Practical Methods for the Analysis of "Big Data"

Philip A. Schrodt

The Pennsylvania State University
schrodt@psu.edu

Workshop at the Odum Institute
University of North Carolina, Chapel Hill
20-21 May 2013

Outline by Modules

- Monday morning: Overview and technology
  - 1a: Introduction
  - 1b: Hardware, software and ethics issues
- Monday afternoon: Natural language processing
  - 2a: processing text for classification and analysis
  - 2b: named entity recognition and automated coding
  - 2c: dictionary-based automated coding: TABARI
- Tuesday morning: supervised clustering; unsupervised topic models; sequence analysis
  - 3a: supervised clustering: SVM and neural networks
  - 3b: latent Dirichlet allocation models; sequence analysis methods
- Tuesday afternoon: Working with large-scale semi-structured data:
  - 4a: clustering and decision-trees
  - 4b: ensemble methods

Daily schedule

- 10-11:10 Lecture
- 11:10-11:20 Break
- 11:20-12:30 Lecture/Active learning

- 12:30-1:30 Lunch

- 1:30-2:40 Lecture/Active learning
- 2:40-2:50 Break
- 2:50-4 Lecture/Active learning

All slides will be available: you are not here to practice stenography

Topics: Module 1

- Overview of the Course
- The Big Picture
- Approach of the course
- Who is this guy?
- Canonical Definition of “Big Data”
- Your turn: Who are you?
- A sufficient—and very common—open source toolkit: R, Weka and Python
- A couple comments on programming
- Parallel Computing and Hadoop
- Legal and Ethical Issues

Primary drivers of change in social science research in the 21st century

Big data

- Effective use of high performance computing
- Wider range of analytical methods
- Wider range of data
- Increased concerns about privacy and intellectual property

Decentralized collaborative environments

- Open source / open access
- Kahn Academy / MOOC
- Increased interest in policy and private-sector applications of cutting-edge techniques

Data!


Science
11 February 2011 | $10
scientific
data
information
climate
research
science
new
analysis
visualization
researchers
access
many
human
knowledge
example
AAAS

Though this may be going a little far...

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

Computing Power

Control Data Corporation 3600
(ca.1965)
32 K (48-bit) RAM memory
1 processor
~1-million operations per second
Output: line printer

Computing Power

Control Data Corporation 3600
(ca.1965)
32 K (48-bit) RAM memory
1 processor
~1-million operations per second
Output: line printer

Penn State High Performance Computing Facility
15 cluster computers
100 to 2000 2.66 Ghz processors in each cluster
~50 Gb RAM accessible to each processor
130 Tb disk space
4 interactive visualization rooms

Computing Power

Control Data Corporation 3600
(ca.1965)
32 K (48-bit) RAM memory
1 processor
~1-million operations per second
Output: line printer

Penn State High Performance Computing Facility
15 cluster computers
100 to 2000 2.66 Ghz processors in each cluster
~50 Gb RAM accessible to each processor
130 Tb disk space
4 interactive visualization rooms

Motorola Razr
16 Gb RAM memory
Dual processor
~500-million operations per sec
540 x 860 color display

Open Source Software

MySQL

python

mozilla Firefox

ubuntu

source forge

T R

perl

Ruby PROGRAMMING Language

WordPress.com

APACHE HTTP SERVER

R

LibreOffice The Document Foundation

GNU run free run GNU

stackoverflow

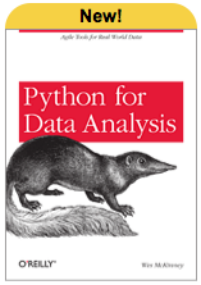
FREE SOFTWARE FOUNDATION

Approach of the course

- Breadth, not depth!
- a guide to vocabulary[ies], approaches and what you need to know
- you can then follow up on all of this material in detail. If I can look it up, you can look it up
- Emphasis on practical applications, not trendy theory: think of it as applied rather than academic.
- This being Odum, the focus is on the social sciences
- This is an experiment
- and it's the departure lounge, not the baggage claim

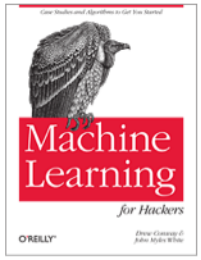
My emphasis is on methods that have been used for a while: note that almost always, the core method gets you 80% to 90% of what you are going to get, and then the little variations give you the final 5%, and not consistently

O'Rielly: Data Science



Python for Data Analysis: is concerned with the nuts and bolts of manipulating, processing, cleaning, and crunching data in Python. It is also a practical, modern introduction to scientific computing in Python, tailored for data-intensive applications. This is a book about the parts of the Python language and libraries you'll need to effectively solve a broad set of data analysis problems. This book is not an exposition on analytical methods using Python as the implementation language.

Ebook: $31.99 Add to Cart

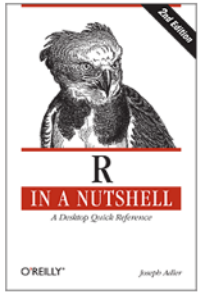


Machine Learning for Hackers: If you're an experienced programmer interested in crunching data, this book will get you started with machine learning—a toolkit of algorithms that enables computers to train themselves to automate useful tasks. Each chapter focuses on a specific problem in machine learning, such as classification, prediction, optimization, and recommendation.

Ebook: $31.99 Add to Cart

R Cookbook: Over 200 recipes for R users, ranging from the basic to the esoteric. Why re-invent the wheel? This collection of concise, task-oriented recipes makes you productive with R immediately, with solutions ranging from basic tasks to input and output, general statistics, graphics, and linear regression.

Ebook: $31.99 Add to Cart

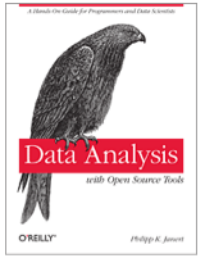


R in a Nutshell, 2nd Edition: The authoritative guide to what's become the de-facto standard for statistical programming. R in a Nutshell provides a quick and practical way to learn this increasingly popular open source language and environment.

Ebook: $35.99 Add to Cart

Bad Data Handbook: What is bad data? Some people consider it a technical phenomenon, like missing values or malformed records, but bad data includes a lot more. In this handbook, data expert Q. Ethan McCallum has gathered 19 colleagues from every corner of the data arena to reveal how they've recovered from nasty data problems.

Ebook: $31.99 Add to Cart

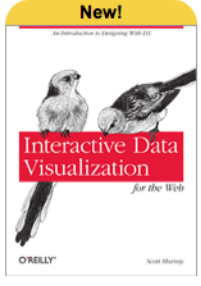


Data Analysis with Open Source Tools: A survey of data analysis from a practitioner – from histograms to machine learning,this book presents the tools you need to make sense with data. You'll learn how to look at data to discover what it contains, how to capture those ideas in conceptual models, and then feed your understanding back into the organization through business plans, metrics dashboards, and other applications.

Ebook: $31.99 Add to Cart

MapReduce Design Patterns: Each pattern is explained in context, with pitfalls and caveats clearly identified to help you avoid common design mistakes when modeling your big data architecture. This book also provides a complete overview of MapReduce that explains its origins and implementations, and why design patterns are so important. All code examples are written for Hadoop.

Ebook: $39.99 Add to Cart



Interactive Data Visualization for the Web: Create and publish your own interactive data visualization projects on the Web, even if you have no experience with either web development or data visualization. It's easy with this hands-on guide. You'll start with an overview of data visualization concepts and simple web technologies, and then learn how to use D3, a JavaScript library that lets you express data as visual elements in a web page.

Ebook: $23.99 Add to Cart

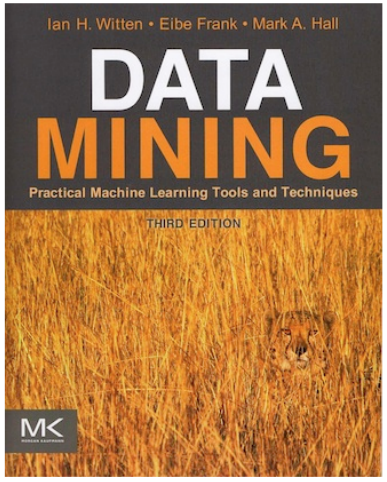
Weka Project: Data Mining

Machine Learning Group at the University of Waikato

Project Software **Book** Publications People Related

Data Mining:
Practical Machine Learning Tools and Techniques

We have written a companion book for the Weka software, now into its third edition, that describes the machine learning techniques that it implements and how to use them. It is structured into three parts. The first part is an introduction to data mining using basic machine learning techniques, the second part describes more advanced machine learning methods, and the third part is a user guide for Weka. The third edition was published in January 2011 by Morgan Kaufmann Publishers (ISBN: 978-0-12-374856-0). **Mark Hall** has joined **Ian Witten** and **Eibe Frank** as co-author for this edition, which has expanded to 629 pages.



Click here to order from Amazon.com

"If you have data that you want to analyze and understand, this book and the associated Weka toolkit are an excellent way to start."

-Jim Gray, Microsoft Research

"The authors provide enough theory to enable practical application, and it is this practical focus that separates this book from most, if not all, other books on this subject."

-Dorian Pyle, Director of Modeling at Numetrics

"This book would be a strong contender for a technical data mining course. It is one of the best of its kind."

-Herb Edelstein, Principal, Data Mining Consultant, Two Crows Consulting

"It is certainly one of my favourite data mining books in my library."

-Tom Breur, Principal, XLNT Consulting, Tiburg, Netherlands

Challenges

- Staying awake and alert for ten hours
  - Apparently the optimal way to do this is to shift approach every fifteen minutes and topic every forty, which we can't quite do
  - Fundamental active learning exercise: how is [whatever I'm talking about] relevant to *your* problems
- Retaining any of this
- Feel free to browse the web on the computers in front of you: do as I do

Data-oriented topics we will not be covering in any depth

- Social network analysis
- Geospatial
- Visualization
- New social media in technical detail

[Each of these would be well worth a course on their own.]

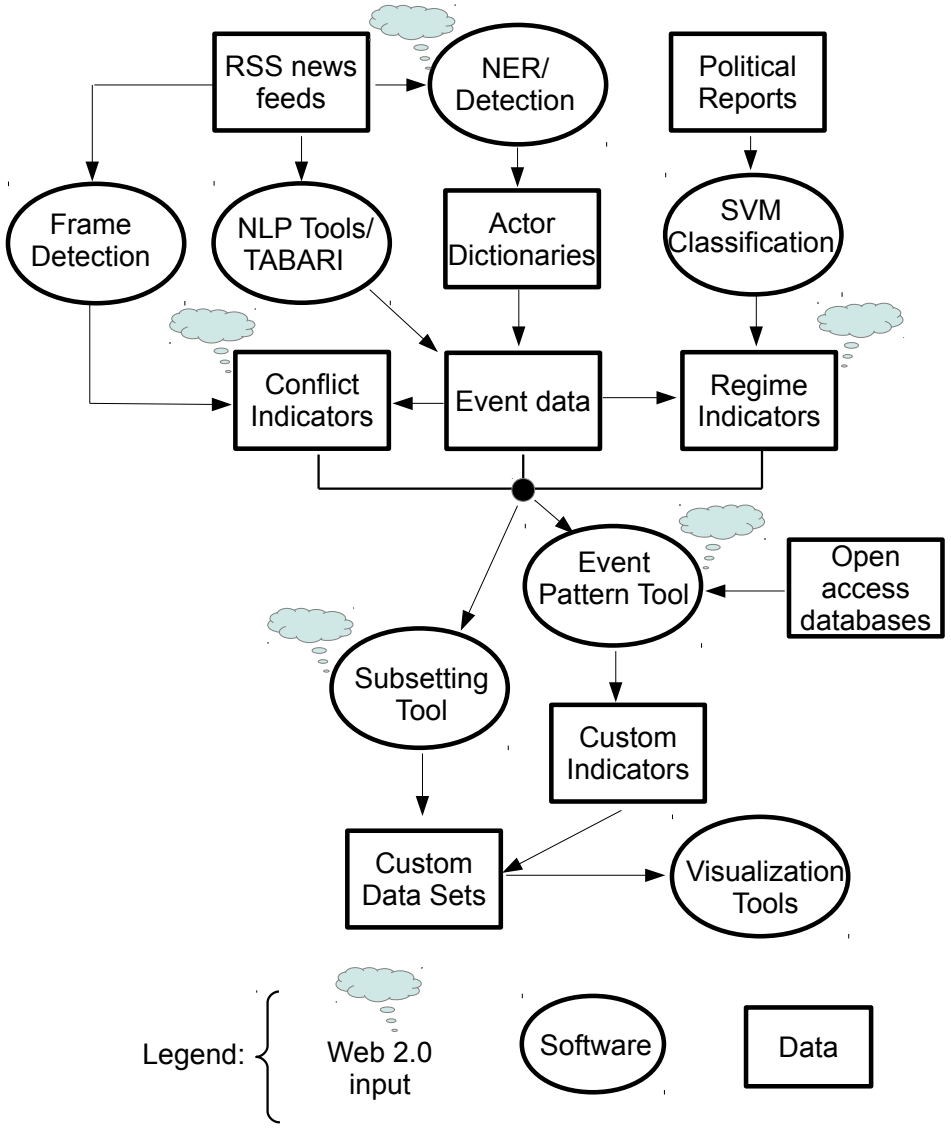
Machine learning topics we will not be covering in any depth

- An assortment of statistical methods such as logistic regression which you probably already know
- Genetic algorithms
- Expert systems
- Case-based reasoning
- Feature selection
- Sentiment analysis

Hey, why should I listen to you?

- B.A. and M.A. in mathematics; Ph.D. in political science
- 45 years of programming experience
- NSF panels
  - political science
  - methods, measurement and statistics
  - various interdiscplinary panels
- past president of Society for Political Methodology
- working with the machine learning methods since 1980s
  - *Patterns, Rules and Learning* [1995]
  - *Analyzing International Event Data* [2002]
- author of KEDS and TABARI automated event data coders
- political forecasting projects
  - Integrated Conflict Early Warning System
  - Political Instability Task Force

How I'm spending my summer "vacation": MADCOW



But a few things I don't really know

- My work has largely been in political science
- I mostly work with text
- I mostly work with time series, not hierarchical data
- I don't work with data on individuals, so I haven't personally dealt with IRB issues

Canonical Definition of “Big Data”

- Volume
- Variety
- Velocity

Common Examples of "Big Data" outside of the social sciences

- Sloan Sky Survey
- Human Genome project and genomics data more generally
- Very large scale customer databases: Walmart, Amazon
- Google searches
- Cell phone data
- Climate data and models
- Remote sensing

Volume

- NSM and other “spontaneously generated” data sources
- very large scale text DB if you can get hold of them, but we run into IP issues, whereas businesses generate these naturally
- otherwise the social sciences do not generally have the instrumentation (yet) to generate really high volume

Variety

- Which is to say, unstructured data, and we've got lots of that
- Nature language is the primary form

Velocity

- Near real-time web scraping, for example news feeds
- Twitter

Definition of "Big Data" from a social science statistics perspective

- Big: sample sizes are sufficiently large that pretty much everything is significant
- Heterogeneous: population consists of distinct sub-populations, and variables within those populations are correlated
- Sparse: values are missing or irrelevant on many variables, in contrast to a rectangular structure

Your turn: Who are you?

- Name, any other useful identifying data
- What sort of research do you do?
- What sort of data are you using?
- What sort of problems are you trying to solve that can't be done with the conventional methods?

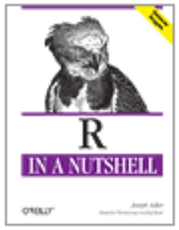
Break

Topics: Module 1

- Overview of the Course
- The Big Picture
- Approach of the course
- Who is this guy?
- Canonical Definition of “Big Data”
- Your turn: Who are you?
- A sufficient—and very common—open source toolkit: R, Weka and Python
- A couple comments on programming
- Parallel Computing and Hadoop
- Legal and Ethical Issues

- Open source
- Widely used in all approaches to statistical analysis and pattern recognition
- CRAN library provides almost immediate access to new methods
- Robust scripting capabilities; easily interfaces with C/C++ when needed
- Skill set is widely available

O'Rielly: R



R in a Nutshell
By Joseph Adler
Publisher: O'Reilly Verlag
Release Date: December 2010
Language: Deutsch
Other Languages: English

R Cookbook
By Paul Teetor
Publisher: O'Reilly Media
Release Date: March 2011
4.3

R Graphics Cookbook
By Winston Chang
Publisher: O'Reilly Media
Release Date: December 2012
4.6

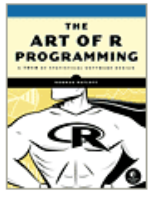
Exploring Everyday Things with R and Ruby
By Sau Sheong Chang
Publisher: O'Reilly Media
Release Date: June 2012
3.7

25 Recipes for Getting Started with R
By Paul Teetor
Publisher: O'Reilly Media
Release Date: January 2011
3.5

Data Mashups in R
By Jeremy Leipzig, Xiao-Yi Li
Publisher: O'Reilly Media
Release Date: March 2011
3.3

Parallel R
By Q. Ethan McCallum, Stephen Weston
Publisher: O'Reilly Media
Release Date: October 2011

Learning R
By Richard Cotton
Publisher: O'Reilly Media
Release Date: September 2013



The Art of R Programming
By Norman Matloff
Publisher: No Starch Press
Release Date: September 2011
2.3

The Essential R Reference
By Mark Gardener
Publisher: Wiley
Release Date: November 2012

Data Analysis with R
By Garrett Grolemund
Publisher: O'Reilly Media
Release Date: October 2013

Weka Project

Machine Learning Group at the University of Waikato

Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.

Getting started

- **Requirements**
- **Download**
- **Documentation**
- **FAQ**
- **Getting Help**

Further information

- **Citing Weka**
- **Datasets**
- **Related Projects**
- **Miscellaneous Code**
- **Other Literature**

Developers

- **Development**
- **History**
- **Subversion**
- **Contributors**

Weka features

Weka's main user interface is the *Explorer*, but essentially the same functionality can be accessed through the component-based *Knowledge Flow* interface and from the command line. There is also the *Experimenter*, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The *Explorer* interface features several panels providing access to the main components of the workbench:

- The *Preprocess* panel has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a so-called *filtering* algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.
- The *Classify* panel enables the user to apply classification and regression algorithms (indiscriminately called *classifiers* in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).
- The *Associate* panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.
- The *Cluster* panel gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
- The *Select attributes* panel provides algorithms for identifying the most predictive attributes in a dataset.
- The *Visualize* panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

Source: http://en.wikipedia.org/wiki/Weka_(machine_learning)

The Problem with Programmers

Old model

> *"We'll just hire a programmer because that will be more efficient than doing it ourselves"*

Reality

- Computer science departments and ExxonMobil can't find enough programmers either
- You take a serious efficiency hit in trying to explain what you want done
- You may take a serious efficiency hit in not doing the task in the best way—NLP (and statistics) are specialized subfields
- Programmers frequently are trained to focus on GUIs, (e.g. Java) which usually just get in the way in research computing

The Problem with Programmers

Old model

> *"We'll just hire a programmer because that will be more efficient than doing it ourselves"*

Reality

- Computer science departments and ExxonMobil can't find enough programmers either
- You take a serious efficiency hit in trying to explain what you want done
- You may take a serious efficiency hit in not doing the task in the best way—NLP (and statistics) are specialized subfields
- Programmers frequently are trained to focus on GUIs, (e.g. Java) which usually just get in the way in research computing

Why people don't want to be programmers

- Programming is a craft, not a science
  - "Between the mathematics that make [the computer] theoretically possible and the electronics that makes it practically feasible lies the programming that makes it intellectually, economically and socially useful. Unlike the extremes, the middle remains a craft, technical rather than technological, mathematical only in appearance."
    Michael Sean Mahoney, Histories of Computing (Harvard University Press)
  - Practice, practice, practice
- Programmer efficiency varies by a factor of 10 to 20, which can be very demoralizing
- Popular perception of programmers

Why Python?

- Open source (of course...tools want to be free...)
- Standardized across platforms and widely available/documented
- Automatic memory management (unlike C/C++)
- Generally more coherent than perl, particularly when dealing with large programs
- Text oriented rather than GUI oriented (unlike Java)
- Extensive libraries but these are optional (unlike Java): seems to be generating very substantial network effects
- C/C++ can be easily integrated by high-performance applications
- Tcl can be used for GUI

Introduction to Computer Science

UDACITY

Beginner

INSTRUCTORS

David Evans

Take the Class

Class Summary

In this course you will learn key concepts in computer science and learn how to write your own computer programs in the context of building a web crawler.

What Should I Know?

There is no prior programming knowledge needed for this course. Beginners welcome!

What Will I Learn?

At the end of this course, you will have learned key concepts in computer science and enough Python programming to be able to write programs to solve problems on your own. This course will prepare you to move on to more intermediate-level courses in CS.

Syllabus

Lesson 1: How to Get Started
Your first program: Extracting a link

Lesson 2: How to Repeat
Procedures, decisions (if), loops; finding all of the links on a page

Additional Lesson: How to Solve Problems
Universal techniques for solving programming problems

Lesson 3: How to Manage Data
Lists; crawling the web

Lesson 4: Responding to Queries
Complex data structures; building a reverse index to do searches; networks

Lesson 5: How Programs Run
Reasoning about cost; hash tables (Dictionary)

Lesson 6: How to Have Infinite Power
Recursive definitions; ranking search results

Lesson 7: Where to Go from Here

Course Instructors

David Evans
Instructor

David Evans is a Professor of Computer Science at the University of Virginia where he teaches computer science and leads research in computer security. He is the author of an introductory computer science textbook and has won Virginia's highest award for university faculty. He has PhD, SM, and SB degrees from MIT.

O'Rielly: Python

Python Cookbook, 3rd Edition
By David Beazley, Brian K. Jones
Publisher: O'Reilly Media
Release Date: May 2013
5.0
Other Editions: 1st Edition, 2nd Edition

Learn More

Programming Computer Vision with Python
By Jan Erik Solem
Publisher: O'Reilly Media
Release Date: June 2012
3.8

Learn More

Python for Data Analysis
By Wes McKinney
Publisher: O'Reilly Media
Release Date: October 2012
4.2

Learn More

Python and AWS Cookbook
By Mitch Garnaat
Publisher: O'Reilly Media
Release Date: October 2011

Learn More

Test-Driven Web Development with Python
By Harry Percival
Publisher: O'Reilly Media
Release Date: March 2013
4.0

Learn More

MongoDB and Python
By Niall O'Higgins
Publisher: O'Reilly Media
Release Date: September 2011
3.5

Learn More

Learning Python, 5th Edition
By Mark Lutz
Publisher: O'Reilly Media
Release Date: June 2013

Learn More

Head First Python
By Paul Barry
Publisher: O'Reilly Media
Release Date: November 2010
3.4

Learn More

Programming Python, 4th Edition
By Mark Lutz
Publisher: O'Reilly Media
Release Date: December 2010
4.1
Other Editions: 1st Edition, 2nd Edition, 3rd Edition

Learn More

Programmieren lernen mit Python
By Allen B. Downey
Publisher: O'Reilly Verlag
Release Date: January 2013
Language: Deutsch

Learn More

Think Python
By Allen B. Downey
Publisher: O'Reilly Media
Release Date: August 2012
3.7

Learn More

High Performance Python
By Micha Gorelick, Andy R. Terrel
Publisher: O'Reilly Media
Release Date: October 2013

Learn More

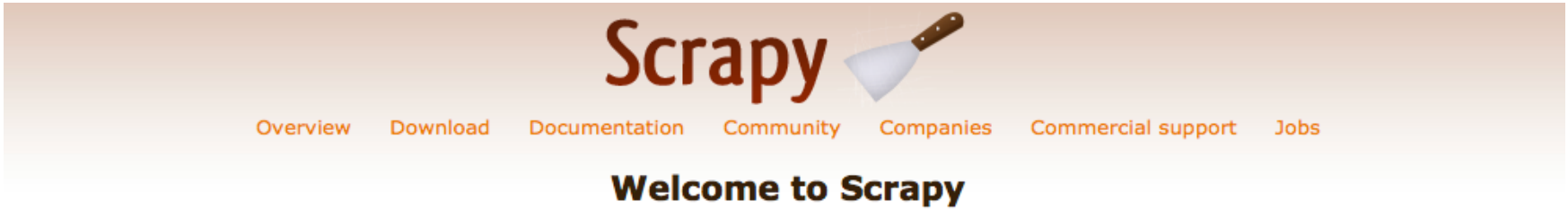
Python for Kids
By Jason R. Briggs
Publisher: No Starch Press
Release Date: November 2012
4.3

Learn More

Introducing Python
By Bill Lubanovic
Publisher: O'Reilly Media
Release Date: November 2013

Learn More

Python: scrapy



Welcome to Scrapy

What is Scrapy?

Scrapy is a fast high-level screen scraping and web crawling framework, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing.

Features

Simple
Scrapy was designed with simplicity in mind, by providing the features you need without getting in your way

Productive
Just write the rules to extract the data from web pages and let Scrapy crawl the entire web site for you

Fast
Scrapy is used in production crawlers to completely scrape more than 500 retailer sites daily, all in one server

Extensible
Scrapy was designed with extensibility in mind and so it provides several mechanisms to plug new code without having to touch the framework core

Portable, open-source, 100% Python
Scrapy is completely written in Python and runs on Linux, Windows, Mac and BSD

Batteries included
Scrapy comes with lots of functionality built in. Check this section of the documentation for a list of them.

Well-documented & well-tested
Scrapy is extensively documented and has an comprehensive test suite with very good code coverage

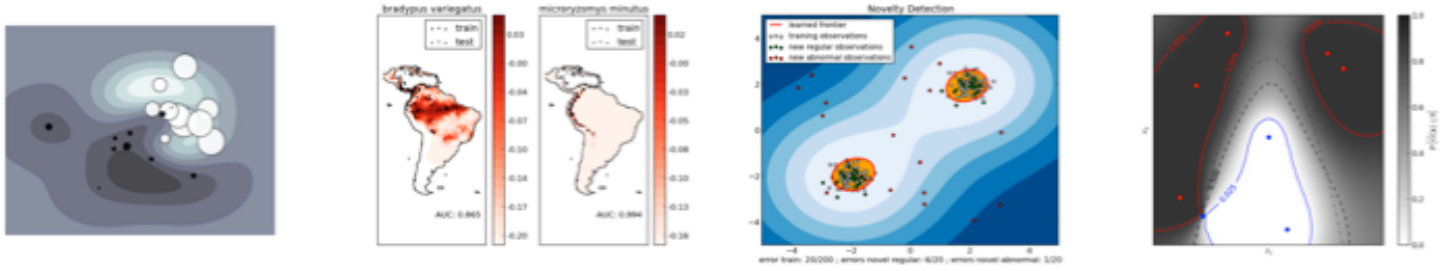
Healthy community
1,500 watchers, 350 forks on Github (link)
700 followers on Twitter (link)
850 questions on StackOverflow (link)
200 messages per month on mailing list (link)
40-50 users always connected to IRC channel (link)

Commercial support
A few companies provide Scrapy consulting and support

Still not sure if Scrapy is what you're looking for?. Check out Scrapy at a glance.

Python: scikit-learn

scikit-learn: machine learning in Python



Easy-to-use and general-purpose machine learning in Python

Scikit-learn integrates **machine learning** algorithms in the tightly-knit scientific **Python** world, building upon numpy, scipy, and matplotlib. As a machine-learning module, it provides versatile tools for data mining and analysis in any field of science and engineering. It strives to be **simple and efficient**, accessible to everybody, and reusable in various contexts.

Supervised learning

Support vector machines, linear models, naive Bayes, Gaussian processes...

Unsupervised learning

Clustering, Gaussian mixture models, manifold learning, matrix factorization, covariance...

And much more

Model selection, datasets, feature extraction... ***See below.***

License: Open source, commercially usable: **BSD license** (3 clause)

Source: http://scikit-learn.org/stable/

Python: other common libraries

- nltk: Natural language Toolkit
- numpy: Numerical scientific processing
- SciPy: statistical functions
- TkInter: graphical user interface
- VTK: visualization toolkit

Other computer languages*

- C/C++: still very widely used, particularly when code needs to be very fast. Frequently used to write libraries, e.g. R and Python
- FORTRAN: one of the earliest languages; highly optimized numerical libraries
- perl: very flexible—as in "too flexible"—character processing language, though largely displaced by Python
- php: interfaces web pages with databases
- Java: standardized language widely used in business application; extensive graphical libraries

With web resources, O'Reilly *Cookbooks*, and the increasing standardization of most control structures, you can pick up new languages fairly quickly once you've learned a couple. Though fundamentally, it is still practice, practice, practice.

* And don't forget *whitespace*, which consists entirely of invisible characters (space, tab, linefeed).

http://en.wikipedia.org/wiki/Whitespace_(programming_language)

Why Unix?

- Stable and compact
- About 20 commands will do almost everything you need; they haven't changed in 30 years
- Core of all major operating systems except Windows
- Linux, OS-X, Android
- Most functions are identical across OS-X and Linux
- Standardized set of compilers, so identical code can run on multiple systems.
- "make" command will compile code on any machine
- Used in most cluster computers
- Research software is more likely to be written for Unix
- Command line is more efficient than mouse/menus in advanced applications

Topics: Module 1

- Overview of the Course
- The Big Picture
- Approach of the course
- Who is this guy?
- Canonical Definition of “Big Data”
- Your turn: Who are you?
- A sufficient—and very common—open source toolkit: R, Weka and Python
- A couple comments on programming
- Parallel Computing and Hadoop
- Legal and Ethical Issues

Parallel computing options

- Dedicated cluster computers
- Networked clustering environments such as Beowulf (Linux) and XGrid (Apply, proprietary)
- Cloud solutions such as Amazon “Elastic Compute Cloud” (EC2) and Google Cloud
- National supercomputer centers
- NSF Teragrid

Go to: Hadoop.Presentation.pdf

Source: Ben Bagozzi and John Beieler, Penn State

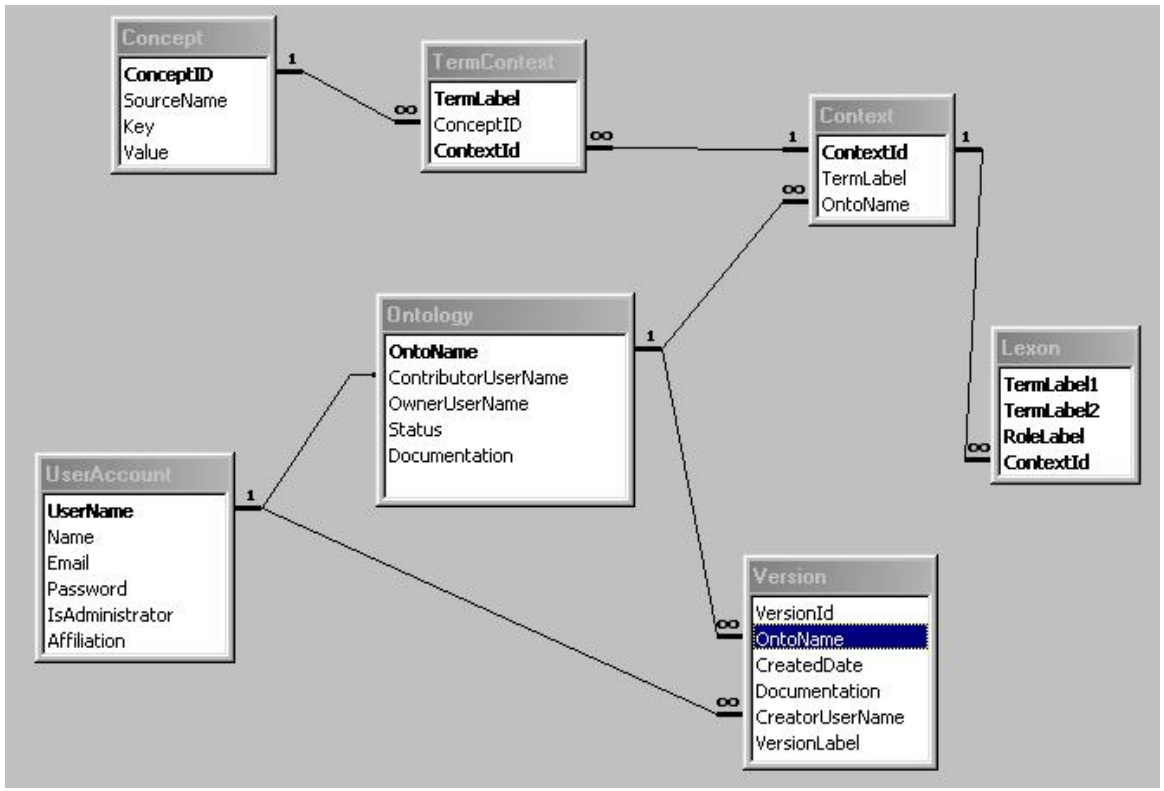
Some additional thoughts on parallel computing

- For many text processing problems, "parallelism on the cheap" consists of simply splitting your files. This is a simple form of “data parallelism.”
- “Task parallelism” is more difficult to program, though to some extent it can be done—and in contemporary processors, is frequently done (“threads”)—automatically
- “Message passing interface” [MPI] is the *de facto* open standard set of simple routines for implementing parallel programs, originally implemented in Fortran and C and since extended to Java and Python

Per the Bagozzi/Beieler experiments, make sure that you actually need parallel implementation before investing a lot of effort in these, particularly when customized programming is involved. Contemporary processors already implement a great deal of parallel processing, and the processors in cluster computers are usually inexpensive, not state-of-the-art.

Aside: Relational Database

A relational database is a database that has a collection of tables of data items, all of which is formally described and organized according to the relational model. [https://en.wikipedia.org/wiki/Relational_database]



Source: http://aldwyn-digitalglossary.blogspot.com/2010/12/relational-database.html

Relational Database

- Basic concept in an RDB is that information is stored once and can be accessed by multiple records
- RDBs also tend to contain heterogeneous data—different cases contain different variables—rather than rectangular data/tables found in most statistics programs
- Most common contemporary implementation is the open source "mySQL", after the ISO standard query language "SQL" ["structured query language"]
- Very mature technology; original development was in the 1970s

Topics: Module 1

- Overview of the Course
- The Big Picture
- Approach of the course
- Who is this guy?
- Canonical Definition of “Big Data”
- Your turn: Who are you?
- A sufficient—and very common—open source toolkit: R, Weka and Python
- A couple comments on programming
- Parallel Computing and Hadoop
- Legal and Ethical Issues

Security Issues in Big Data

Privacy

- Legal requirements, which are evolving and not always consistent
- Ethical requirements
- Keeping current with new challenges such as deanonymization

Intellectual property

- Adhering to legitimate claims while defending legitimate rights of educational and research institutions.
- Information does not necessary want to be free, but neither is all information privately owned.
- Increased interest in policy and private-sector applications of cutting-edge techniques

System security

- Keeping current with best practices, not urban legends
- Convincing users to become part of the solution rather than challenging them to cleverly evade you (which they will). Remember the story of Liu Bang (ca. 210 BCE)
- Accurately assessing the degree of threat to your data

Just a reminder...

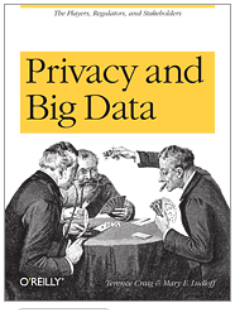
I am not a lawyer!

You knew that...

Problems with personal data

- If you are wondering whether you need IRB approval, you probably do.
- Subpopulations may be so small that individuals can be identified
- Multiple datasets, while individually anonymous, may contain sufficient information to identify individuals when commonly keyed
  - There are apparently quite a few very clever ways of doing this
- Analysis of large scale data may be reveal patterns of characteristics the individual does not want known
  - Apparently non-urban-legend story of purchasing data leading to identification of a pregnant teenager...before she told her parents
- Informed consent
  - Social media for individuals under 18?
- Securing large scale data against unauthorized use
- US vs. European approaches to privacy

Browse Contents



Google Preview Larger Cover

Privacy and Big Data

The Players, Regulators, and Stakeholders

By Terence Craig, Mary E. Ludloff

Publisher: O'Reilly Media
Released: September 2011
Pages: 108

5.0 Tweet 123 +1 11 Like 78

Read 1 Review | Write a Review

Description | Table of Contents | Product Details | About the Author

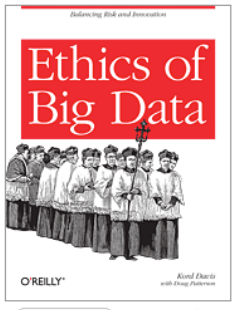
Much of what constitutes Big Data is information about us. Through our online activities, we leave an easy-to-follow trail of digital footprints that reveal who we are, what we buy, where we go, and much more. This eye-opening book explores the raging privacy debate over the use of personal data, with one undeniable conclusion: once data's been collected, we have absolutely no control over who uses it or how it is used.

Personal data is the hottest commodity on the market today—truly more valuable than gold. We are the asset that every company, industry, non-profit, and government wants. Privacy and Big Data introduces you to the players in the personal data game, and explains the stark differences in how the U.S., Europe, and the rest of the world approach the privacy issue.

You'll learn about:

- **Collectors:** social networking titans that collect, share, and sell user data
- **Users:** marketing organizations, government agencies, and many others
- **Data markets:** companies that aggregate and sell datasets to anyone
- **Regulators:** governments with one policy for commercial data use, and another for providing security

Browse Contents



Google Preview Larger Cover

Ethics of Big Data

Balancing Risk and Innovation

By Kord Davis

Publisher: O'Reilly Media
Released: September 2012
Pages: 82

0.0 Tweet 0 +1 13 Like 30

Write a Review

Description | Table of Contents | Product Details | About the Author

What are your organization's policies for generating and using huge datasets full of personal information? This book examines ethical questions raised by the big data phenomenon, and explains why enterprises need to reconsider business decisions concerning privacy and identity. Authors Kord Davis and Doug Patterson provide methods and techniques to help your business engage in a transparent and productive ethical inquiry into your current data practices.

Both individuals and organizations have legitimate interests in understanding how data is handled. Your use of data can directly affect brand quality and revenue—as Target, Apple, Netflix, and dozens of other companies have discovered. With this book, you'll learn how to align your actions with explicit company values and preserve the trust of customers, partners, and stakeholders.

- Review your data-handling practices and examine whether they reflect core organizational values
- Express coherent and consistent positions on your organization's use of big data
- Define tactical plans to close gaps between values and practices—and discover how to maintain alignment as conditions change over time
- Maintain a balance between the benefits of innovation and the risks of unintended consequences

Intellectual Property

Copyright law is generally open to "fair use" in research and education (but not commercial applications). However, institutional contracts with data providers are more limited

- Information does not necessarily want to be free

A lot of the legal issues, particularly involving content on the web, are still very open

- You probably do not want to be a test case: due to various laws passed to benefit the entertainment industry, these can be prosecuted as criminal violations, not just civil violations. That means jail.
- Just because someone claims IP rights doesn't mean they actually have those rights
- You still probably do not want to be a test case
- Public institutions have considerable protection due to sovereign immunity, though many have been wimps in asserting this.
- Educational institutions have very different policies: check locally

Your turn: ethical issues

Consider our original V3 definition of "Big Data": volume, variety and velocity. Are any of these introducing new issues in your field—or in your own work—that were not present ten years ago.

Anyone have some interesting stories of where IRBs (or other university institutions) have had a difficult time dealing with these? Anyone with experience with new policies? Did those make sense?

How should we go about dealing with these [rapidly] evolving issues as professionals?

Let's eat!

End of Module 1