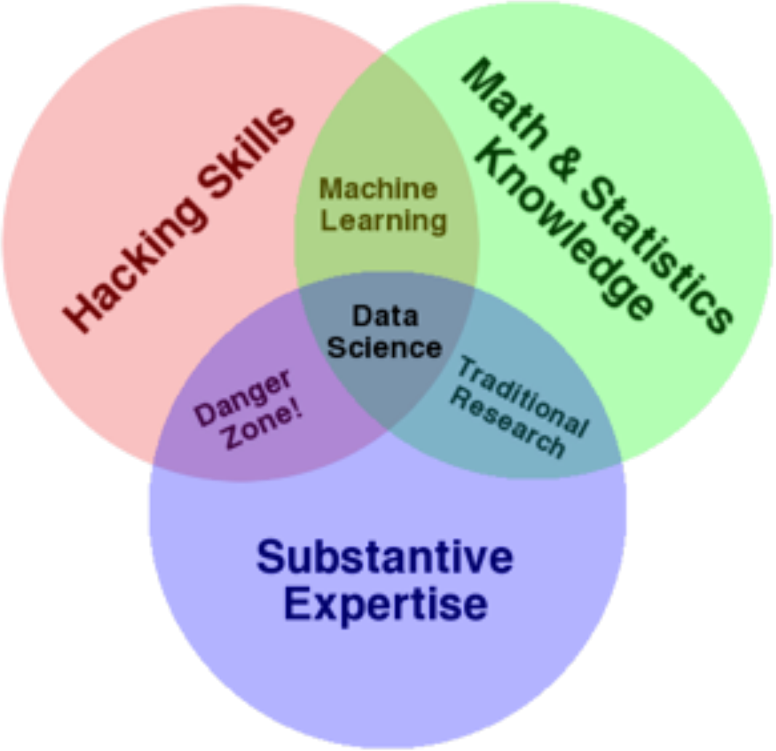
Training Strategies for Data Science

Data Science Meet-Up

Judith Mueller, PhD



6. May 2015, contact: judith.mueller1@gmail.com

My Background

14 years in Research and Development in Semiconductor Industry at Motorola and Freescale (Austin, France, Burlington)

- Data Analysis, Technology Validation (CMOS technology of advanced nodes)

- Device Model Parameter Extraction (Regression, Neutral Nets), Electrical Circuit Simulations

- Scripting and Automation

PhD in Physics at McGill University (Montreal): Stress-Induced Morphological Instabilities

- Computational Physics: Algorithm, Model development, Monte Carlo Simulations

- Statistical Physics, Non-linear Dynamics (Chaos), and Pattern Formation

Diplom/Masters in Physics at RWTH Aachen (Germany): Simulation of Stochastic Differential Equation with Colored Noise

- Stochastic Processes, Markoff Chain, Monte Carlo Simulations, Algorithm

- Dynamical Systems, Time Series Analysis, Autocorrelation

Data Science: Why now?

Availability and Storage of Data:

- Internet (faceBook, Google, etc),
- mobile devices,
- sensors, etc
- cloud and availability of inexpensive memory

Technology advances:

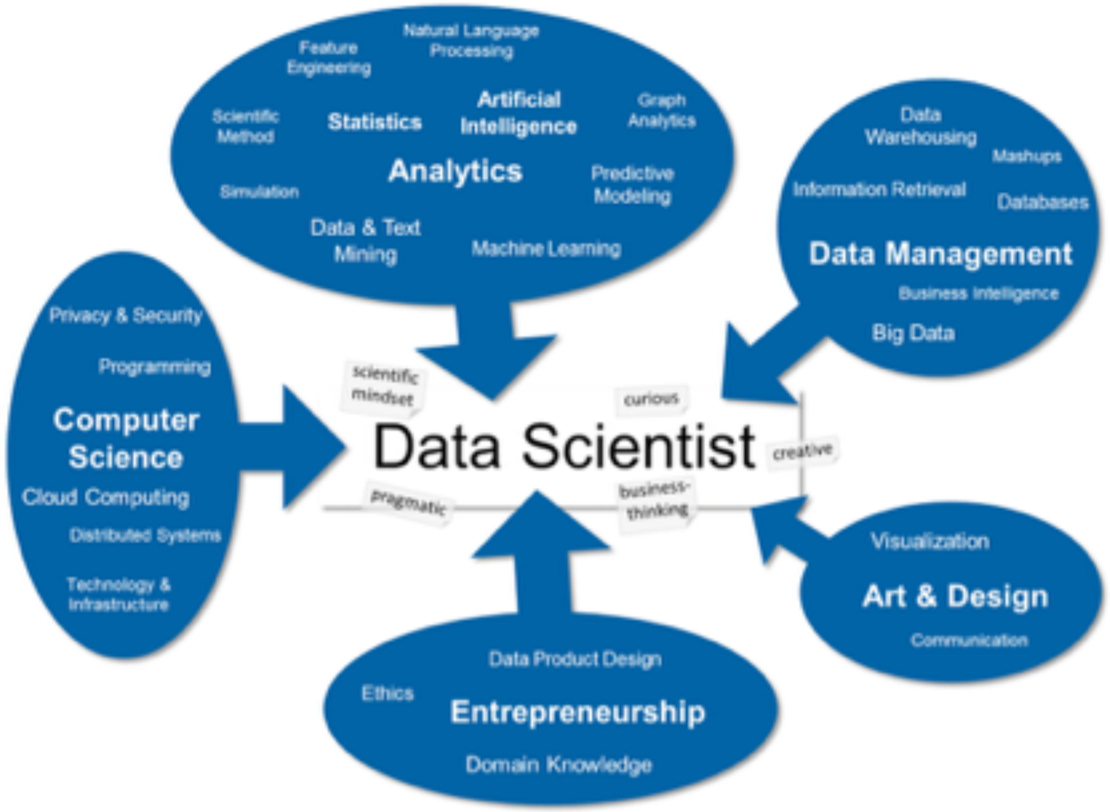
- hardware: computing power, speed, price, and memory size
- data management: dataBases, cloud
- software: algorithms (machine learning, Hadoop ecosystem)

Big Data Timeline by WinShuttle

Infograpic: Big Data just beginning to explode by CSC

13 TED talks: Making sense out of too much data

What is a Data Scientist?



A data scientist is someone who is
better at statistics than any software engineer and
better at software engineering than any statistician.

Turns data into actionable results.

Broad definition, and generally recognized that a team with different expertise is required.

Useful links:

- Data Science Central - Online Resource for Big Data Practitioners.
- Masters in Data Science in the real world. - Overview on history and challenges by industry: biotech, heath care, pharmaceutical, retail, manufacturing, telecommunication, insurance, government, energy etc)

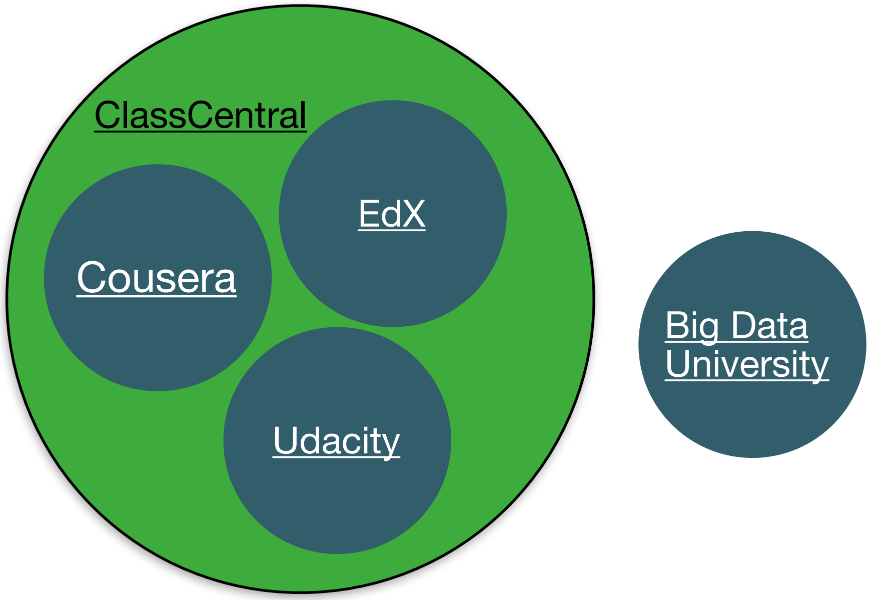
How to become a Data Scientist?

Masters of Data Science: formal, expensive

Course at UVM:

- Introduction to Data Science and Visualization (James Bagrow)
- Computer Science Deprtment: Machine Learning, Artificial Intelligence, etc

Free Online classes: MOOC (Massive Open Online Course):



ClassCentral

EdX

Cousera

Udacity

Big Data University

Pay for Certification

Useful links:

- Blog DataCamp: How to become a data scientist?
- DataScienceCentral: How to become a data scientist for free
- The Field Guide to Data Science by Booz Allan

Combination of theory, techniques, tools, and experience is required.

Introduction to Data Science

- Introduction to Data Science (Cousera) by Bill Howe (University of Washington)

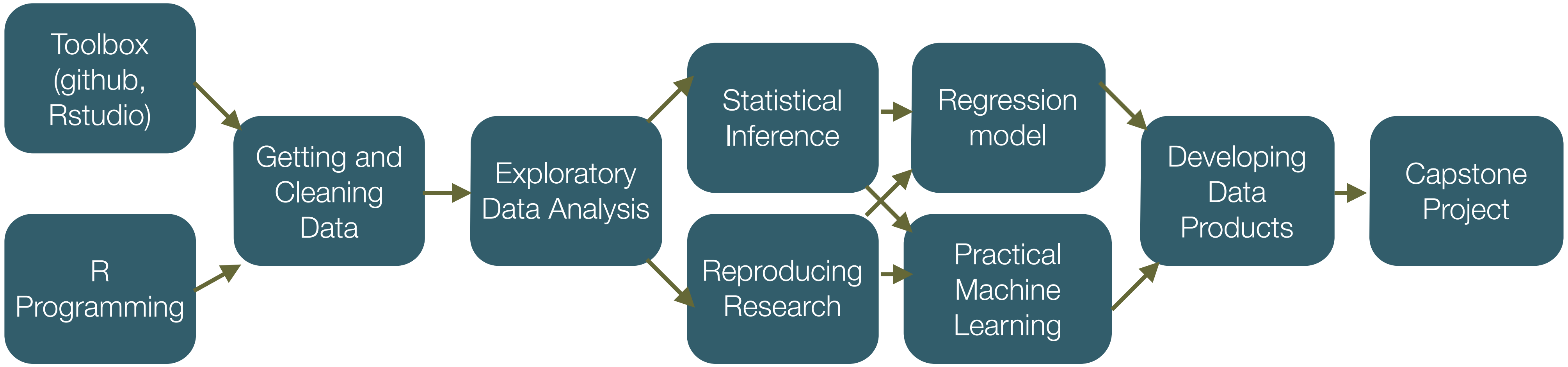
Data Manipulation at Scale
Databases and the relational algebra
Parallel databases, parallel query processing, in-database analytics
MapReduce, Hadoop, relationship to databases,
algorithms, extensions, languages
Key-value stores and NoSQL; tradeoffs of SQL and NoSQL

Analytics
Topics in statistical modeling: basic concepts, experiment design, pitfalls
Topics in machine learning:
 supervised learning (rules, trees, forests, nearest neighbor, regression),
 optimization (gradient descent and variants),
 unsupervised learning

Communicating Results
Visualization, data products, visual data analytics
Provenance, privacy, ethics, governance

Special Topics
Graph Analytics: structure, traversals, analytics, PageRank,
community detection, recursive queries, semantic web

- Data Science Specialization (Cousera) by Jeff Leek, Roger Peng, and Brian Caffo from John-Hopkins University



Provides overview of flow.
Provides How to but little theoretical background.
Good starting point.

Learn to Code

- Scripting Language: Python, Perl
- Learning Python:
  - CodeAcademy
  - PythonGuide (introduction)
- Statistical program: R
  - R-Bloggers
  - DataCamp
  - TryR CodeSchool
- Basic database concepts: SQL, NoSQL, HDFS
- File Formats: XML, JSON
- help: stackoverflow

Basic Statistics Course

- Data Analysis and Statistical Inference (Cousera) by Mine Cetinkaya-Rundel (Duke University)

  Basic Introduction to data analysis
  Probability and distributions,
  Inference: Central Limit theorem, confidence intervals, hypothesis tests, ANOVA (analysis of variance)
  Linear and multiple regression

- Statistics One (Cousera) by Andrew Conway (Princeton)

  Comprehensive introduction to fundamental concepts in statistics.
  Distributions, summary statistics, correlation,
  regression, binary logistic regression
  Null Hypothesis Significance Tests (NHST), Central limit theorem, Confidence interval, Multiple regression, ANOVA ,
  Generalized Linear Model

- Statistics: Making Sense of Data (Cousera) by Alison Gibbs (University of Toronto)

  An applied statistics course that teaches the complete pipeline of statistical analysis.
  Data collection and sampling
  Probability: Probability models, the normal distribution, the Central Limit Theorem, sampling distributions.
  Confidence Intervals: Confidence intervals and sample size estimation for proportions and means.
  Tests of significance: estimates of significance, power and sample size estimation for proportions and means

- Statistics Blog

Machine Learning
Data Mining, Pattern Recognition and Predictive Analytics

Algorithmic approach to data analysis: Classification, clustering, feature learning, etc.

General info: How do I learn Machine Learning

- Statistical Learning by Trevor Hastie, Robert Tibshirani, Jerome Friedman (Stanford)

This is an introductory-level course in supervised learning with a focus on regression and classification methods.
linear and polynomial regression, logistic regression and linear discriminant analysis; nonlinear models, splines and generalized additive models;
cross-validation and the bootstrap, model selection and regularization methods (ridge and lasso);
classification methods: tree-based methods, random forests and boosting; support-vector machines.
Some unsupervised learning methods are discussed: principal components and clustering (k-means and hierarchical):

- Machine Learning (Cousera) by Pedro Domingos (University of Washington)

The course covers the main supervised learning techniques, and two main classes of unsupervised learning methods.
Supervised Learning: decision trees, rules, instances, Bayesian techniques, neural networks, model ensembles, and support vector machines.
Unsupervised learning methods: clustering and dimensionality reduction.

- Machine Learning (Cousera) by Andrew Ng (Stanford)

This course provides a broad introduction to machine learning, data mining, and statistical pattern recognition.
(i) Supervised learning: parametric/non-parametric algorithms, support vector machines, kernels, neural networks;
(ii) Unsupervised learning: clustering, dimensionality reduction, recommender systems, deep learning,
(iii) Best practices in machine learning (bias/variance theory; innovation process in machine learning and AI).

Recommendation: Breath and Depth.

Implement an algorithm yourself, at least once, instead of using a black box.

Big Data, Hadoop Ecosystem

Big Data: Volume, Velocity, Variety

- Mining Massive Datasets (Cousera) by Jure Leskovec, Anand Rajaman, Jeff Ullman (Stanford)

  MapReduce, Link Analysis -- PageRank, Locality-Sensitive Hashing, Distance Measures, Nearest Neighbors
  Data Stream Mining, Analysis of Large Graphs, Recommender Systems, Dimensionality Reduction, Clustering
  Computational Advertising, Support-Vector Machines, Decision Trees, MapReduce Algorithms

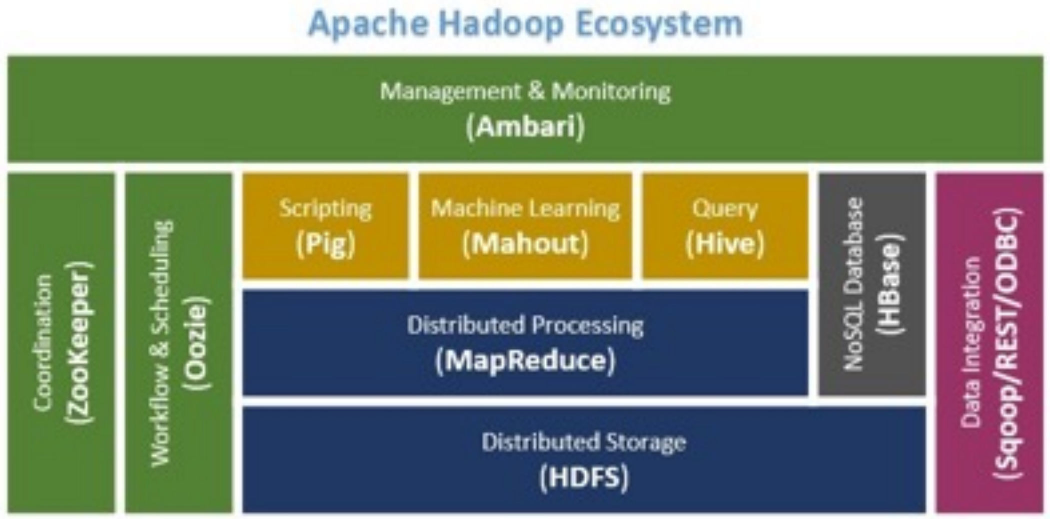
- Introduction to Hadoop and MapReduce (Udacity) built by Cloudera

  The basics of HDFS and Hadoop ecosystem, MapReduce and Hadoop cluster.
  Writing MapReduce programs to answer questions about data.
  MapReduce design patterns.

- Hadoop (Apache)

- Big Data University courses:
  - Big Data Fundamentals
  - Hadoop Fundamentals
  - Introduction to MapReduce Programming
  - Introduction to Spark Fundamentals

- Conference: Big Data TechCon in Boston

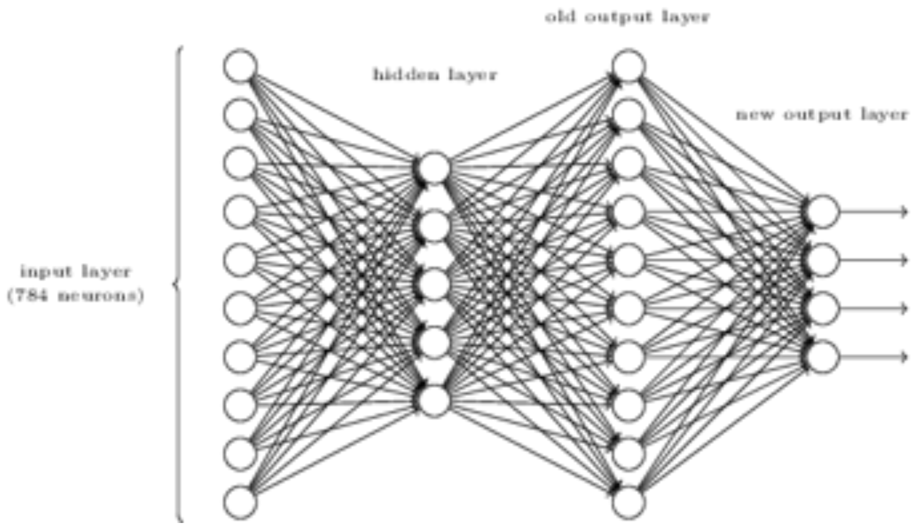


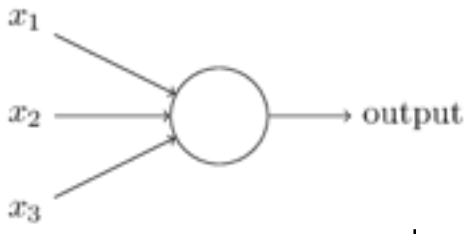
Deep Learning (AI)

Used for speech and object recognition, image segmentation, modeling language and human motion, etc

Based on Artificial Neural networks with many more levels and connections than in the 80s.

- Neural Networks for Machine Learning (Cousera) by Geoffrey Hinton (University of Toronto)
- Stanford panel: Deep Learning: Intelligence from Big Data Clarifai, Google Brain
- TED talk by Fei-Fei Li (Stanford) : How we are teaching computers to understand pictures





perceptron

Visualization and Story Telling

- Data Visualization (Cousera) by John C. Hart (University of Illinois at Urbana Champlain)

  Visualization Infrastructure (graphics programming and human perception)
  Basic Visualization (charts, graphs, animation, interactivity)
  Visualizing Relationships (hierarchies, networks)
  Visualizing Information (text, databases)

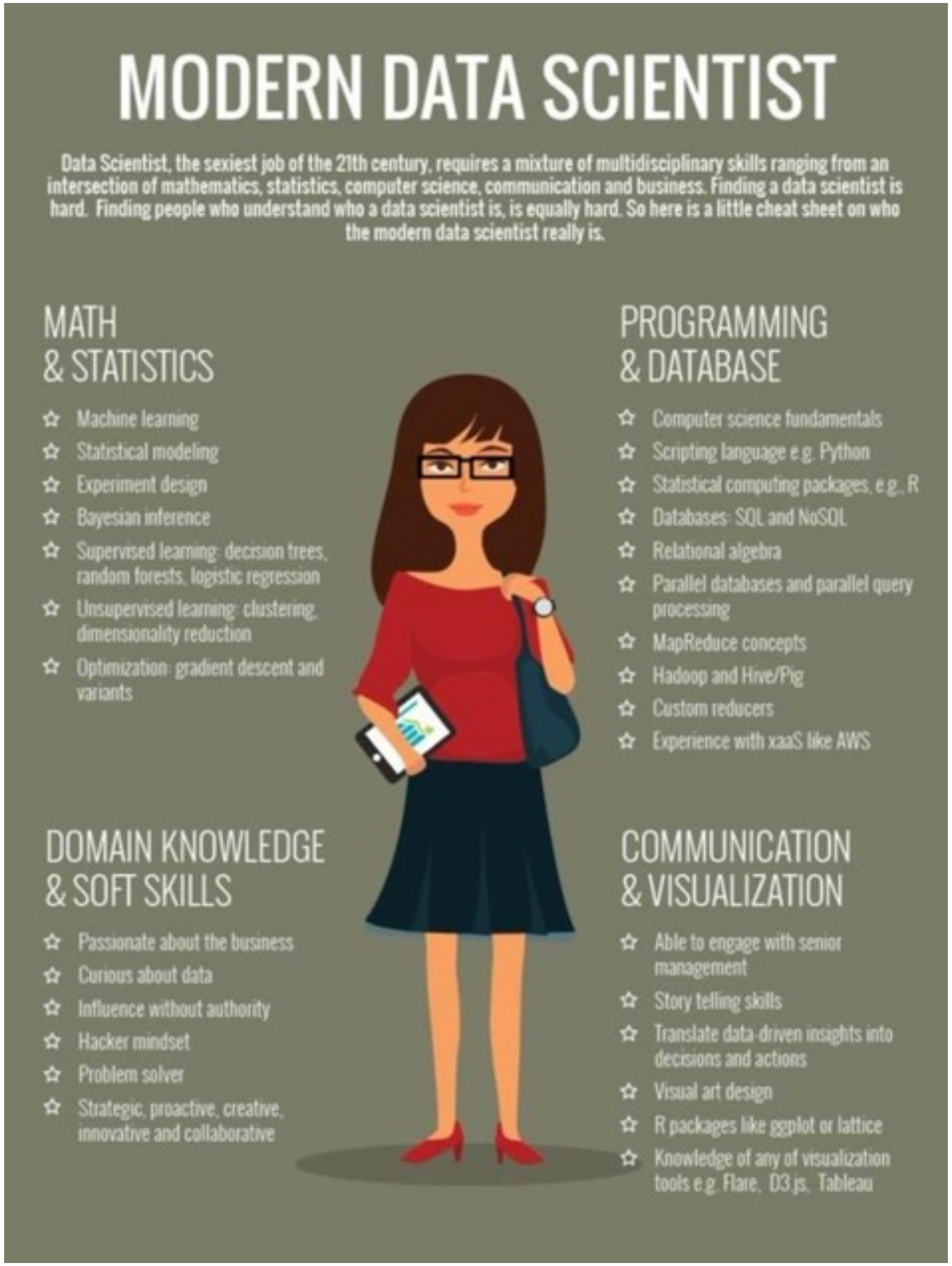
- d3 tutorial (d3 - a library used to create interactive data visualizations in the browser)
- Burlington JS: Introduction to d3 (Burlington JavaScript meeting: May 13th at 6pm)
- Check out Shiny from Rstudio (Turns your analyses into interactive web applications)
- Hans Rosling:
  - TED talk: The best Stats you have ever seen, and
  - documentary: The Joy of Stats

Get your hands dirty

- Understand capability and limitation of different approaches/tools,
- Start simple and with few data first (variety more important than volume),
- Understanding algorithms on a small scale,
- Add volume and complexity later,
- Play with data set,
    - https://www.data.gov/
    - http://www.statsci.org/datasets.html
    - http://www.lib.jmu.edu/resources/statistics_datasets.aspx
    - http://bagrow.com/dsv/datasets.html
- understand and engage with your data,
- Tell a story with your data.
- Participate in Kaggle Competition kaggle

Connect with Community

- Participate in Meet-ups in Burlington, maybe Montreal
- Attend conferences: 9 Big Data Conferences
- Attend seminars of Complex Systems Group at UVM
- Subscribe to newsletters: DataScienceCentral
- Follow VT Tech Jam and Vermont Technology Alliance



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau