Data Science and Data Engineering

Quick Overview

datasciencedojo
unleash the data scientist in you

Data Science & Machine Learning

- ~3 days
- 60% theory. 40% Hands-on Exercises
- Math/Theory is minimal but not trivial
- Primary tools: R and Azure ML Studio
- Mentored Kaggle participation
- Emphasis on best practices

datasciencedojo
unleash the data scientist in you

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.

datasciencedojo
unleash the data scientist in you

Hack Day

- Gather temperature and humidity data in real-time
- Use message queues, stream processors to get real time analytics

datasciencedojo
unleash the data scientist in you

Logistics

- Eight hours of pre-bootcamp work
- Bootcamp: May 25th – May 29th **(8:30 am – 5:00 pm)**
- Hackday: May 30th
- Slides, sample code and other resources are consolidated in a git repository:
  - Send your github id to yuhui@datasciencedojo.com
- Office hours. Kaggle. LinkedIn group

datasciencedojo
unleash the data scientist in you

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

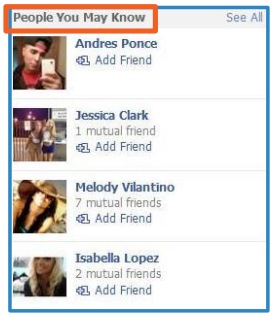
Healthcare

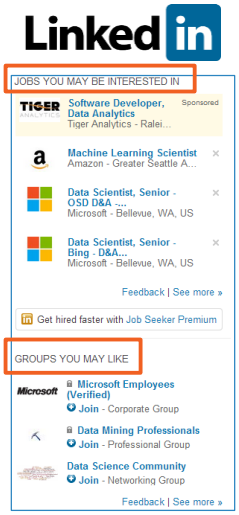
datasciencedojo
unleash the data scientist in you

Online Shopping

datasciencedojo
unleash the data scientist in you

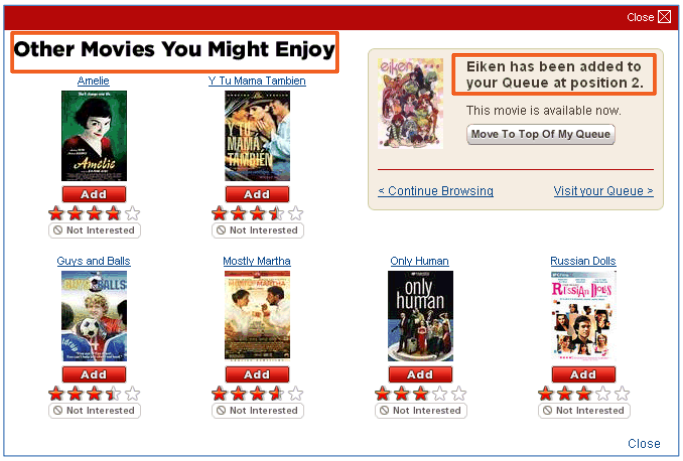
Social Networks





datasciencedojo
unleash the data scientist in you

Online Entertainment



datasciencedojo
unleash the data scientist in you

Brainstorming

- What are some other applications?

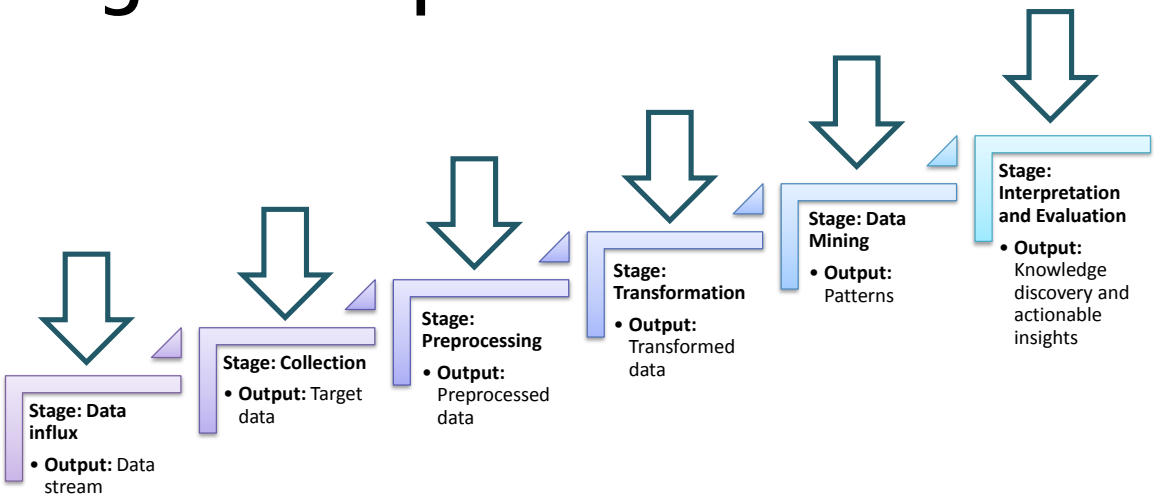
datasciencedojo
unleash the data scientist in you

Connecting the Dots

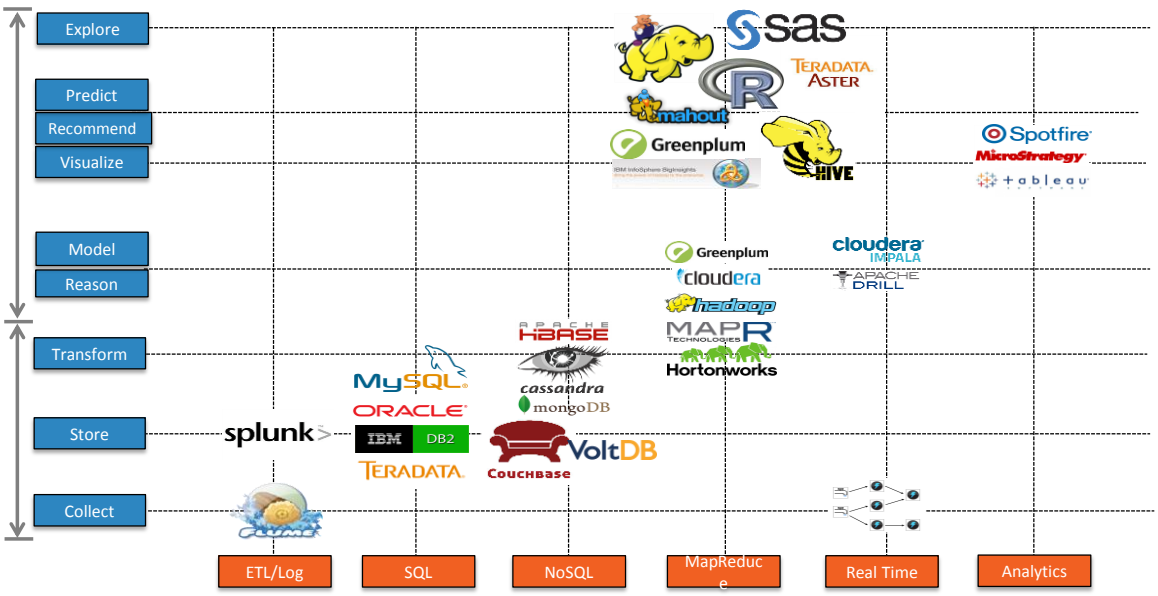
- The underlying magic behind what we saw is 'big data' and 'predictive analytics'

datasciencedojo
unleash the data scientist in you

Big Data Pipeline



datasciencedojo
unleash the data scientist in you



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

datasciencedojo
unleash the data scientist in you

Traffic Management

Descriptive [Informing Role]:

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

datasciencedojo
unleash the data scientist in you

Traffic Management

Predictive [Informing and Warning Role]:

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

datasciencedojo
unleash the data scientist in you

Traffic Management

Prescriptive [Informing, Warning, and Advisory Role]:

Take action so traffic jam does not happen

OR

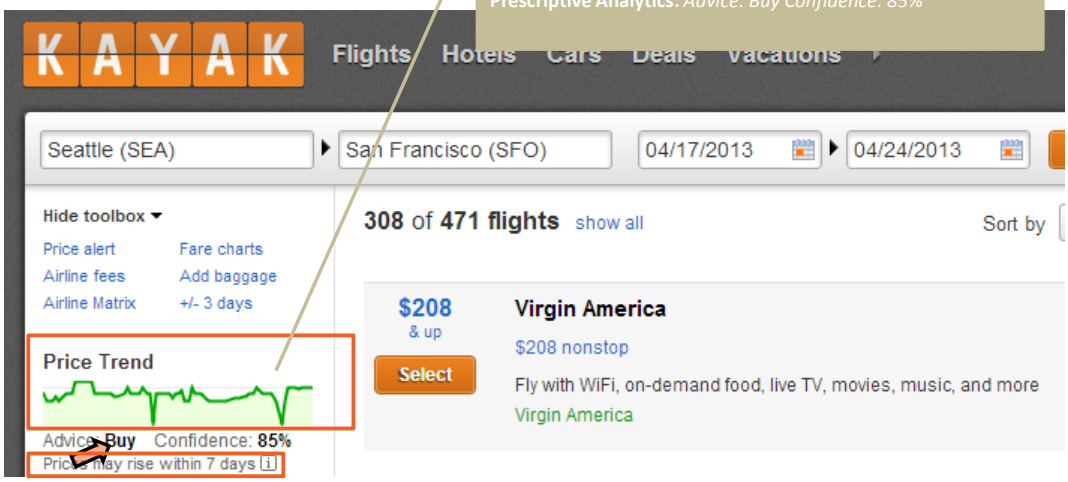
Traffic jam is about to happen in the next 30 minutes and you could possibly take the following courses of action:

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

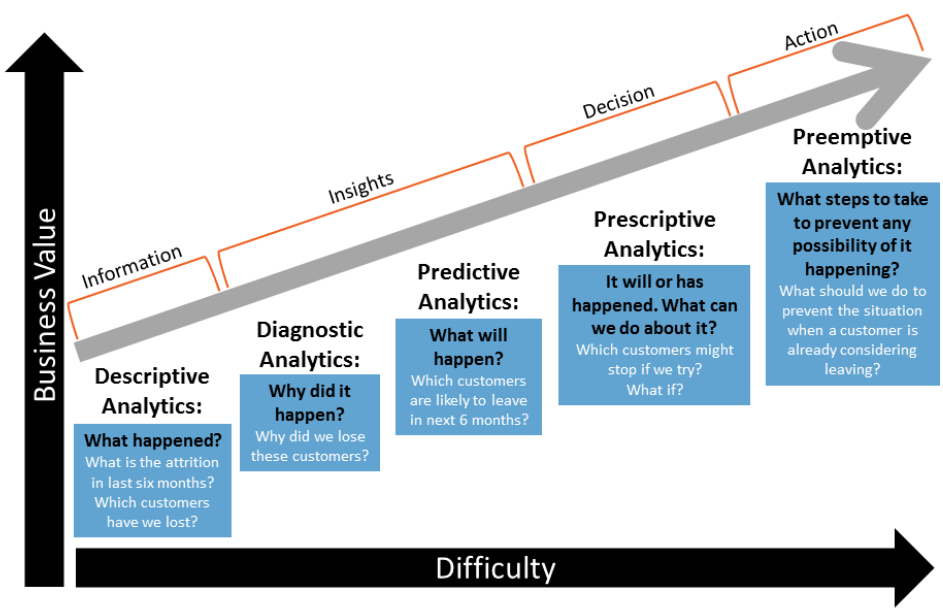
datasciencedojo
unleash the data scientist in you

Online Travel

Descriptive Analytics: Historical price trend and variation
Predictive Analytics: Price may rise in next 7 days
Prescriptive Analytics: *Advice: Buy Confidence: 85%*



unleash the data scientist in you



datasciencedojo
unleash the data scientist in you

Data Mining and Predictive Analytics

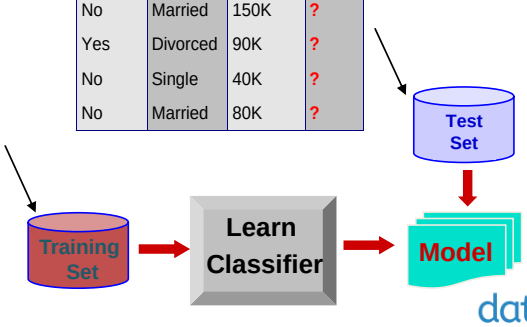
In the next few slides, we will take a look at some of the most common data mining tasks.

datasciencedojo
unleash the data scientist in you

Classification: A Simple Example

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
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	?



datasciencedojo
unleash the data scientist in you

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.

datasciencedojo
unleash the data scientist in you

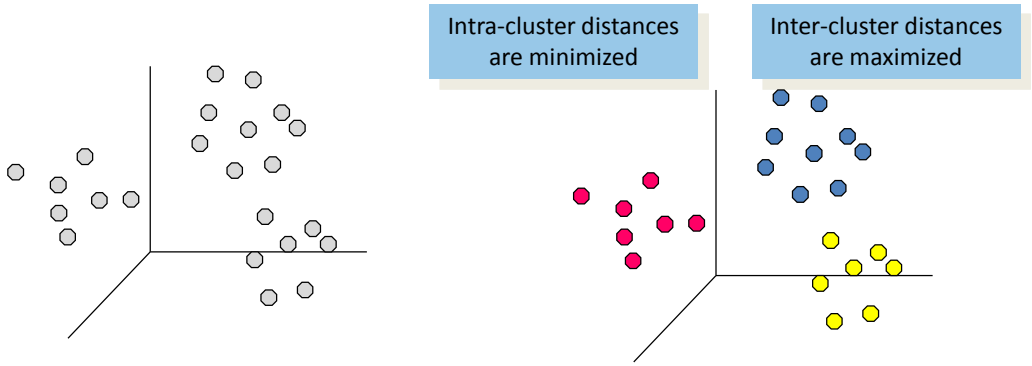
Classification: More Examples

- 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

datasciencedojo
unleash the data scientist in you

Clustering: An Illustration

Clustering in 3-D space using Euclidean distance



datasciencedojo
unleash the data scientist in you

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

datasciencedojo
unleash the data scientist in you

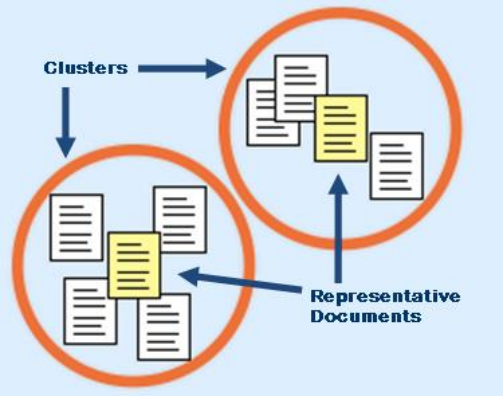
Clustering: Similarity Measures

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

datasciencedojo
unleash the data scientist in you

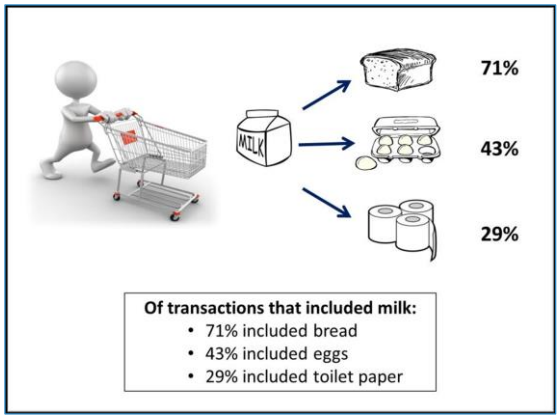
Clustering: Examples

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



datasciencedojo
unleash the data scientist in you

Association Analysis



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

datasciencedojo
unleash the data scientist in you

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}

datasciencedojo
unleash the data scientist in you

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

datasciencedojo
unleash the data scientist in you

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

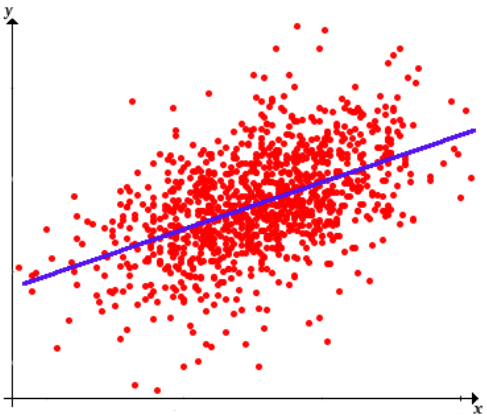
datasciencedojo
unleash the data scientist in you

Regression Example: Predict Housing Prices

datasciencedojo
unleash the data scientist in you

Regression

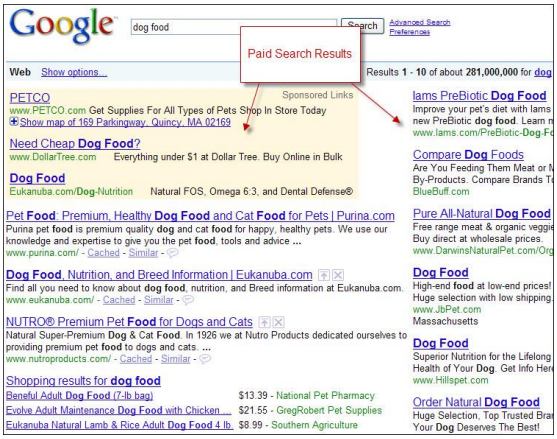
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency



datasciencedojo
unleash the data scientist in you

Regression: Ad Clicks

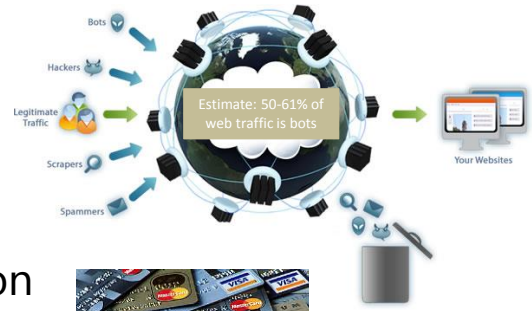
Predict the probability of whether or not an ad will be clicked



datasciencedojo
unleash the data scientist in you

Deviation/Anomaly Detection

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



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	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

datasciencedojo
unleash the data scientist in you

Questions?

datasciencedojo
unleash the data scientist in you