


Data Analysis / Data Science on Hadoop

Agenda

- Data Science and Data Scientist
- High level understanding of Machine Learning
- Introduction to Data Analysis using Pig and Hive
- Data Science using Python
- Hands on Tutorials

Data Science, Business Intelligence, and Data Scientist

Google transformed advertising with AdWords

Google 

[Web](#) [Maps](#) [Videos](#) [Images](#) [Shopping](#) [More ▾](#) [Search tools](#)


About 2,310,000 results (0.28 seconds)

Ads related to **beginner yoga classes** ⓘ

[Beginners Yoga Classes - yogaglo.com](#)
[www.yogaglo.com/BeginnerYoga ▾](#)
Find the Best **Beginner Yoga Classes** at YogaGlo® Today! Free Trial Offer
YogaGlo has 1,263 followers on Google+
[How It Works](#) [Start Your Free Trial](#)
[Beginners Center](#)

[Online Yoga Videos - Try Online Yoga For Beginners](#)
[www.myogaworks.com/ ▾](#)
Free 14 Day Trial - Start Today!

[Yoga Online - DailyBurn - Beginner To Advanced Yoga Workouts](#)
[www.dailyburn.com/Yoga ▾](#)
Start Your Free 30-Day Trial Today.
Be At Peace with Pricing - Namaste from DailyBurn - Discover the Yogi In You

[Yoga for Complete Beginners - Yoga Class 20 Minutes - YouTube](#)
 [www.youtube.com/watch?v... ▾](#) YouTube ▾
Dec 6, 2010 - Uploaded by YogaVidyaEnglish
Yoga for complete **beginners**. 20 minute gentle **yoga class** to give you greater relaxation, more energy and ...

[DoYogaWithMe.com: Free Online Yoga Videos - Classes and Poses](#)
[www.doyogawithme.com/ ▾](#)
Online yoga videos from DoYogaWithMe.com. ... Browse **Yoga Class** Videos By: ... This is a great transition class from **beginner** ashtanga to the full primary ...

Ads ⓘ

[Zumba® Class by Zip Code](#)
[www.zumba.com/FindAClass ▾](#)
A Fun, Fast & Effective Workout.
Find a Zumba® **Class** Near You Today!

[Learn Yoga for Free Today](#)
[www.alison.com/Free-Yoga ▾](#)
Improve Core Strength & Flexibility
Free, Online, Self-Paced Course

[Yoga Classes for Beginners](#)
[www.hulu.com/plus ▾](#)
Try Hulu Plus! More TV Shows & Movies. Get 1 Week Free Now.

[Yoga Videos For Beginners](#)
[www.gaiamtv.com/FreeTrial ▾](#)
Practice **Yoga** On Your Own Time.
Sign Up for Your Free Trial Today!










[Beginners Yoga Course](#)
[webcrawler.com/beginners+yoga+course ▾](#)
Find more facts, sites and tools.
Get more **beginners yoga** course

[Yoga Certification NYC](#)
[www.atmananda.com/ ▾](#)
200 hr **Yoga** Alliance, all Summer
Inc Free **Yoga** Classes







LinkedIn network growth: People You May Know

People You May Know beta

See people from different parts of your professional life




All Suggestions / LinkedIn (12) Connect All

 <p>Brad Mauney <small>2nd</small> Senior Product Manager, Search & Social Graph at LinkedIn Mountain View, California</p> <p>Connect 5 shared connections</p>	 <p>Albert Wang <small>2nd</small> Senior User Experience Designer at LinkedIn Mountain View, California</p> <p>Connect 127 shared connections</p>
 <p>Sam Shah <small>2nd</small> Principal Engineer at LinkedIn Mountain View, California</p> <p>Connect 22 shared connections</p>	 <p>Tan Nhu <small>2nd</small> Senior Web Developer at LinkedIn Mountain View, California</p> <p>Connect 16 shared connections</p>
 <p>Vinodh Jayaram <small>2nd</small> Software Engineering Manager at LinkedIn Mountain View, California</p> <p>Connect 10 shared connections</p>	 <p>Andy Chen <small>2nd</small> Software Engineer at LinkedIn Mountain View, California</p> <p>Connect 78 shared connections</p>

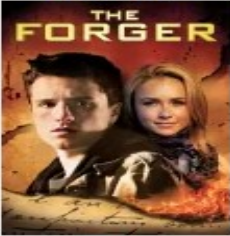
Netflix: 75% of streaming video from recommendations

Because you added **The Hunger Games**

The Hunger Games
has been added to My List



The Forger




Play

★★★★☆

Not Interested

Red Dawn




Play

★★★★☆

Not Interested

Pirates of the Caribbean: Black Pearl

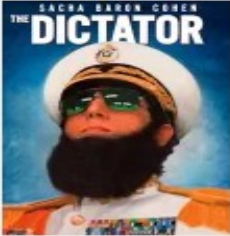


Play

★★★★☆

Not Interested

The Dictator




Play

★★★★☆

Not Interested

Hansel & Gretel: Witch Hunters




Play

★★★★☆

Not Interested

Olympus Has Fallen




Play

★★★★☆

Not Interested

The Grey




Play

★★★★☆

Not Interested

Mission: Impossible - Ghost Protocol




Play

★★★★☆

Not Interested

Skyfall




Play

★★★★☆

Not Interested

Jack Reacher



Play

★★★★☆

Not Interested

Amazon: 35% of sales from product recommendations

Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#)

Even Faster Web Sites: Performance... (Paperback) by Steve Souders
★★★★★ (7) \$23.10
[Fix this recommendation](#)

Simply JavaScript (Paperback) by Kevin Yank
★★★★☆ (19) \$26.37
[Fix this recommendation](#)

The Art & Science of JavaScript (Paperback) by Robert Iler
★★★★★ (5)
[Fix this recommendation](#)

Any Category Algorithms Boxed Sets Business & Culture Java
Graphic Design Microsoft Networking Networks, Protocols & APIs New SQL

Retail: Market Basket Analysis

Question

Which products are very frequently purchased together

Solution approach

Determine frequent item sets from transaction logs

Use to design physical store layout accordingly

Use to design special offers and coupon strategy

Example:

The urban legend of Diapers & Beer

Finance: Customer Profiling

Question

How likely is this customer to pay back the loan?

Solution approach

Collect information about customer (e.g. FICO score)

Build predictive model based on historical data

Use model to price loan products based on risk

Finance: Fraud Detection

Question

How can I detect fraudulent credit card activity?

Solution approach

Track spending habits

Build model of typical spending

Alert when abnormal transaction occurs

Example:

A transaction for the purchase of \$10K worth of Jewellery
from Mongolia

What is a Data Scientist?

“I keep saying that the sexy job in the next 10 years will be statisticians,” said **Hal Varian**, chief economist at Google. “And I’m not kidding”

The Data Science Skillset Continuum

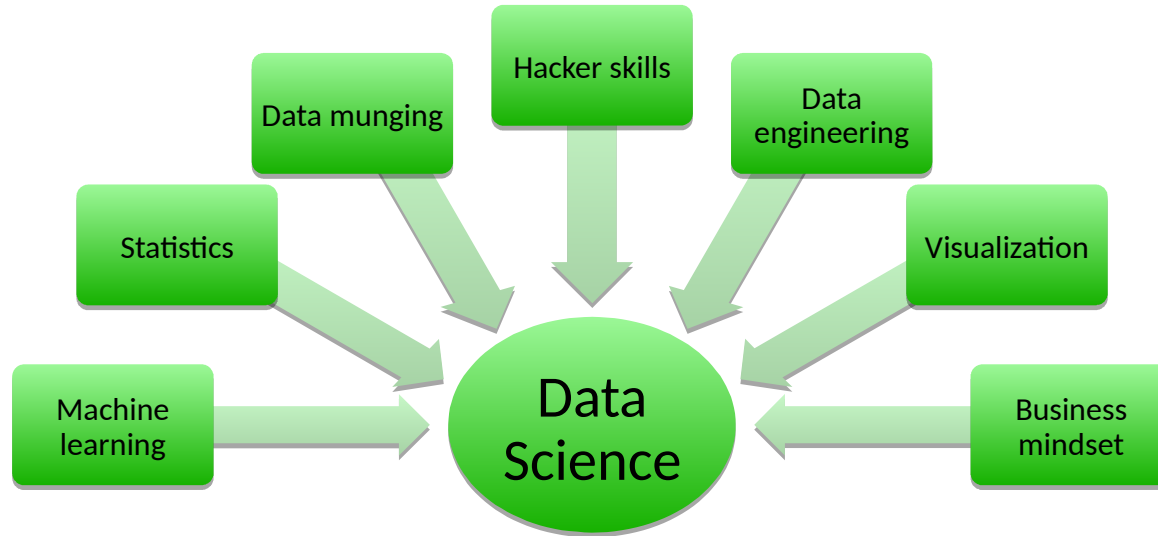


Role	Data Engineer	Applied Scientist
Function	Build production-grade data products	Find signal/meaning in the data Applies statistical/ML models and tunes the algorithm
Good at....	Data and Systems architecture Hadoop, PIG/HIVE, MapReduce, ops/admin Java, Python, Perl, SQL, C++, NoSQL (Hbase, Cassandra, Mongo)	Statistics, Machine learning Text processing, NLP R, Python, Matlab, SAS, SQL Scripting Visualization / telling the story

Skill set for a data scientist

Our perspective:

- It's a hybrid role: data engineer + applied scientist
- Combines many disciplines:



Machine Learning

"The study of systems that can learn from data"

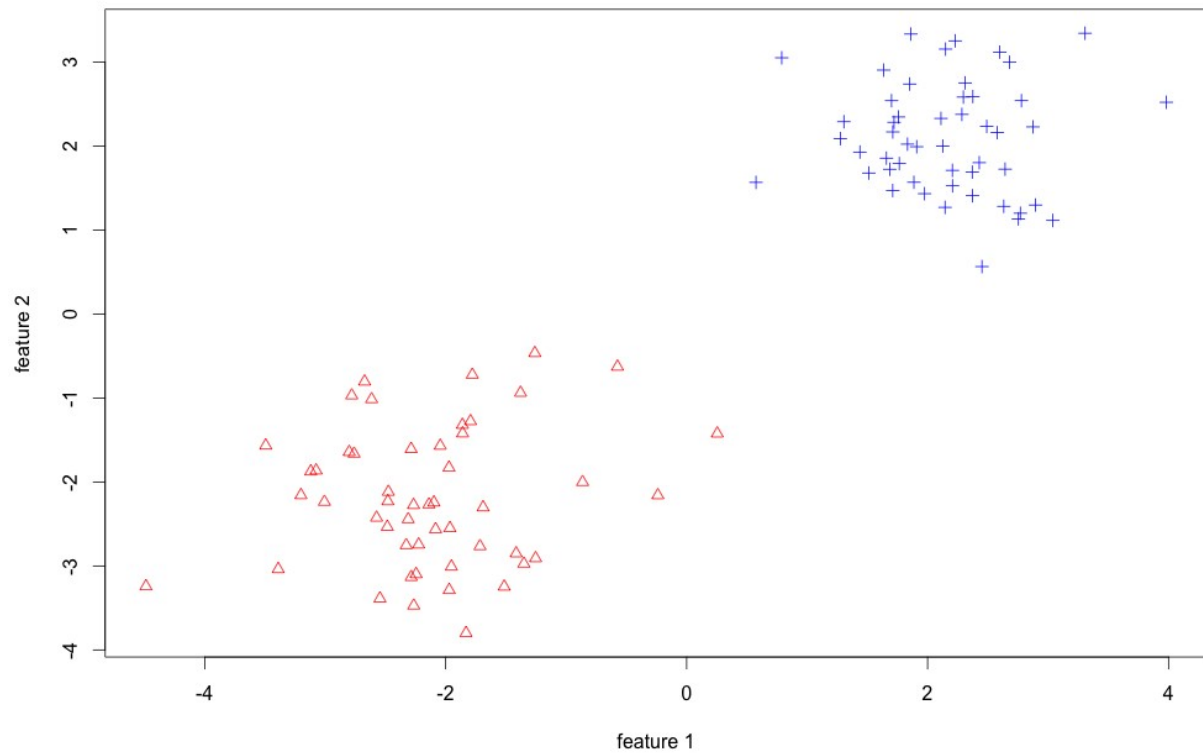
Machine Learning and Big Data

- Learning from large datasets (terabytes, petabytes)
- Distributed algorithms for machine learning, compatible with MapReduce


Six Machine Learning Tasks

- Unsupervised
 - Clustering
 - Outlier Detection
 - Affinity Analysis
- Supervised
 - Classification
 - Regression
 - Recommendation

Clustering



Clustering Example

[web](#) [news](#) [images](#) [maps](#) [blogs](#) [wikipedia](#) [jobs](#) [more »](#)

[advanced preferences](#)

[clouds](#) [sources](#) [sites](#) [time](#)

All Results (169)

Scientists (12)

Social (6)

NSF, Gov (6)

Project, Fair (6)

International (22)

Journal (7)

Research (19)

Summaries (6)

Health (7)




Preserving (2)




[more](#) | [all clouds](#)




find in clouds:




remix




Top 169 results of at least 266,000,000 retrieved for the query **data science** ([details](#))

[CODATA, The Committee on Data for Science and Technology](#)   
International Council for **Science** : Committee on **Data for Science** and ... CODATA, The Committee on **Data for Science** and Technology. 5 rue Auguste ... The mission of CODATA is to strengthen international **science** for the benefit ...
[www.codata.org](#) - [cache] - Additional Sources, Yippy Sources II

[Science & Data - SCAR \(Scientific Committee on Antarctic ...](#)   
You are in: Home » **Science & Data** . **Science & Data** . Antarctic Climate Change and the Environment. ACCE Report published by SCAR; ACCE Expert Group; **Scientific** ...
[www.scar.org/researchgroups](#) - [cache] - Additional Sources

[CODATA Data Science Journal](#)   
The **Data Science** Journal is a peer-reviewed electronic journal publishing papers on the management of **data and databases in Science** and Technology.
[www.codata.org/dsj/index.html](#) - [cache] - Additional Sources

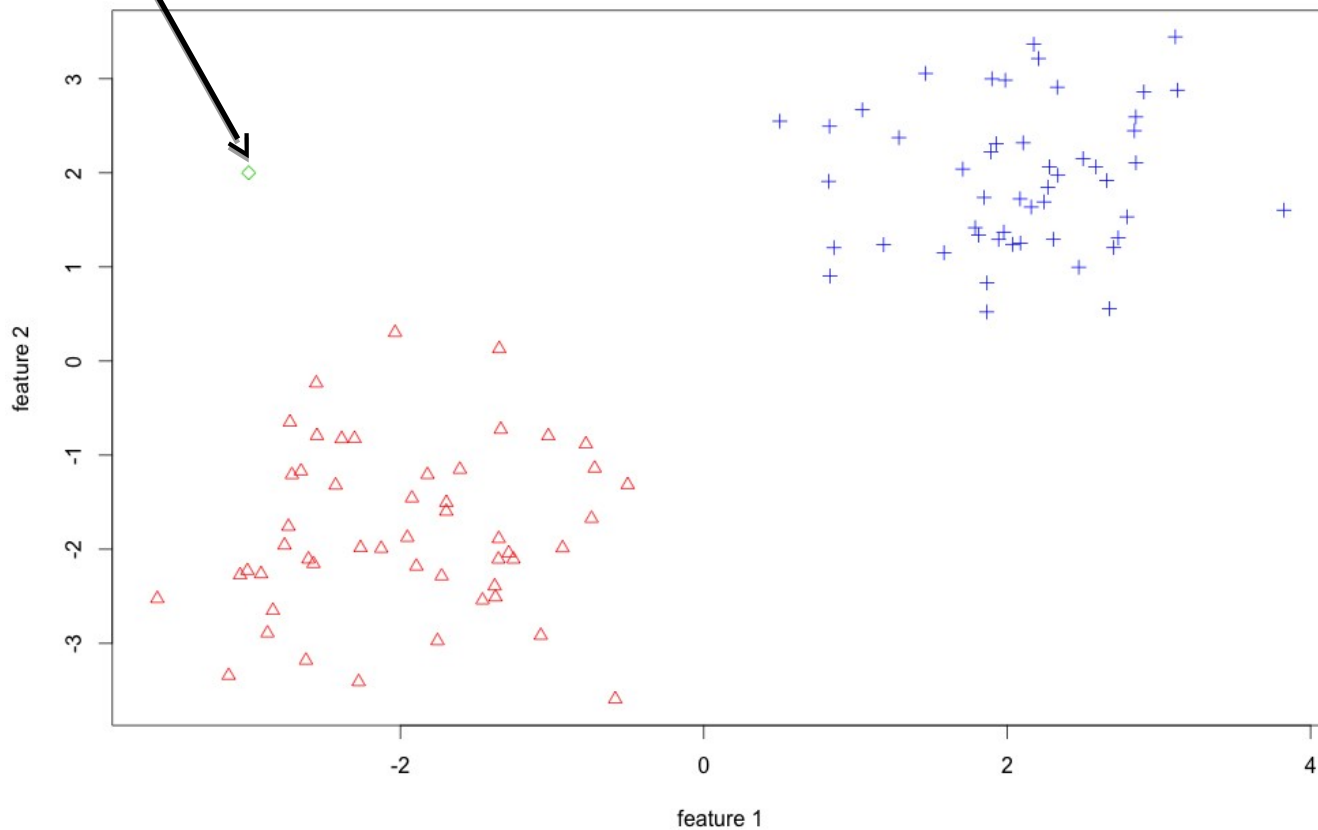
[Building Data Science Teams: DJ Patil: Amazon.com: Kindle Store](#)   
Start reading Building **Data Science** Teams on your Kindle in under a minute. Don't have a Kindle? Get your Kindle here.
[www.amazon.com/Building-Data-Science-Teams-ebook/dp/B005O4U3ZE](#) - [cache] - Additional Sources, Yippy Sources

[nsf.gov - National Center for Science and Engineering ...](#)   
Data and Tools SESTAT WebCASPAR **Data** Files: Errata: Site Features: ... The National **Science** Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, ...
[www.nsf.gov/statistics](#) - [cache] - Additional Sources

Font size:

Outlier Detection

Outlier point



Example: Credit Card Fraud Detection



Affinity Analysis

	Item 1	Item 2	Item 3	Item 4	Item 5	..
Tx 1	Y	N	N	Y	N	
Tx 2	Y	N	N	Y	N	
Tx 3	Y	Y	N	Y	N	
Tx 4	N	N	Y	Y	Y	
Tx 5						
...						



	Item 1	Item 2	Item 3	Item 4	Item 5	..
Tx 1	Y	N	N	Y	N	
Tx 2	Y	N	N	Y	N	
Tx 3	Y	Y	N	Y	N	
Tx 4	N	N	Y	Y	Y	
Tx 5						
...						

Example: Yahoo! Search Assist

YAHOO!

[Web](#) [Images](#) [Video](#) [Local](#) [Shopping](#) [Apps](#) [News](#) [More ▾](#)

Oscars

Search

oscars 2013

oscars

oscars 2012

oscars 2011

oscars 2009

oscars predictions

oscars 2008

oscars 2007

oscars 2006

oscars fish

OSCARS 2013

LATEST NEWS

Oscars 2013: Barbra Streisand to Perform for First ...
The Academy Award winner will give "a very special ...



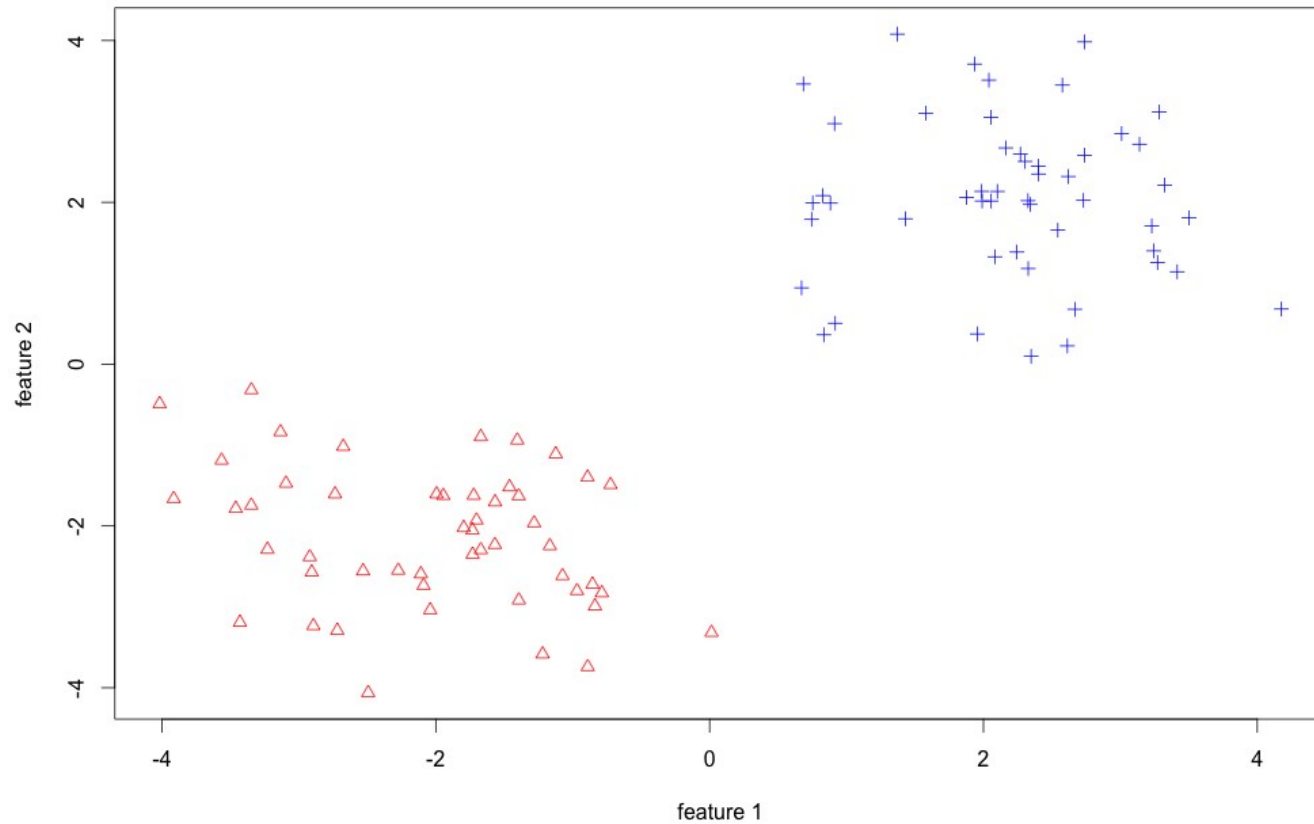
Barbra Streisand will sing at Oscars

Oscars 2013: 'Argo,' 'Lincoln' wage distinct campaigns

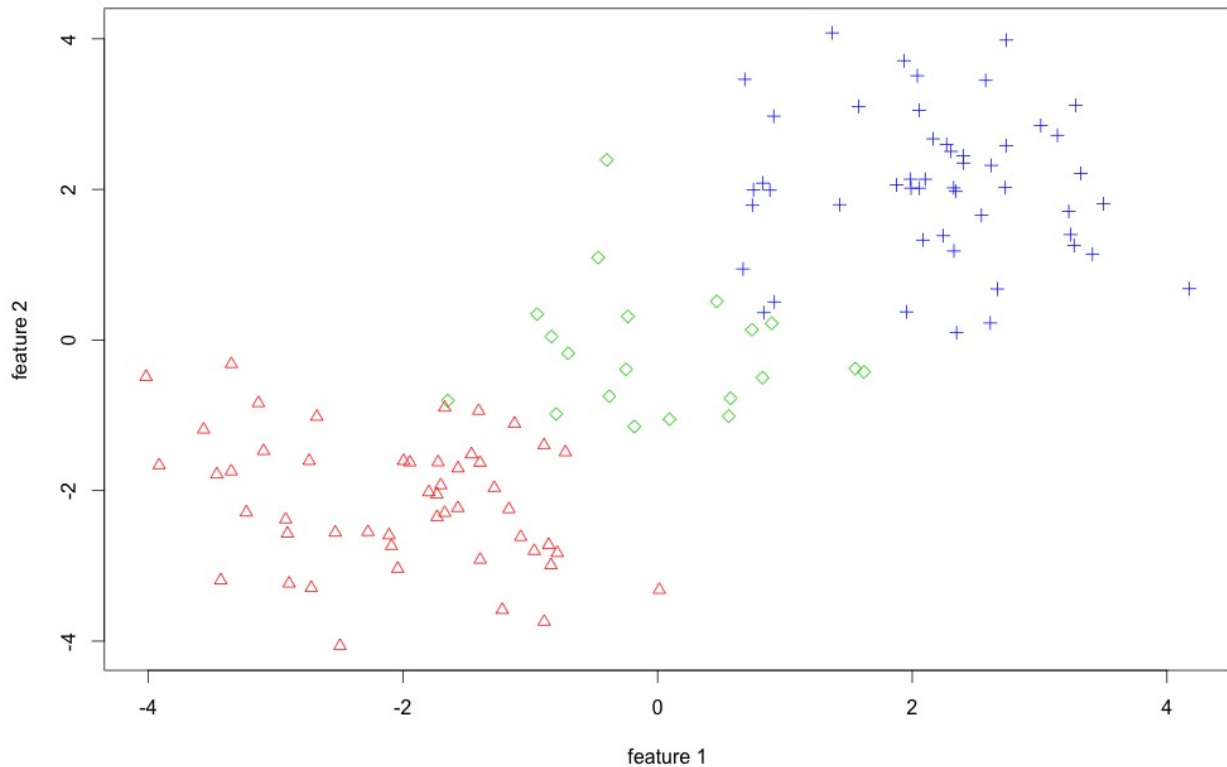
[View Photos](#) • [View Videos](#) • [Twitter Results](#)

YAHOO! NEWS

Classification



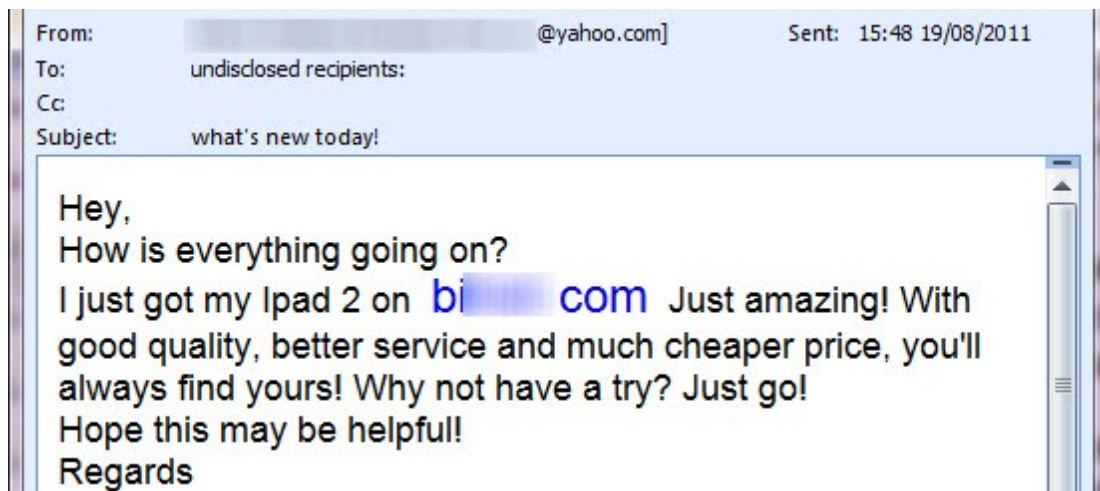
Using the model to classify new points



Classification Techniques

- Naive Bayesian
- Support Vector Machines
- Logistic regression
- ... many more ..

Example: Email Spam Detection



2 classes: Spam or Not-Spam

Features: words that appear (or not) in the email text

Regression Analysis

"In statistics, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables."

Regression Example: Click-Through Rate Prediction

flights to Miami

Web Images Maps Shopping News More Search tools

4 personal results. 30,900,000 other results.

Ads related to flights to miami

Miami @ \$147 Round Trip
www.cheapoair.com/Miami-Cheap-Flights
★★★★★ 435 reviews for cheapoair.com
Low Fares Available on Flights. Book Now & Save Big. Limited Seats!
Under \$150 Round Trip Flights Top 25 Flight Deals
Under \$199 Round Trip Flights Best Domestic Flights Deals

Flights To Miami - Low Fares on American Airlines - AA.com
www.aa.com/Miami
Book on AA.com Today & Save!
1,066,185 people +1'd or follow American Airlines
Book Flights - Discount Flight Deals - Lowest Price Guarantee - Travel Deals

Find Flights To Miami
www.kayak.com/
Don't Overpay For Your Airfare. Compare NYC to Miami Airfare
531,363 people +1'd or follow KAYAK

Sponsored

Flights from San Francisco, CA (SFO) to Miami, FL (MIA)
www.google.com/flights

Depart	Sat February 16	Return	Wed February 20
Nonstop	American	5h 50m	from \$523
	Alaska	5h 50m	from \$624
All flights	American	6h 40m+	from \$338
	United	8h 15m+	from \$429
	Other airlines	6h 45m+	from \$354

More Google flight search results

Ads

JetBlue - Official Site
www.jetblue.com/
Winter deals from \$59 one way.
See all flight deals & book now!
201 people +1'd this page

Find Flights to Miami, FL
www.united.com/
Get United's Guaranteed Lowest Fare to Miami, Florida. Book Now.

\$49 Miami Flights
flights.cheapflightnow.com/Miami
Cheapest Miami Flights
Up To 65% + \$12 Off. Limited Offers

Miami Flights From \$99
www.onetravel.com/Miami-Flights
★★★★★ 229 reviews for onetravel.com
Cheap Round Trip Flights to Miami.
Save up to \$15 Off Fees. Book Now !

Flights to Miami from \$136
www.travelzoo.com/miami
★★★★★ 119 reviews for travelzoo.com
Compare Top Travel sites & Airlines
Find Cheap flights to Miami
597 people +1'd or follow Travelzoo

\$49* to Miami
www.farespotter.net/Miami+Flights
Promo for Miami Flights.
Today Only! Tickets from \$49.
183 people +1'd or follow
Farespotter.net

Rank = bid * CTR

Predict CTR for each ad to determine placement, based on:

- Historical CTR
- Keyword match
- Etc...

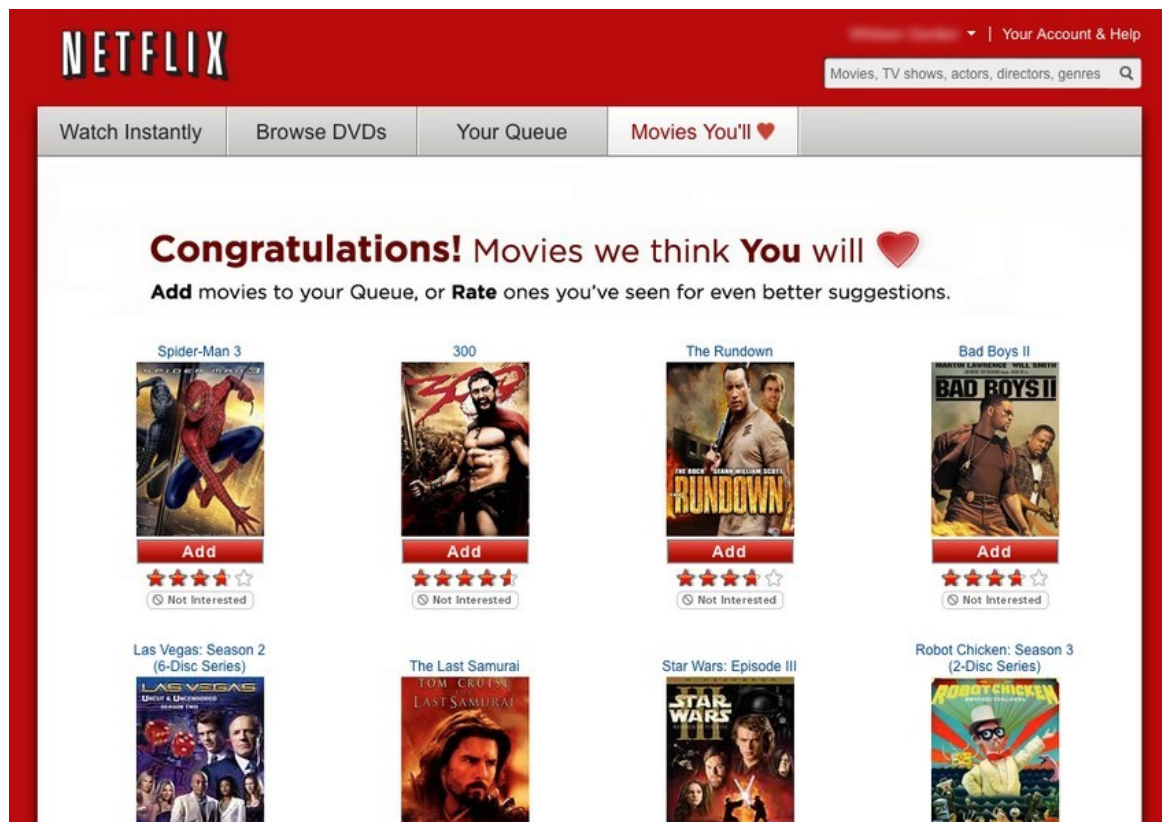
Recommendation Engine

	Harry potter	X-Men	Hobbit	Argo	Pirates
101	5	2	4	?	?
102	?	?	5	2	?
103	1	2	?	?	3
104					
105					
...					



	Harry potter	X-Men	Hobbit	Argo	Pirates
101	5	2	4	1	3
102	4	1	5	2	3
103	1	2	4	1	3
104					
105					
...					

Example: Netflix Movie Recommendations



Machine Learning and Hadoop

- Parallel training of multiple models:
 - We want to build multiple models
 - Example: build CTR model per keyword
- Ensemble training models:
 - We want to build one model by the voting of many sub-models
 - Example: random forest
- Distributed learning algorithms:
 - We want to build a compute-heavy model using MapReduce
 - Good when: few iterations, and not much data b/w iterations
 - Examples: linear regression, Naïve Bayes, k-means clustering, ALS, frequent item-set mining
 - Mahout implements many of these

Mahout Algorithms

```
[root@sandbox ~]# mahout
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.8.0.2.0.6.0-76-job.jar
An example program must be given as the first argument.
Valid program names are:
  arff.vector: : Generate Vectors from an ARFF file or directory
  baumwelch: : Baum-Welch algorithm for unsupervised HMM training
  canopy: : Canopy clustering
  cat: : Print a file or resource as the logistic regression models would see it
  cleansvd: : Cleanup and verification of SVD output
  clusterdump: : Dump cluster output to text
  clusterpp: : Groups Clustering Output In Clusters
  cmdump: : Dump confusion matrix in HTML or text formats
  concatmatrices: : Concatenates 2 matrices of same cardinality into a single matrix
```


Introduction to Apache Pig



Hadoop Essentials Working with Pig



© Hortonworks Inc. 2012

Youtube: Hadoop Tutorial : Apache Pig

Introduction to Hive



Hadoop Essentials Programming with Hive



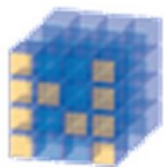
© Hortonworks Inc. 2012

Youtube: Hadoop Tutorial : Apache Hive

Python and Data Analysis

The Scientific Python Ecosystem

- The following commonly-used Python modules are a part of an ecosystem referred to as **SciPy**:



NumPy
Base N-dimensional
array package



SciPy library
Fundamental library
for scientific
computing



Matplotlib
Comprehensive 2D
Plotting



IPython
Enhanced Interactive
Console



Sympy
Symbolic
mathematics



pandas
Data structures &
analysis

Overview of NumPy



- ***NumPy*** (Numerical Python) is the fundamental package for scientific computing with Python. It contains:
 - N-dimensional array object called ***ndarray***
 - Routines for working with arrays
 - Linear algebra routines
 - Other capabilities like Fourier transforms, financial functions, random sampling, and matrices

Overview of pandas

- ***pandas*** is an open-source library with easy-to-use data structures and functions that simplifies data analysis and modeling in Python, including:
 - **DataFrame** and **Series** data structures
 - tools for reading and writing data in CSV format, Excel, and SQL databases
 - “group by” operator on data sets
 - merging and joining data sets
- pandas data structures are used in many other Python libraries, so it is a good library to be familiar with.

The SciPy Library

- **SciPy** is a collection of mathematical algorithms and functions built on NumPy, including:
 - **scipy.cluster**: clustering algorithms
 - **scipy.interpolate**: spline functions and interpolation classes
 - **scipy.odr**: orthogonal distance regression
 - **scipy.sparse**: for working with sparse matrices
 - **scipy.spatial**: spatial algorithms and data structures
 - **scipy.stats**: large collection of statistical functions

Options for Running Python on Hadoop

- If you are going to develop Data Science solutions in Python, then you need the ability to run those solutions on your Hadoop cluster.
- This unit demonstrates two techniques:
 - **Pig User Defined Functions (UDFs)**: Invoking Python code from within a Pig script
 - **Pig Streaming**: Using the Pig **STREAM** command

Overview of Pig UDFs

- Pig provides extensive support for user defined functions (UDFs) as a way to specify custom processing.
- UDFs can be written in:
 - Java
 - Jython: a Python implementation that runs on a JVM
 - Python (Python UDFs do not work on Hadoop 2.x)
 - JavaScript, Ruby and Groovy (experimental)
- HDP 2.x uses Jython for UDFs written in Python

UDF Libraries

- There are some open source Pig UDFs that are packaged and available for use, including:
 - **Piggy Bank**: a popular collection of commonly-used functions for Pig.
 - **DataFu**: created by LinkedIn, DataFu is a collection of Pig UDFs for data analysis on Hadoop.
 - **ElephantBird**: Twitter's open source library of InputFormats, OutputFormats, Writables, Pig LoadFuncs, Hive SerDe, and more.

An Example of Using a UDF

```
register 'datafu-1.2.0.jar';  
define Quintile datafu.pig.stats.Quantile('5');
```

```
temps_filter = FILTER temperatures  
               BY hightemp is not null;  
temps_group = GROUP temps_filter  
               BY location;  
quintiles = FOREACH temps_group {  
    sorted = ORDER temps_filter BY hightemp;  
    GENERATE group AS location,  
              Quintile(sorted.hightemp) AS quant;  
}
```

Hadoop Map Reduce Streaming

- A Streaming job is a MapReduce job defined in the **hadoop-streaming.jar** file:

```
hadoop jar $HADOOP_HOME/lib/hadoop-streaming.jar  
-input input_directories  
-output output_directories  
-mapper mapper_script  
-reducer reducer_script
```

Tutorials