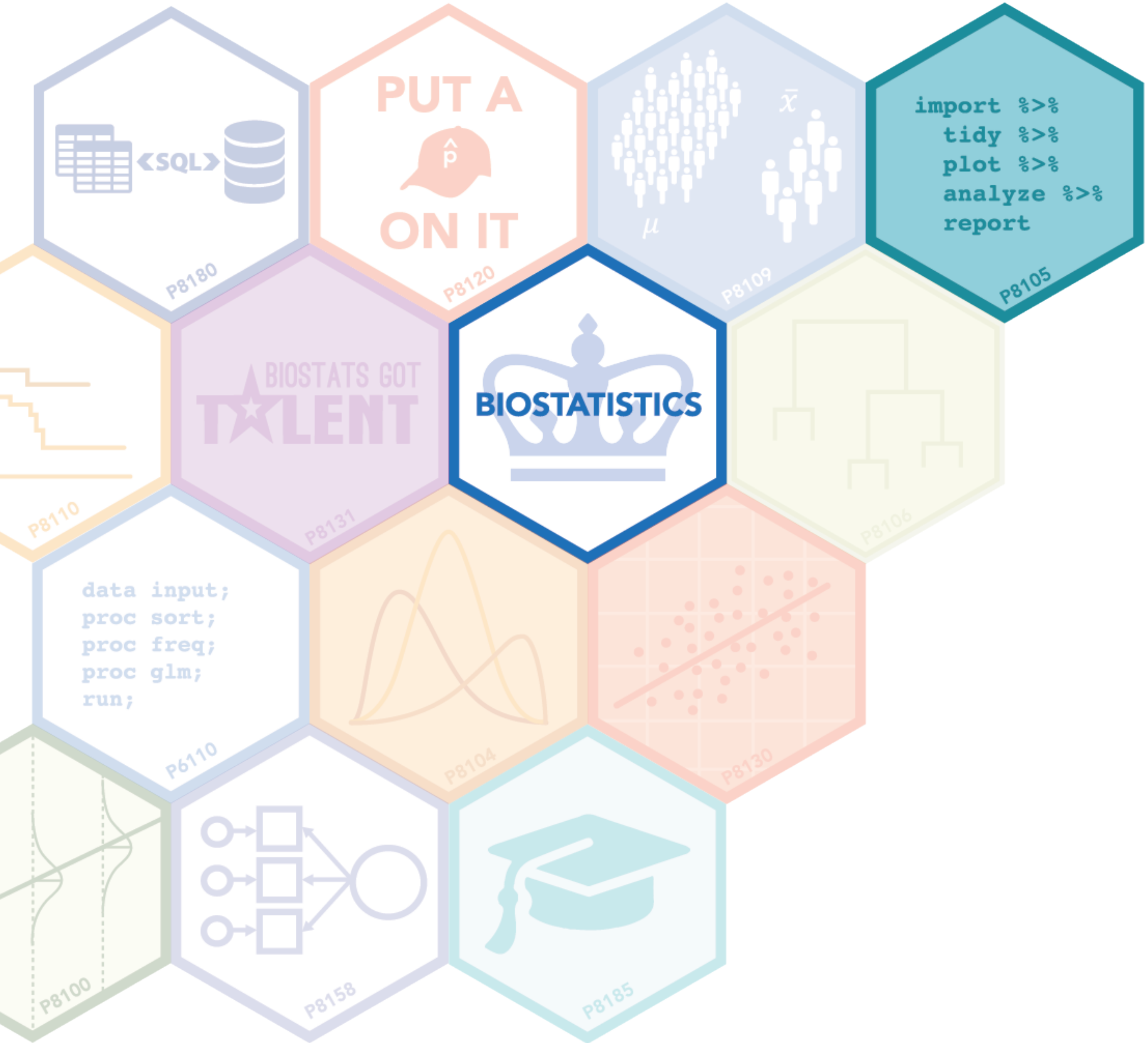


“WHAT IS DATA SCIENCE?” REVISITED

Jeff Goldsmith, PhD

Department of Biostatistics

Some not great definitions

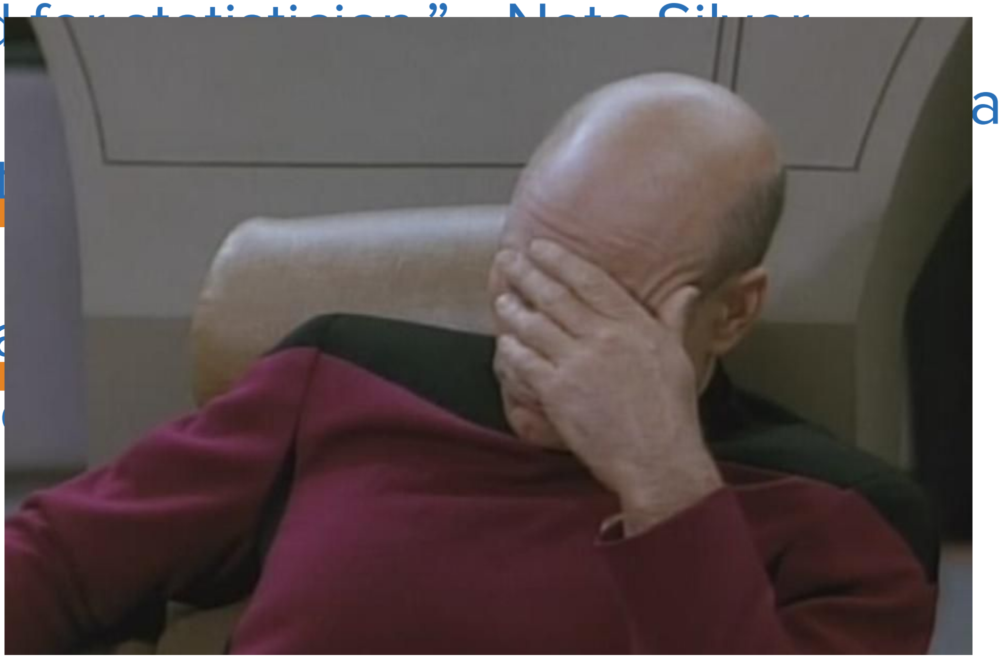
- Data science = statistics
- Data science = computer science
- Data science = machine learning
- Data science = statistics + computer science + machine learning
- Data scientists are big data wranglers
- “A data scientist is just a sexier word for statistician.” –Nate Silver
- “A data scientist is a better computer scientist than a statistician and is a better statistician than a computer scientist.”
- “A data scientist is a statistician who is useful” – Hadley Wickham
- A data scientist is a good statistical analyst
- A data scientist is a statistician who codes in python

Some not great definitions

- Data science = statistics
- Data science = computer science
- Data science = machine learning
- Data science = statistics + computer science + machine learning
- Data scientists are big data wranglers
- “A data scientist is just a sexier word for statistician.” –Nate Silver
- “A data scientist is a better computer scientist than a statistician and is a better statistician than a computer scientist.”
- “A data scientist is a statistician who is useful” – Hadley Wickham
- A data scientist is a good statistical analyst
- A data scientist is a statistician who codes in python

Some not great definitions

- Data science = statistics
- Data science = computer science
- Data science = machine learning
- Data science = statistics + computer science + machine learning
- Data scientists are big data wranglers
- “A data scientist is just a sexier word for statistician.” - Nate Silver
- “A data scientist is a better compute a better statistician than a computer scientist
- “A data scientist is a statistician who
- A data scientist is a good statistical a
- A data scientist is a statistician who



Maybe pictures will help?

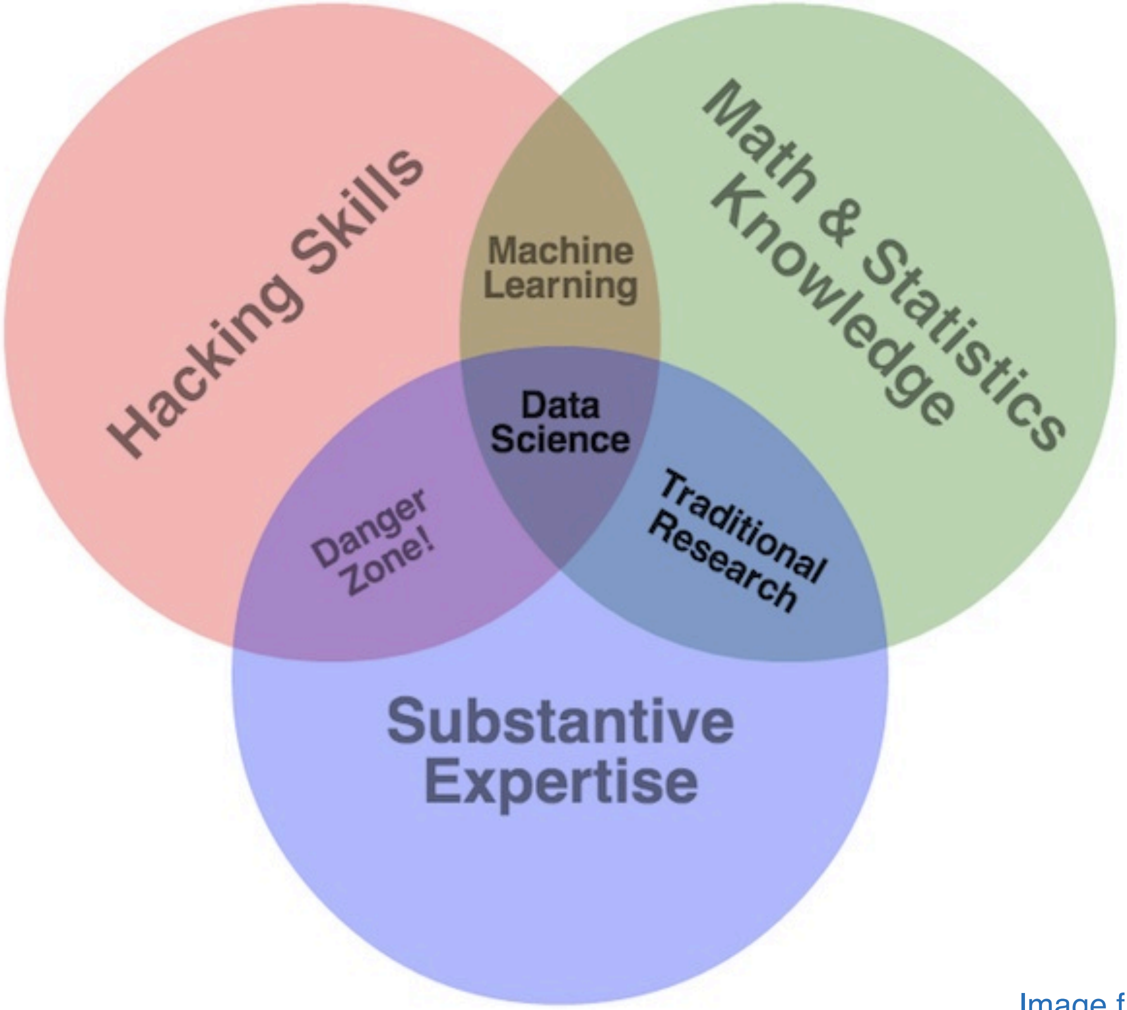
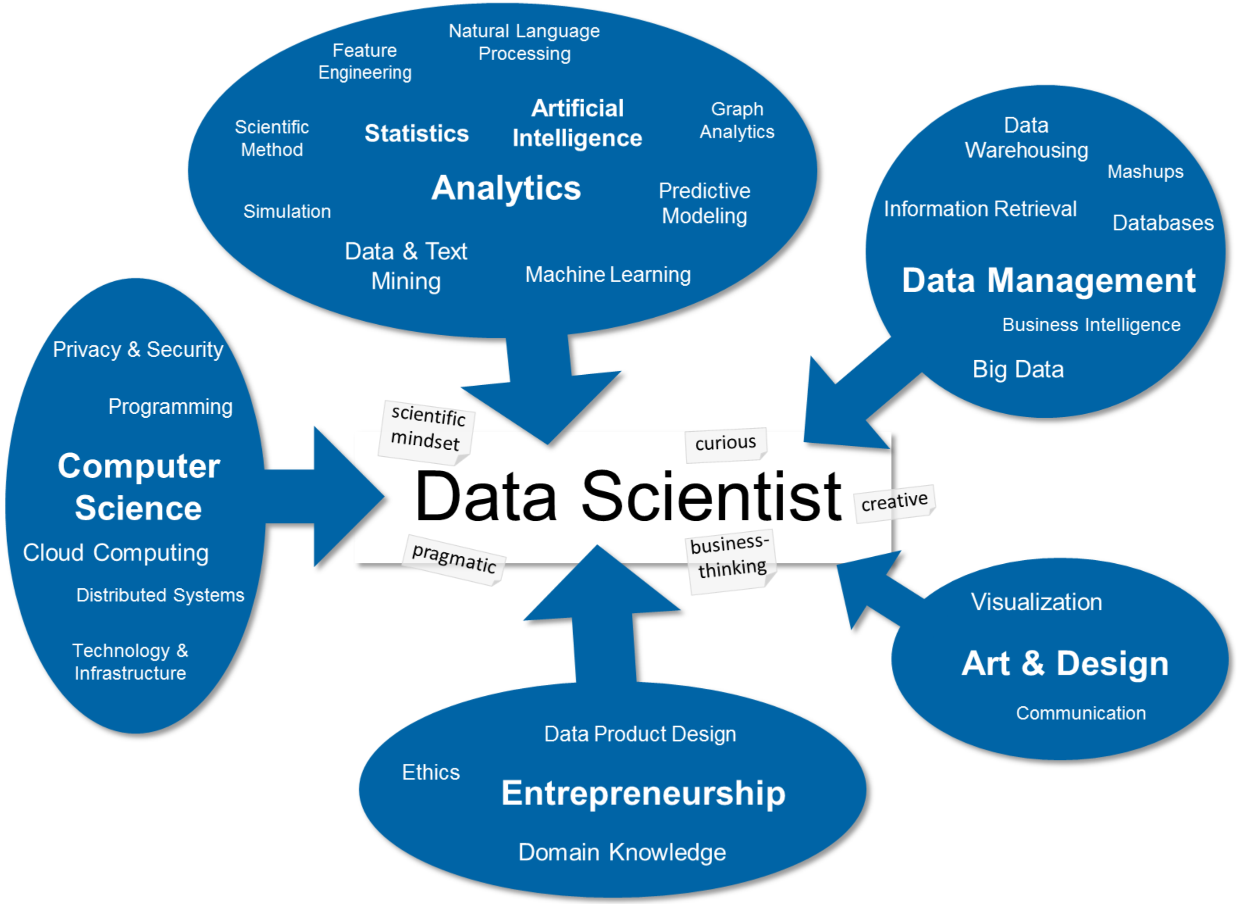


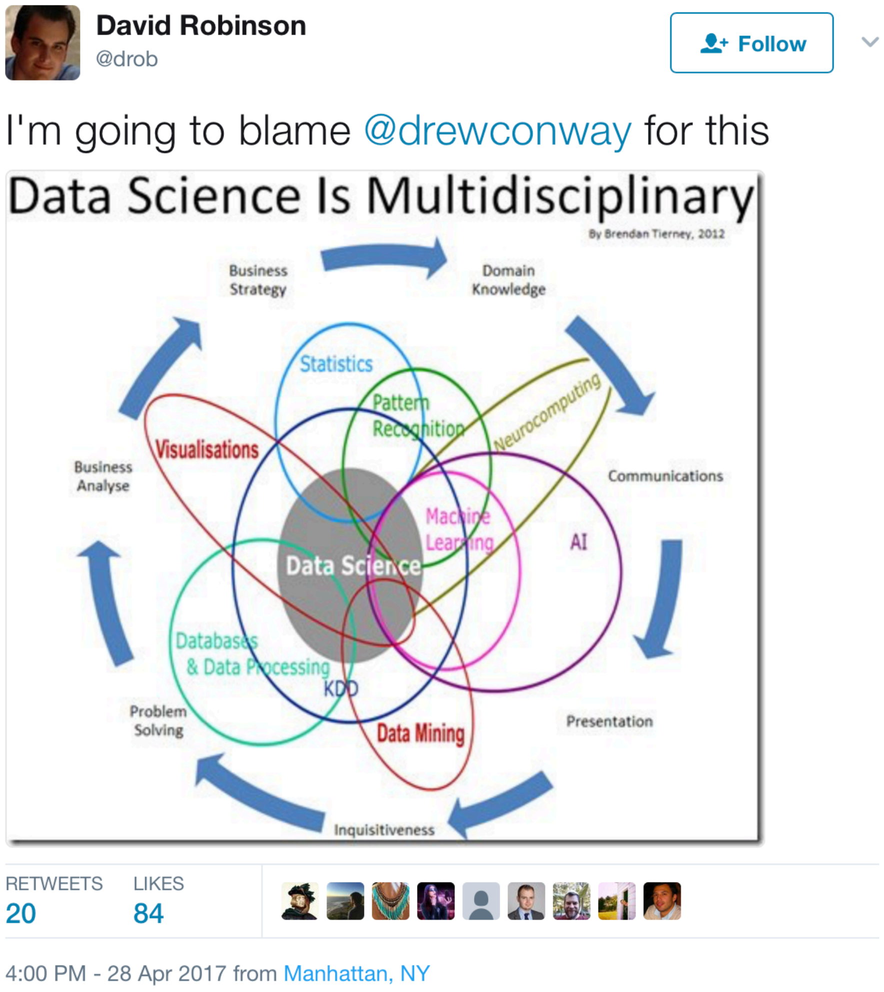
Image from Drew Conway

Maybe pictures will help?



https://blog.zhaw.ch/datascience/the-data-science-skill-set/

Maybe pictures will help?



David Robinson
@drob

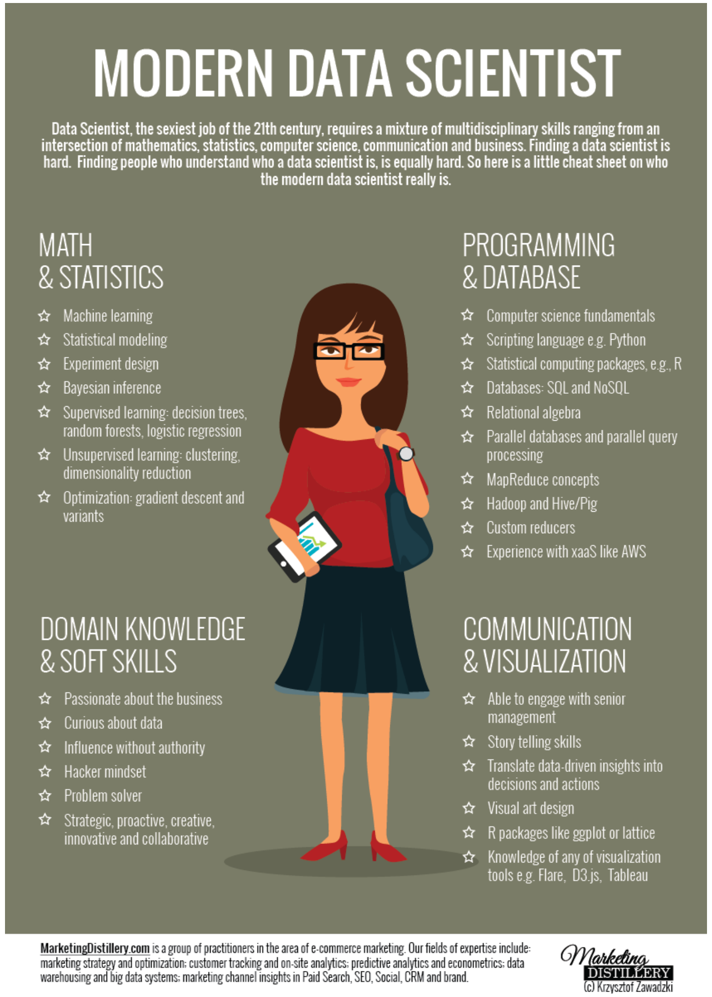
I'm going to blame @drewconway for this

RETWEETS 20 LIKES 84

4:00 PM - 28 Apr 2017 from Manhattan, NY

From twitter

Maybe pictures will help?



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

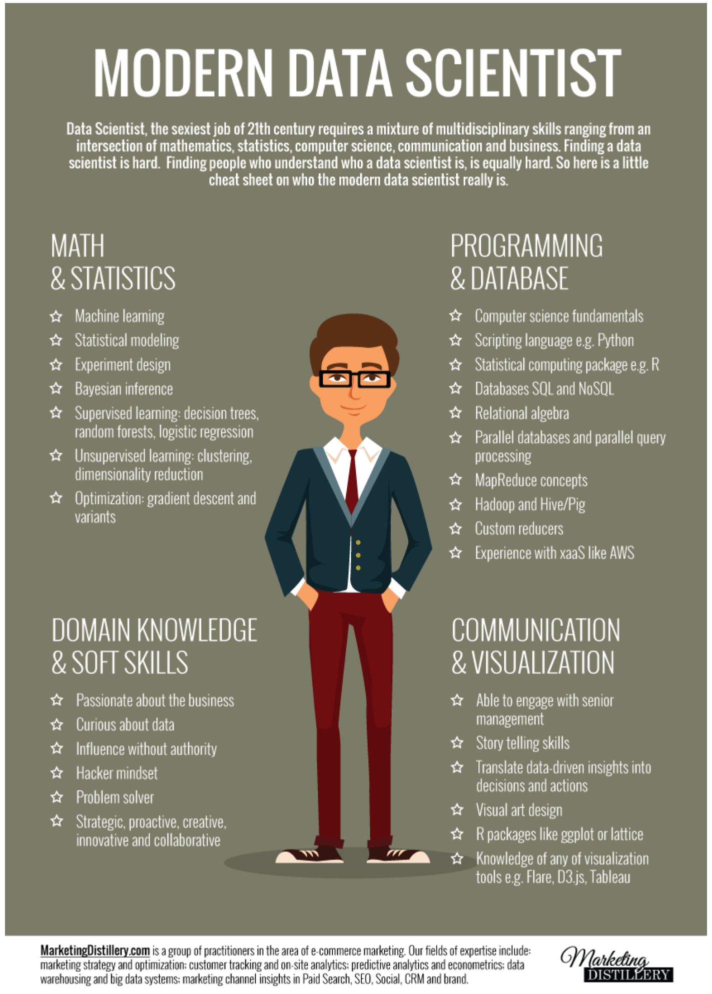
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

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

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY
(c) Krzysztof Zawadzki



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 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 package 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

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

Recurring themes

- You need “data skills”
  - Data wrangling
  - Reproducibility
  - Communication
  - Analytics and modeling
- These have been the focus of this course and others, and will continue to be the focus

- You also need a mindset
  - Intellectual curiosity
  - Ability to solve problems
  - Interest in domain, even empathy with collaborators

Problem solving

“I’ve interviewed a lot of people over the years…. Recently, when people have an interview, I ask a single question that I think tries to get at the point of problem solving. The question I ask is along the lines of ‘[Imagine you had access to a database of 100 million mobile devices.] What questions would you ask? What types of things do you think you could learn, and how would you go about doing it?’”

From “How Industry Views Data Science Education in Statistics Departments”, Chris Volinsky’s JSM 2015 talk

Practice problem solving

- You can (and should) practice having a mindset, or a style of thinking
  - Make a habit of asking yourself what you would like to do with a data resource
  - Think about how you would accomplish it

- Be on the lookout for cool projects, and learn from them
  - Pay attention to the thought process, not just the specific tools

- Many projects need overlapping skill sets
  - You don’t have to be a domain expert yourself, but you may need to work with one
  - You’ll also have to communicate effectively with that person, which means at least taking an interest