A Survey of Data Science Landscape

Data Science Indy
May 9, 2013

Pradeep Gowda
ENthEnergy & CS@IUPUI
@btbytes

Engineers needed to fight online fraud with machine learning. Join us

Cloud computing

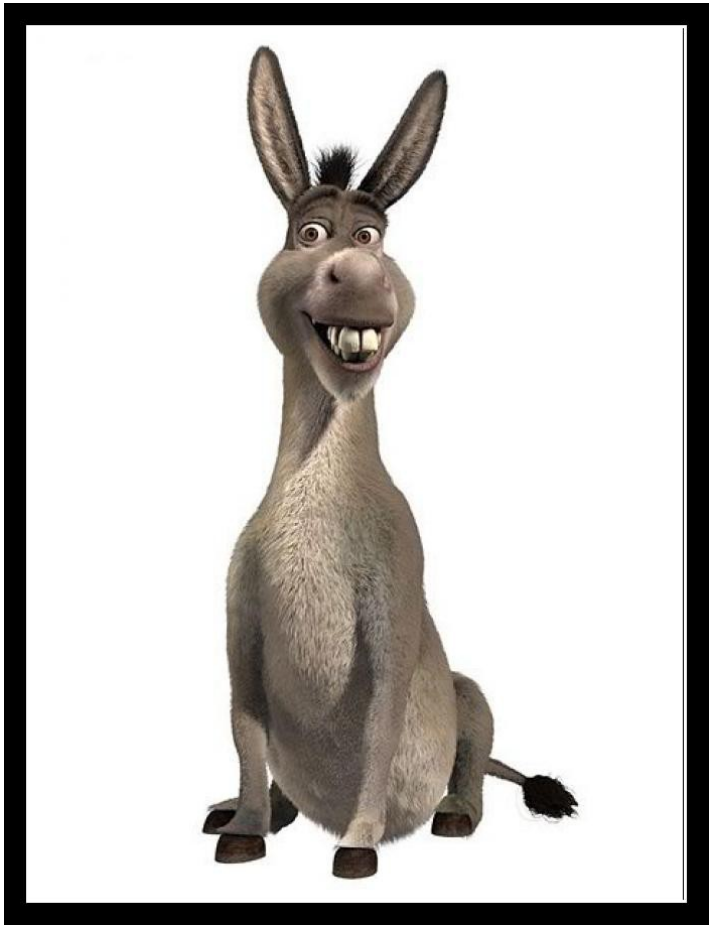
Rockstars

Big Data

Ninja

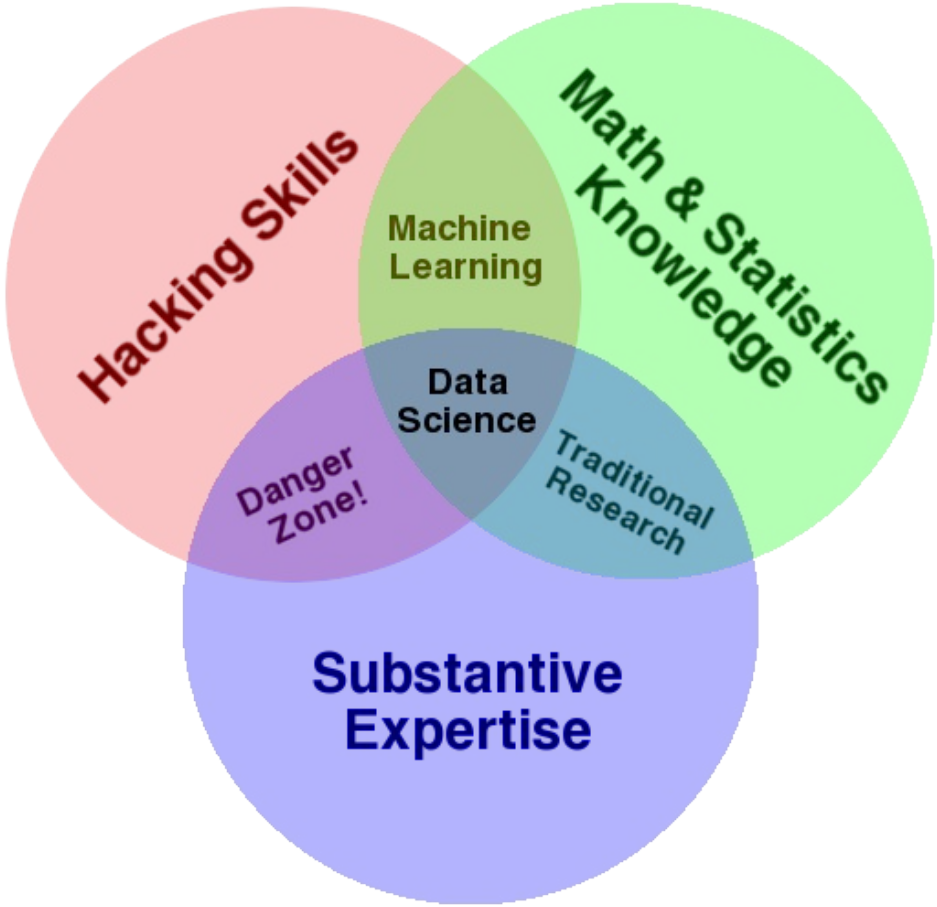
What the world thinks a Data scientist looks like

An Aspiring Data Scientist



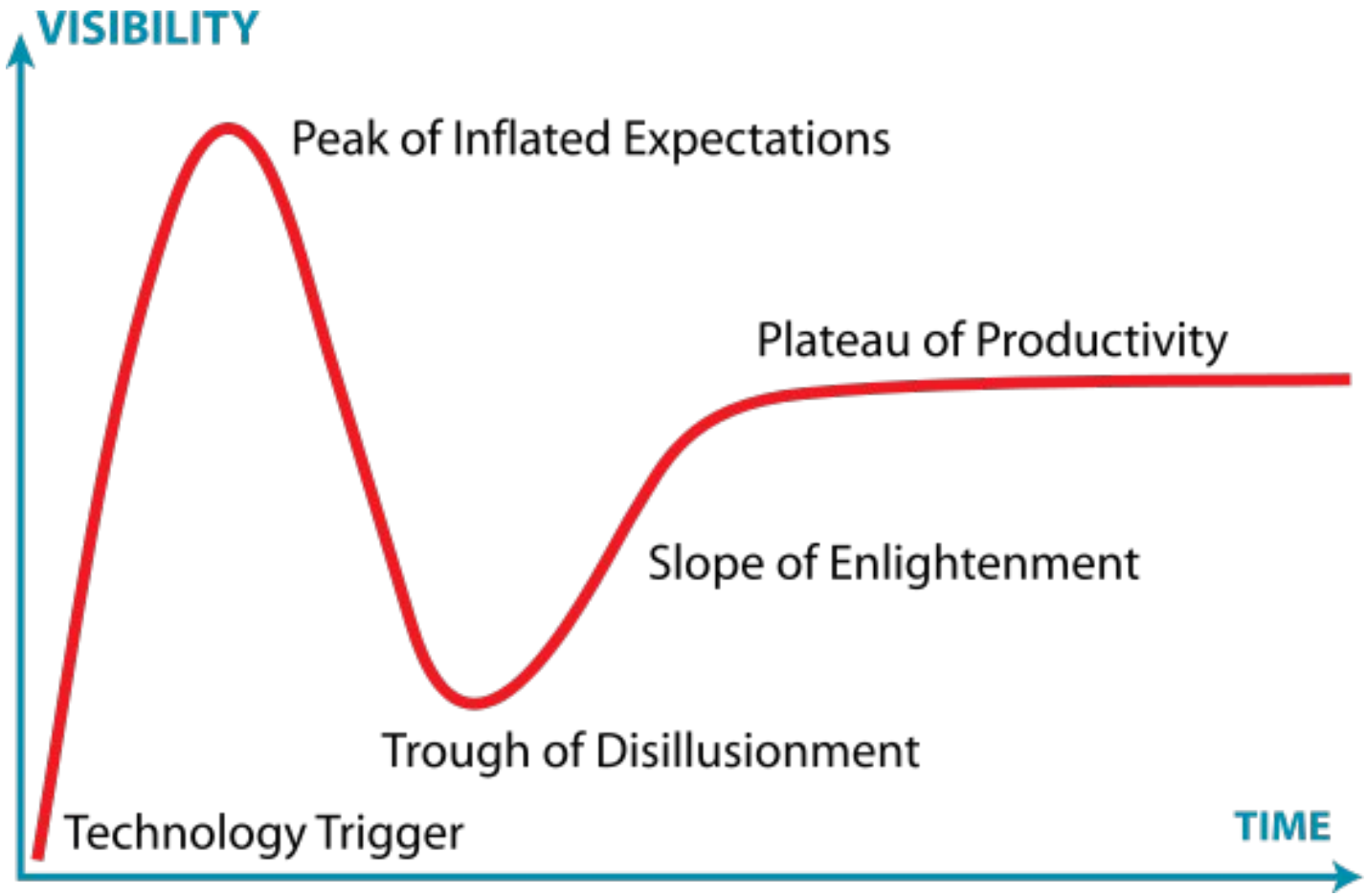
What is Data science?

- A Multidisciplinary... discipline, comprising:
  - Mathematics
  - Statistics
  - Machine learning and Pattern Recognition
  - Data Engineering
  - Algorithms
  - Visualisation
  - Data warehousing
  - High performance Computing
  - <this slot is left intentionally left blank>

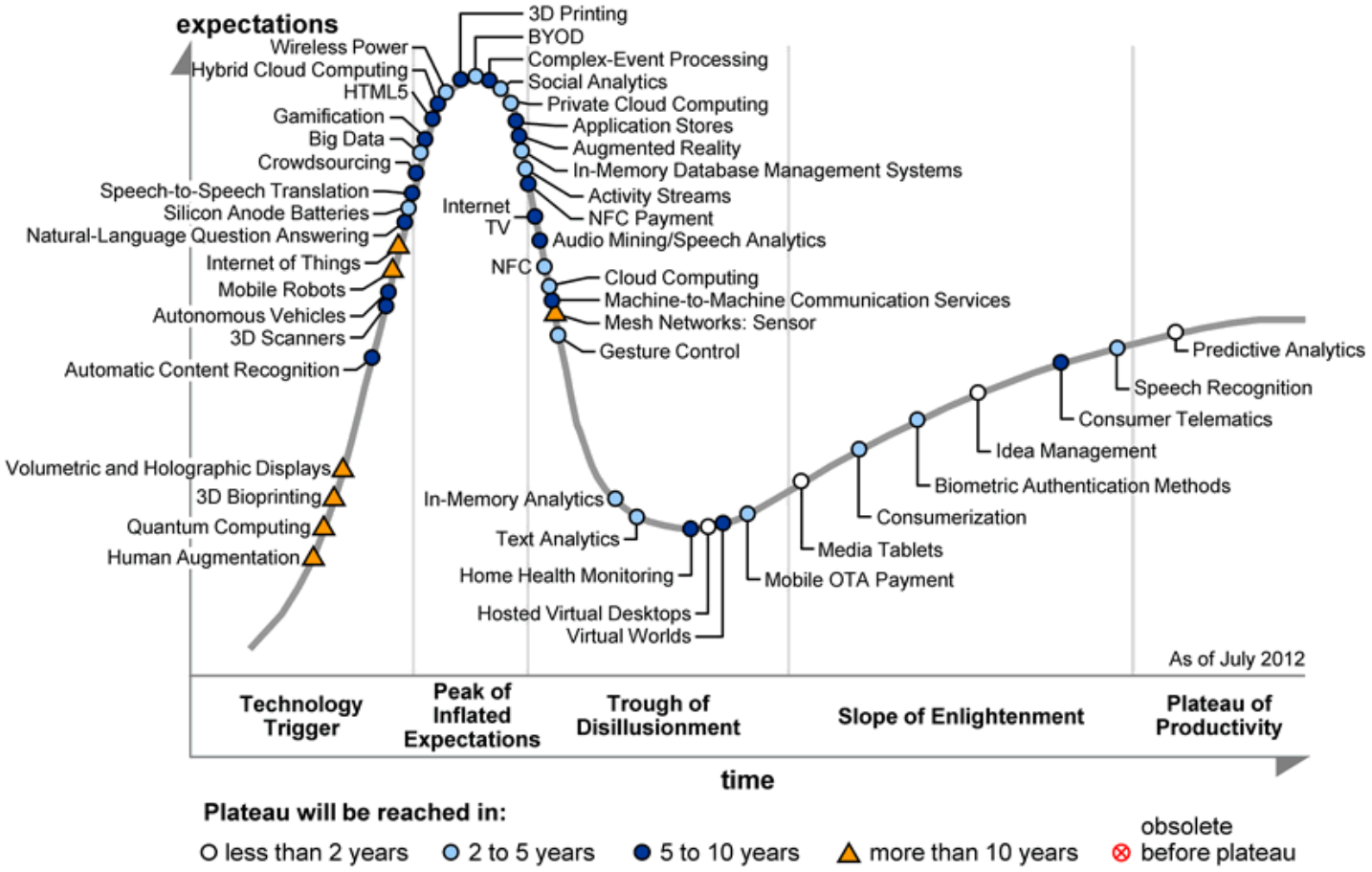


Drew Conway. IQT Quarterly, Data science in the intelligence community, pg24-27, Vol2. No.4

Hype Cycle


VISIBILITY
Peak of Inflated Expectations
Plateau of Productivity
Slope of Enlightenment
Trough of Disillusionment
Technology Trigger
TIME

Hype Cycle 2012



Data + Science

- Science has only two legs
  - Theory
  - Experimentation

What has changed is the scale of computation.

Doing theory today requires highly sophisticated Computational-Science techniques carried out on cutting-edge high-performance computers

Computation is the universal enabler of science, supporting both theory and experimentation. Today the two legs of science are thoroughly computational!

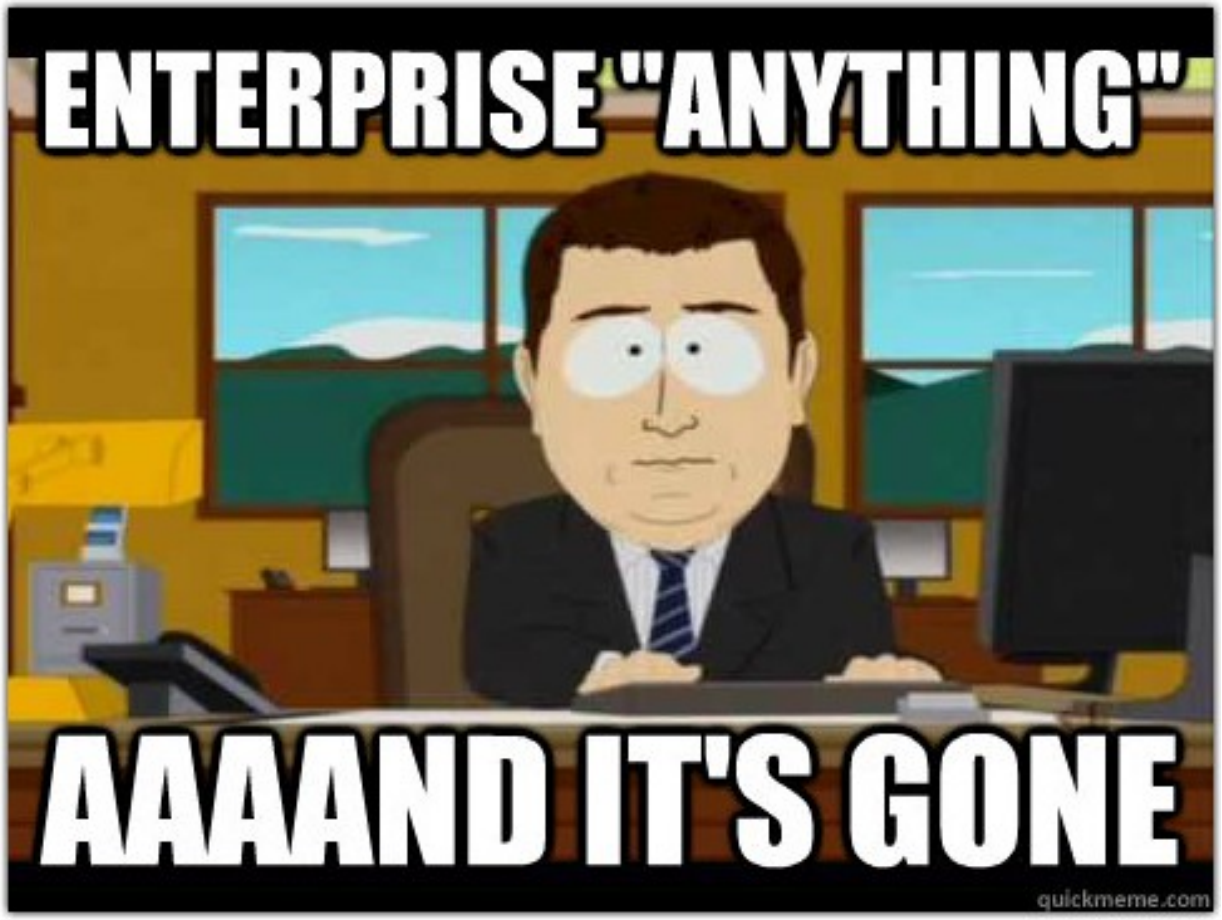
cf: DOI:10.1145/1820000.1810892, Verdi. Communications of the ACM. Vol 5. Sept 2010.

How is Data Science different from...

Don't Panic!
Here comes the buzzword train

- Business Intelligence
- Data Mining
- Data Analytics
- Cloud Computing
- SaaS/Paas/IaaS

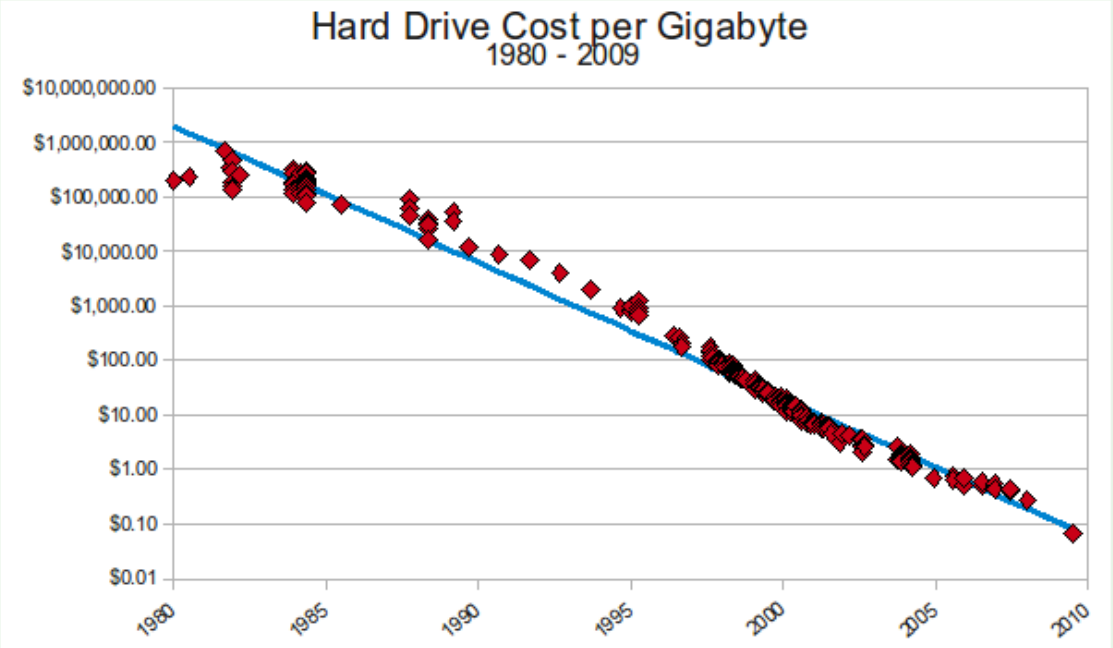
Goody! I don't like _____, let's me start replacing it with Data Science!


ENTERPRISE "ANYTHING"
AAAAND IT'S GONE
quickmeme.com

So what is new?

- Volume of data generated every second

Cheaper storage and computation Cycles



The data confirms it: there is a very strong exponential correlation in space/cost ratio (r=0.9916). Over the last 30 years, **space per unit cost has doubled roughly every 14 months** (increasing by an order of magnitude every 48 months). The regression equation is given by:

$$\text{cost} = 10^{-0.2502(year-1980)+6.304}$$

Ref:http://www.mkomo.com/cost-per-gigabyte

Data from-to everywhere

- Data generated, collected from everywhere
  - vs. Data is internal (BI, enterprise analytics)
  - Internal data overlaid with external ones.
    - Eg: GIS, Social networkin

So, what's new?

- Consumed anywhere
  - Contrast “Executive dashboard”/tool oriented software building
  - Web/mobile/tablet/Augmented Reality?

- Democratisation of data
  - Not limited to Data silos
  - Not just DBAs and “analysts”
  - Not just for the “C” suite
- Low cost, High-performance computing
  - “Cloud” is a credit card away.
- Quick turn-around times for many applications
  - Eg: Elections
  - Marketing campaigns


I DON'T ALWAYS TALK ABOUT BIGDATA
BUT WHEN I DO, I QUOTE @BIGDATABORAT
memegenerator.net

Top Data Scientist must be Don Draper in board room and Sheldon Cooper in server room.

@BigDataBorat

BigDataBorat Jr ask what is #bigdata. I say when person monitor database it #smalldata, when database monitor person it #bigdata.

Data Science is 99% preparation, 1% misinterpretation.

Big data

- aka Anything that doesn't fit in Excel?
- Big data has come to mean
  - Map reduce
  - Hadoop
  - Non-RDBMS datastores
  - Non-SQL query languages
  -

Most BIG data is really just big

- In the case of Facebook, most of the jobs engineers ask their clusters to perform are in the “megabyte to gigabyte” range
- at Yahoo, where it appears the median task size handed to Yahoo’s cluster is 12.5 gigabytes.
- “Nobody ever got fired for buying a cluster,”
- Big data has become a synonym for “data analysis,” which is confusing and counter-productive.
- In some cases, big data is as likely to confuse as it is to enlighten.

Ref: http://qz.com/81661/

Small Data

- “Small data is the amount of data you can conveniently store and process on a single machine, and in particular, a high-end laptop or server”
  - democratisation of data
  - large-scale distributed community of data wranglers
  - working collaboratively
- Size in itself doesn’t matter – what matters is having the data, of whatever size, that helps us solve a problem or address the question we have.

Machine Learning

- Focuses on “prediction” based on known properties learned from training data
- Vs “Discovery” in data mining.
- Many times the two phrases are used interchangeably
- Algorithm types
  - Supervised learning (eg:classification)
  - Unsupervised learning (eg:clustering)

Machine Learning

- Probabilistic models
- Regression
- Classification
- Clustering
- Neural networks
- Graphical models
- Sequential data (Markov models)
- Combining models (Bagging and boosting)

Machine Learning

- Fuzzy Logic, Support Vector Machines
- Genetic algorithms and Genetic programming (Optimisation)
- Ant-colony optimisation, Particle Swarm Optimisation
- Baysian Statistics

The list is endless.

Algorithms

- Why learn them? *I'll just use the library right?*
- Complexity.
- Suitability to the problem at hand
- Some popular examples
  - Bloom filters
  - Skiplists
  - Hyperloglog
  - Probabilistic data structures

http://ivory.idyll.org/blog/2013-pycon-awesome-big-data-algorithms-talk.html

The cost of not understanding complexity...

- Is like not understanding Compound Interest

Distributed and Concurrent systems

- A lot of machine generated data may not fit in single system any more
- Distributed Databases
- Distributed computation
  - Map reduce
  - Take data to computation vs Take computation to data

Data stores

- Column oriented Databases – Cassandra, Vertica, MonetDB, HBase
- Key value stores – DynamoDB, Riak, Riak, Tokyo C/T, Memcache, Voldemort
- Document Stores - CouchDB, MongoDB
- In-memory data stores - Redis
- RDBMS – Vertica,
- GraphDB – Neo4J, AllegoGraph,

Programming Langauges

- R
- Python
- Julia
  - PL by CS for technical computing
- FORTRAN/C++/Java
  - Performance, Libraries
- F#
- SAS/MATLAB/SPSS etc.,

R

- F/OSS environment for Statistical computing and graphics
- Used by statisticians/data miners
- Some techinques:
  - Linear and non-linear modeling
  - Statistical tests,
  - Time series analysis,
  - Classification, clustering
- New features are added via “packages”.
- Very active contributor community
- Excellent for the mathematically literate end of Datascientist spectrum

Python

- Universal language
- Large Scientific, Data analysis community
- User friendly
- Readability, and grokability as a design feature
- Accessible to programmers and non-traditional programmers
- Opens up ML/Stats/Math/Viz to Programmers
- Opens up “regular programming” /Software Engineering/web development /data-munging/DBconnectivity/System Administration to “non-programmers”

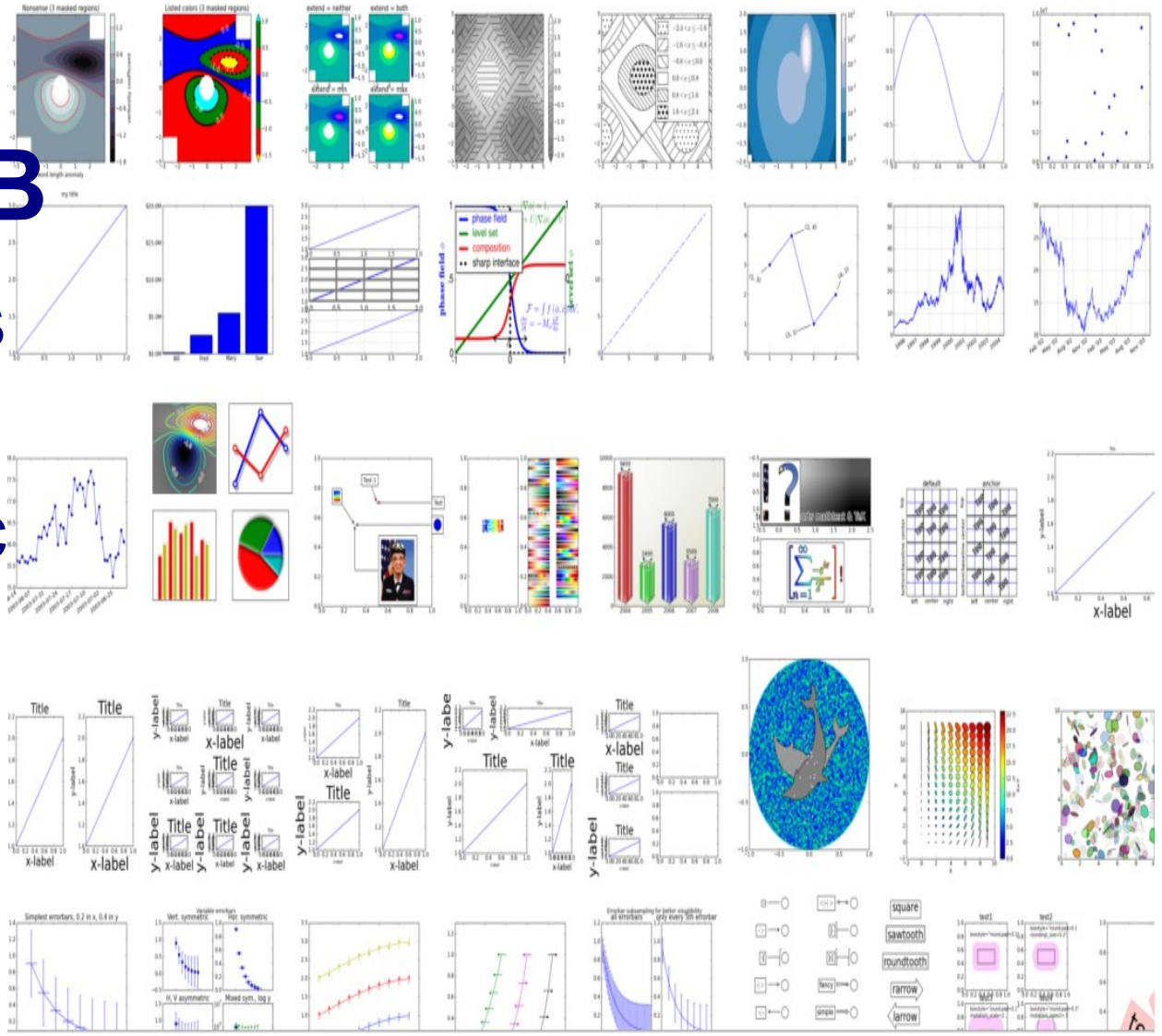
Python DS ecosystem

- Numpy
  - Backed by high performance FORTRAN/C libs
  - Linear algebra functions
- SciPy
  - Statistics
  - Optimisation
  - Signal processing
  - Scikit-learn – machine learning
- R/Py – python bridge to R
- Python Foreign Fuction Interface, C/C++ bindings.

Python DS ecosystem (2)

MATPLOTLIB

- Scientific plots
- Excellent for print/web static reporting

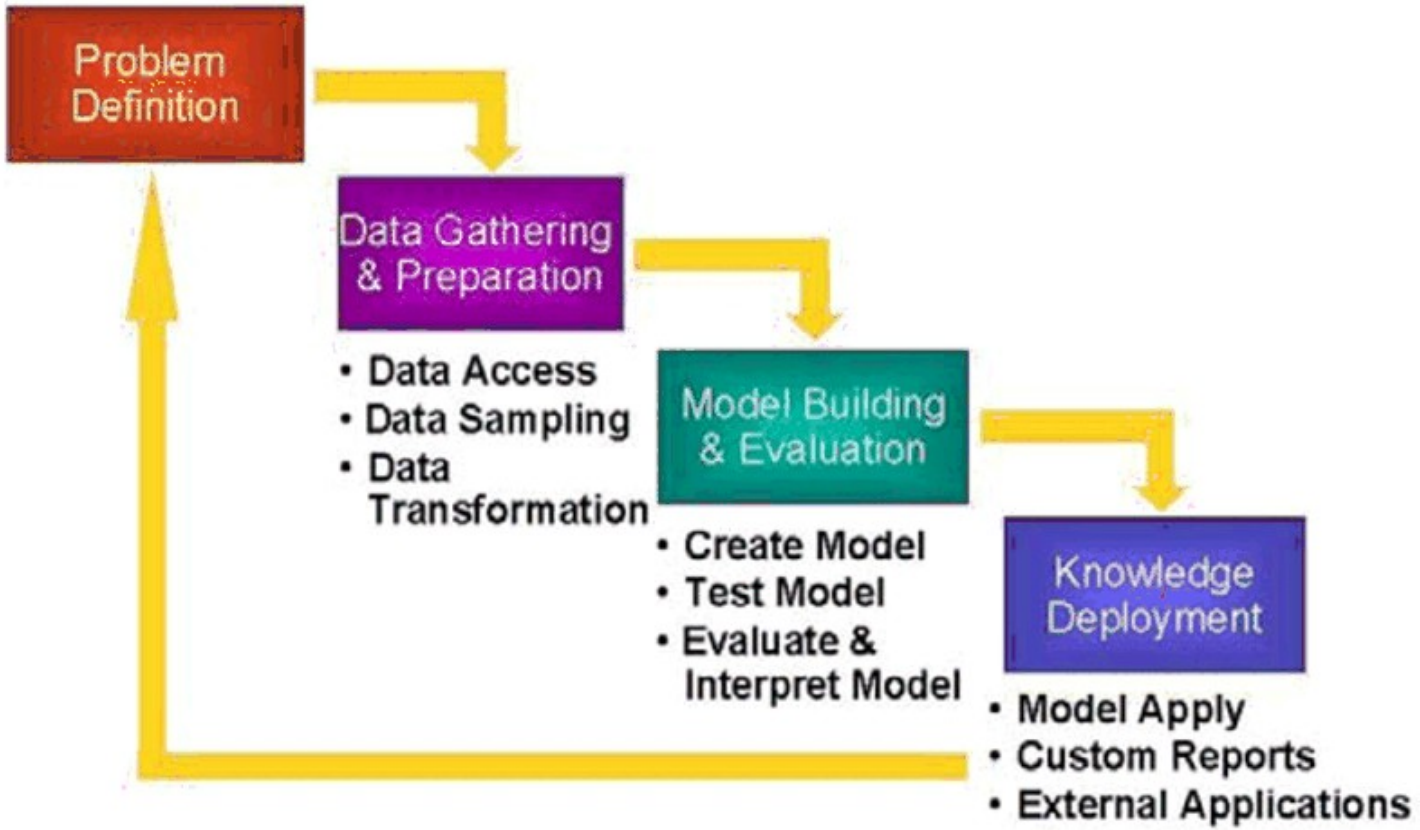


The Process

- Collecting
- Cleaning
- Analysing
- Visualising
- Publishing

http://www.slideshare.net/hhamalai/python-for-data-science

Data Mining



Credit: Oracle.com

Collecting

- Scraping the web
- Scraping printed material
- Deploying new sensors
- Subscribing to data feeds

Data sources

- Twitter data
- Social network data
- Geographical Information System data
- Personal Data – GPS, interaction with apps, user input, passive interaction
- Demographical information
- Government/Other agencies(eg: NOAA)
  - Executive Order -- Making Open and Machine Readable the New Default for Government Information
  -

Data Cleaning

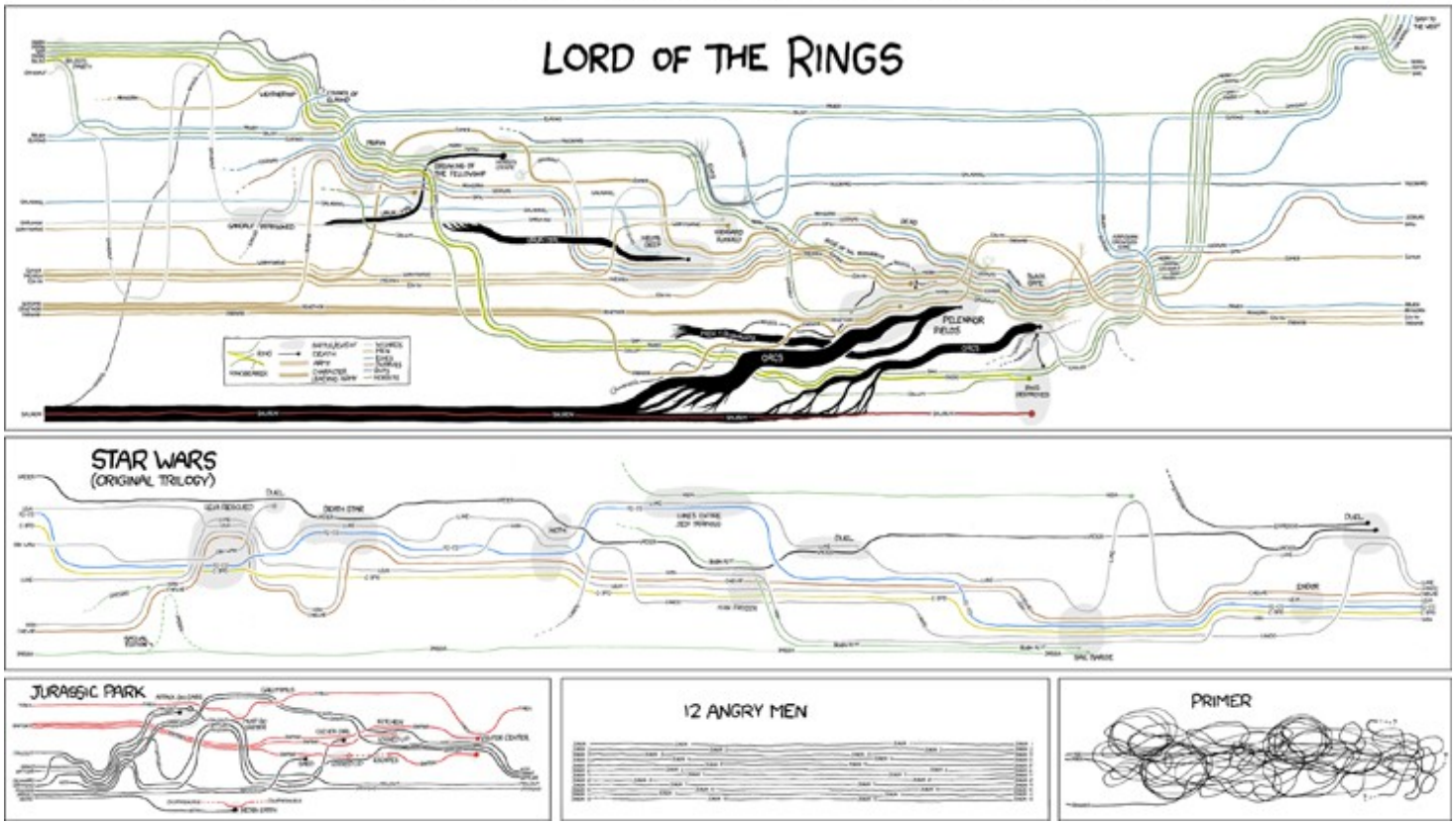
- Writing parsers
- Format conversion.
  - HTML,JSON, XML, CSV, structured data
- Text mining algorithms
- Natural language toolkit

Ghosts of the ETL past (now with zombies!!)

Analysing

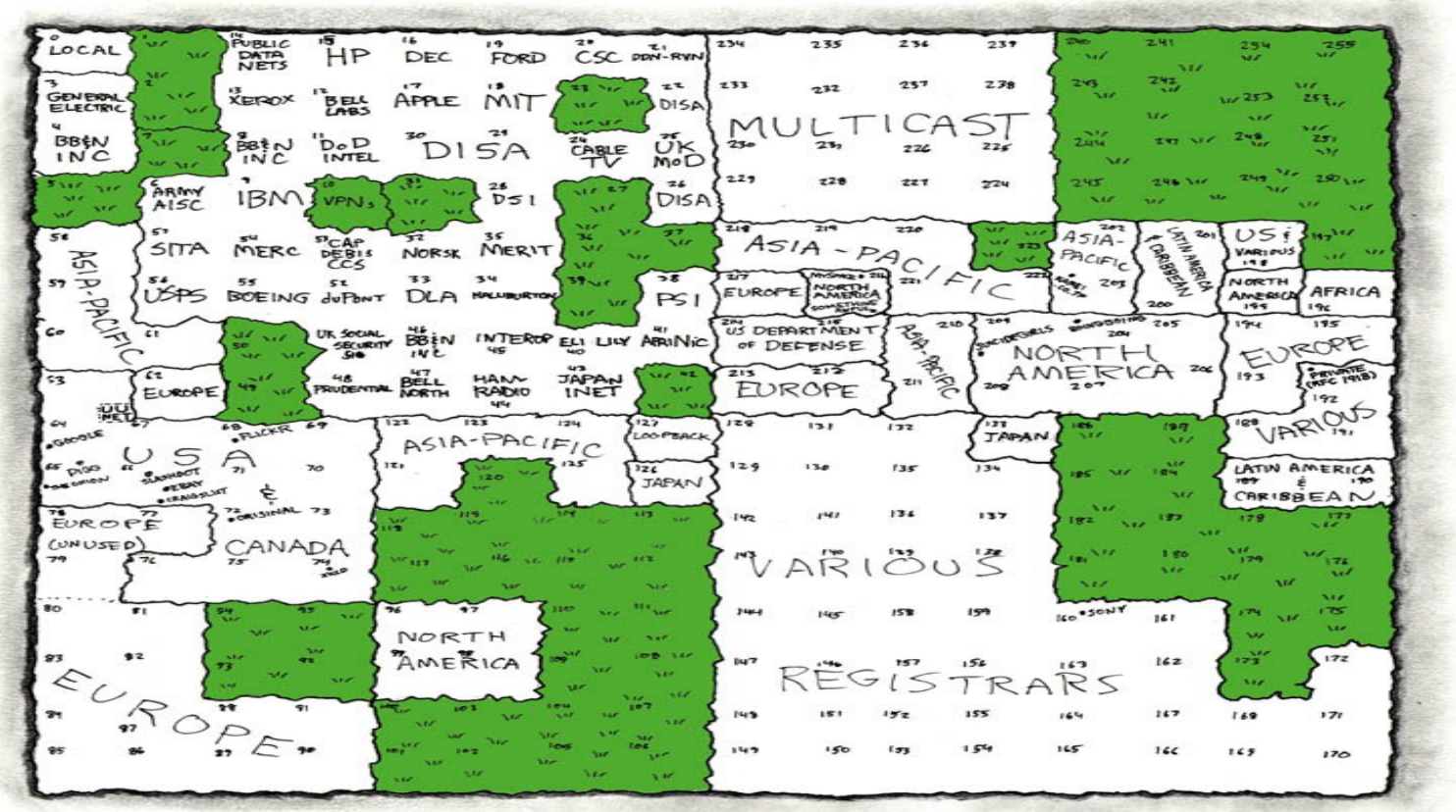
- Sampling
- “Playing”
- Iterative exploration

A good data science endeavour is about “Stories”



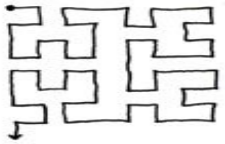
MAP OF THE INTERNET

THE IPv4 SPACE, 2006



THIS CHART SHOWS THE IP ADDRESS SPACE ON A PLANE USING A FRACTAL MAPPING WHICH PRESERVES GROUPING -- ANY CONSECUTIVE STRING OF IPs WILL TRANSLATE TO A SINGLE COMPACT, CONTIGUOUS REGION ON THE MAP. EACH OF THE 256 NUMBERED BLOCKS REPRESENTS ONE /8 SUBNET (CONTAINING ALL IPs THAT START WITH THAT NUMBER). THE UPPER LEFT SECTION SHOWS THE BLOCKS SOLD DIRECTLY TO CORPORATIONS AND GOVERNMENTS IN THE 1990's BEFORE THE RIRs TOOK OVER ALLOCATION.

0	1	14	15	16	19 →
3	2	13	12	17	18
4	7	8	11		
5	6	9	10		



= UNALLOCATED BLOCK

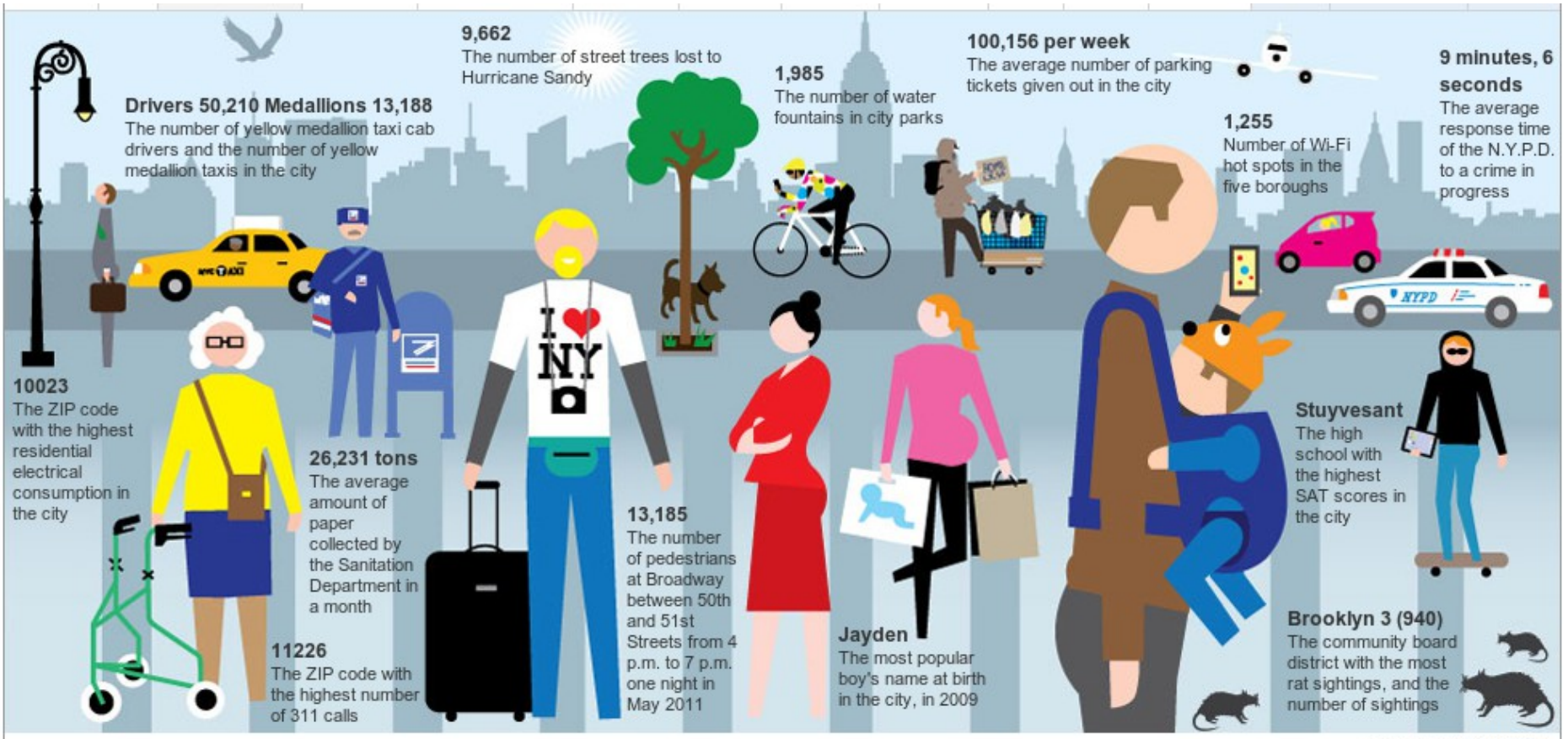


Illustration by QuickHoney

The Mayor's Geek Squad

By ALAN FEUER
Published: March 23, 2013

MOST E-MAILED | RECOMMENDED FOR YOU

Source: NYTimes, http://www.nytimes.com/2013/03/24/nyregion/mayor-bloombergs-geek-squad.html

Visualising

- Standard plots – learn to be amazed at how much line/bar charts can convey
- Domain specific plots
  - Can get technical, but have higher information density
- Multiple Channel output – web, interactive, mobile, print
  - Keep it simple.
  - Avoid “junk charts”
- cf Edward Tufte, Stephen Few, Rosling
- Less is more.
- “Infographic” - don't go overboard

Remember: The use of Visual tools is to highlight similarities, differences and “changes” and not bedazzling the user.

Many hats

- Visual artist
- Data engineer
- Programmer
- Scientist
- Mathematician
- Domain Expert

Can you be an expert in all of these?

- Unlikely
- You will still need to specialise
  - Math+Stat/CS/SysAdmin/Vis/Comm...
- Build a team around you
- Bring in expertise when you need it
- Evolve the team around you

Learning – On-line courses

- Machine Learning – Andrew Ng/Stanford/Coursera
- Data Science – Bill Howe/UoW/Coursera
- Probabilistic Graphical Models – Kohler/Coursera
- Algorithms I and II
- Statistics

The challenge is no longer about access, but about finding the time to
Learn
Find new and interesting (& profitable) problems
And apply them

Issues

- Privacy
- Control over “crown jewels”
- Legal

- Confusing the map with the territory
- Data-in, garbage out

Big Data = Big errors?

- Modernity provides too many variables, but too little data per variable. So the spurious relationships grow much, much faster than real information.
- Big data may mean more information, but it also means more false information.
- Big-data researchers have the option to stop doing their research once they have the right result.

Src: NN Taleb. http://www.wired.com/opinion/2013/02/big-data-means-big-errors-people/