Introduction to Data Science

A drink from the fire hose so you can start your own exploration.

Bryan Nehl
@k0emt
dbBear.com

What is Data Science?

- Turns out, that's a good question!
- Skills
- Sourcing
- Extract, Transform and Load
- Databases
- Languages and Libraries
- Business Knowledge
- Visualization
- What now?

Skills

- Data Journal – Engineer's Notebook
  - http://soloso.blogspot.com/2014/05/engineers-notebook.html
- Regular Expressions, Markdown
- Analysis (math/stats is part of this!)
- LINUX
  - Putty
- Version Control
  - http://git-scm.com/
- Messaging – RabbitMQ
  - http://rabbitmq.com

Virtualization

- Why?
  - Time
    - Pre-built images
  - Cost
  - On Demand
- How/Where?
  - Microsoft Azure
  - Amazon Elastic Cloud
  - Google Compute
  - Heroku
  - Your Own Machine
    - Oracle VirtualBox - `www.virtualbox.org`
    - Parallels
    - VMWare

Common Virtual Machines (VMs)

LAMP/WAMP

- Linux/Windows
- Apache
- MySQL
- PHP

MEAN

- MongoDB
- Express
- Angular
- Node.JS
- mean.io

MongoDB is the leading NoSQL database, empowering businesses to be more agile and scalable.

Express is a minimal and flexible node.js web application framework, providing a robust set of features for building single and multi-page, and hybrid web applications.

AngularJS lets you extend HTML vocabulary for your application. The resulting environment is extraordinarily expressive, readable, and quick to develop.

Node.js is a platform built on Chrome's JavaScript runtime for easily building fast, scalable network applications.

Data Sourcing

Sourcing the data

- Locate it
  - Provided
  - Search for it
    - Manually
    - Automated
  - Networking
- Get it
  - ftp
  - Scraping
  - Database
  - Web services
- Work with it

Some dataz

- data.gov
- data.mo.gov
- data.kcmo.gov
- data.gov.uk
- data.worldbank.org
- gutenberg.org
- http://bitly.com/bundles/hmason/1
- soloso.blogspot.com/2011/07/getting-enron-mail-database-into.html

ETL

Extract, Transform and Load (the data!)

Extract, Transform & Load Data Formats

- XLS
- CSV
- Text
  - Delimited
  - Fixed format
- JSON – json.org
- XML & HTML
- Mail files
- SQL scripted INSERTs
- PDF
- Character sets

`https://docs.python.org/2/howto/unicode.html`

Data Analysis ???

- What do I expect to see?
- What are the field types?
- Does the field type change?
- What are the range of values?
- How frequently do those values occur?
  - *Can I get a graph please?*
- Are there nulls?
- How big is my sample set?
  - *Is it significant?*
- How big do I expect the real data to be?
- Are there holes in the data?
- What constitutes a *good* record?
- Where are the trends/ clusters in the data?

- openrefine.org
  - Explore
  - Clean and Refine
  - Matching
  - Exporting

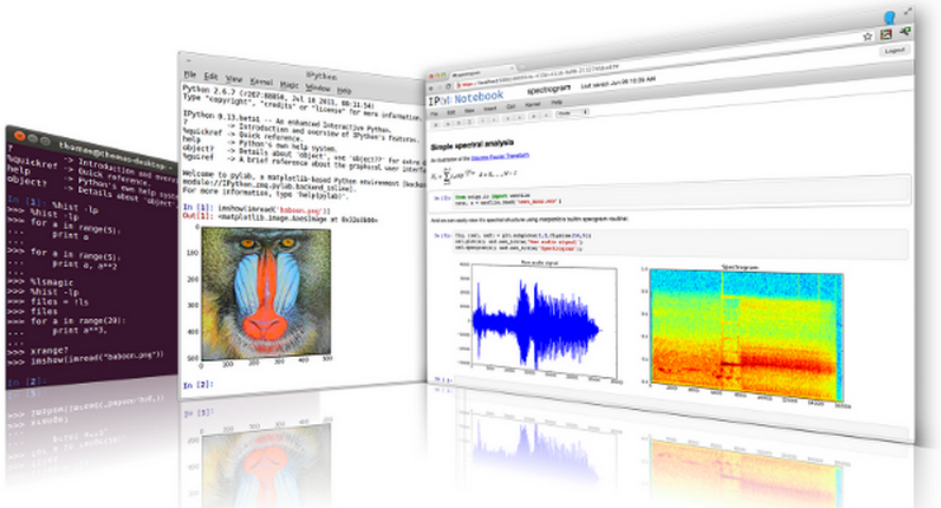
  - Plug-in architecture
  - Works with web services
  - Works with freebase.com

Languages and Libraries

Tools to get the work done. Don’t reinvent the wheel.

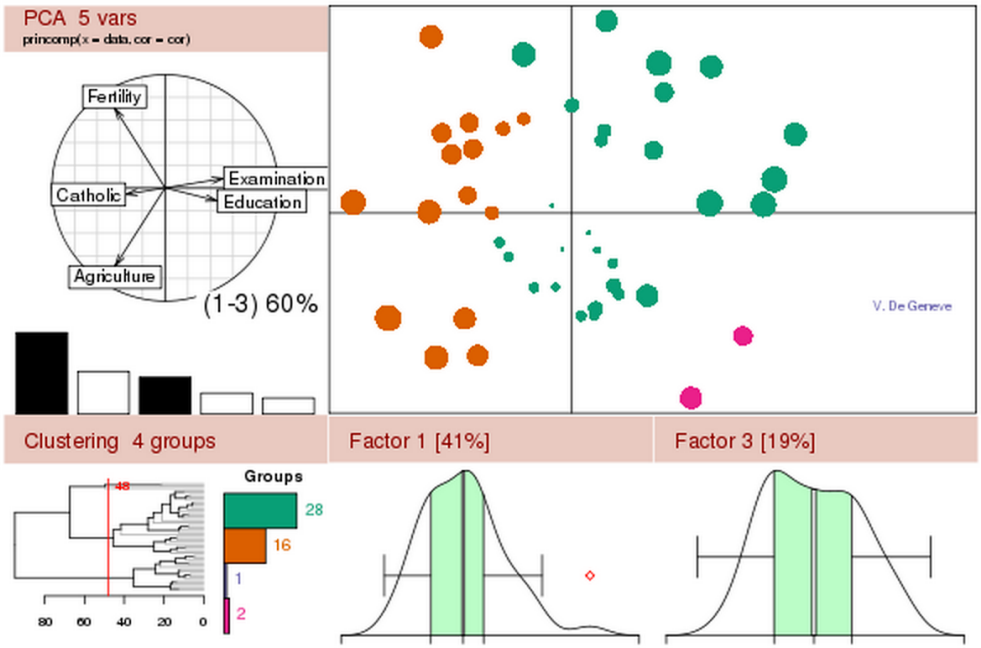
Languages

- Python –python.org
- IPython – ipython.org



Languages

- R - www.r-project.org
- R Studio - www.rstudio.com



Languages

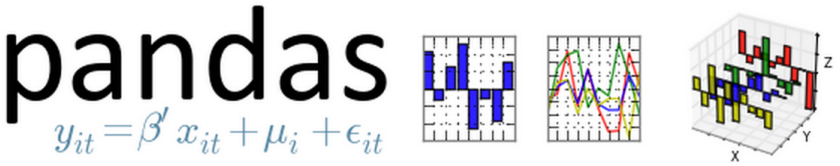
- Julia – `julialang.org`
  - Very fast numeric operations
  - Close to FORTRAN in speed
  - Designed for parallelism
  - IJulia works with IPython
- Javascript
  - jquery

Libraries for Excel

- It is everywhere
- Python Libraries:
  - xlrd
  - XlsxWriter
- Apache Project – Office Open XML file formats
  - `http://poi.apache.org/`
    - Excel
    - Word
    - PowerPoint

Libraries

- SciPy
  - scipy.org
- NumPy
  - numpy.org
- Pandas - Python data analysis library
  - pandas.pydata.org

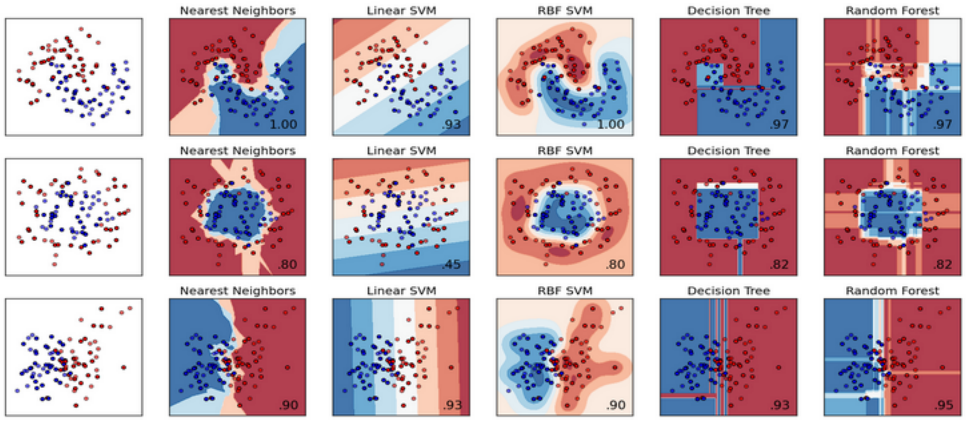


Libraries

- lxml
  - `lxml.de`
- pymongo
  - `pypi.python.org/pypi/pymongo/`
- pika – AMQP
  - `pypi.python.org/pypi/pika`
- nose – unit test framework extension
  - `nose.readthedocs.org`

scikit-learn.org

- Machine Learning
  - Clustering
  - Classification
- Data Mining



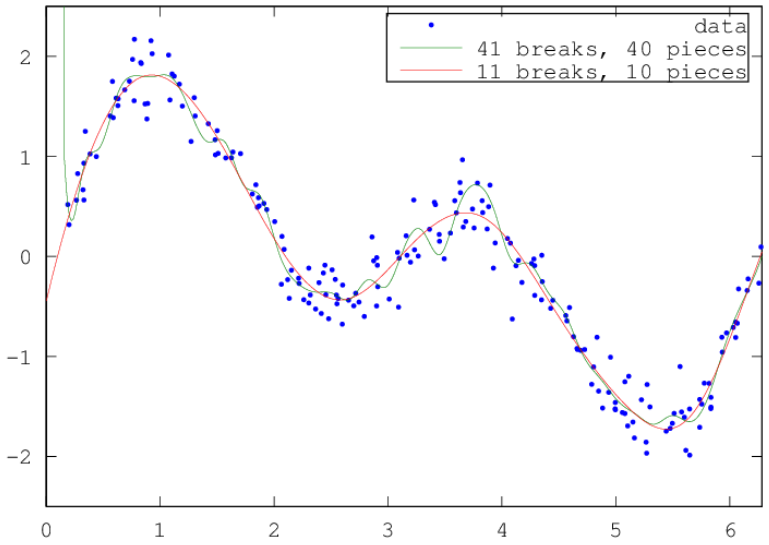
Packages

- Anaconda Scientific Python development environment
  - Getting IPython set up by hand is a pain—Anaconda is a must on Windows machines.
  - `https://store.continuum.io/cshop/anaconda/`

- wakari.io web based Python data analysis

Other Tools

- Processing.org
- Octave
  - www.gnu.org/software/octave/



Databases

Choose the right one(s) for the job!

Polyglot Persistence
http://martinfowler.com/bliki/PolyglotPersistence.html

Relational - SQL

- MySQL
  - open source
- Oracle
- Microsoft SQL Server
- Express Editions
- Microsoft Access

- ODBC / JDBC

NOSQL

- Definition

- MongoDB – JSON/BSON documents
  - mongodb.org
- neo4j – graph
  - neo4j.org
- PostgreSQL – object-relational
  - postgresql.org

NOSQL

- Aerospike – Key-Value
  - aerospike.com
- redis – Key-Value
  - redis.io
- CouchDB – JSON docs, HTTP
  - couchdb.apache.org
- Couchbase – JSON, memcached
  - couchbase.com

- Distributed framework for processing large datasets
- MongoDB and other databases can be used to feed it

- hadoop.apache.org

Example Polyglot System

Lobbyist Influence

- Relationships:
  - Politicians
  - Lobbyists
  - Legislation

- PDF Versions of the bill

- Financial information

Business Knowlege

I have data. Now what?

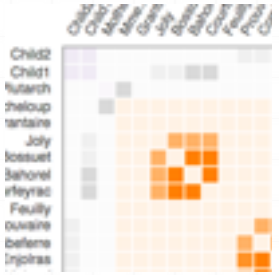
Numbers need context

Visitors	1M
Page Views	5.2M
72%	Conversion Rate
42	Customer average age
1	Top Referrer.com

Analysis

Techniques

- Adjacency Matrix
- pivot and fold operations on tables
- hexagonal binning
- confusion matrix
- predictive modeling fundamentals
- machine learning
- The work of John Tukey (Statistics)
  - wikipedia.org/wiki/John_Tukey



What can I do with this data that will benefit the business?

- Is there some insight I can bring?
- Can I generalize from this data? (global)
- Can I ascertain local area insights?
- Are there natural partitions in the data?
  - Gender, race, age, location?
- Is there some business pain I can relieve?
- Can I enhance an existing data set?
- Can I bring in the data product with a shorter cycle time?

Business Intelligence

- wikipedia.org/wiki/Business_intelligence

- Pentaho - pentaho.com
- Tableau - tableausoftware.com
- Microsoft Power BI for Office 365
  - microsoft.com/en-us/powerbi/

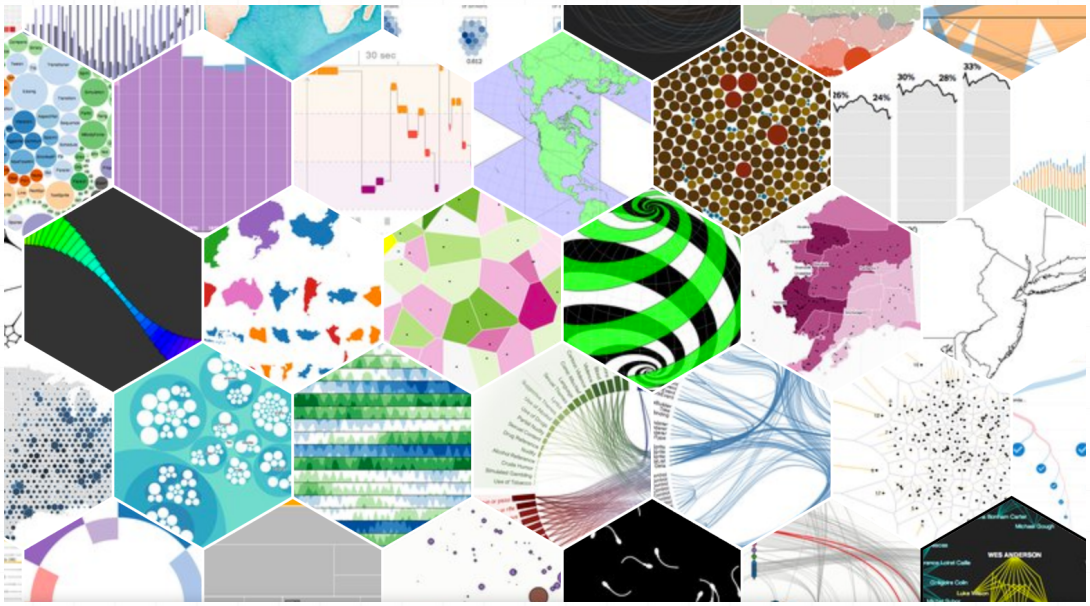
Visualization

Use your pixels!

HTML Tools & Libraries

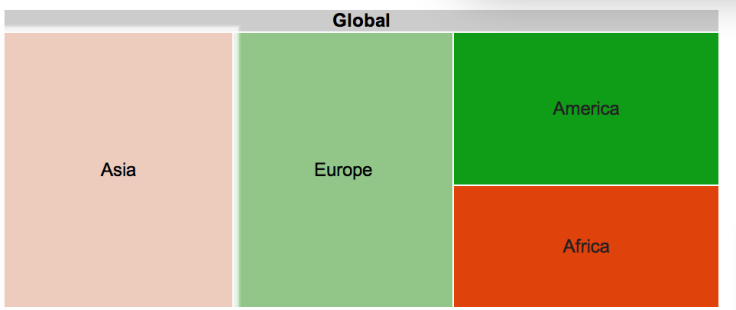
- HTML5 / CSS3
- Javascript
- **d3.js – `d3js.org`**

- HTML 5 canvas charts
  - chartjs.org
  - canvasjs.com



Google Tools & Libraries

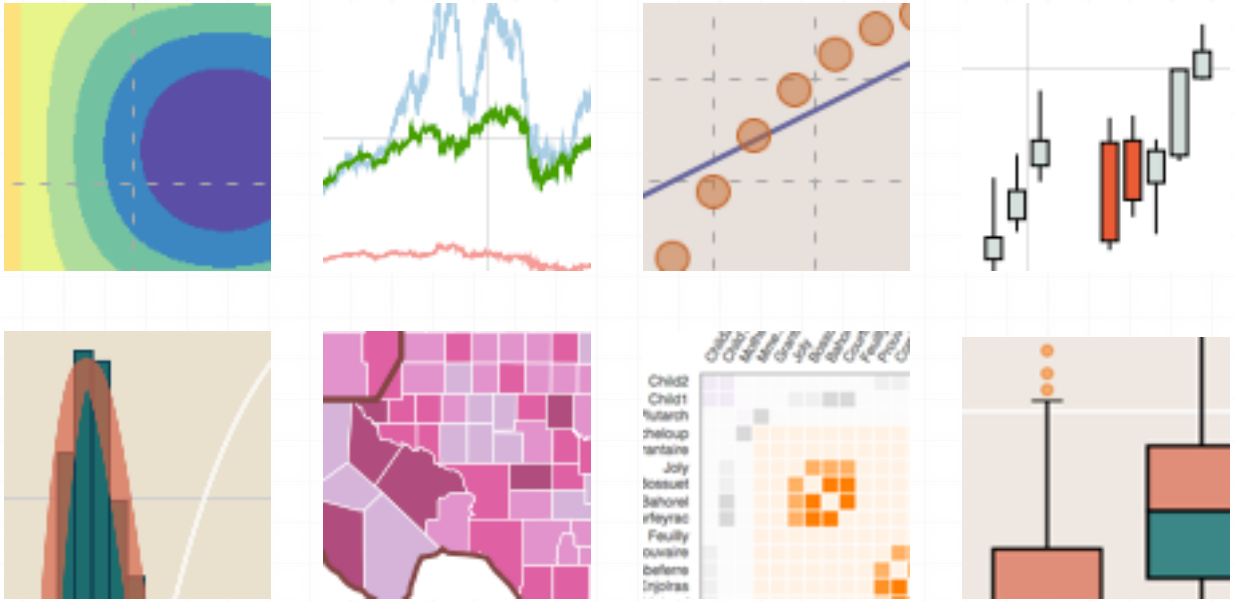
- Google Charts
  - developers.google.com/chart/



- Google Fusion Tables
  - Now integrated with Google Drive

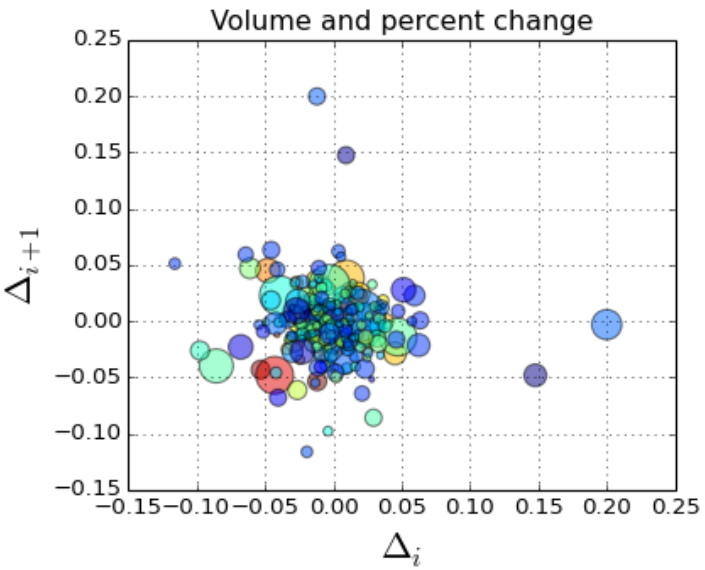
Python Tools & Libraries

0 bokeh - `bokeh.pydata.org`

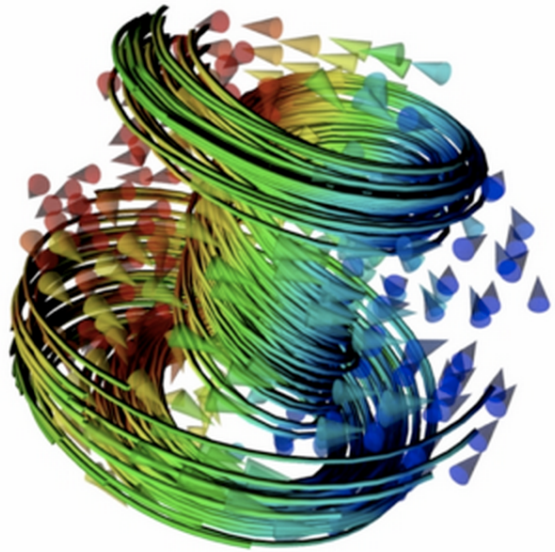


Python Tools & Libraries

- Matplotlib
  - matplotlib.org

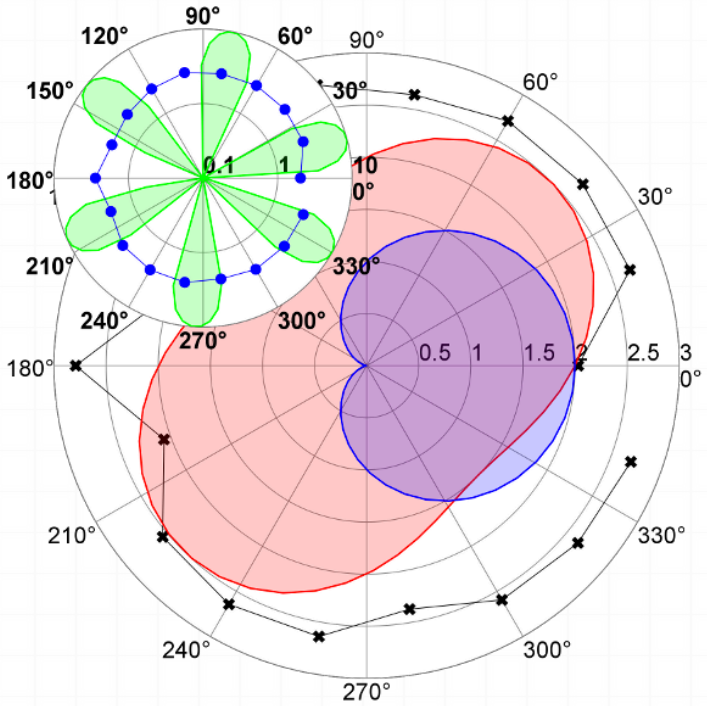


- Mayavi 2
  - code.enthought.com/projects/mayavi/

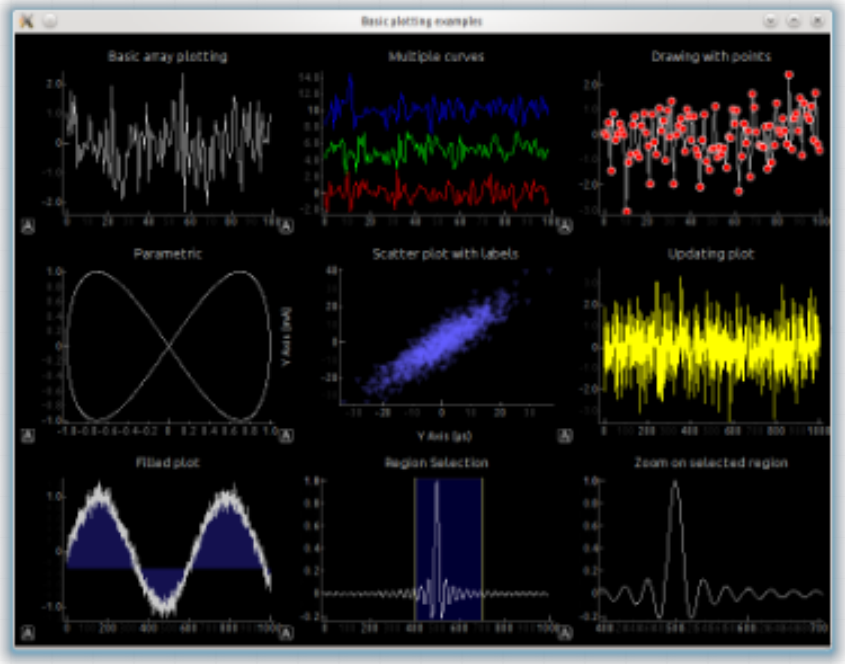


Python Tools & Libraries

- home.gna.org/veusz/



- pyqtgraph.org



Excel: data quality check

0.335857	0.733451	0.599874	0.335857	0.733451	0.599874
0.398299	0.193938	0.572766	0.398299	0.193938	0.572766
0.71445	0.22316	0.360831	0.71445	0.22316	0.360831
0.821805	0.568467	0.858095	0.821805	0.568467	0.858095
0.069867	0.434296	0.730381	0.069867	0.434296	0.730381
0.206457	0.918653	0.377569	0.206457	0.918653	0.377569
0.04397	0.908735	0.801125	0.04397	0.908735	0.801125
0.952784	0.213182	0.621818	0.952784	0.213182	0.621818
0.305901	0.528717	0.545583	0.305901	0.528717	0.545583
0.732739	0.579152	0.202078	0.732739	0.579152	0.202078

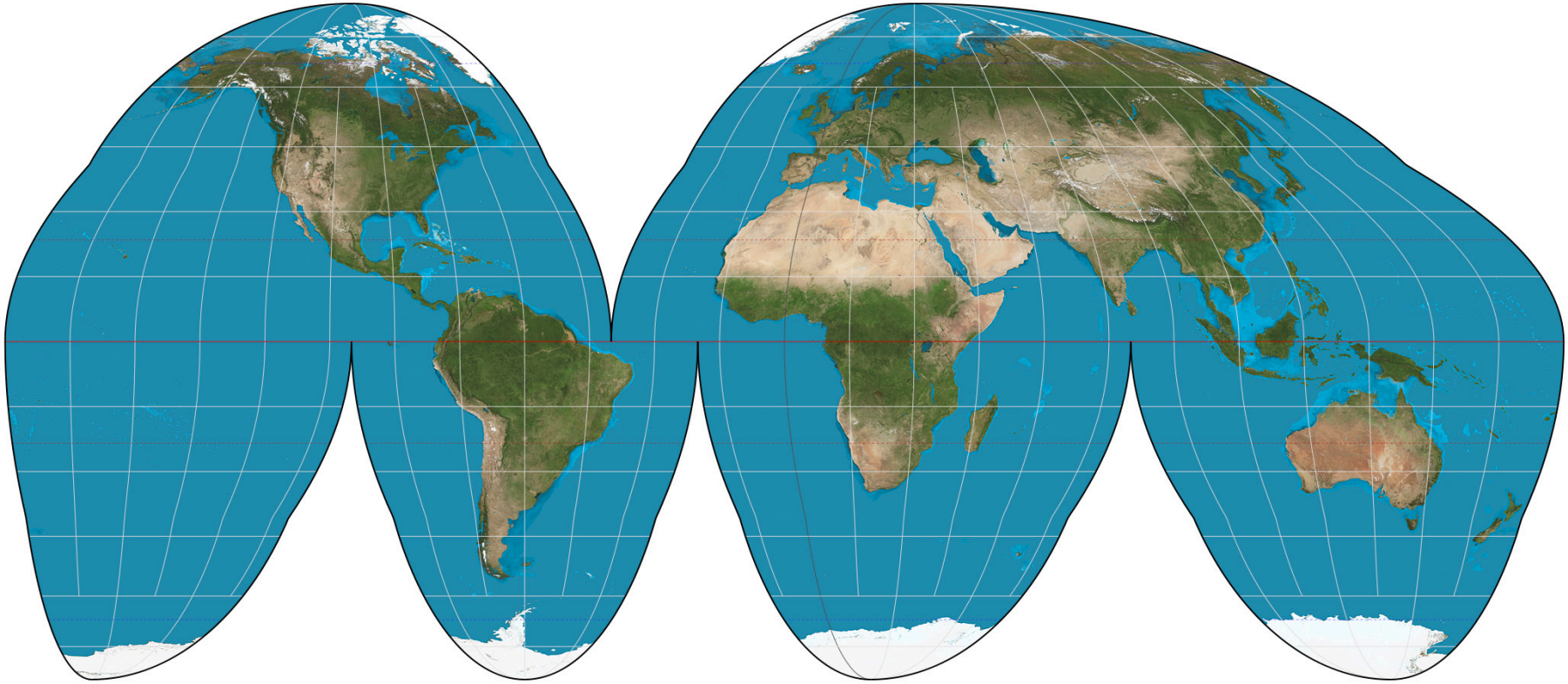
Map Projections

- http://en.wikipedia.org/wiki/List_of_map_projections
- http://xkcd.com/977/
- http://www.jasondavies.com/maps/countries-by-area/

Mercator

http://upload.wikimedia.org/wikipedia/commons/f/f4/Mercator_projection_SW.jpg
http://commons.wikimedia.org/wiki/File:Australia-Greenland_size_comparison.svg

Goode homolosine projection



Equal area projection map

http://en.wikipedia.org/wiki/File:Goode_homolosine_projection_SW.jpg

That was a

***BIG** drink*

from the fire hose!

What now?

Teams

Decide your direction

0 Personal Tech Radar

http://nealford.com/memeagora/2013/05/28/build_your_own_technology_radar.html

Training

- Tutorials and sample files that come with software.
- Local courses
- Online Education from vendors
  - MongoDB University
    - `university.mongodb.com`
- Other online education
  - Coursera – `coursera.org`
  - iTunes University (iTunes U)
  - O'Reilly Safari, books, videos and free publications
  - `oreilly.com/data/free`

Experiment

- Set up a development environment
- Create a Virtual Machine
- Try out stuff
  - Work related
  - Something you are passionate about
- Share your experiences
  - blog, tweet, present
  - GitHub and Gists

Contests

- kaggle.com
- www.kdnuggets.com/competitions/
- www.crowdanalytix.com
- www.innocentive.com
- tunedit.org

- Tips for winning
  - http://www.allanalytics.com/author.asp?doc_id=268513

Mentors & Community

- Google+
  - Data Science
  - Statistics and R
  - Artificial Intelligence
  - Machine Learning
  - Python
  - MongoDB
- LinkedIN
- Twitter
- IRC
- People within your company

Conferences

- Strata
  - strataconf.com
- PyData
  - pydata.org
- PyCon
  - us.pycon.org
- Big Data Summit KC
  - BigDataSummitKC.org
- Investigative Reporters & Editors Conference
  - www.ire.org/conferences/

Introduction to Data Science

Questions?

Bryan Nehl

@k0emt

dbBear.com

(Links to Twitter, blog, GitHub, gists)