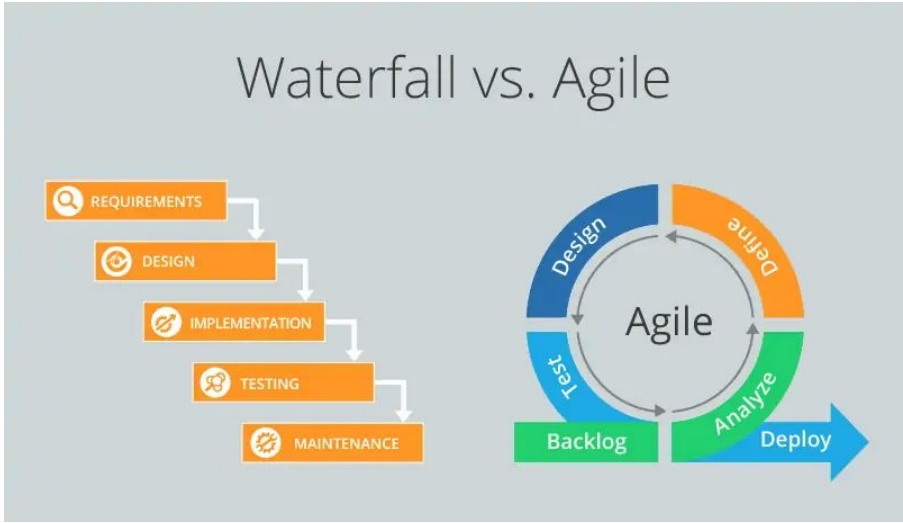
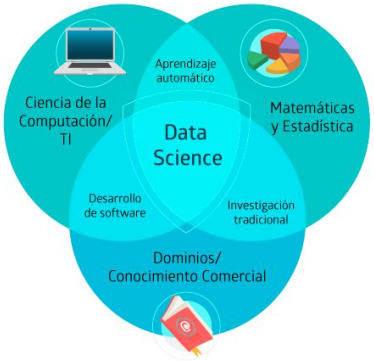
Why **Agile Data Science**?

Agenda

- Agile Data Science
- Introduction to Data Science
- Data Product Chain
- Data Product Lifecycle
- Machine Learning Lifecycle

Agile methodologies….





… for Data Science projects

Why?


Enjoyable
Higher Revenue
Increased Productivity
Happy Stakeholder
AGILE
guidance
Time-to-Market
efits
of
AGILE
ALL THE COMPANIES ARE WORKING IN THIS WAY

Introduction to Data Science

Data…


3 Important Statistics About How Much Data Is Created Every Day
FinancesOnline
REVIEWS FOR BUSINESS
1 How much data is generated every minute?
Source: Domo
41,666,667
messages shared by WhatsApp users
1,388,889
video / voice calls made by people worldwide
404,444
hours of video streamed by Netflix users
347,222
stories posted by Instagram users
150,000
messages shared by Facebook users
147,000
photos shared by Facebook users
2 Estimated Data Consumption from 2021 to 2024
Source: IDC / Statista
200
150
100
50
74 ZETTABYTES
94 ZETTABYTES
118 ZETTABYTES
149 ZETTABYTES
2021
2022
2023
2024
3 Data Growth in 2021
Sources: TechJury, Internet Live Stats, Cisco, PurpleSec
2 TRILLION
searches on Google by the end of 2021
1.134 TRILLION MB
volume of data created every day
3,026,626
emails sent every second, 67% of which are spam
278,108 PETABYTES
global IP data per month by the end of 2021
230,000
new malware versions created every day
82%
share of video in total global internet traffic at the end of 2021

…and 1 ZB is ???? Bytes… how many zeros?

Multiple-byte units V·T·E

Decimal			Binary				
Value	Metric		Value	IEC		Legacy	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte	TB	terabyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte		–
1000^6	EB	exabyte	1024^6	EiB	exbibyte		–
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte		–
1000^8	YB	yottabyte	1024^8	YiB	yobibyte		–

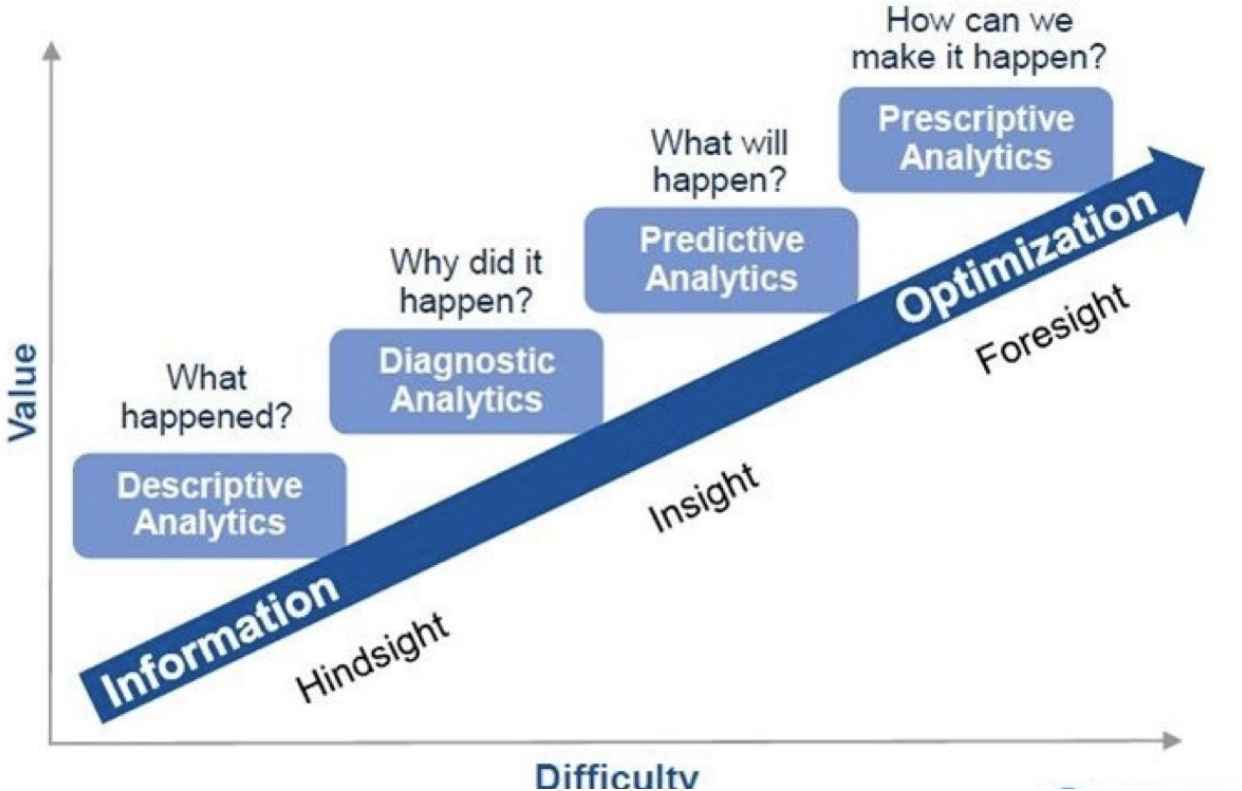
Orders of magnitude of data

1 ZB = 10000000000000000000 Bytes

So we have a lot of data, now…

… we need to know what are we going to do with it

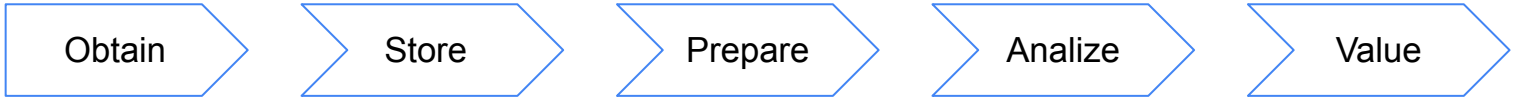
Data Value Chain



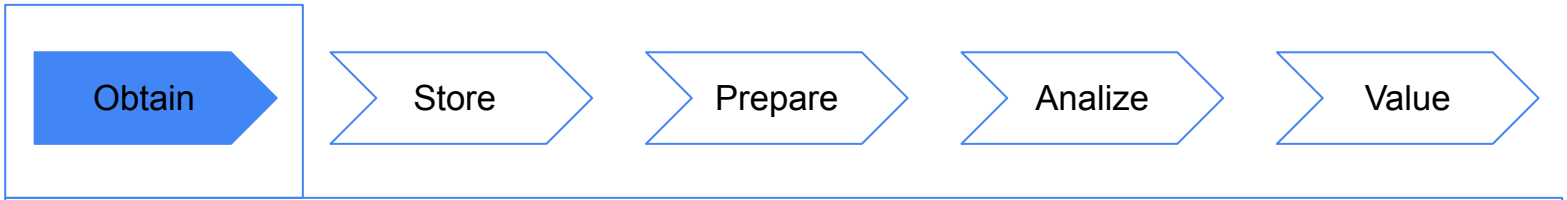
Can we define what is a Data Product?

- for instance…

Data Product chain



Data Product chain



We need to obtain the data from the different data sources, like:

- ❏ Structured internal/external data (customers, employees, products, sales, …)
- ❏ Unstructured internal/external data (documents, e-mails, videos, photos, …)
- ❏ Open Data (financieros, metereológicos, demográficos, …)
- ❏ IoT Data (from devices in the field…)

Every data source has their own characteristics!

Data Product chain

We need to store the data:

- ❏ File systems
- ❏ Relational Databases (MS SQL Server, Oracle, …)
- ❏ Data Warehouses
- ❏ Data Lakes

And for each one, we have to manage the security and availability!

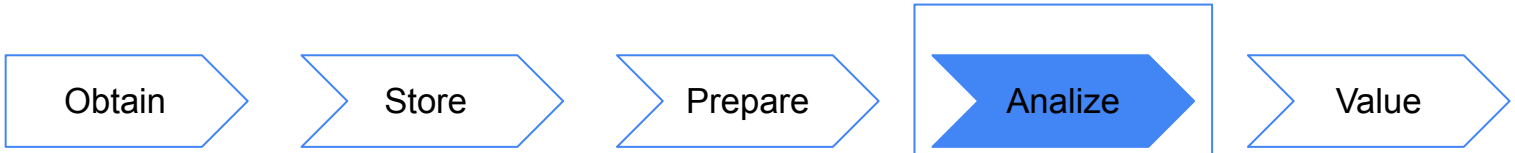
Data Product chain

Usually, we can not use the data as-is, we need to prepare it:

- ❏ Delete duplicates
- ❏ Impute missing values
- ❏ Normalize numerical features
- ❏ Add meta-data

This is the most time-consuming process but it is absolutely necessary. Remember: Garbage, In Garbage Out. If we are not using clean data, we can not trust the insights.

Data Product chain

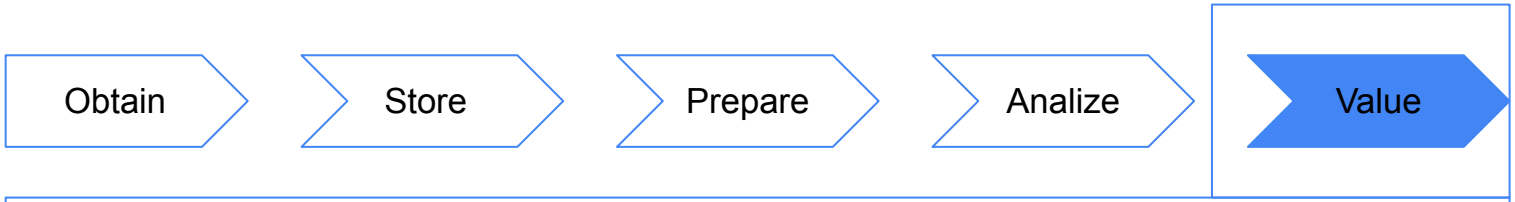


Once we have the data ready, we can start to use it to create value:

- ❏ Doing exploratory and descriptive analytics
- ❏ Creating Predictive models
- ❏ Clustering data
- ❏ …

Every process that could help us to understand the data and generate value.

Data Product chain



Finally the last step and our main target, generate value! It could be:

- Reports
- Dashboards
- ML Models
- Recommendation engines
- …

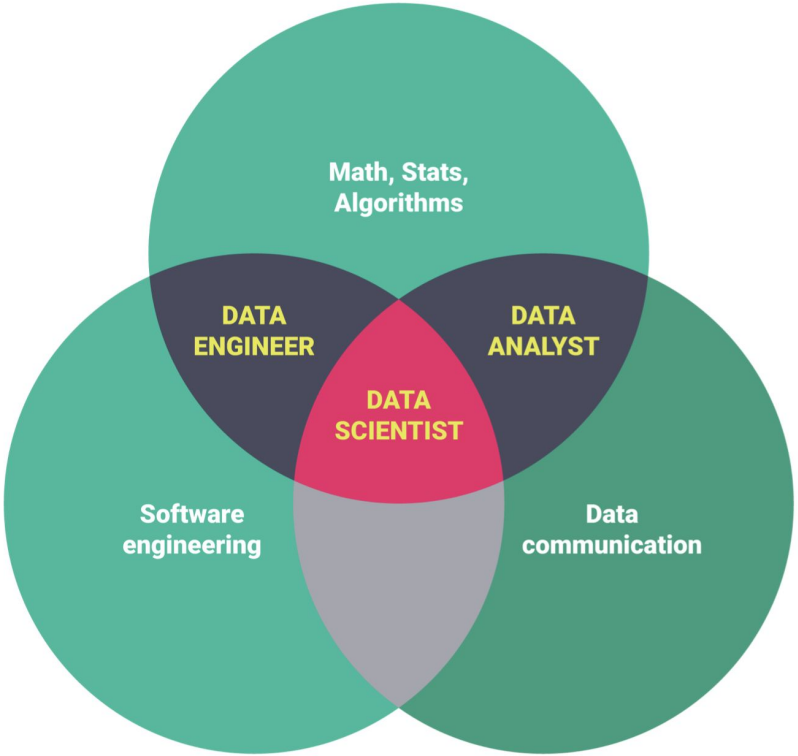
Every possible product that generates value with data (cost reduction, process optimization or new business models) is a valid one!

Data Science, the way to create Data Products

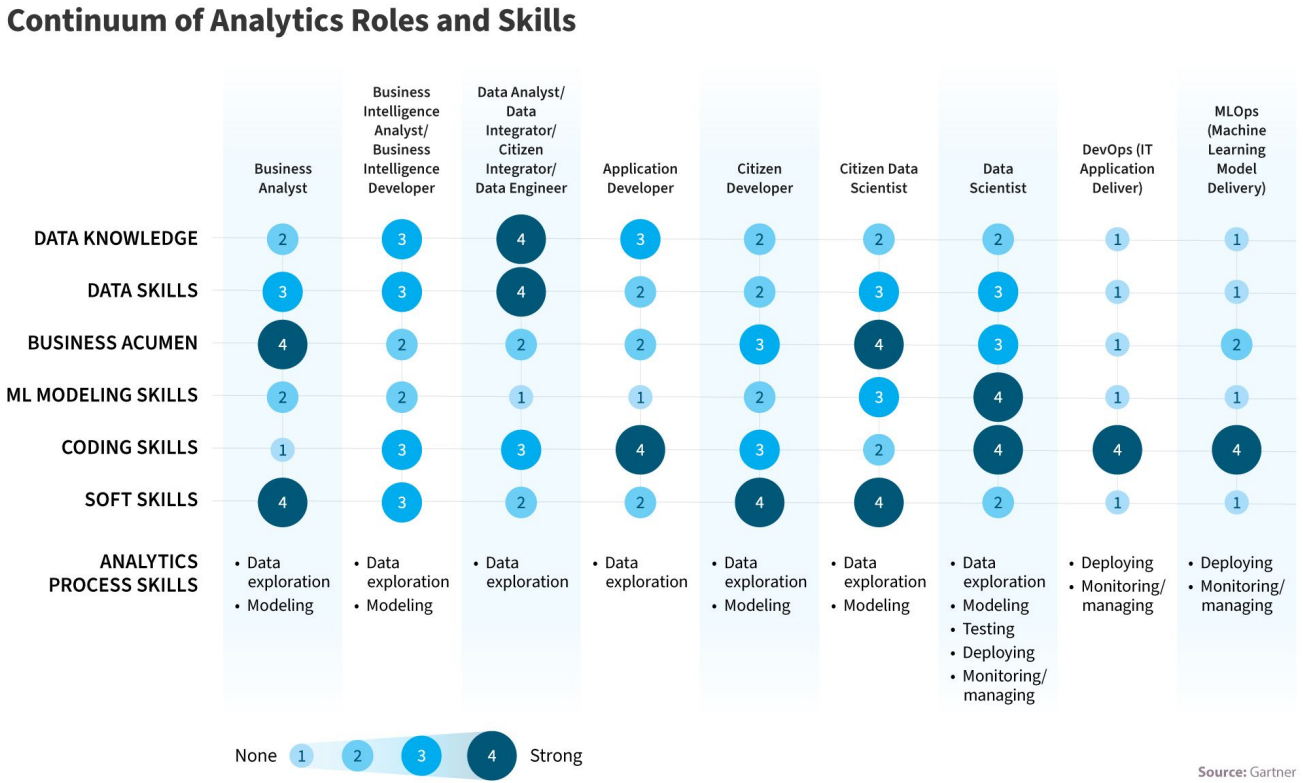
Data Science is a **methodology** to define:

1. what we want to do with data,
2. how do we evaluate our actions,
3. what decisions can be grounded on data,
4. how do we combine evidences from several sources.

Data Product development is a team sport

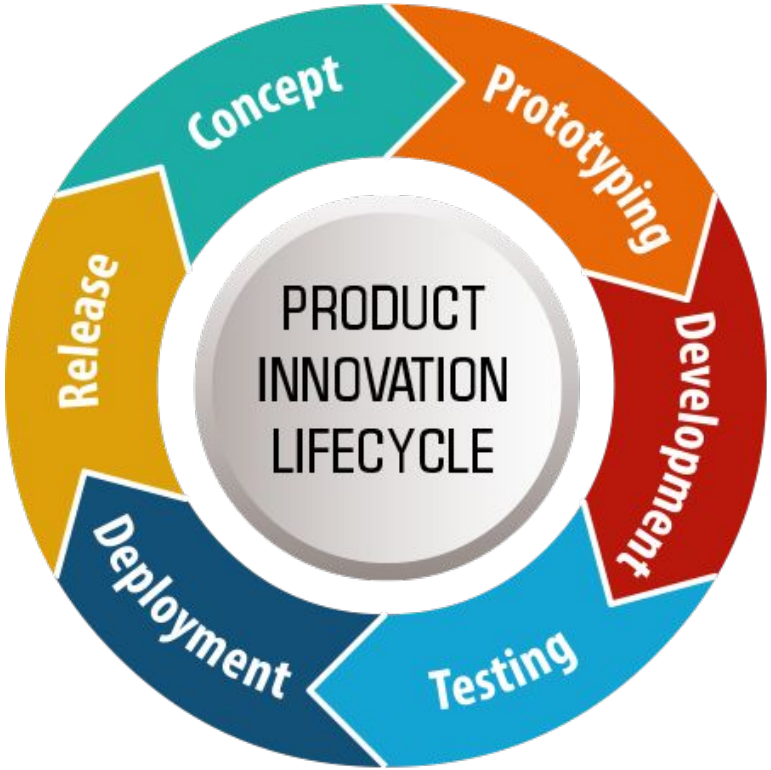


Much more complex that you can imagine…

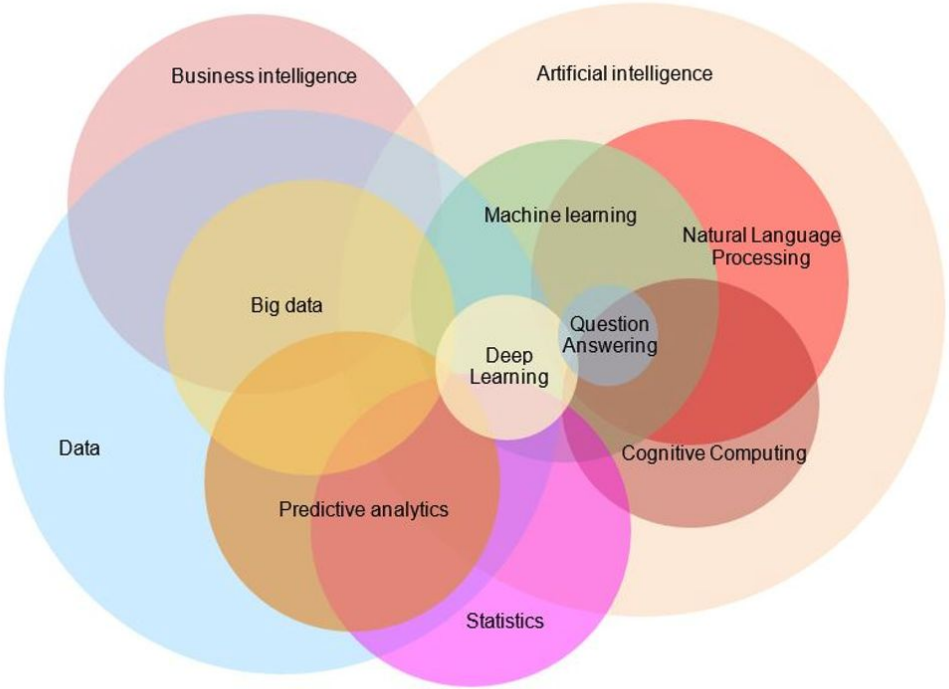


why? The Data Product Lifecycle

Build data products is not any more just to run a Notebook to train a model in python… there are much more steps



Machine Learning… (aka AI in the Enterprises)



Machine Learning… 2 cents

- **Machine learning** is a subfield of artificial intelligence.
- Its goal is ***to enable computers to learn on their own***.
- A machine's learning algorithm enables computers
  - ***to identify patterns in observed data***,
  - build models that explain the world, and
  - predict ***things without having explicit pre-programmed rules*** and models

Machine Learning… 2 cents

"If we are ever to make a machine that will speak, understand or translate human languages, solve mathematical problems with imagination, practice a profession or direct an organization, either we must reduce these activities to a science so exact that we can tell a machine precisely how to go about doing them or we must develop a machine that can do things without being told precise."

R.M. Friedberg – IBM Journal of Research - 1958

Machine Learning… 2 cents

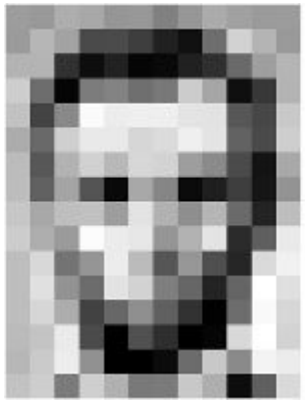
- Four legs
- Snout
- Tail
- Hair

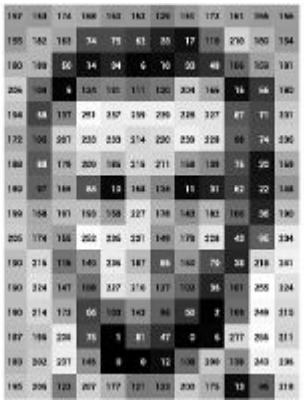
Machine Learning… 2 cents

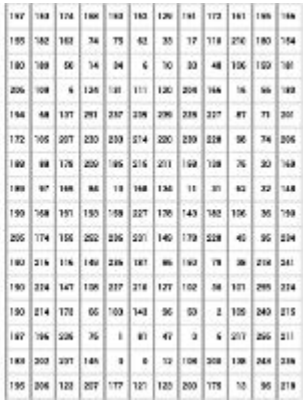
- Soft fur
- Long ears

Machine Learning… 2 cents

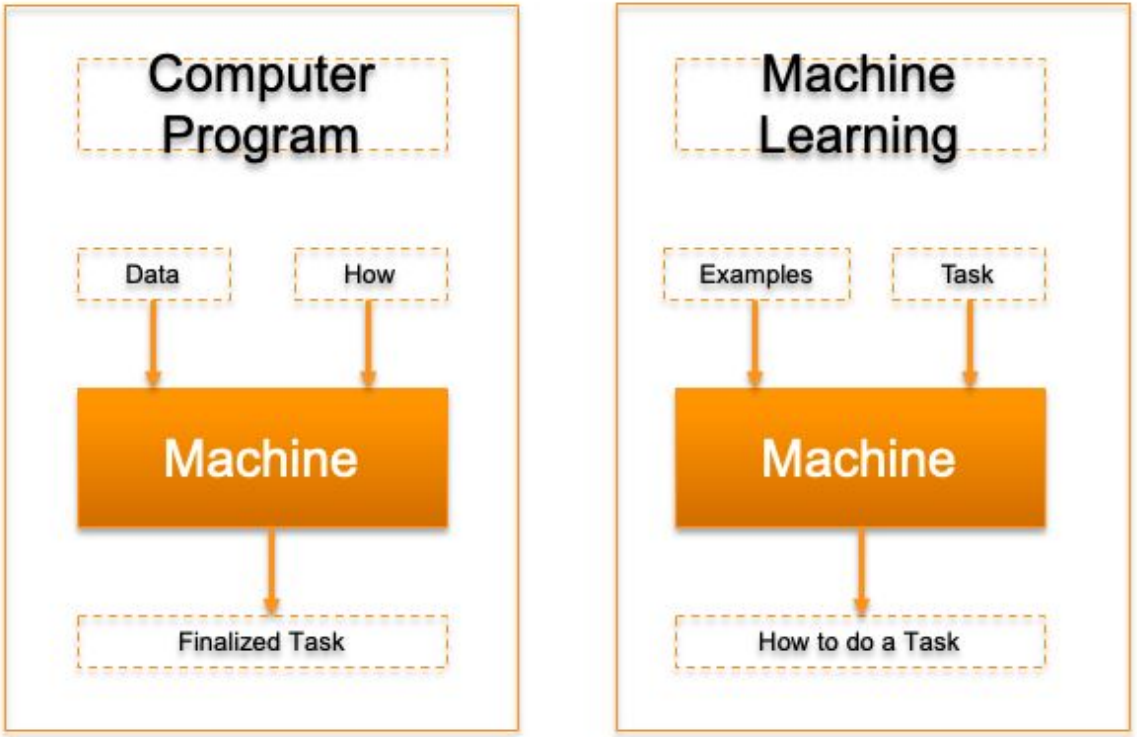
Machine Learning… 2 cents



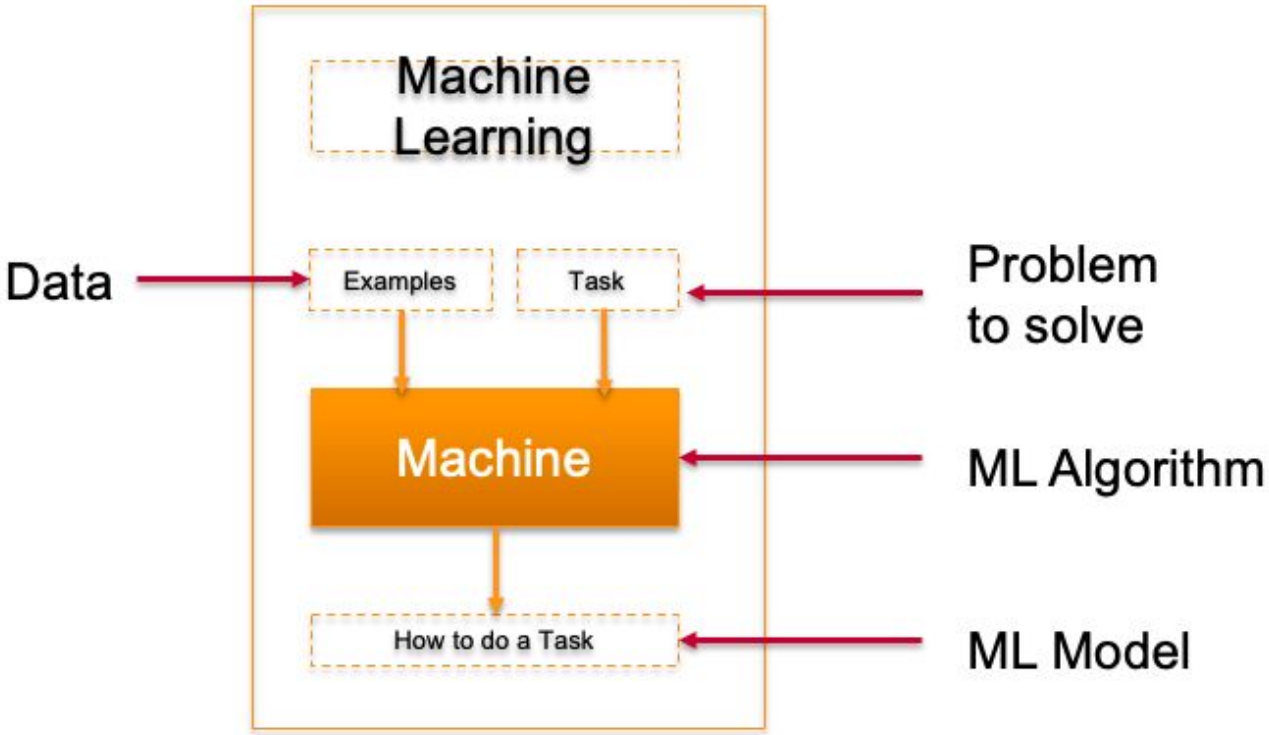




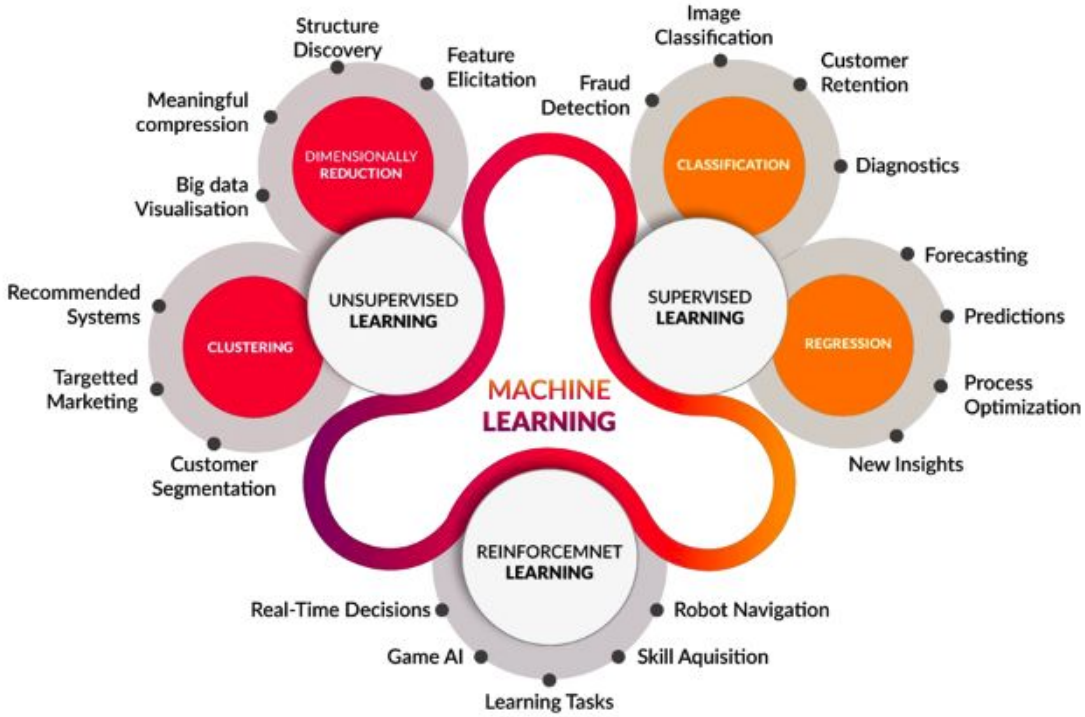
Machine Learning… 2 cents



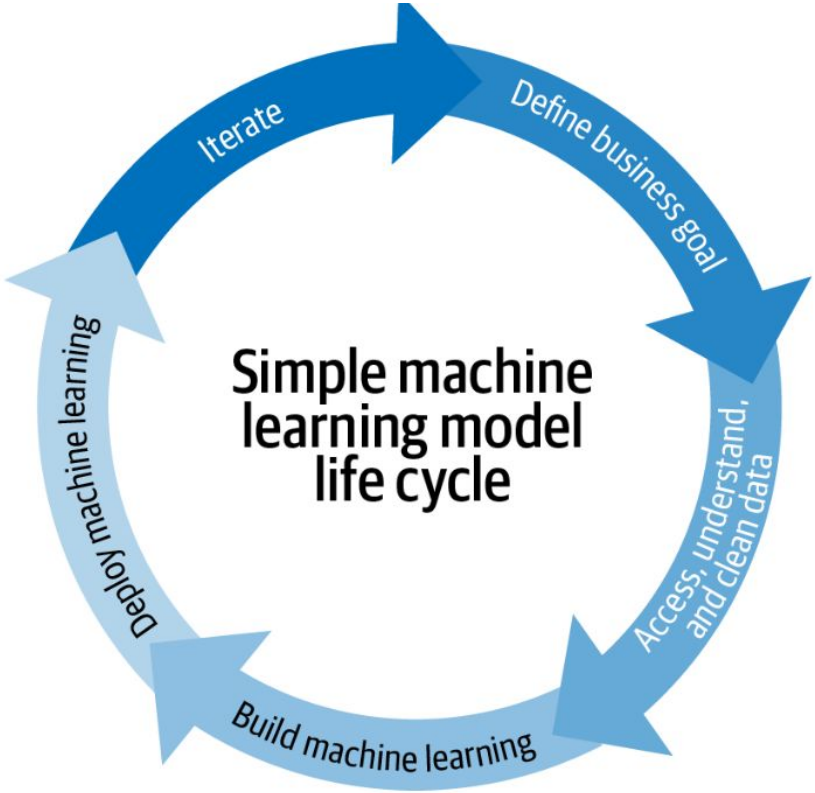
Machine Learning… 2 cents



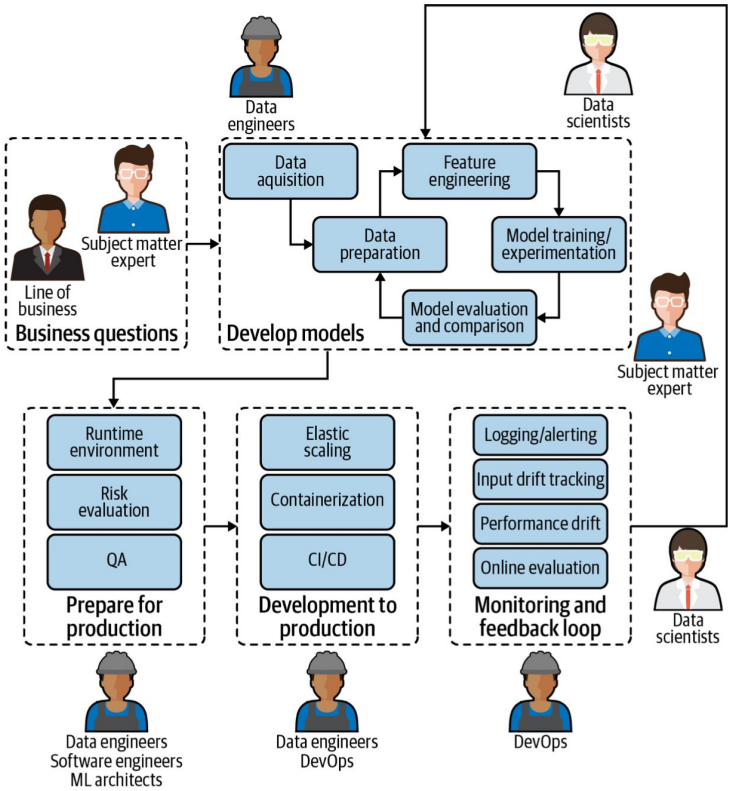
Machine Learning… 2 cents



The ML Model Lifecycle (simple version)



The ML Model Lifecycle (real version)



so… why Agile Data Science?

- Data products development is a team game
- A lot of things may change, we need an agile mindset and agile methodologies
- ML is tricky, we need a way to deal with the complexities
- ML lifecycle needs to be managed, so you need to understand the basic concepts and tools
- Enterprises and companies are working in this way, so you need to understand the basic concepts and tools