

Data Science & the Data Cycle

Girl Geeks Waterloo

May 2015

Jennifer Nguyen

Why do we need data?

To inform our decisions

Why do we need data?

Everyday examples

- To decide what to wear in the morning



Why do we need data?

Everyday examples

- To book a flight



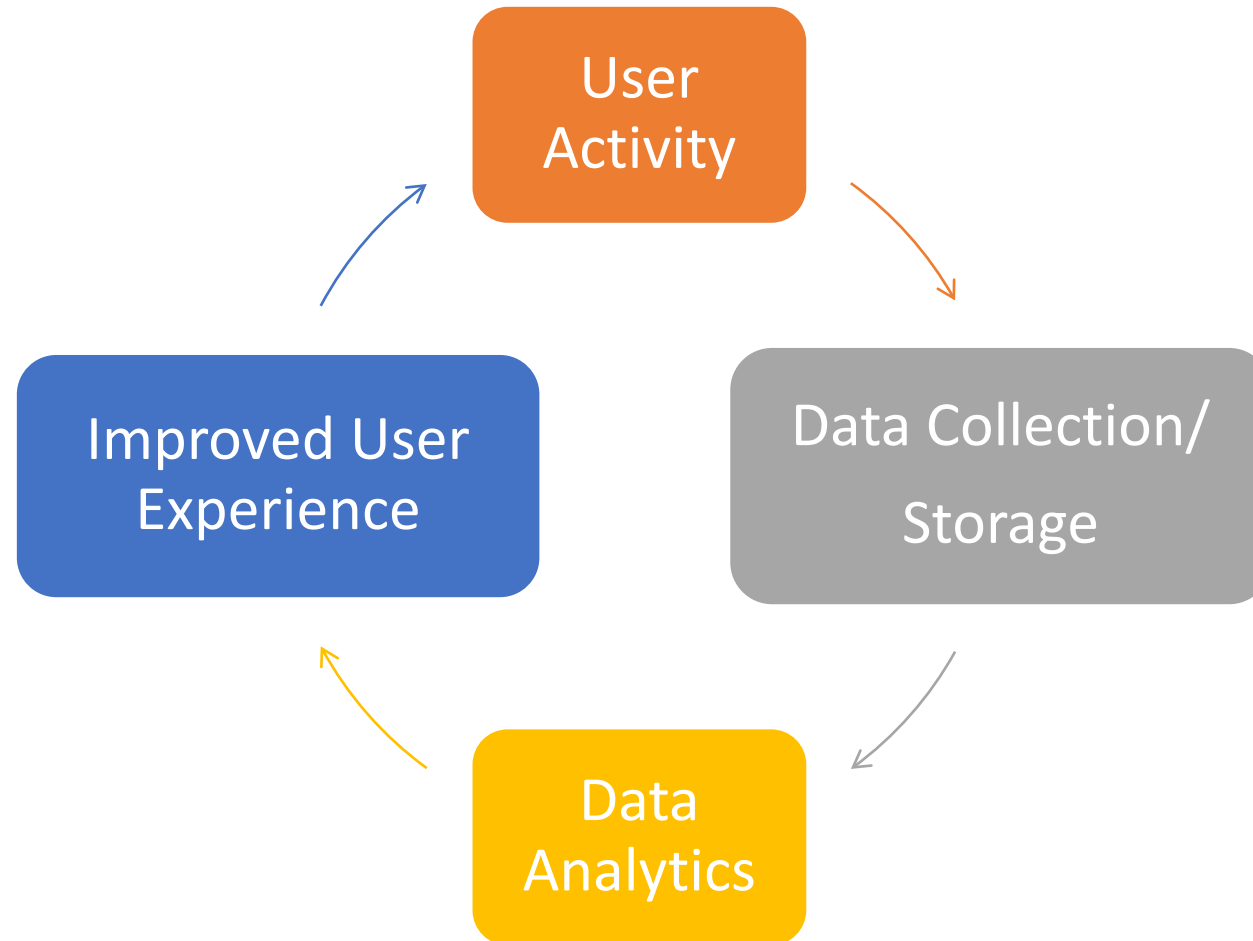
Why do we need data?

Everyday examples

- To negotiate a salary



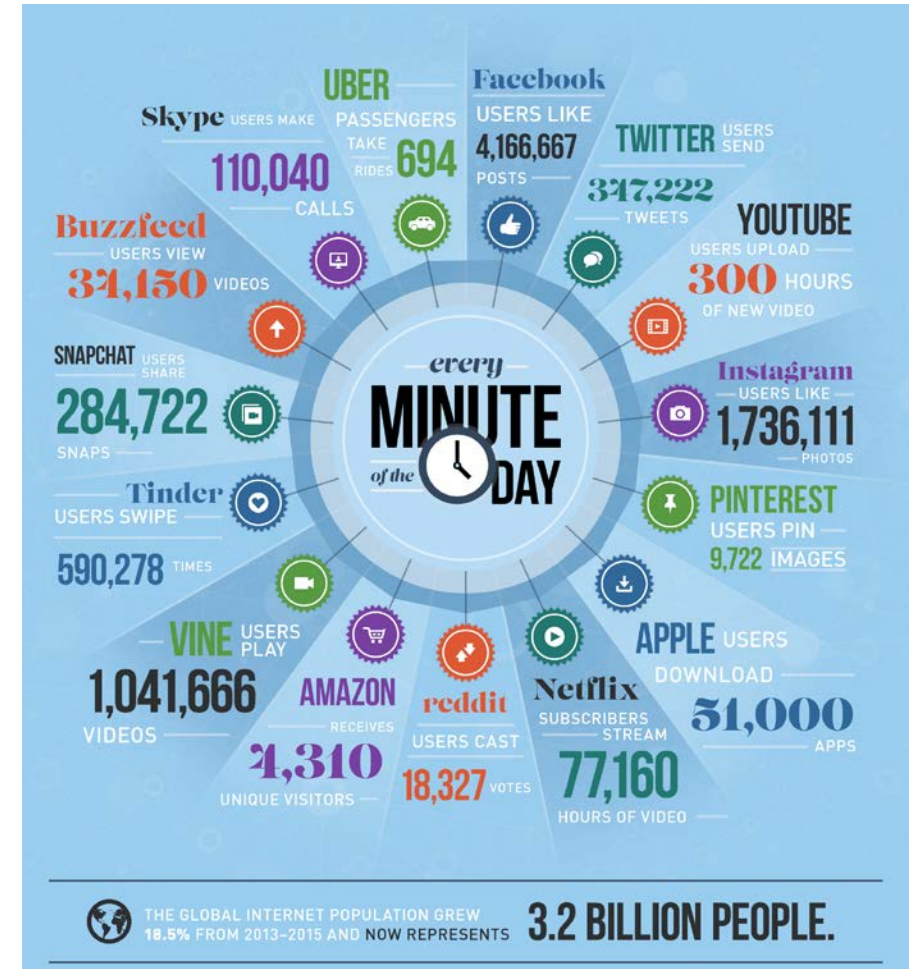
Data Cycle



How do we create data?

How do we create data?

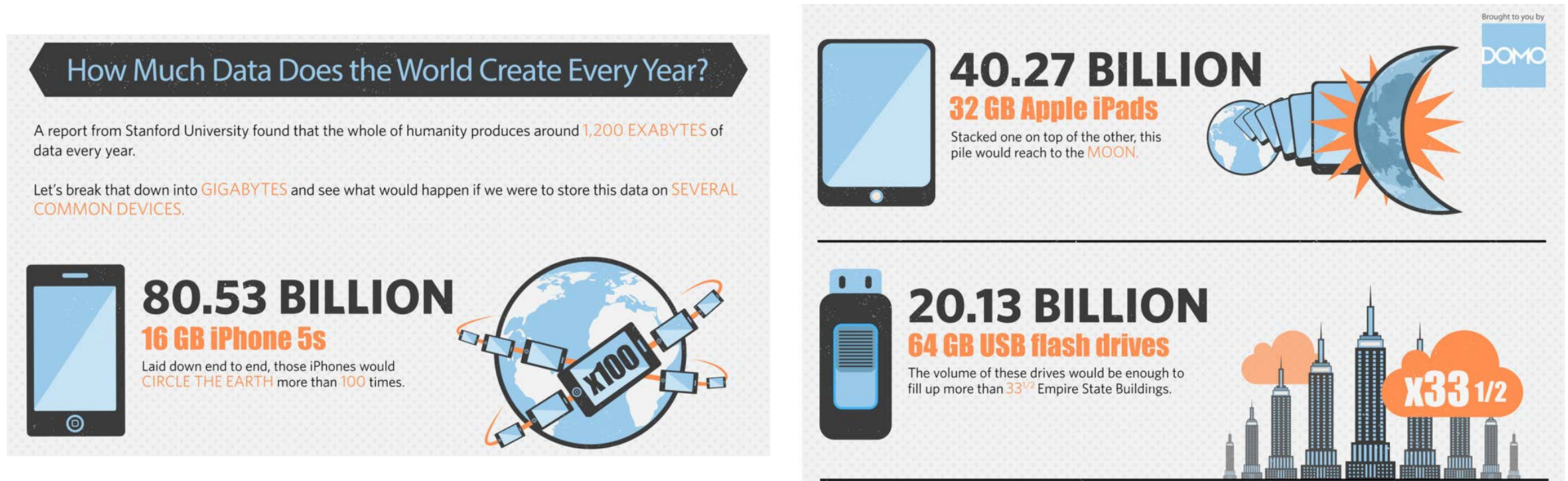
- Every interaction on any platform is collected
 - Clicks, mouse hovers, scrolls
- User generated data
 - photos, videos, “likes”
 - purchases
- Direct feedback
 - Registration information



Source: DOMO

Data is collected and stored

- Data is collected and stored in large databases to be analyzed

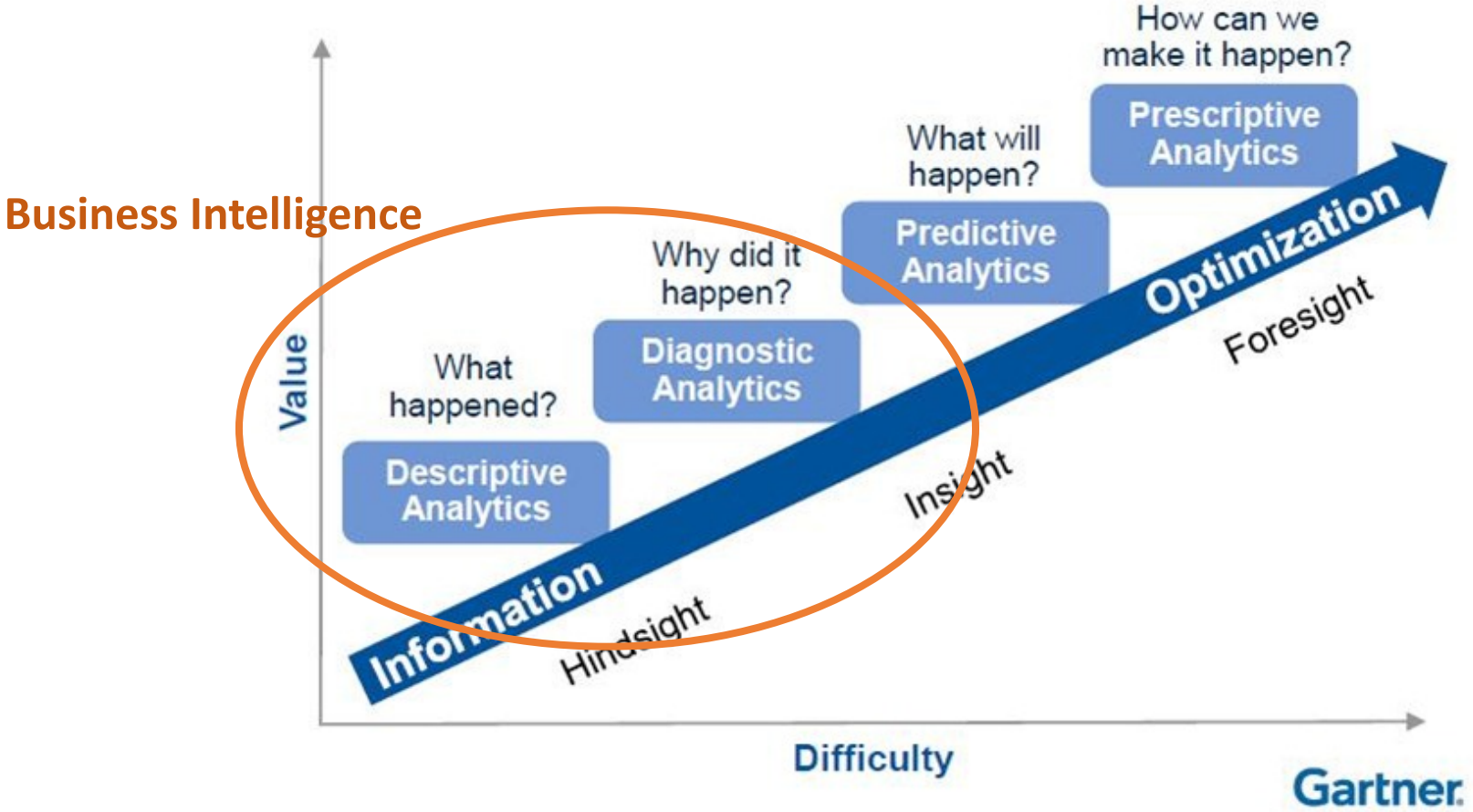


Source: DOMO

What is it used for?

*“Not only are we doing more with data, data is doing more with us” – Jer Thorp,
DOMO*

Data Analytics Spectrum



Gartner.

Source: Gartner

Business Intelligence

a.k.a. reactive analytics

- **Reactive analytics** answers questions related to **past/current events**
- E.g., “How are we doing?”, “What went well?”
- Answers are used to inform business decisions



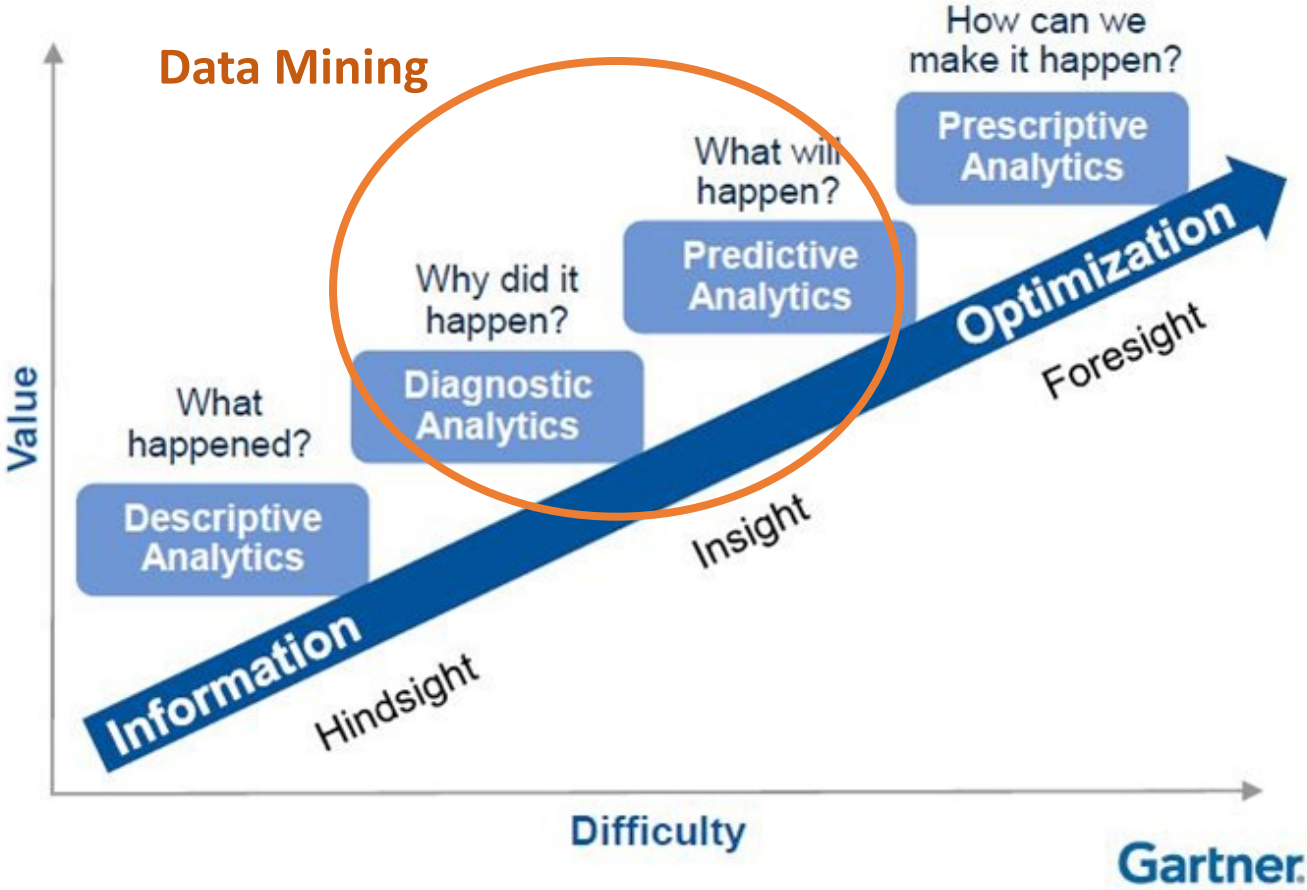
Business Intelligence

a.k.a. reactive analytics

- Measuring web traffic
 - How did the volume of traffic change since yesterday/last week/last month/last year?
 - Where are users coming from?
 - Which topics resonated with readers?



Data Analytics Spectrum



Gartner.

Source: Gartner

Data Mining

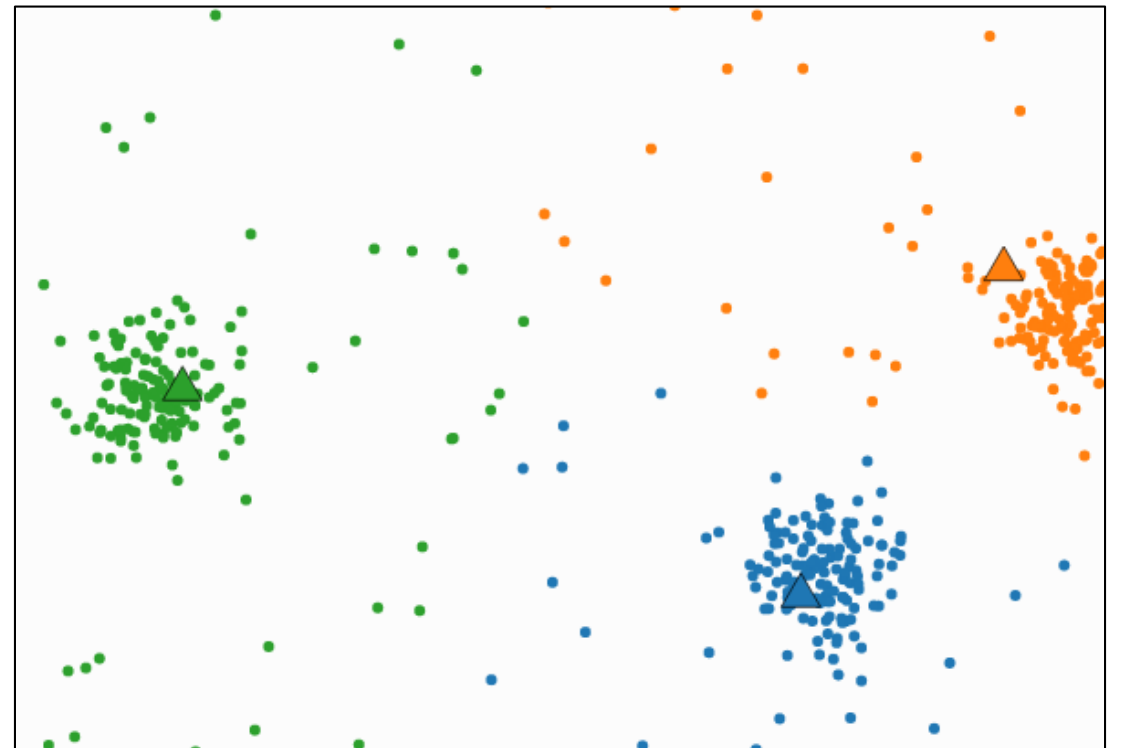
Knowledge Discovery

- Used to find patterns, trends, and insights from data
- Techniques include
 - Anomaly detection
 - Association rule learning
 - Clustering

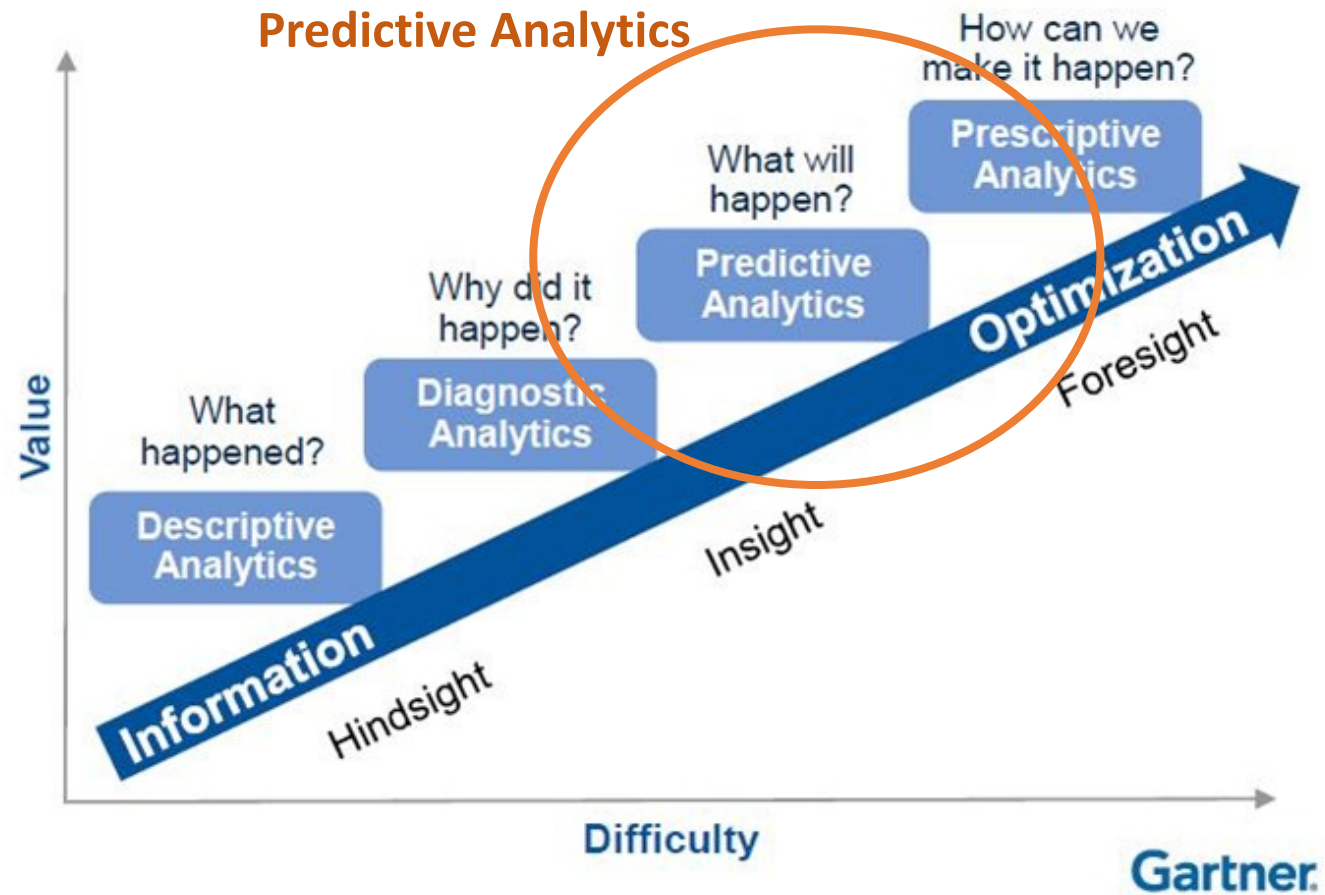


Data Science Technique: Clustering

- How to segment users?
 - [K-Means](#) Clustering
- Clustering can be used to discover communities of users
- Other applications:
 - Find similar items, movies



Data Analytics Spectrum



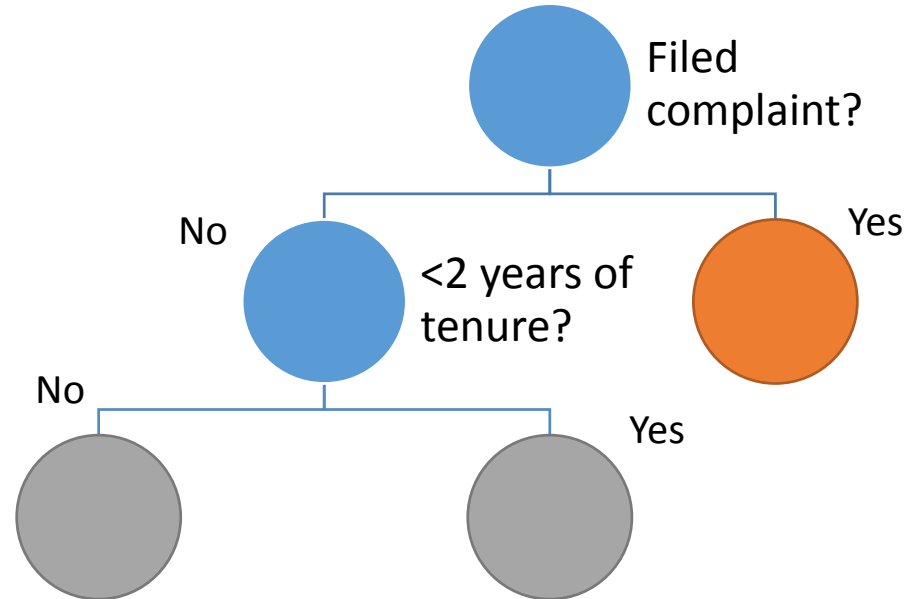
Predictive Analytics

- **Predictive analytics** answers questions related to **future events**
- E.g.:
 - How likely will this student drop out?
 - What would this reader like to read next?
 - Will a customer churn?



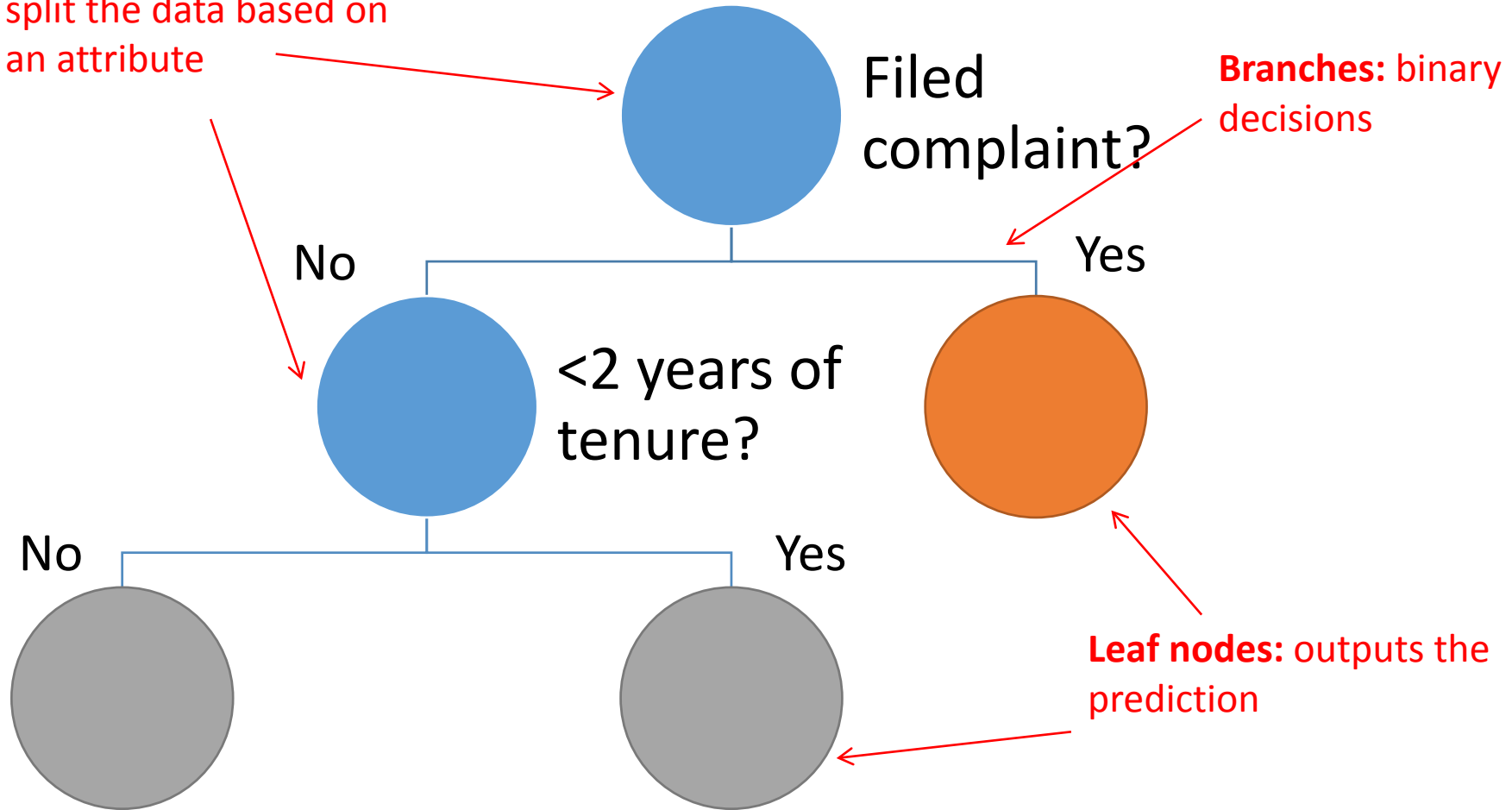
Data Science Technique: Classification

- How to predict customer churn?
 - Logistic Regression
 - **Decision Trees**
 - Random Forests
- Results can be used to “save” a customer



What is a decision tree?

Root/internal nodes:
split the data based on
an attribute



How to build a decision tree

- Given the following data set:

<2 years of tenure?	Filed Complaint?	Churned?
N	N	N
Y	N	Y
N	N	N
N	N	N
N	Y	Y
Y	N	N
N	Y	N
N	Y	Y
Y	N	N
Y	N	N

Example adapted from Michael S. Lewicki, Artificial Intelligence: Learning and Decision Trees,
<http://www.cs.cmu.edu/afs/cs/academic/class/15381-s07/www/slides/041007decisionTrees1.pdf>

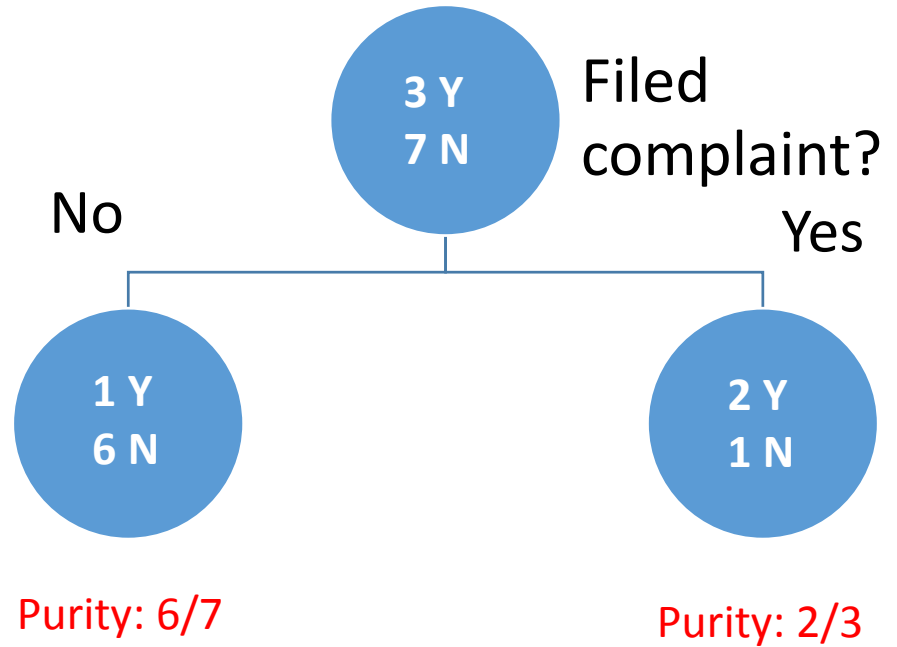
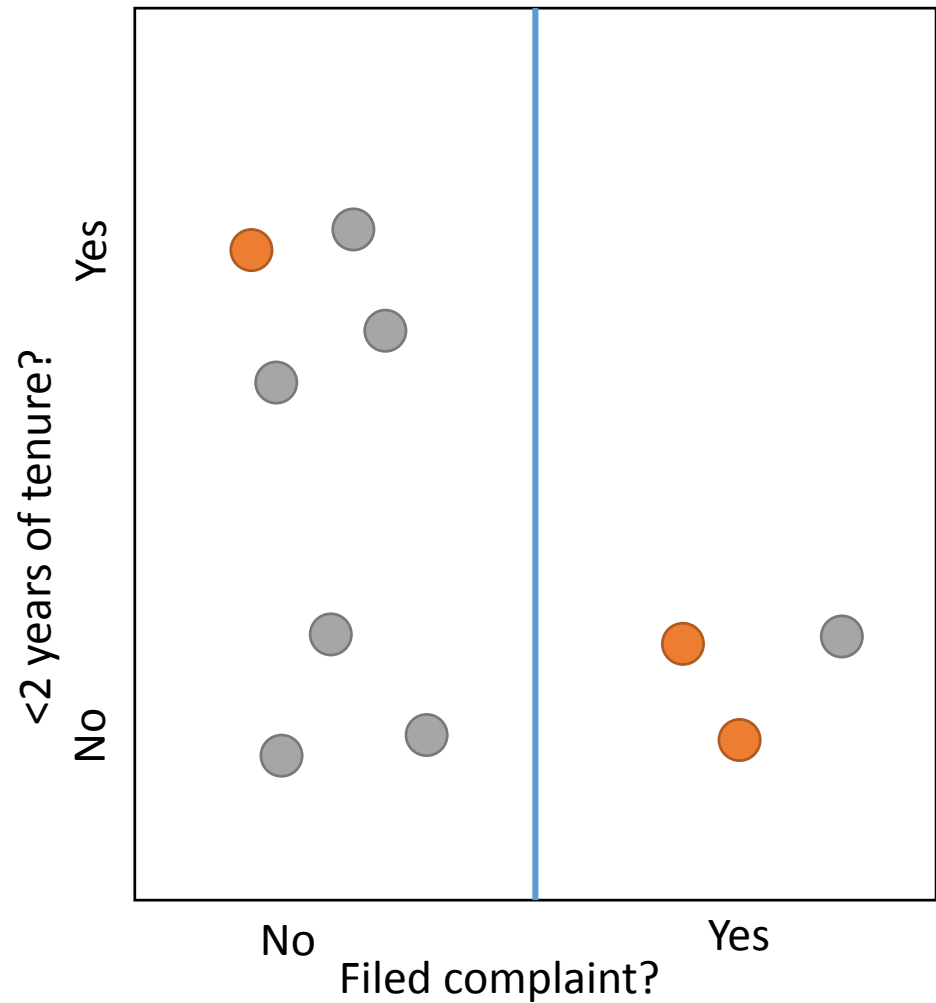
How to build a decision tree

- Goal:
 - Split the data in such a way to achieve high classification accuracy
- Requires knowing which attributes to use and in which order
- Use “greedy” algorithm:
 - Choose attribute that gives best split at each level of tree

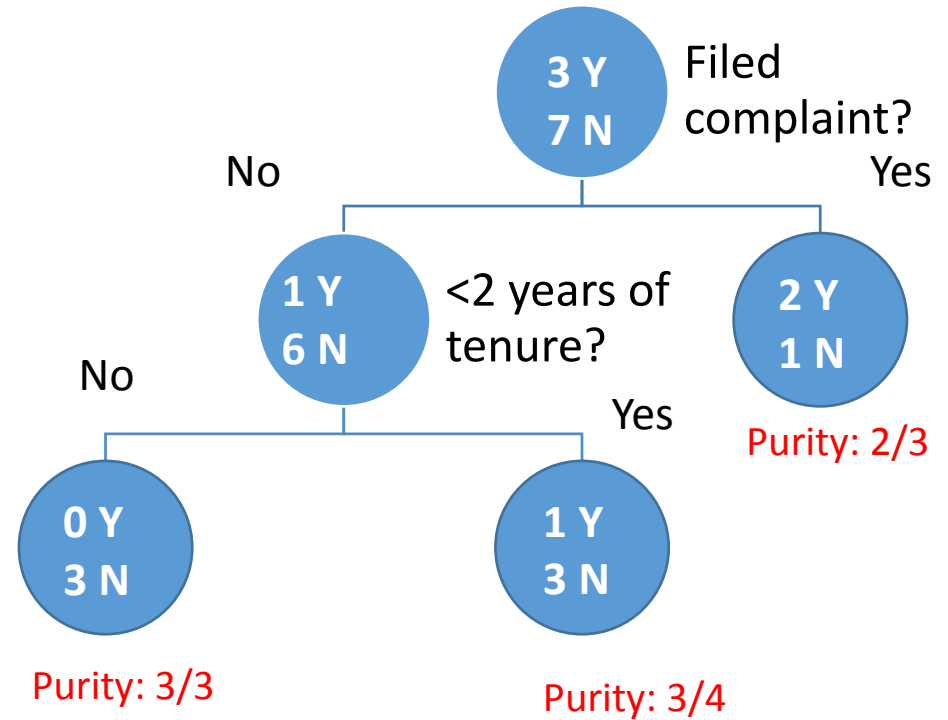
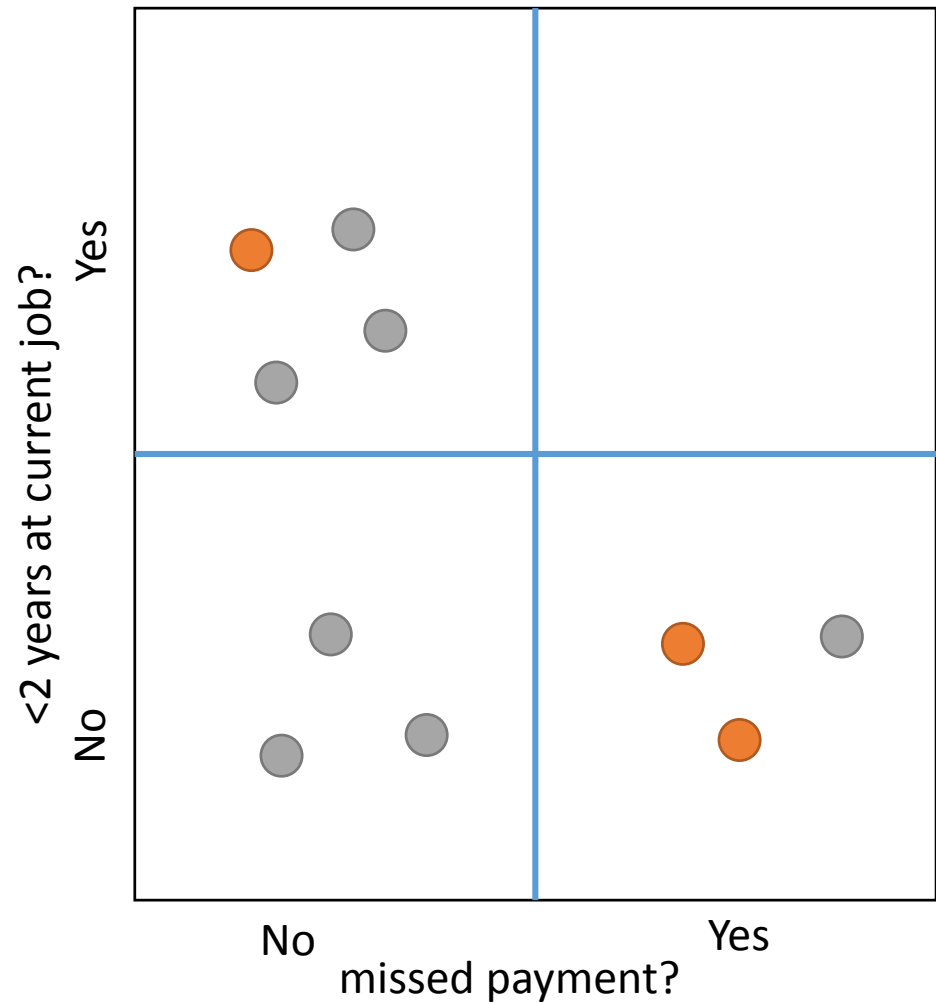
Recursive Algorithm

1. Start with all data
2. Find query that gives best split.
3. Create child nodes
4. Recurse until stopping criterion:
 - Node consists of one dominant class, considered the node's "purity"

Building a Decision Tree



Building a Decision Tree

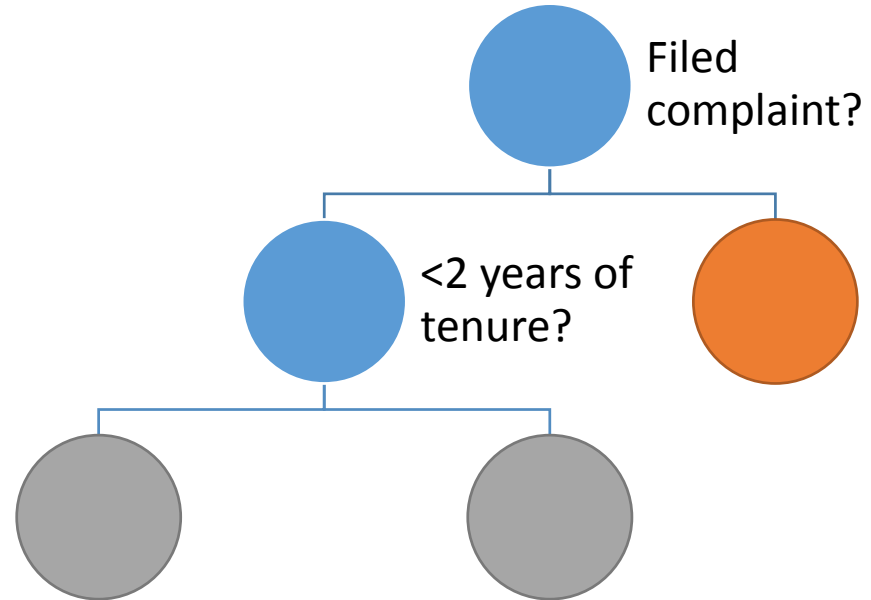


How to use a decision tree

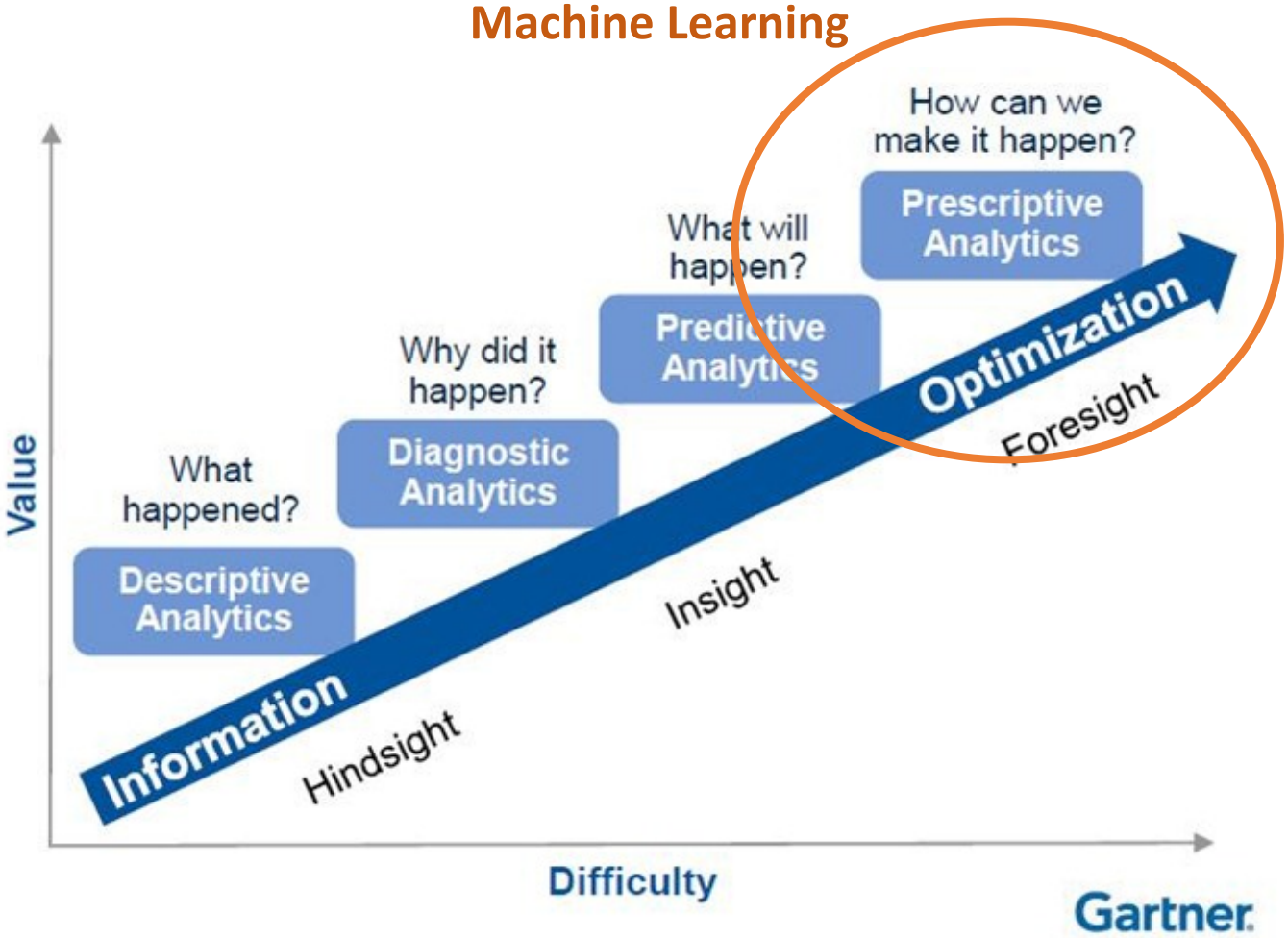
- Given new data samples, how to predict if the individual will churn?

<2 years of tenure?	Filed complaint?	Churned?
N	N	?
Y	N	?
N	Y	?

- Use recursive tree traversal algorithm to find the corresponding leaf node



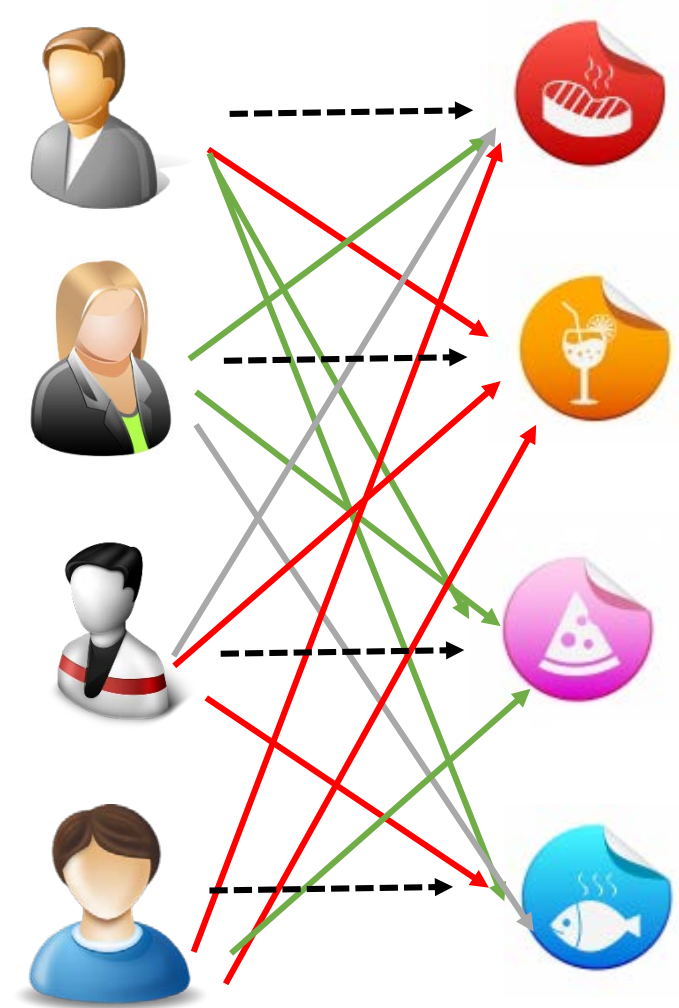
Data Analytics Spectrum















Source: Gartner

Data Science Technique: Collaborative Filtering

- Used in simple recommendation engines to recommend items to users
- Items can be news articles, movies, clothing, friends



Data Mining Technique: Collaborative Filtering













	Item 1	Item 2	Item 3	Item 4
User A				
User B				
User C				
User D				

(a) Known ratings, r_{ui}

Data Mining Technique: Collaborative Filtering

















- Goal is to fill in the missing empty cells with a prediction
- Items with positive predictions are recommended to users
- Predictions are influenced by ratings from other users
- The more similar two users are with respect to their ratings, the more they will influence the prediction

Data Mining Technique: Collaborative Filtering

	Item 1	Item 2	Item 3	Item 4
User A				
User B				
User C				
User D				

(a) Known ratings, r_{ui}

Data Mining Technique: Collaborative Filtering

	Item 1	Item 2	Item 3	Item 4
User A				
User B				
User C				
User D				

(b) Predictions, \hat{r}_{ui}

Data Mining Technique: Collaborative Filtering

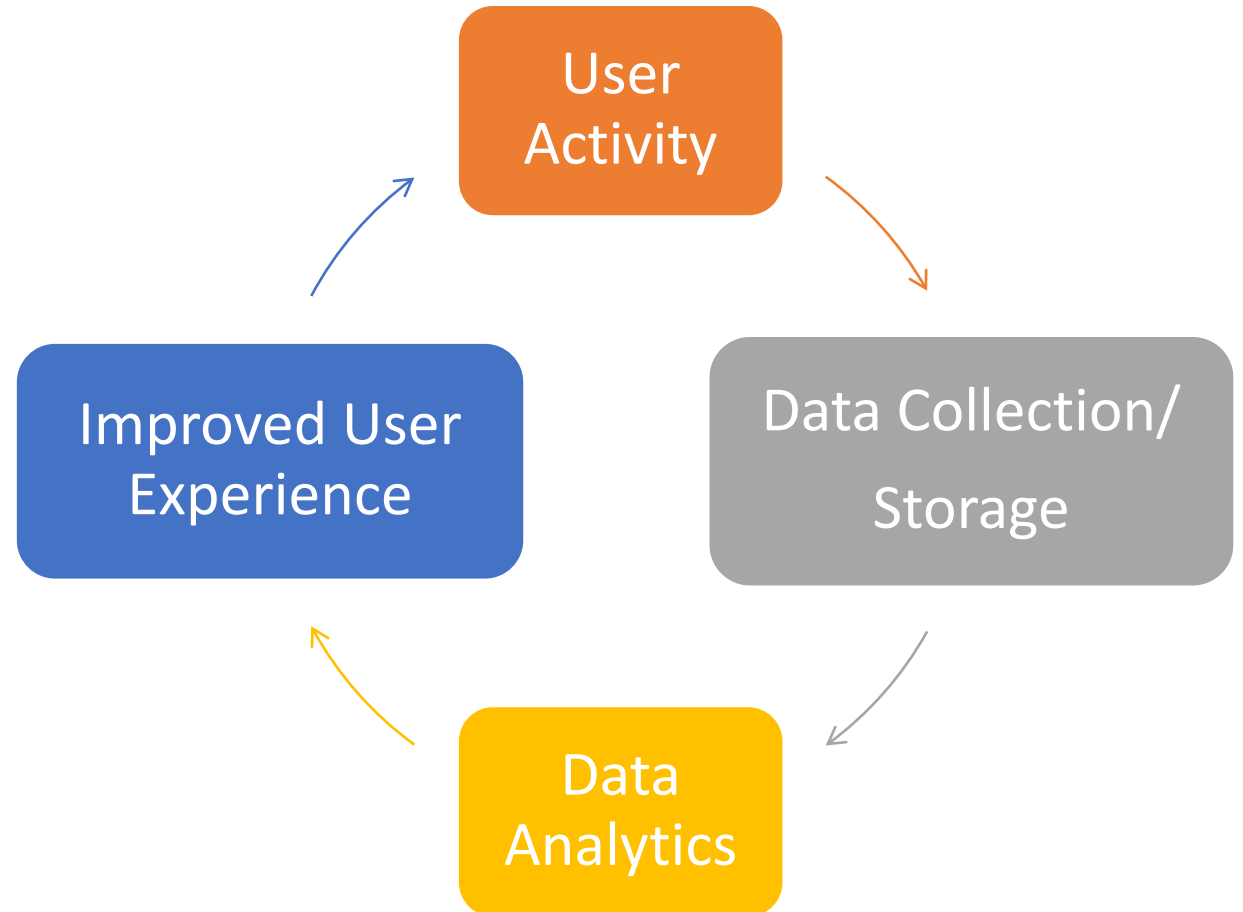
- Disadvantages of this method:
 - Users do not understand why some recommendations are made by the engine
 - Users may not receive recommendations they like because other users have not liked them
- Advanced recommendation methods exist to address these shortcomings

How do users benefit from data science?

- Users get a more personalized experience that is tailored to their interests
 - E.g., Nest Thermostat, PC Plus
- Users save time from not having to sift through the vast sea of options
 - E.g., Netflix, online retailers, LinkedIn

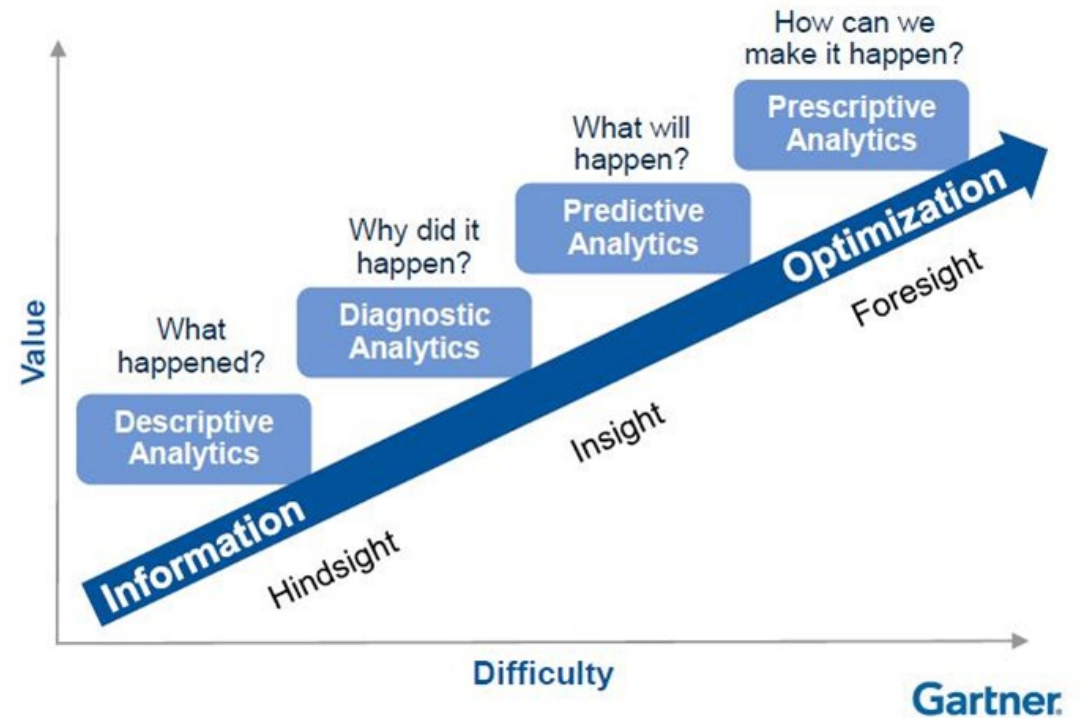
Key Takeaways

- Data science is part of an iterative process to continually improve the user experience



Key Takeaways

- There are many flavours of data analytics
 - Which one to use depends on the questions you want answered



Questions?