COEN 169

Data Science

Yi Fang

Department of Computer Engineering

Santa Clara University

Netflix Prize of COEN 169

0.742817271384009	**Choulos, Alexander**
0.746100804465605	**Velcich, Kevin**
0.749425381710721	**Zhang, Tian**

http://www.cse.scu.edu/~yfang/coen169/ranklist.pl

Summary of Recommendation Assignment

- 3 below 0.75, 5 below 0.78, 6 below 0.8, 14 below 0.84
- Various programming languages (Java, Python, C++, C, Perl, etc.)
- Average number of uploads per student: 22

Administrative Stuff

- Final exam on June 9 (Thursday), 1:30pm-3:30m, EC 325.

✓ Comprehensive with more weight on the second half of the course

✓ Closed books

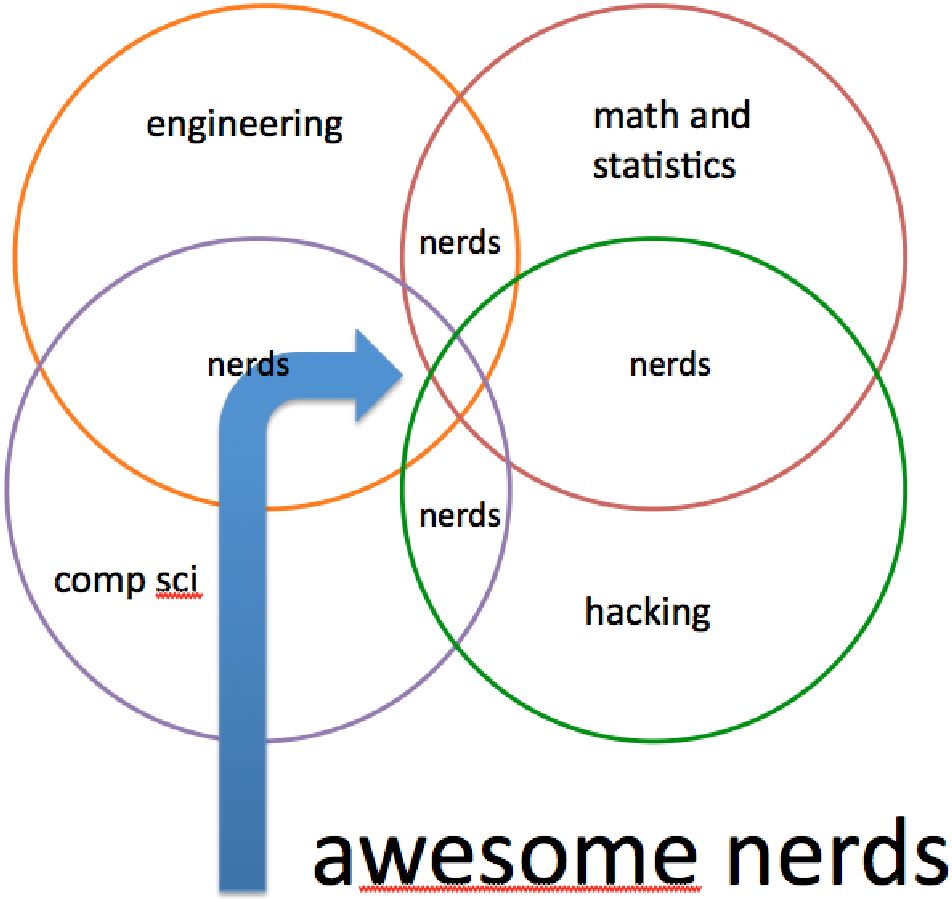
✓ A calculator and one (regular letter size) cheat-sheet are allowed

✓ Format and style are similar to the midterm exam

✓ Office hours: Friday 1-2pm, Tuesday 2-3pm, Wed 4-6pm

What is Data Science?

Data scientists?



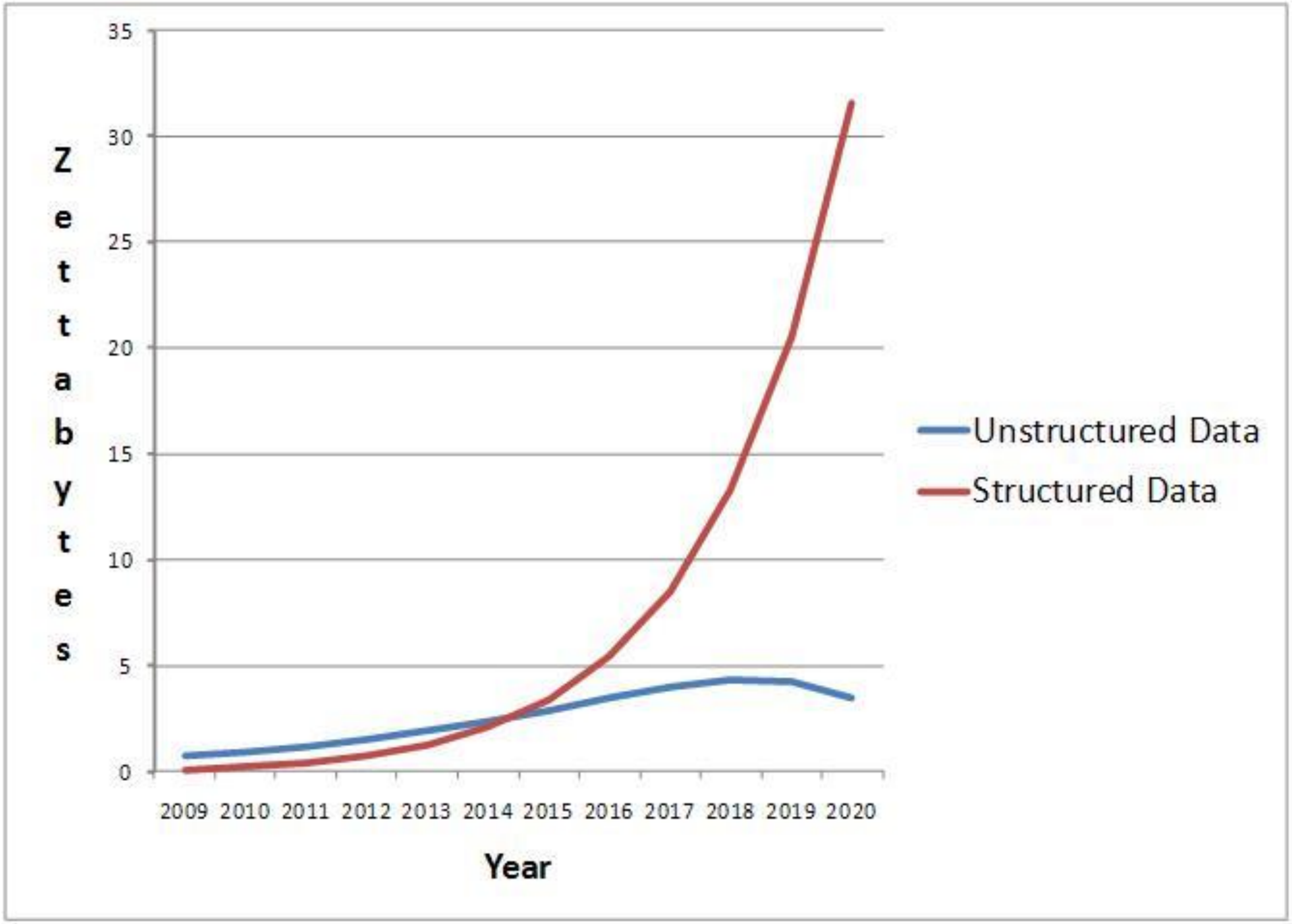
awesome nerds

Why is data science different from other fields?

Unstructured Data

- Documents
- Webpages
- Images
- Audio
- Video
- More…

Growth



http://www.emc.com/leadership/programs/digital-universe.htm

Big Data

Any dataset where the size or speed of incoming data causes difficulties in processing

- Volume
- Velocity
- Variety

Law of Data

18 Months

the amount of time for digital data to double

Why do you care?

“Every single industry will be totally revolutionized by big data”

- Joe Tucci, EMC

Big Data Examples

- Google: > 100 PB; > 1T indexed URLs
- Facebook: 1 billion users; 40 billion photos
- YouTube: > 750 PB
- Twitter: > 55 billion tweets/year;

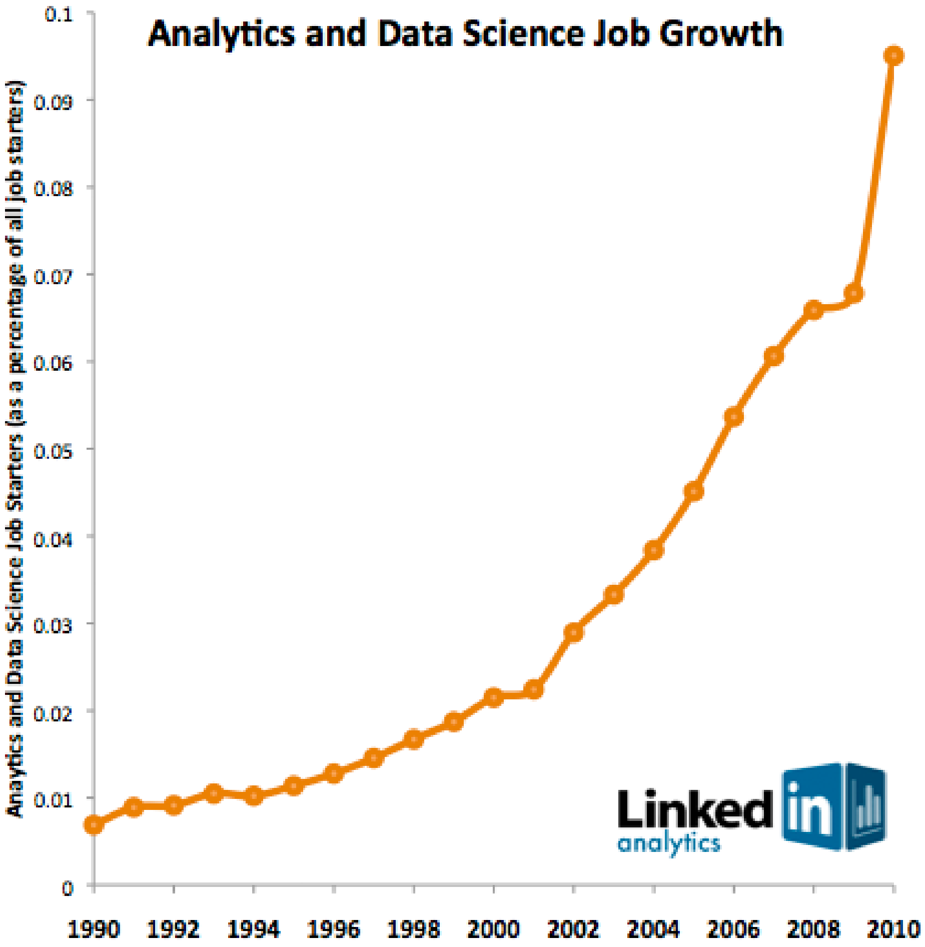
  > 150 million/day; 1700/second
- Text messages: 6.1 T/year; 876/person/year
- US cell calls: 2.2 T minutes/year;

  19 minutes/person/day

  ~ size of a YouTube

Sensors and The Internet of Things

Data Science Job Listing



Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

The World’s Most Powerful Data Scientists

“The success of companies like Google, Facebook, Amazon, and Netflix, not to mention Wall Street firms and industries from manufacturing to retail and healthcare, is increasingly driven by better tools for extracting meaning from very large quantities of data.”

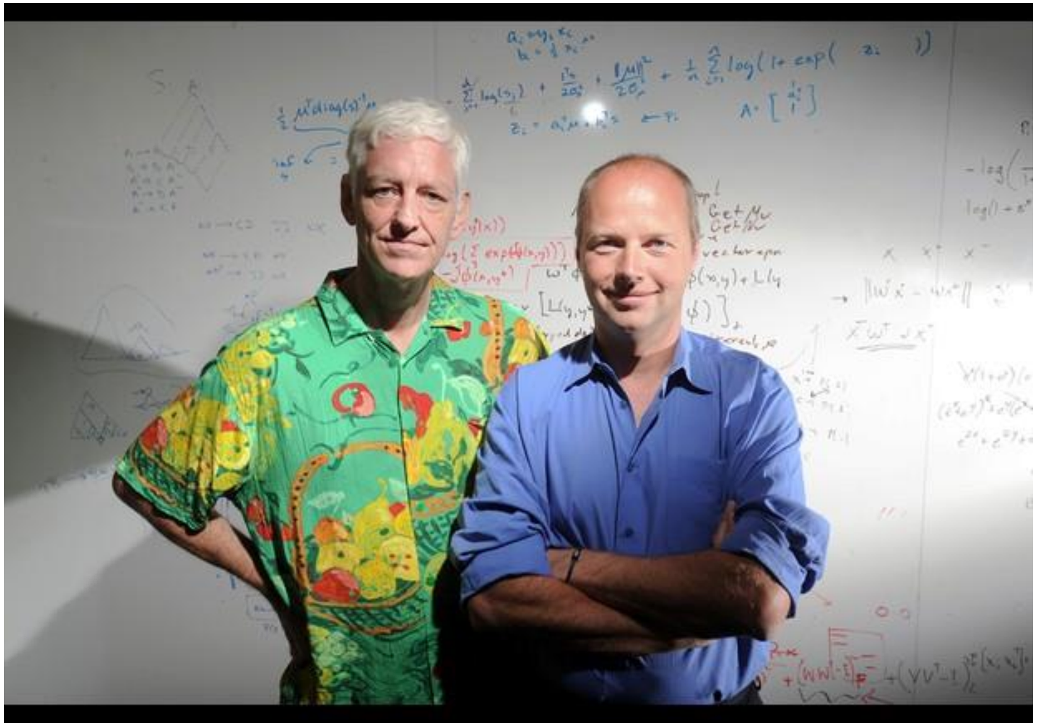
- Tim O'Reilly

#1 Larry Page, founder, Google

#2

- Jeff Hammerbacher, Chief Scientist, Cloudera
- DJ Patil, U.S. Chief Data Scientist

#3



- Peter Norvig, Director of Research, Google
- Sebastian Thrun, Professor, Stanford University

My Own List

Michael Jordan

Andrew Ng

Hilary Mason

Jeff Dean

Companies

Companies with the best data science teams:

LinkedIn, Facebook, Twitter, AT&T, Yahoo! Research, IBM Research, Google Research, Microsoft Research, Disney Research, HP Labs

Goldman Sachs, Renaissance Technologies, D.E. Shaw

SAS, MathWorks, Wolfram|Alpha

What skills you need to have?

1. Machine Learning
2. Statistics
3. Information Retrieval
4. Algorithms
5. Programming Languages
6. Distributed Systems
7. Hacking
8. Curiosity

Where to learn?

STANFORD
COMPUTER SCIENCE

Berkeley
UNIVERSITY OF CALIFORNIA

Santa Clara
University

Open Online Courses

coursera

edX

UDACITY

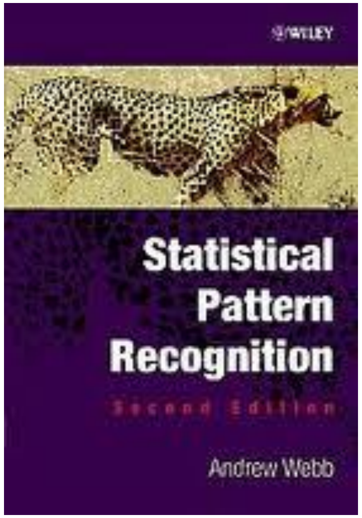
BigData University

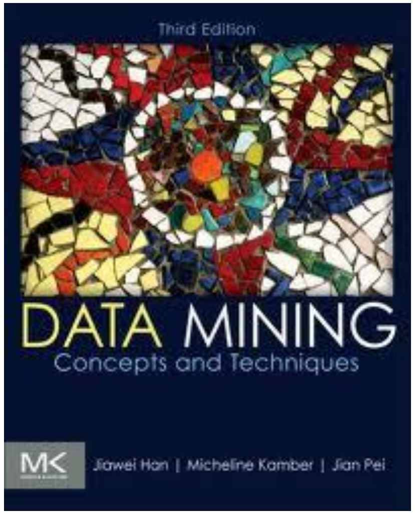
Learn about Machine Learning

- Prof. Andrew Ng’s course on Coursera

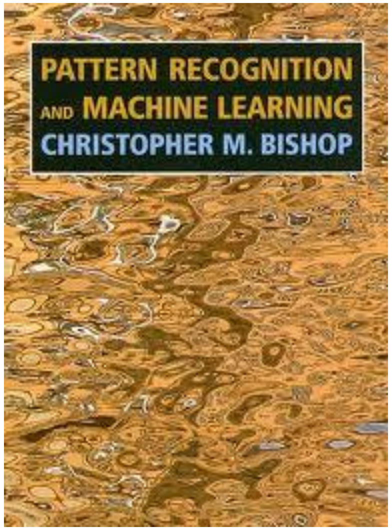
https://www.coursera.org/learn/machine-learning

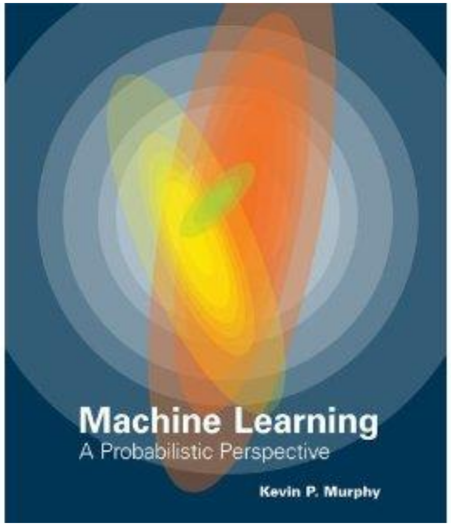
- Books

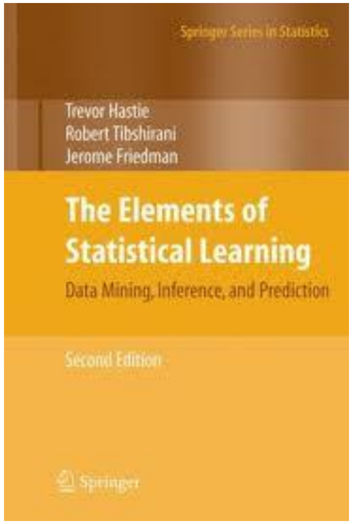




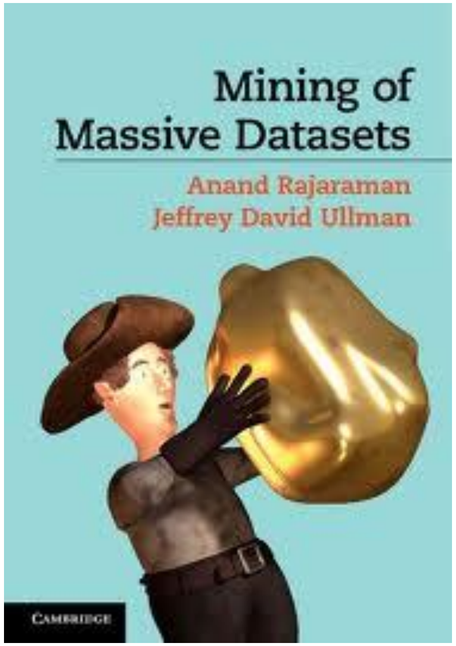
More Theoretical Books

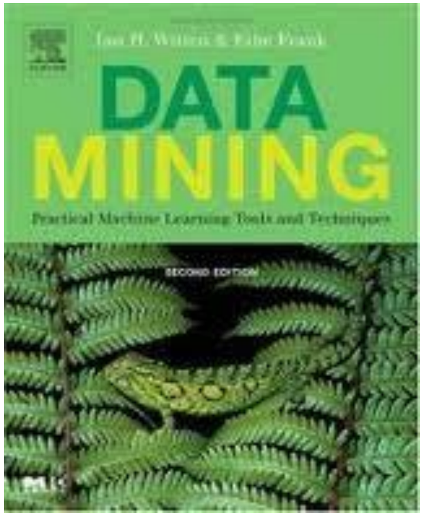





Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

More Practical Books




Ian H. Witten & Eibe Frank
DATA
MINING
Practical Machine Learning Tools and Techniques
SECOND EDITION

Learn about Distributed Systems

- MapReduce
- BigTable
- Google File System

Open Source Implementations

Apache Hadoop project

- MapReduce => Hadoop
- BigTable => HBase
- Google File System => HDFS (Hadoop Distributed File System)

A Great Book on MapReduce

Copyrighted Material

MORGAN & CLAYPOOL PUBLISHERS

Data-Intensive Text Processing with MapReduce

Jimmy Lin
Chris Dyer

Synthesis Lectures on Human Language Technologies

Graeme Hirst, Series Editor

Copyrighted Material

http://lintool.github.com/MapReduceAlgorithms/MapReduce-book-final.pdf

Open Source Projects

- Could9

http://lintool.github.com/Cloud9/

- Hadoop

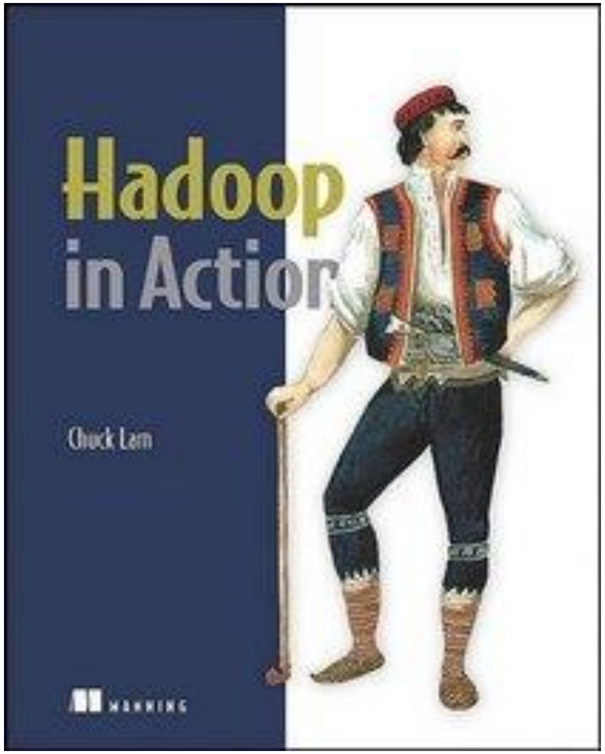
http://hadoop.apache.org/

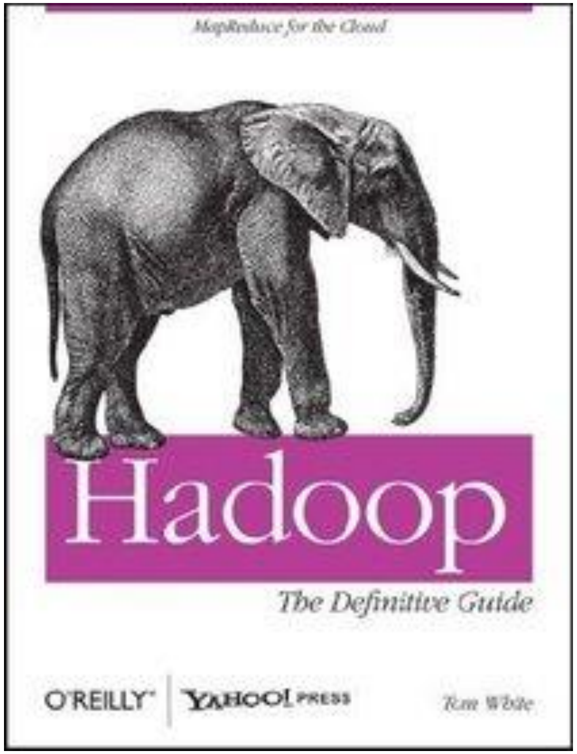
- Apache Spark

http://spark.apache.org/

Machine Learning algorithms implemented on Hadoop

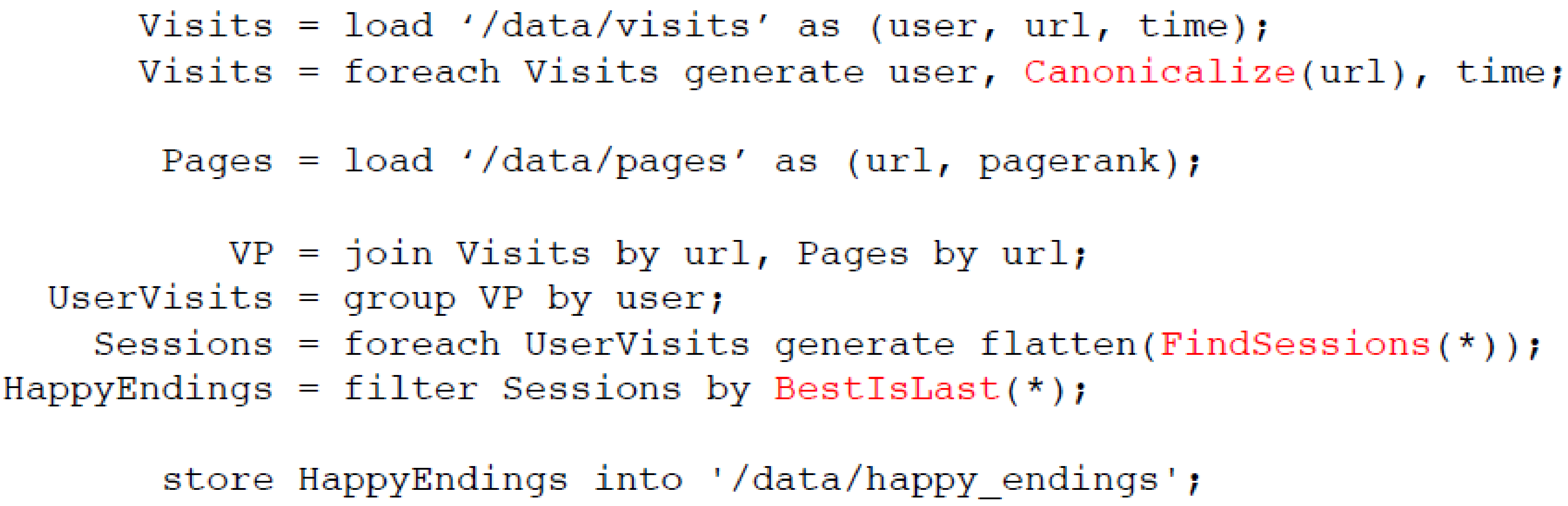
References on Hadoop





Apache Pig Latin

- A high-level programming language for creating MapReduce programs used with Hadoop



```
     Visits = load ‘/data/visits’ as (user, url, time);
     Visits = foreach Visits generate user, Canonicalize(url), time;

      Pages = load ‘/data/pages’ as (url, pagerank);

         VP = join Visits by url, Pages by url;
 UserVisits = group VP by user;
   Sessions = foreach UserVisits generate flatten(FindSessions(*));
HappyEndings = filter Sessions by BestIsLast(*);

      store HappyEndings into '/data/happy_endings';
```

http://pig.apache.org/

Developed At Yahoo!

Learn about Programming Languages

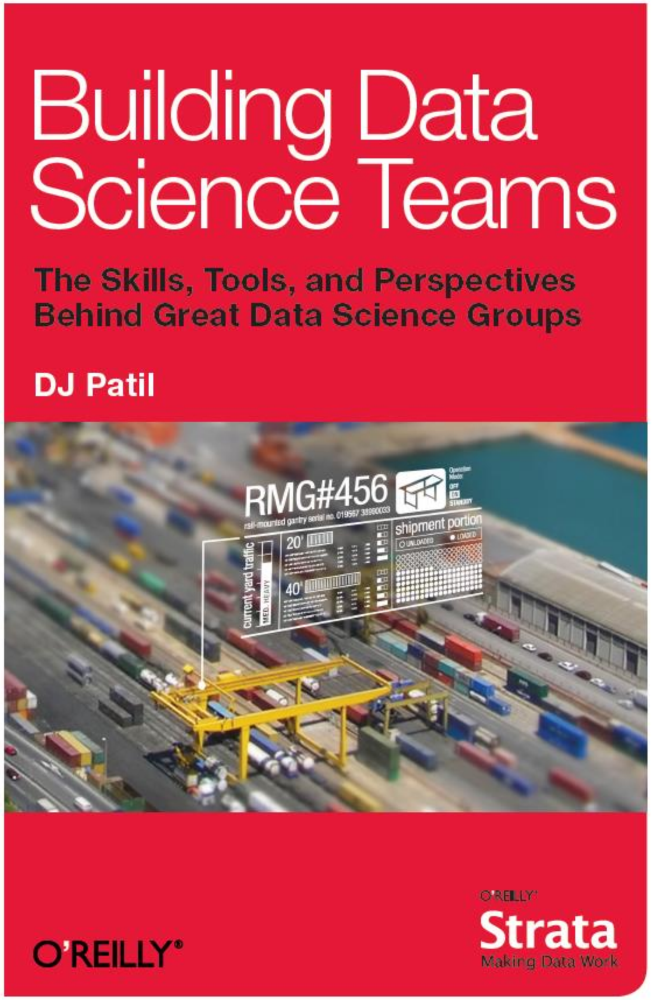
- Python or Ruby or Perl
- Java or C#
- C/C++
- R or Matlab
- Haskell or Erlang or Scheme or ML

Participate in Data Competitions

- Netflix prize II?
- Text REtrieval Conference (TREC)
- ACM KDD-CUP
- Kaggle
- Heritage Health Data Analysis Prize

  ($3 million prize, 2013)

Building Data Science Teams



Building Data Science Teams

The Skills, Tools, and Perspectives Behind Great Data Science Groups

DJ Patil

O'REILLY®

O'REILLY® Strata Making Data Work

Final Note

Curiosity