

I. Data Science & Big Data: Basics

Introduction to Data Science

Patricio Mallea González¹

¹Data Scientist

May 6, 2020





Overview I

1. Why Data Science and Big Data?
2. Facets of Data
3. The Data Science Process
4. BigData Ecosystem: Frameworks and Tools



1. Why Data Science and Big Data?



1.1 Basics

Big Data

Blanket term for any collection of datasets so large or complex that it becomes difficult to process them using traditional Data Management.

Data Science

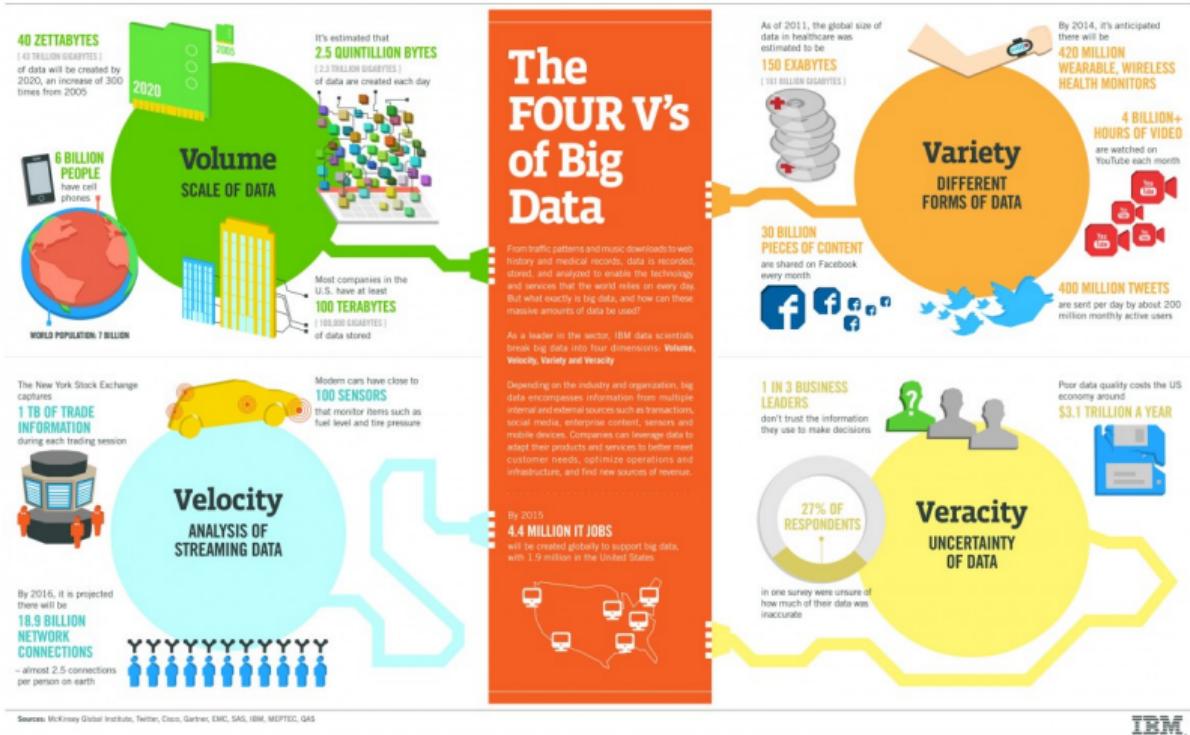
Involves using methods to analyze massive amounts of data and extract the knowledge it contains.

Challenges: Data capture, curation, storage, search, sharing, transfer, and visualization.

Skills of a DS: Good Statistics basis, Big Data understand, Machine Learning models, Algorithms Buildings (Hadoop, Pig, Spark, R, Python, Java, etc.), Databases (SQL and NoSQL).



1.1 Basics



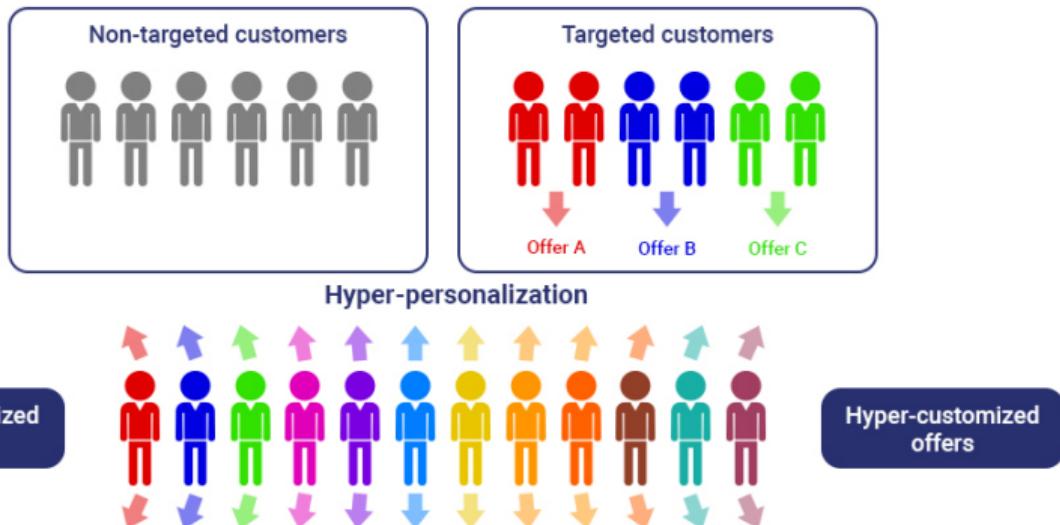


1.2 Uses of Data Science

(i) Commercial Companies: Hyper Personalization

What is hyper-personalization?

Hyper-personalization is an advanced and real-time customization of offerings, content and customer experience at an individual level. Designed to perfectly match a customer, hyper-personalization leverages Big Data to deliver such tailor-made solutions in real time.





1.2 Uses of Data Science

(ii) Commercial Companies: People Analytics

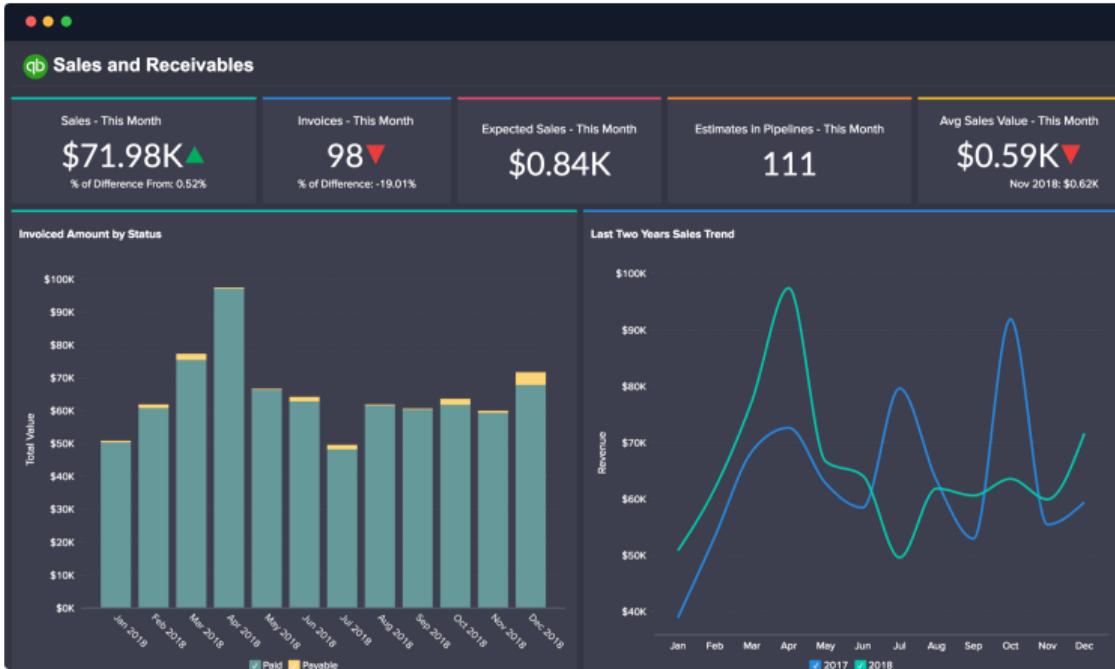




DataOwl

1.2 Uses of Data Science

(iii) Commercial Companies: Financial Analytics





1.2 Uses of Data Science

(iv) Governmental Organizations





1.2 Uses of Data Science

(v) Nongovernmental Organizations

Humanitarian Data Scientist

Ever wondered about the skills of a data specialist during a crisis? And how to become one?



Data Management

Represent and refine data
Data visualization
Data collection
Data system management
Research methodologies
Report design
Curiosity
Ability to translate technology



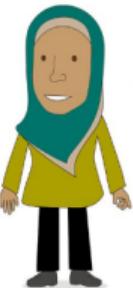
Humanitarian Business

Inter-cluster coordination
Needs Assessment
Humanitarian response planning
Appeal Prep & Response
CERF Request preparation
Financial Tracking of Humanitarian Aid
Donor relations
Advocacy (e.g. humanitarian principles)
Access monitoring & negotiations



Programming/Databases

Scripting Language like Python
Java
Filter & Mine Data
Machine Learn
Database SQL & NoSQL
Hadoop & Hive/Pig



Information Management

GIS & mapping
Survey methodology
Infographics & visualisation
Web development
Convene & coordinating Network



Statistics

Probability
Algorithms
Mathematics
R
Data Analysis
Statistics for programming like R



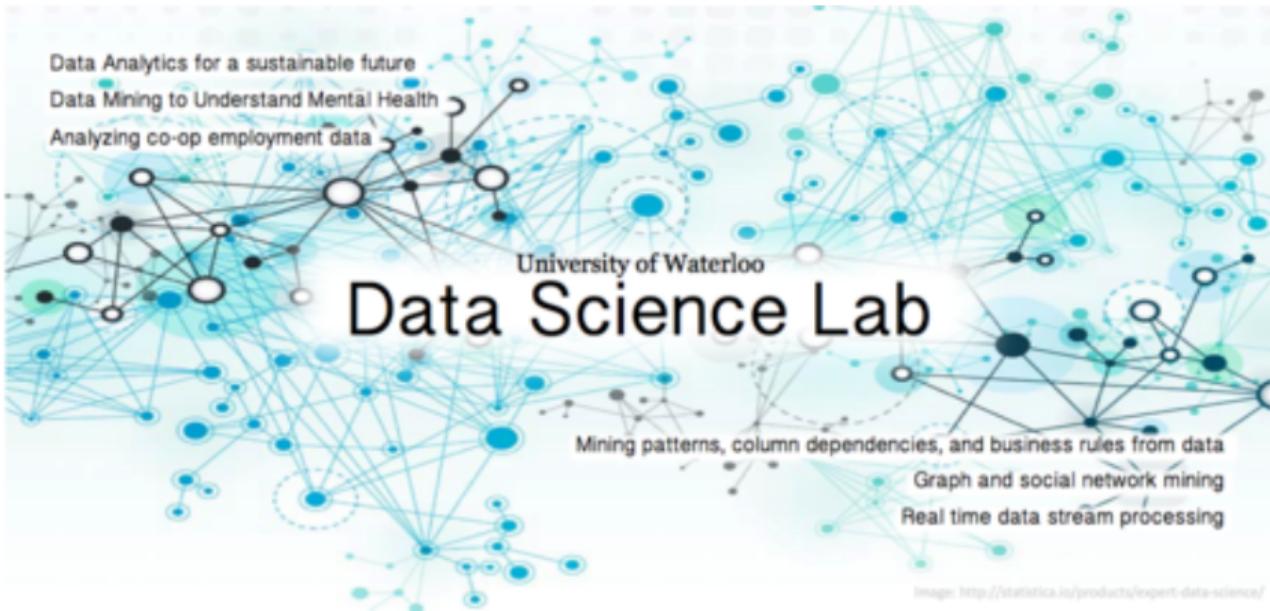
Inter Disciplinary

Meeting facilitation
Training
Staff management
Office management



1.2 Uses of Data Science

(vi) Universities



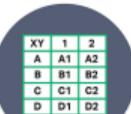


2. Facets of Data



Structured Data vs Unstructured Data

Can be displayed
in rows, columns and
relational databases



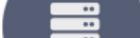
Numbers, dates
and strings



Estimated 20% of
enterprise data (Gartner)



Requires less storage



Easier to manage
and protect with
legacy solutions



vs

Unstructured Data

Cannot be displayed
in rows, columns and
relational databases



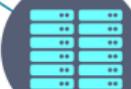
Images, audio, video,
word processing files,
e-mails, spreadsheets



Estimated 80% of
enterprise data (Gartner)



Requires more storage



More difficult to
manage and protect
with legacy solutions





(i) Structured Data: Tables

Category	This Year Sales Status	Average Unit Price	Last Year Sales	This Year Sales	This Year Sales Goal
010-Womens	●	\$7.30	\$2,680,662	\$1,787,958	\$2,680,662
020-Mens	🟡	\$7.12	\$4,453,133	\$4,452,421	\$4,453,133
030-Kids	🟡	\$5.30	\$2,726,892	\$2,705,490	\$2,726,892
040-Junior	●	\$7.00	\$3,105,550	\$2,930,385	\$3,105,550
050-Shoes	🟡	\$13.84	\$3,640,471	\$3,574,900	\$3,640,471
060-Intimate	●	\$4.28	\$955,370	\$852,329	\$955,370
070-Hosiery	●	\$3.69	\$573,604	\$486,106	\$573,604
080-Accessories	🟢	\$4.84	\$1,273,096	\$1,379,259	\$1,273,096
090-Home	🟢	\$3.93	\$2,913,647	\$3,053,326	\$2,913,647
100-Groceries	🟢	\$1.47	\$810,176	\$829,776	\$810,176
Total	🟡	\$5.49	\$23,132,601	\$22,051,952	\$23,132,601

(ii) Unstructured Data: Natural Language

Capítulo 14

Crecimiento económico con ahorro óptimo*

En los capítulos anteriores hemos analizado el crecimiento asumiendo que la tasa de ahorro es constante e igual a s . Aunque en una primera aproximación esta es una buena idea, tiene también algunas limitaciones. La primera es que el crecimiento al final depende de lo que pase con el crecimiento de la productividad y otros factores, todo lo cual debiera incidir en la tasa de ahorro. Solo podemos especular acerca de cómo cambia la tasa de ahorro sin mayores fundamentos. Y en segundo lugar, desde el punto de vista de tener una buena teoría de crecimiento que nos permita analizar el bienestar, se debe tener un modelo bien especificado, que incluya la utilidad de los hogares.

(iii) Unstructured Data: Machine Generated Data

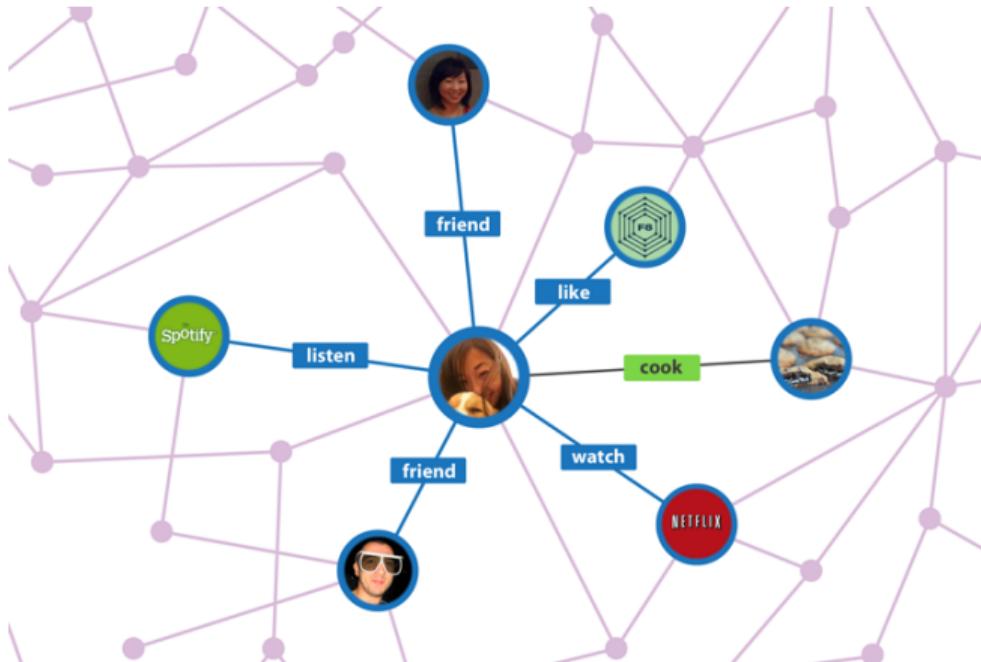
```

[08/04/00 14:28:27](0)      Server: PSNT sleeping for 15 seconds
[08/04/00 14:28:41](0)      Server: PSNT looking for work
[08/04/00 14:28:41](5)      Server: PSNT checking status...
[08/04/00 14:28:41](5)      Server action mode: Ok (looking for requests)
[08/04/00 14:28:41](5)      Checking Process cancels...
[08/04/00 14:28:41](4)      Checking status of active processes...
[08/04/00 14:28:41](5)      Process 9836 is still running as Session ID 711
[08/04/00 14:28:41](5)      Process 9837 is still running as Session ID 634
[08/04/00 14:28:41](5)      Process 9838 is still running as Session ID 703
[08/04/00 14:28:41](5)      Info for array of Request(s) associated with a Job slated to be submitted
[08/04/00 14:28:41](5)          Size of array: 1
[08/04/00 14:28:41](5)      Info for array of Active Processes
[08/04/00 14:28:41](5)          Size of array: 3
[08/04/00 14:28:41](5)              Crystal : Active: 3 Max: 3
[08/04/00 14:28:41](5)      Server: PSNT checking status...
[08/04/00 14:28:41](5)      Server action mode: Submitting request
[08/04/00 14:28:41](5)      Number of New Process Request(s) To Start: 1
[08/04/00 14:28:41](1)      Process Instance: 9843 started (PID: 645)
[08/04/00 14:28:41](4)      Starting process:: 9843
[08/04/00 14:28:41](4)          Command Line: Y:\BIN\CLIENT\WINX86\PSSQR.EXE
[08/04/00 14:28:41](4)          Farm List: -CT ORACLE -CS -CD E800R21B -CA %ACCESSID% -CAP %ACCESSPSWD%
[08/04/00 14:28:41](4)          Working Dir: c:\apps\db\oracle8i\bin
[08/04/00 14:28:41](4)              Session Id: 645
[08/04/00 14:28:41](0)      Server: PSNT sleeping for 14 seconds
[08/04/00 14:28:55](0)      Server: PSNT looking for work
[08/04/00 14:28:55](5)      Server: PSNT checking status...
[08/04/00 14:28:55](5)      Server action mode: Ok (looking for requests)
[08/04/00 14:28:55](5)      Checking Process cancels...
[08/04/00 14:28:55](4)      Checking status of active processes...
[08/04/00 14:28:55](5)      Process 9836 is still running as Session ID 711
[08/04/00 14:28:55](5)      Process 9837 is still running as Session ID 634
[08/04/00 14:28:55](5)      Process 9838 is still running as Session ID 703
[08/04/00 14:28:55](5)      Info for array of Request(s) associated with a Job slated to be submitted
[08/04/00 14:28:55](5)          Size of array: 0
[08/04/00 14:28:55](5)      Info for array of Active Processes
[08/04/00 14:28:55](5)          Size of array: 3
[08/04/00 14:28:55](5)              Crystal : Active: 3 Max: 3
[08/04/00 14:28:55](5)      Info for array of Queued Request(s) found in Process Request table
[08/04/00 14:28:55](5)          Size of array: 16

```



(iv) Unstructured Data: Graph-based or Network Data



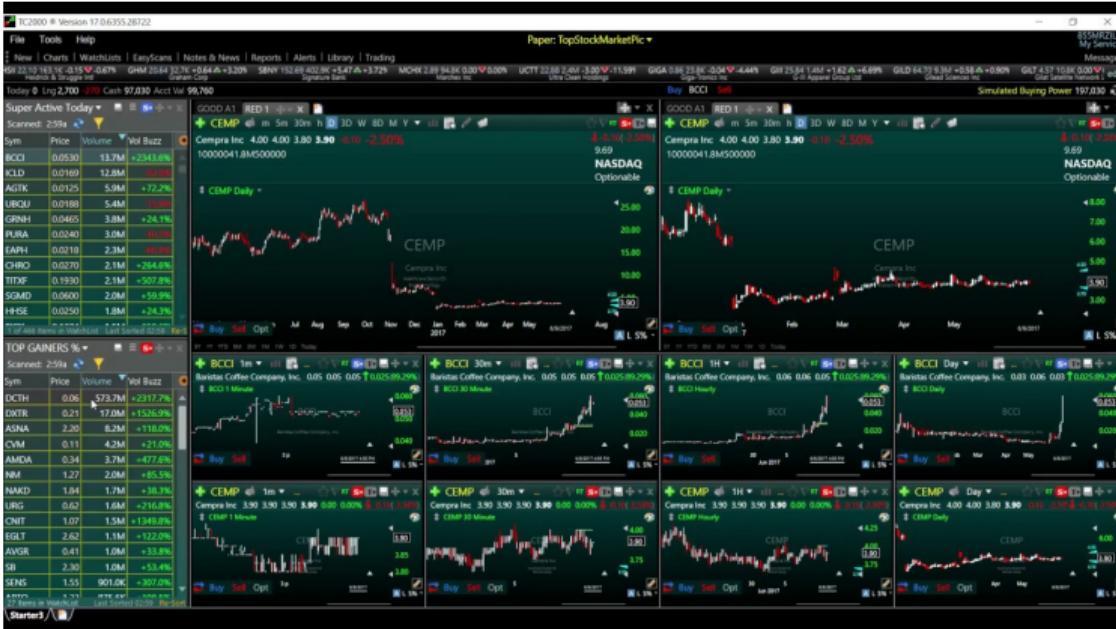
(v) Unstructured Data: Audio, Image, and Video





DataOwl

(vi) Structured/Unstructured Data: Streaming Data

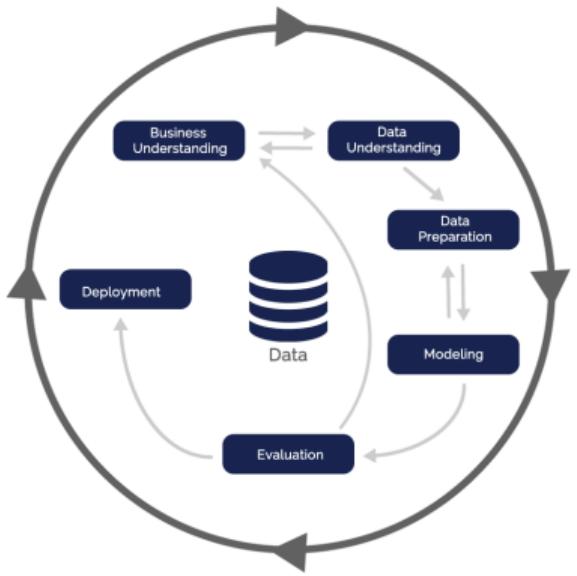




3. The Data Science Process

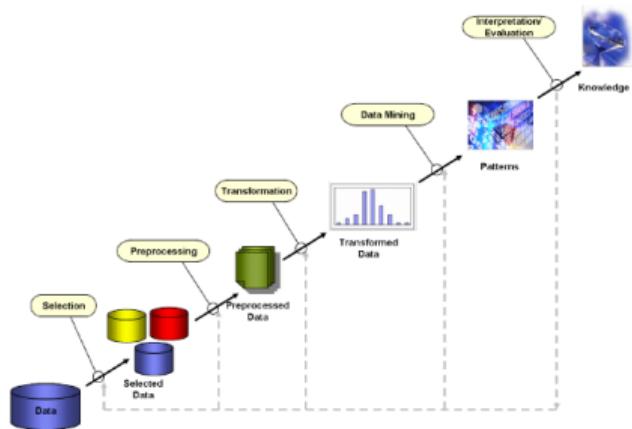


3. The Data Science Process: An approaching



The **CRISP-DM** methodology (*Cross-Industry Standard Process for Data Mining*) is a lifecycle process. Consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary.

3. The Data Science Process: An approaching



KDD (Knowledge Discovery in Databases) is an iterative process where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.

3. The Data Science Process: An approaching



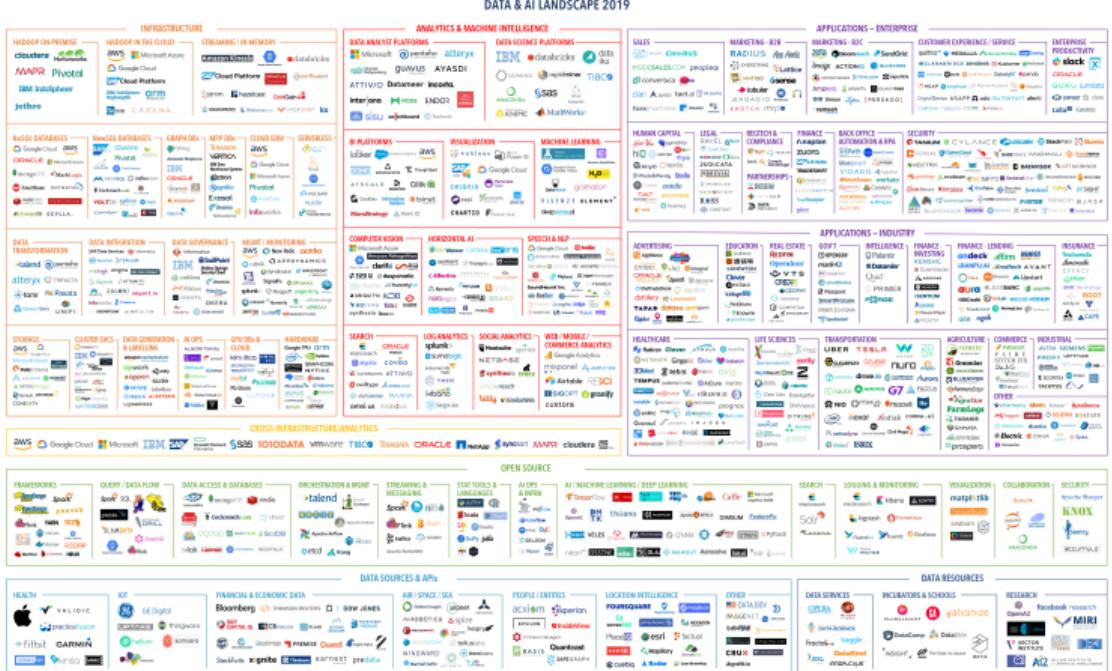
SAS Institute defines data mining as the process of Sampling, Exploring, Modifying, Modeling, and Assessing (SEMMA) large amounts of data to uncover previously unknown patterns which can be utilized as a business advantage.



4. BigData Ecosystem: Frameworks and Tools



4. The Big Data Ecosystem



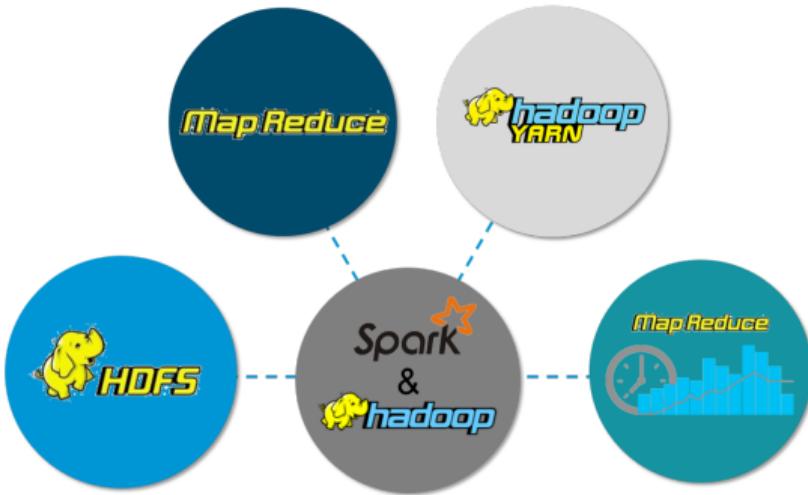
July 16, 2019 - FINAL 2019 VERSION

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap) mattturck.com/data2019

FIRSTMARK
EARLY STAGE VENTURE CAPITAL



4.1 Distributed File Systems





4.2 Distributed Programming Framework





4.3 Machine Learning Frameworks



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



4.4 NoSQL Databases

ORACLE®



mongoDB



cassandra



A. Tools for the Course



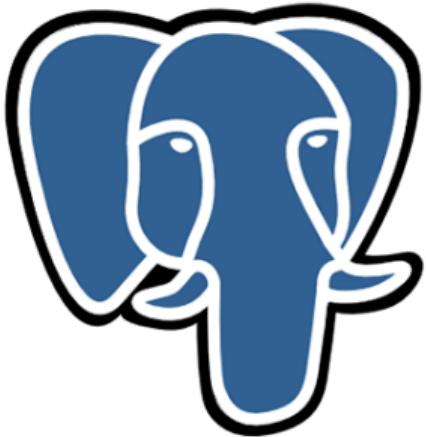
A.1 Anaconda



Anaconda is a distribution of Python (and R). It is free and open-source and makes package management and deployment simpler. It is the standard platform for Python data science and open-source machine learning.



A.2 PostgreSQL



PostgreSQL

PostgreSQL, also known as Postgres, is a free and open-source relational database management system (RDBMS) emphasizing extensibility and SQL compliance.

I. Data Science & Big Data: Basics

Introduction to Data Science

Patricio Mallea González¹

¹Data Scientist

May 6, 2020

