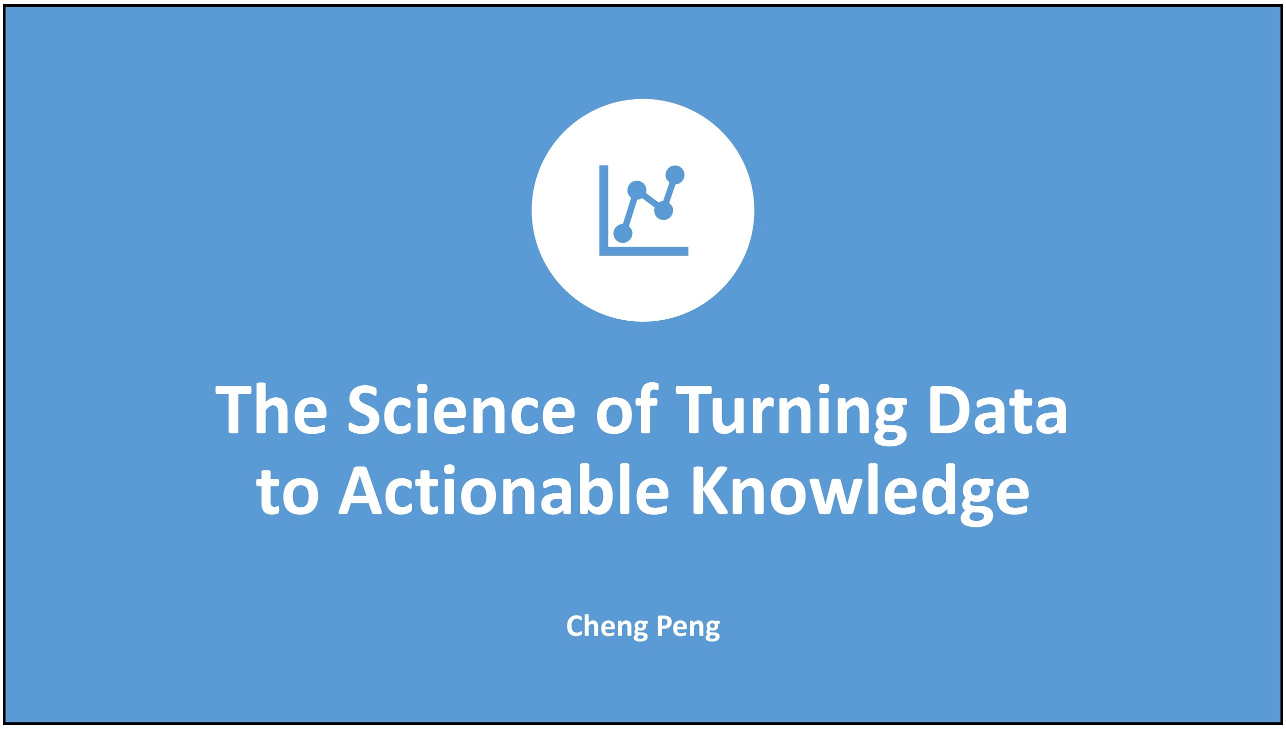
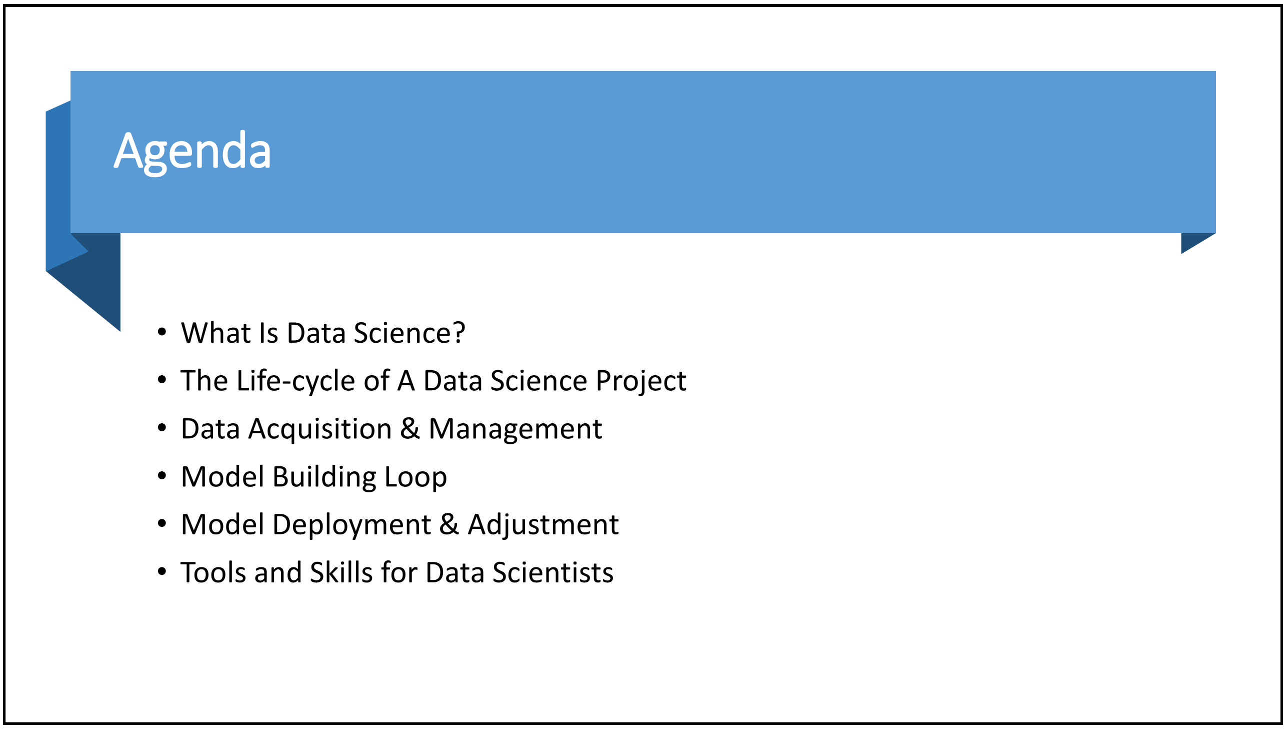


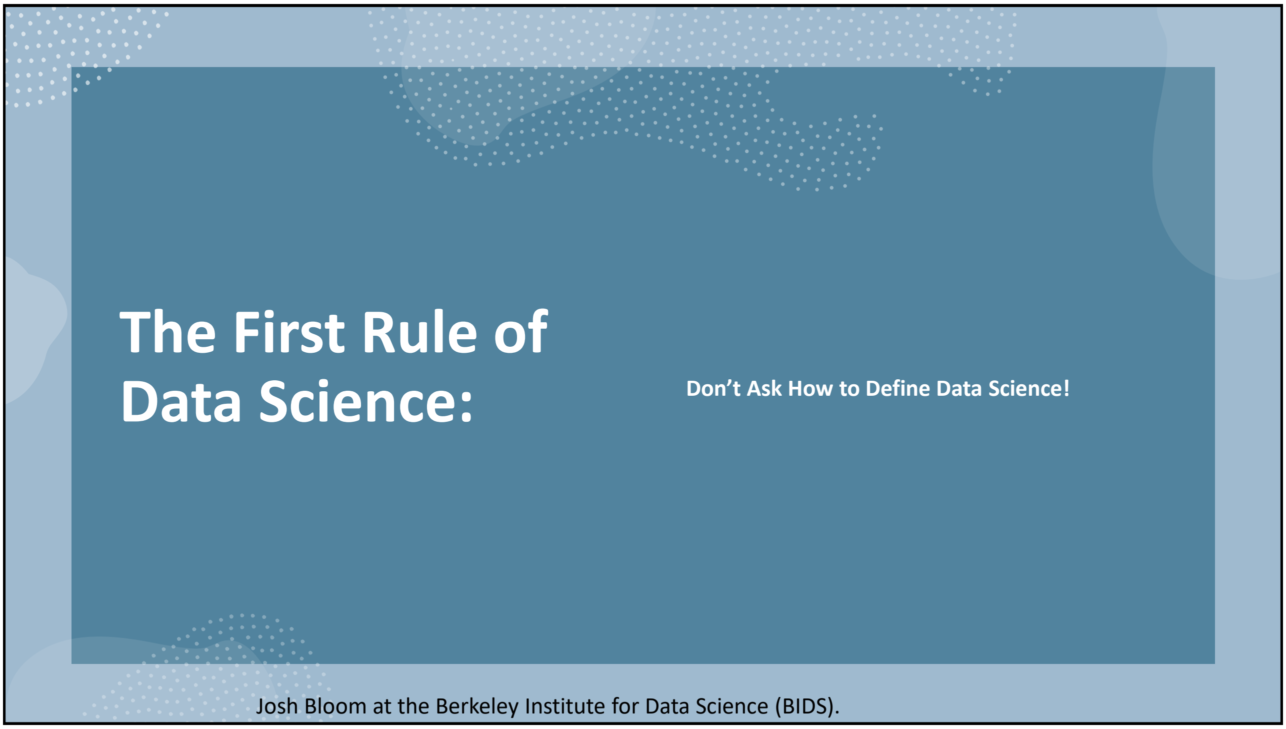
The Science of Turning Data to Actionable Knowledge

Cheng Peng



Agenda

- What Is Data Science?
- The Life-cycle of A Data Science Project
- Data Acquisition & Management
- Model Building Loop
- Model Deployment & Adjustment
- Tools and Skills for Data Scientists

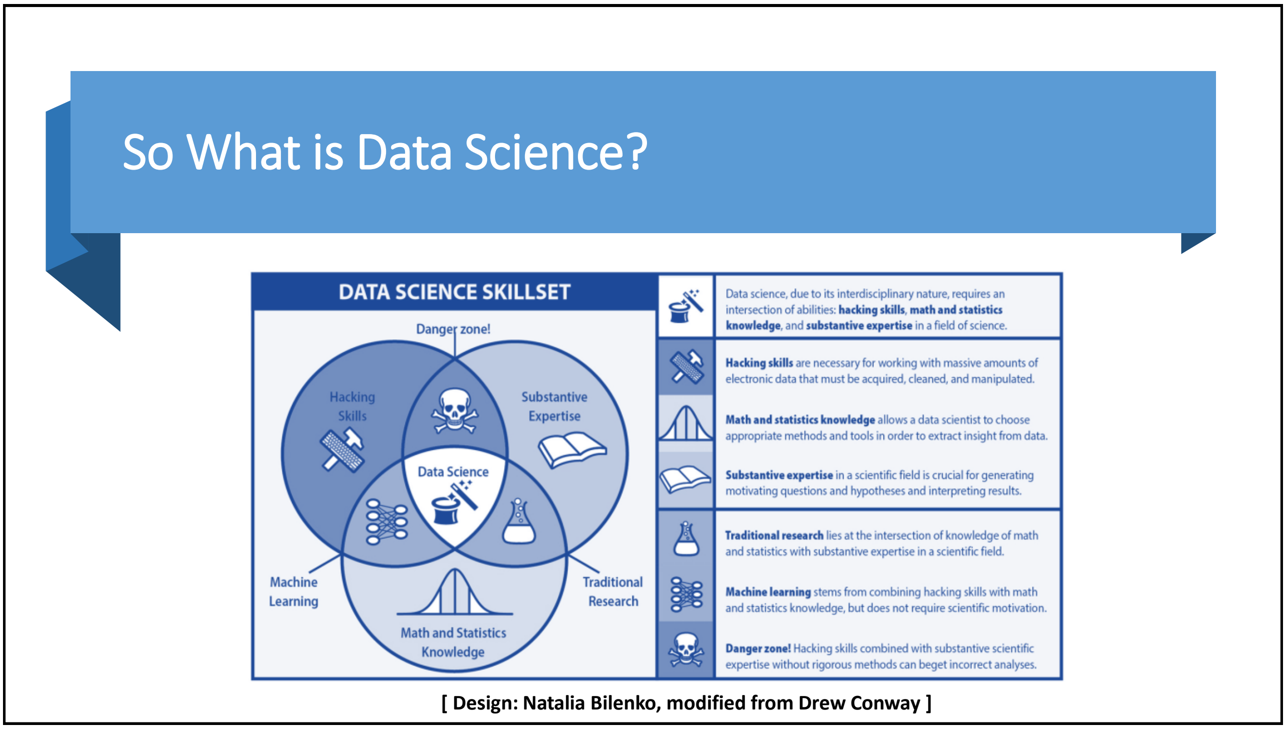


The First Rule of Data Science:

Don’t Ask How to Define Data Science!

Josh Bloom at the Berkeley Institute for Data Science (BIDS).

3



So What is Data Science?

Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills**, **math and statistics knowledge**, and **substantive expertise** in a field of science.

Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

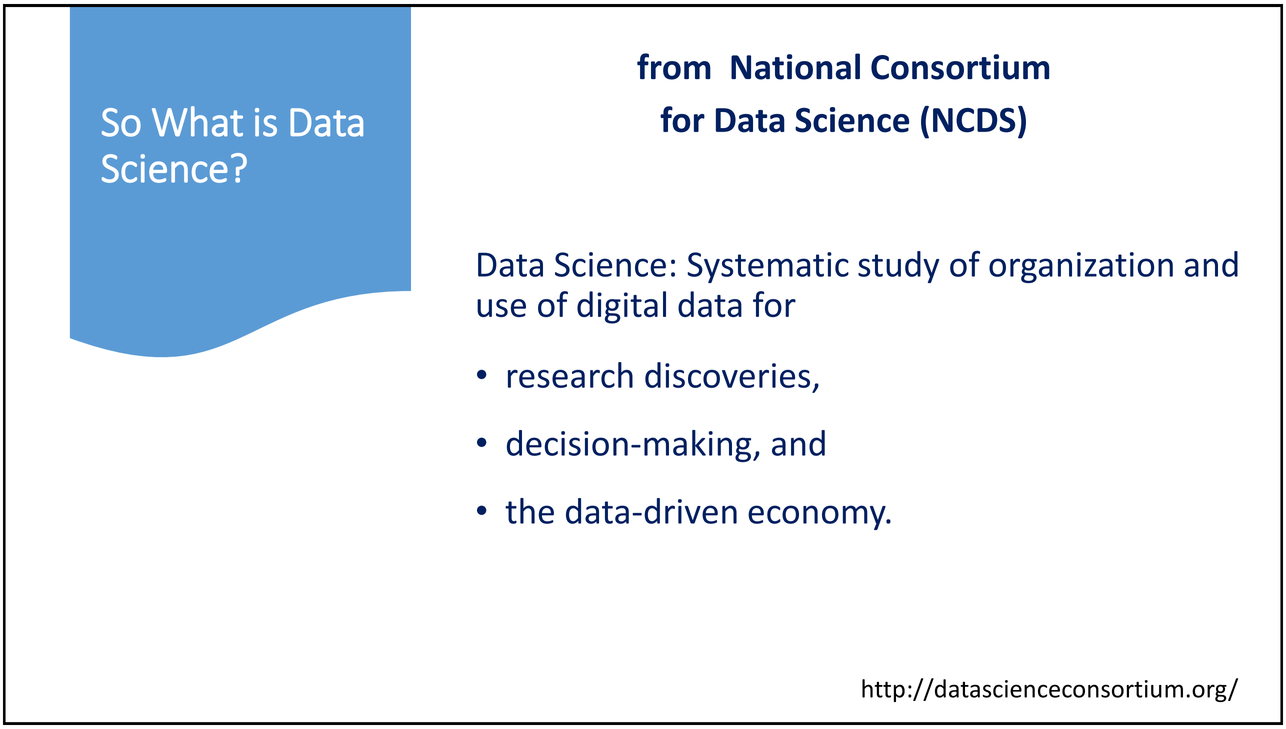
Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

[Design: Natalia Bilenko, modified from Drew Conway]

4



So What is Data Science?

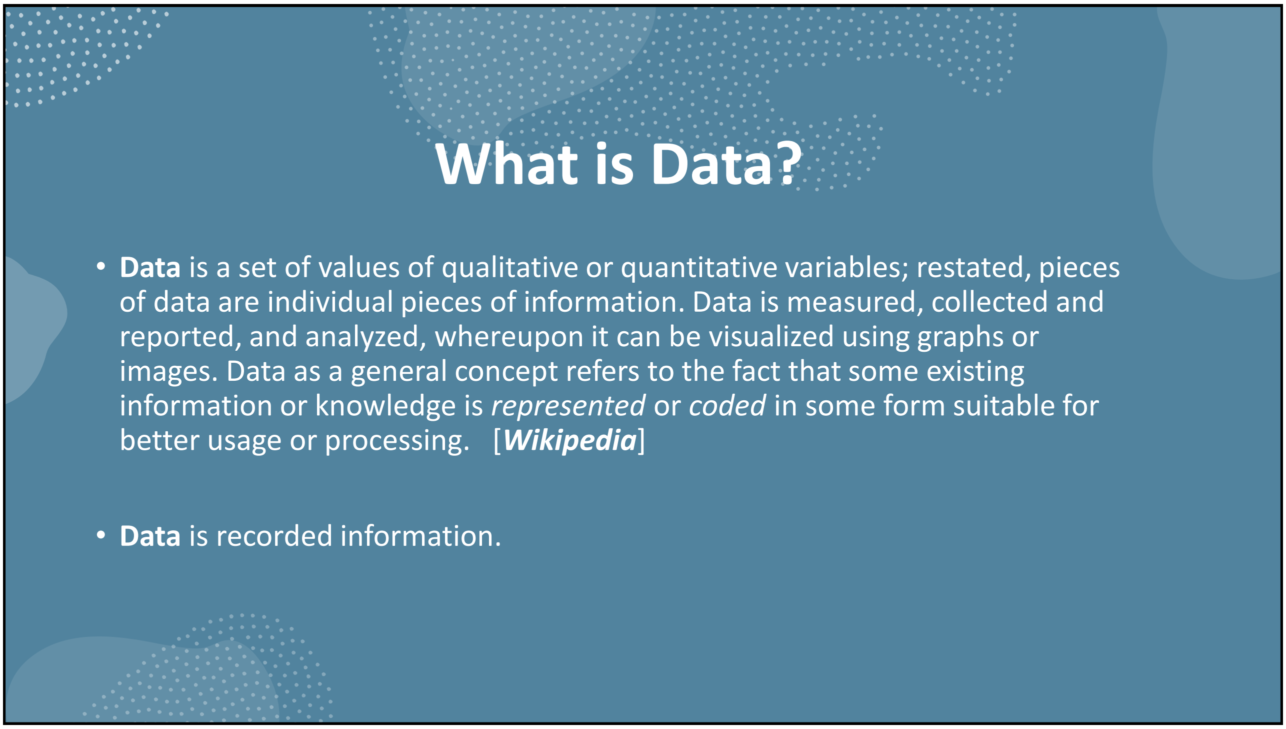
from National Consortium for Data Science (NCDS)

Data Science: Systematic study of organization and use of digital data for

- research discoveries,
- decision-making, and
- the data-driven economy.

http://datascienceconsortium.org/

5

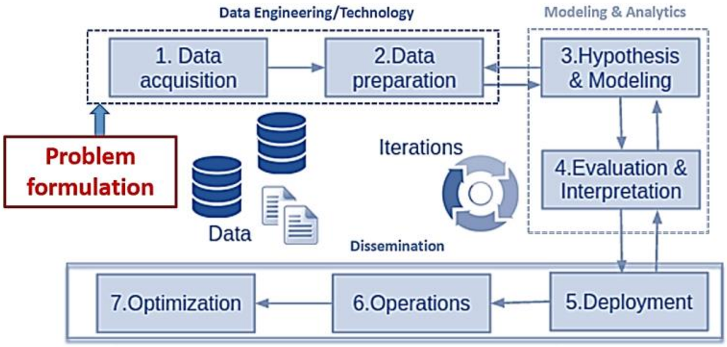


What is Data?

- **Data** is a set of values of qualitative or quantitative variables; restated, pieces of data are individual pieces of information. Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs or images. Data as a general concept refers to the fact that some existing information or knowledge is *represented* or *coded* in some form suitable for better usage or processing. [***Wikipedia***]
- **Data** is recorded information.

6

DS Project Lifecycle

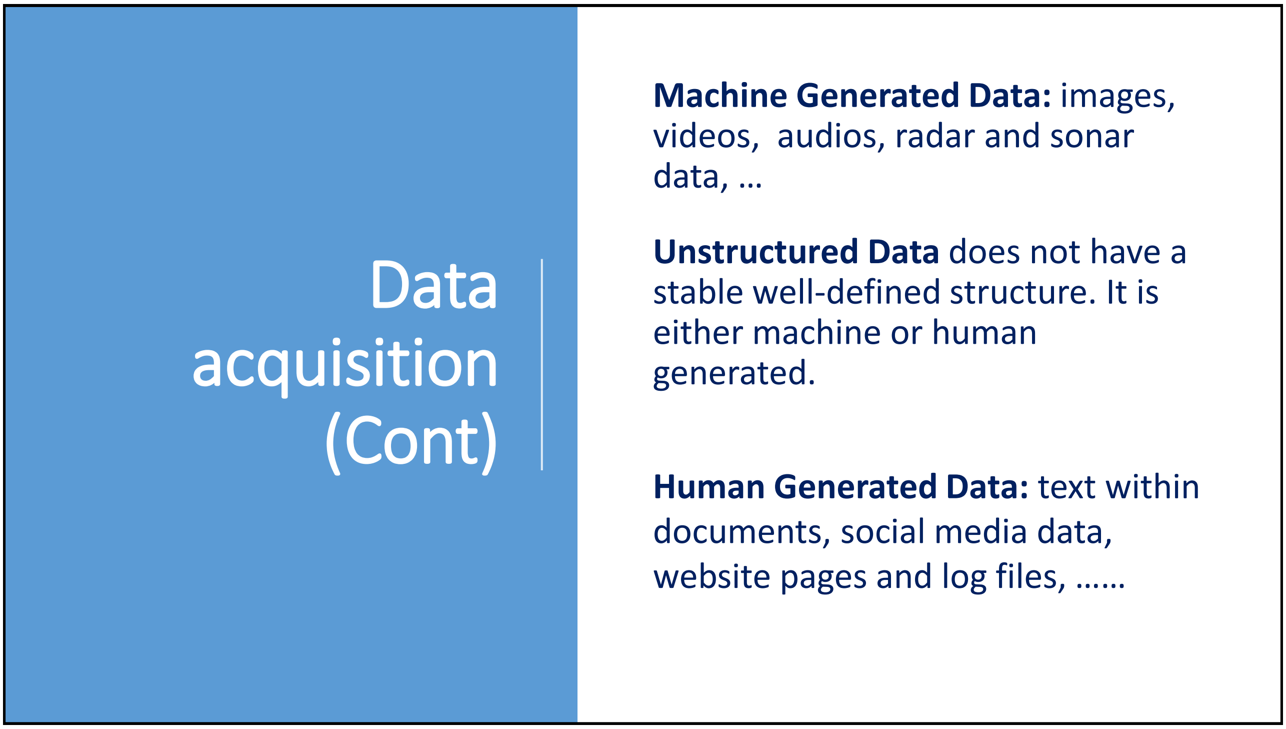


[http://www.datasciencecentral.com/profiles/blogs/the-data-science-project-lifecycle]

Data acquisition:

- **Data acquisition** is a set of processes and programs that extracts data for the data warehouse and operational data store from the operational systems

Basic Types of Data 1) Structured Data
2) Unstructured Data
3) Semi-structured Data



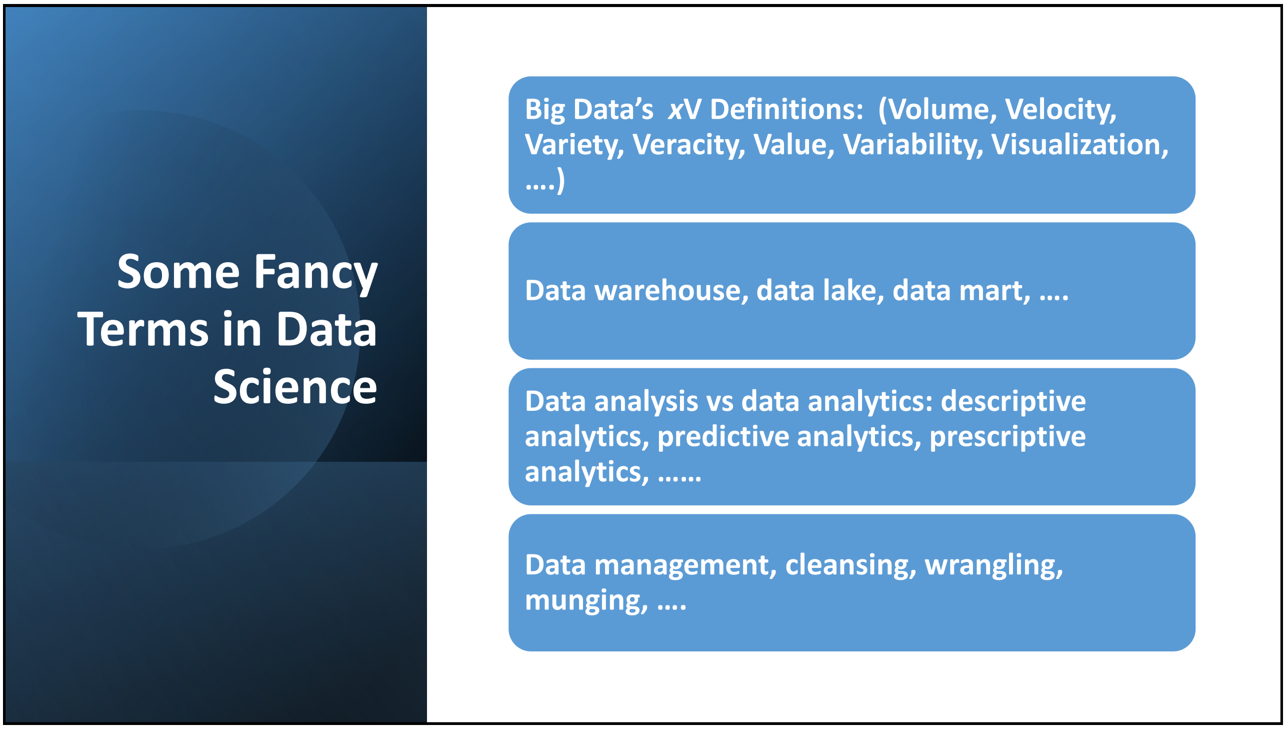
Data acquisition (Cont)

Machine Generated Data: images, videos, audios, radar and sonar data, …

Unstructured Data does not have a stable well-defined structure. It is either machine or human generated.

Human Generated Data: text within documents, social media data, website pages and log files, ……

9



Some Fancy Terms in Data Science

Big Data's *x*V Definitions: (Volume, Velocity, Variety, Veracity, Value, Variability, Visualization, ….)

Data warehouse, data lake, data mart, ….

Data analysis vs data analytics: descriptive analytics, predictive analytics, prescriptive analytics, ……

Data management, cleansing, wrangling, munging, ….

10

Big Data, New Challenges!

Challenge #1: Data Storage – New Infrastructures

Challenge #2: Data Processing – New Tools and Methods for Information Extraction

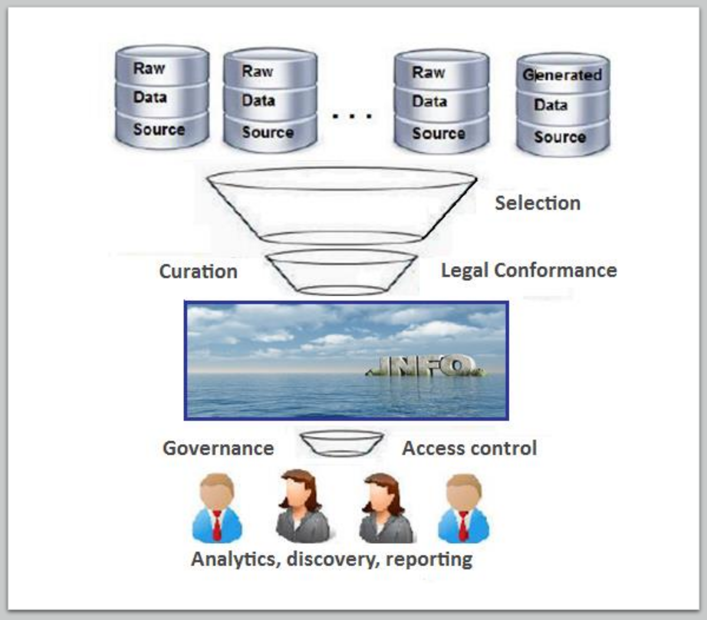
Challenge #3: Analytical Modeling – New Methods with a Sound Theoretical Foundation

Challenge #4: Lack of Data Science Talent!

11

Data Lake

Data Wrangling: The Challenging Journey from the Wild to the Lake.



Source: Ignacio Terrizzano, Peter Schwarz, Mary Roth, John E. Colino from IBM Research

12

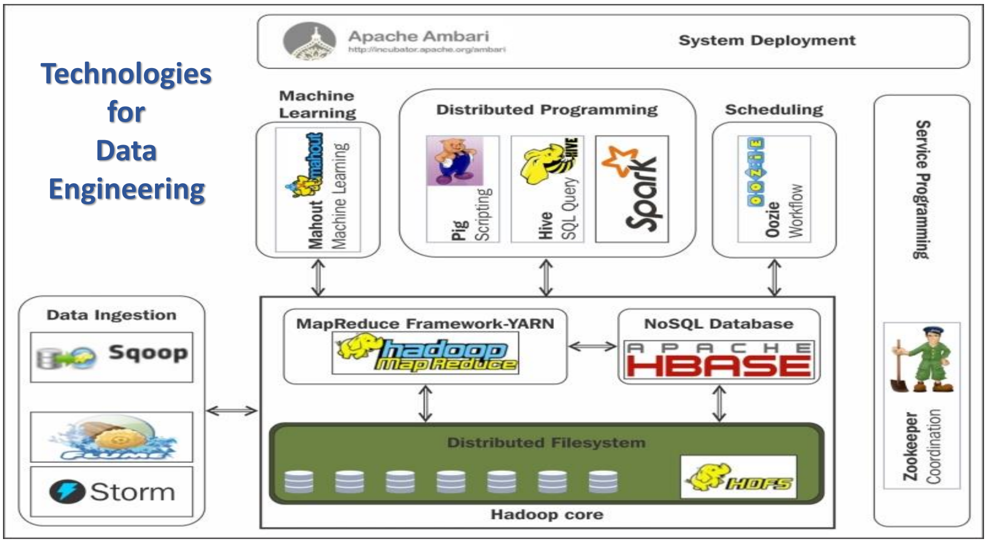
LET'S DIVE IN

Data Acquisition & Management

http://ebooks-store.com/dealing-big-data-ascendency-data-lakes/

13

Data Technologies – An At-a-glance View



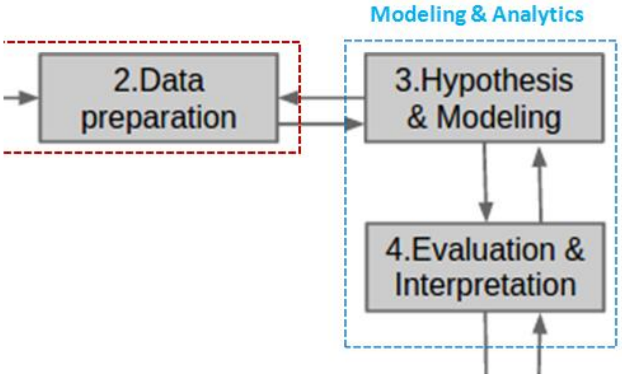
Technologies for Data Engineering

https://www.safaribooksonline.com/library/view/hadoop-essentials/9781784396688/ch02s05.html

14

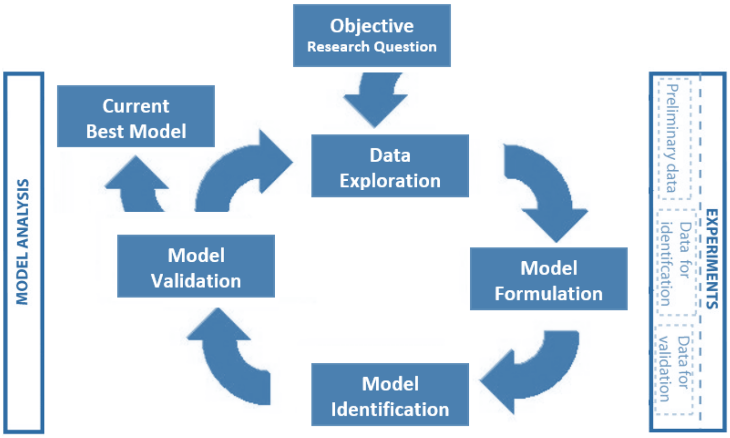
Modeling Building Loop:

A Recursive Process



15

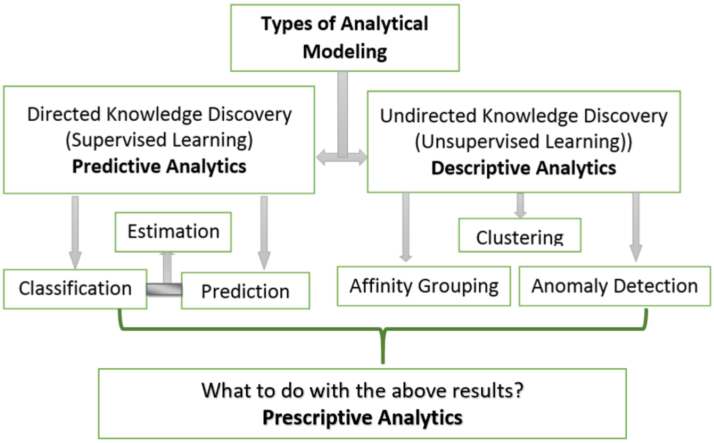
Modeling Building Loop: A Recursive Modeling Process



16

Model Building Loop:

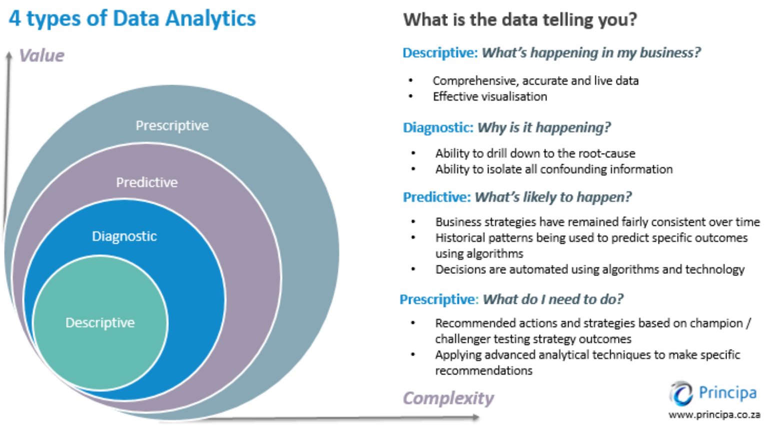
- Basic Types of Data
- Science Modeling Problems


Types of Analytical Modeling
Directed Knowledge Discovery (Supervised Learning) **Predictive Analytics**
Undirected Knowledge Discovery (Unsupervised Learning)) **Descriptive Analytics**
Estimation
Classification
Prediction
Clustering
Affinity Grouping
Anomaly Detection
What to do with the above results? **Prescriptive Analytics**

17

Model Building Loop

- Description of
- Science Modeling Problems



What is the data telling you?

Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

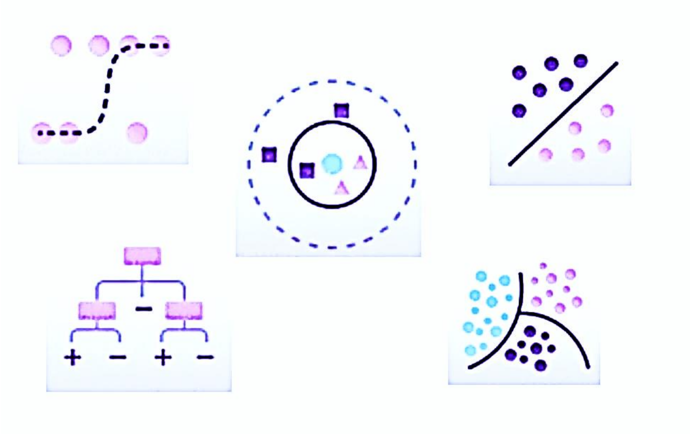
Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

18

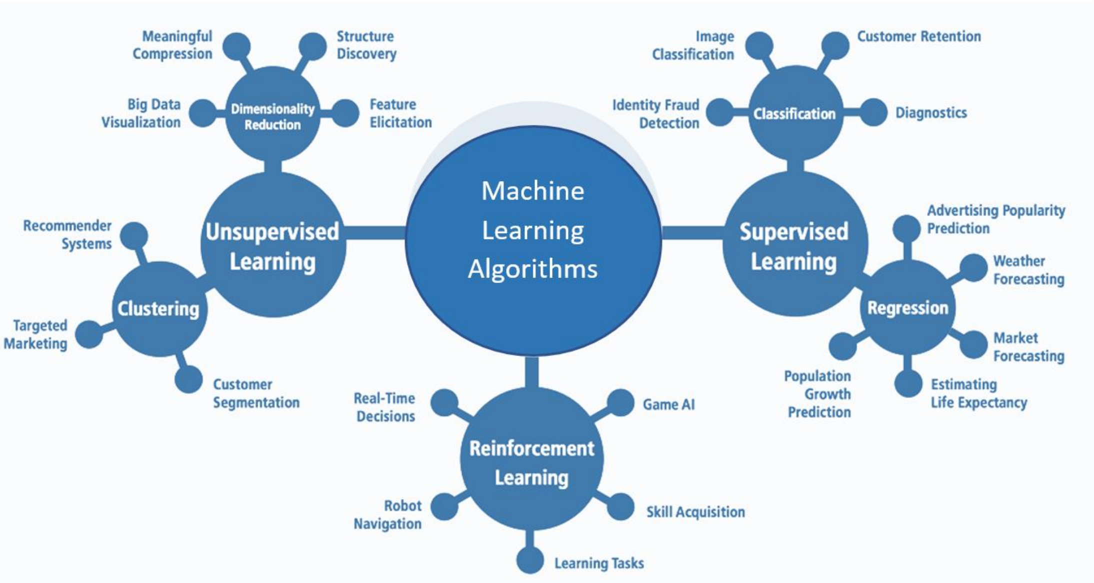
Model Building Loop:

- **Classification Models**
- **(partial list)**



19

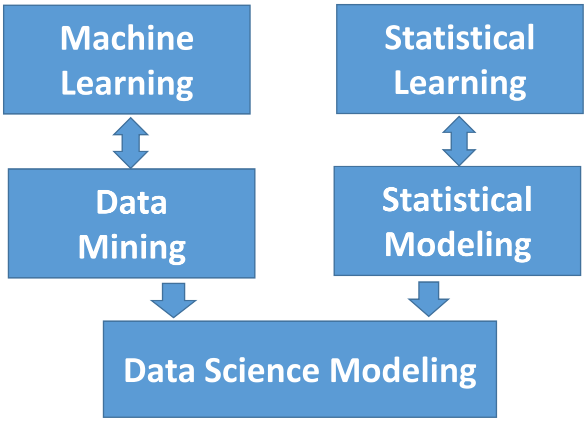
Model Building Loop: Machine Learning Algorithms



20

Model Building Loop:

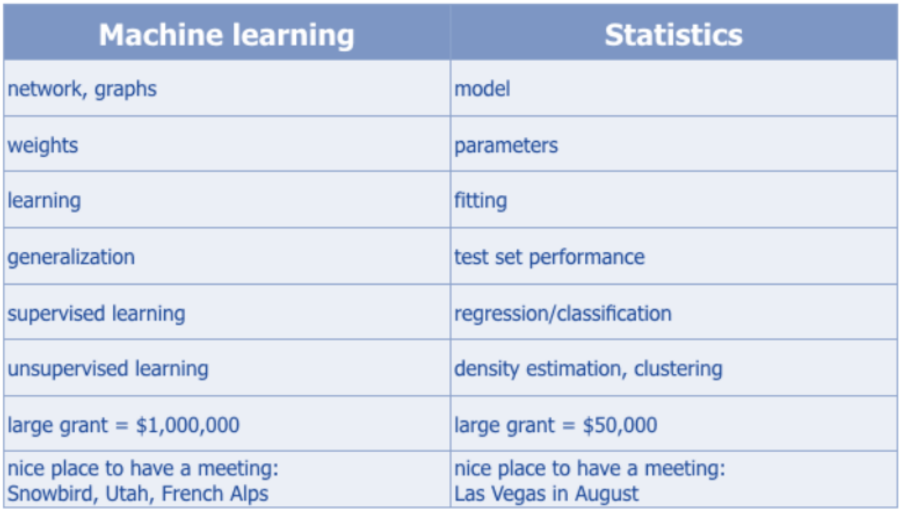
Approaches to Data Science Models



21

Model Building Loop:

Some Comparisons Between ML & Stats

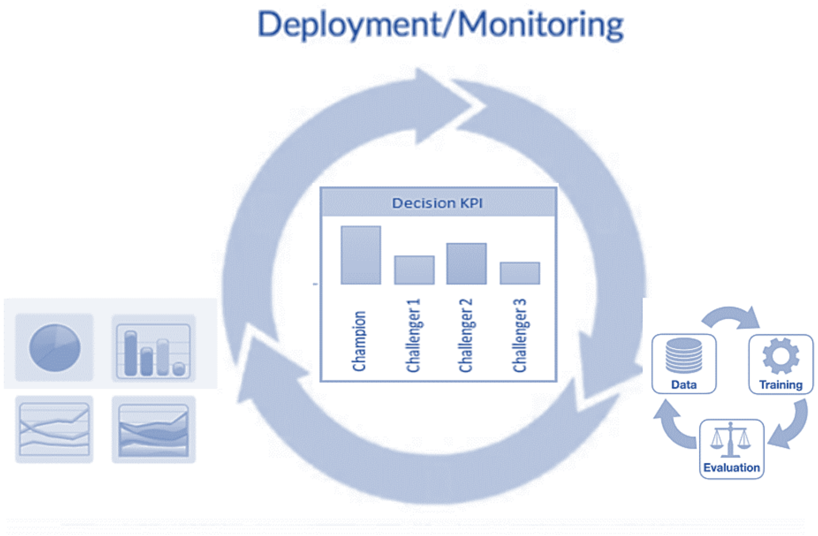


Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant = $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

[From a note of Robert Tibshirani]

22

Deployment & Monitoring: Iterative Process



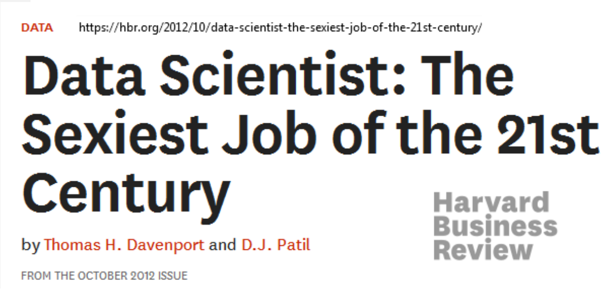
23

Data Science Tools and Technology

Platform	Apache Spark, hadoop
Tools	tableau, jupyter, RStudio, QlikView
Frameworks	Spark MLlib, scikit learn, TensorFlow™
Language	python™, R, Java, Scala
Format	Parquet, pandas
SaaS	Google Machine Learning, Amazon Machine Learning, Azure Machine Learning

24

Tools and Skills for Data Scientists



DATA https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

Data Scientist: The Sexiest Job of the 21st Century

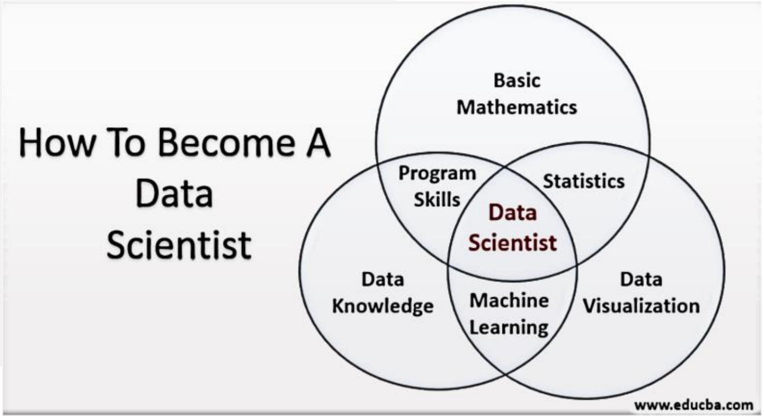
by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Harvard Business Review

What is a data scientist?

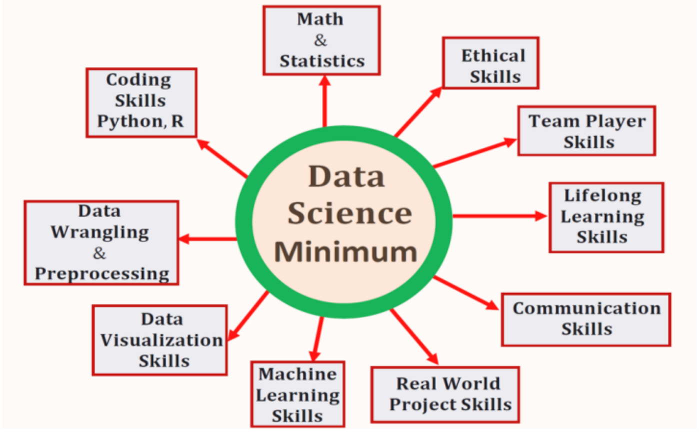
25



What is a data scientist?

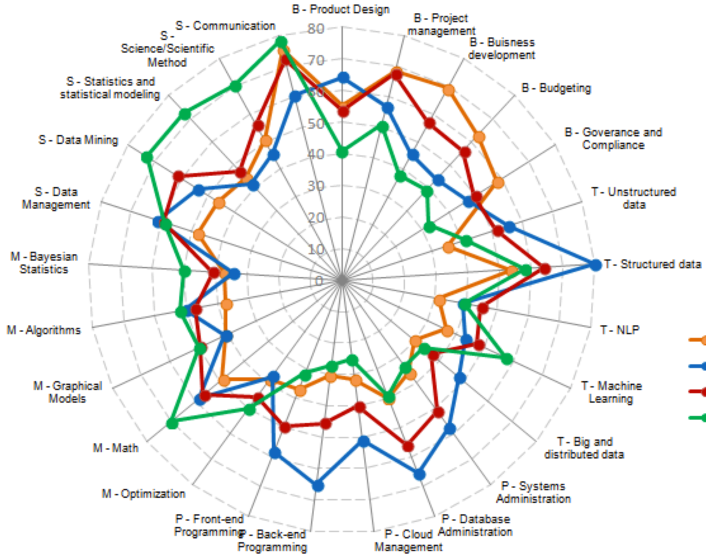
26

Tools and Skills for Data Scientists



Source: https://towardsdatascience.com/

Tools and Skills for Data Scientists



DS Skills by Job Role

DS Skill Category

B – Business
T – Technology
P – Programming
M – Mathematics & Modeling
S – Statistics

DS Role /Types of Data Scientists

- 🟧 Business Management (e.g., leader, business person, entrepreneur)
- 🟦 Developer (e.g., developer, engineer)
- 🟥 Creative (e.g., Jack of all trades, artist, hacker)
- 🟩 Researcher (e.g., researcher, scientist, statistician)

AnalyticsWeek
BUSINESS BROADWAY

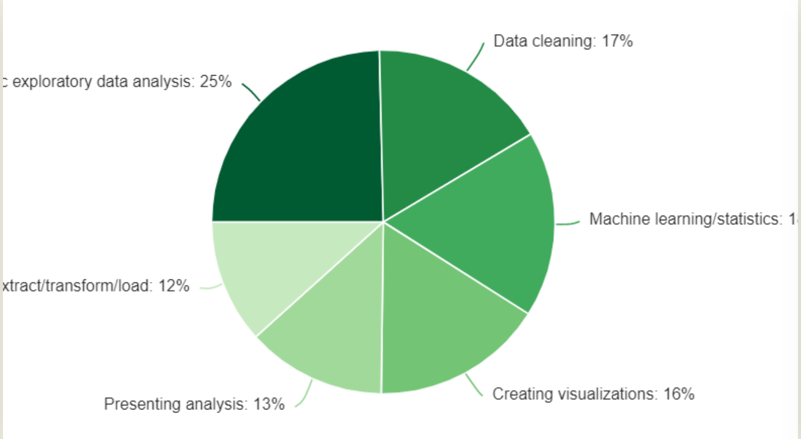
http://businessoverbroadway.com/wp-content/uploads/2015/09/datascienceblogroleskills.png

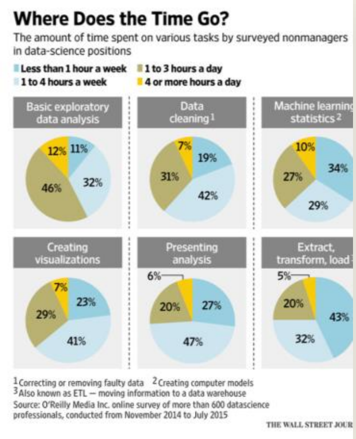
Analytical Tasks for Statisticians and Data Scientists

	Statistician	Data Scientist
Image	Baseball (Cricket)	HBR Sexiest Job of 21st Century
Mode	Reactive	Consultative
Works	Solo	In a team
Inputs	Data File, Hypothesis	A Business Problem
Data	Pre-prepared, clean	Distributed, messy, unstructured
Data Size	Kilobytes	Gigabytes
Tools	SAS, Mainframe	R, Python, awk, Hadoop, Linux, …
Nouns	Tables	Data Visualizations
Focus	Inference (why)	Prediction (what)
Output	Report	Data App / Data Product
Latency	Weeks	Seconds
Stars	G.E.P Box Trevor Hastie	Hilary Mason Nate Silver

29

What does a data scientist do in a typical workday?





Where Does the Time Go?

The amount of time spent on various tasks by surveyed nonmanagers in data-science positions

1 Correcting or removing faulty data 2 Creating computer models
3 Also known as ETL — moving information to a data warehouse

Source: O'Reilly Media Inc. online survey of more than 600 datascience professionals, conducted from November 2014 to July 2015

THE WALL STREET JOUR

30