Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Discuss homework
- What is data science?

HW1

- How many were submitted?
    - 79 people of 89 registered.

HW1

- Specific assignment in following email, will include:
    - Install Jupyter and Python
    - Demo that it is working by designing a tutorial that covers the topics of strings, lists, sorting, and dicts in a Jupyter notebook.
        - Your goals are: to cover core competencies in creating and manipulating these variables
        - You may not: copy tutorials.
        - You may: Reference existing tutorials / talk to friends, but you should produce interestingly novel examples.
        - You must: Cite any sources that you reference. Hand in your own work.
        - You will: be graded based on the rubric provided. Note that this rubric requires interpretation in order to be applied to specific assignments. We will talk more about this on Monday.

HW1 - Rubric

- Excellent
- Good
- Fair
- Poor

Interpretation *Ability to explain information presented in mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Provides accurate explanations of information presented in mathematical forms. Makes appropriate inferences based on that information. *For example, accurately explains the trend data shown in a graph and makes reasonable predictions regarding what the data suggest about future events.*
Representation *Ability to convert relevant information into various mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Skillfully converts relevant information into an insightful mathematical portrayal in a way that contributes to a further or deeper understanding.
Calculation	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem. Calculations are also presented elegantly (clearly, concisely, etc.)

Application / Analysis *Ability to make judgments and draw appropriate conclusions based on the quantitative analysis of data, while recognizing the limits of this analysis*	Uses the quantitative analysis of data as the basis for deep and thoughtful judgments, drawing insightful, carefully qualified conclusions from this work.
Assumptions *Ability to make and evaluate important assumptions in estimation, modeling, and data analysis*	Explicitly describes assumptions and provides compelling rationale for why each assumption is appropriate. Shows awareness that confidence in final conclusions is limited by the accuracy of the assumptions.
Communication *Expressing quantitative evidence in support of the argument or purpose of the work (in terms of what evidence is used and how it is formatted, presented, and contextualized)*	Uses quantitative information in connection with the argument or purpose of the work, presents it in an effective format, and explicates it with consistently high quality.

HW1 - Rubric

- Excellent
- Good
- Fair
- Poor

What is data science?

Data science

From Wikipedia, the free encyclopedia

Not to be confused with information science.

Data science, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to data mining.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

When Harvard Business Review called it "The Sexiest Job of the 21st Century" [6] the term became a buzzword, and is now often applied to business analytics,[7] or even arbitrary use of data, or used as a sexed-up term for statistics.[8] While many university programs now offer a data science degree, there exists no consensus on a definition or curriculum contents.[7] Because of the current popularity of this term, there are many "advocacy efforts" surrounding it.[9]

What is data science?

- Method: Check out universities
  - Berkeley

What is Data Science?

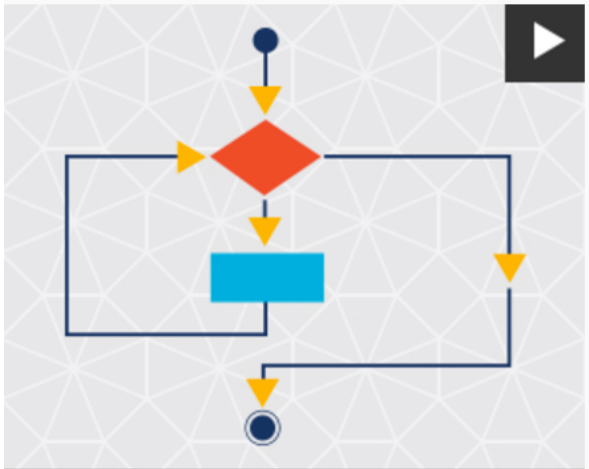
A New Field Emerges

There is significant and growing demand for data-savvy professionals in businesses, public agencies, and nonprofits. The supply of professionals who can work effectively with data at scale is limited, and is reflected by rapidly rising salaries for data engineers, data scientists, statisticians, and data analysts.

A recent study by the McKinsey Global Institute concludes, "a shortage of the analytical and managerial talent necessary to make the most of Big Data is a significant and pressing challenge (for the U.S.)." The report estimates that there will be four to five million jobs in the U.S. requiring data analysis skills by 2018, and that large numbers of positions will only be filled through training or retraining. The authors also project a need for 1.5 million more managers and analysts with deep analytical and technical skills "who can ask the right questions and consume the results of analysis of big data effectively."

The statistics listed below represent this significant and growing demand for data scientists.

#16	3,433	$105,395	#1

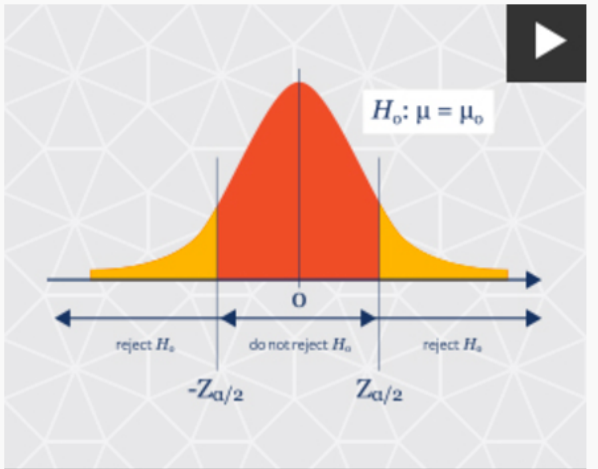


Python for Data Science

3 UNITS

Research Design and Application for Data and Analysis

3 UNITS



Statistics for Data Science

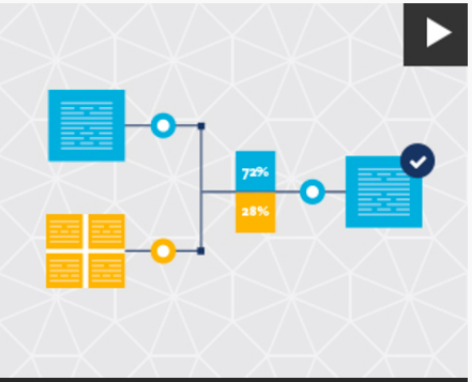
3 UNITS

Storing and Retrieving Data

3 UNITS

Applied Machine Learning

3 UNITS

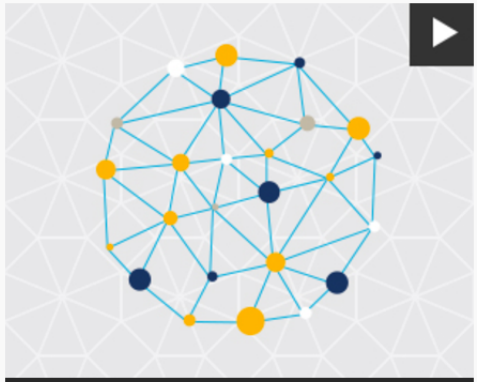


Experiments and Causality

3 UNITS

Behind the Data: Humans and Values

3 UNITS

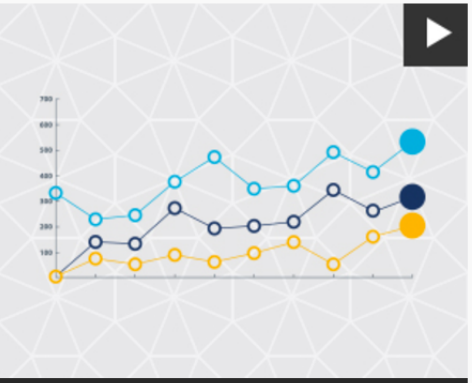


Scaling Up! Really Big Data

3 UNITS

Data Visualization

3 UNITS



Statistical Methods for Discrete Response, Time Series, and Panel Data

3 UNITS

Machine Learning at Scale

3 UNITS

Natural Language Processing with Deep Learning

3 UNITS

What is data science?

- Method: Check out universities
    - NYU
- http://datascience.nyu.edu/what-is-data-science/
- http://datascience.nyu.edu/academics/programs/

What is Data Science?

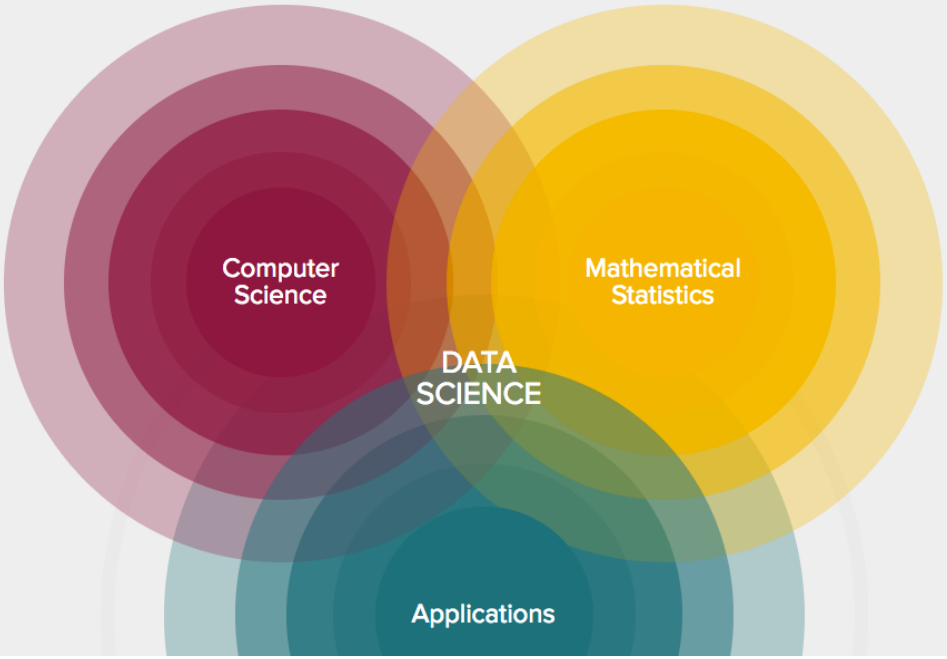
There is much debate among scholars and practitioners about what data science is, and what it isn't. Does it deal only with big data? What constitutes big data? Is data science really that new? How is it different from statistics and analytics?

One way to consider data science is as an evolutionary step in interdisciplinary fields like business analysis that incorporate computer science, modeling, statistics, analytics, and mathematics.

At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them. With such automated methods turning up everywhere from genomics to high-energy physics, data science is helping to create new branches of science, and influencing areas of social science and the humanities. The trend is expected to accelerate in the coming years as data from mobile sensors, sophisticated instruments, the web, and more, grows. In academic research, we will see an increasingly large number of traditional disciplines spawning new sub-disciplines with the adjective "computational" or "quantitative" in front of them. In industry, we will see data science transforming everything from healthcare to media.

50x

in 2020 the world will generate 50 times the amount of data than in 2011 Source: emc.com



Data-Driven Discovery

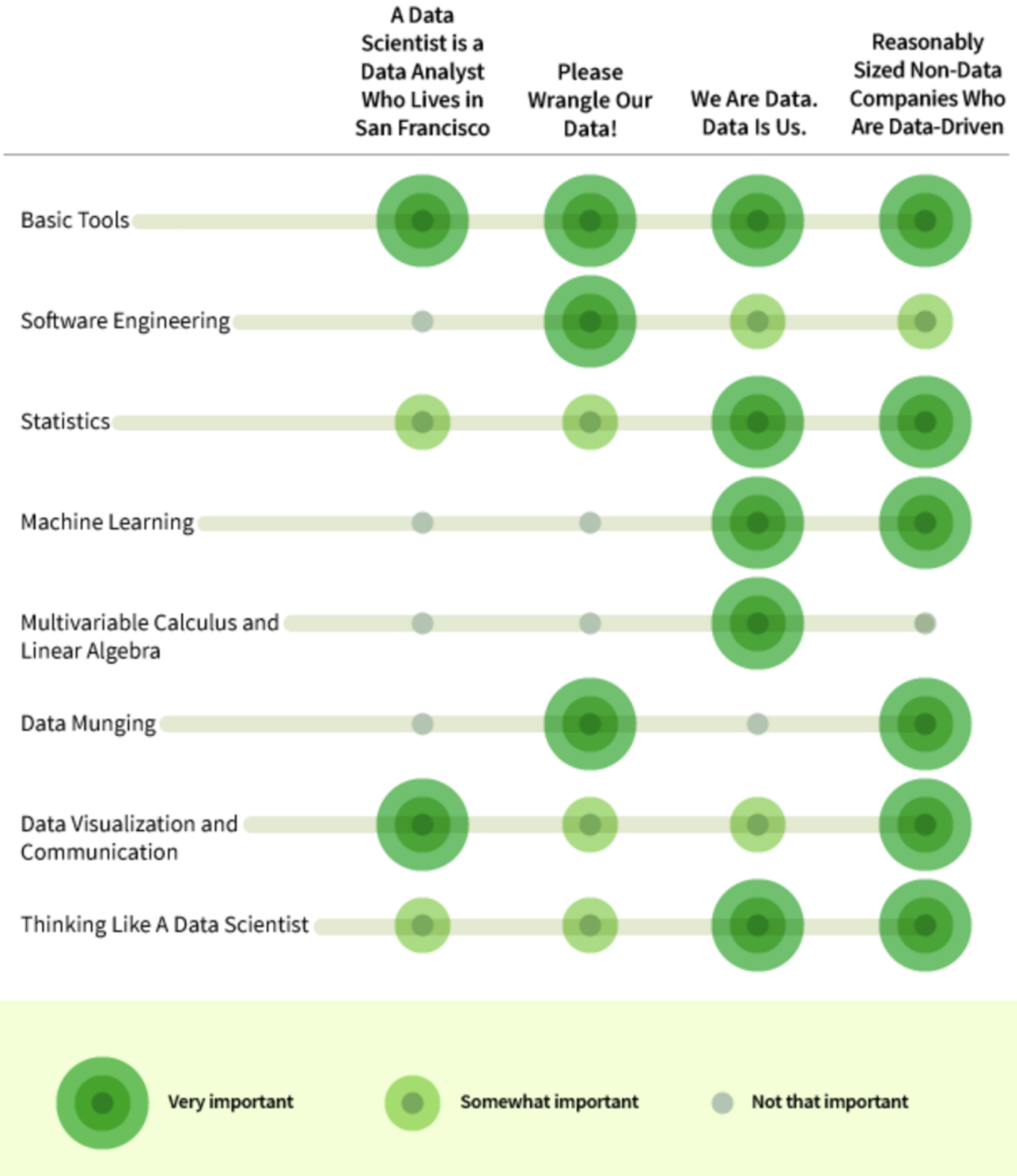
WHAT DATA SCIENCE MEANS FOR RESEARCH

In virtually all areas of intellectual inquiry, data science offers a powerful new approach to making discoveries. By combining aspects of statistics, computer science, applied mathematics, and visualization, data science can turn the vast amounts of data the digital age generates into new insights and new knowledge.

Click on the icons to the left to see how social scientists, medical researchers, and many others are using data science to advance their fields.

What is data science?

- Method: Check online courses
    - Udacity


A Data Scientist is a Data Analyst Who Lives in San Francisco
Please Wrangle Our Data!
We Are Data. Data Is Us.
Reasonably Sized Non-Data Companies Who Are Data-Driven
Basic Tools
Software Engineering
Statistics
Machine Learning
Multivariable Calculus and Linear Algebra
Data Munging
Data Visualization and Communication
Thinking Like A Data Scientist
Very important
Somewhat important
Not that important

What is data science?

- Method: Check online courses
    - Coursera (JHU)

- The Data Scientist's Toolbox
- R Programming
- Getting and Cleaning Data
- Exploratory Data Analysis
- Reproducible Research
- Statistical Inference
- Regression Models
- Practical Machine Learning
- Developing Data Products
- Data Science Capstone

What is data science?

- Method: Ask a statistician
    - Larry Wasserman

Data Science: The End of Statistics?

As I see newspapers and blogs filled with talk of “Data Science” and “Big Data” I find myself filled with a mixture of optimism and dread. Optimism, because it means statistics is finally a sexy field. Dread, because statistics is being left on the sidelines.

The very fact that people can talk about data science without even realizing there is a field already devoted to the analysis of data — a field called statistics — is alarming. I like what Karl Broman says:

When physicists do mathematics, they don’t say they’re doing “number science”. They’re doing math.

If you’re analyzing data, you’re doing statistics. You can call it data science or informatics or analytics or whatever, but it’s still statistics.

Well put.

Maybe I am just pessimistic and am just imagining that statistics is getting left out. Perhaps, but I don’t think so. It’s my impression that the attention and resources are going mainly to Computer Science. Not that I have anything against CS of course, but it is a tragedy if Statistics gets left out of this data revolution.

Two questions come to mind:

1. Why do statisticians find themselves left out?

2. What can we do about it?

I'd like to hear your ideas. Here are some random thoughts on these questions. First, regarding question 1.

1. Here is a short parable: A scientist comes to a statistician with a question. The statistician responds by learning the scientific background behind the question. Eventually, after much thinking and investigation, the statistician produces a thoughtful answer. The answer is not just an answer but an answer with a standard error. And the standard error is often much larger than the scientist would like.

   The scientist goes to a computer scientist. A few days later the computer scientist comes back with spectacular graphs and fast software.

   Who would you go to?

   I am exaggerating of course. But there is some truth to this. We statisticians train our students to be slow and methodical and to question every assumption. These are good things but there is something to be said for speed and flashiness.

2. Generally, speaking, statisticians have limited computational skills. I saw a talk a few weeks ago in the machine learning department where the speaker dealt with a dataset of size 10 billion. And each data point had dimension 10,000. It was very impressive. Few statisticians have the skills to do calculations like this.

What is data science?

- Method: Check industry sources
- KDNuggets

Technical Skills: Analytics

1. **Education** – Data scientists are highly educated – 88% have at least a Master's degree and 46% have PhDs – and while there are notable exceptions, a very strong educational background is usually required to develop the depth of knowledge necessary to be a data scientist. Their most common fields of study are Mathematics and Statistics (32%), followed by Computer Science (19%) and Engineering (16%).
2. **SAS and/or R** – In-depth knowledge of at least one of these analytical tools, for data science R is generally preferred.

Technical Skills: Computer Science

3. **Python Coding** – Python is the most common coding language I typically see required in data science roles, along with Java, Perl, or C/C++.
4. **Hadoop Platform** – Although this isn't always a requirement, it is heavily preferred in many cases. Having experience with Hive or Pig is also a strong selling point. Familiarity with cloud tools such as Amazon S3 can also be beneficial.
5. **SQL Database/Coding** – Even though NoSQL and Hadoop have become a large component of data science, it is still expected that a candidate will be able to write and execute complex queries in SQL.
6. **Unstructured data** – It is critical that a data scientist be able to work with unstructured data, whether it is from social media, video feeds or audio.

Non-Technical Skills

7. **Intellectual curiosity** – No doubt you've seen this phrase everywhere lately, especially as it relates to data scientists. Frank Lo describes what it means, and talks about other necessary "soft skills" in his guest blog posted a few months ago.
8. **Business acumen** – To be a data scientist you'll need a solid understanding of the industry you're working in, and know what business problems your company is trying to solve. In terms of data science, being able to discern which problems are important to solve for the business is critical, in addition to identifying new ways the business should be leveraging its data.
9. **Communication skills** – Companies searching for a strong data scientist are looking for someone who can clearly and fluently translate their technical findings to a non-technical team, such as the Marketing or Sales departments. A data scientist must enable the business to make decisions by arming them with quantified insights, in addition to understanding the needs of their non-technical colleagues in order to wrangle the data appropriately. Check out our recent flash survey for more information on communication skills for quantitative professionals.

What is data science?

- Method: Google images!

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

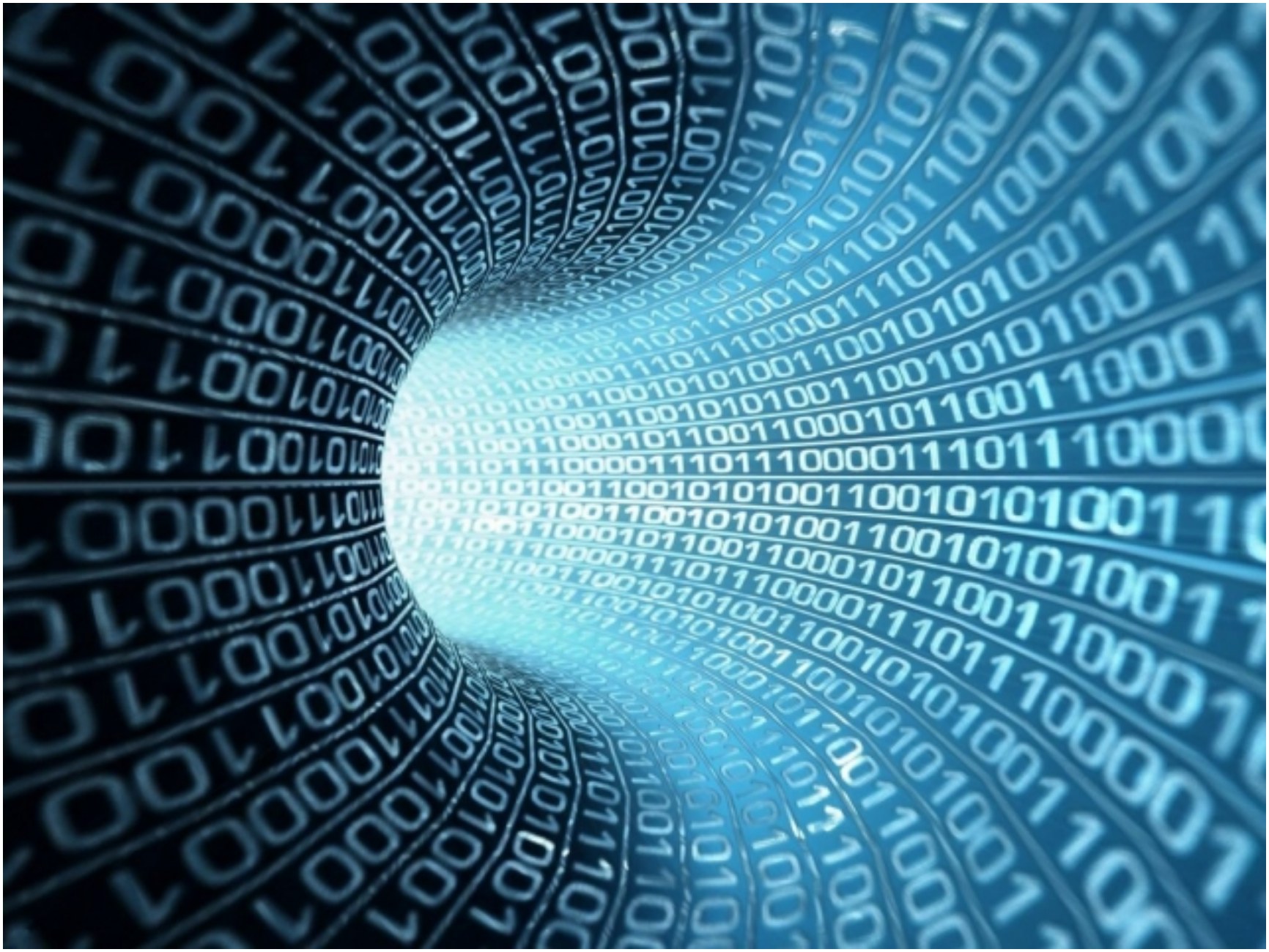
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

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.

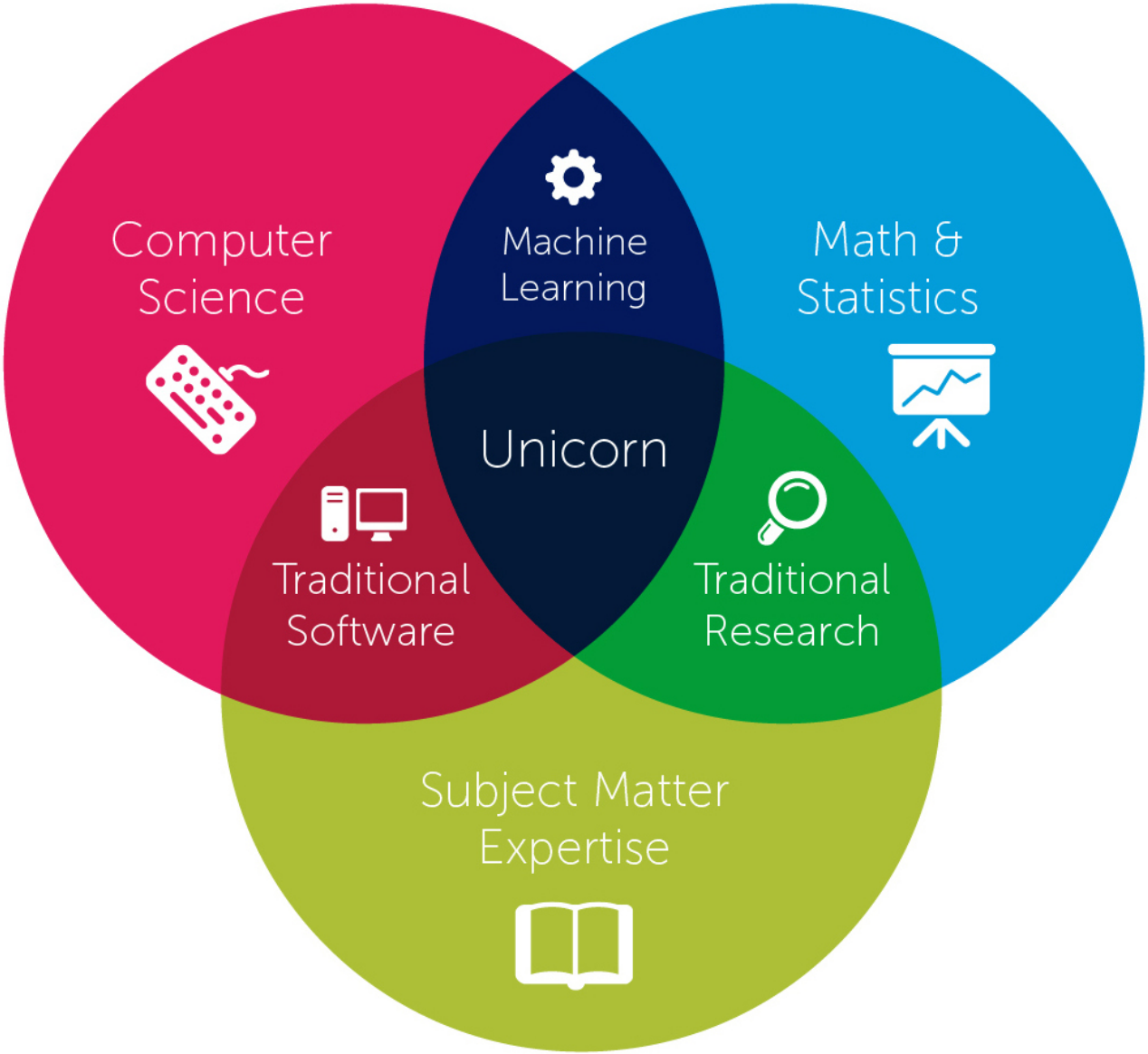

Product
IDEA


BIG
DATA
STORAGE
VOLUME
PROCESSING
INFORMATION
SHARING
COLLECTION
ANALYSIS
TOOLS
DECISION
SPEED
OPTIMIZATION
SOFTWARE
BUSINESS
VARIETY
VELOCITY
STATISTICS
DIFFICULT
DATABASE
PROCESS
SEARCH
SOURCES
SYSTEMS
MANAGE
CAPTURE
SHARING
ENHANCED
TRIGGER
LOGS
TYPES
AREAS
DENSITY
MAKING
PETABYTES
AMOUNT
RELATIONAL
VISUALIZATION

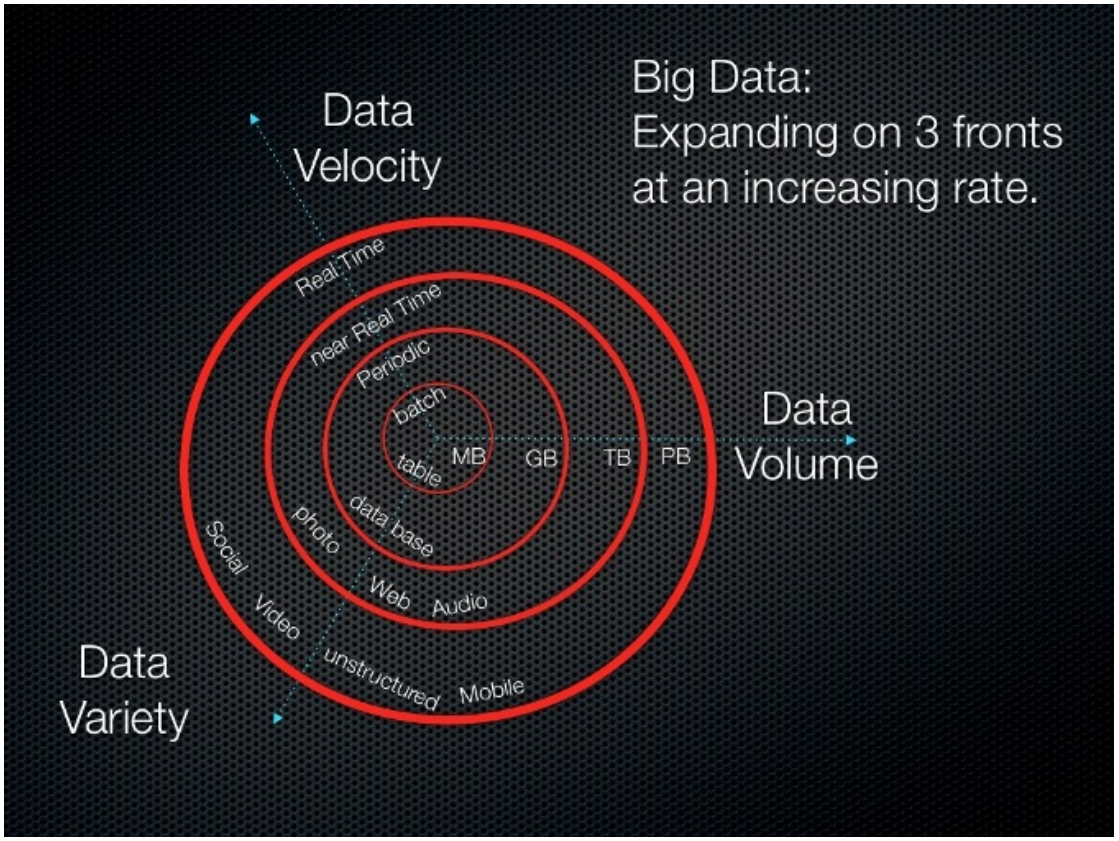

0110
0100
1001
0110
0100
1001
0110
0100
1001
1001



Data Science



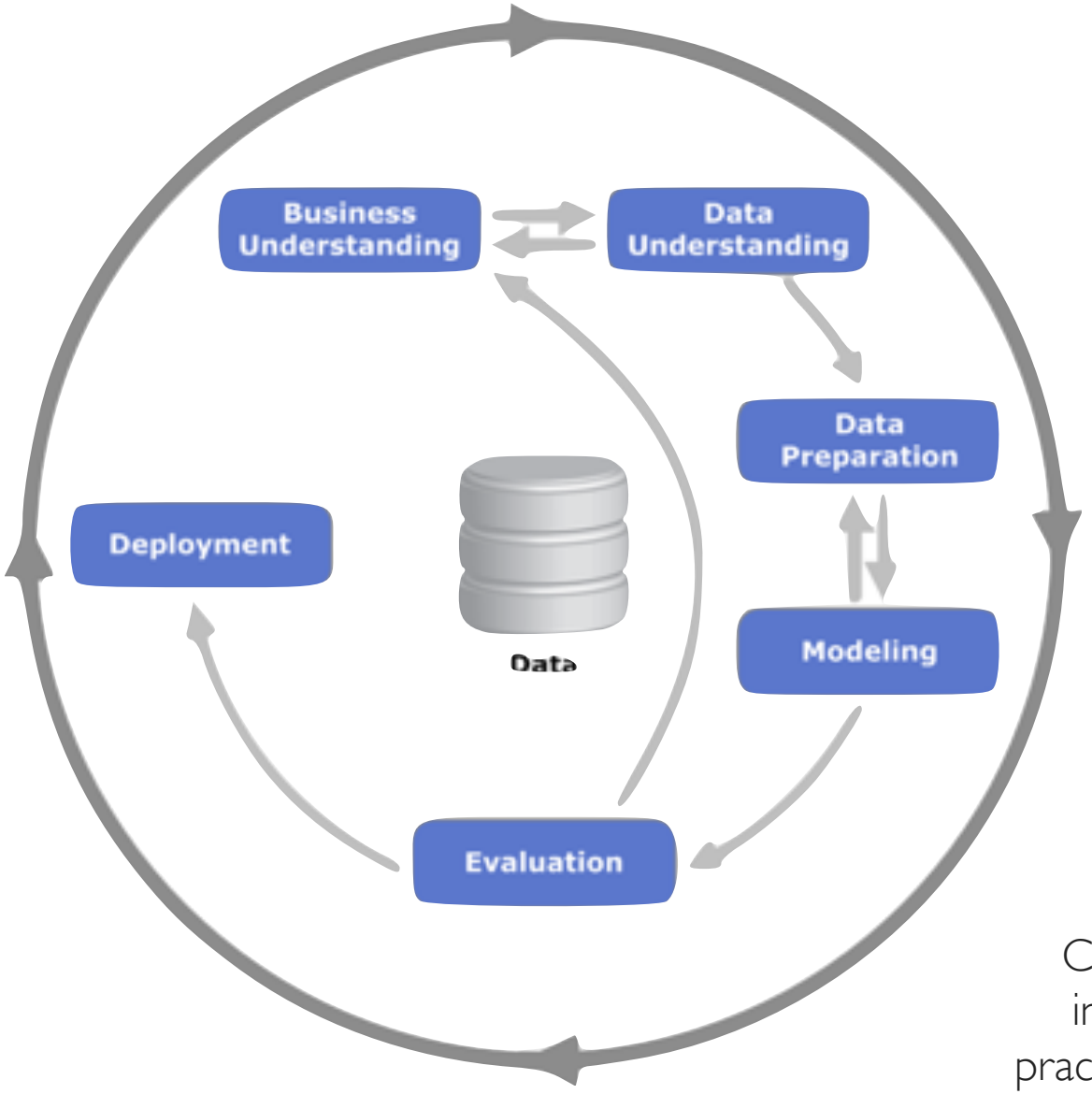
"Big data": Promise and problems



Promise: Personalization of prediction, service, education, healthcare, …

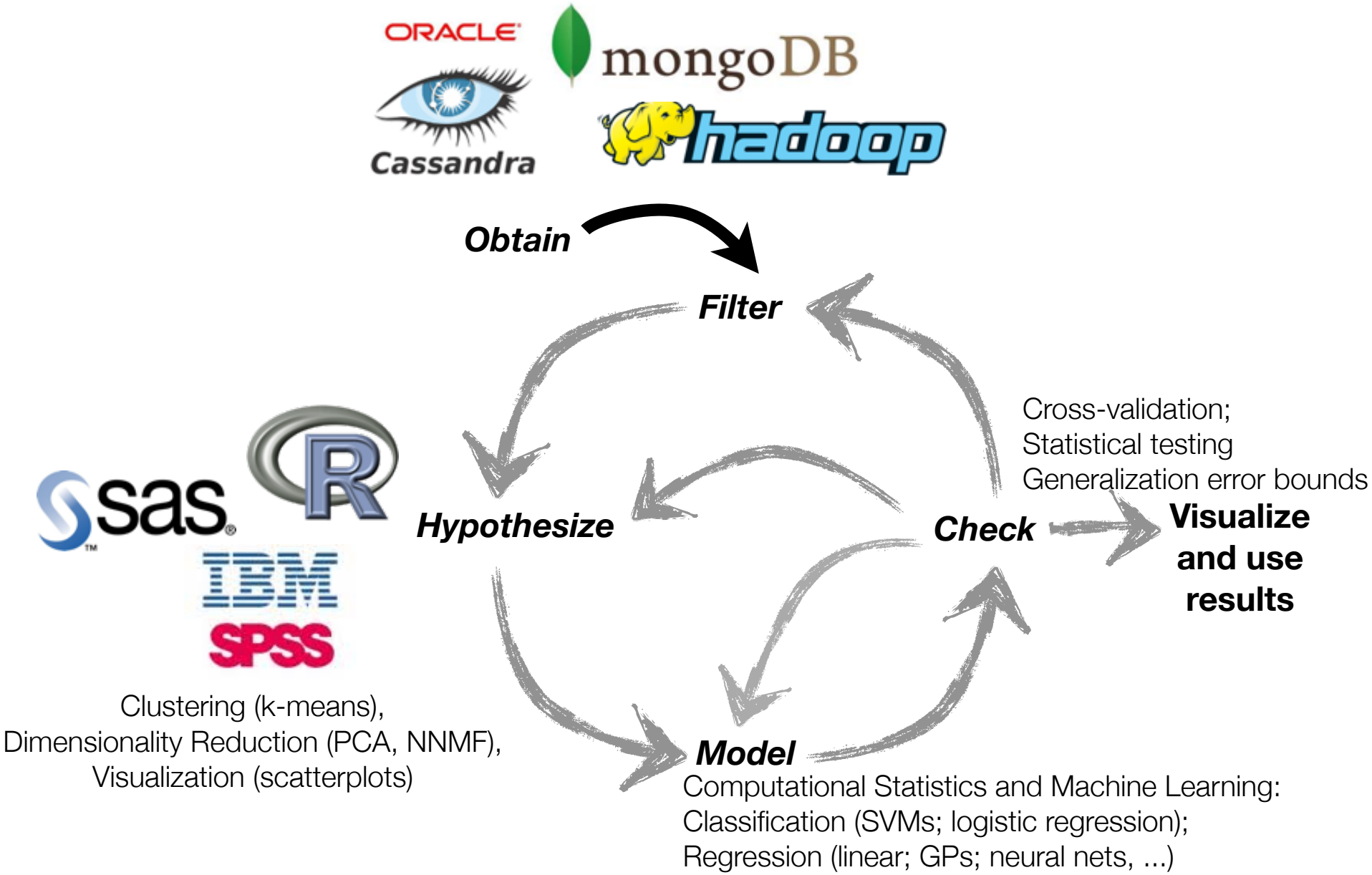
Problems: High dimensional, heterogeneously typed data; missing values; noisy variables

Big Data: Acquisition, processing, and analysis cycle

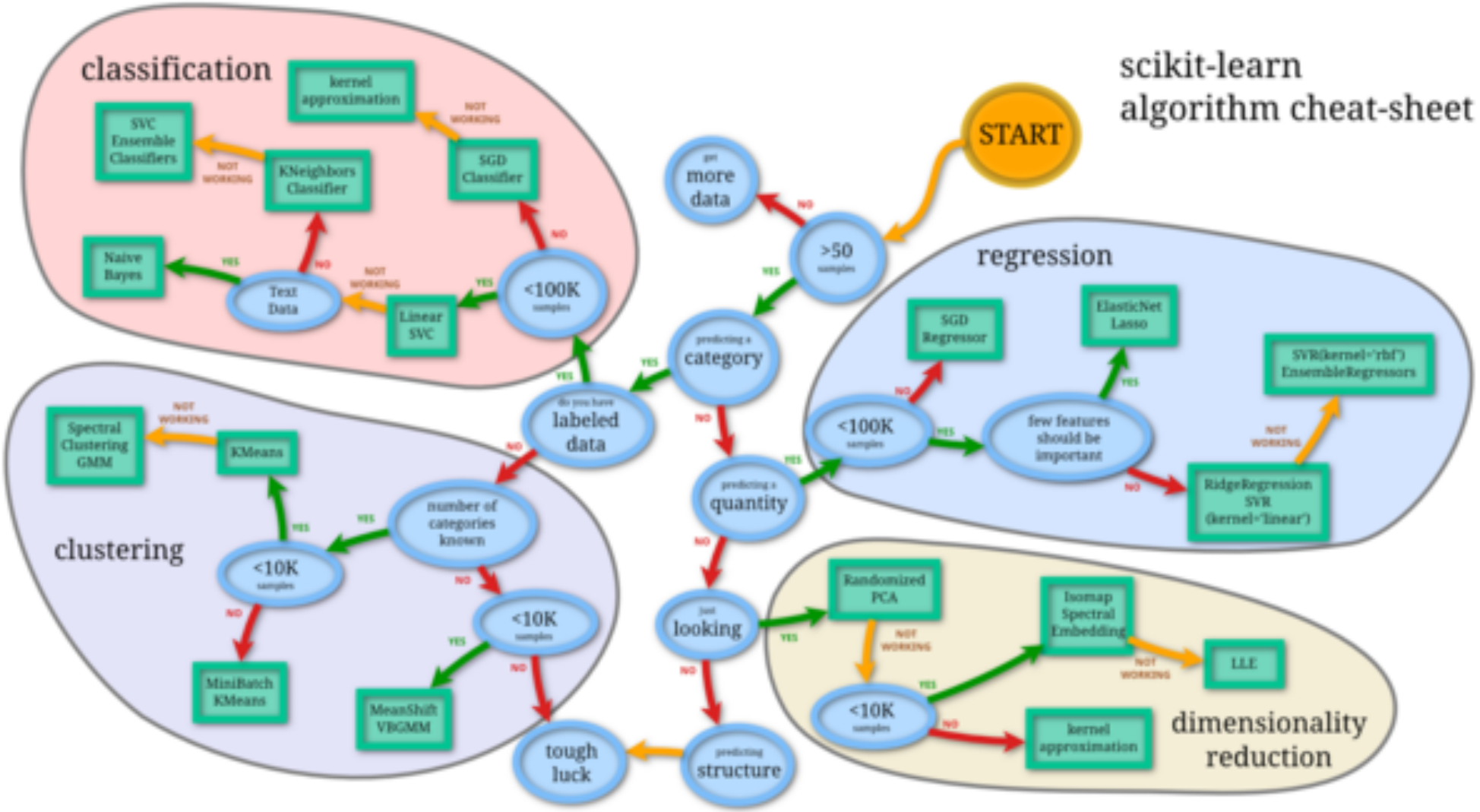


CRISP-DM: Cross-industry standard practice for data mining

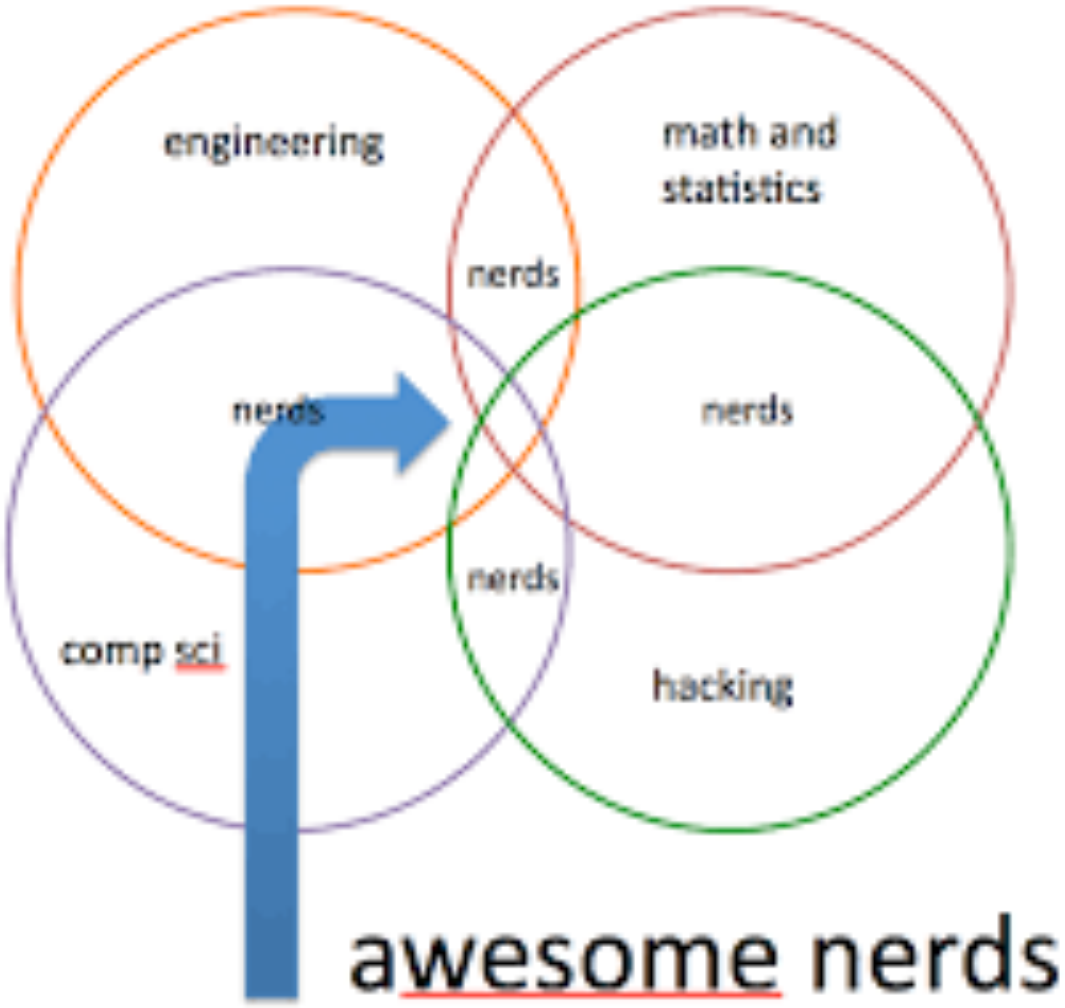
Big Data: Acquisition, processing, and analysis cycle



A "simple" graphic to help with modeling…



Data scientists?



Hilary Mason, of Fast forward labs
(formerly of bitly)

The De Finetti Theorem

- An infinite sequence of random variables $(\theta_1, \theta_2, \ldots)$ is called infinitely exchangeable if the distribution of any finite subsequence is invariant to permutation
- *Theorem*: infinite exchangeability if and only if

$$p(\theta_1, \theta_2, \ldots, \theta_n) = \int \prod_i p(\theta_i|G) P(dG)$$

for some random measure G

"The Michael Jordan of Machine Learning"

HW

- Do your peer grading!
    - Due Tuesday by Midnight