TU 257 – Fundamentals of Data Science

Data Analytics

L1 - Introduction

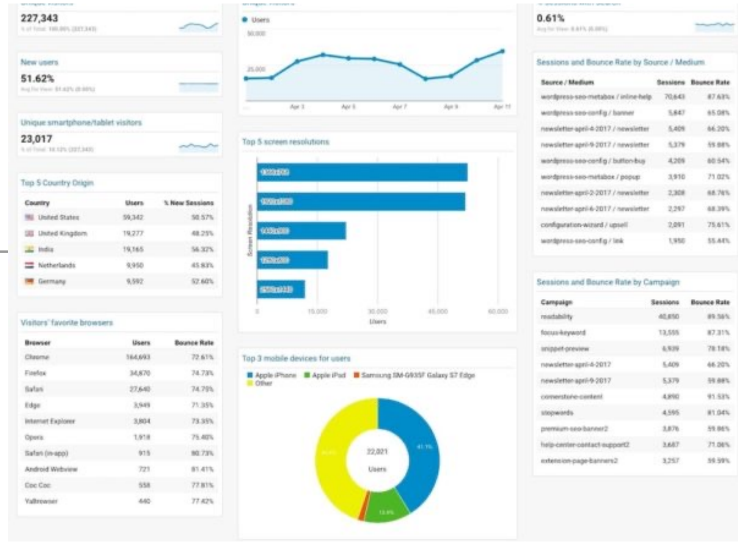
Brendan Tierney

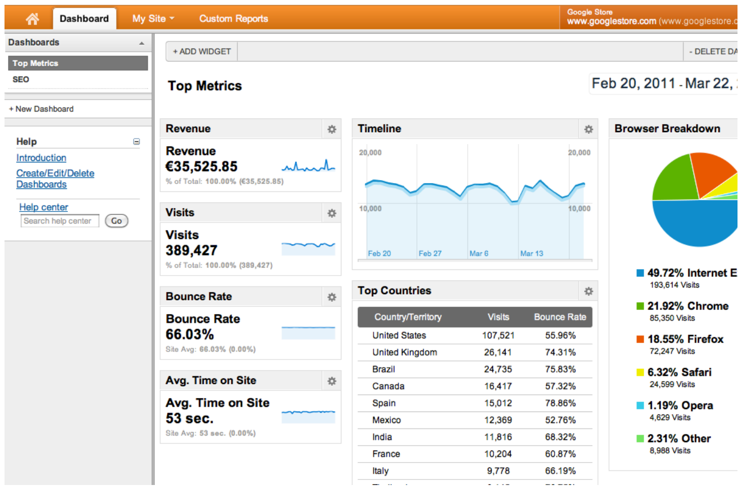
Agenda

- Introduction
- Different types/stages of Analytics
- Application areas
- No Free Lunch
- Where Machine Learning fits in
- Analysing Data Challenge
- What product/tool/language/package to use

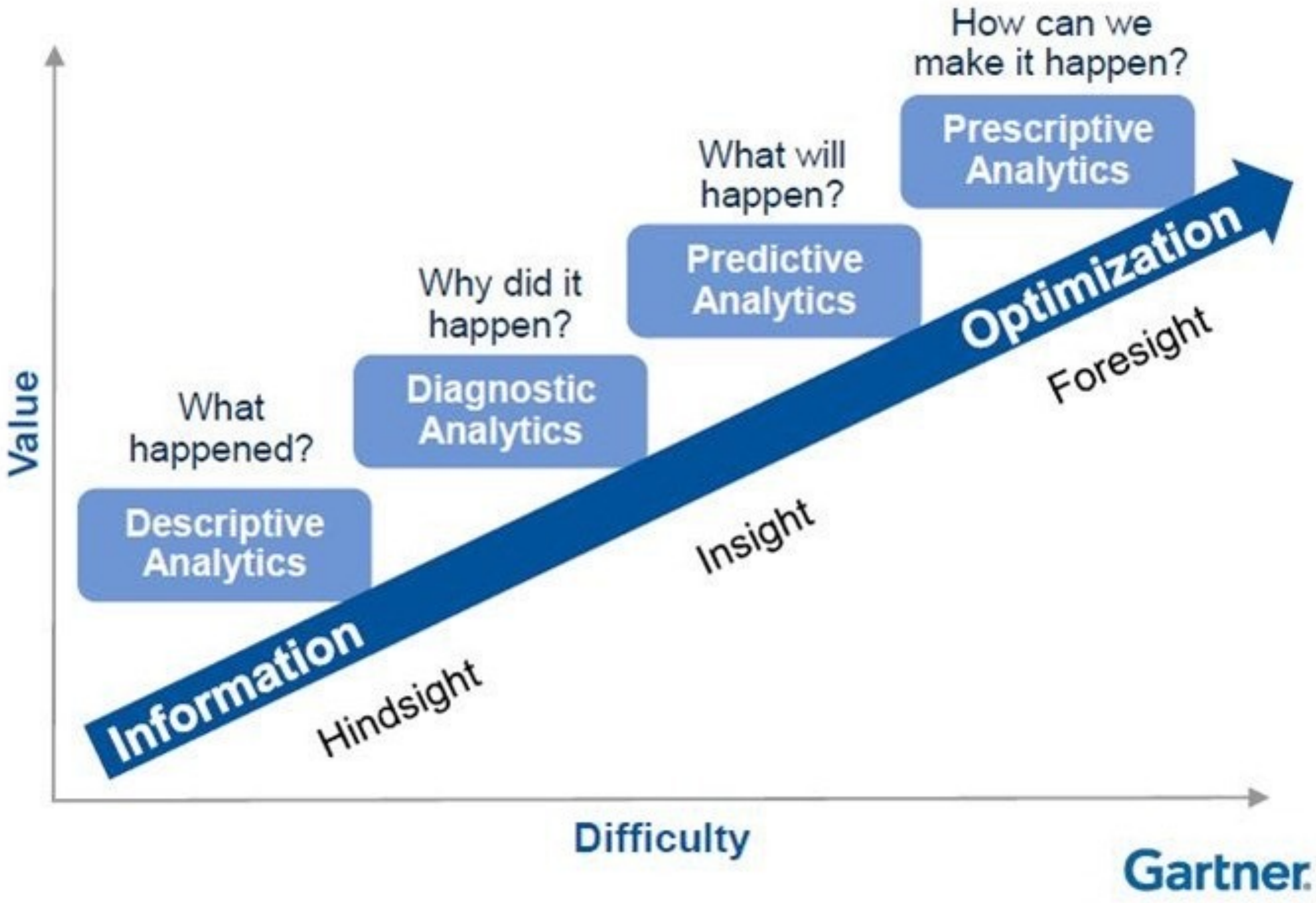
The Analytics Challenge

- We have a wide variety of Tools to help use Analyse Data
  - BI Tool
  - SQL
  - Every programming language
  - Observing, viewing, inspecting
- We can use the Dashboard to see & understand what is happening
- We can make predictions on what we see (visual predictive analytics)
- But, Humans can only process a certain amount of information at the one time
  - Maybe 3 or 4 attributes x 3 or 4 values
- We cannot see complex patterns in our data -> we need other tools

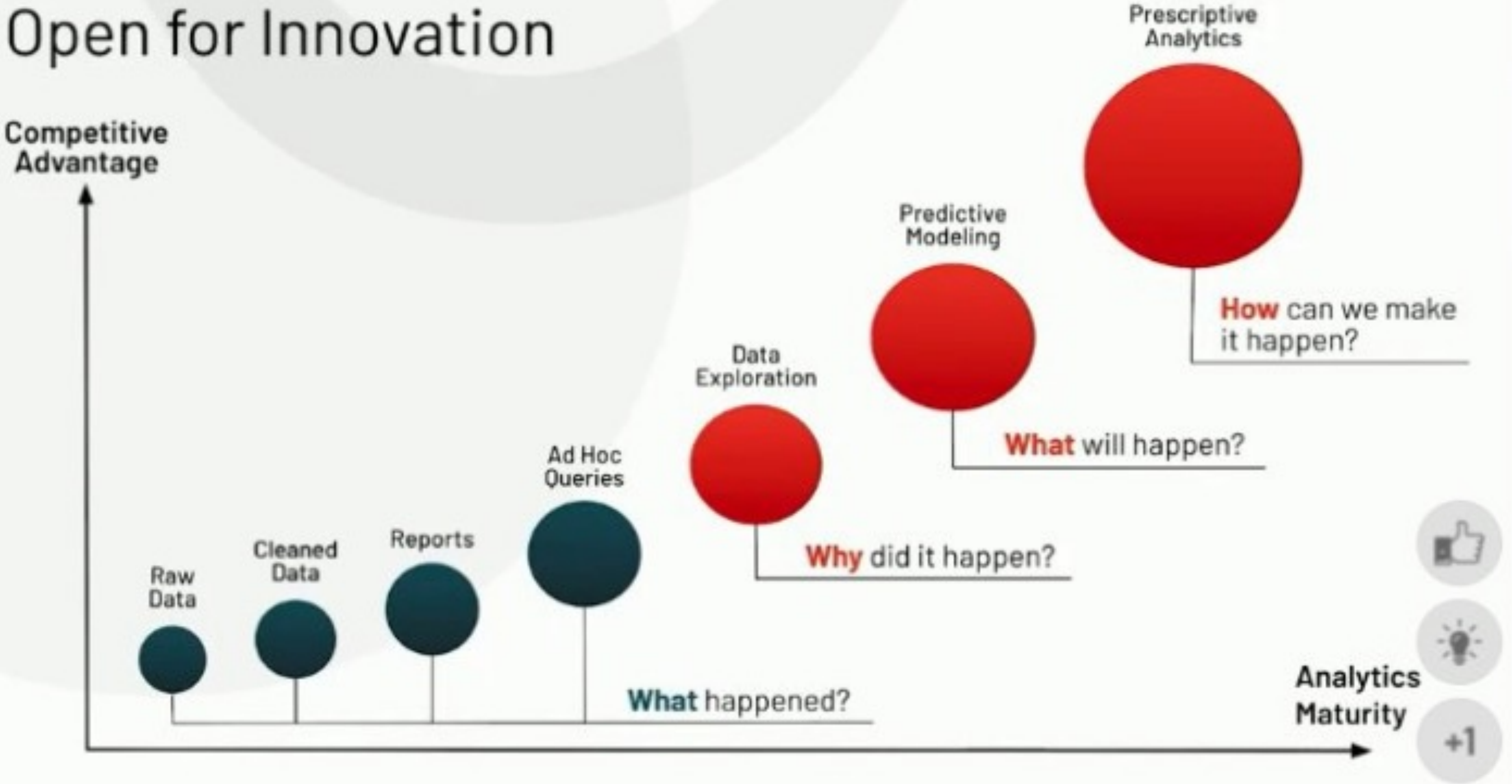




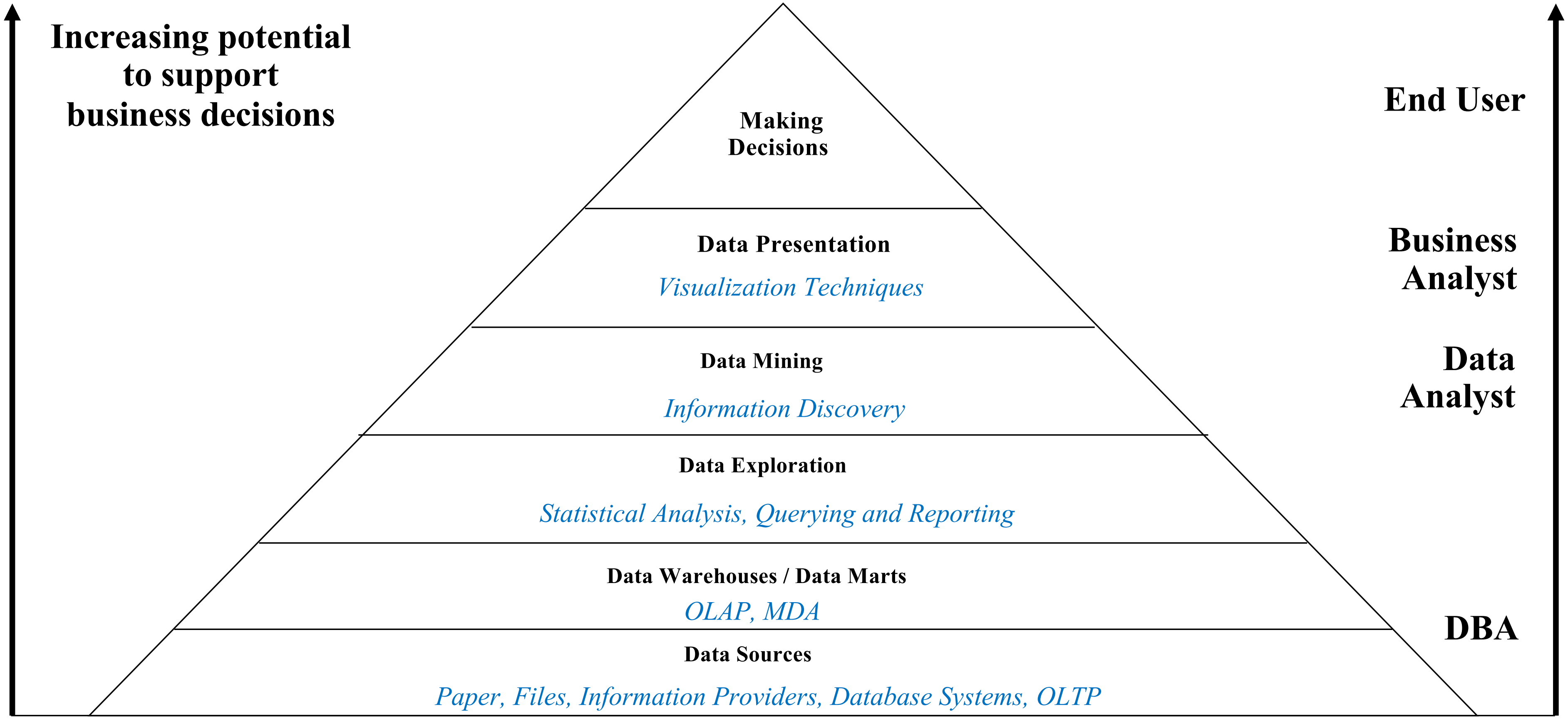
Different Types



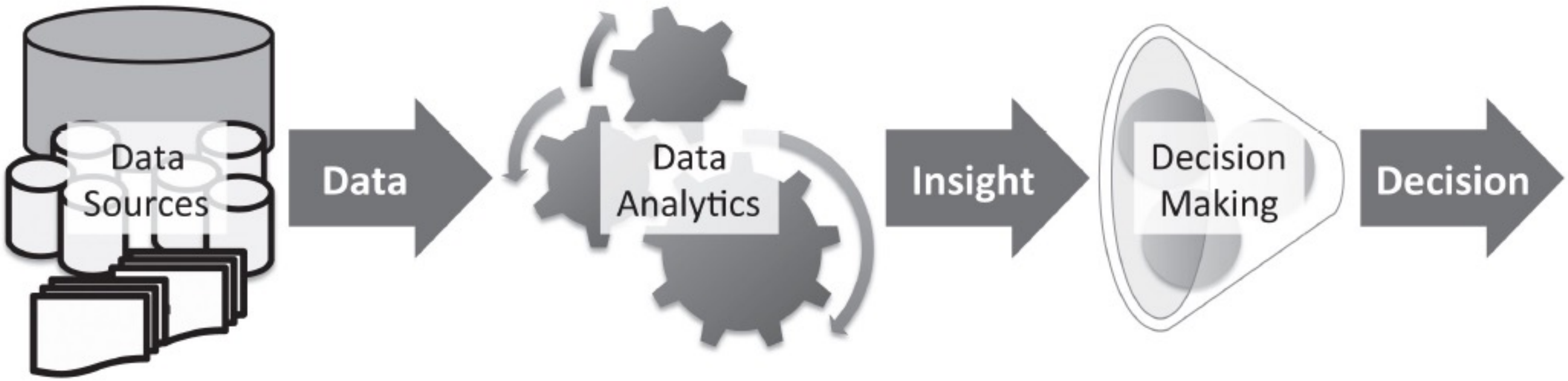
Open for Innovation



Data Analytics



- Predictive Data Analytics encompasses the business and data processes and computational models that enable a business to make data-driven decisions



Data Analytics moving from Data to Insights to Decisions

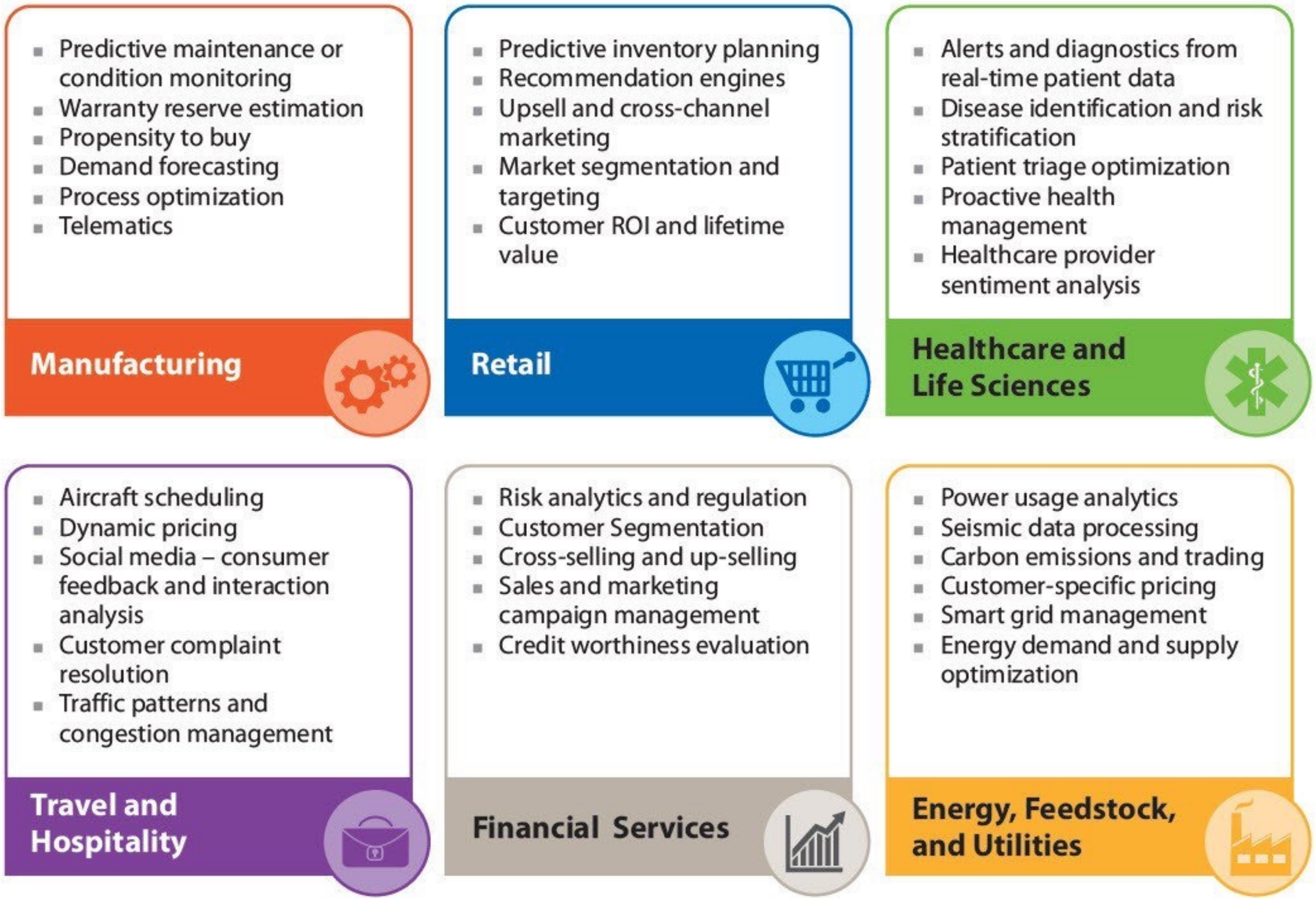
Application Areas

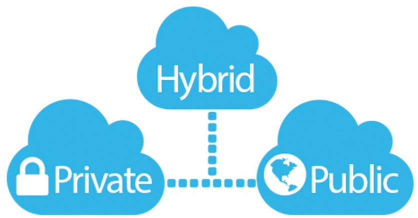
- It's Everywhere !

- And if it isn't being used now, it will soon.

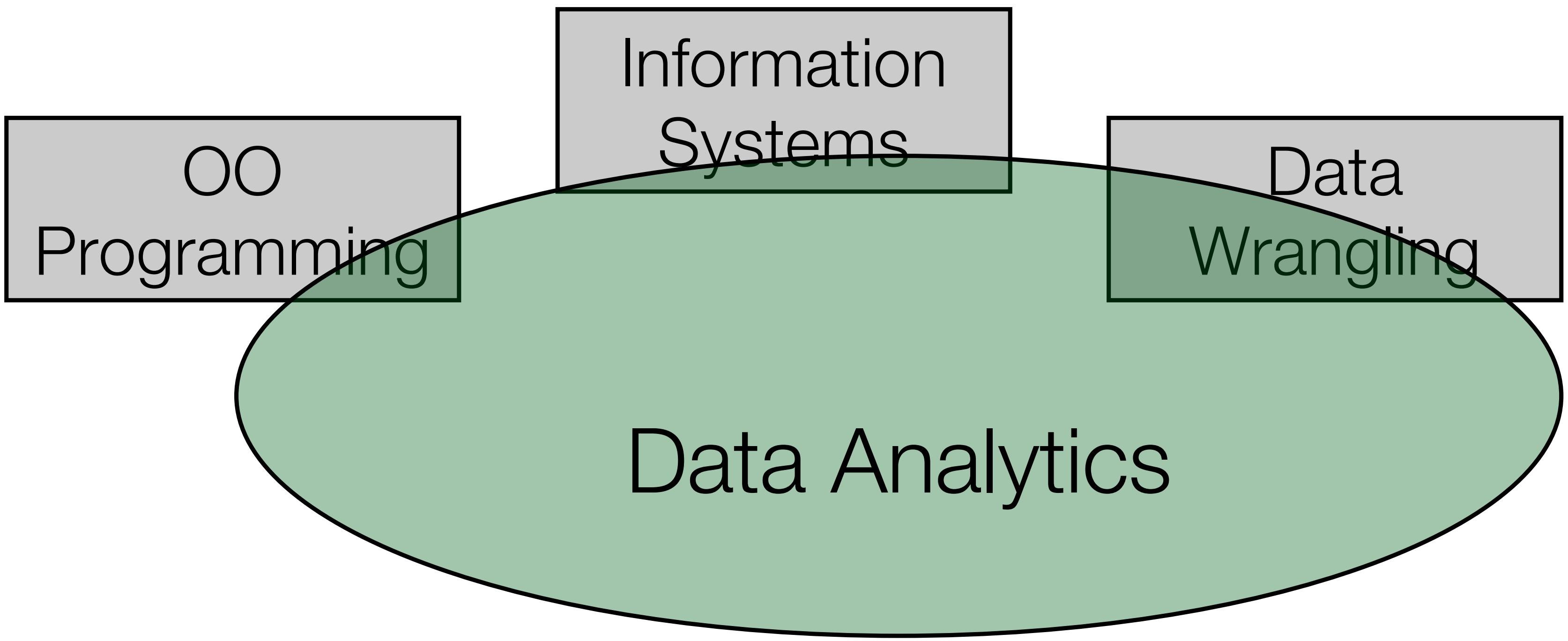
- Take a minute to think about where Analytics (in the wider sense of its meaning) is used
    - In your daily work, in your team
    - Within your Department, Section, Area, etc
    - Is there potential to introduce new/additional Analytics to Improve decision making

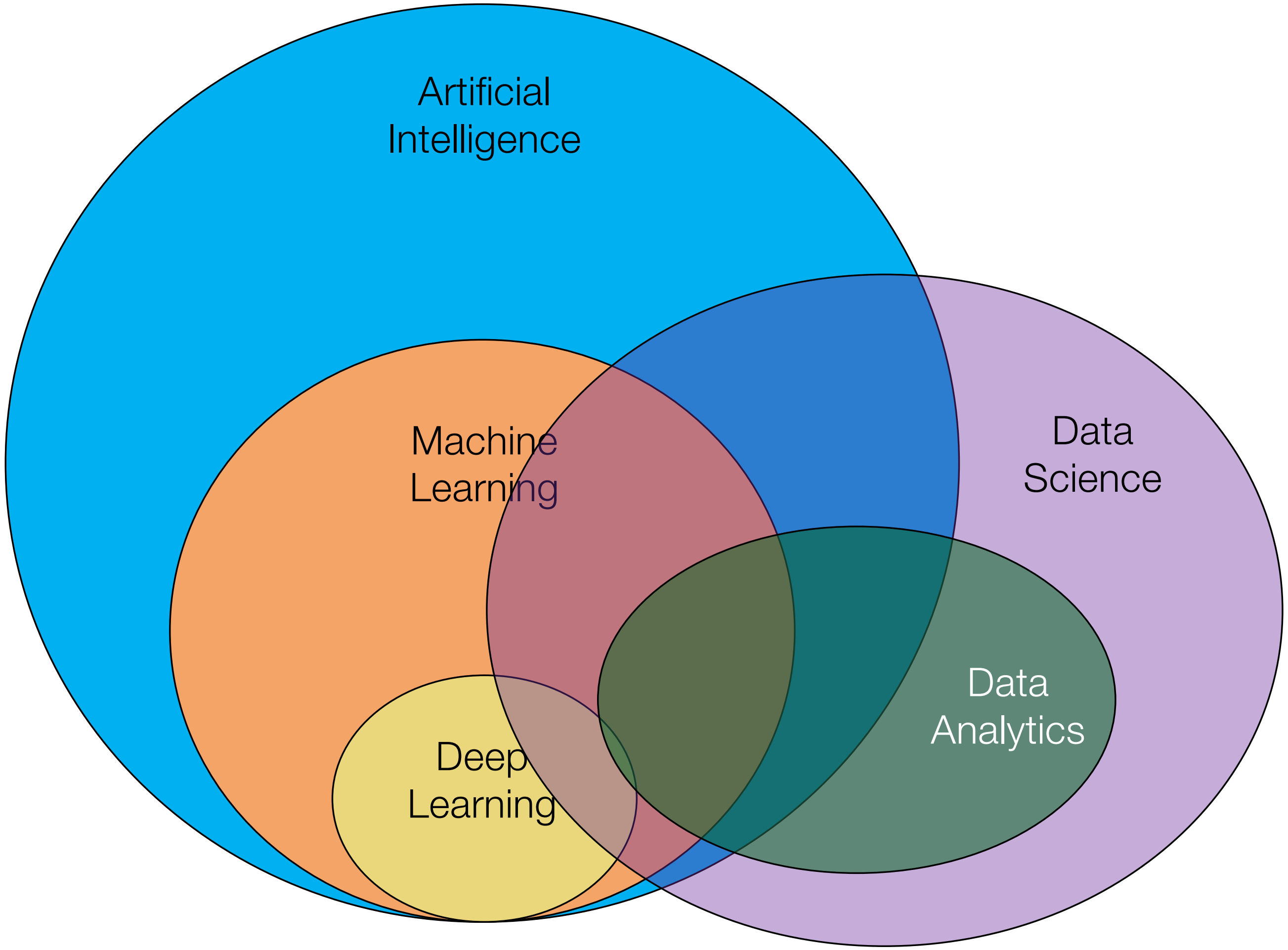
Examples of Application Areas

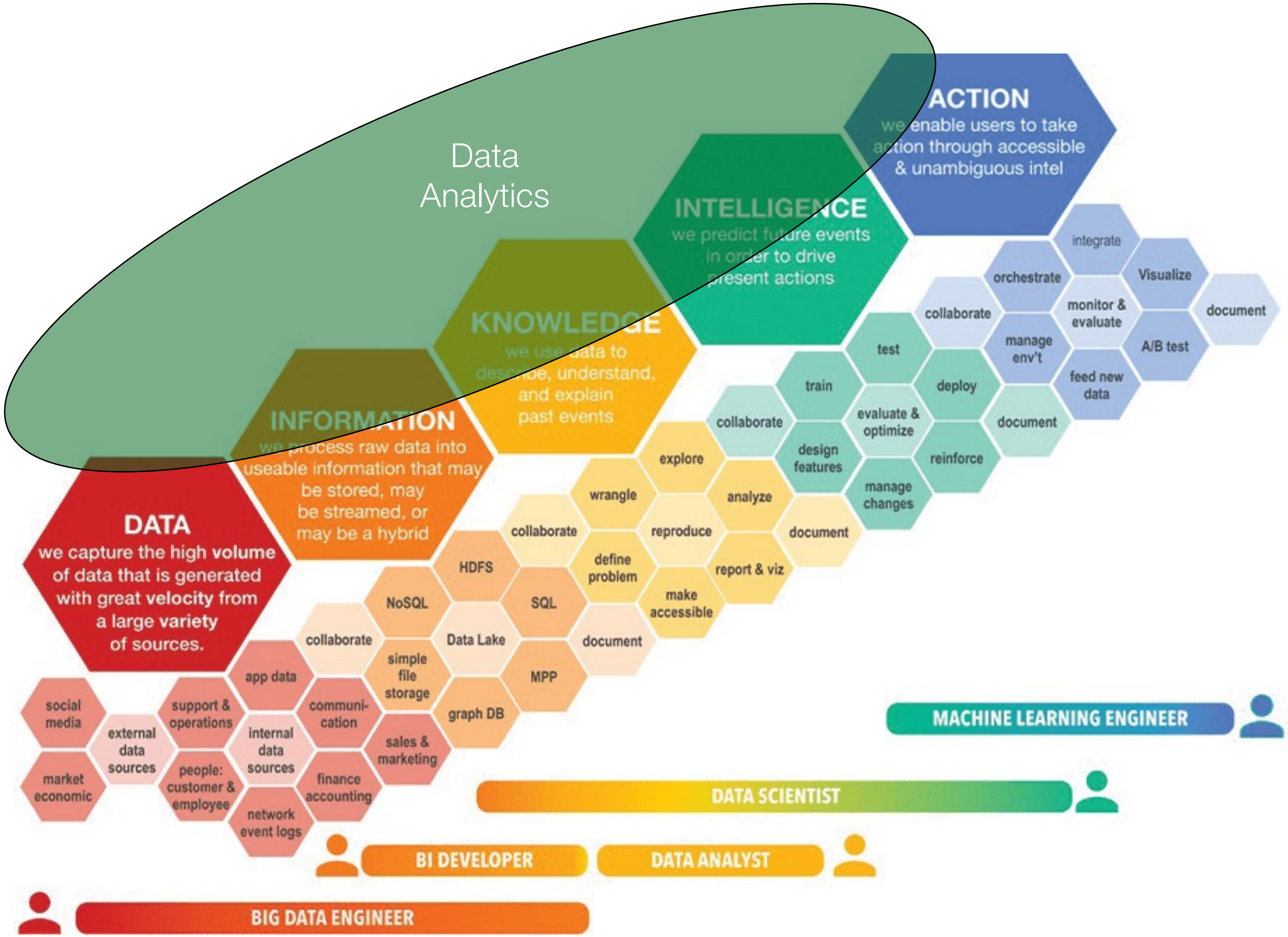


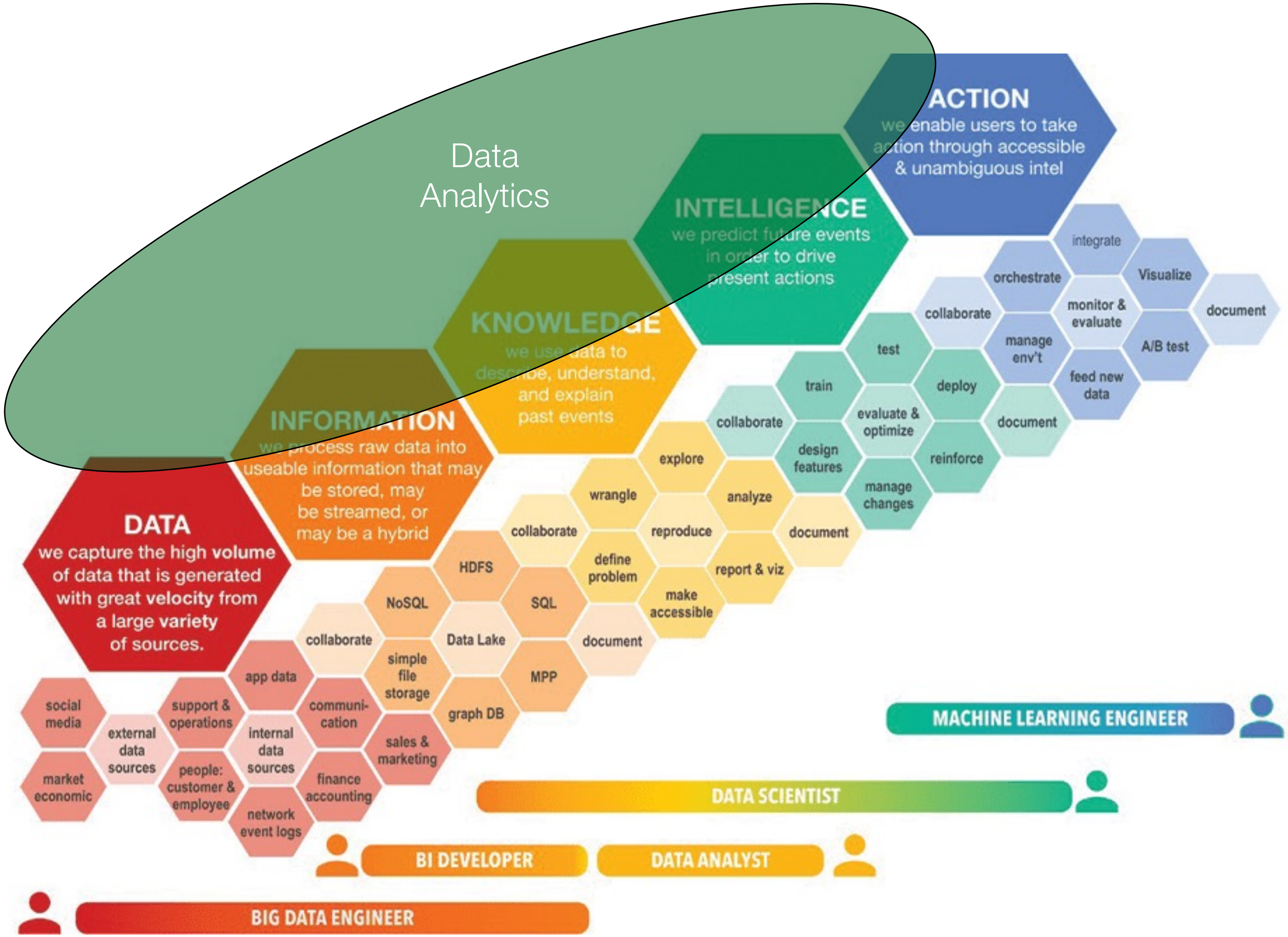


Where have you learned Analytics



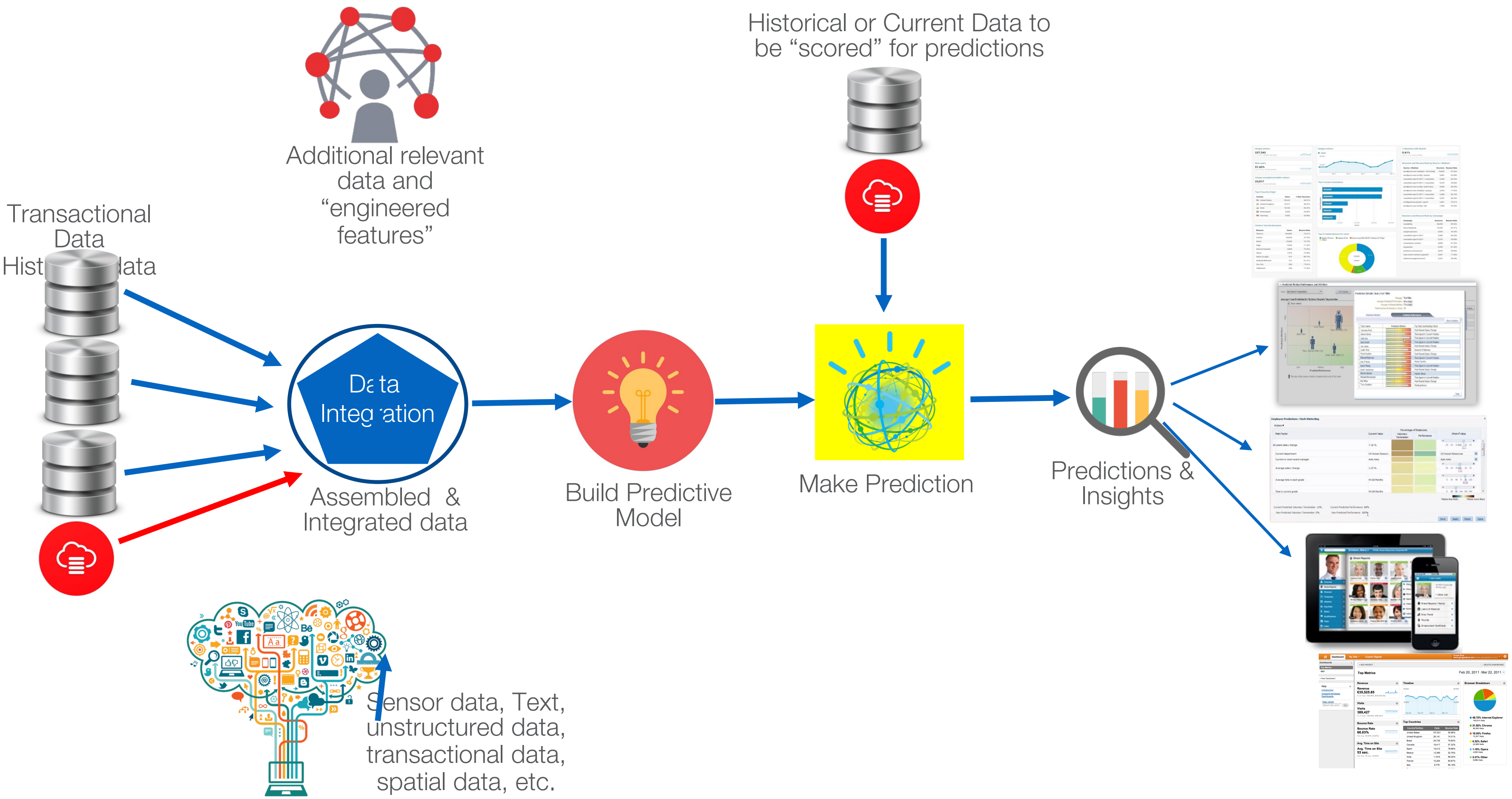

Artificial Intelligence
Machine Learning
Deep Learning
Data Science
Data Analytics


Data Analytics
ACTION
we enable users to take action through accessible & unambiguous intel
INTELLIGENCE
we predict future events in order to drive present actions
KNOWLEDGE
we use data to describe, understand, and explain past events
INFORMATION
we process raw data into useable information that may be stored, may be streamed, or may be a hybrid
DATA
we capture the high volume of data that is generated with great velocity from a large variety of sources.
social media
external data sources
market economic
support & operations
people: customer & employee
app data
internal data sources
network event logs
communi-cation
finance accounting
collaborate
sales & marketing
NoSQL
simple file storage
HDFS
Data Lake
graph DB
SQL
MPP
collaborate
define problem
document
wrangle
reproduce
make accessible
explore
analyze
report & viz
collaborate
design features
document
train
test
evaluate & optimize
manage changes
collaborate
deploy
reinforce
orchestrate
manage env't
document
integrate
monitor & evaluate
feed new data
Visualize
A/B test
document
MACHINE LEARNING ENGINEER
DATA SCIENTIST
BI DEVELOPER
DATA ANALYST
BIG DATA ENGINEER



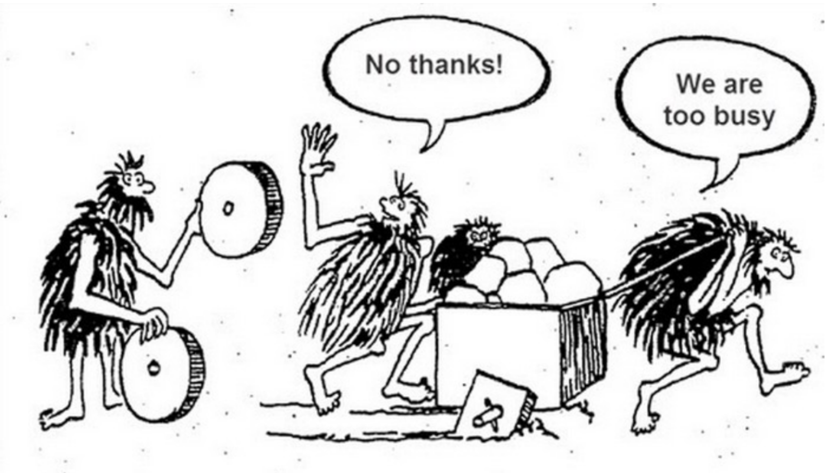
Data Analytics
ACTION
we enable users to take action through accessible & unambiguous intel
INTELLIGENCE
we predict future events in order to drive present actions
KNOWLEDGE
we use data to describe, understand, and explain past events
INFORMATION
we process raw data into useable information that may be stored, may be streamed, or may be a hybrid
DATA
we capture the high volume of data that is generated with great velocity from a large variety of sources.
social media
external data sources
market economic
support & operations
people: customer & employee
app data
internal data sources
network event logs
communi-cation
finance accounting
sales & marketing
collaborate
simple file storage
NoSQL
HDFS
Data Lake
graph DB
SQL
MPP
collaborate
wrangle
define problem
document
explore
reproduce
make accessible
analyze
report & viz
document
collaborate
train
design features
test
evaluate & optimize
manage changes
deploy
reinforce
collaborate
orchestrate
manage env't
document
integrate
monitor & evaluate
feed new data
Visualize
A/B test
document
MACHINE LEARNING ENGINEER
DATA SCIENTIST
BI DEVELOPER
DATA ANALYST
BIG DATA ENGINEER

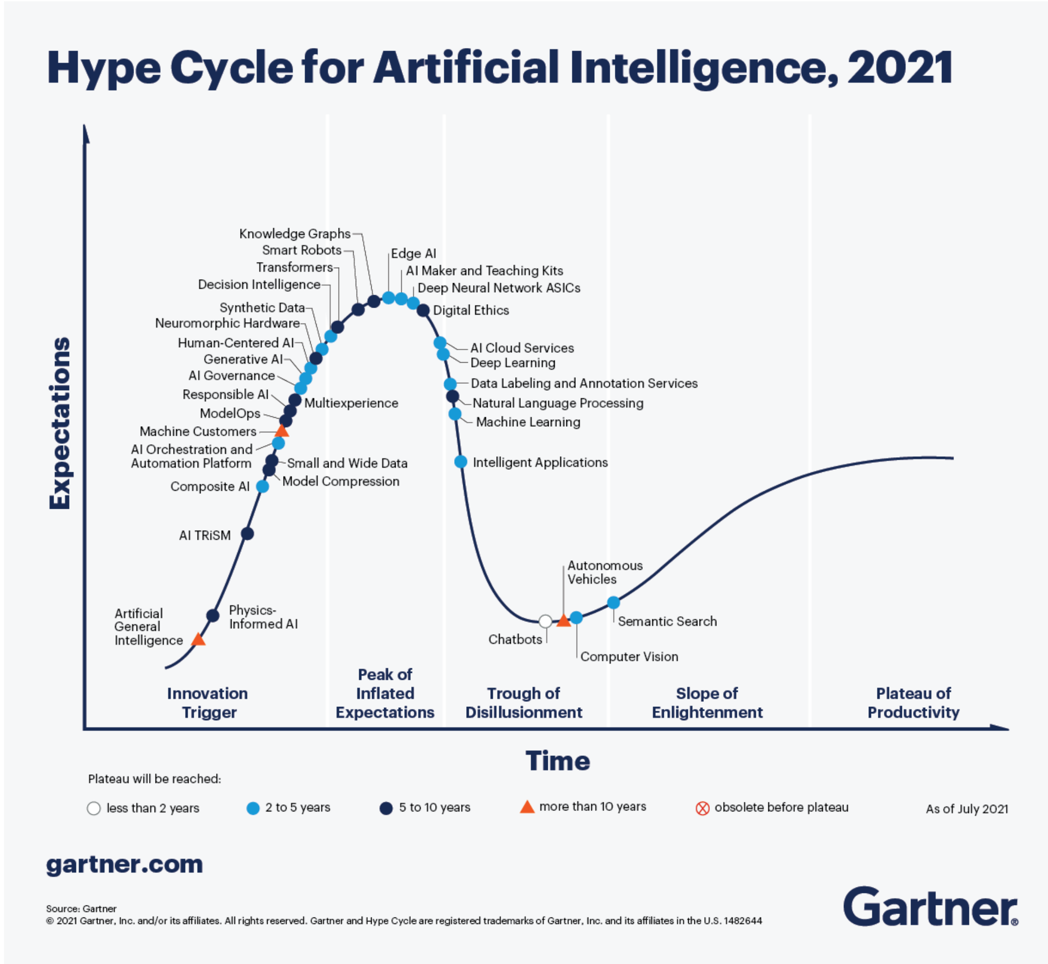
From Data to Deployment

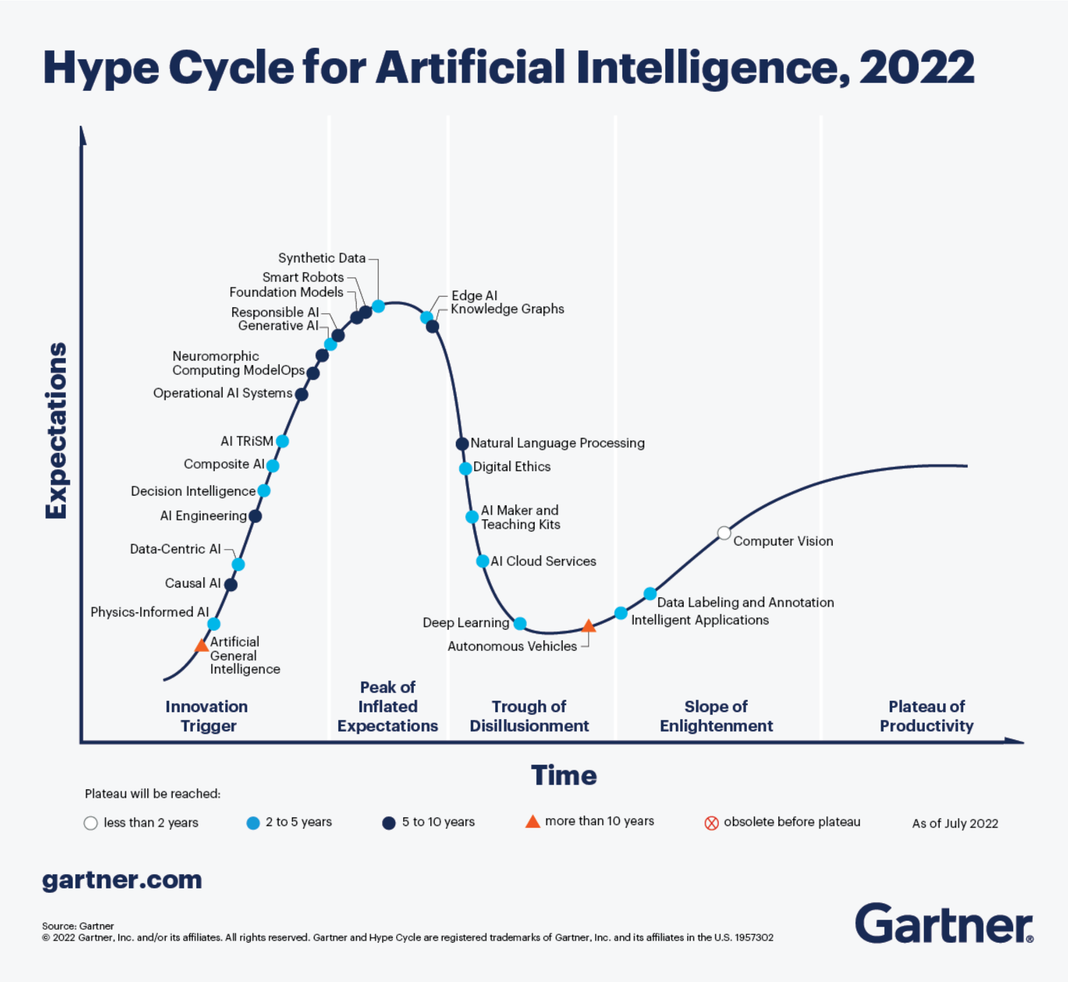


What Analytics should you use?

- You could follow the latest trends
- You could use what people are saying is the best tool/language/package/API, etc
- You could re-invent the wheel

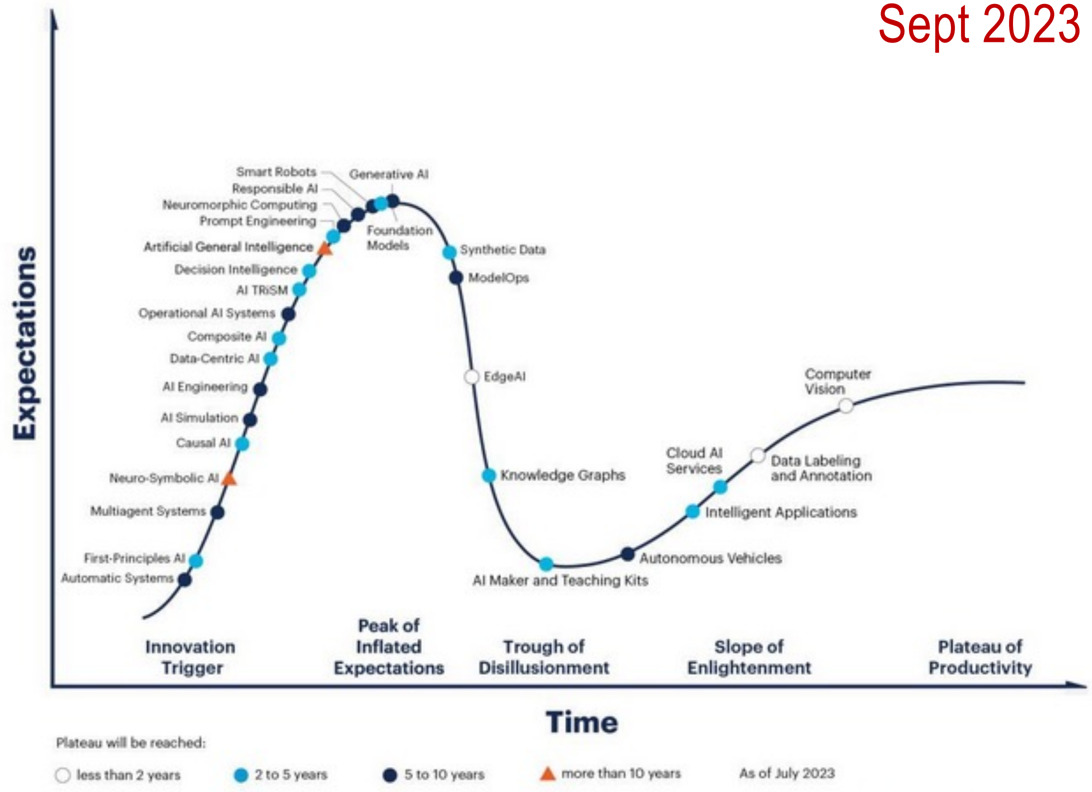






Hype Cycle for Artificial Intelligence, 2022
Expectations
Synthetic Data
Smart Robots
Foundation Models
Responsible AI
Generative AI
Neuromorphic Computing
ModelOps
Operational AI Systems
AI TRiSM
Composite AI
Decision Intelligence
AI Engineering
Data-Centric AI
Causal AI
Physics-Informed AI
Artificial General Intelligence
Edge AI
Knowledge Graphs
Natural Language Processing
Digital Ethics
AI Maker and Teaching Kits
AI Cloud Services
Deep Learning
Autonomous Vehicles
Data Labeling and Annotation
Intelligent Applications
Computer Vision
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
Plateau will be reached:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
As of July 2022
gartner.com
Source: Gartner
© 2022 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1957302
Gartner

Sept 2023



Expectations
Smart Robots
Responsible AI
Neuromorphic Computing
Prompt Engineering
Generative AI
Foundation Models
Artificial General Intelligence
Decision Intelligence
AI TRiSM
Operational AI Systems
Composite AI
Data-Centric AI
AI Engineering
AI Simulation
Causal AI
Neuro-Symbolic AI
Multiagent Systems
First-Principles AI
Automatic Systems
Synthetic Data
ModelOps
EdgeAI
Knowledge Graphs
AI Maker and Teaching Kits
Autonomous Vehicles
Intelligent Applications
Cloud AI Services
Data Labeling and Annotation
Computer Vision
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
Plateau will be reached:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
As of July 2023

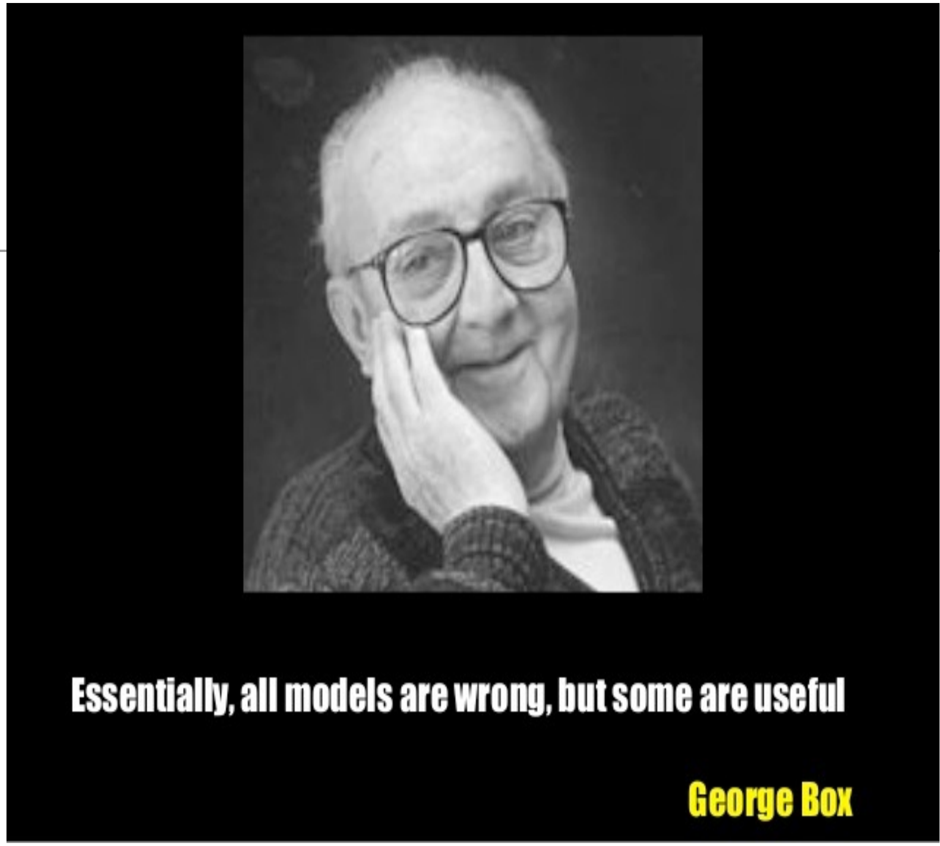
What Analytics should you use?

- Use KISS approach
    - Keep it Simple S….d!

- Build your knowledge and experience

- Prove what you are doing works
    - Learn to crawl before learning to walk, before learning to run, …

- A lot of Analytics algorithms and approaches can be 10s to 100s of years old
    - Naïve Bayes theorem 1763
    - Babylonian algorithms 2000-1700BC – multiplication algorithms
    - Logarithms – 1614
    - Nearest Neighbor – 1967
    - First Neural Network machine - 1951



A model is a simplification or approximation of reality and hence will not reflect all of reality.

His paper was published in the Journal of the American Statistical Association, 1976
Book *Empirical Model-Building and Response Surfaces*, 1987

What Algorithm Should you use?

- The **“No Free Lunch” theorem** states that there is no one model that works best for every problem.
- The assumptions of a great model for one problem may not hold for another problem, so it is common in machine learning to try multiple models and find one that works best for a particular problem.
- Depending on the problem, it is important to assess the trade-offs between **speed**, **accuracy**, and **complexity** of different models and algorithms and find a model that works best for that particular problem.

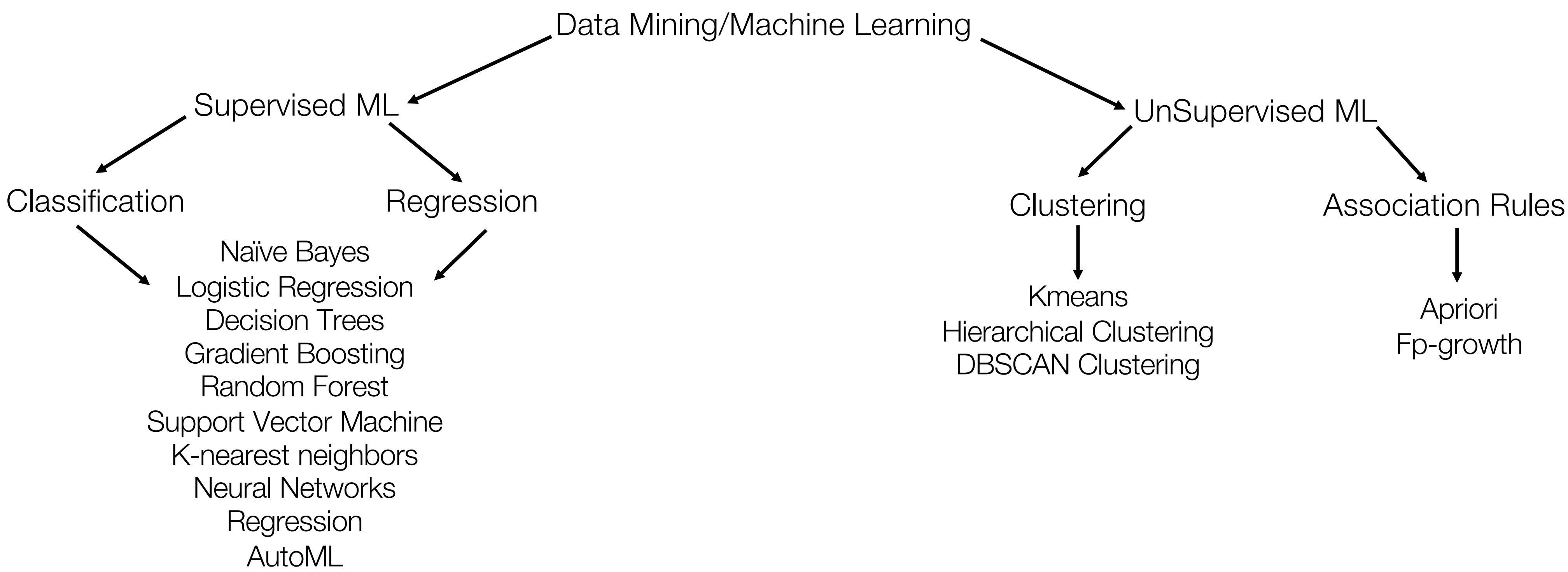
⇒Try lots of algorithms (and not just one)

⇒What’s trendy today? vs what really works

⇒ Prove it

Machine Learning

- the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data.



Let's look at an example

- Supervised Machine Learning techniques automatically learn a model of the relationship between a set of descriptive features and a target feature from a set of historical examples

ID	OCCUPATION	AGE	LOAN-SALARY RATIO	OUTCOME
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

- What is the relationship between the descriptive features (OCCUPATION, AGE, LOAN-SALARY RATIO) and the target feature (OUTCOME)?

ID	Occupation	Age	Loan-Salary Ratio	Outcome
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

```
if LOAN-SALARY RATIO > 3 then
    OUTCOME='default'
else
    OUTCOME='repaid'
end if
```

ID	Occupation	Age	Loan-Salary Ratio	Outcome
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

```
if LOAN-SALARY RATIO > 3 then
    OUTCOME='default'
else
    OUTCOME='repaid'
end if
```

ID	Occupation	Age	Loan-Salary Ratio	Outcome
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

```
if LOAN-SALARY RATIO > 3 then
    OUTCOME='default'
else
    OUTCOME='repaid'
end if
```

This is an example of a prediction model

This is also an example of a consistent prediction model

Notice that this model does not use all the features and the feature that it uses is a derived feature (in this case a ratio): feature design and feature selection are two important topics that we will return to again and again.

Where do you start?

Low hanging fruit   vs   The difficult ones to pick

Data Analytics

- Analysing data to discover patterns in the Data
- Can use these patterns to explain behaviours in the Data
    - Explain behaviours in our Customers
    - Explain behaviours in our Manufacturing
    - Explain behaviours in our Products
    - Explain behaviours in our Services
    - Etc

- We can use these patterns in different ways
    - Reports
    - Production
    - Decision making

Data Analytics

- But

- These are simple patterns !
  - We can See these patterns
  - We can see and discover these by exploring the data
  - By applying our Business (Domain) Knowledge to understand these
  - Apply a meaning to them, explain them within certain events in the business
    - Certain things happen at different times of year

- What about when our Data before Bigger? (Big Data)
  - Data doesn’t have to be Big to get value from it
  - We need additional tools/algorithms/infrastructure etc to help us explore the data & find patterns

Another Example

- The real value of Data Analytics machine learning becomes apparent in situations like this when we want to build prediction models from large datasets with multiple features.
- What is the relationship between the descriptive features/variables (Amount, Salary, Loan-Salary Ratio, Age, Occupation, House, Type) and the target feature/variable (Outcome)?
- This is a little bit more difficult!

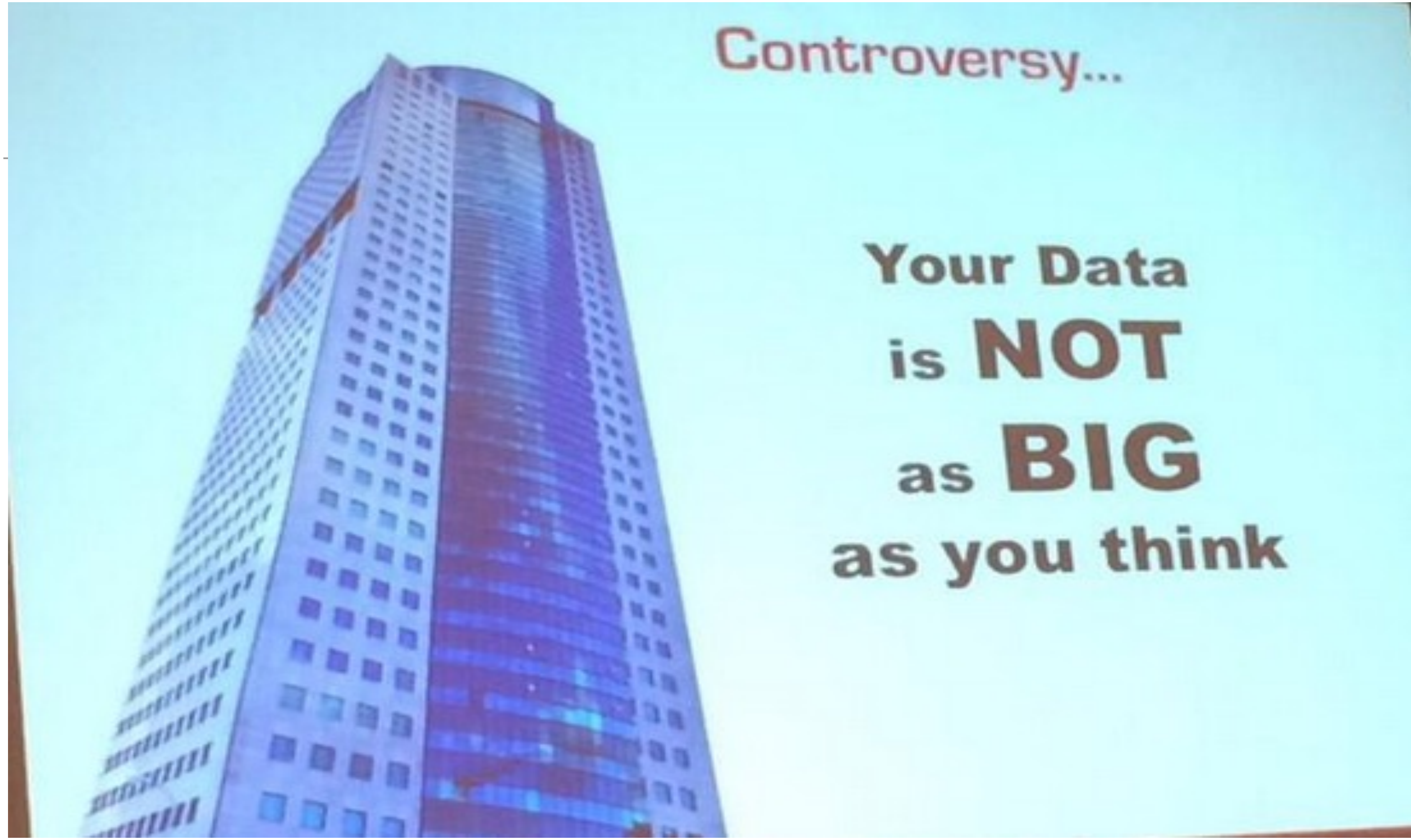
ID	Amount	Salary	Loan-Salary Ratio	Age	Occupation	House	Type	Outcome
1	245,100	66,400	3.69	44	industrial	farm	stb	repaid
2	90,600	75,300	1.2	41	industrial	farm	stb	repaid
3	195,600	52,100	3.75	37	industrial	farm	ftb	default
4	157,800	67,600	2.33	44	industrial	apartment	ftb	repaid
5	150,800	35,800	4.21	39	professional	apartment	stb	default
6	133,000	45,300	2.94	29	industrial	farm	ftb	default
7	193,100	73,200	2.64	38	professional	house	ftb	repaid
8	215,000	77,600	2.77	17	professional	farm	ftb	repaid
9	83,000	62,500	1.33	30	professional	house	ftb	repaid
10	186,100	49,200	3.78	30	industrial	house	ftb	default
11	161,500	53,300	3.03	28	professional	apartment	stb	repaid
12	157,400	63,900	2.46	30	professional	farm	stb	repaid
13	210,000	54,200	3.87	43	professional	apartment	ftb	repaid
14	209,700	53,000	3.96	39	industrial	farm	ftb	default
15	143,200	65,300	2.19	32	industrial	apartment	ftb	default
16	203,000	64,400	3.15	44	industrial	farm	ftb	repaid
17	247,800	63,800	3.88	46	industrial	house	stb	repaid
18	162,700	77,400	2.1	37	professional	house	ftb	repaid
19	213,300	61,100	3.49	21	industrial	apartment	ftb	default
20	284,100	32,300	8.8	51	industrial	farm	ftb	default
21	154,000	48,900	3.15	49	professional	house	stb	repaid
22	112,800	79,700	1.42	41	professional	house	ftb	repaid
23	252,000	59,700	4.22	27	professional	house	stb	default
24	175,200	39,900	4.39	37	professional	apartment	stb	default
25	149,700	58,600	2.55	35	industrial	farm	stb	default

Another Example

```
if Loan-Salary Ratio < 1.5 then
    outcome = 'repaid'
else if Loan-Salary Ratio > 4 then
    outcome = 'default'
else if Age < 40 and Occupation = 'industrial' then
    outcome = 'default'
else
    outcome = 'repaid'
end if
```

Challenging!

ID	Amount	Salary	Loan-Salary Ratio	Age	Occupation	House	Type	Outcome
1	245,100	66,400	3.69	44	industrial	farm	stb	repaid
2	90,600	75,300	1.2	41	industrial	farm	stb	repaid
3	195,600	52,100	3.75	37	industrial	farm	ftb	default
4	157,800	67,600	2.33	44	industrial	apartment	ftb	repaid
5	150,800	35,800	4.21	39	professional	apartment	stb	default
6	133,000	45,300	2.94	29	industrial	farm	ftb	default
7	193,100	73,200	2.64	38	professional	house	ftb	repaid
8	215,000	77,600	2.77	17	professional	farm	ftb	repaid
9	83,000	62,500	1.33	30	professional	house	ftb	repaid
10	186,100	49,200	3.78	30	industrial	house	ftb	default
11	161,500	53,300	3.03	28	professional	apartment	stb	repaid
12	157,400	63,900	2.46	30	professional	farm	stb	repaid
13	210,000	54,200	3.87	43	professional	apartment	ftb	repaid
14	209,700	53,000	3.96	39	industrial	farm	ftb	default
15	143,200	65,300	2.19	32	industrial	apartment	ftb	default
16	203,000	64,400	3.15	44	industrial	farm	ftb	repaid
17	247,800	63,800	3.88	46	industrial	house	stb	repaid
18	162,700	77,400	2.1	37	professional	house	ftb	repaid
19	213,300	61,100	3.49	21	industrial	apartment	ftb	default
20	284,100	32,300	8.8	51	industrial	farm	ftb	default
21	154,000	48,900	3.15	49	professional	house	stb	repaid
22	112,800	79,700	1.42	41	professional	house	ftb	repaid
23	252,000	59,700	4.22	27	professional	house	stb	default
24	175,200	39,900	4.39	37	professional	apartment	stb	default
25	149,700	58,600	2.55	35	industrial	farm	stb	default


Controversy...
Your Data
is NOT
as BIG
as you think

How to approach a Data Science Project

Find me something interesting in my data is a question from hell.

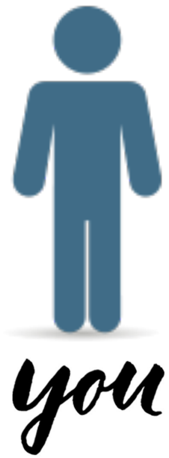
Analytics should be guided by business goals

Before you can measure something you really need to lay down a very concrete definition of what you're measuring

Focus hard on Business Question (and the relevant variables) that captures the essence of the question.


data

+


you

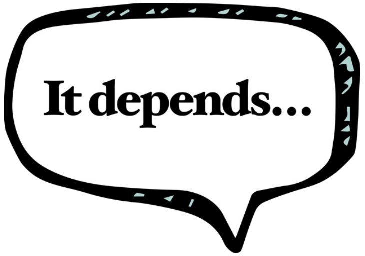
=


iNSiGHT

What Language or Tool

- There are lots and lots of Languages and Tools you can use
- They all do the same thing !!!
- No-Code Tools

Which one is the best?



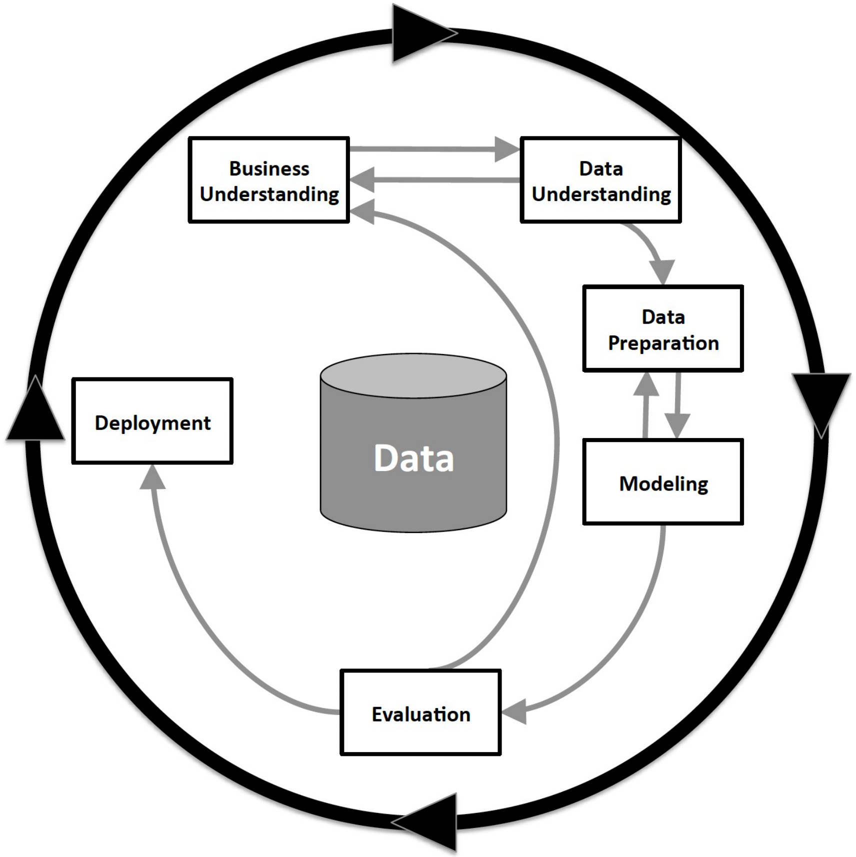
- Excel
- Tableau
- QlikViiew
- Power BI
- SAS
- SAP Business Object
- Google Data Studio
- IBM Cognos
- Looker

- Python
- R
- SQL
- Spark
- Scala
- Java
- Julia
- C / C++
- Go
- SAS

Next Week

- We will look at the Data Analytics & Data Science Lifecycle

- We'll start analysing data



Any Questions ?

What Now/Next ?