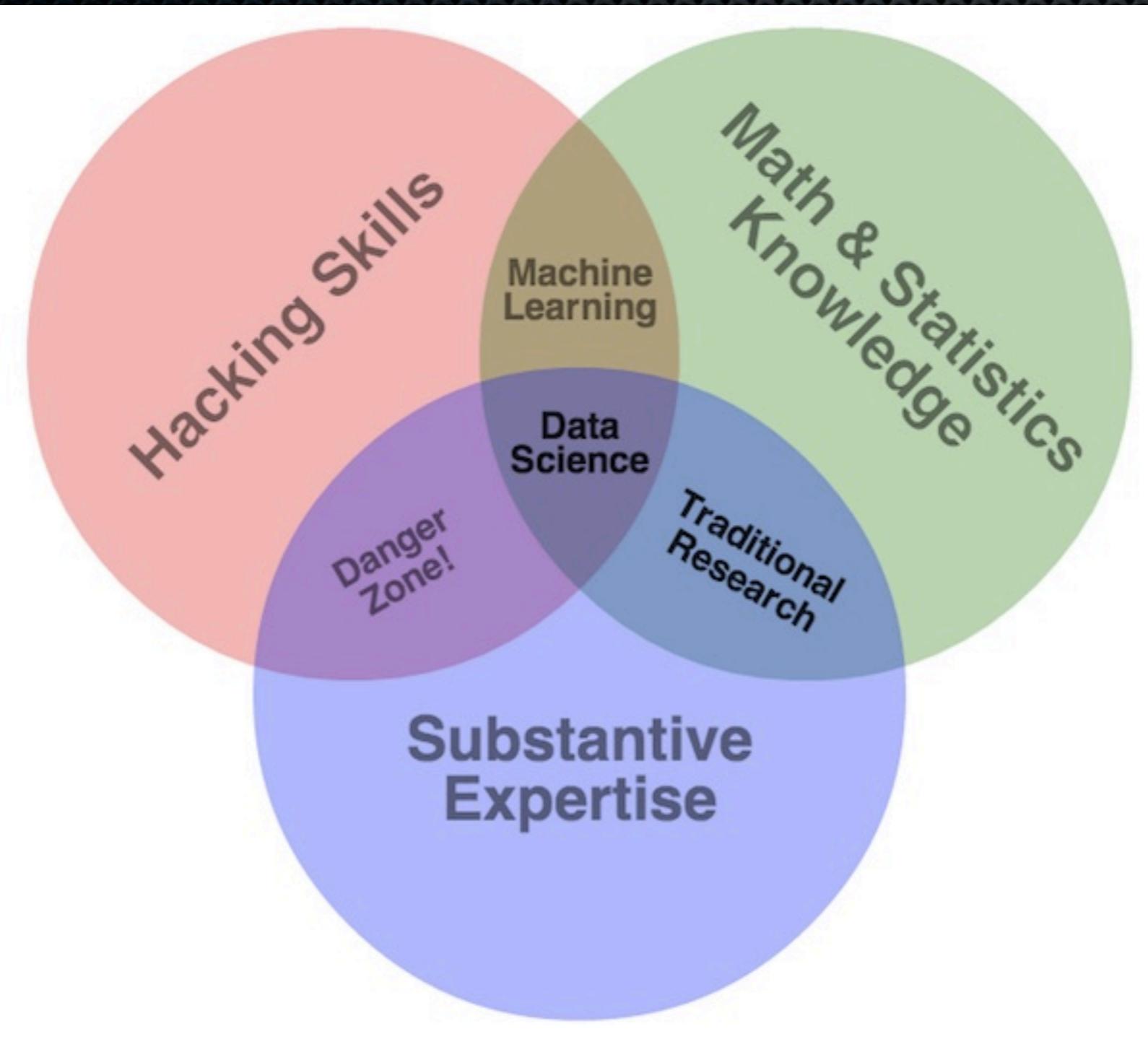


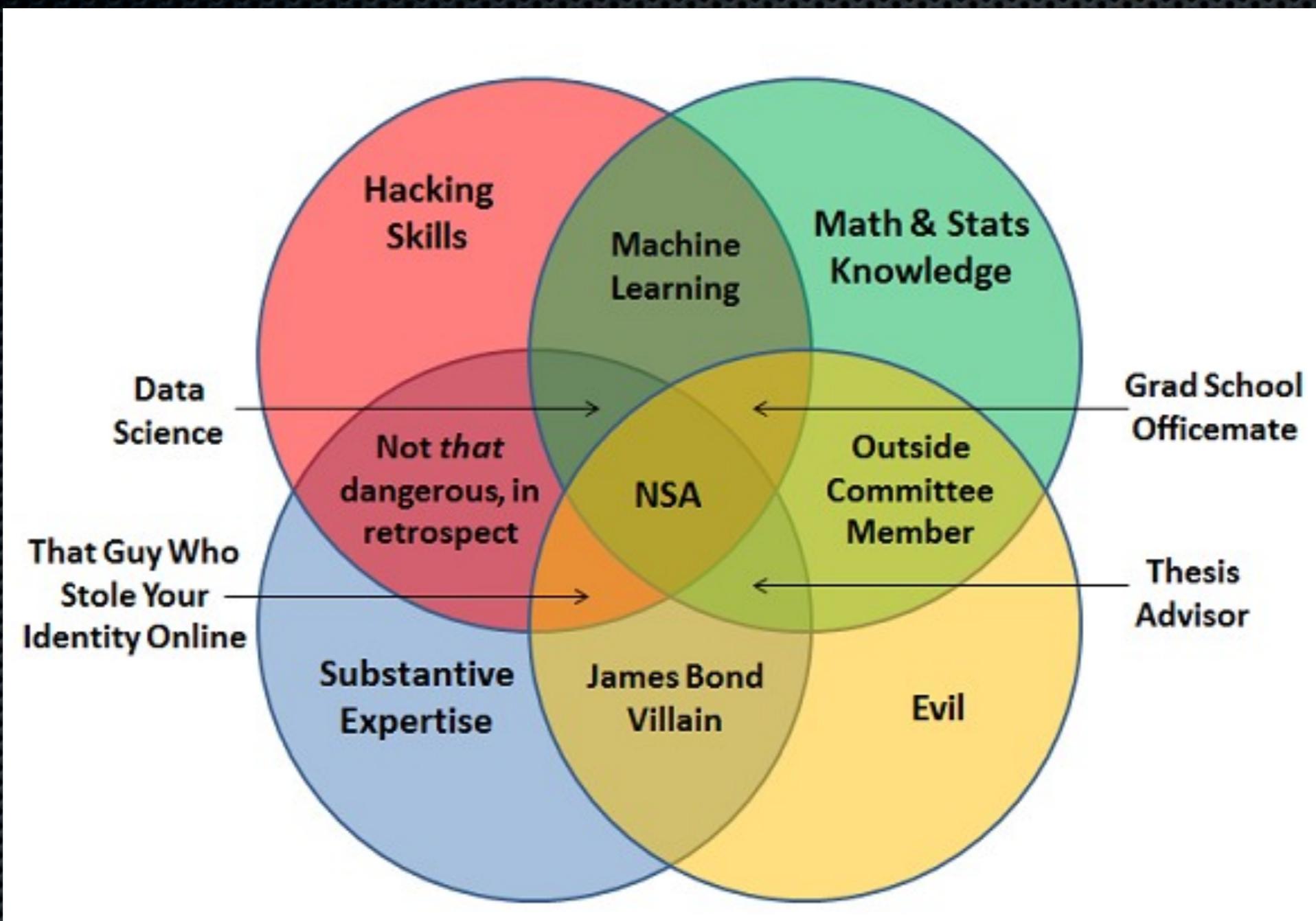
Practical Data Science

May 2014

@jattenberg

<http://practical-data-science.tumblr.com/>





What is a Data Scientist?



Guest

Subscribe today and get access to all current articles and HBR online archive.

THE MAGAZINE

October 2012

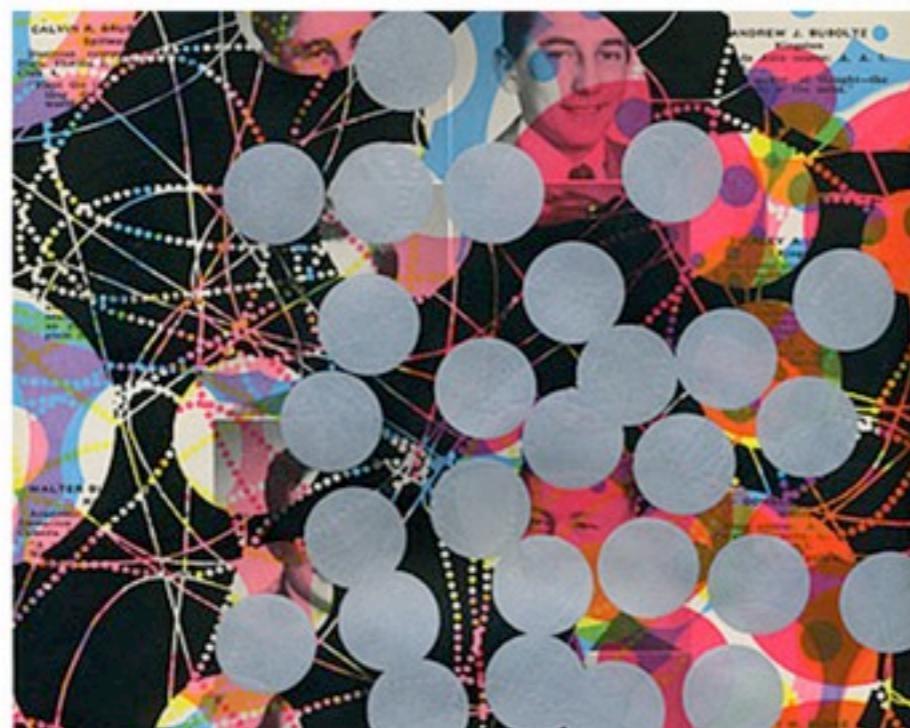


ARTICLE PREVIEW To read the full article, [sign-in](#) or [register](#). HBR subscribers, click [here to register for FREE access »](#)

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (96)



Artwork: Tamar Cohen, *Andrew J Buboltz*, 2011, silk screen on a page from a high school

Source: <http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1>

What is a Data Scientist?

- Hadoop / Database Engineer
- Statistician / Analyst
- Data Visualization Expert
- Applied Machine Learning / Data Mining Researcher
- Data Mining Product Developer

(depends on who you ask)

Example Data Science Applications

- Churn modeling
- Medical Diagnoses
- Content Clustering
- Recommender Systems
- Customer Segmentation
- AB Experimentation / Bandit Experimentation

Example Data Science Applications

- CLV Estimation
- Targeted Marketing
- Classification (eg, spam filtering)
- Search / Information Retrieval
- Parameter Estimation
- etc

A Sample DS Workflow

- Gather data
- Understand data
- Opportunity sizing
- Product Planning
- Modeling
- System deployment
- Analysis & evaluation

A Sample DS Workflow

- Gather data
- Understand data
- Opportunity sizing
- Product Planning
- Modeling
- System deployment
- Analysis & evaluation

Implementation!

A Data Scientist's Toolbox

- A terminal & operating system
- Databases
- General programming languages
- Machine learning and statistical libraries
- Web APIs
- Hadoop & Distributed systems

Why should I care?

I only work with data scientists

- Product / Project managers
- Analysis / Research
- Marketing
- Management
- etc

Why should I care?

I only work with data scientists

- Answer your own questions: hypothesis validation
- Independence
- Opportunity sizing
- Understanding / appreciation of techniques: ideas for new data products
- Faster iteration!

Practical Data Science: Implementation

- Unix command line: operating system scripts and utilities
- Git / Github: version control, hosting of source code, collaborative editing
- Python: a general programming language
- Understanding data: objects, relationships & information
- Using regular expressions to for information extraction, data cleaning
- SQL / Relational databases
- Data visualization / plotting
- Tools for predictive modeling
- Distributed computing in hadoop

Practical Data Science: Implementation

- Unix command line: operating system scripts and utilities
- Git / Github: version control, hosting of source code, collaborative editing
- Python: a general programming language
- Understanding data: objects, relationships & information
- Using regular expressions to for information extraction, data cleaning
- SQL / Relational databases
- Data visualization / plotting
- Tools for predictive modeling
- Distributed computing in hadoop



A sampler

Principles for Success

- Try to learn about your as much as you can about your problem at each step
- Make hypotheses. Devise experiments to test these hypotheses
- Machine learning is hard!
- Always try the stupidest (simplest) thing first

Principles for Success

- Log everything. You might need it later.
- Use data to make decisions; don't let the numbers lie.
Stats are your friend.
- Be skeptical. Dig deep for the truth.
- Always over-communicate.

Principles for Success*

*for this course

- Google your error messages or problems. Then try stackoverflow.
- When in doubt, read the documentation.
- Don't be afraid to read the source code, if it's available.
- The only way to learn to code is to write it.

Principles for Success*

*for this course

- Fledgeling programming skills will go away unless exercised.
- Think of a project inspired by this course, work through it on your own.
- Help your classmates when you know an answer or can figure it out. The best way to learn is to teach!