A Contextual introduction to Data Science

Kick starting your own exploration!

Bryan Nehl

@k0emt

dbBear.com

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/*NIX/ OS X/macOS
  - putty
- Version Control
  - http://git-scm.com/
- Messaging
  - http://rabbitmq.com
  - http://zeromq.org

Virtualization/Containers

- Why?
  - Time
    - Pre-built images
  - Cost
  - On Demand

- How/Where?
  - Microsoft Azure
    - Data Science VM
  - Amazon Elastic Cloud
  - Google Compute
  - Your Own Machine
    - Oracle VirtualBox - www.virtualbox.org
    - Docker
      - hub.docker.com

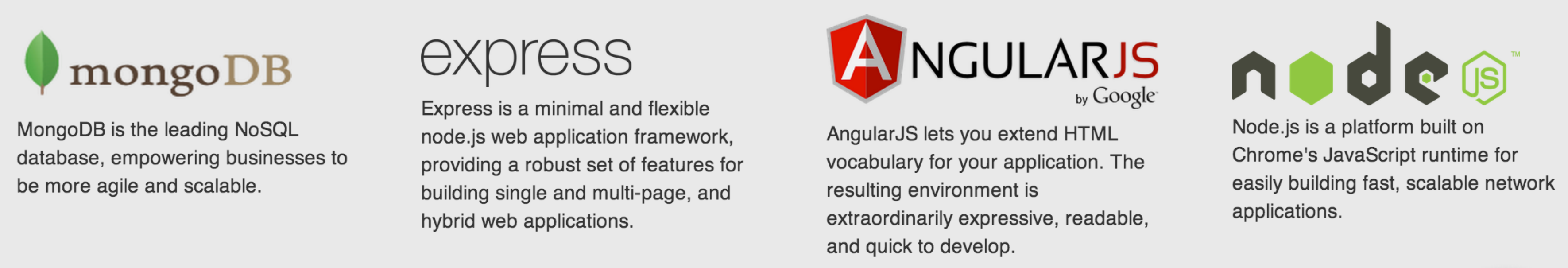
Common Virtual Machines (VMs)

LAMP/WAMP

- Linux/Windows
- Apache
- MySQL
- PHP

MEAN

- MongoDB
- Express
- Angular
- Node.JS
- meanjs.org
- meteor.com



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.

Azure Data Science VMs

- *Windows* Based VM
- Microsoft R Server Developer Edition
- Anaconda Python
- Jupyter notebooks
  - Python & R
- Visual Studio CE
  - Python & R Tools
- Power BI desktop
- SQL Server Express
- Machine Learning Tools

- *Linux* Based VM
- Microsoft R Open
- Anaconda Python
- Jupyter notebooks
  - Python & R
- Postgres Database
- Azure Tools
- Machine Learning Tools

Journal

- Linux, Linux VM, mac

- Workstation

Sourcing the data

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

Some data

- data.gov
- data.mo.gov
- data.kcmo.gov
- data.gov.uk
- data.worldbank.org
- aws.amazon.com/datasets
- gutenberg.org
- https://gist.github.com/k0emt/63f19f828561c074f119
- soloso.blogspot.com/2011/07/getting-enron-mail-database-into.html

Journal

- Sourcing Data
  - XML file layout example

NEVER

trust

the data!

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?

Journal

- Upfront Analysis
  - File layout
  - File description

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`

Languages and Libraries Extract &*Transform*

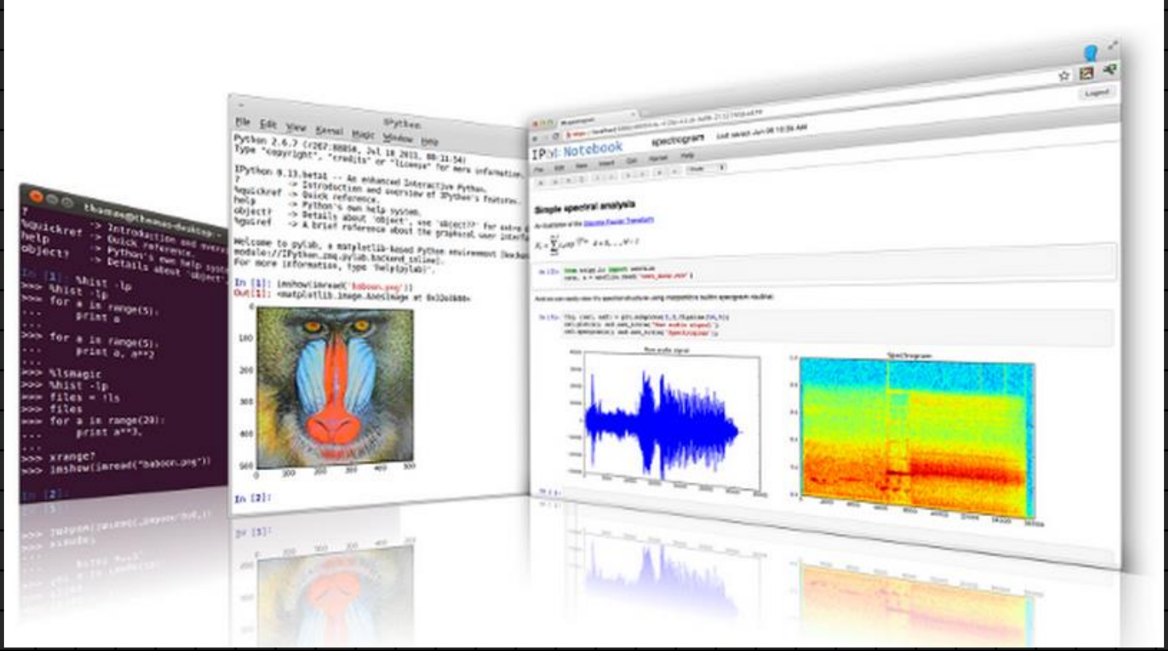
Tools to get the work done. Don’t reinvent the wheel.

Languages

- Python –python.org
- IPython – ipython.org
- http://nbviewer.jupyter.org/

python™

IP[y]: IPython Interactive Computing

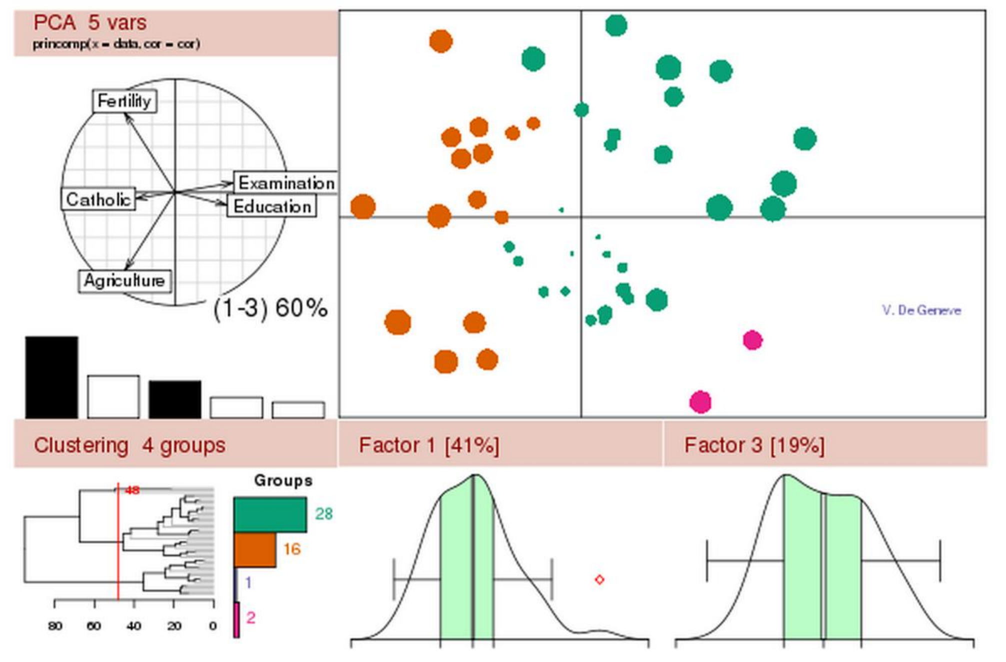


```
?
%quickref -> Introduction and overvi
help      -> Quick reference.
object?   -> Python's own help syste
          -> Details about 'object'
In [1]: %hist -lp
>>> %hist -lp
>>> for a in range(5):
...     print a
...
>>> for a in range(5):
...     print a, a**2
...
>>> %lsmagic
>>> %hist -lp
>>> files = !ls
>>> files
>>> for a in range(20):
...     print a**3,
...
>>> xrange?
>>> imshow(imread("baboon.png"))
In [2]:
```

```
Python 2.6.7 (r267:88850, Jul 18 2011, 08:11:54)
Type "copyright", "credits" or "license" for more information.

IPython 0.13.beta1 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra d
%guiref   -> A brief reference about the graphical user interfa

Welcome to pylab, a matplotlib-based Python environment [backe
module://IPython.zmq.pylab.backend_inline].
For more information, type 'help(pylab)'.

In [1]: imshow(imread('baboon.png'))
Out[1]: <matplotlib.image.AxesImage at 0x33e6880>

In [2]:
```

Languages

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



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

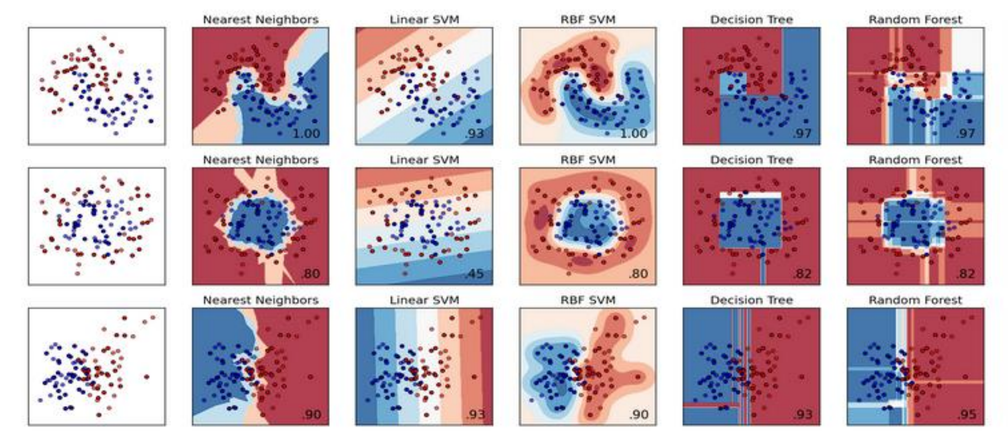
- SciPy
  - scipy.org
- NumPy
  - numpy.org
- Pandas - Python data analysis library
  - pandas.pydata.org

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://www.continuum.io/why-anaconda
- wakari.io web based Python data analysis

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
- OrientDB – document & graph
  - orientdb.com
- PostgreSQL – object-relational
  - postgresql.org

hadoop

- Distributed framework for processing large datasets
- MongoDB and other databases can be used to feed it
- MapReduce
- hadoop.apache.org

Spark

Lightning-fast cluster computing

- In Memory MapReduce
- spark.apache.org

drill.apache.org

Journal

- DOC ETL
  - sample data file
  - Counties
  - Python Program for translating into JSON
  - error file
  - import
- Analysis of available data
  - the top 5

Business Context

I have data. Now what?

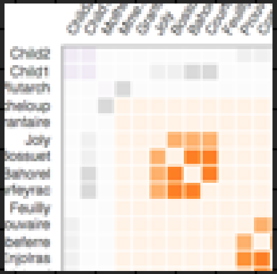
Numbers need context

Visitors	1M	Last Year 2M
Page Views	5.2M	Last Year 7.2M
72%	Conversion Rate	
42	Customer average age	
1	Top Referrer.com	

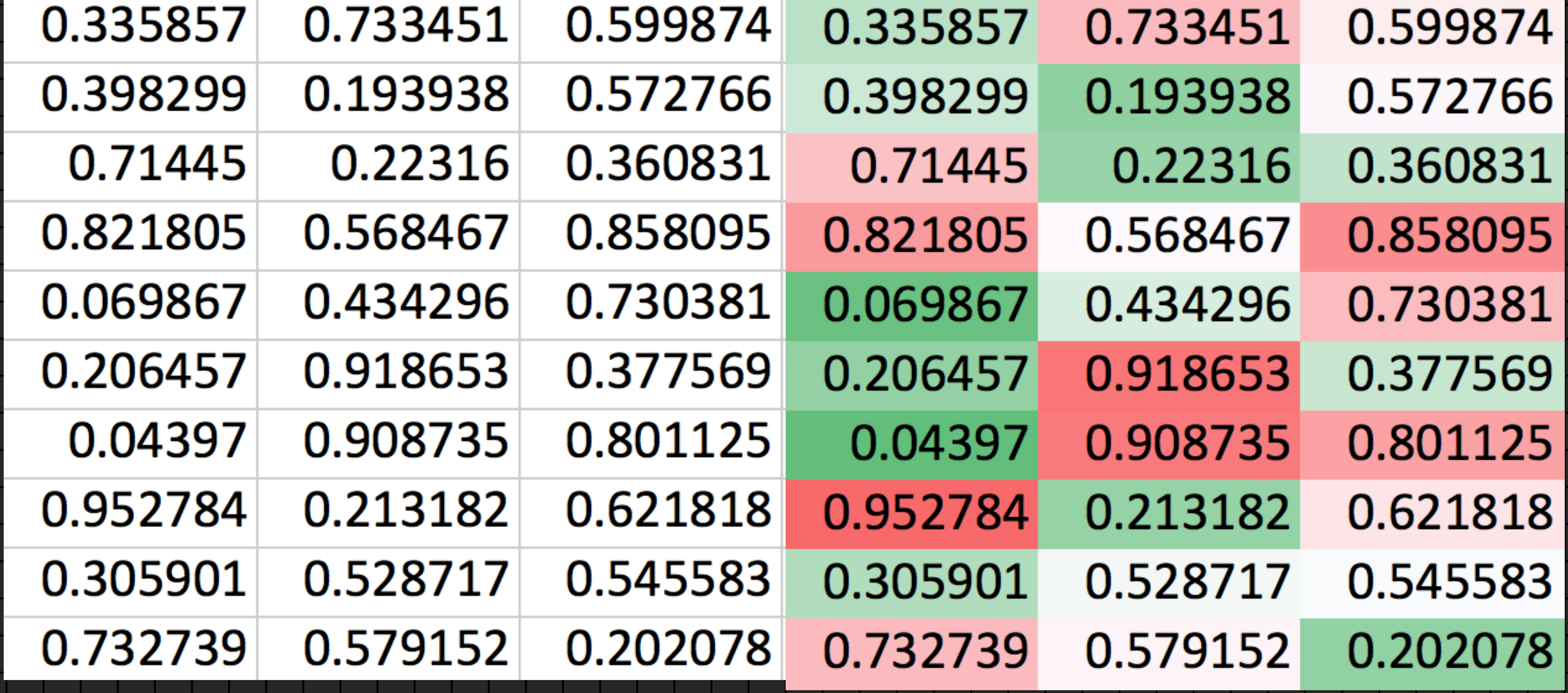
Analysis

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



Looking at numbers



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

Conditional Formatting – Color Scales

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?

The Science Part

- Ask a question
- Form a hypothesis
- Do the research

Journal

- Questions
- Data Work to Answer the questions
  - population buckets

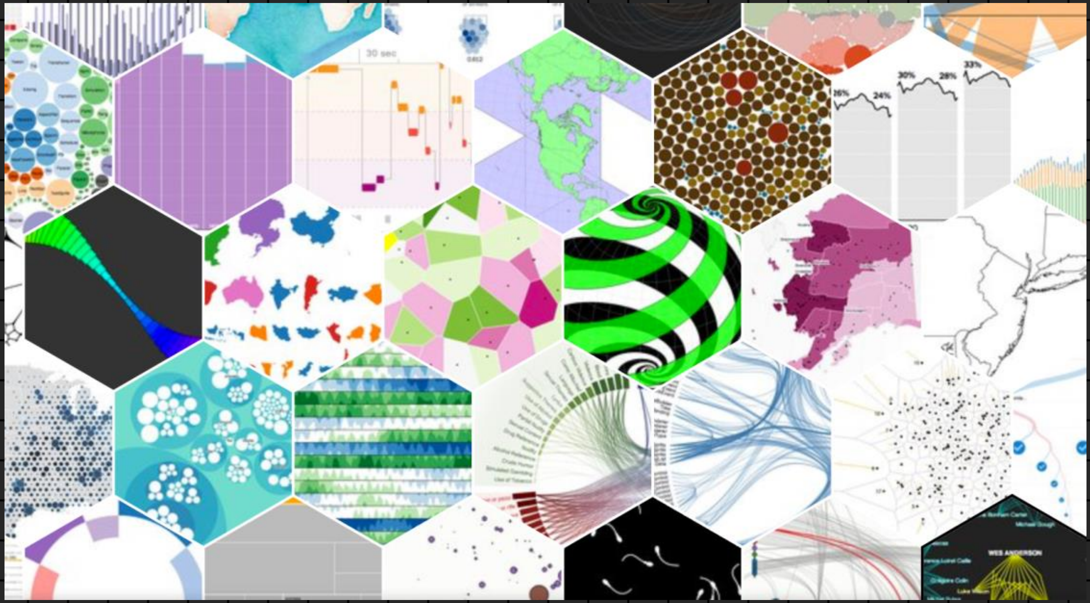
Visualization

Use your pixels!

HTML Tools & Libraries

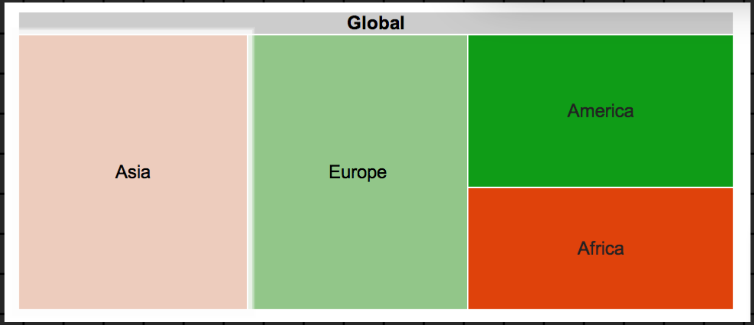
- HTML5 / CSS3
- Javascript
- **D3 – `d3js.org`**

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

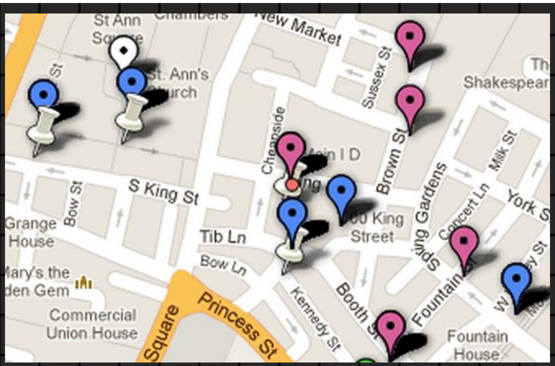


Google Tools & Libraries

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

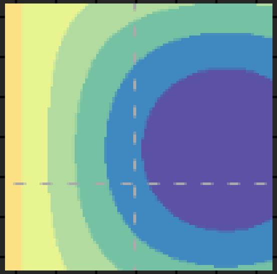
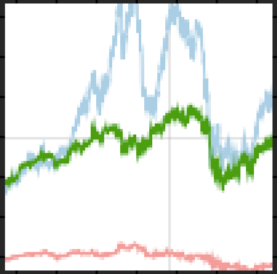
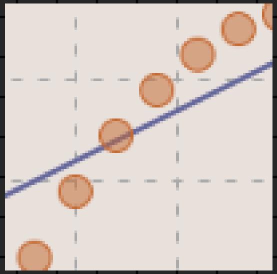
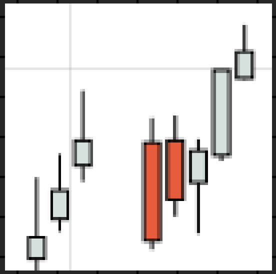
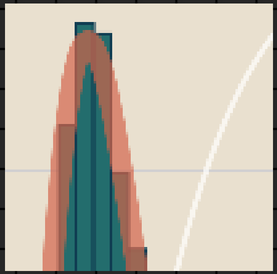
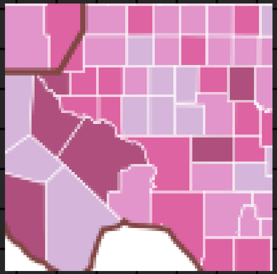
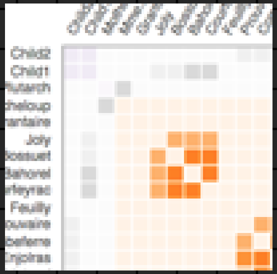


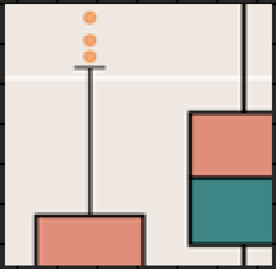
- Google Fusion Tables
  - Now integrated with Google Drive



Python Tools & Libraries

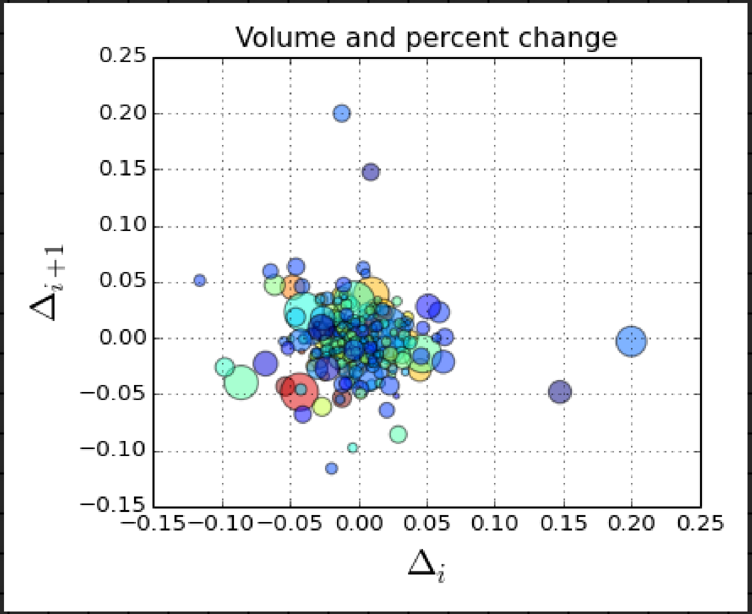
- bokeh - `bokeh.pydata.org`

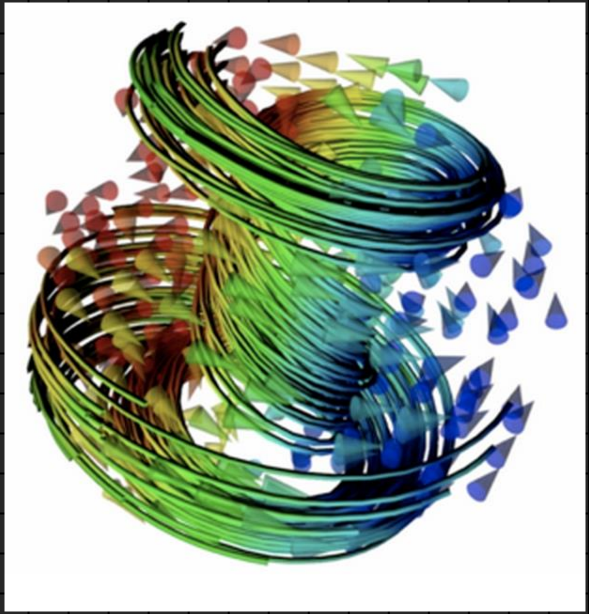


Python Tools & Libraries

- Matplotlib
  - matplotlib.org



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



Journal

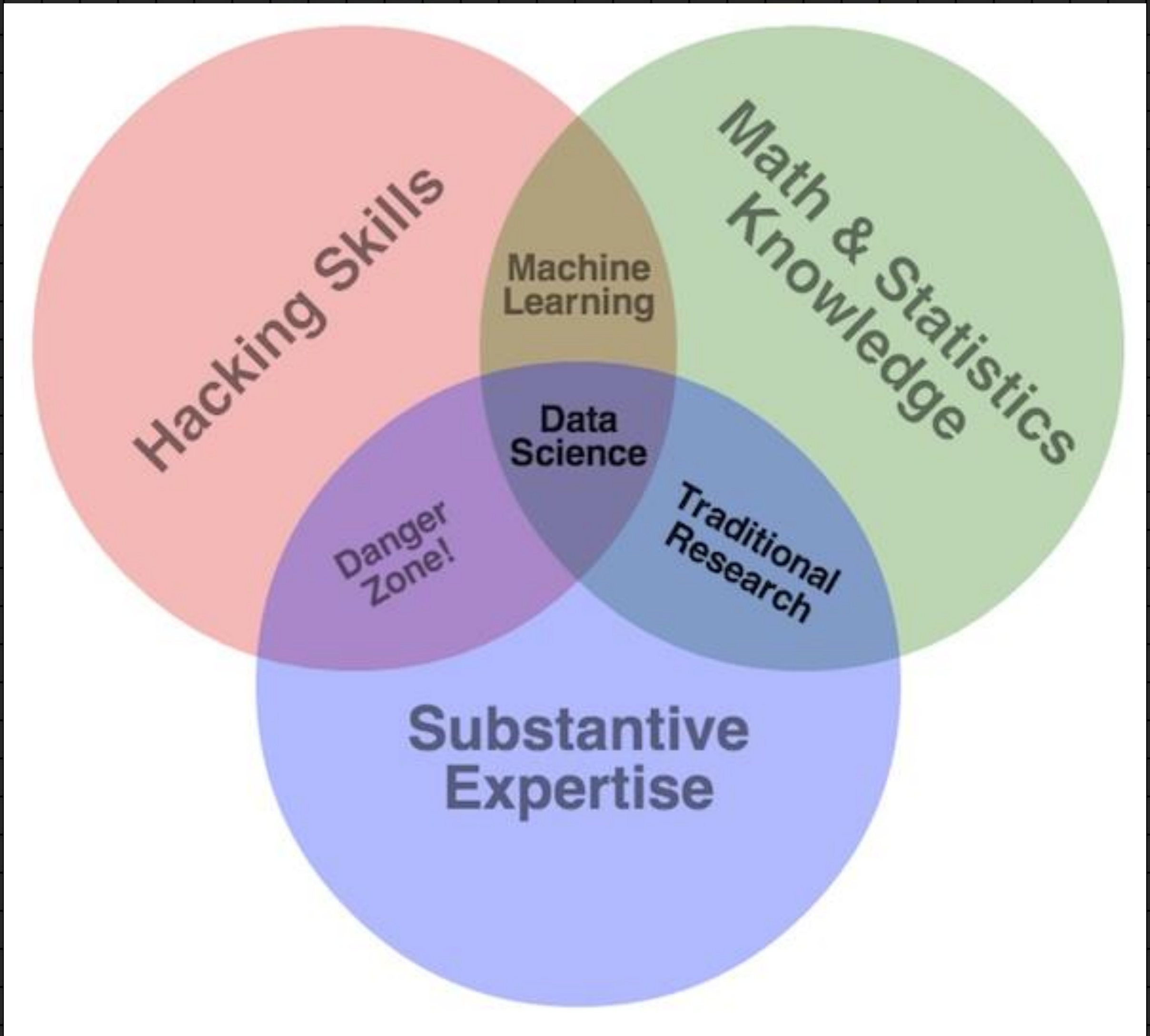
- Visualization
  - Excel
  - R
    - Box Plots - base
    - Violin Plots – ggplot2

- Anaconda & bokeh

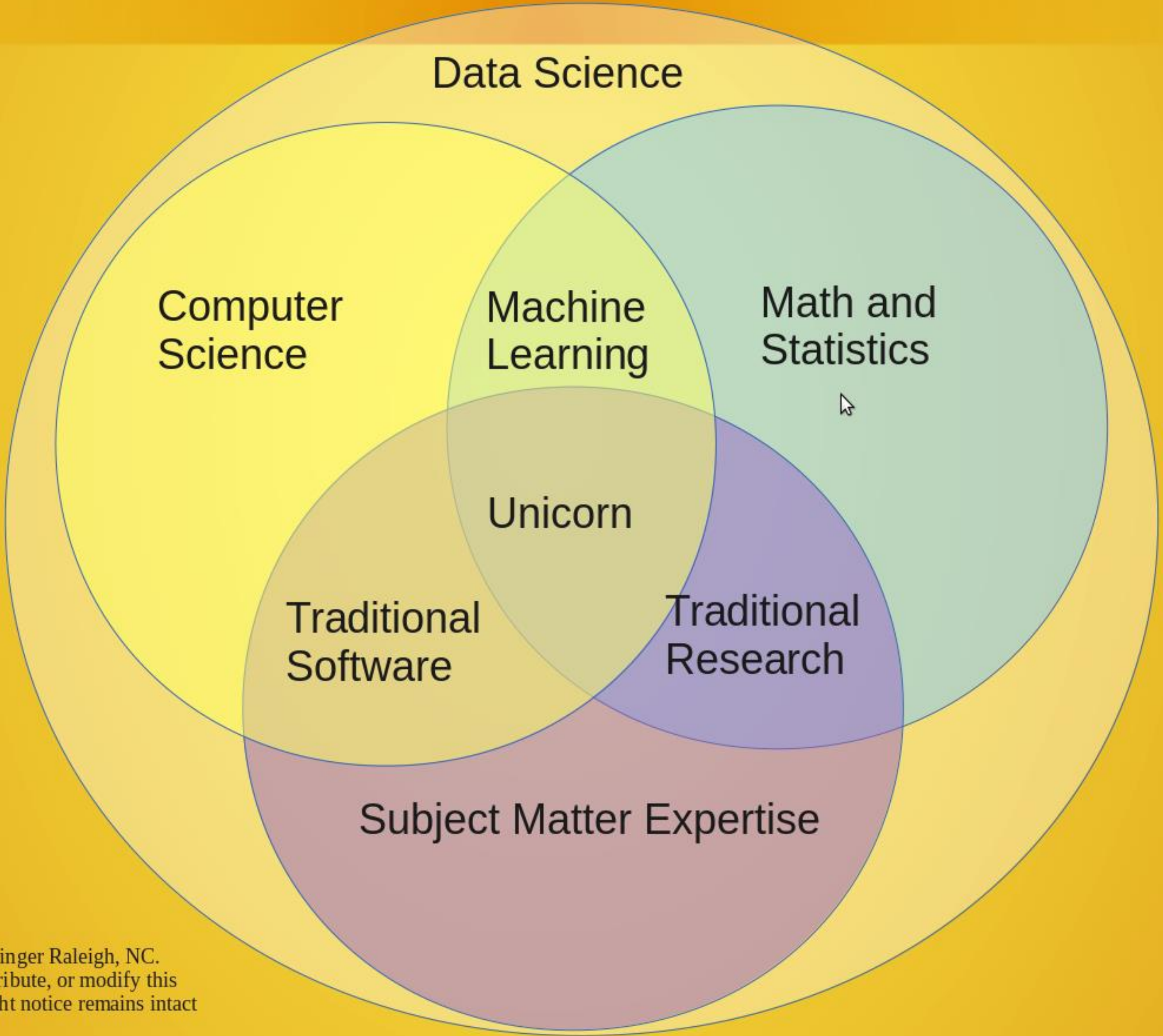
Don’t jump to conclusions!

What now?

Drew Conway's Diagram



Data Science Venn Diagram v2.0



Teams

Decide your direction

- Personal Tech Radar

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

Conferences

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

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
  - YouTube
  - Open Source Data Science Masters
    - datasciencemasters.org

Mentors & Community

- Google+
  - Data Science
  - Statistics and R
  - Artificial Intelligence
  - Machine Learning
  - Python
  - MongoDB

- LinkedIN
- Twitter
- IRC
- People within your company
- BecomingADataScientist.com
- Reddit /r/datascience
- Datascience.stackexchange.com

Contests

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

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

Experiment

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

Enjoy your journey!

Bryan Nehl – **@k0emt** – `dbBear.com`

https://github.com/k0emt/Presentations

https://github.com/k0emt/**corrections**