Resources

2014 HealthDataPalooza

The public reaction to the recent CMS data dump has been met with mixed reactions. While transparency has generally been valued, many are concerned about the ethical implications of providing so much information to the general public without the means necessary to understand, critically evaluate, or use it meaningfully. To us, that just sounds too much like the doctor telling the patient "don't bother with independent research, you wouldn't understand."

Answering the question "How can the healthcare consumer benefit" from such data dumps is two-fold. First, as data scientists, we have an obligation to present information in an accurate manner. Unfortunately, this is often not the case, and wildly popular data visualizations from well known publications are still missleading. The statistical background necessary for critical analysis is lacking in our community. The first step to helping the public benefit is to leverage data dumps like these to spur interest in statistics and consumer education. Second, it's a great platform to dissect analysis methodology, and walk consumers through the process so that they may start with their own analysis. True data transparency is more than just raw numbers. Then can we start talking about visualizations.

Education is more democratic now than ever, and information is more accessible, to the point of overload. Neat data visualizations will only go so far if the consumer can't evaluate them critically. In this spirit, we are blogging real time about our exploration, and curating a list of resources for those interested in understanding the foundation of data science. This is an ongoing project, beyond HDPalooza. If you know of good resources, please let us know!

Statistics

Data scientists are often described as those who are better at software engineering than a statistician and better at statistics than any software engineer. As such, statistical inference underpins much of the theory behind data analysis and a solid foundation of statistical methods and probability serves as a stepping stone into the world of data science. There is an abundance of information freely availible for those needing a refresher in stats and statistical programming.

Courses:

edx: Introduction to Statistics: A basic introductory statistics course.

Stanford: Statistical Learning: Introductory-level course with a focus on regression and classification methods.

Coursera: Statistics, Making sense of Data: A applied Statistics course that teaches the complete pipeline of statistical analysis.

MIT: Statistical Thinking and Data Analysis: Introduction to probability, sampling, regression, common distributions, and inference.

edx: Explore Statistics with R: Followup to edx's intro course, this gives a deaper overview of statistical programming with R.

Books:

A trustworthy, practical, well structured statistical reference book is a great supplement for online dynamic information, particularly when it comes to the fundamentals. These may be particularly useful if you already have some knowledge of the subject or just need to fill in some gaps in your understanding. Work through one, mark it up, break the binding and show print some love.

Think Bayes: An simple introduction to Bayesian Statistics with Python code examples.

OpenIntro: Statistics: Introductory text book with supplementary exercises and labs in an online portal.

Lecture notes for Introduction to Probability: Compiled lecture notes of above textbook, complete with exercises.

Introduction to Probability: Textbook for Berkeley's Stats 134 class, an introductory treatment of probability with complementary exercises.

O'Reilly: Think Stats: An introduction to Probability and Statistics for Python programmers.

O'Reilly: R Cookbook: A collection of R "recipes" to jump start statistical analysis.

Programming

Programming serves a dual function when it comes to data science: it is first and foremost the tool used for statistical analysis, but it is also the foundation for communication and education. Python is a great multi purpose language capable of handling both analysis and platform develovelopment, but familiarity with other languages and frameworks strengths and weaknesses can help narrow down the decision making process for more advanced operations.

While R is the de facto standard for statistical programming, it has quite a high learning curve. For those new to both programming and statistical analysis, Python is well suited for a number of specific problem domains, in addition to general web development. Many of the functionalities of R can be replicated with NumPy, SciPy, IPython, Matplotlib, Pandas and many more statistical libraries.

The "learn to code" movement is a booming business, and as such resources for those just starting out can be overwhelming, but below (no particular order) are some of the more tried and true books, lessons, tutorials and resources focused on programming basics.

Rosetta Code: A programming chrestomathy site where solutions to the same task are presented in as many different languages as possible.

Learn Code The Hard Way: Lessons for Python, Ruby, C, SQL, Regular Expression and an active support community.

How to Think Like a Computer Scientist: Ebook programming resource focusing on problem sovling with Python.

Getting Started with Data Science: A data science programming tutorial using Python and real data from NTSB.

Udacity: CS101: Popular MOOC intro to computer science course, widely considered the most comprehensive CS intro MOOC.

Crowdsourced CS Courses: A large list of beginner CS online courses, resources and tutorials crowdsourced by Reddit.

Discouraged Resources: Sometimes knowing what resources to avoid is just as important as knowing the good ones.

GoogleDev: Intro to R: A 21-video series introducing users to basic R concepts, reshaping data and function writing.

Coursera: Computing for Data Analysis: Part of the Data Science certificate program, focuses on data analysis fundamentals in R.

Julia: A LLVM-based just-in-time compiled programming language for technical computing, with impressive benchmarks.

ROOT: A Data Analysis Framework For those comfortable with compiled languages, the ROOT framework allows for largescale data analysis in C++.

Python Opensource Library: A large collection of ebooks from Google, O'Reilly, PSF, and developers covering all aspects of development.

O'Reilly: Python for Data Analysis: Bridges the gap between programming and stastical analysis.

Data Visualization

Now the fun part. D3js is one of the most popular options for data visualisations, but there are many more resources available. Keep in mind color schemes and scales, normalization and labels as they can be inadvertently misleading. Additionally, R can map many of these visualisations with fewer dependencies.

FlowingData: Nathan Yau's tutorial offers examples of basic plotting in R, including downloadable source code.

HighCharts: A charting library written in pure HTML5/JavaScript, offering intuitive, interactive charts for web applications.

D3.js: An small, flexible and efficient library to create and manipulate interactive documents based on data.

Circos: A software package in Perl for visualizing data in a circular layout.

CartoDB: A web service for mapping, analyzing and building applications with data.

MapBox: A web platform for hosting custom designed map tiles and a set of open source tools to produce them.

Raphael: A small JavaScript library that should simplify your work with vector graphics on the web.

PaperJS: A vector graphics scripting framework in a well designed, consistent and clean programming interface.

Polymaps: A library for making dynamic, interactive maps with image- and vector-based tiles.

DataWraongler: An interactive web application for data cleaning and transformation.

Flare: A set of software tools for creating rich interactive data visualizations in ActionScript.

Healthcare Data

This section may never be "done." Listed here is some base information one may need to get up and running in the healthcare analytics space, and familiar with the many standards, data types, regulations, and policies in place.

Hacking Healthcare: A candid assessment of the US healthcare system as it ramps up its use of healthIT to comply with regualtions.

The Information Technology Fix for Health: A focus on how personal data fits into the enterprise healthcare space.

HL7: Health Level 7 - Provides the standard for exchanging clinical information in pipe delimited format.

HL7 CDA: Clinical Document Architecture – provides an exchange model (XML-based) for clinical documents (such as discharge summaries and progress notes); recently known as the Patient Record Architecture (PRA).

CCR: Continuity of Care Record – responds to the need to organize and make transportable a set of basic information about a patient's health care that is accessible to clinicians and patients.

X12: Provides for electronic exchange of business transactions-electronic data interchange (EDI). The American National Standards Institute (ANSI) chartered the Accredited Standards Committee (ASC) X12 to develop uniform standards.

CCOW: Clinical Context Management Specification – allows clinical applications to share information at the point of care.

NCPDP: National Council for Prescription Drug Programs – governs prescription transactions.

CPT: Current Procedural Terminology - Registered Tradement of the AMA, used to report medical procedures and services under public and private health insurance programs.

Direct Project: Specifies a simple, secure, scalable, standards-based way for participants to send authenticated, encrypted health information directly to known, trusted recipients over the Internet.

Health Information Exchanges: Allows health care professionals and patients to appropriately access and securely share a patient's vital medical information electronically between otherwise unrelated vendors.

Blue Button: The ability for patients to get records in a human-readable and machine-readable format; and to send them where they choose. Several different flavors, BB+, BlueButtonDirect, Push, Pull, API, etc.

SNOMED: Systematized Nomenclature of Medicine Clinical Terms – provides comprehensive computerized clinical terminology covering clinical data for diseases, clinical findings, and procedures.

ICD-9/10: International Classification of Diseases - The cornerstone of classifying diseases, injuries, health encounters and inpatient procedures in morbidity settings.

HCPCS: Healthcare Common Procedure Coding System - Set of health care procedure codes based on the American Medical Association's Current Procedural Terminology (CPT)

DICOM: Digital Imaging and Communications in Medicine - Provides for handling, storing, printing, and transmitting information in medical imaging.

Meaningful Use: The Medicare and Medicaid EHR Incentive Programs provide financial incentives for the "meaningful use" of certified EHR technology to improve patient care.

HITECH Act/HIPAA Omnibus: Enacted as part of the American Recovery and Reinvestment Act of 2009, strengthens the privacy and security protections for health information established under the Health Insurance Portability and Accountability Act of 1996 (HIPAA).

ACOs: Accountable Care Organizations - are groups of doctors, hospitals, and other health care providers, who come together voluntarily to give coordinated high quality care to their Medicare patients.

Miscellaneous Tools

One of the great things about statistical programming in recent years is the abundance of tools availible to the average person. Forums like DataTau and /r/DataIsBeautiful are a good central starting place to stay current with new libraries, datasets and tools. Additionally, there are a number of web platforms for analyzing smaller data sets and collaborating with others. As of now, a number of tools which help fill in webdev information gaps, from version control to CSS, and hosting, are also listed in this section.

Tableau Public: Free cloud-based data visualization tool focused on community collaboration.

DataNitro: Microsoft Excel spreadsheet plug-in for Python scripting and data manipulation.

Plotly: An online collaborative data analysis and graphing tool to make and share interactive charts.

Many Eyes: A web application to build, share and discuss graphic representation of user uploaded data.

ColorBrewer: Color advice for cartography and map data visualizations.

MrDataConverter: Convert MSExcel data into one of several web-friendly formats, including HTML, JSON and XML.

BeautifulSoup: Provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree.

Codepen: An HTML, CSS, and JavaScript code editor in your browser with instant previews of the code you see.

jsFiddle Another useful online sandbox to explore web development parameters in a collaborative manner.

Git Ready: A set of git (version control software) tutorials. If you want to stick around and code then you should get to know version control.

HTMLDog: At some point you'll want to package it all together and put it on the web, and this is the tutorial for that.

Public Datasets

Half the fun of datascience is figuring out what information can be mashed together for new insights. Both StackOverflow and Quora have extensive lists of publically availible data. Below are some of the more comprehensive and interesting health data libraries availible to the public. Additional information can be mined through web scrapping, crowdsourcing and integrating service APIs.

LinkedData: Connected distributed data across the semantic web.

Data.Medicare: Download, explore, and visualize availible Medicare data.

Data.gov: Data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, based on open US data.

Amazon: Public Data Sets on AWS providing a centralized repository of public data sets that can be seamlessly integrated.

Figshare: Store, share, discover, research and manage cloud based data.

CDC.gov: National Health and Nutrition Examination Survey data repository.

CMS.gov: Center for Medicare and Medicaid Services public data sets.

HealthData.gov: A broader collection of HHS provided data and population statsitics.

EPA Data: Access to EPA's data sources organized by subject as well as access to the Environmental Dataset Gateway.