Data Science

Carolina A. de Lima Salge
Assistant Professor
Terry College of Business
University of Georgia

Business Intelligence
Spring 2021

October 2012 Issue

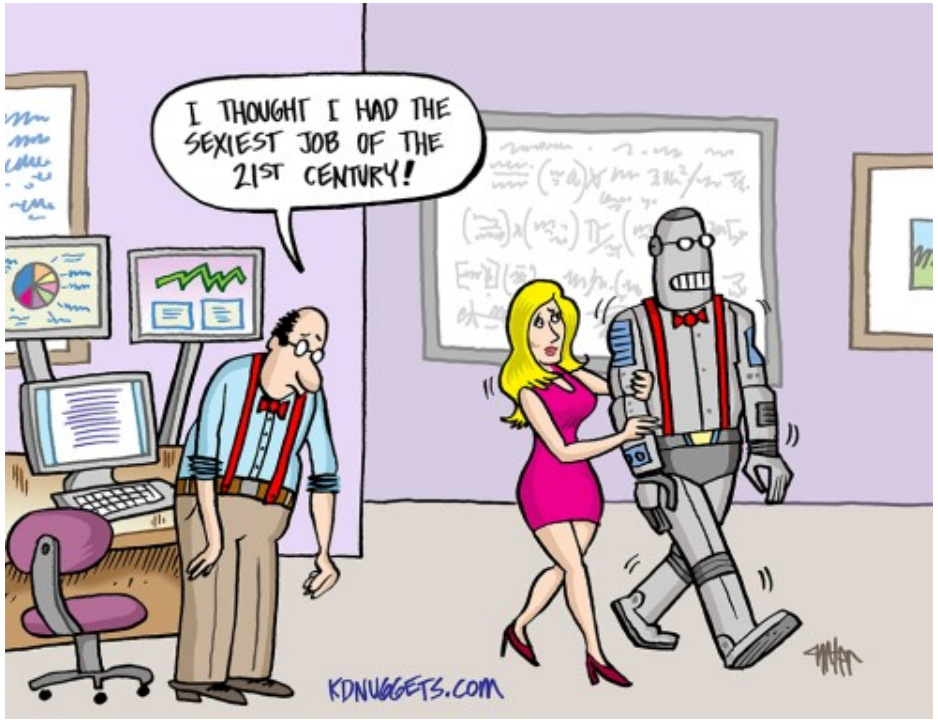
DATA

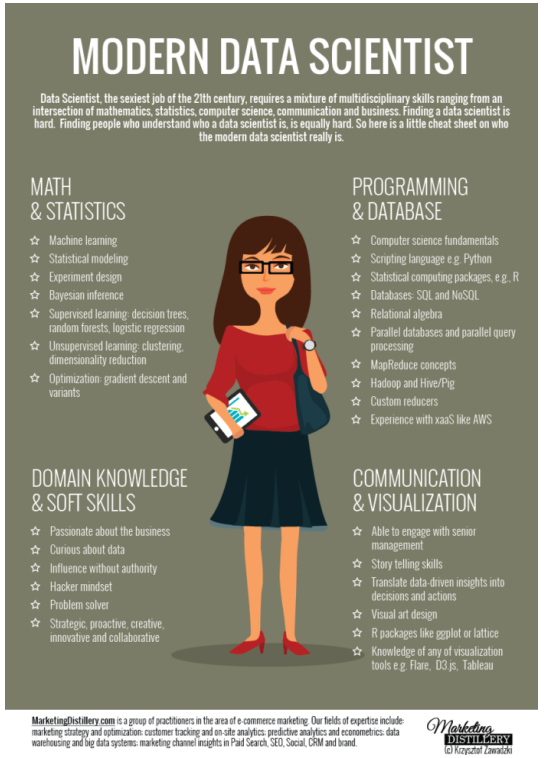
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

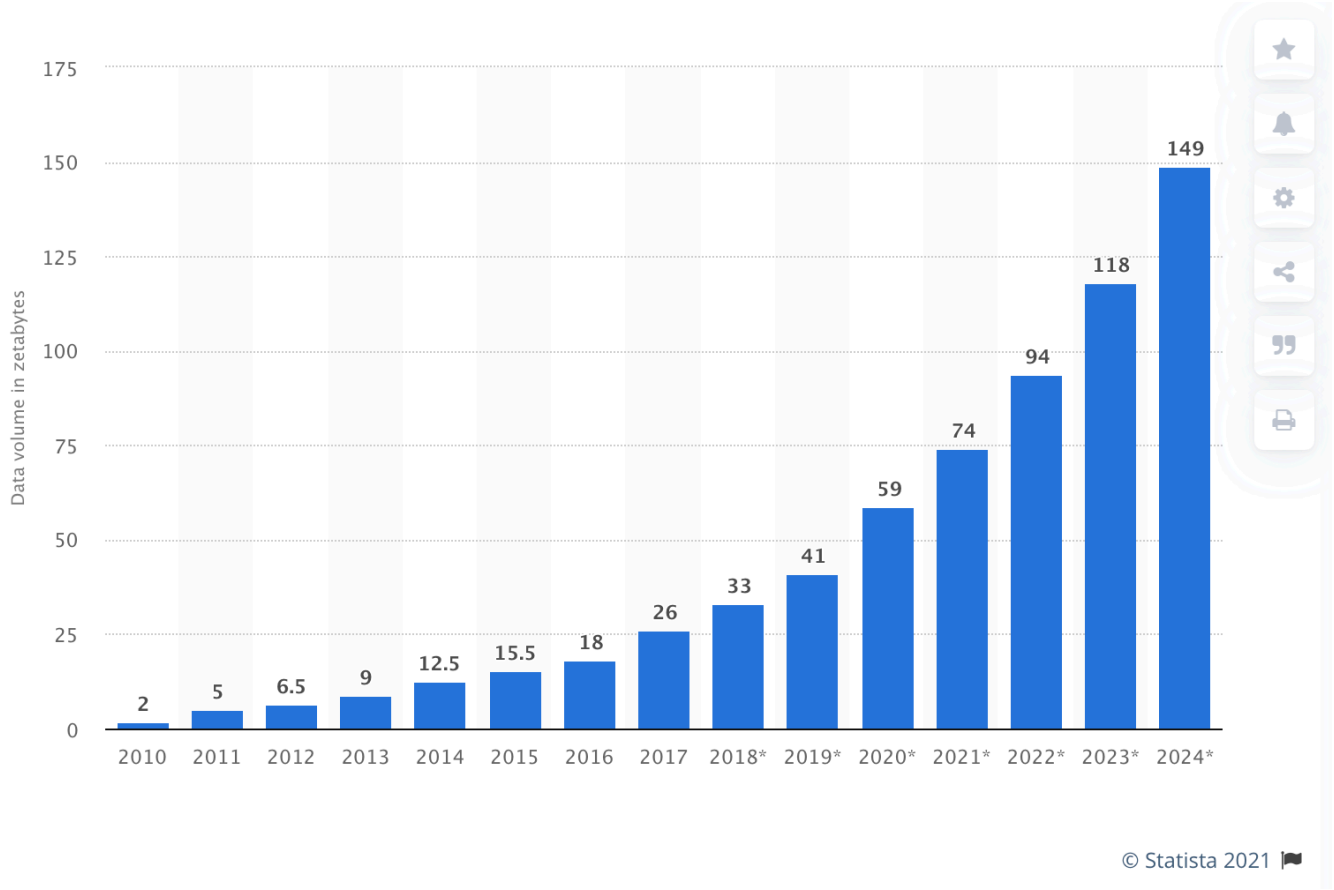
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."





Terry College of Business
UNIVERSITY OF GEORGIA

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024 *(in zettabytes)*



A **zettabyte** is a measure of storage capacity and is 2 to the 70th power bytes, also expressed as 10^{21} (1,000,000,000,000,000,000,000 bytes) or 1 sextillion bytes. One **Zettabyte** is approximately equal to a thousand Exabytes, a billion Terabytes, or a trillion Gigabytes

https://www.statista.com/statistics/871513/worldwide-data-created/

https://blogs.cisco.com/sp/the-zettabyte-era-officially-begins-how-much-is-that

Harvard Business Review

Diversity Latest Magazine Popular Topics Podcasts Video Store Bo

Data

Companies Are Failing in Their Efforts to Become Data-Driven

by Randy Bean and Thomas H. Davenport

February 05, 2019

Big Data Is Getting *Bigger*

Big data and business analytics revenue worldwide 2015-2022

Published by Statista Research Department, Jan 11, 2021

The global big data and business analytics market was valued at 168.8 billion U.S. dollars in 2018 and is forecast to grow to 274.3 billion U.S. dollars by 2022, with a five-year compound annual growth rate (CAGR) of 13.2 percent.

Big data

High volume, high velocity and high variety: one or more of these characteristics is used to define big data, the kind of data sets that are too large or too complex for traditional data processing applications. Fast-growing mobile data traffic, cloud computing traffic, as well as the rapid development of technologies such as artificial intelligence (AI) and the Internet of Things (IoT) all contribute to the increasing volume and complexity of data sets. By 2021, the global cloud data center IP traffic is forecast to reach around 19.5 zettabytes (ZBs). Connected IoT devices are projected to generate 79.4 ZBs of data in 2015.

Business analytics

Advanced analytics tools, such as predictive analytics and data mining, help to extract value from the data and generate business insights. The size of the business intelligence and analytics software application market is forecast to reach around 14.5 billion U.S. dollars in 2022. Banks make the most use of big data and business analytics technology – the banking industry contributed to 13.9 percent of the market revenue in 2018.

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

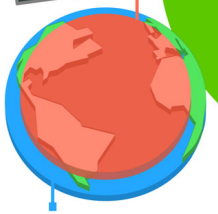
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

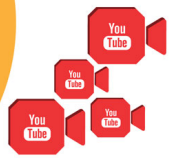
Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

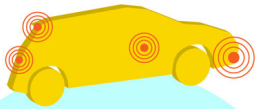
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session



Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

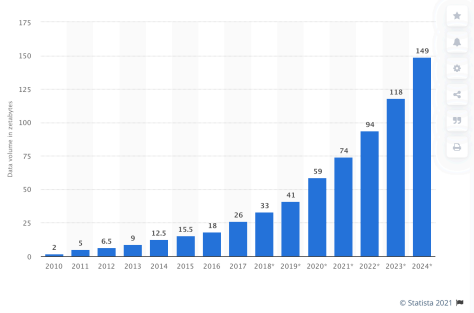
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

"… ***a wealth of information creates a poverty of attention*** …" — Herbert Simon (1971)

'Designing Organizations for an Information-Rich World' in Martin Greenberger (ed.) *Computers, Communications, and the Public Interest* (1971)



Big data and business analytics revenue worldwide 2015-2022

Published by Statista Research Department, Jan 11, 2021

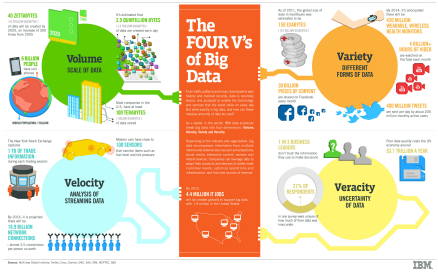
The global big data and business analytics market was valued at 168.8 billion U.S. dollars in 2018 and is forecast to grow to 274.3 billion U.S. dollars by 2022, with a five-year compound annual growth rate (CAGR) of 13.2 percent.

Big data

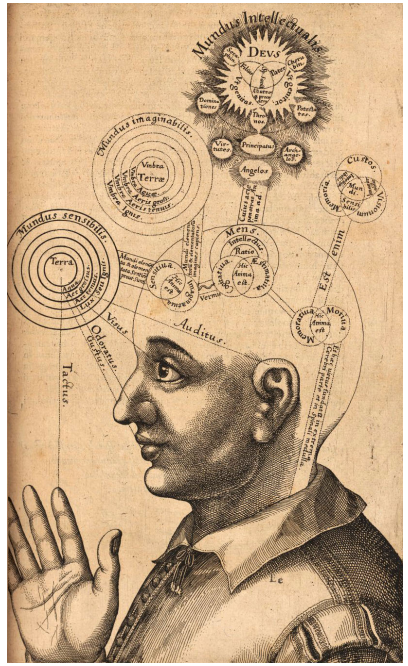
High volume, high velocity and high variety: one or more of these characteristics is used to define big data, the kind of data sets that are too large or too complex for traditional data processing applications. Fast-growing mobile data traffic, cloud computing traffic, as well as the rapid development of technologies such as artificial intelligence (AI) and the Internet of Things (IoT) all contribute to the increasing volume and complexity of data sets. By 2021, the global cloud data center IP traffic is forecast to reach around 19.5 zettabytes (ZBs). Connected IoT devices are projected to generate 79.4 ZBs of data in 2015.

Business analytics

Advanced analytics tools, such as predictive analytics and data mining, help to extract value from the data and generate business insights. The size of the business intelligence and analytics software application market is forecast to reach around 14.5 billion U.S. dollars in 2022. Banks make the most use of big data and business analytics technology – the banking industry contributed to 13.9 percent of the market revenue in 2018.



+



=

Data Science

- A set of **methods for extracting useful information** and knowledge from data (Provost & Fawcett 2013)
- A set of **methodologies for drawing meaningful conclusions** from thousands of forms of data that are available to us today (DataCamp, Data Science for Business Course)
- … *many more definitions out there!*

Wait, but this is a **Business Intelligence** class!

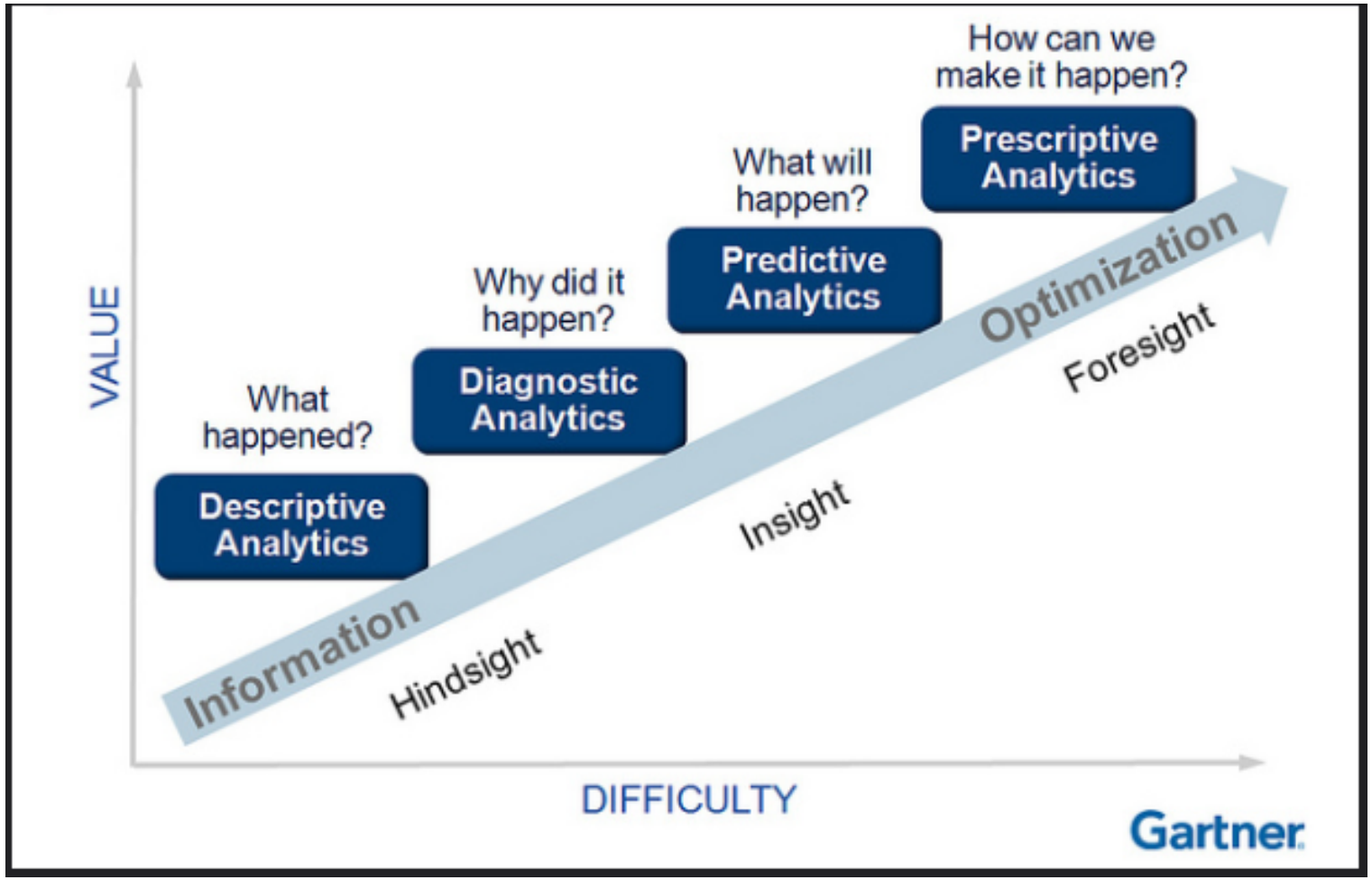
Defining the Terms: Data Science vs. Business Intelligence (BI)

It is important to begin with some basic definitions of the two terms, taking a deeper look at the two distinct (though closely allied) fields within Data Analytics. Data Science, as used in business, is intrinsically data-driven, where many interdisciplinary sciences are applied together to extract meaning and insights from available business data, which is typically large and complex. On the other hand, Business Intelligence or BI helps monitor the current state of business data to understand the historical performance of a business.

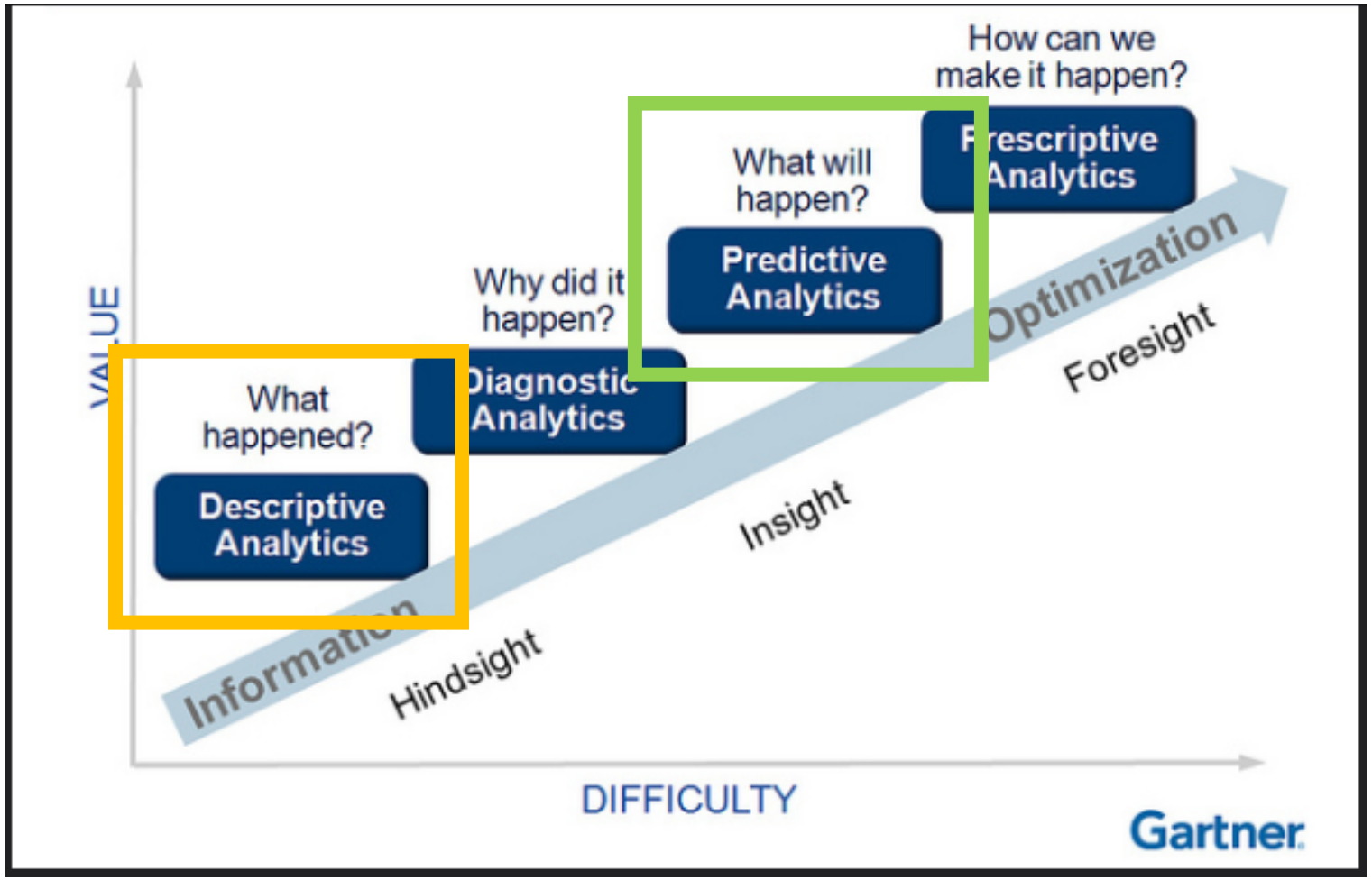
So, in nutshell, while BI helps interpret past data, Data Science can analyze the past data (trends or patterns) to make future predictions. BI is mainly used for reporting or Descriptive Analytics ; whereas Data Science is more used for Predictive Analytics or Prescriptive Analytics .

https://www.dataversity.net/data-science-vs-business-intelligence/#

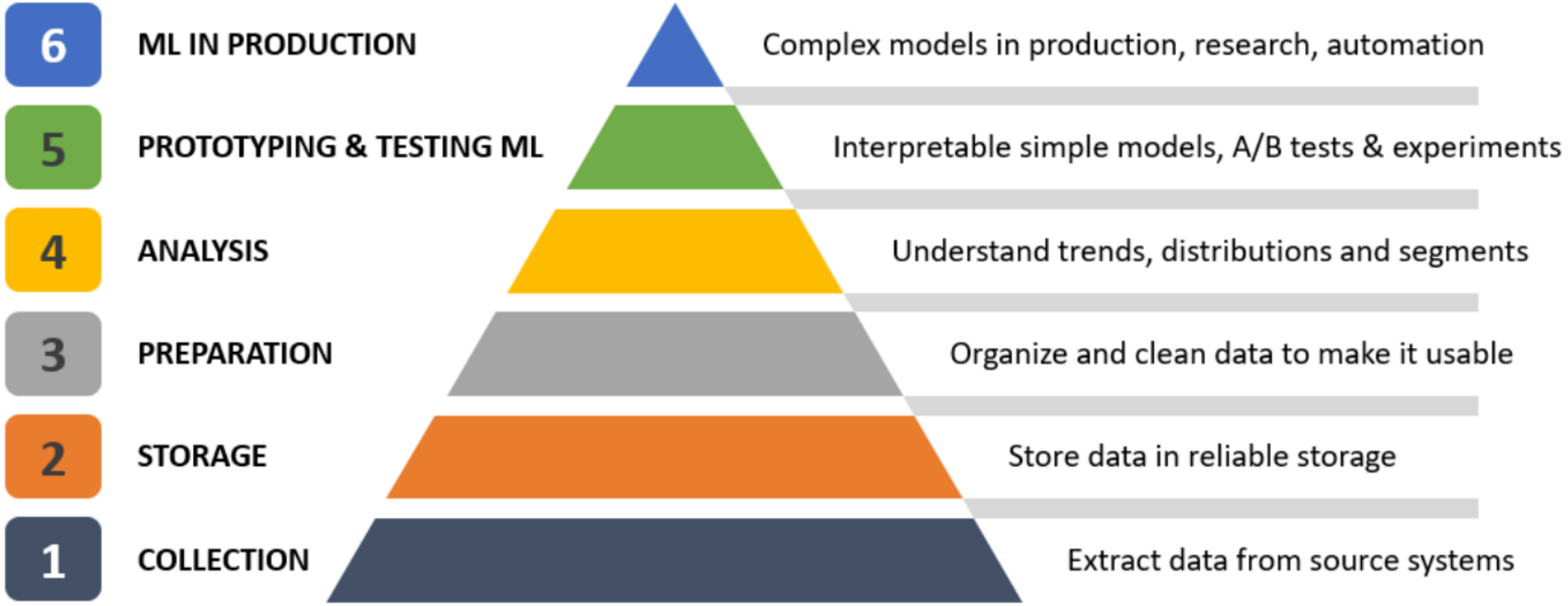
Creating Value With Data



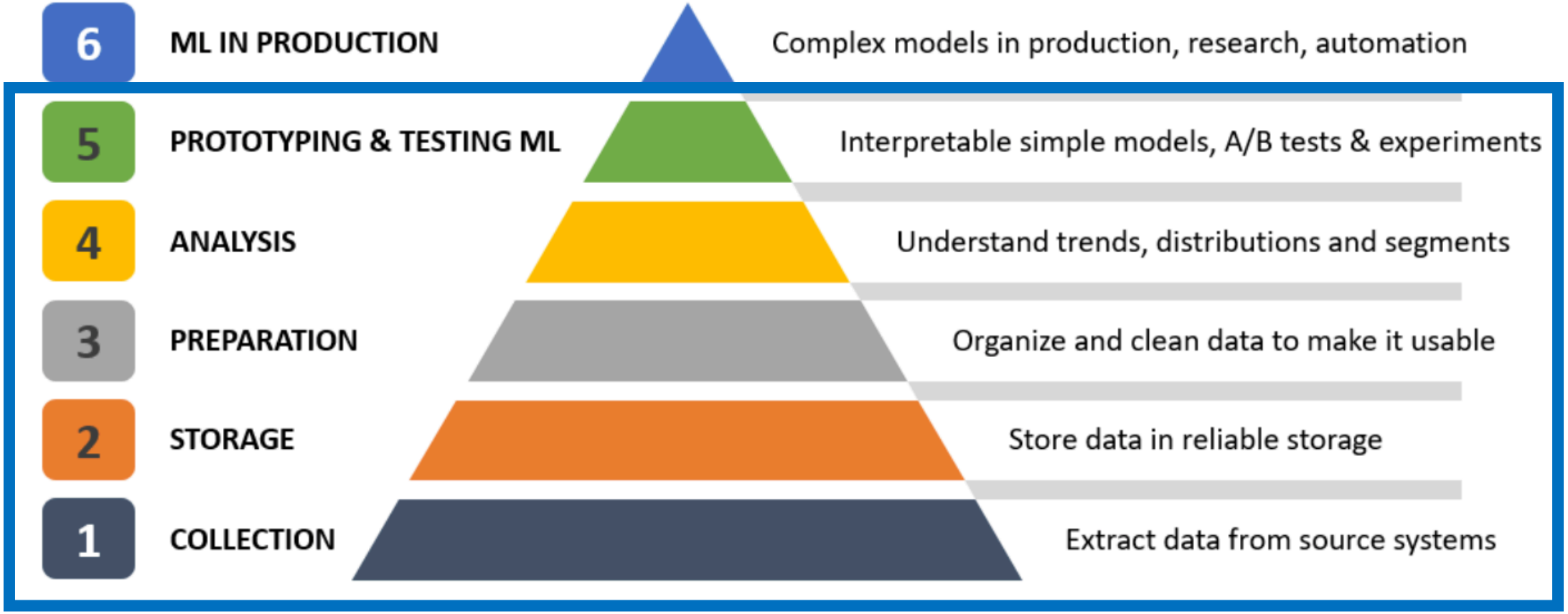
Creating Value With Data



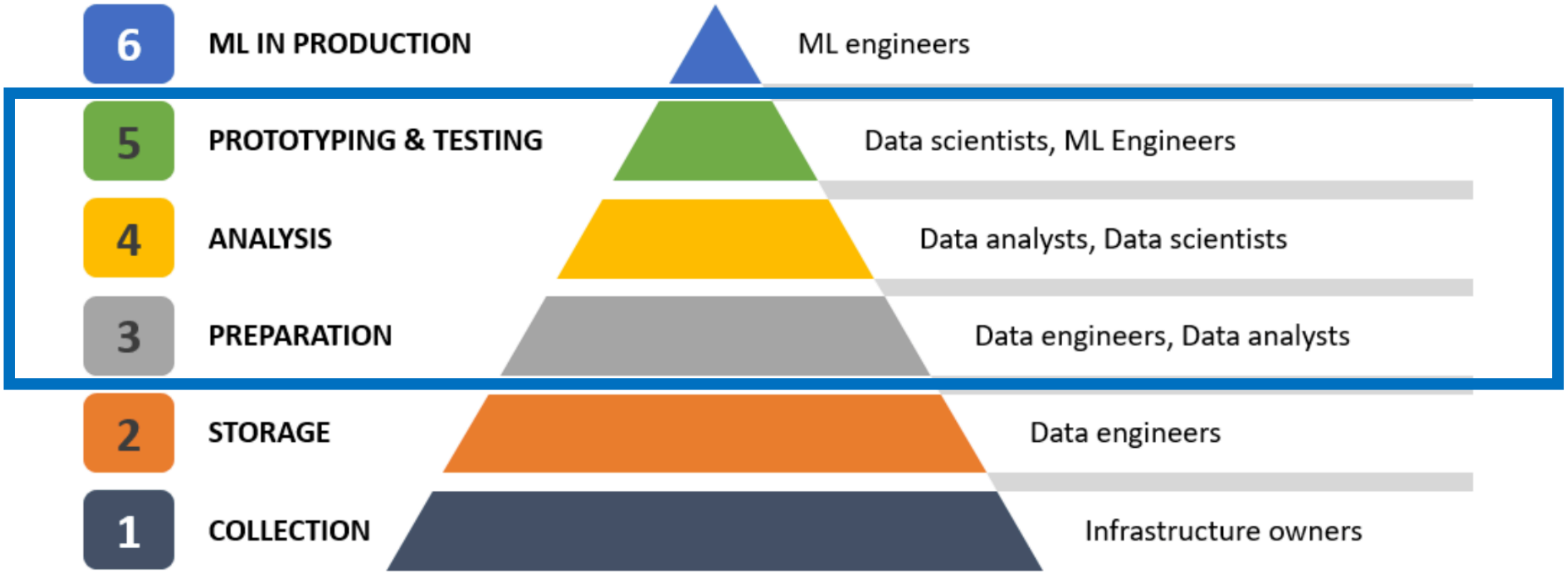
Creating Value With Data

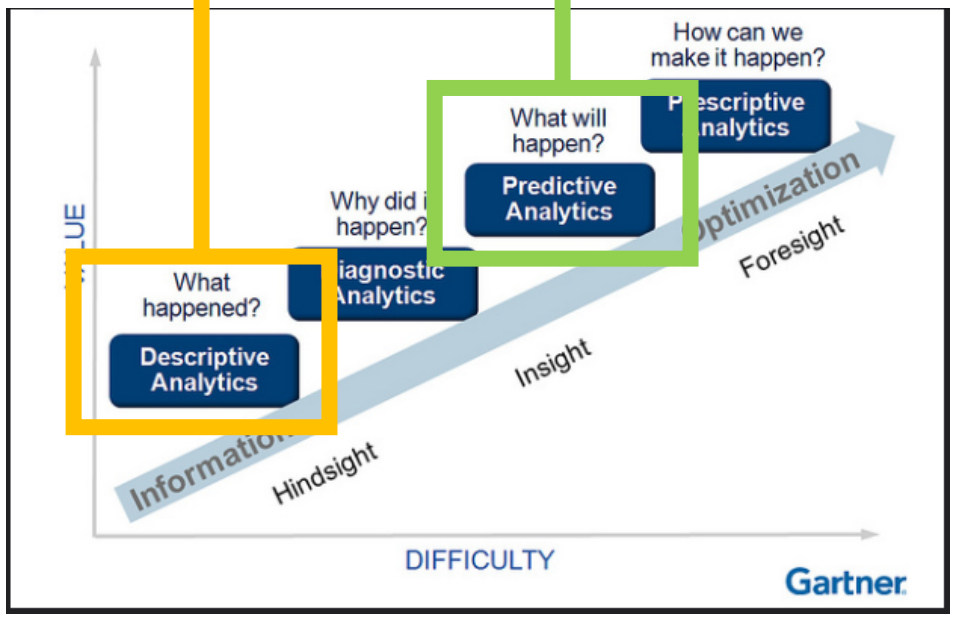


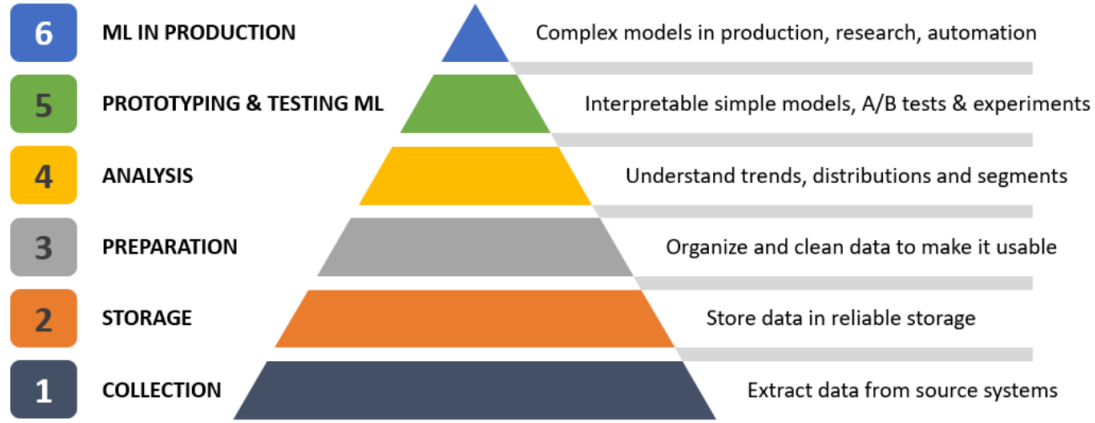
Creating Value With Data

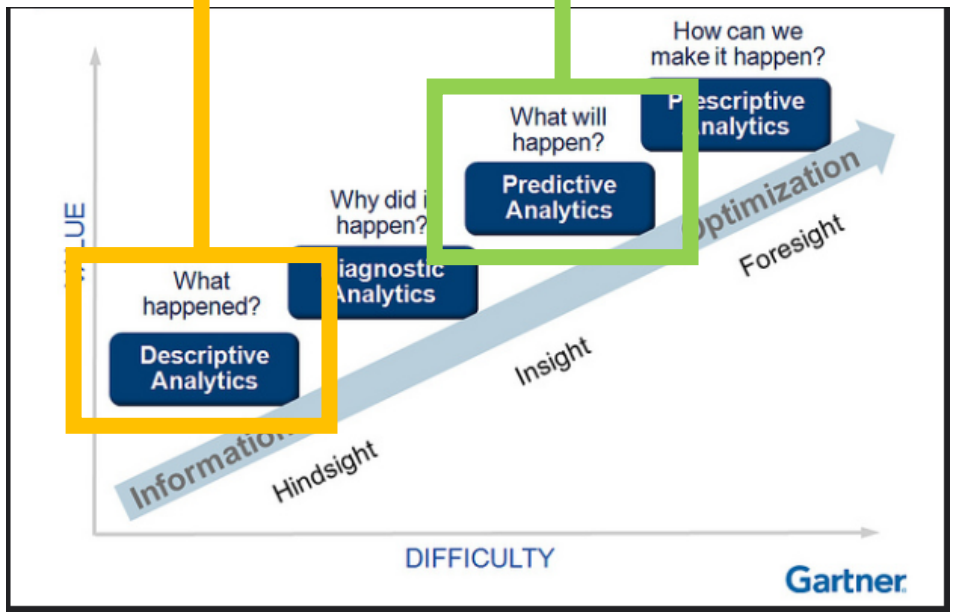


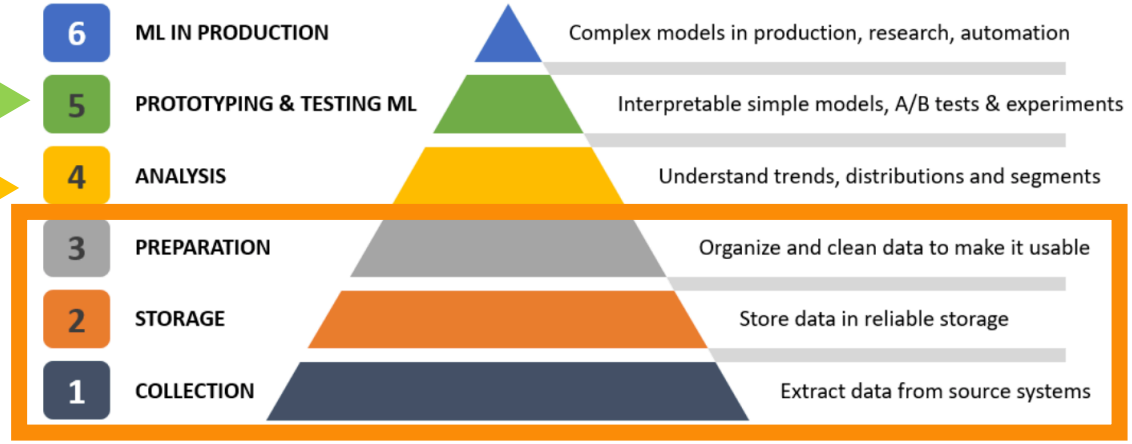
Roles in Value Creation

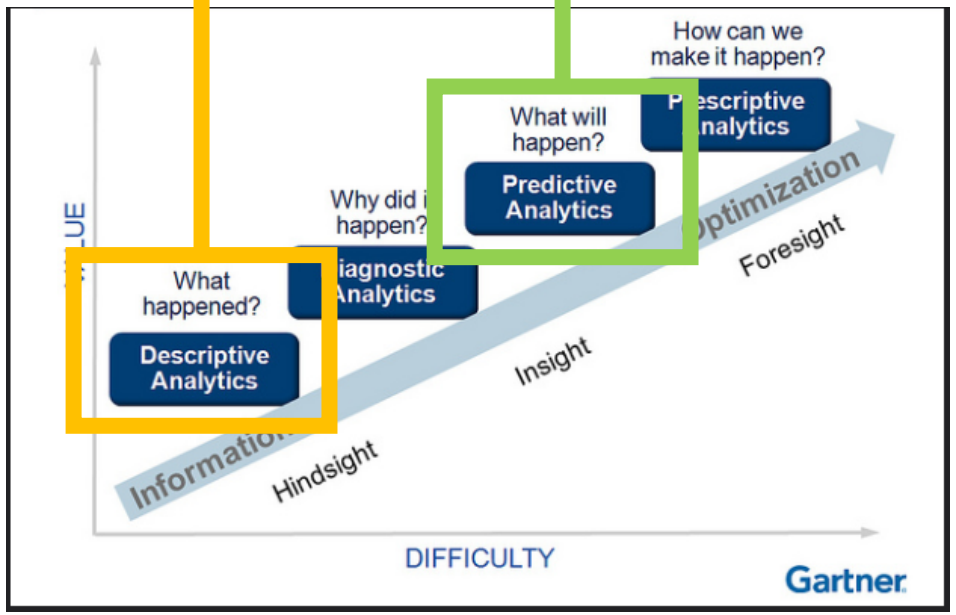


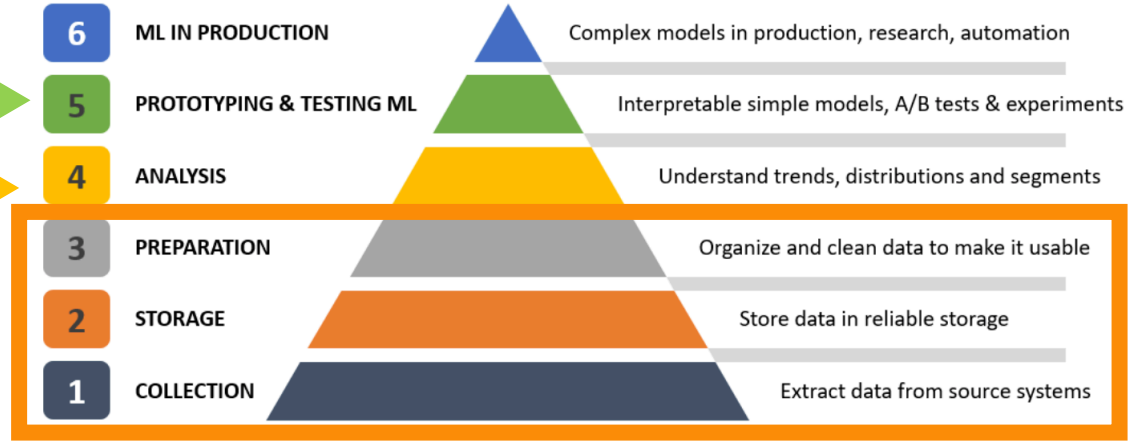

How can we make it happen?
What will happen?
Predictive Analytics
Why did it happen?
What happened?
Descriptive Analytics
Insight
Foresight
Hindsight
DIFFICULTY
Gartner


6
ML IN PRODUCTION
Complex models in production, research, automation
5
PROTOTYPING & TESTING ML
Interpretable simple models, A/B tests & experiments
4
ANALYSIS
Understand trends, distributions and segments
3
PREPARATION
Organize and clean data to make it usable
2
STORAGE
Store data in reliable storage
1
COLLECTION
Extract data from source systems

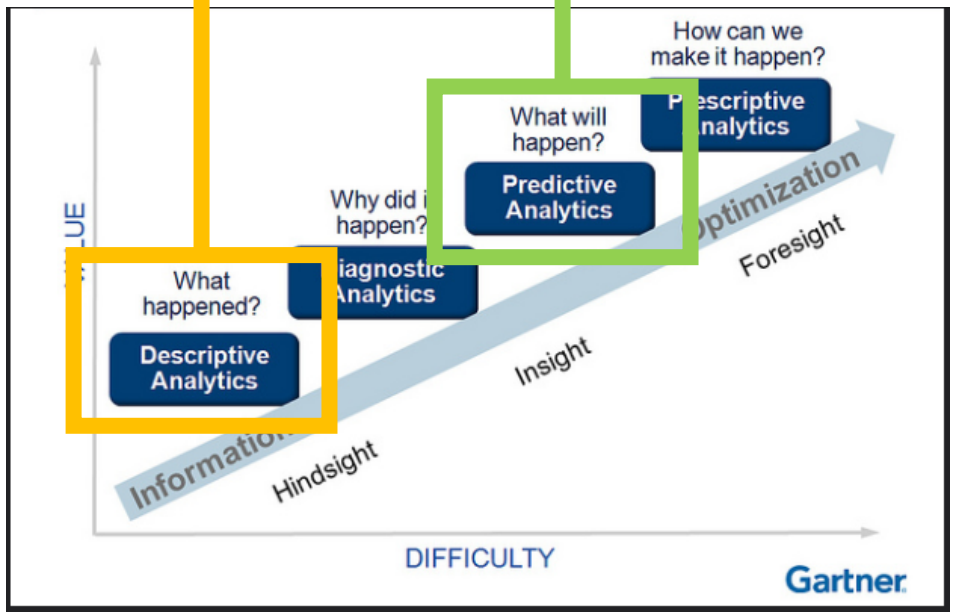

How can we make it happen?
What will happen?
Predictive Analytics
Why did it happen?
What happened?
Descriptive Analytics
Optimization
Foresight
Insight
Hindsight
DIFFICULTY
Gartner

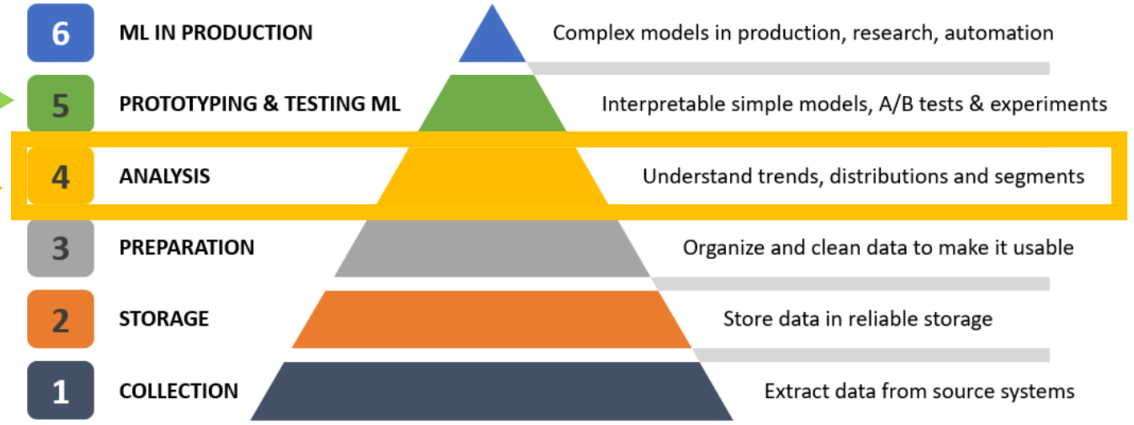

6
ML IN PRODUCTION
Complex models in production, research, automation
5
PROTOTYPING & TESTING ML
Interpretable simple models, A/B tests & experiments
4
ANALYSIS
Understand trends, distributions and segments
3
PREPARATION
Organize and clean data to make it usable
2
STORAGE
Store data in reliable storage
1
COLLECTION
Extract data from source systems



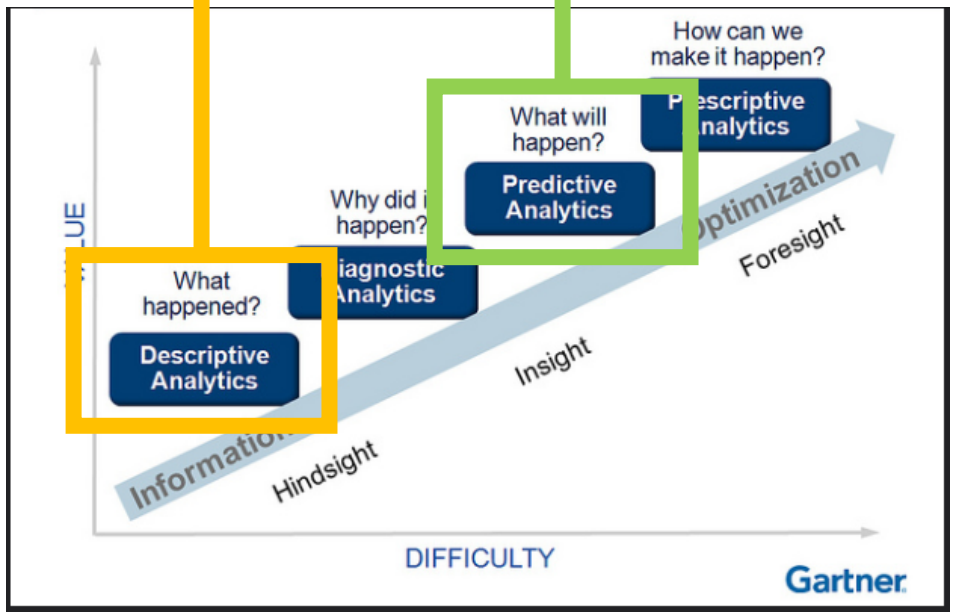


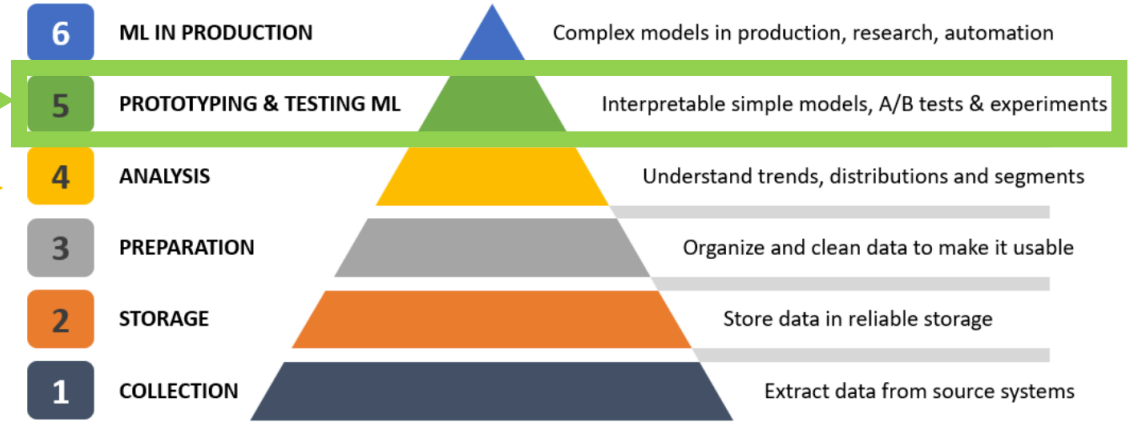
1-28-21	RMarkdown & Import
2-2-21	Web Scraping I
2-4-21	Web Scraping II
2-9-21	Wrangling I
2-11-21	Wrangling II
2-16-21	Transformation





2-16-21	Transformation
2-18-21	Visualization I
2-23-21	Visualization II
2-25-21	Visualization III
3-2-21	Visualization IV
4-1-21	Dimensionality Reduction
4-6-21	Clustering





3-4-21	Model Basics
3-16-21	Model Fitting I
3-18-21	Model Fitting II
3-23-21	Model Fitting III
3-25-21	Overfitting I
3-30-21	Overfitting II
4-13-21	Model Performance I
4-15-21	Model Performance II

Quiz "Question"

- Building a customer service chatbot
  - Kim and her data team are working on a customer service chatbot. They will use transcripts from over 300,000 customer service interactions to train a chatbot to answer customer questions
    - Classify each action below that Kim's team will take as either **Data Collection**, **Preparation and Analysis (or Descriptive Analytics)**, or **Prototyping and Testing ML (or Predictive Analytics)**

- User a Markov model to predict responses for each question
- Gather customer information for each conversation
- Plot the number of conversations vs. the time of day
- Create a bar chart of the number of conversations of each type
- Get the timestamps for each transcript
- Create an algorithm that classifies the initial customer question

Creating Value with Data: Fraud Detection

Suppose you work in fraud detection at a large bank. You'd like to use data to determine the probability that the transaction is fake. In other words, the question you want to answer is: **What is the probability that this transaction is fraudulent?**

Where should you and your team start if you are to answer this question?

Creating Value with Data: Fraud Detection

Suppose you work in fraud detection at a large bank. You'd like to use data to determine the probability that the transaction is fake

- By **gathering information about each purchase**, such as the amount, date, location, purchase type, and card-holder's address. You'll need many examples of transactions, including this information, as well as a label that tells you whether each transaction is valid or fraudulent. You likely have this information in a database, so you can use SQL to get it

Now that you have the data, what should you do next?

Creating Value with Data: Fraud Detection

Suppose you work in fraud detection at a large bank. You'd like to use data to determine the probability that the transaction is fake

- **Store the records in a reliable location** and **make sure the data are tidy, clean, and ready** for analysis. This can take a lot of time depending on the nature of the data. It will require further SQL skills and it may involve the use of regular expression (if you have text data that you want to use in the analysis)

The data are tidy, clean, and ready, now what?

Creating Value with Data: Fraud Detection

Suppose you work in fraud detection at a large bank. You'd like to use data to determine the probability that the transaction is fake

- Time to **understand your data!** How many transactions do you have? Out of these, how many are fraudulent? What is the distribution of fraudulent transactions? What is the correlation between the variables in your dataset? These are sample questions. Here, you will move away from SQL toward Statistics

I have a pretty good understanding of the data, what to do next?

Creating Value with Data: Fraud Detection

Suppose you work in fraud detection at a large bank. You'd like to use data to determine the probability that the transaction is fake

- Use your data insights and contextual knowledge to **build a predictive model.** Since the task at hand is of classification (categorize transactions into fraudulent versus valid), you will build some type of classification model (e.g., logistic regression). A model is a simplified representation of reality created for a specific purpose. Competency in ML will be important

The model is created, and I have results, am I done yet?

Creating Value with Data: Fraud Detection

Suppose you work in fraud detection at a large bank. You'd like to use data to determine the probability that the transaction is fake

- Not really. You still need to **evaluate the performance of your model**. Chances are your first model will not be your best model, and some tweaking will be necessary. Again, competency in ML (e.g., cross-validation) will be important

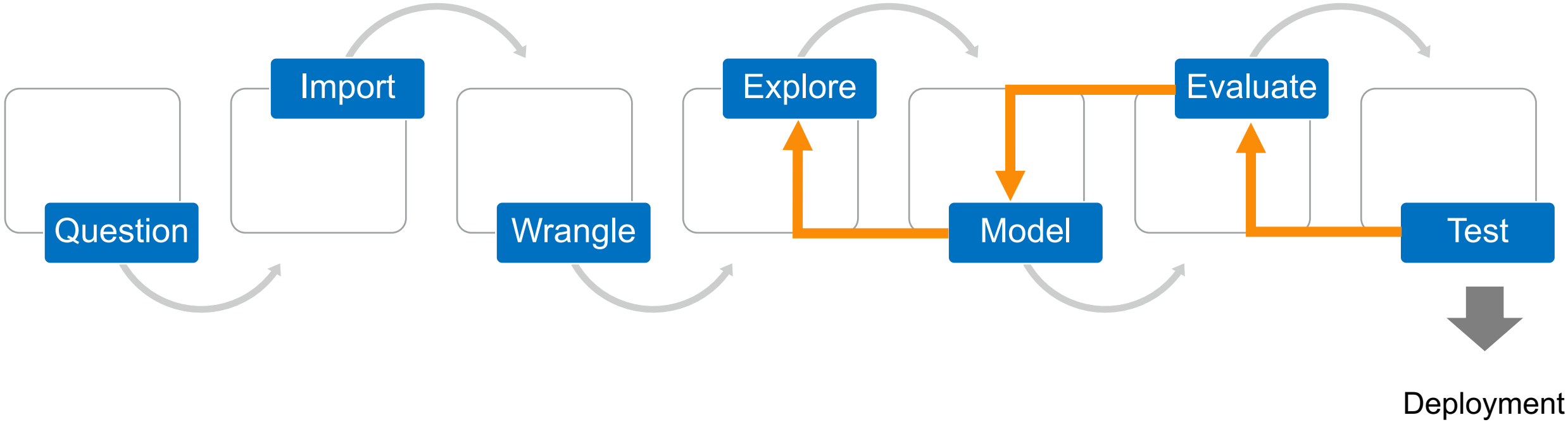
I have developed the best model ever! How to deploy it?

Creating Value with Data: Fraud Detection

Suppose you work in fraud detection at a large bank. You'd like to use data to determine the probability that the transaction is fake

- First, are you sure that you have developed the best model, ever? Have you **tested it with new data? If not, then you need to do that!** If the results are good, then you can start thinking about production and deployment so that the bank you work for can start using the model you and your team developed to detect fraud in real-time

Data Value Creation Model



Database → Statistics → Machine Learning

What Is *Your* Role?

- (Business) Data Scientist
  - Does the actual modeling
  - Applied statistician X Computer scientist (X Business expert)
- Collaborator in a data-centric project
  - Translates from business to the execution and back
- Manager of a data-centric project
  - Understands the potential
  - Has the ability to evaluate a proposal and the execution
  - Has the ability to interface with a broad variety of people
- Strategist, Investor, Entrepreneur, …
  - Envisions opportunities, produces novel ideas, designs data projects conceptually

At Home Exercise

Think of a topic of interest and write down …

- … a well-defined question you have about it
- … what kind of data you will need, and where you can find such data
- … how you may need to wrangle the data
- … the ways you may explore and better understand the data
- … the model you may need to build, evaluate, and test
- … the practical value of your model, and how to deploy it

Thank You!