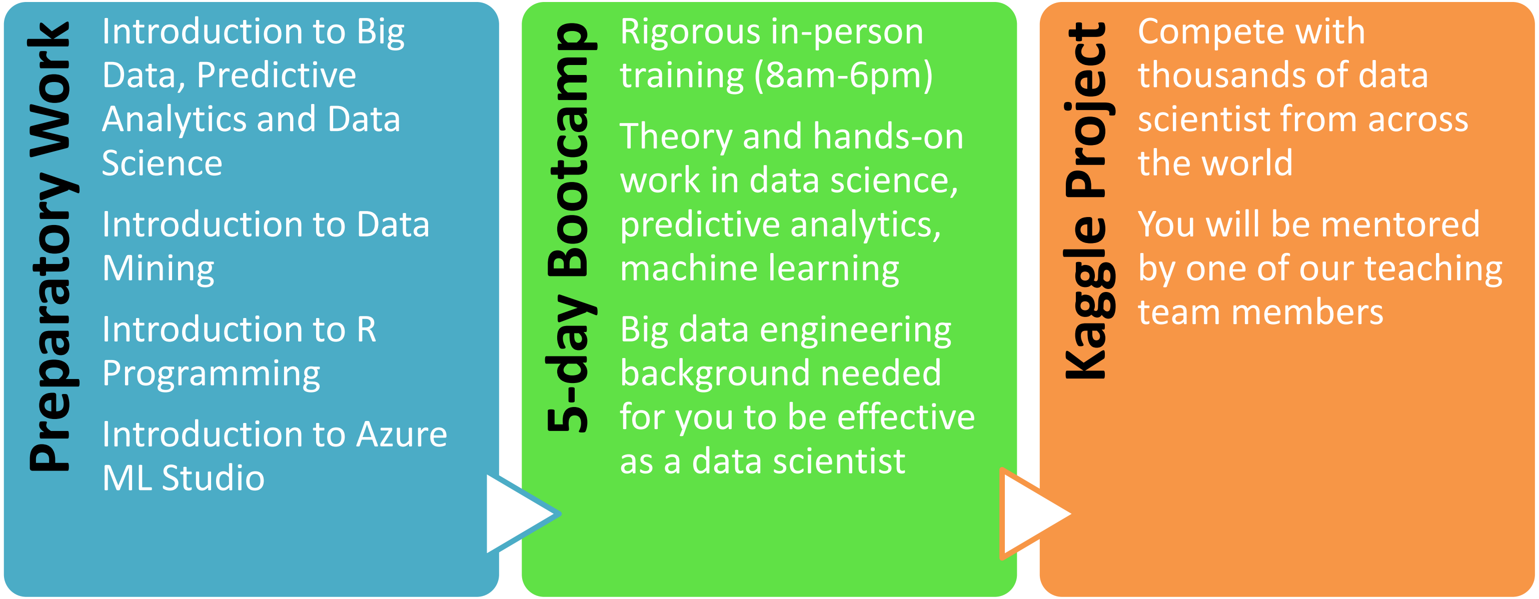
Data Science and Data Engineering

Curriculum Overview

datasciencedojo

unleash the data scientist in you

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

datasciencedojo

unleash the data scientist in you

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

datasciencedojo

unleash the data scientist in you

Big Data and Data Science Everywhere

Web search and online ads

Insurance

Telcos

Online Education

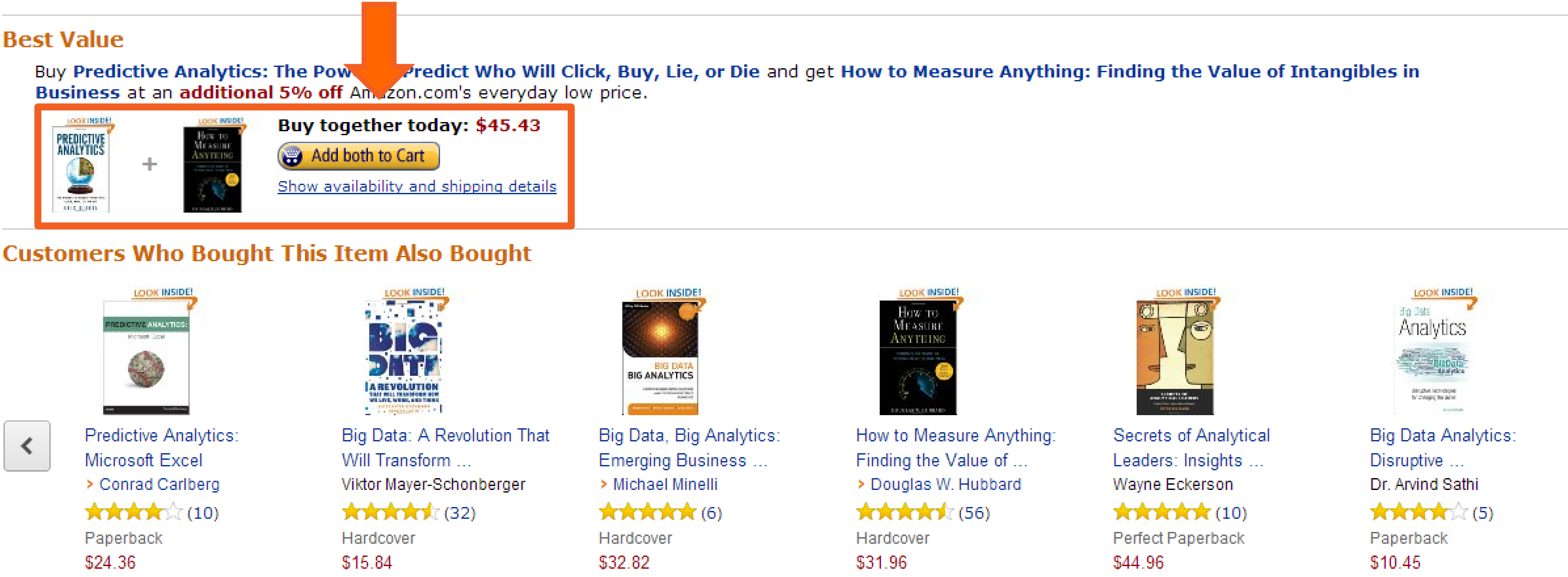
Online Retail

Social Networks

Entertainment

Healthcare

Online Shopping



Best Value

Buy **Predictive Analytics: The Power to 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

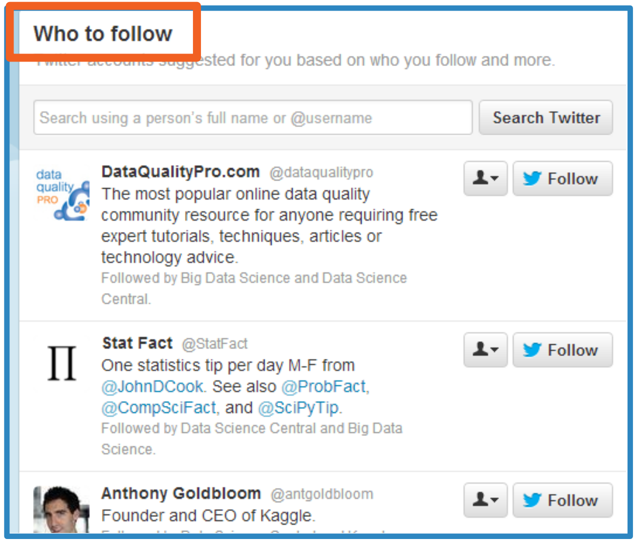
Show availability and shipping details

Customers Who Bought This Item Also Bought

Predictive Analytics: Microsoft Excel	Big Data: A Revolution That Will Transform ...	Big Data, Big Analytics: Emerging Business ...	How to Measure Anything: Finding the Value of ...	Secrets of Analytical Leaders: Insights ...	Big Data Analytics: Disruptive ...
› Conrad Carlberg	Viktor Mayer-Schonberger	› Michael Minelli	› Douglas W. Hubbard	Wayne Eckerson	Dr. Arvind Sathi
★★★★☆ (10)	★★★★½ (32)	★★★★★ (6)	★★★★½ (56)	★★★★★ (10)	★★★★☆ (5)
Paperback	Hardcover	Hardcover	Hardcover	Perfect Paperback	Paperback
$24.36	$15.84	$32.82	$31.96	$44.96	$10.45

Social Networks

twitter



Who to follow

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

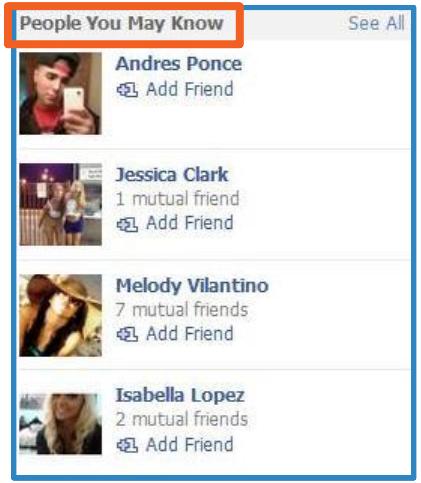
Search using a person's full name or @username | Search Twitter

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.

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.

Anthony Goldbloom @antgoldbloom
Founder and CEO of Kaggle.

facebook



People You May Know See All

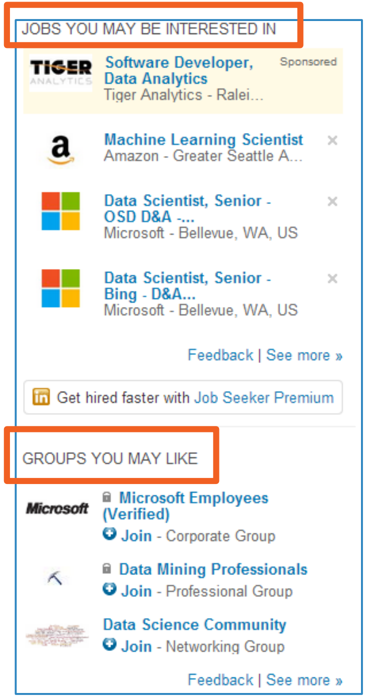
Andres Ponce
Add Friend

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

Software Developer, Data Analytics — Sponsored
Tiger Analytics - Ralei...

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 Employees (Verified)
Join - Corporate Group

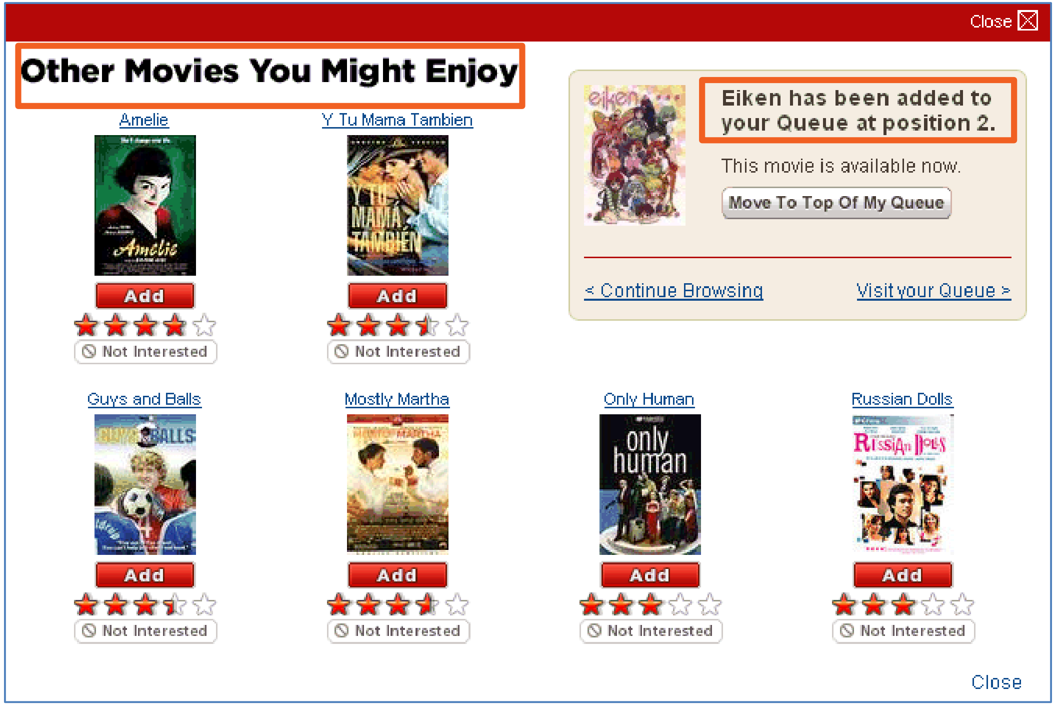
Data Mining Professionals
Join - Professional Group

Data Science Community
Join - Networking Group

Feedback | See more »

datasciencedojo
unleash the data scientist in you

Online Entertainment



Other Movies You Might Enjoy

Amelie

Y Tu Mama Tambien

Guys and Balls

Mostly Martha

Only Human

Russian Dolls

Eiken has been added to your Queue at position 2.

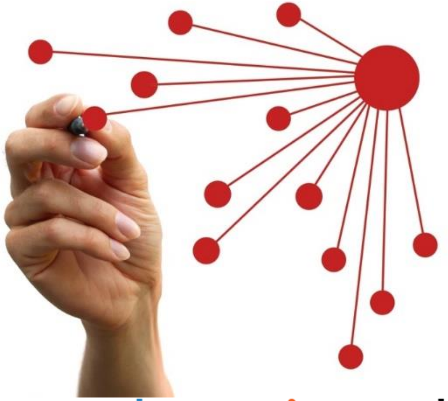
This movie is available now.

Brainstorming

- What are some other applications?

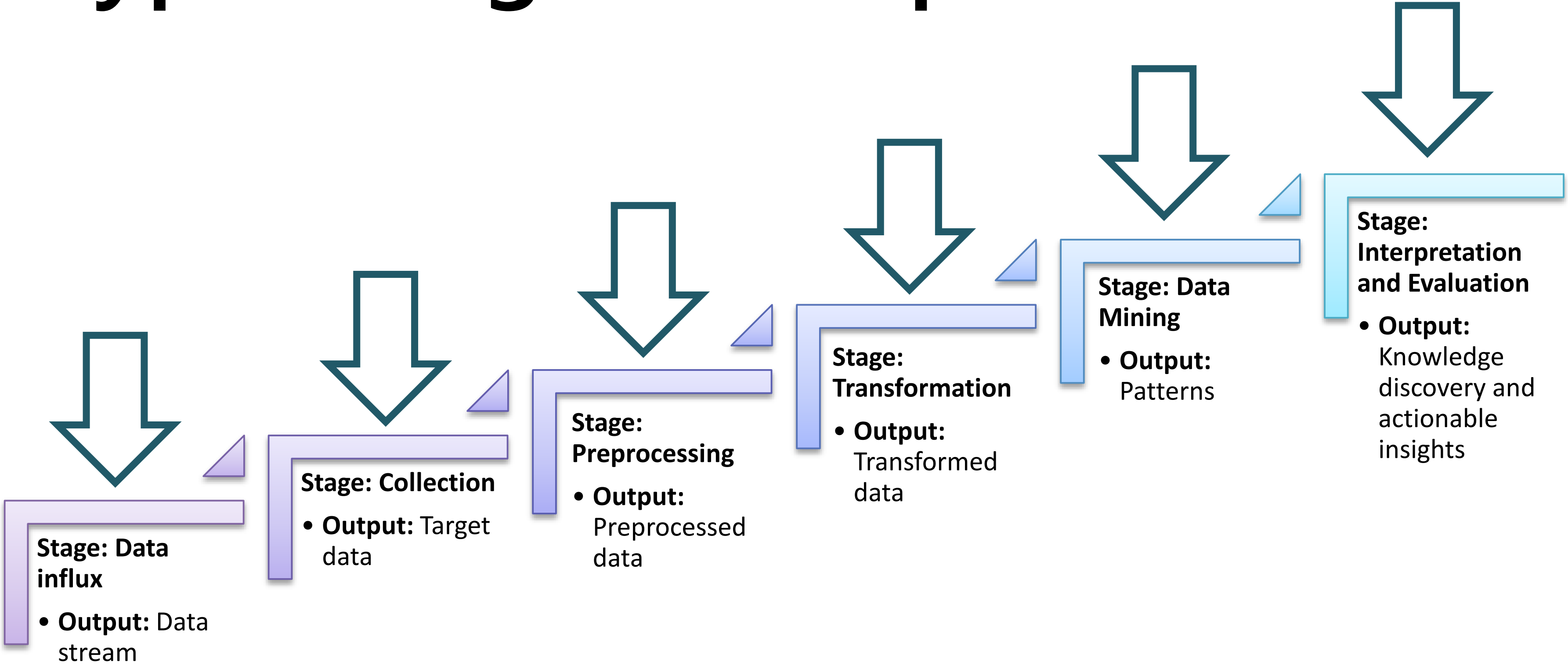
Connecting the Dots

- 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



Stage: Data influx

- **Output:** Data stream

Stage: Collection

- **Output:** Target data

Stage: Preprocessing

- **Output:** Preprocessed data

Stage: Transformation

- **Output:** Transformed data

Stage: Data Mining

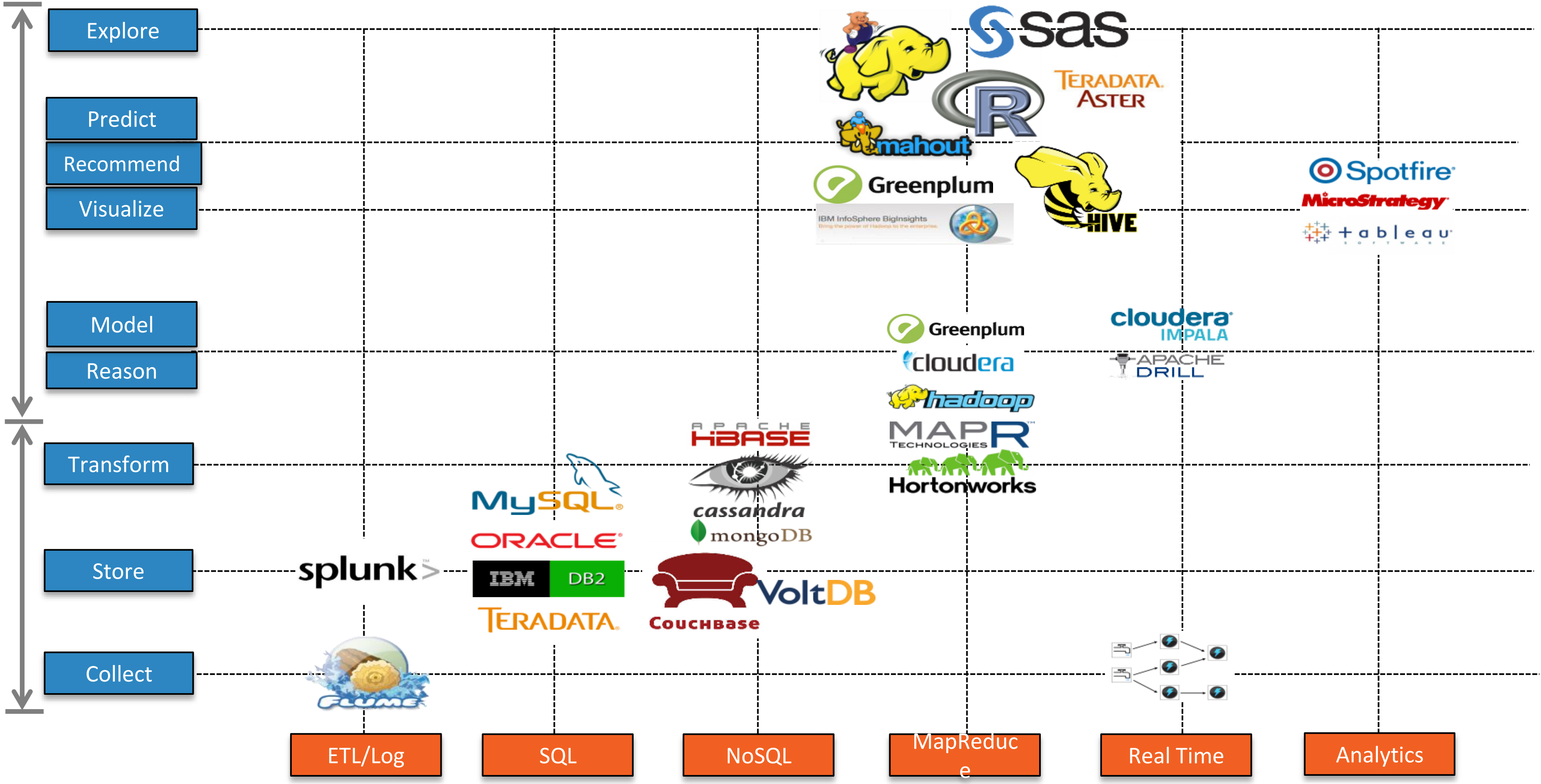
- **Output:** Patterns

Stage: Interpretation and Evaluation

- **Output:** Knowledge discovery and actionable insights

datasciencedojo

unleash the data scientist in you



Data Science

Explore

Predict

Recommend

Visualize

Model

Reason

Data Management

Transform

Store

Collect

ETL/Log

SQL

NoSQL

MapReduc
e

Real Time

Analytics

Big Data – Technology, Platforms & Products

datasciencedojo
unleash the data scientist in you

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

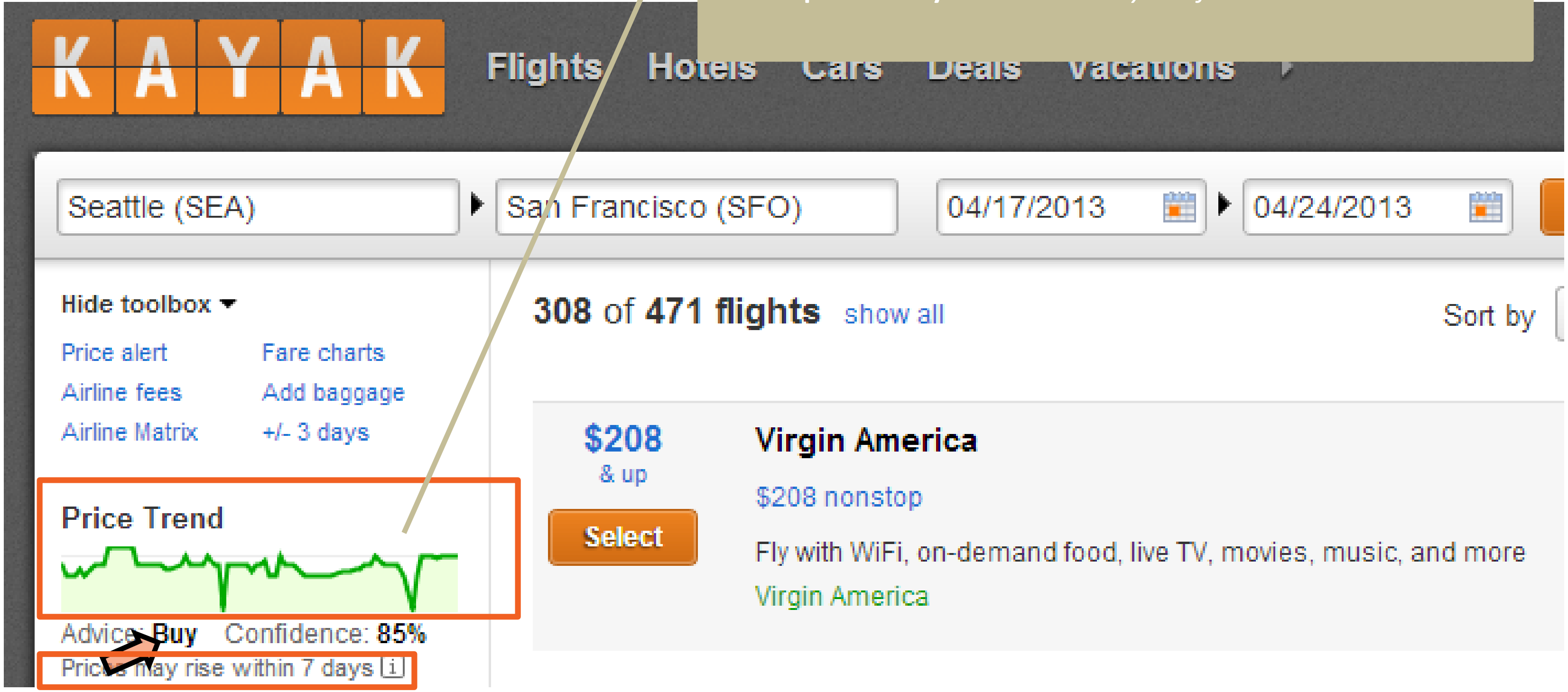
OR

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%



Seattle (SEA) ▸ San Francisco (SFO) 04/17/2013 ▸ 04/24/2013

308 of 471 flights show all

$208 & up

Virgin America

$208 nonstop

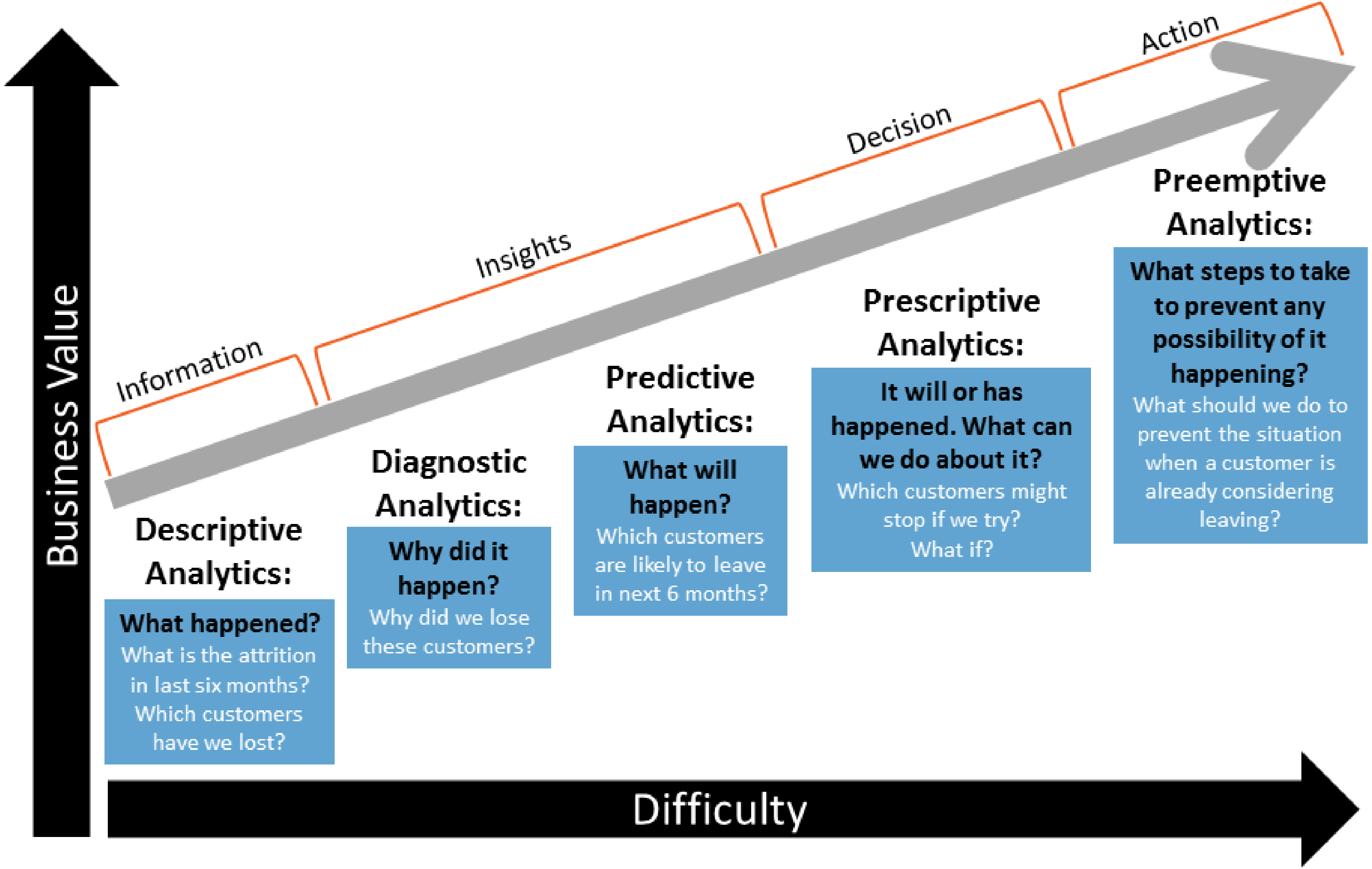
Fly with WiFi, on-demand food, live TV, movies, music, and more

Virgin America

Price Trend

Advice: Buy Confidence: 85%

Prices may rise within 7 days



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?

datasciencedojo
unleash the data scientist in you

Data Mining and Predictive Analytics

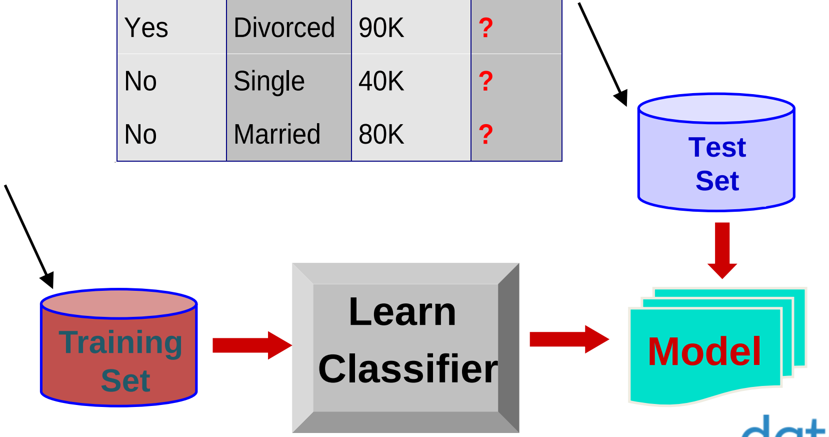
In the next few slides, we will take a look at some of the most common data mining tasks.

Classification: A Simple Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification

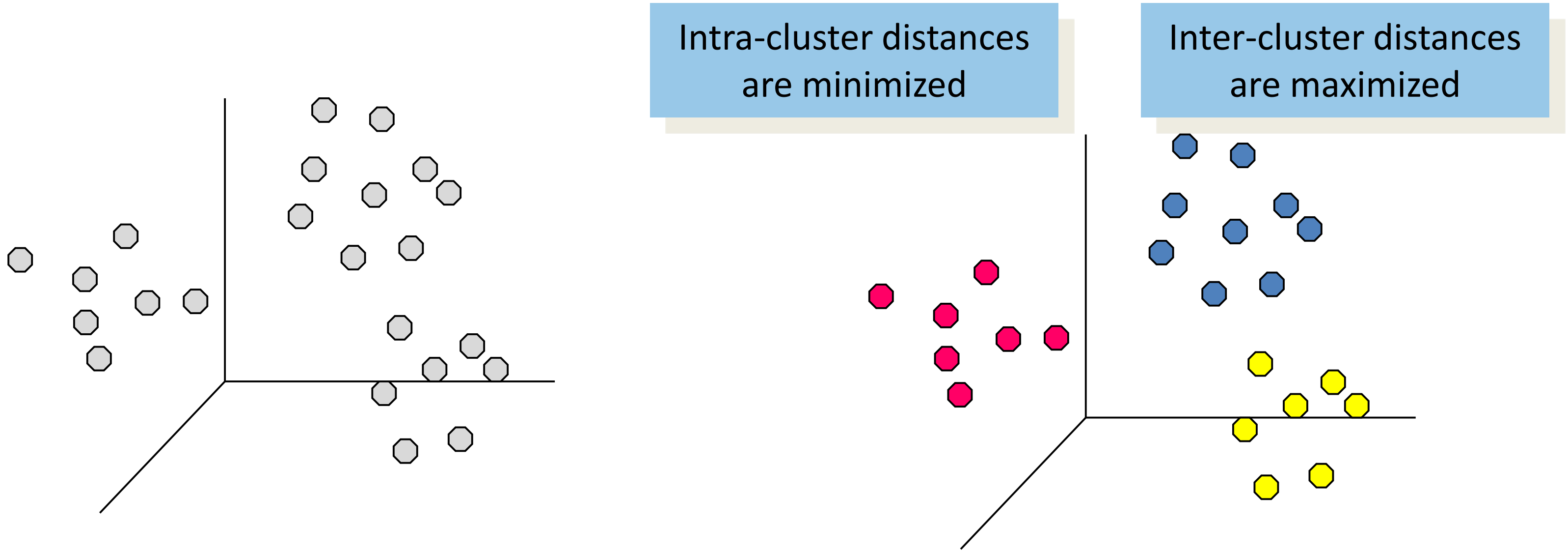
- Given a collection of records (training set)
  - Each record contains a set of *attributes*; one of the attributes is the *class label*.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as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

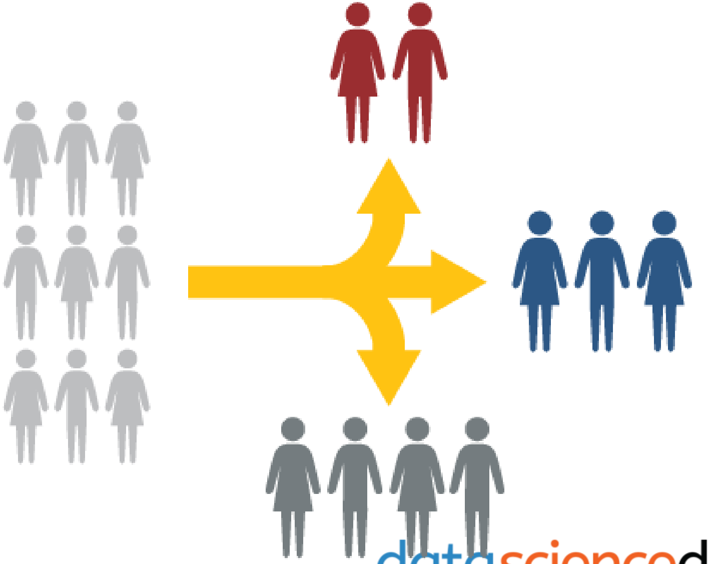


Intra-cluster distances are minimized

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



datasciencedojo
unleash the data scientist in you

Clustering

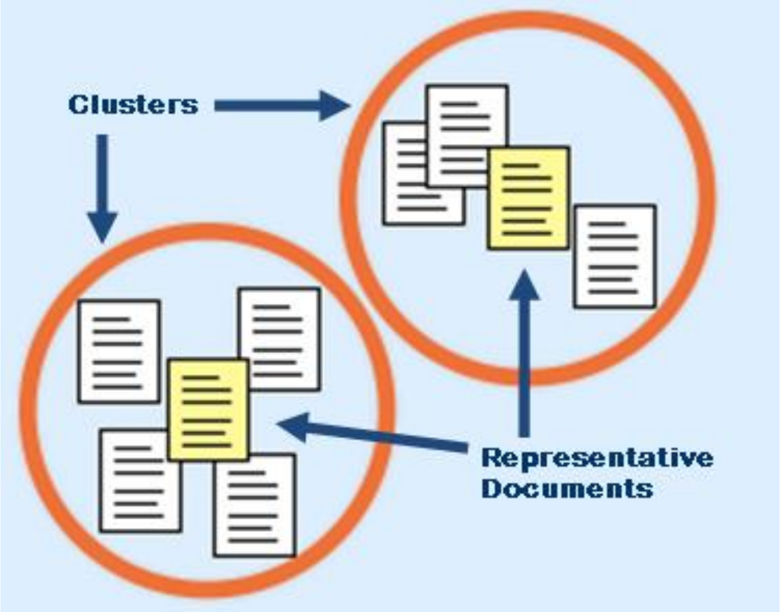
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that:
  - Data points within a cluster have more similarities with one another
  - Data points in different clusters have less similarities with one another

Clustering: Similarity Measures

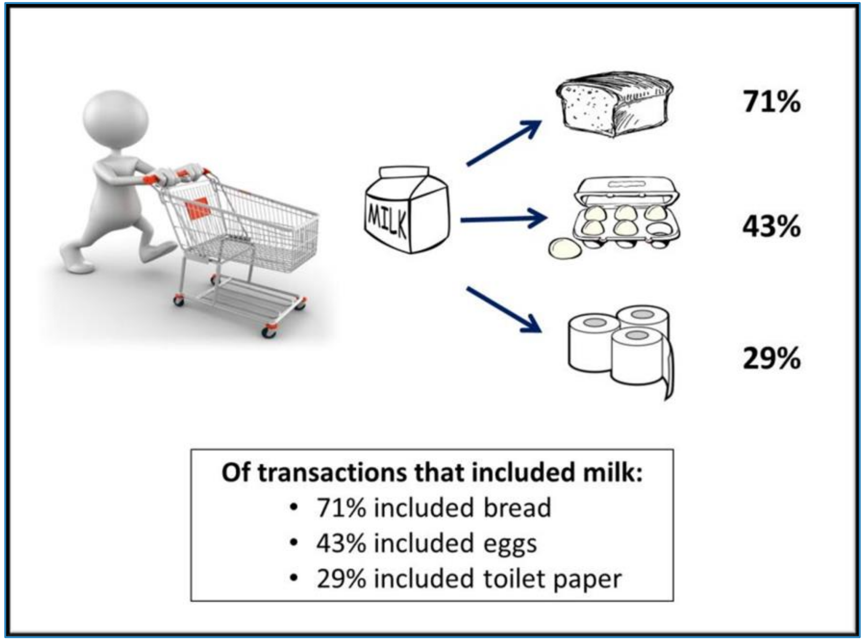
- Similarity Measures:
  - Euclidean Distance if attributes are continuous
  - Other problem-specific measures
  - Example: If a particular word occurs in two documents or not

Clustering: Examples

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



Association Analysis



Of transactions that included milk:

- 71% included bread
- 43% included eggs
- 29% included toilet paper

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}

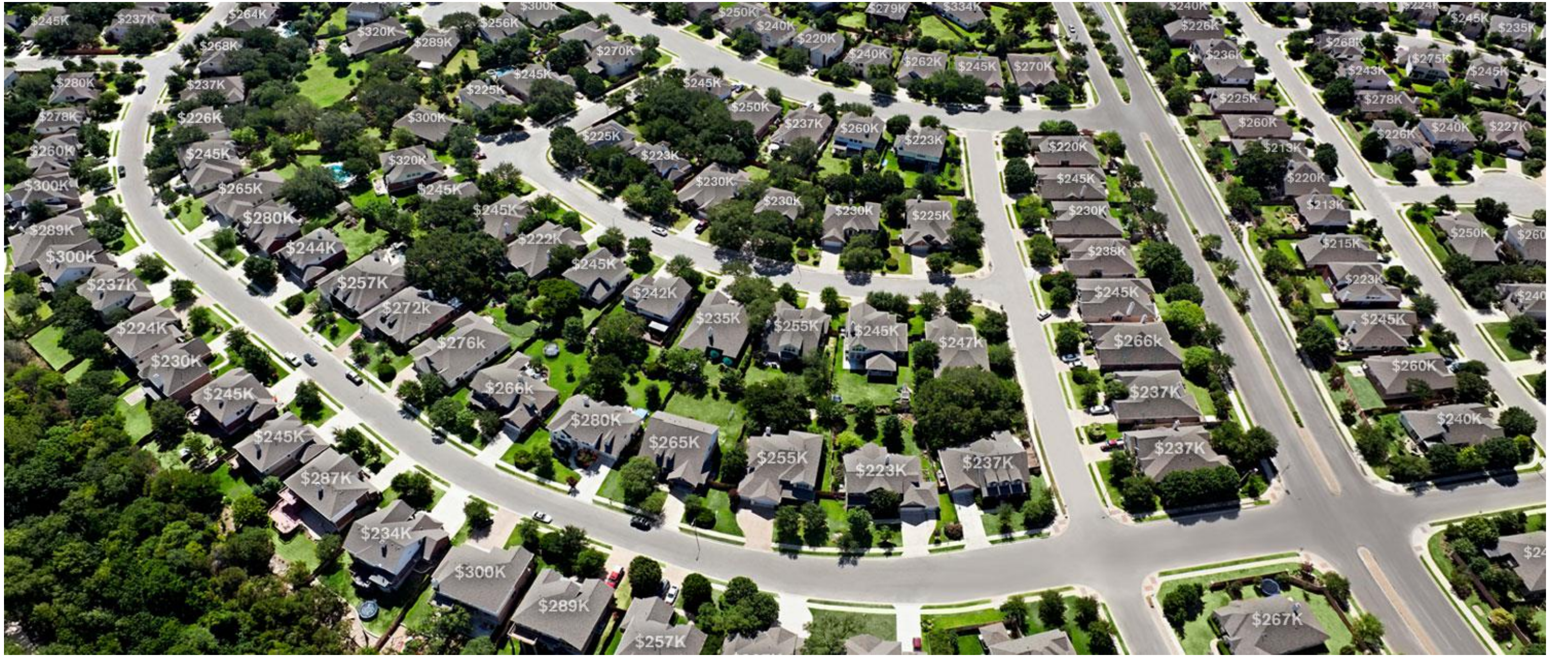
Association Analysis: Supermarket Shelf Management

- 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

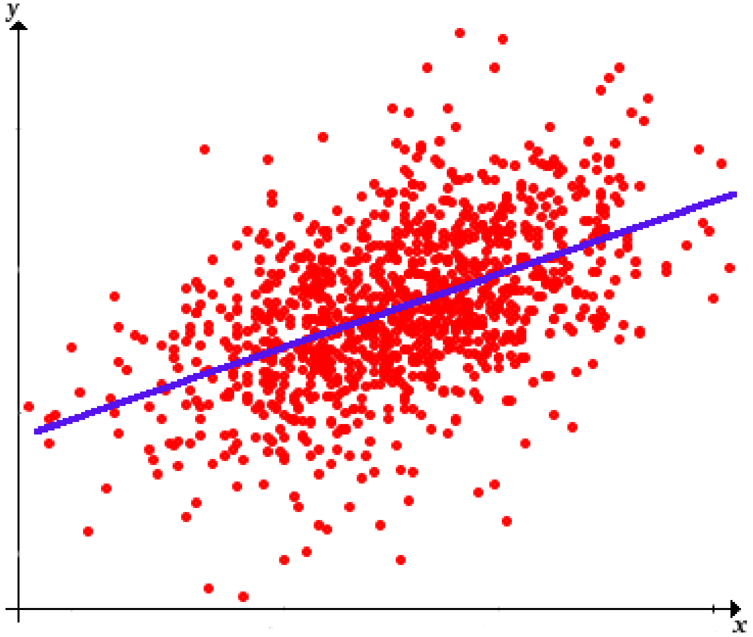
Regression Example: Predict Housing Prices



datasciencedojo
unleash the data scientist in you

Regression

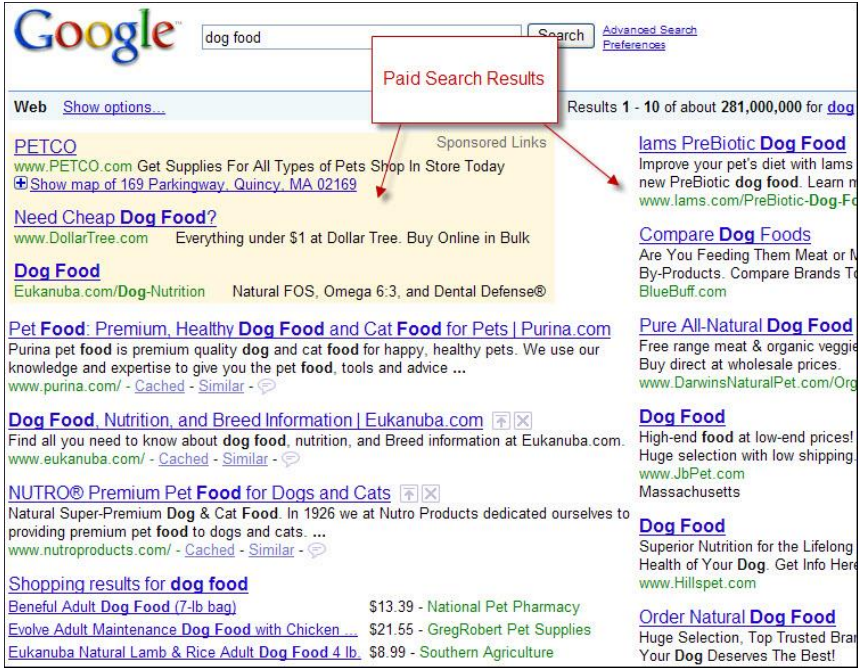
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency



datasciencedojo
unleash the data scientist in you

Regression: Ad Clicks

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



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

Sponsored Links

PETCO
www.PETCO.com Get Supplies For All Types of Pets Shop In Store Today
⊞ Show map of 169 Parkingway, 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

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

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

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

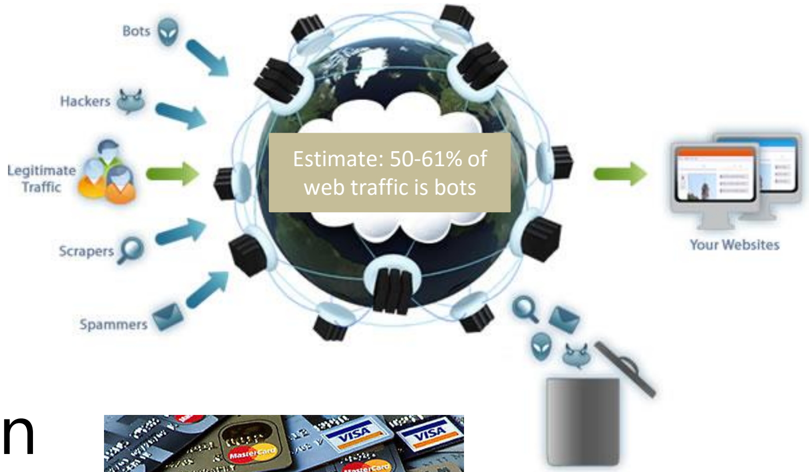
Dog Food
High-end **food** at low-end prices!
Huge selection with low shipping.
www.JbPet.com
Massachusetts

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 Bra
Your **Dog** Deserves The Best!

Deviation/Anomaly Detection

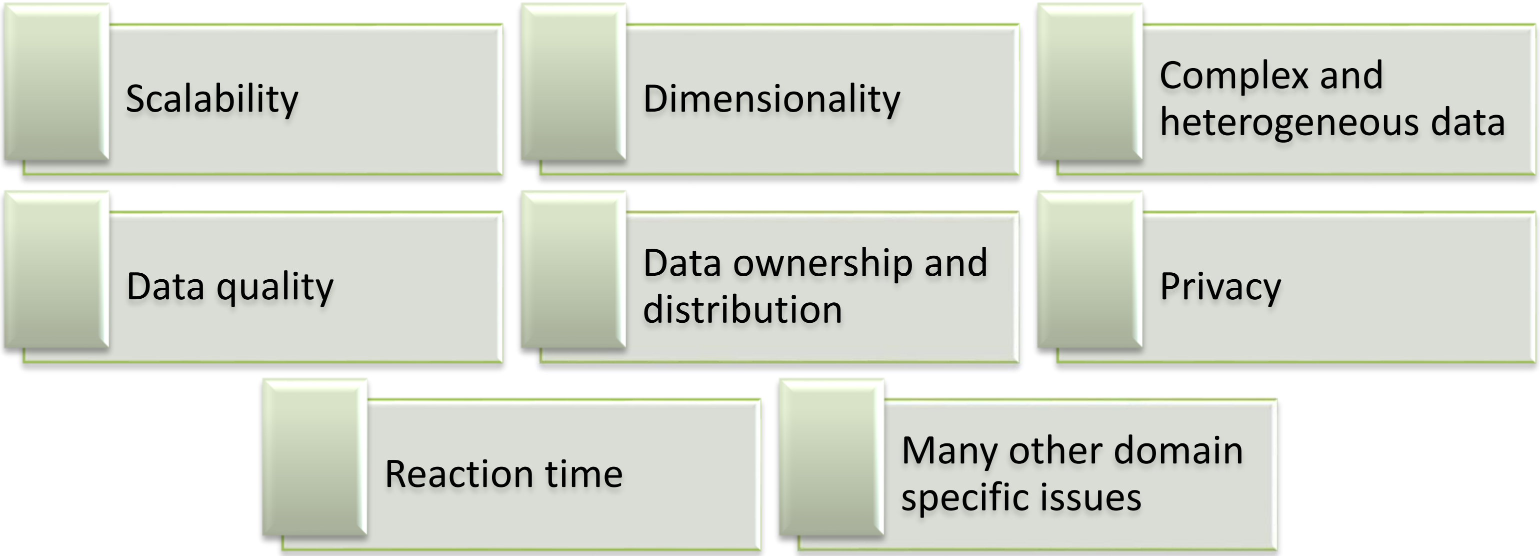
- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Bot detection in web traffic



Estimate: 50-61% of web traffic is bots

datasciencedojo
unleash the data scientist in you

Challenges in Data Mining

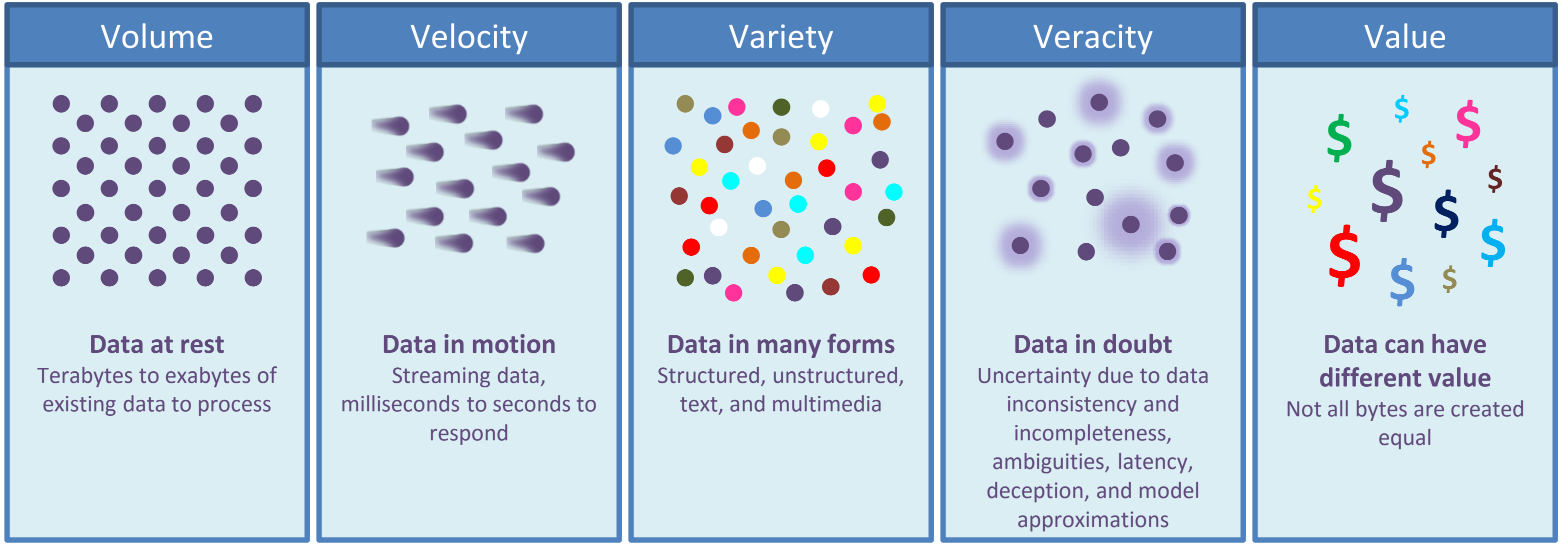


- Scalability
- Dimensionality
- Complex and heterogeneous data
- Data quality
- Data ownership and distribution
- Privacy
- Reaction time
- Many other domain specific issues

datasciencedojo

unleash the data scientist in you

5 Vs Of Big Data



Volume

Data at rest

Terabytes to exabytes of existing data to process

Velocity

Data in motion

Streaming data, milliseconds to seconds to respond

Variety

Data in many forms

Structured, unstructured, text, and multimedia

Veracity

Data in doubt

Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception, and model approximations

Value

Data can have different value

Not all bytes are created equal

datasciencedojo

unleash the data scientist in you

Questions?