Introduction to Big Data, Predictive Analytics, and Data Science

Big Data and Data Science Everywhere

Web search and online ads

Insurance

Telcos

Online Education

Healthcare

Online Retail

Social Networks

Entertainment

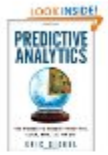
Big Data and Data Science Everywhere

AND many other places....

Online Shopping

Best Value

Buy **Predictive Analytics: The Pow… Predict 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

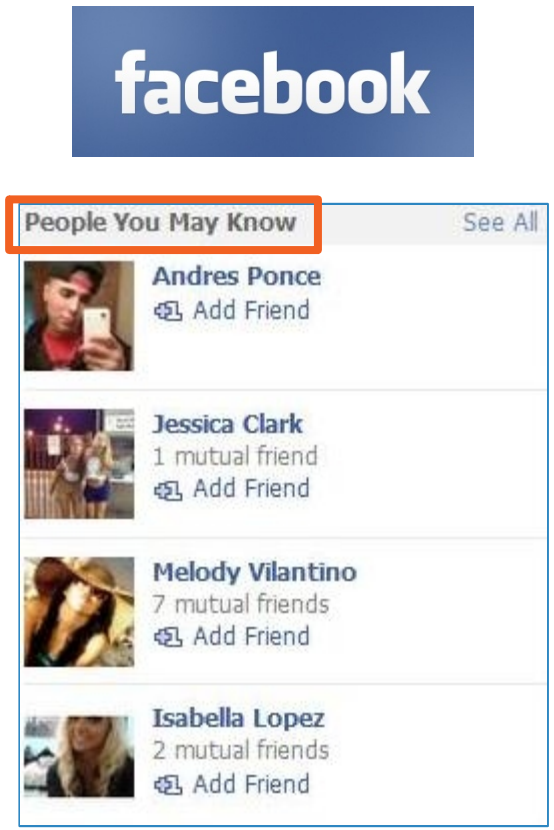
Big Data, Big Analytics: Emerging Business …
› Michael Minelli
(6)
Hardcover
$32.82

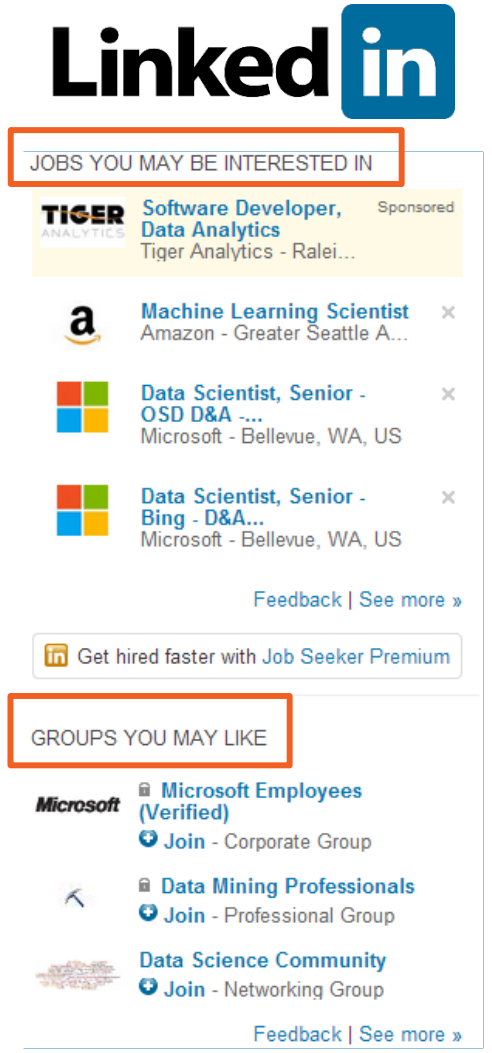
How to Measure Anything: Finding the Value of …
› Douglas W. Hubbard
(56)
Hardcover
$31.96

Secrets of Analytical Leaders: Insights …
Wayne Eckerson
(10)
Perfect Paperback
$44.96

Big Data Analytics: Disruptive …
Dr. Arvind Sathi
(5)
Paperback
$10.45

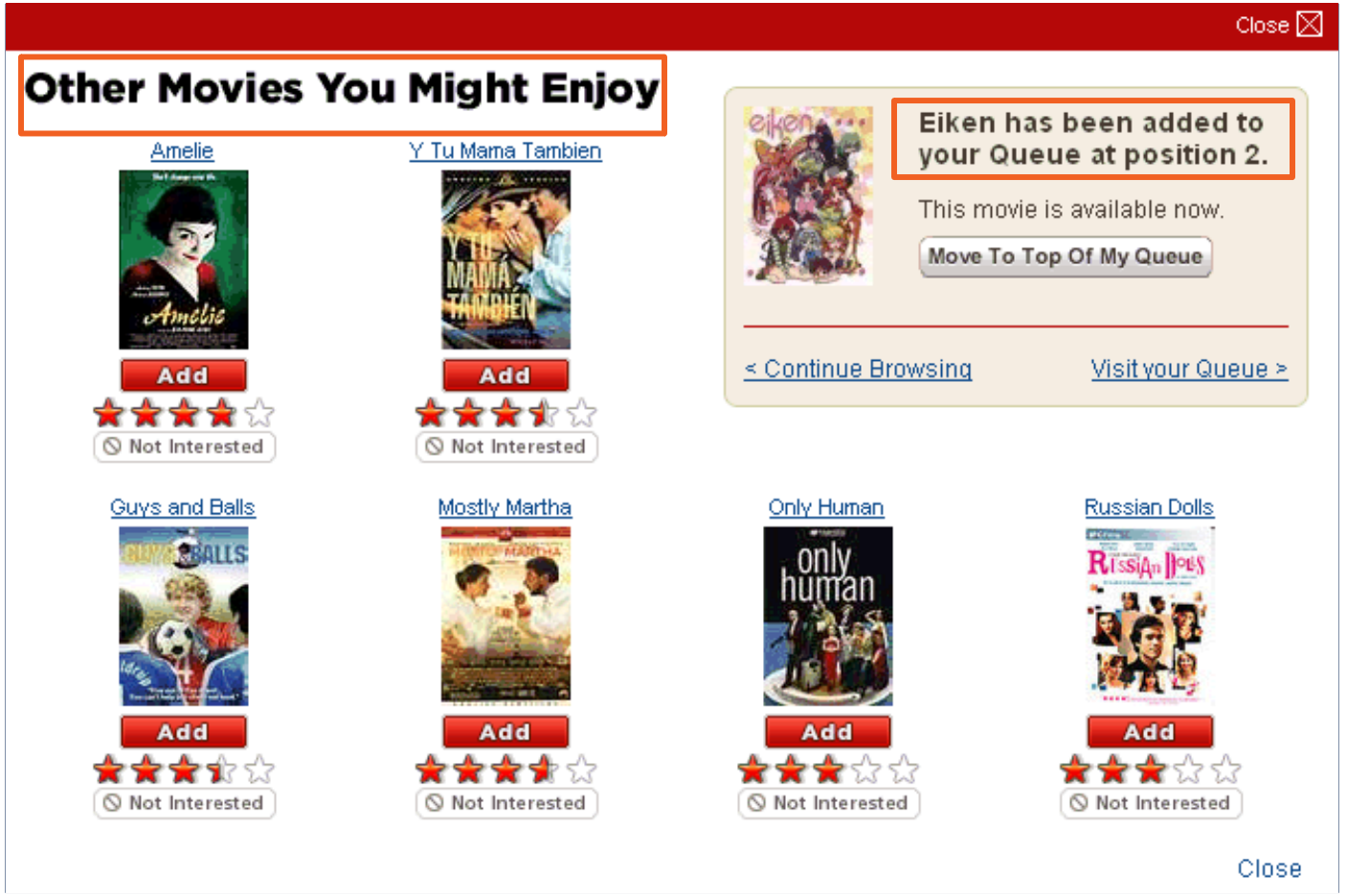
Social Networks



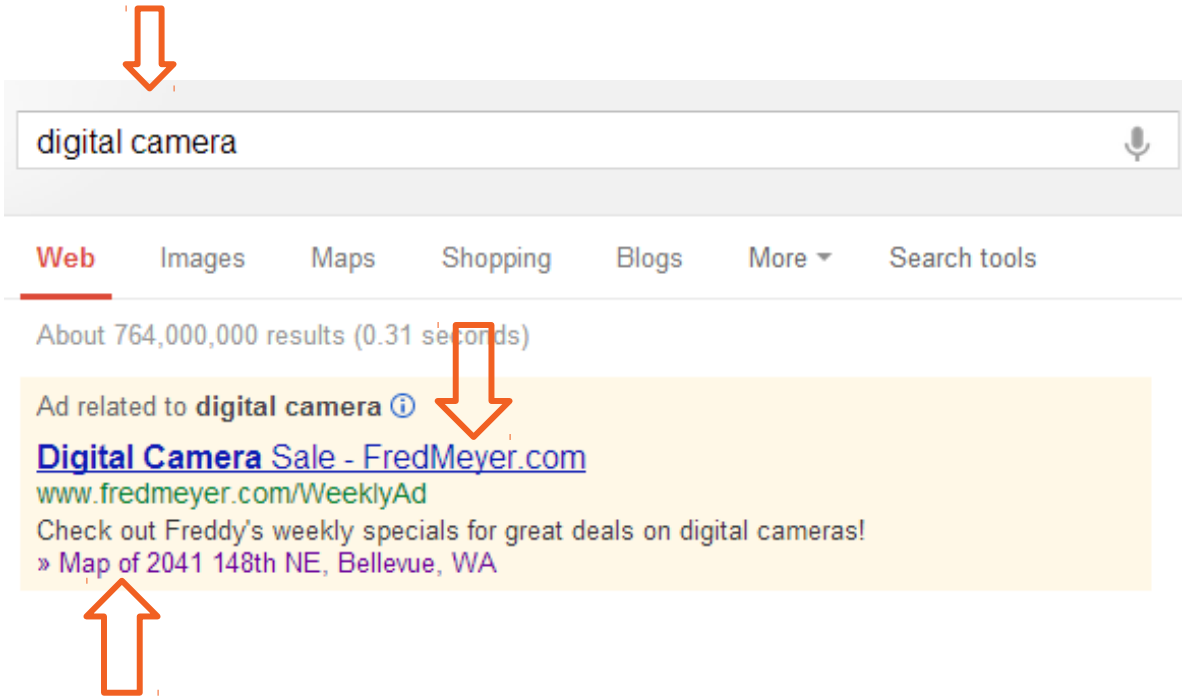


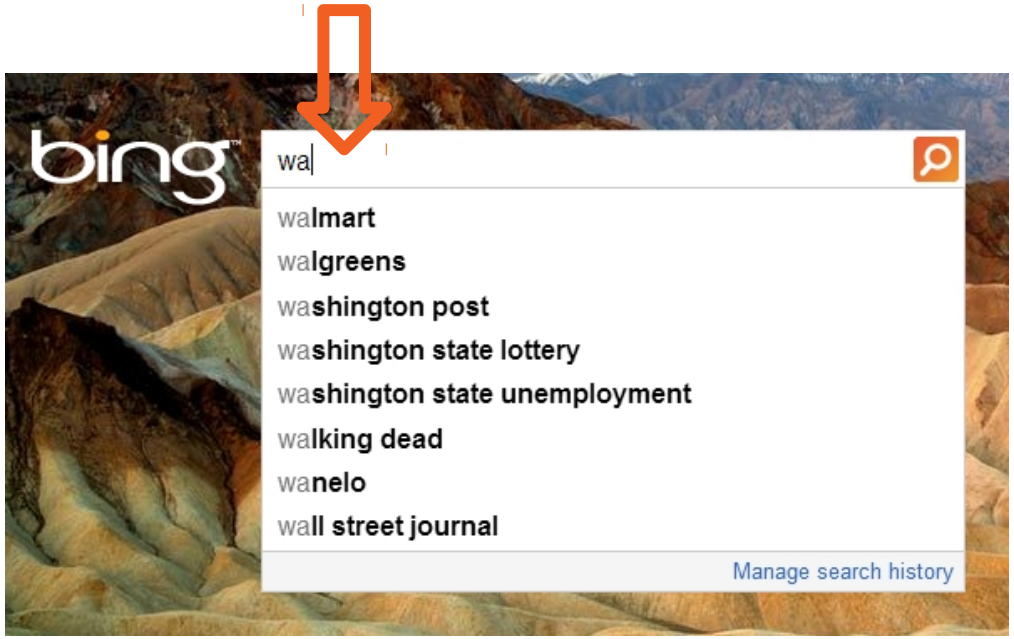
Online Entertainment

NETFLIX



Web Search


digital camera
Web
Images
Maps
Shopping
Blogs
More
Search tools
About 764,000,000 results (0.31 seconds)
Ad related to digital camera
Digital Camera Sale - FredMeyer.com
www.fredmeyer.com/WeeklyAd
Check out Freddy's weekly specials for great deals on digital cameras!
» Map of 2041 148th NE, Bellevue, WA

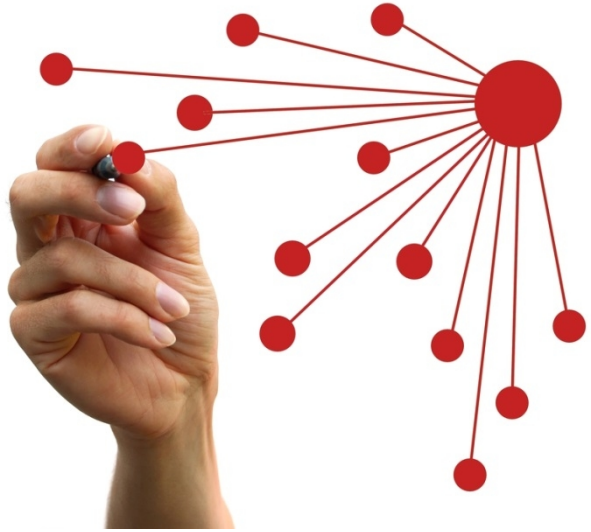

bing
wa
walmart
walgreens
washington post
washington state lottery
washington state unemployment
walking dead
wanelo
wall street journal
Manage search history

Brainstorming

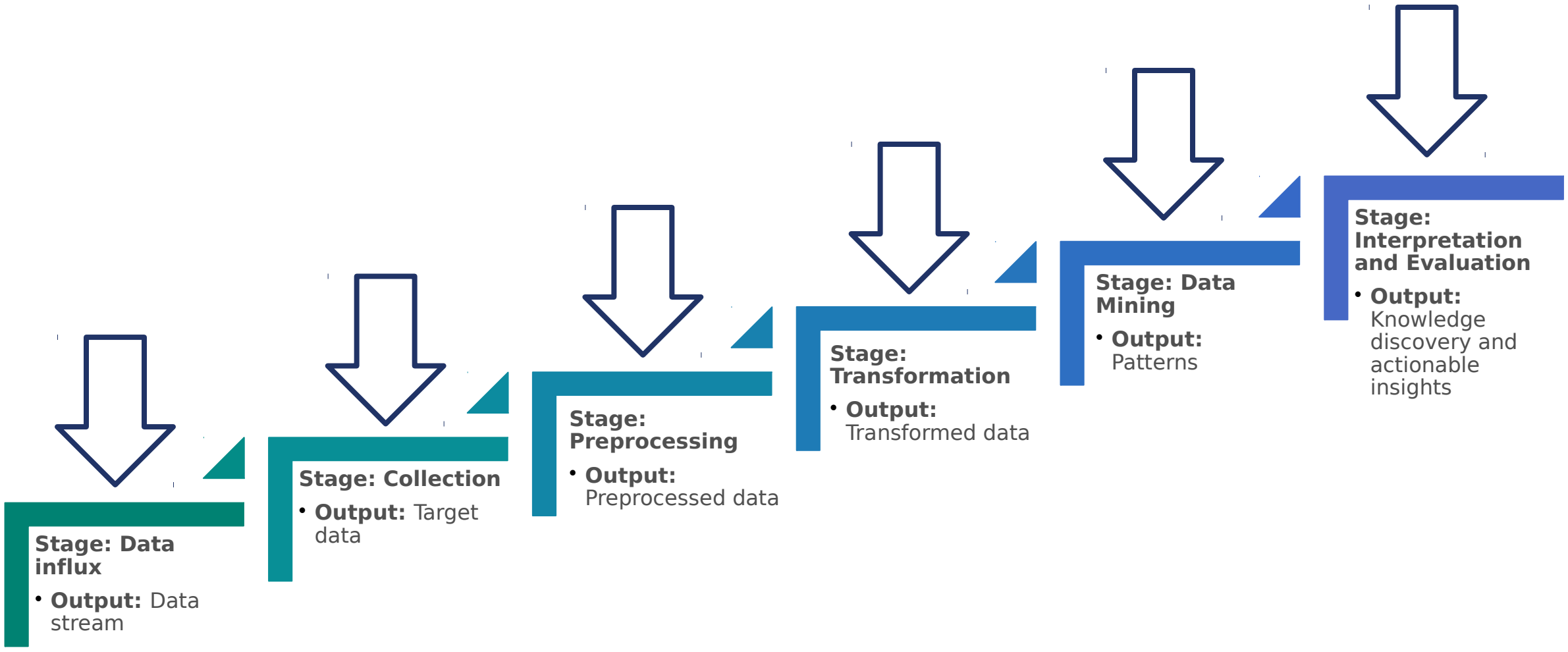
- What are some other applications?

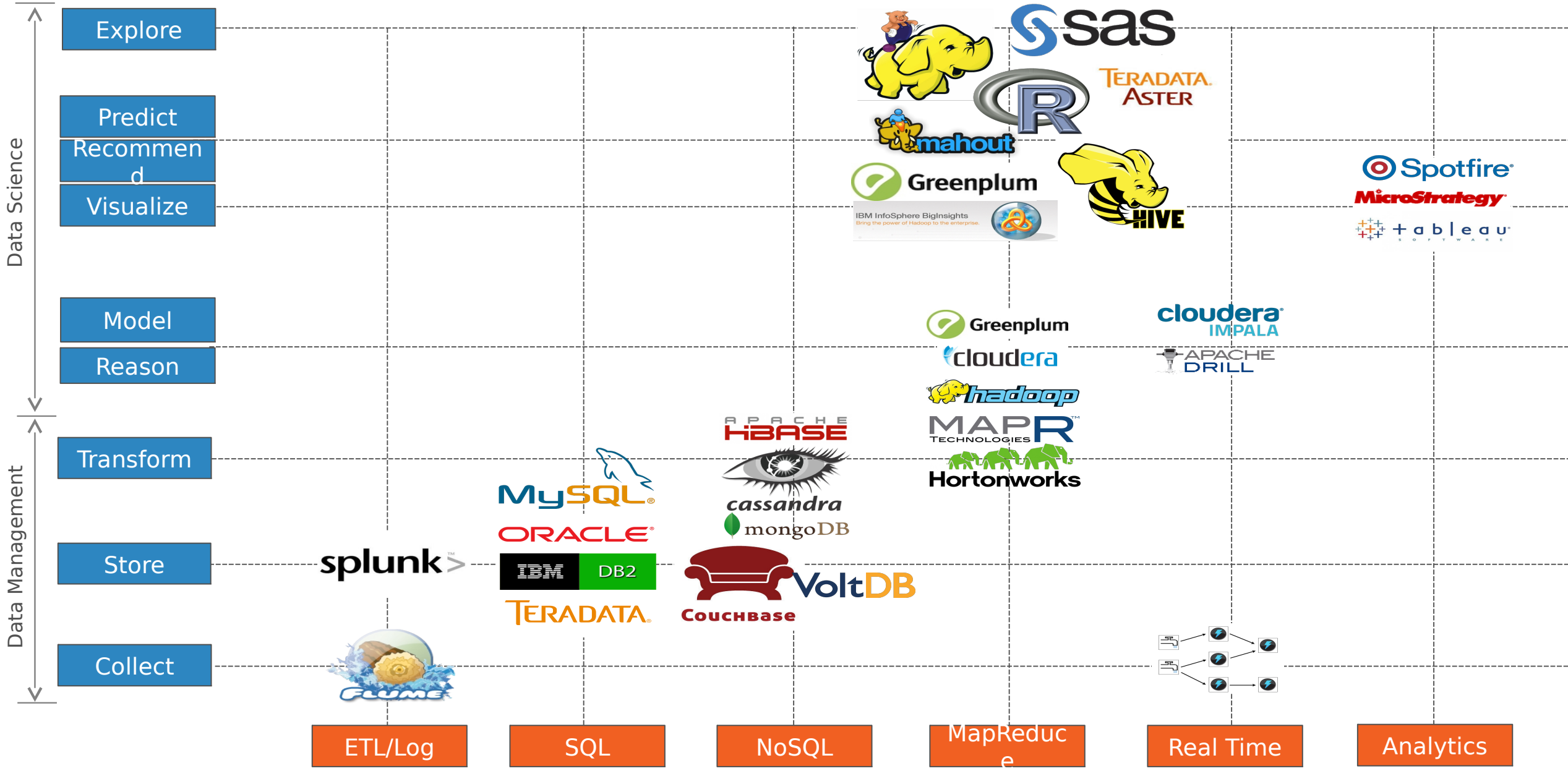
Connecting the Dots

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



Big Data Pipeline




Explore
Predict
Recommend
Visualize
Model
Reason
Transform
Store
Collect
Data Science
Data Management
sas
TERADATA ASTER
R
mahout
Greenplum
IBM InfoSphere BigInsights
Bring the power of Hadoop to the enterprise.
HIVE
Spotfire
MicroStrategy
tableau
Greenplum
cloudera
hadoop
MAPR TECHNOLOGIES
Hortonworks
cloudera IMPALA
APACHE DRILL
APACHE HBASE
cassandra
mongoDB
MySQL
ORACLE
IBM DB2
TERADATA
splunk>
VoltDB
CouchBase
FLUME
ETL/Log
SQL
NoSQL
MapReduce
Real Time
Analytics

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

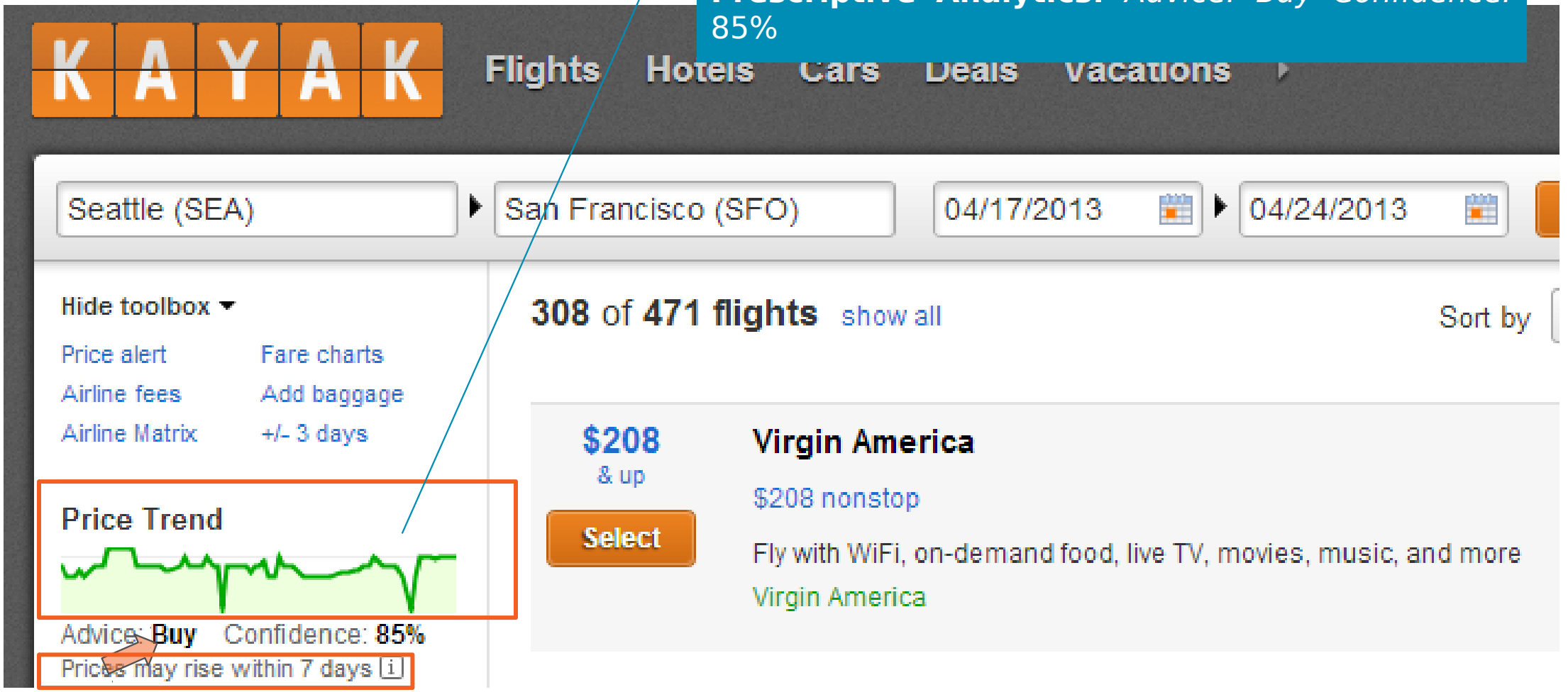
OR

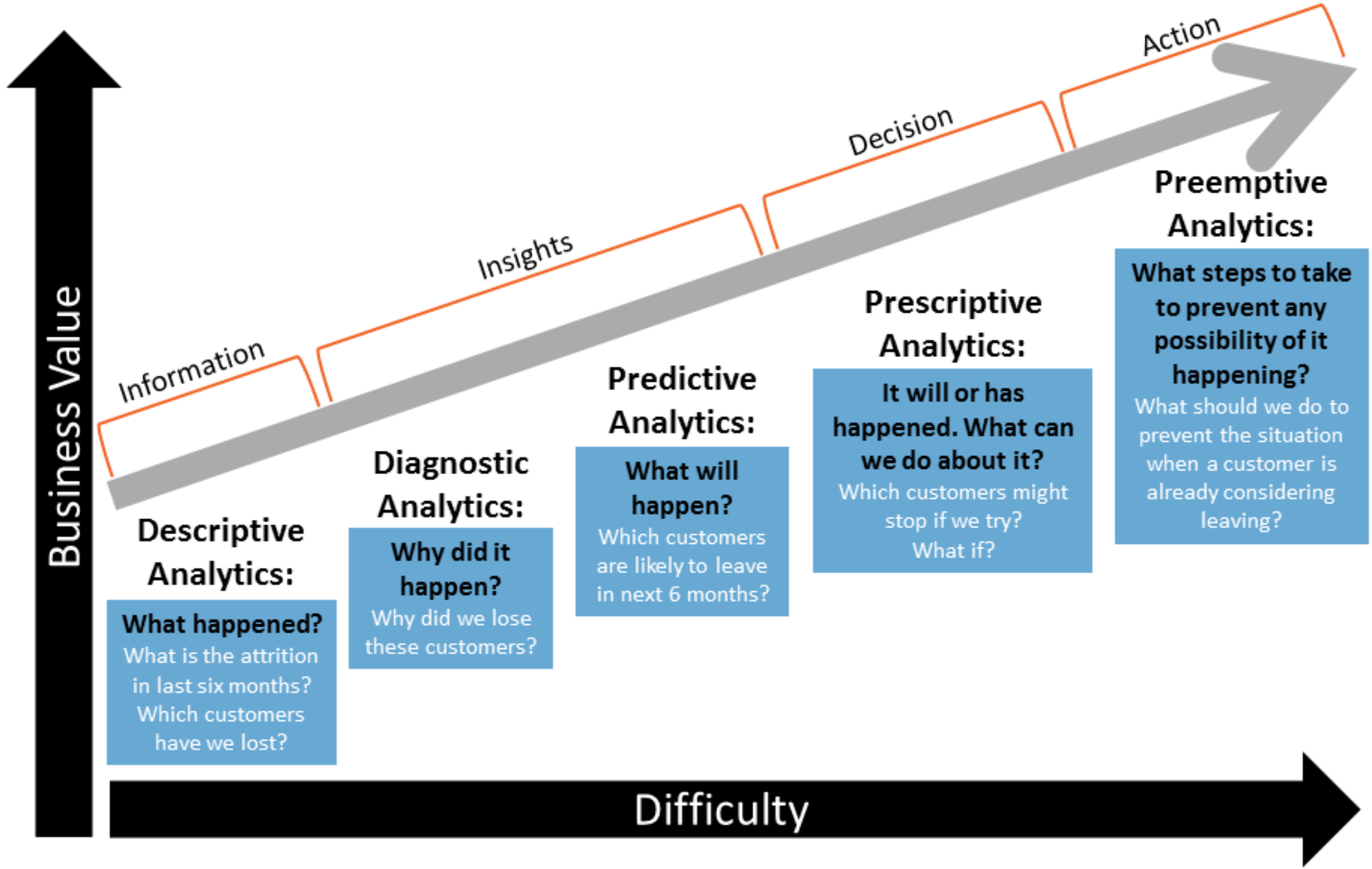
Traffic jam is about to happen in the next 30 minutes and you could possibly take the following courses of action:

- 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%

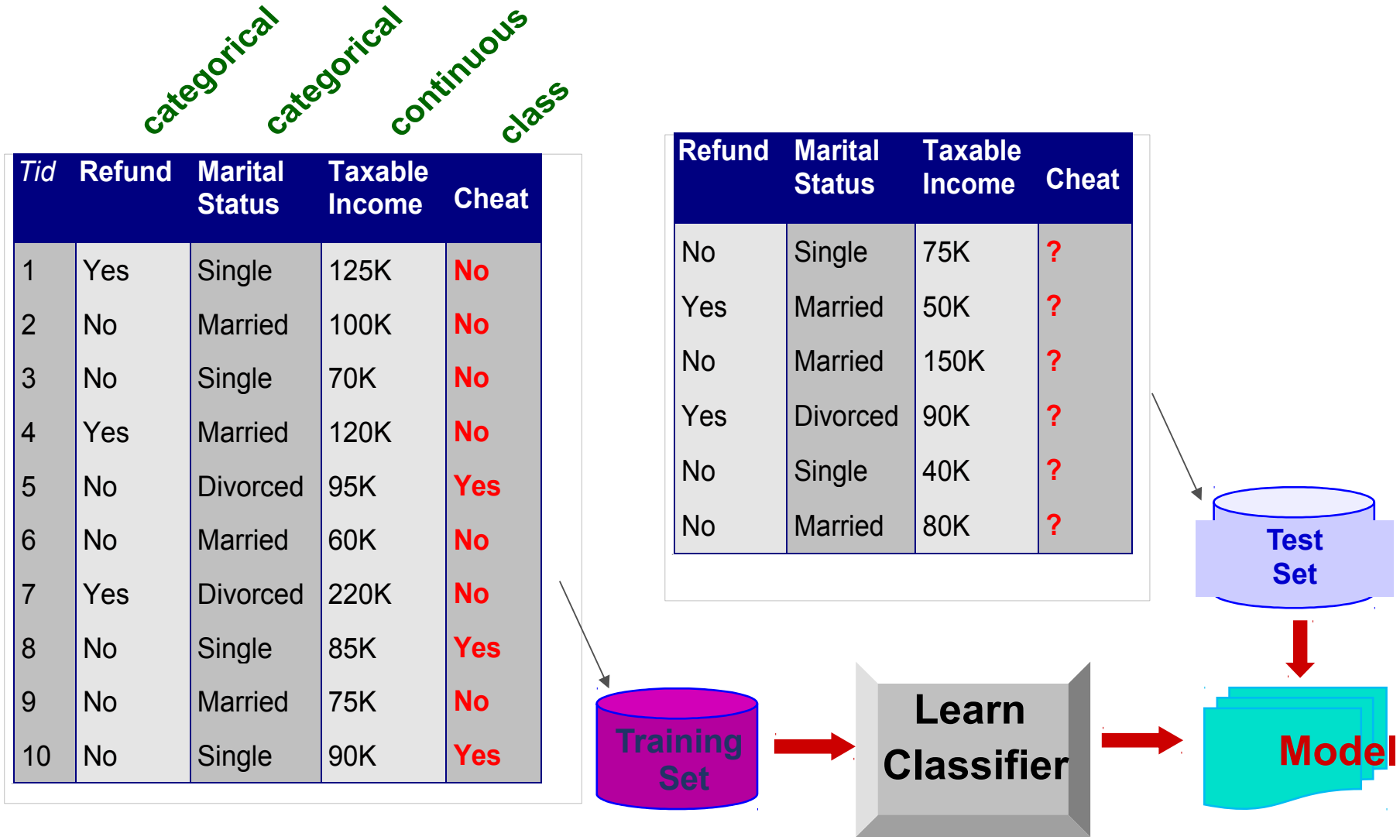



Business Value
Difficulty
Information
Insights
Decision
Action
Descriptive Analytics:
What happened?
What is the attrition in last six months? Which customers have we lost?
Diagnostic Analytics:
Why did it happen?
Why did we lose these customers?
Predictive Analytics:
What will happen?
Which customers are likely to leave in next 6 months?
Prescriptive Analytics:
It will or has happened. What can we do about it?
Which customers might stop if we try? What if?
Preemptive Analytics:
What steps to take to prevent any possibility of it happening?
What should we do to prevent the situation when a customer is already considering leaving?

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

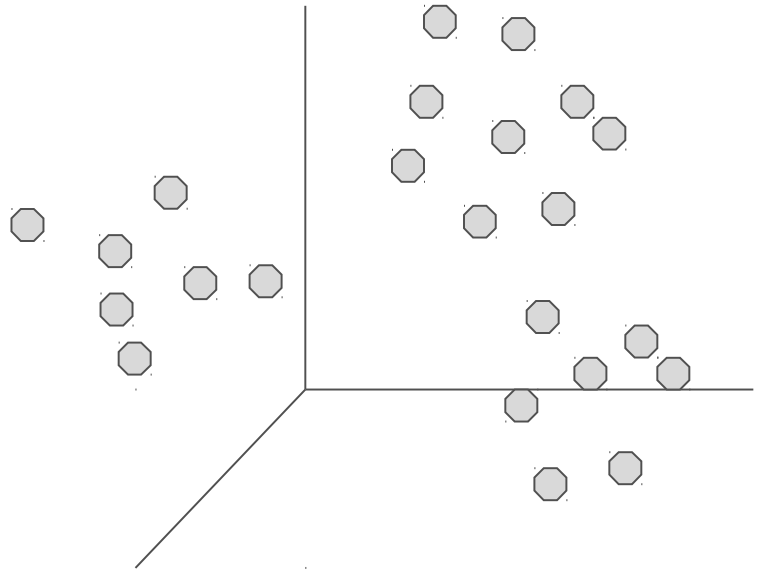
- Given a collection of records (training set)
  - Each record contains a set of *attributes*; one of the attributes is the *class label*.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as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

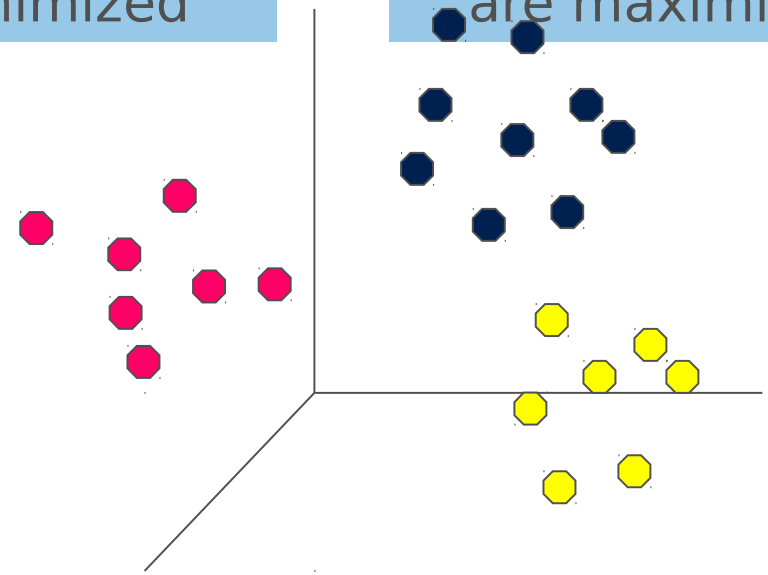
Clustering: An Illustration

Clustering in 3-D space using Euclidean distance



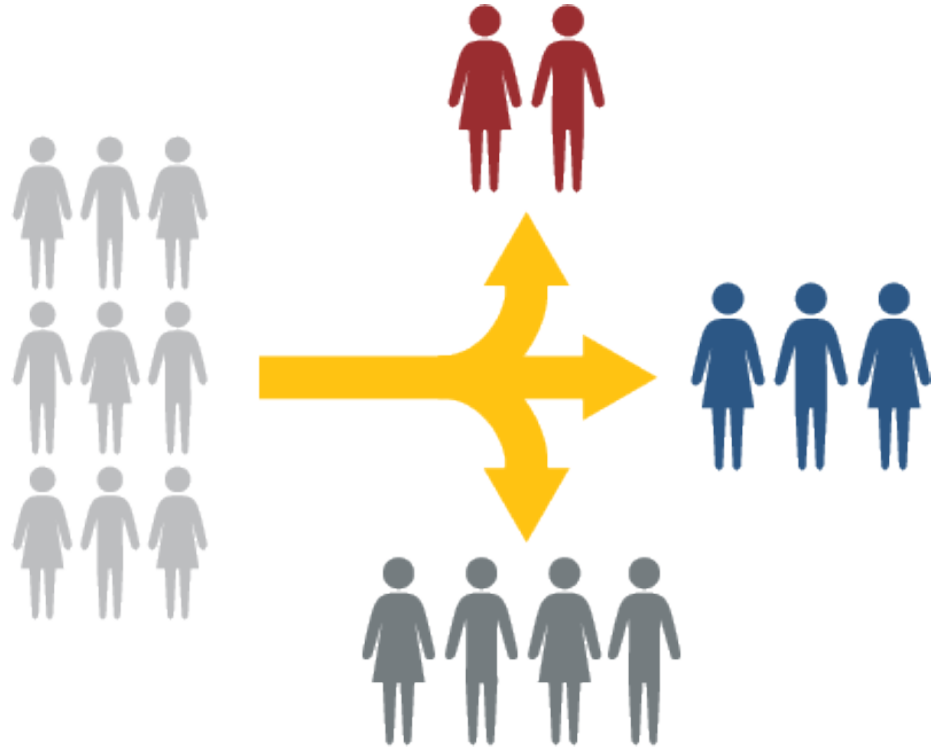
Intra-cluster distances are minimized

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

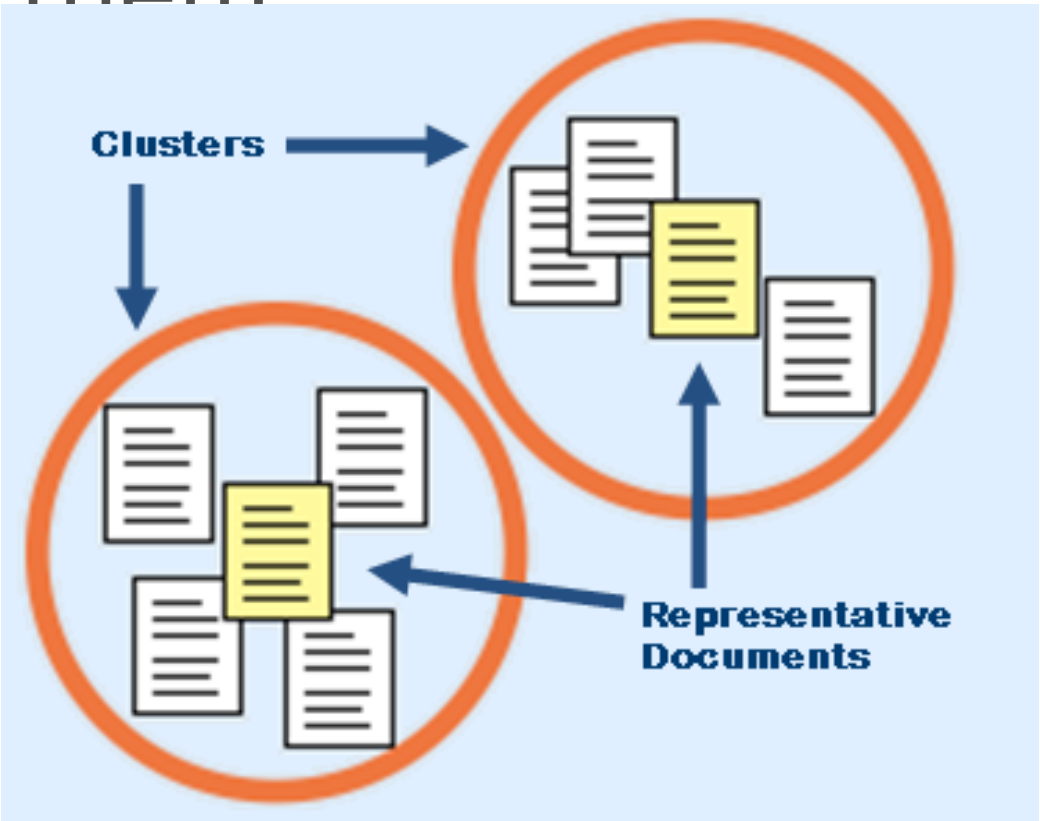
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that:
  - Data points within a cluster have more similarities with one another
  - Data points in different clusters have less similarities with one another

Clustering: Similarity Measures

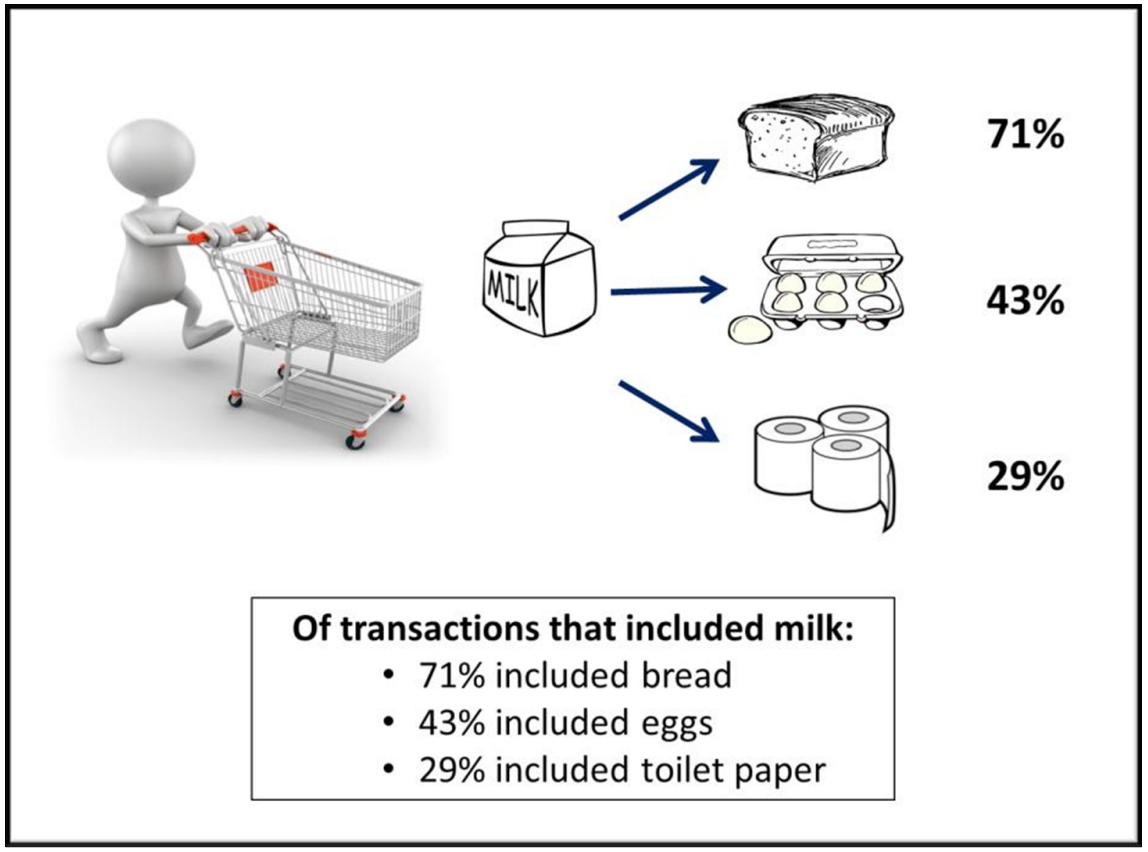
- Similarity Measures:
  - Euclidean Distance if attributes are continuous
  - Other problem-specific measures
  - Example: If a particular word occurs in two documents or not

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

TID	*Items*
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

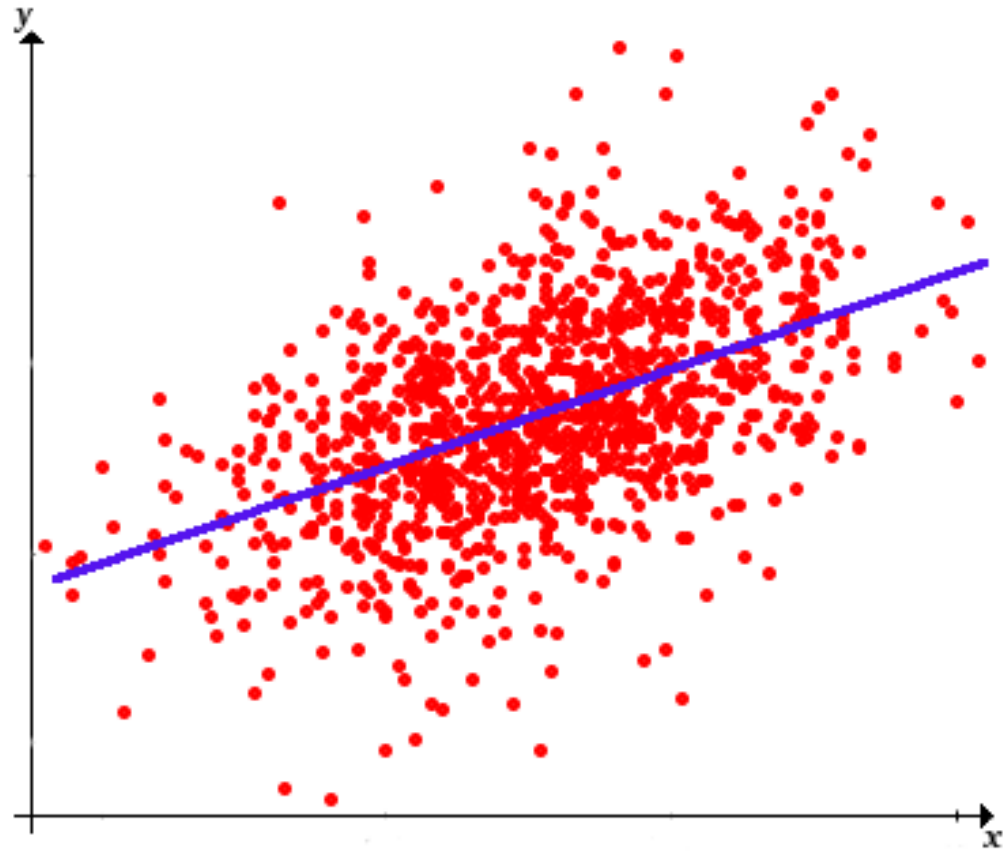
Association analysis examples

- 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

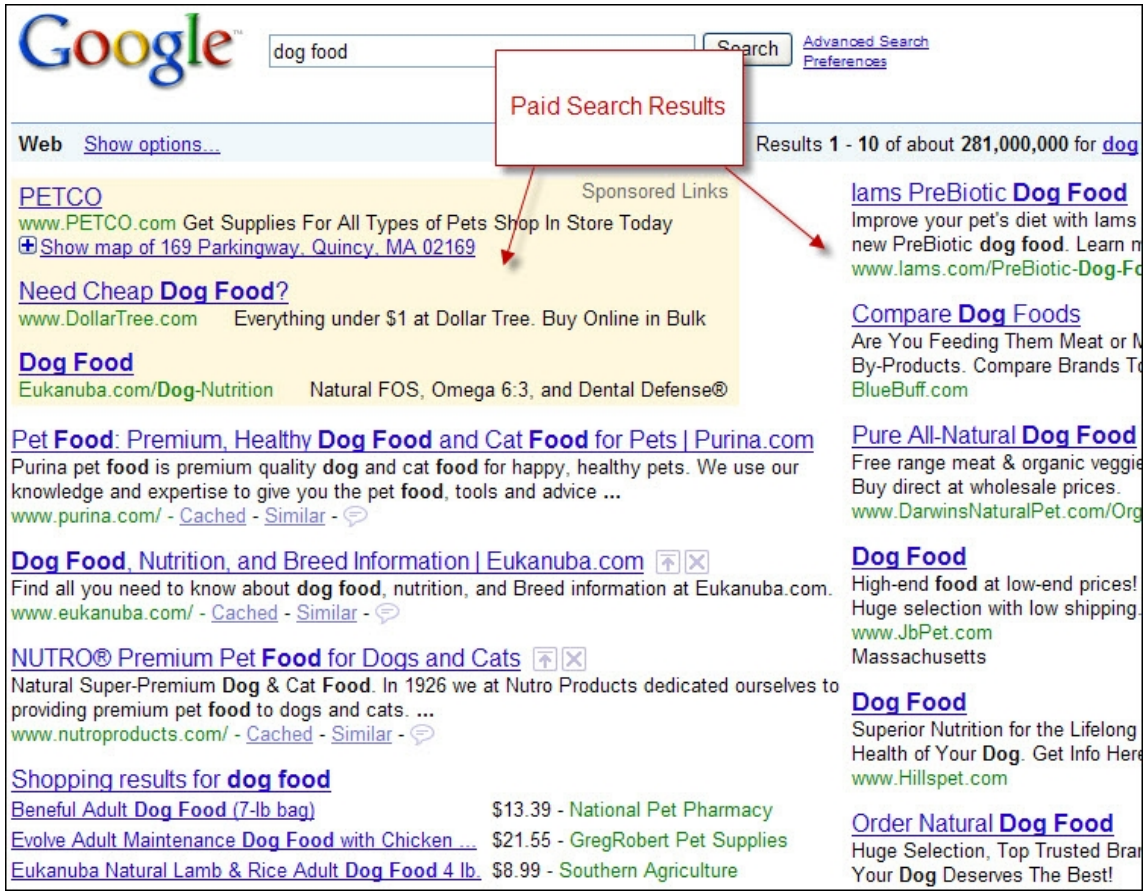
Regression

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



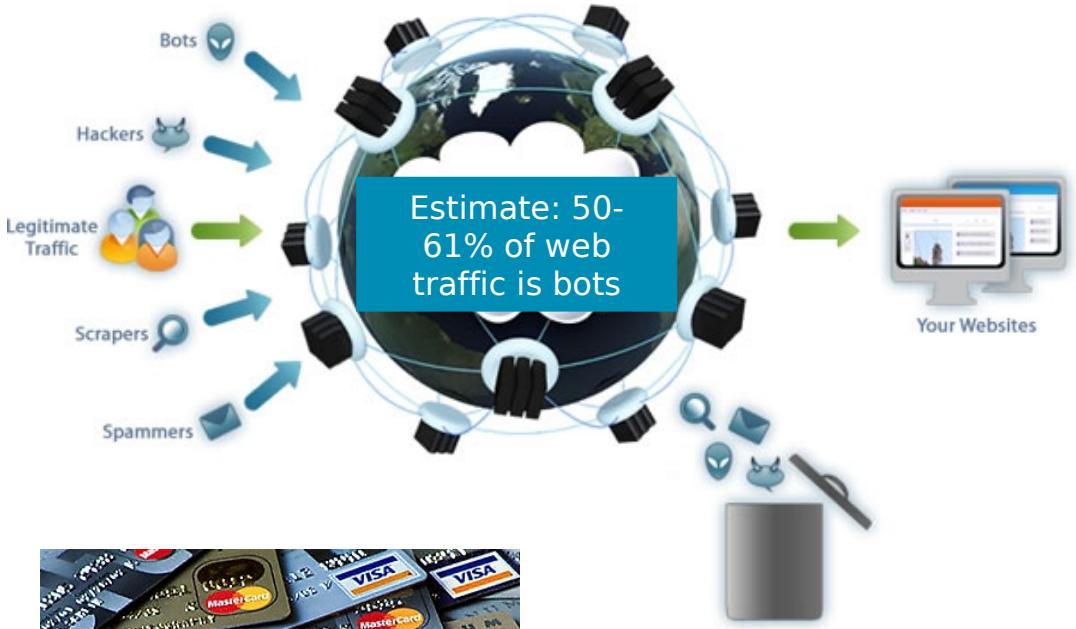
Regression: Ad Clicks

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



Deviation/Anomaly Detection

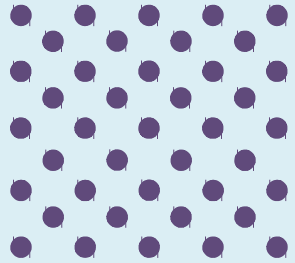
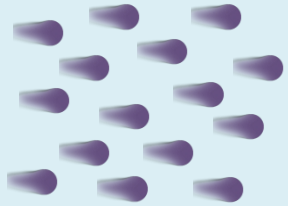
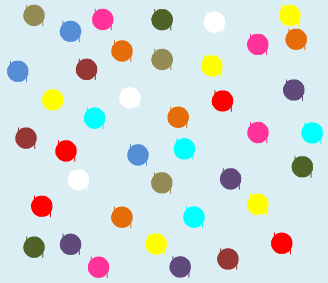
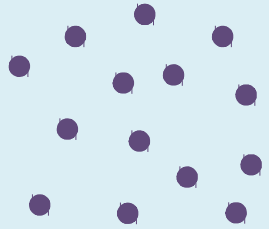
- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Bot detection in web traffic



Challenges in Data Mining

- Scalability
- Dimensionality
- Complex and heterogeneous data
- Data quality
- Data ownership and distribution
- Privacy
- Reaction time
- Many other domain specific issues

5 Vs Of Big Data

Volume	Velocity	Variety	Veracity	Value
				
Data at rest Terabytes to exabytes of existing data to process	**Data in motion** Streaming data, milliseconds to seconds to respond	**Data in many forms** Structured, unstructured, text, and multimedia	**Data in doubt** Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception, and model approximations	**Data can have different value** Not all bytes are created equal

Questions?