

# Data Science and Data Engineering

## Curriculum Overview

# Learning Process

## Preparatory Work

Introduction to Big Data, Predictive Analytics and Data Science

Introduction to Data Mining

Introduction to R Programming

Introduction to Azure ML Studio

## 5-day Bootcamp

Rigorous in-person training (8am-6pm)

Theory and hands-on work in data science, predictive analytics, machine learning

Big data engineering background needed for you to be effective as a data scientist

## Kaggle Project

Compete with thousands of data scientist from across the world

You will be mentored by one of our teaching team members

# Data Science

- Emphasis on the process and best practices and not on covering as many topics as possible
- Data exploration, visualization, feature engineering, machine learning and predictive analytics
- 50% theory. 50% Hands-on Exercises
- Math/Theory is minimal but not trivial
- Primary tools: R and Azure ML Studio

# Data Engineering

- Teach enough data engineering skills to be effective data scientist
- 20% theory. 80% hands-on
- Handle volume, variety and velocity of data
- Internet of Things (IoT) hack day.

# Hack Day

- Gather temperature and humidity data in real-time
- Use message queues, stream processors to get real time analytics
- Answer questions like:
  - What was the average temperature in last 5 seconds?
  - How often did the temperature exceed the allowed threshold?



# Story Behind No Prerequisites

- You **must** attend all the pre-bootcamp webinars to be ready for the 5-day in person training

# Logistics

- ~8 hours of pre-bootcamp work
- Bootcamp: 5-days. 8am-6pm daily
- Slides, sample code and other resources are consolidated in a git repository
- Office hours. Kaggle. LinkedIn group

# Please keep the session interactive

- Interrupt and ask questions often.



# Introduction to Big Data, Predictive Analytics, and Data Science

# Big Data and Data Science Everywhere



Web search and  
online ads



Insurance



Telcos



Online Education



Online Retail



Social Networks



Entertainment



Healthcare

# Online Shopping

## Best Value

Buy **Predictive Analytics: The Power of Prediction Who Will Click, Buy, Lie, or Die** and get **How to Measure Anything: Finding the Value of Intangibles in Business** at an **additional 5% off** Amazon.com's everyday low price.



**Buy together today: \$45.43**

[Add both to Cart](#)

[Show availability and shipping details](#)

## Customers Who Bought This Item Also Bought



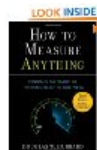
Predictive Analytics:  
Microsoft Excel  
› Conrad Carlberg  
★★★★☆ (10)  
Paperback  
\$24.36



Big Data: A Revolution That  
Will Transform ...  
› Viktor Mayer-Schonberger  
★★★★☆ (32)  
Hardcover  
\$15.84



Big Data, Big Analytics:  
Emerging Business ...  
› Michael Minelli  
★★★★★ (6)  
Hardcover  
\$32.82



How to Measure Anything:  
Finding the Value of ...  
› Douglas W. Hubbard  
★★★★☆ (56)  
Hardcover  
\$31.96



Secrets of Analytical  
Leaders: Insights ...  
› Wayne Eckerson  
★★★★★ (10)  
Perfect Paperback  
\$44.96



Big Data Analytics:  
Disruptive ...  
› Dr. Arvind Sathi  
★★★★☆ (5)  
Paperback  
\$10.45

# Social Networks

twitter

## Who to follow

Twitter accounts suggested for you based on who you follow and more.

Search using a person's full name or @username

Search Twitter



**DataQualityPro.com** @dataqualitypro

The most popular online data quality community resource for anyone requiring free expert tutorials, techniques, articles or technology advice.

Followed by Big Data Science and Data Science Central.



Follow

II

**Stat Fact** @StatFact

One statistics tip per day M-F from @JohnDCook. See also @ProbFact, @CompSciFact, and @SciPyTip.

Followed by Data Science Central and Big Data Science.



Follow



**Anthony Goldbloom** @antgoldbloom  
Founder and CEO of Kaggle.



Follow

facebook

## People You May Know

See All



**Andres Ponce**

Add Friend



**Jessica Clark**

1 mutual friend

Add Friend



**Melody Vilantino**

7 mutual friends

Add Friend



**Isabella Lopez**

2 mutual friends

Add Friend

LinkedIn

## JOBS YOU MAY BE INTERESTED IN

**TIGER**  
ANALYTICS

**Software Developer,  
Data Analytics**  
Tiger Analytics - Raleigh...

Sponsored



**Machine Learning Scientist**  
Amazon - Greater Seattle A...

×



**Data Scientist, Senior -  
OSD D&A ...**  
Microsoft - Bellevue, WA, US

×



**Data Scientist, Senior -  
Bing - D&A...**  
Microsoft - Bellevue, WA, US

×

[Feedback](#) | [See more »](#)



Get hired faster with **Job Seeker Premium**

## GROUPS YOU MAY LIKE

**Microsoft**

**Microsoft Employees  
(Verified)**

Join - Corporate Group



**Data Mining Professionals**

Join - Professional Group



**Data Science Community**

Join - Networking Group

[Feedback](#) | [See more »](#)


# Online Entertainment

X  
I  
T  
E  
N

Close

### Other Movies You Might Enjoy

[Amélie](#)




**Add**

★★★★☆

☐ Not Interested

[Y Tu Mama Tambien](#)




**Add**

★★★★☆

☐ Not Interested

[Guys and Balls](#)




**Add**

★★★★☆

☐ Not Interested

[Mostly Martha](#)

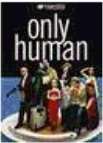


**Add**

★★★★☆

☐ Not Interested

[Only Human](#)




**Add**

★★★★☆

☐ Not Interested


[Russian Dolls](#)



**Add**

★★★★☆

☐ Not Interested



**Eiken has been added to your Queue at position 2.**

This movie is available now.

**Move To Top Of My Queue**

[< Continue Browsing](#)

[Visit your Queue >](#)

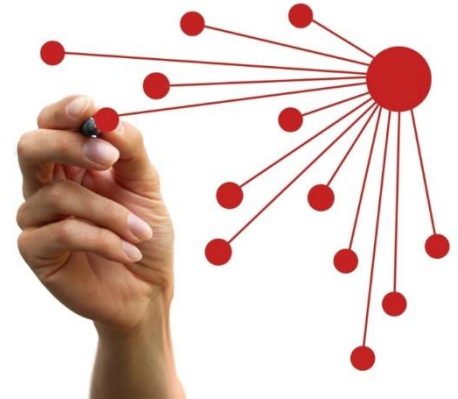
Close

# Brainstorming

- What are some other applications?

# Connecting the Dots

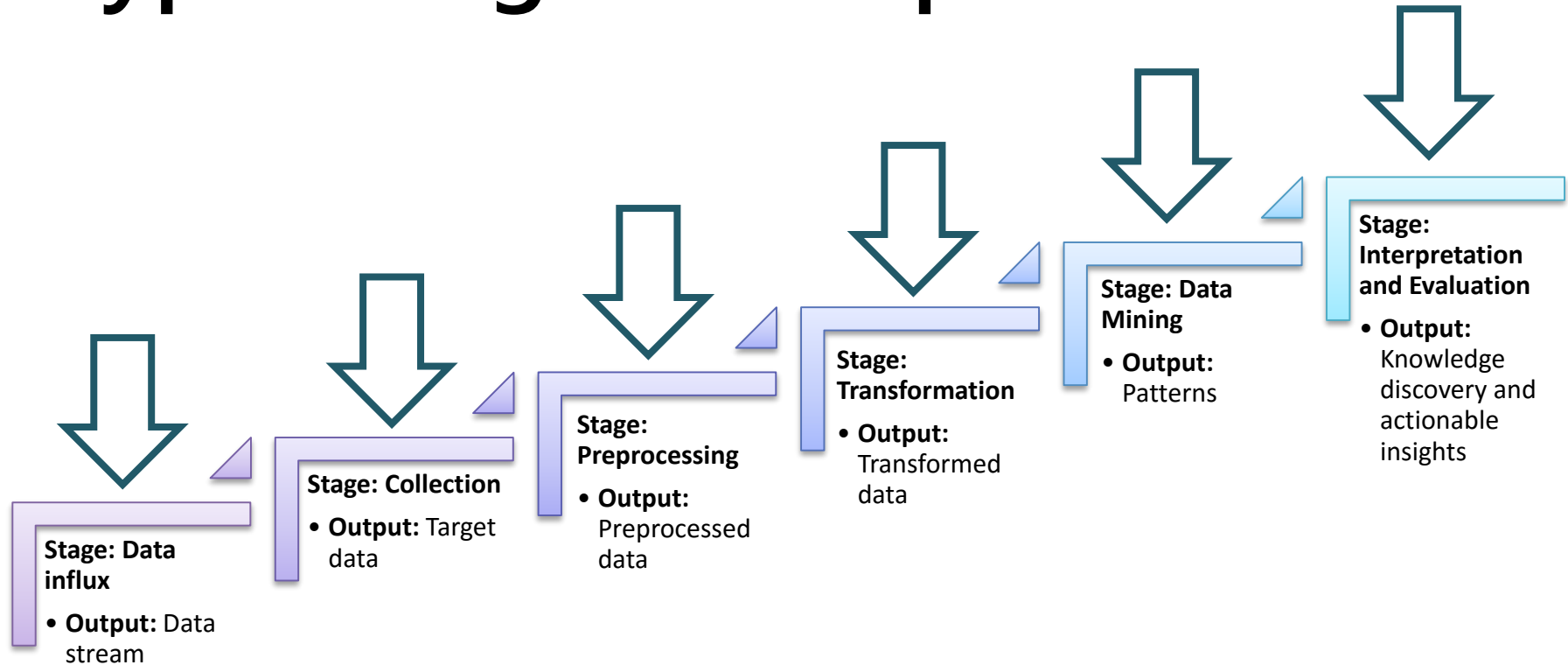
- The underlying magic behind what we saw is 'big data' and 'predictive analytics'

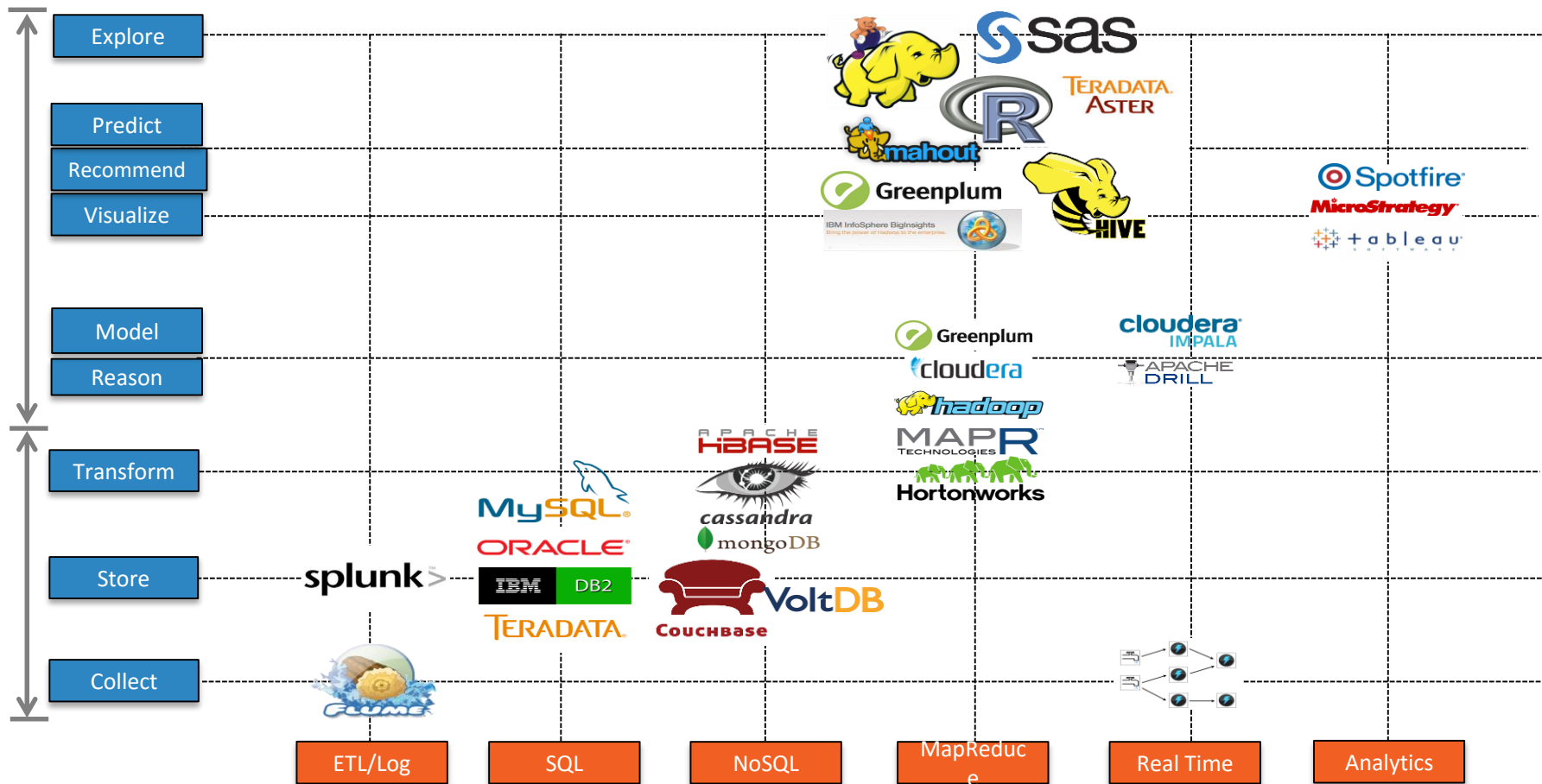


# Let's take a look at a big data pipeline



# Typical Big Data Pipeline





# Data Mining Tasks

## ■ Descriptive Methods:

- Find human-interpretable patterns that describe the data
- Techniques: Clustering, Association Analysis, x-point summaries

## ■ Predictive Methods:

- Use available data to build models that can predict the outcome of future data
- Techniques: Classification, Regression, Anomaly, and Deviation Detection

## ■ Prescriptive Methods:

- Predict future outcomes and suggest actions that may prevent or mitigate the impact of the predicted outcomes
- Techniques: Various optimization techniques

# Traffic Management



## Descriptive [Informing Role]:

- Traffic jam has happened already.
- [Implicit: Do something about it.]

# Traffic Management



## Predictive [Informing and Warning Role]:

- Traffic jam is about to happen in the next 30 minutes.
- [Implicit: Do something before it happens.]

# Traffic Management



Prescriptive [Informing,  
Warning, and Advisory Role]:

Take action so traffic jam does not happen

OR

Traffic jam is about to happen in the next  
30 minutes and you could possibly take  
the following courses of action:

- Route traffic to service road near I-5
- Block more traffic from entering the WA-520 bridge

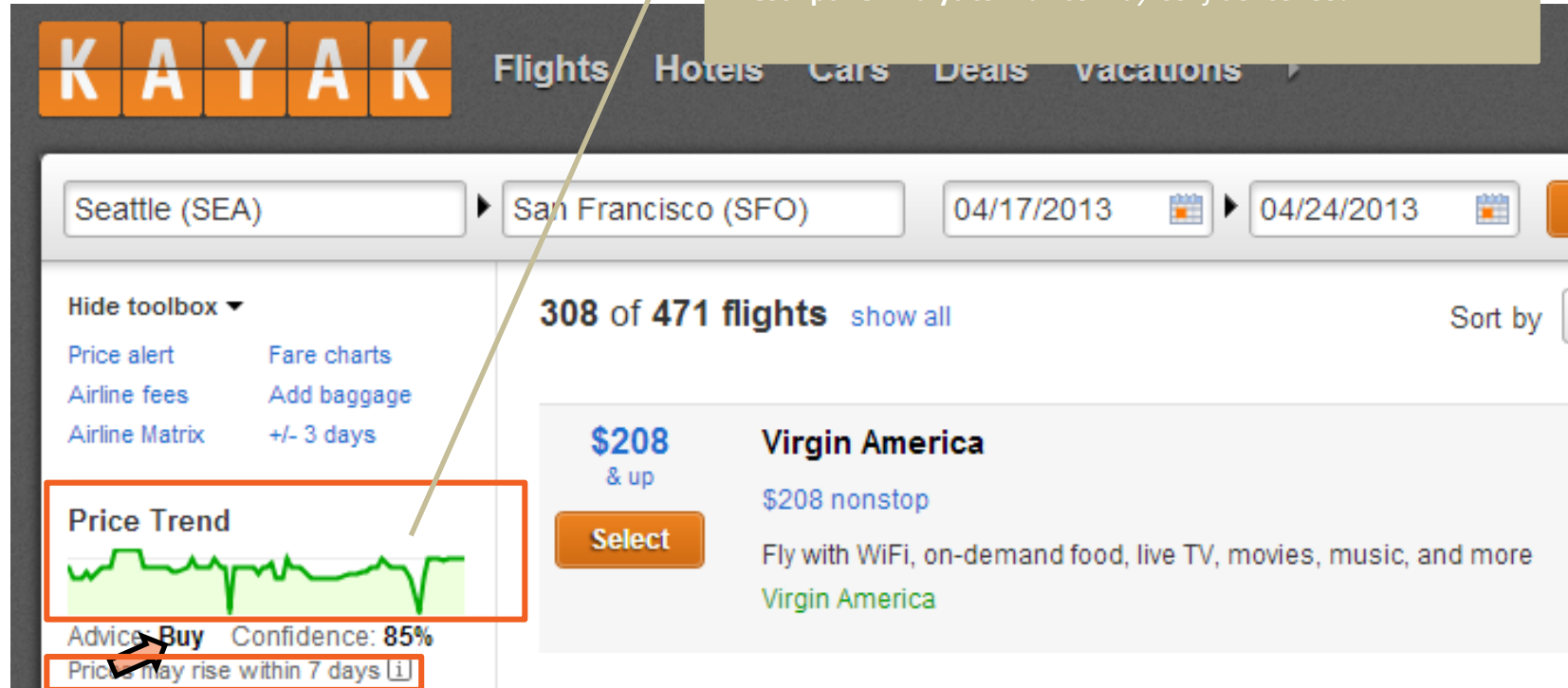


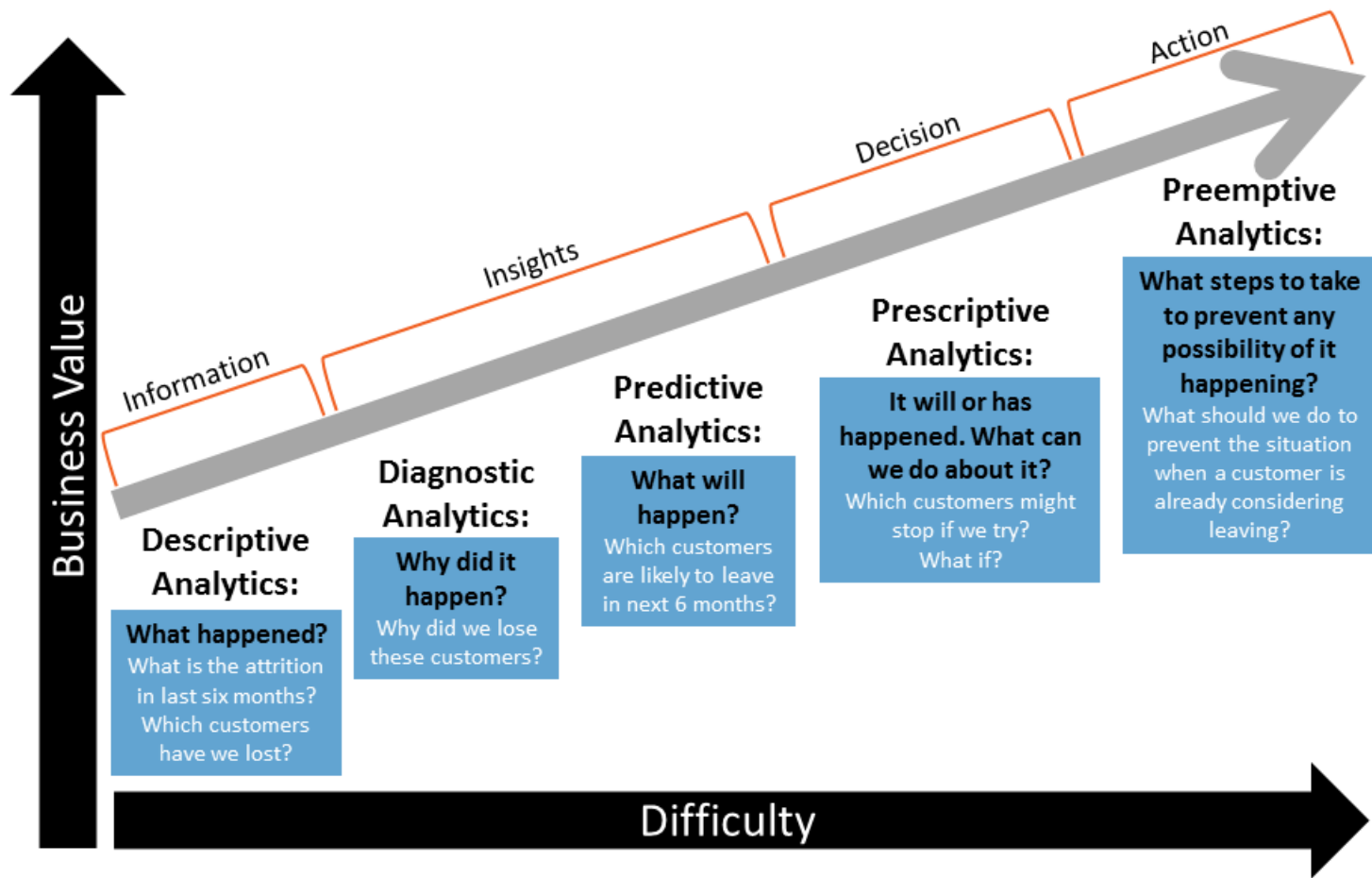
# Online Travel

Descriptive Analytics: Historical price trend and variation

Predictive Analytics: Price may rise in next 7 days

Prescriptive Analytics: Advice: Buy Confidence: 85%



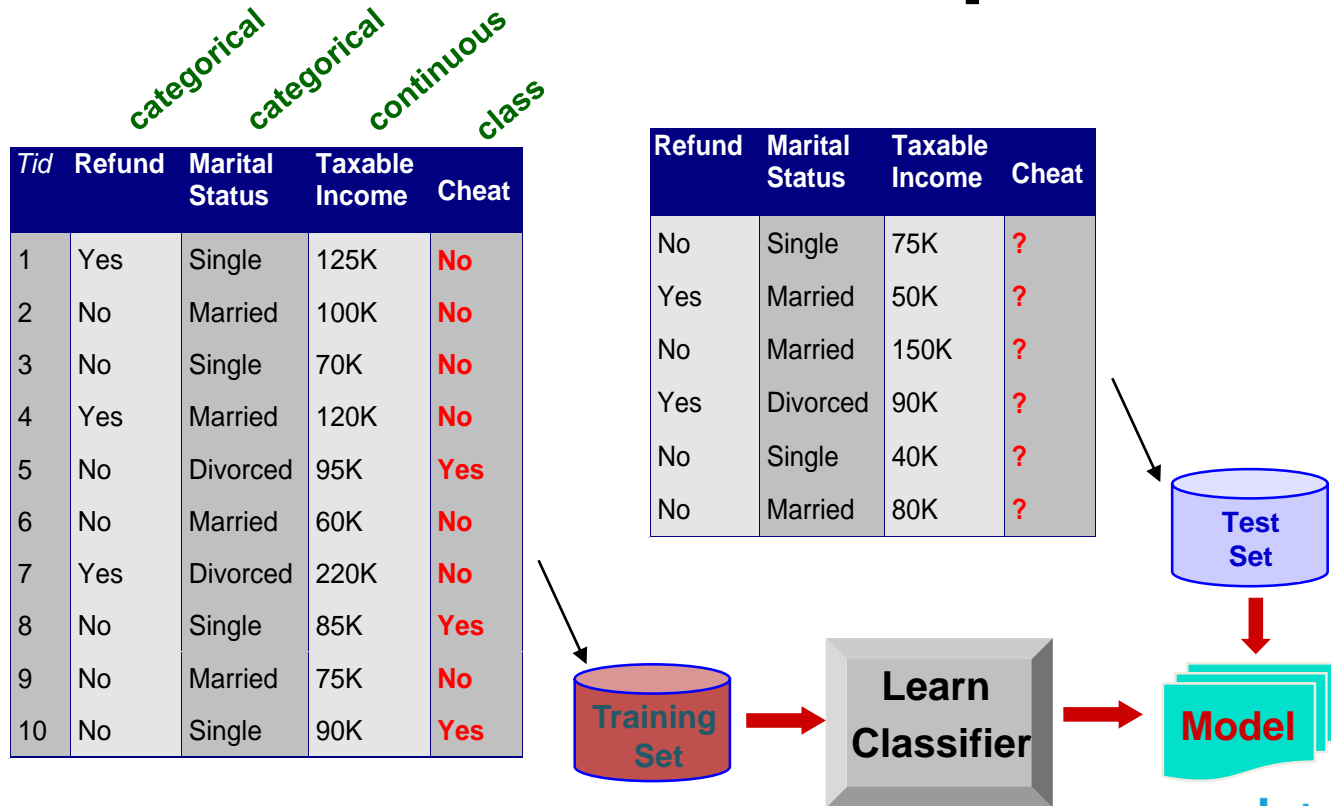




# Data Mining and Predictive Analytics

In the next few slides, we will take a look at some of the most common data mining tasks.

# Classification: A Simple Example



# Classification

- Given a collection of records (**training set**)
  - Each record contains a set of *attributes*; one of the attributes is the *class label*.
- Find a **model** for class attribute as a function of the values of other attributes.
- **Goal:** previously unseen records should be assigned a class as accurately as possible.

# Classification: More Examples

## ■ Direct Marketing

- Goal: reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product

## ■ Fraud Detection

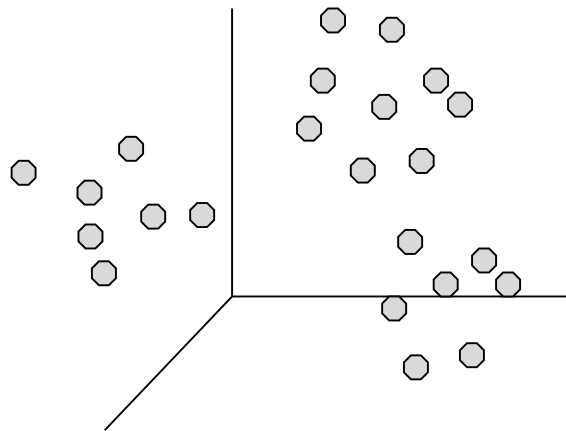
- Goal: predict fraudulent cases in credit card transactions

## ■ Customer Attrition/Churn

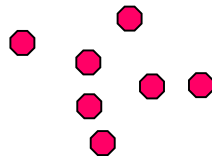
- Goal: predict whether a customer is likely to be lost to a competitor

# Clustering: An Illustration

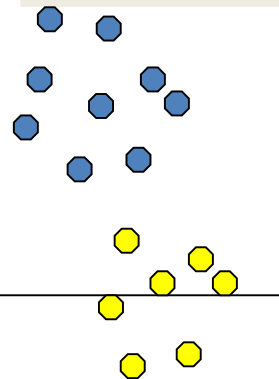
Clustering in 3-D space using Euclidean distance



Intra-cluster distances  
are minimized



Inter-cluster distances  
are maximized



# Clustering: Examples

- Subdivide the market into distinct subsets of customers where any subset may conceivably be selected as a segment to be reached with a particular offer



# Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that:
  - Data points within a cluster have more similarities with one another
  - Data points in different clusters have less similarities with one another

# Clustering: Similarity Measures

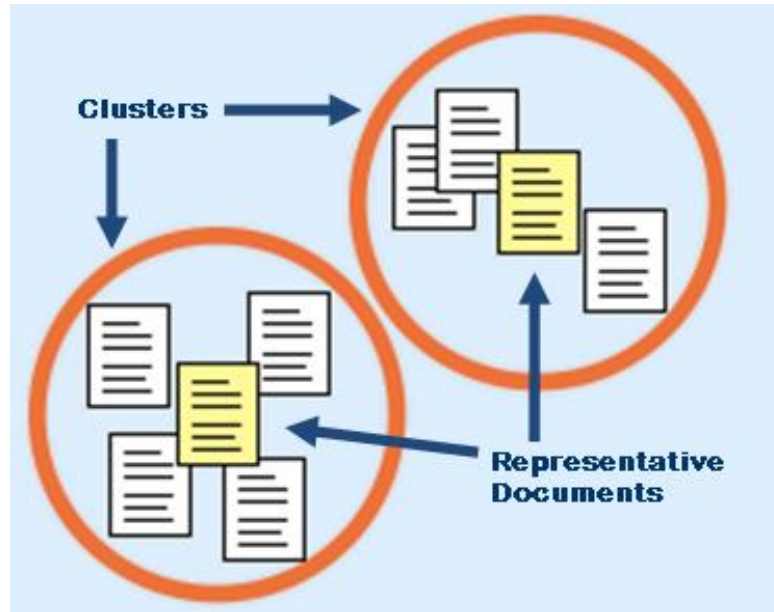
- Similarity Measures:

- Euclidean Distance if attributes are continuous
- Other problem-specific measures
- **Example:** If a particular word occurs in two documents or not

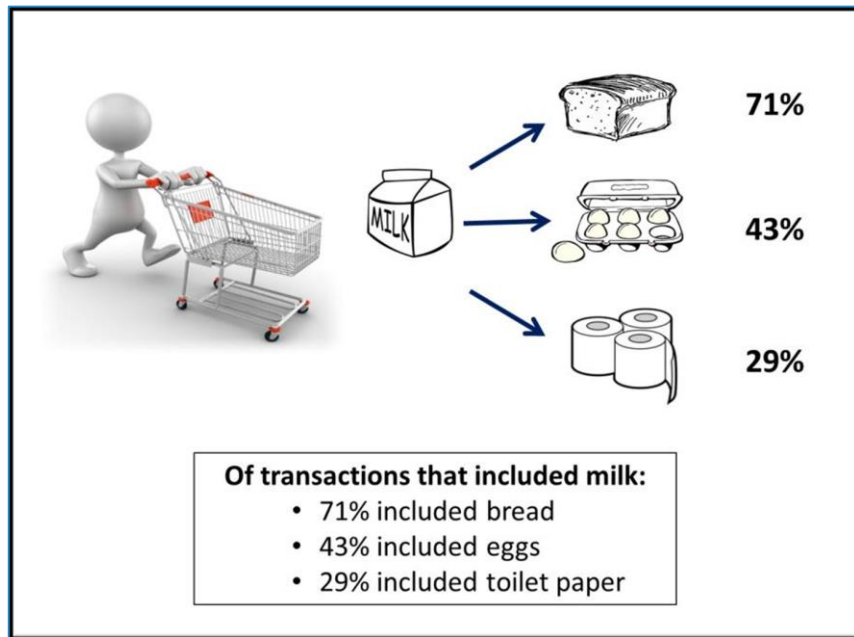


# Clustering: Examples

- To find groups of documents that are similar to each other based on the important terms appearing in them



# Association Analysis



Your behavior is  
being predicted,  
not by studying  
you, but by  
**studying others.**

# Association Rule Discovery

- Given a set of records each of which contain some number of items from a given collection:
  - Produce dependency rules which will predict the occurrence of an item based on the occurrences of other items

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

**{Milk} --> {Coke}**

**{Diaper, Milk} --> {Beer}**

# Association Analysis: Supermarket Shelf Management

- Goal: To identify items that are bought together by a sufficient amount of customers
- Place the items close to each other on supermarket shelves



# Association analysis examples

## ■ Marketing and sales promotion:

- Users who buy item A usually also buy item B
- If users bought item A, suggest item B or even offer discount on item B

## ■ Inventory management:

- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with the right parts to reduce the number of visits to consumer households

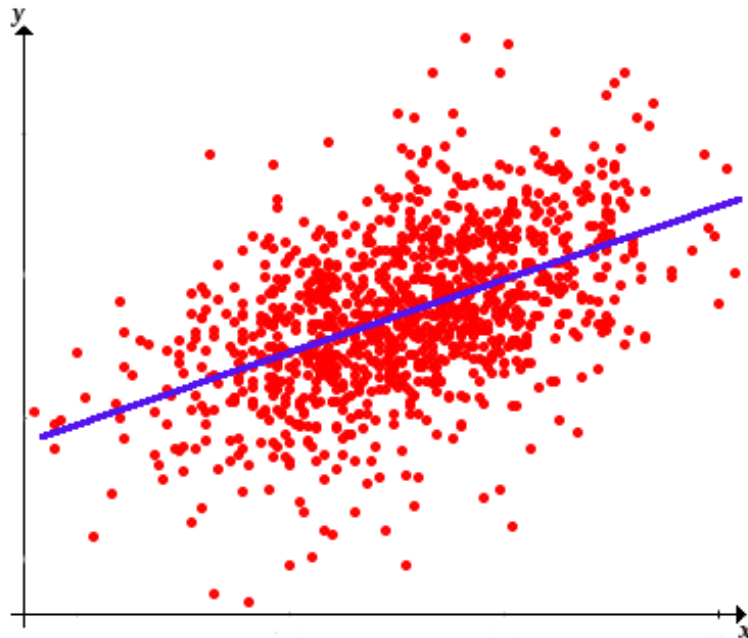
# Regression Example: Predict Housing Prices





# Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency



# Regression: Ad Clicks

Predict the probability of whether or not an ad will be clicked

The image shows a screenshot of a Google search results page for the query "dog food". The search bar at the top contains "dog food" and the Google logo is on the left. A red box labeled "Paid Search Results" with arrows points to two sponsored links. The first sponsored link is from PETCO, titled "PETCO" and "Get Supplies For All Types of Pets Shop In Store Today", with a URL "www.PETCO.com". The second sponsored link is from DollarTree, titled "Need Cheap Dog Food?" and "Everything under \$1 at Dollar Tree. Buy Online in Bulk", with a URL "www.DollarTree.com". Below the sponsored links are organic search results, including "Dog Food" from Eukanuba.com, "Pet Food: Premium, Healthy Dog Food and Cat Food for Pets | Purina.com", "Dog Food, Nutrition, and Breed Information | Eukanuba.com", "NUTRO® Premium Pet Food for Dogs and Cats", and "Shopping results for dog food" which lists products like "Beneful Adult Dog Food (7-lb bag)" and "Evolve Adult Maintenance Dog Food with Chicken".

Google

dog food

Search

Advanced Search  
Preferences

Web Show options...

Results 1 - 10 of about 281,000,000 for **dog**

**Paid Search Results**

Sponsored Links

**PETCO**  
www.PETCO.com Get Supplies For All Types of Pets Shop In Store Today  
Show map of 169 Parkway, Quincy, MA 02169

**Need Cheap Dog Food?**  
www.DollarTree.com Everything under \$1 at Dollar Tree. Buy Online in Bulk

**Dog Food**  
Eukanuba.com/Dog-Nutrition Natural FOS, Omega 6:3, and Dental Defense®

**Pet Food: Premium, Healthy Dog Food and Cat Food for Pets | Purina.com**  
Purina pet food is premium quality dog and cat food for happy, healthy pets. We use our knowledge and expertise to give you the pet food, tools and advice ...  
www.purina.com/ - Cached - Similar

**Dog Food, Nutrition, and Breed Information | Eukanuba.com**  
Find all you need to know about dog food, nutrition, and Breed information at Eukanuba.com.  
www.eukanuba.com/ - Cached - Similar

**NUTRO® Premium Pet Food for Dogs and Cats**  
Natural Super-Premium Dog & Cat Food. In 1926 we at Nutro Products dedicated ourselves to providing premium pet food to dogs and cats. ...  
www.nutroproducts.com/ - Cached - Similar

**Shopping results for dog food**

Beneful Adult Dog Food (7-lb bag)	\$13.39 - National Pet Pharmacy
Evolve Adult Maintenance Dog Food with Chicken	\$21.55 - GregRobert Pet Supplies
Eukanuba Natural Lamb & Rice Adult Dog Food 4 lb.	\$8.99 - Southern Agriculture

**lams PreBiotic Dog Food**  
Improve your pet's diet with lams new PreBiotic dog food. Learn more  
www.lams.com/PreBiotic-Dog-Food

**Compare Dog Foods**  
Are You Feeding Them Meat or Meat By-Products. Compare Brands Today  
BlueBuff.com

**Pure All-Natural Dog Food**  
Free range meat & organic veggies. Buy direct at wholesale prices.  
www.DarwinsNaturalPet.com/Organic

**Dog Food**  
High-end food at low-end prices! Huge selection with low shipping.  
www.JbPet.com  
Massachusetts

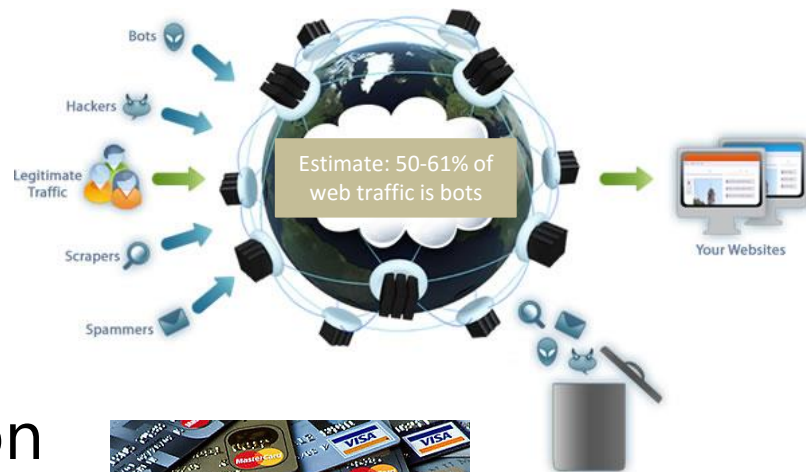
**Dog Food**  
Superior Nutrition for the Lifelong Health of Your Dog. Get Info Here  
www.Hillspet.com

**Order Natural Dog Food**  
Huge Selection, Top Trusted Brands. Your Dog Deserves The Best!

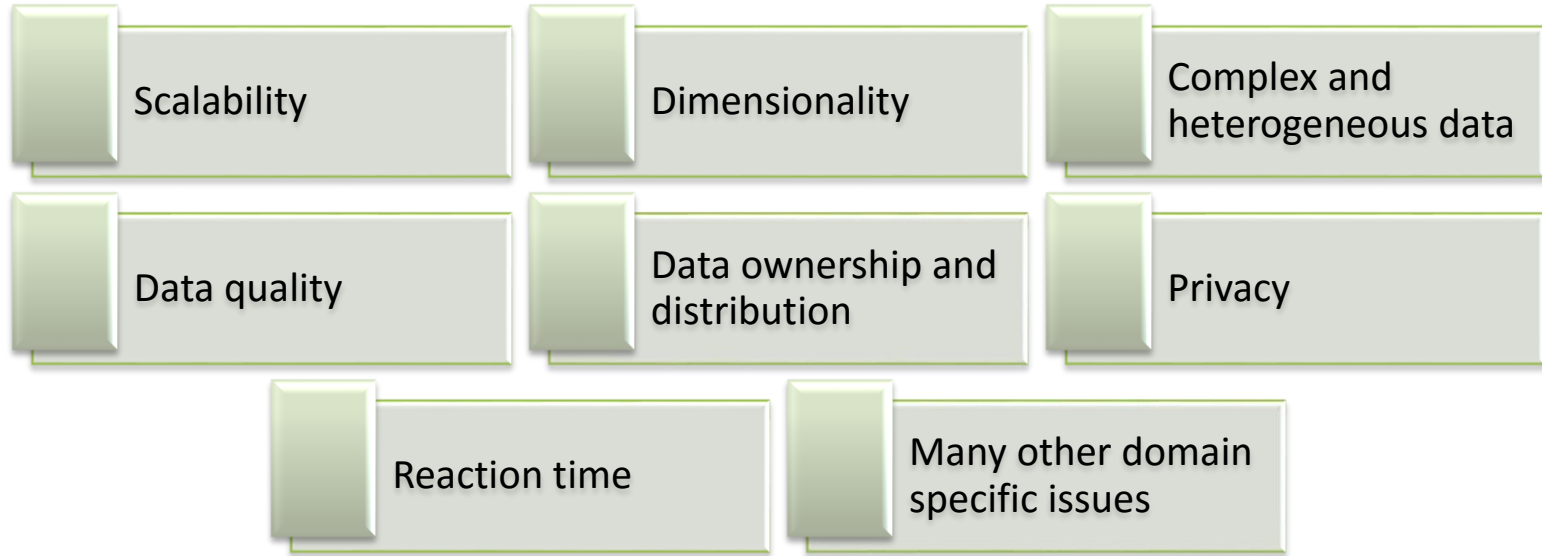


# Deviation/Anomaly Detection

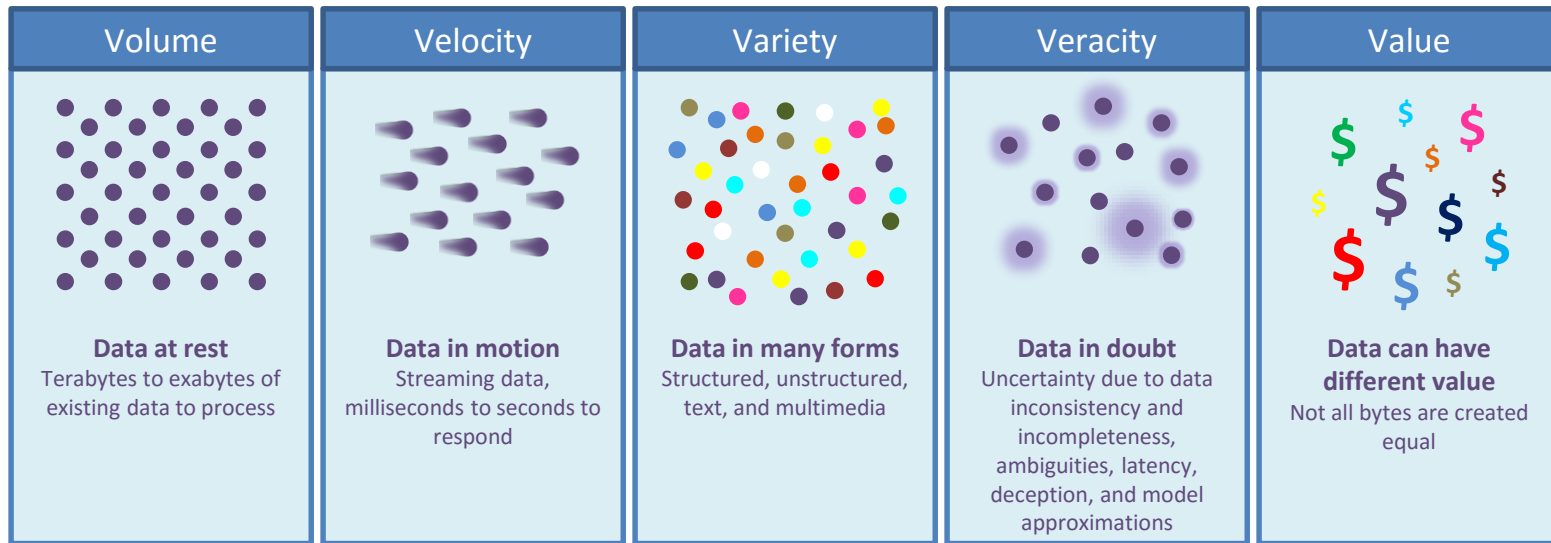
- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Bot detection in web traffic



# Challenges in Data Mining



# 5 V<sub>s</sub> Of Big Data



# Questions?