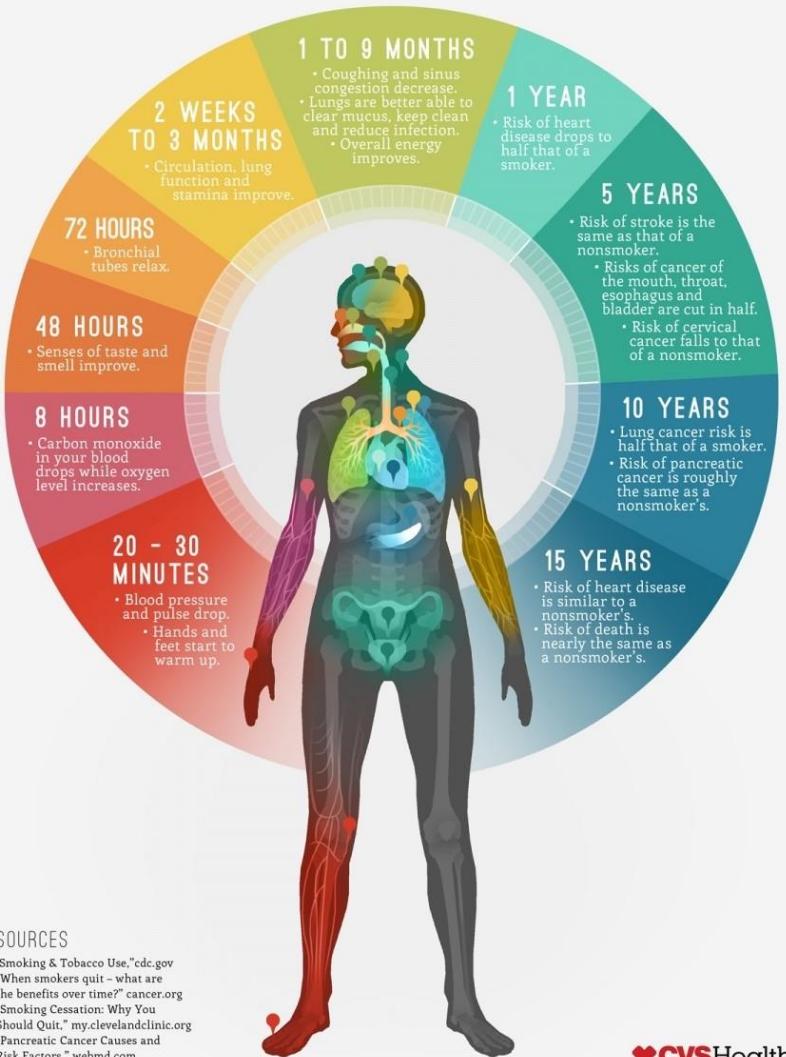
Don't keep all your findings to yourself.
Do share your analysis and results with the world!

The Importance of Visualization

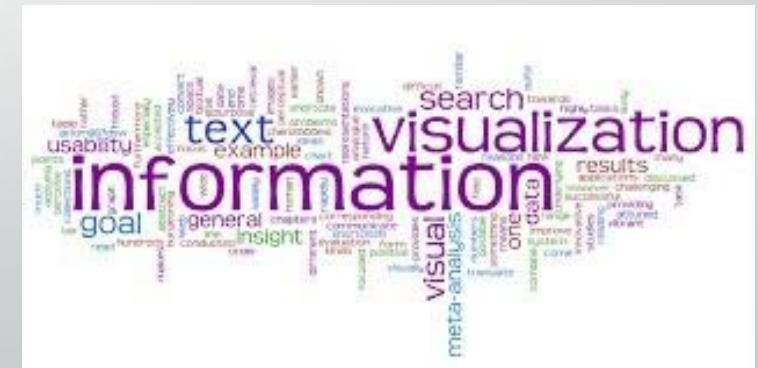


HOW QUITTING SMOKING CHANGES YOUR BODY

Here's what happens to your body after your last cigarette:



Infographics - a great way to convey information about data.



Case Study: Target

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

“ [Pole] ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole’s colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date.

Resources

- The Open-Source Data Science Masters - <http://datasciencemasters.org/>
- Data Science Certificate - <https://www.coursera.org/specializations/jhu-data-science>
- Data Science Courses – bigdatauniversity.com
- Data Science Association – www.datascienceassn.org

JOURNAL TITLE

[EJP Data Science](#)

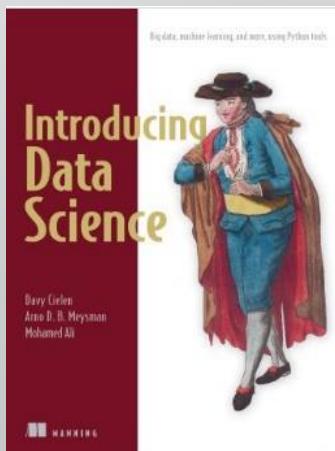
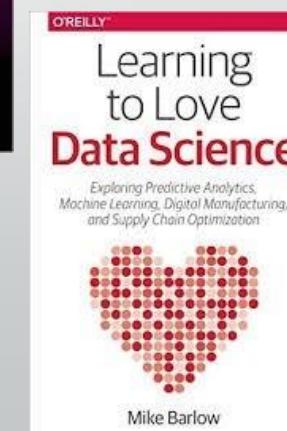
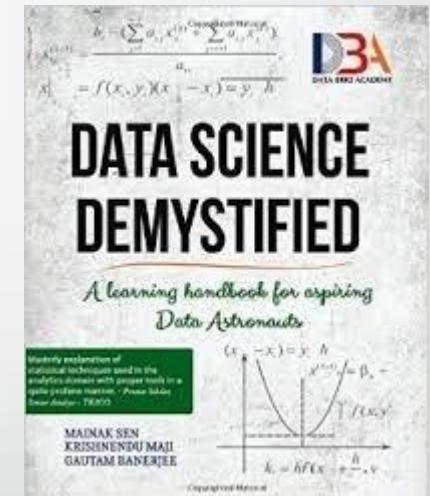
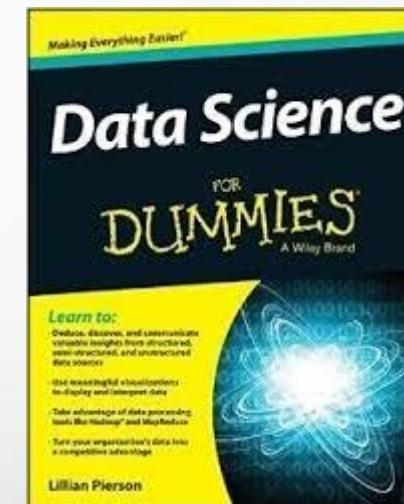
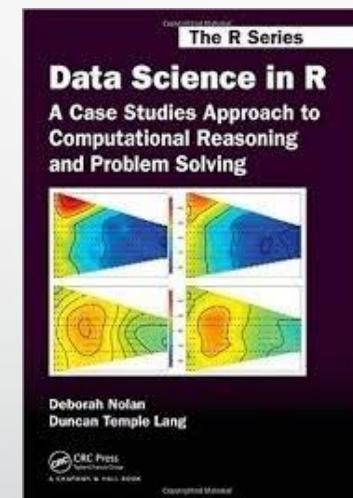
[CODATA Data Science Journal](#)

[Journal of Data Science Online](#)

[Big Data Journal](#)

[JDS](#) (focuses heavily on the applications of data.)

[GigaScience](#) (focuses on life and biomedical science)



24 Data Science Resources

to Keep Your Finger on the Pulse

Trends & Happenings

- [Flowing Data](#)
- [FiveThirtyEight](#)
- [Simply Statistics](#)
- [R-Bloggers](#)
- [Edwin Chen](#)
- [Hunch](#)

Learn More

- [Learn Data Science](#)
- [Open Source Data Science Masters](#)

Join a Community

- [DataTau](#)
- [Cross Validated](#)
- [Reddit Machine Learning Subreddit](#)
- [Metaoptimize](#)
- [Kaggle Competitions](#)

Data Science News

- [Data Science Weekly](#)
- [KD Nuggets](#)

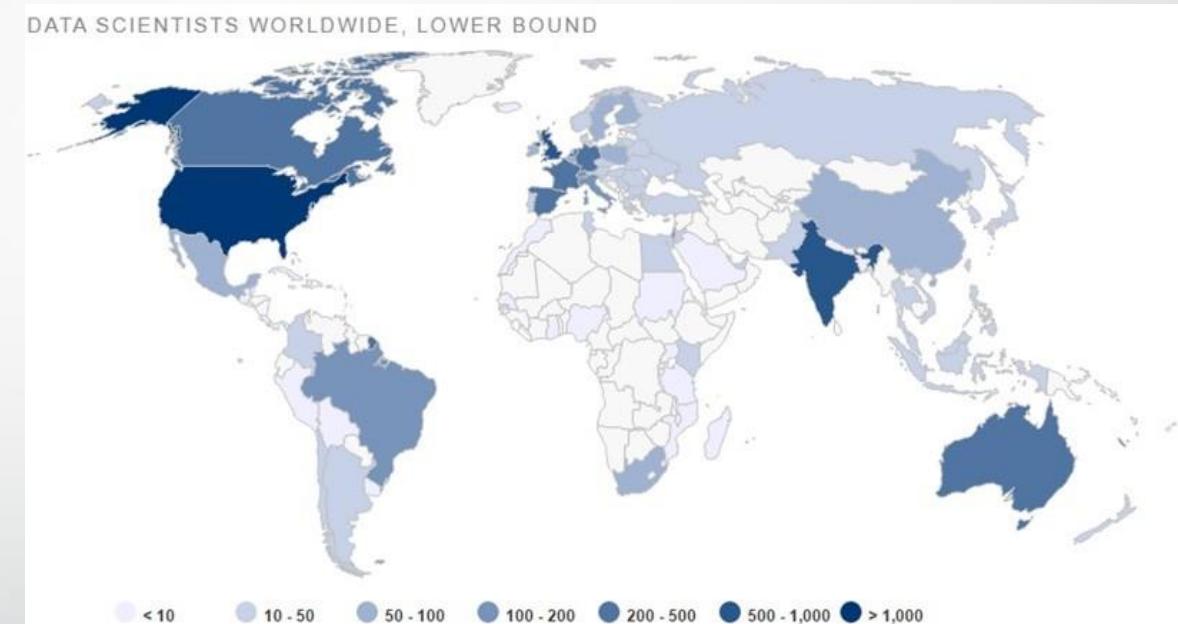
- [@peteskromoroch](#)
- [@hackingdata](#)

Brilliant Minds

- [@drewconway](#)
- [@flowingdata](#)
- [@dpatil](#)
- [@hmason](#)

Conclusion

- Data science is a **growth area**. (*The number of data scientists has doubled over the last 4 years*)
- The future belongs to the companies and people that turn **data into products**. (The Information Technology and Services industry employs the largest number of data scientists.)
- Top skills (The top five skills listed by data scientists are: Data Analysis, R, Python, Data Mining, and Machine Learning)
- Education level (Over 79% of data scientists that list their education have earned a graduate degree, and 38% have earned a PhD.)



Source: <https://rjmetrics.com/resources/reports/the-state-of-data-science/>



Data science for the environment | Dan Hammer | TEDxBerkeley
<https://www.youtube.com/watch?v=ph439t-kTIE>

Go Ahead and Do Now

Own definition of data science based
on own understanding

Rationale why data science
matters today



Identify famous data scientist

List domains of DS applications and
examples of data products.

Provide methodology, process, lifecycle, or
workflow for data science