

# Machine Learning Techniques

## Lesson 1: Data Science Process

Hang Zhang, PhD  
October 1<sup>st</sup>, 2018


# Big Data and Data Science the case for Advanced Analytics in DW

“Big Data: The Next Frontier for Innovation, Competition and Productivity.” – McKinsey Report

---





- Need 140,000 to 190,000 more people with “deep analytical” skills, typically experts in statistical methods and data-analysis technologies
- Need 1.5 million more data-literate managers, whether retrained or hired
- \$300 billion – potential annual value to US health care – more than double the total annual health care spending in Spain
- €250 billion potential annual value to Europe's public sector administration – more than the GDP of Greece
- \$600 billion potential annual consumer surplus from using personal location data globally
- 60% potential increase in retailers’ operating margins possible with big data


# Instructors



**Hang Zhang**  
Principal Data & Applied Scientist at Microsoft  
Redmond, Washington





[Add profile section](#) [More...](#)

-  Microsoft
-  Rutgers University
-  See contact info
-  See connections (500+)



**Wee Hyong Tok** • 1st  
Principal Data Science Manager, Office of CTO AI | Microsoft  
at Microsoft  
Greater Seattle Area

[Message](#) [More...](#)

-  Microsoft
-  National University of Singapore
-  See contact info
-  See connections (500+)

# Set Your Expectations

- Be able to adopt a standard data science process to tackle data science / machine learning challenges
- Get you familiar with and be able to use some of the most popular data science / machine learning concepts and methodologies/algorithms
- Set a good starting point for your data science / machine learning professions

# Course Content

## What We **Will** Cover

- How to do data science project: data science process
- Data acquisition from different data sources.
- Data Preparation
- Practical Techniques of Data Science
- Machine Learning over Data
- Experimentation
- Brief Introduction of Deep Learning
- Literature, what are people writing and saying about data science

# Course Content

## What We **Will Not** Cover

- Probability and Statistics, *covered in second course of the program...*
- NoSQL, *covered in the first course of the program...*
- Machine Learning Theory, *though we will discuss ML techniques...*
- Data Mining, *again we will discuss techniques from data mining...*

# Course Content

Data Science Vocabulary List, *the language of data science (quick poll)...*

- Machine learning
- Supervised learning
- Unsupervised learning
- Training set or training sample
- Test set or test sample
- K-nearest neighbors
- Confusion matrix
- Classification
- Prediction
- Overfitting
- K Fold Cross-validation, why use it?
- Loss functions
- Labels
- Bias, variance, bias variance trade-off
- ROC Curve

# The Way to Learn in this Class

- In class:
  - Be present in classroom.
  - Be active in in-classroom discussion.
  - Bring your laptops with Python (Anaconda) 3 installed.
  - Practice the labs in Jupyter Notebooks
- After class:
  - Do your assignments and capstone project, and turn in on time.
  - Read references to have deeper understanding of the techniques
  - Accumulate your experience in data science / machine learning by applying what learned in class on real projects



# Your Scores

- Attendances (10%)
- Assignments (40%)
  - Totally 5 assignments
  - Due at midnight (11:59pm) the day before the next class after assignment
  - Submit a Jupyter Notebook (.ipynb file) with results
  - Everyone gets 1 late exception (1 week) to accommodate your urgency or travel plan.
  - You should complete the assignments independently, no collaboration allowed.
- Capstone Project (50%)
  - Topic to be determined. Will announce it around the 4<sup>th</sup> class.
  - Mid-term report at 7<sup>th</sup> class.
  - Project report due midnight of one day before the last class
  - Capstone project in teams.
- Final score < 60 will fail the class

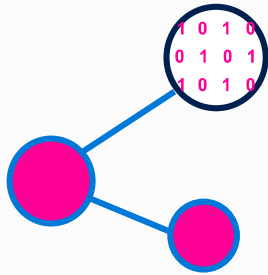
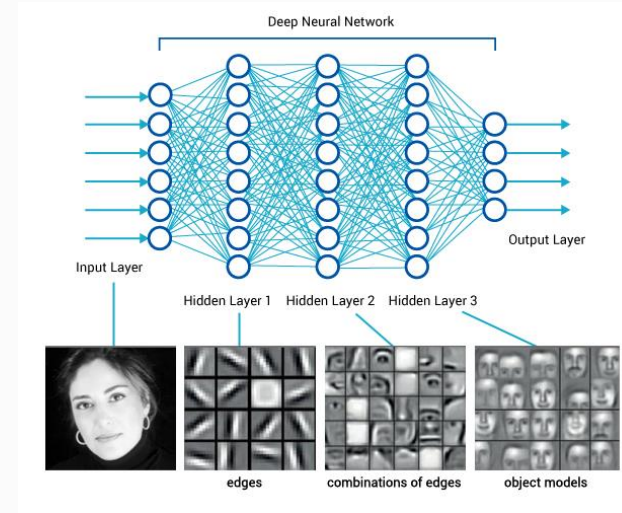
# Resources to Use for Jupyter Notebook

- Install Anaconda with Python 3 on your laptop
  - <https://www.anaconda.com/download/>
  - Run *jupyter notebook* in shell command to start Jupyter Notebook
- Online Jupyter Notebook service (free):
  - <https://notebooks.azure.com/>
  - Anaconda runs in the backend

# The 4<sup>th</sup> industrial revolution



AI



Big Data and IoT



Cloud Computing

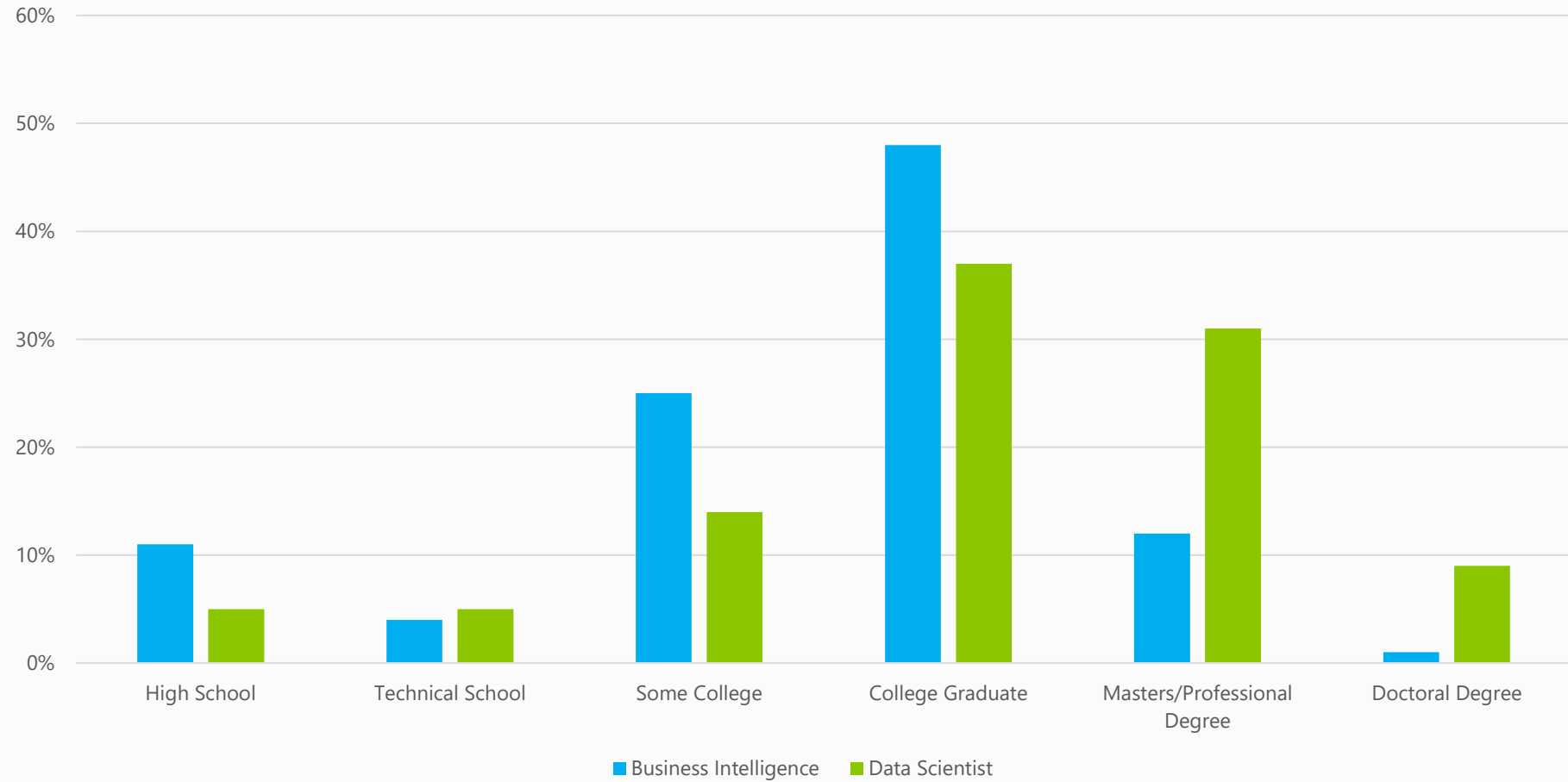
# Data Scientist: The Sexiest Job of the 21<sup>st</sup> Century

- Based on Harvard Business Review in Oct 2012
- 65% of enterprises feel they have a **strategic shortage of data scientists**, a role many did not even know existed 12 months ago...

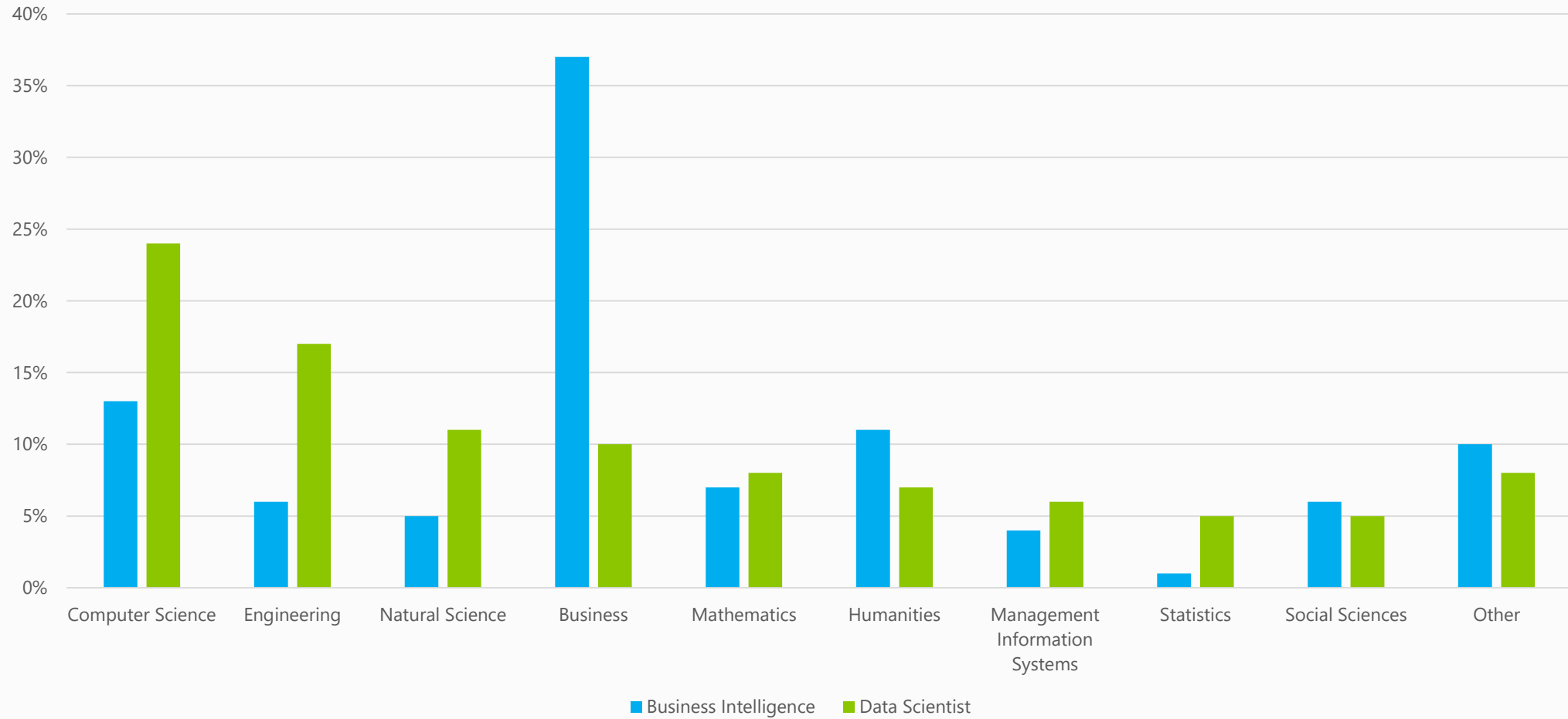
# Where Can We Get New Data Scientists?

Sources of Talents	Percentage
Students Studying Computer Science	34%
Professionals in disciplines other than computer science	27%
Students studying in fields other than computer science	24%
Today's business intelligence professionals	12%
Others	2%

## Data Scientists Are More Likely to Have Higher Degrees than Business Intelligence



Distribution of Educational Fields of Business Intelligence and Data Scientists



# Will Deep Learning/AI Make Data Scientists Lose Their Jobs?

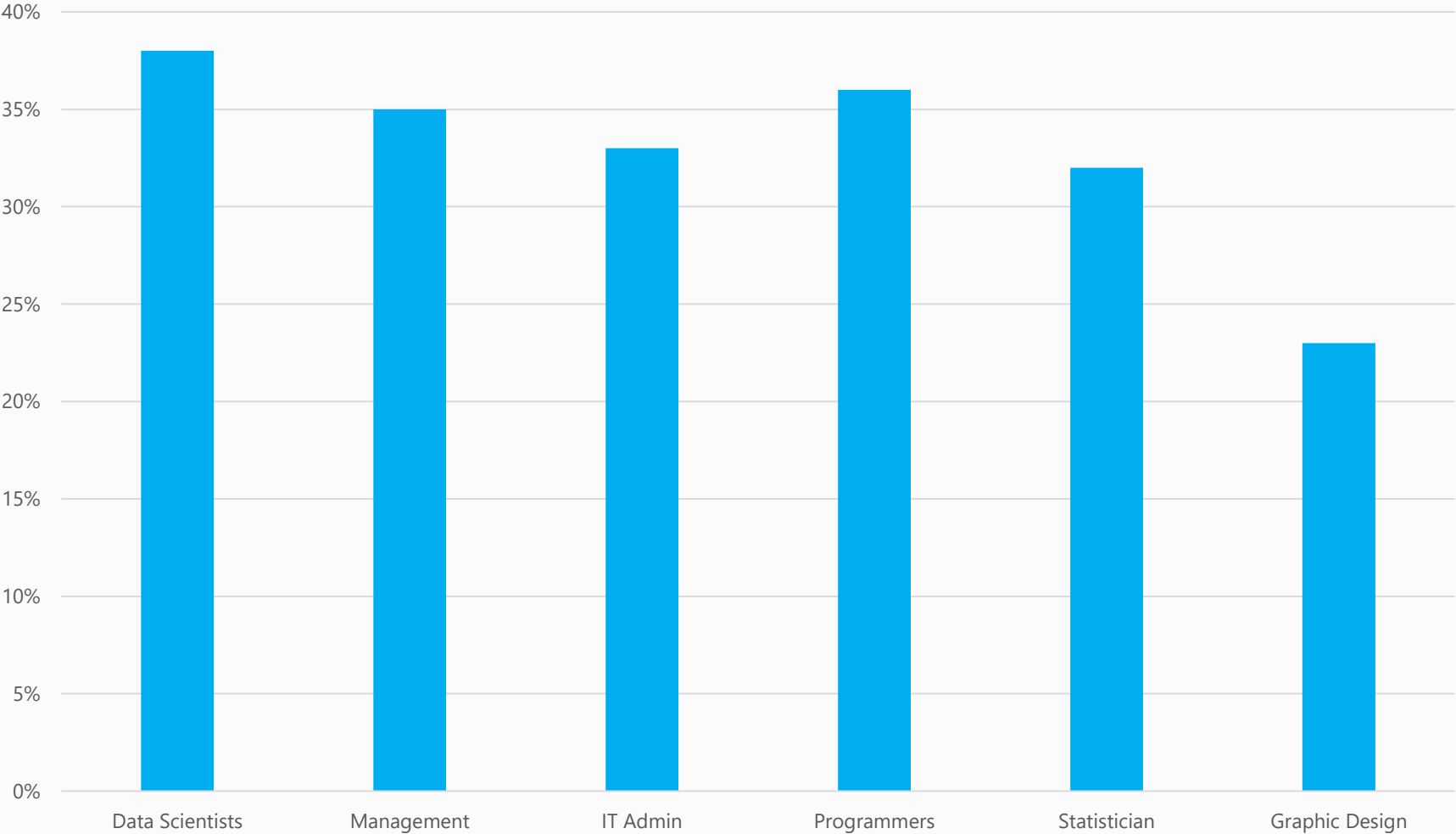
- The most time consuming and most interesting part of data science is feature engineering/selection
- Deep Learning: looks like it does not need feature engineering/selection, just pass the raw data in, and get a model out.



Deep Learning w/o Careful Data Scientist Supervision



Data Scientists Collaborate with A Big Variety of Personnel



# Sobering statistics

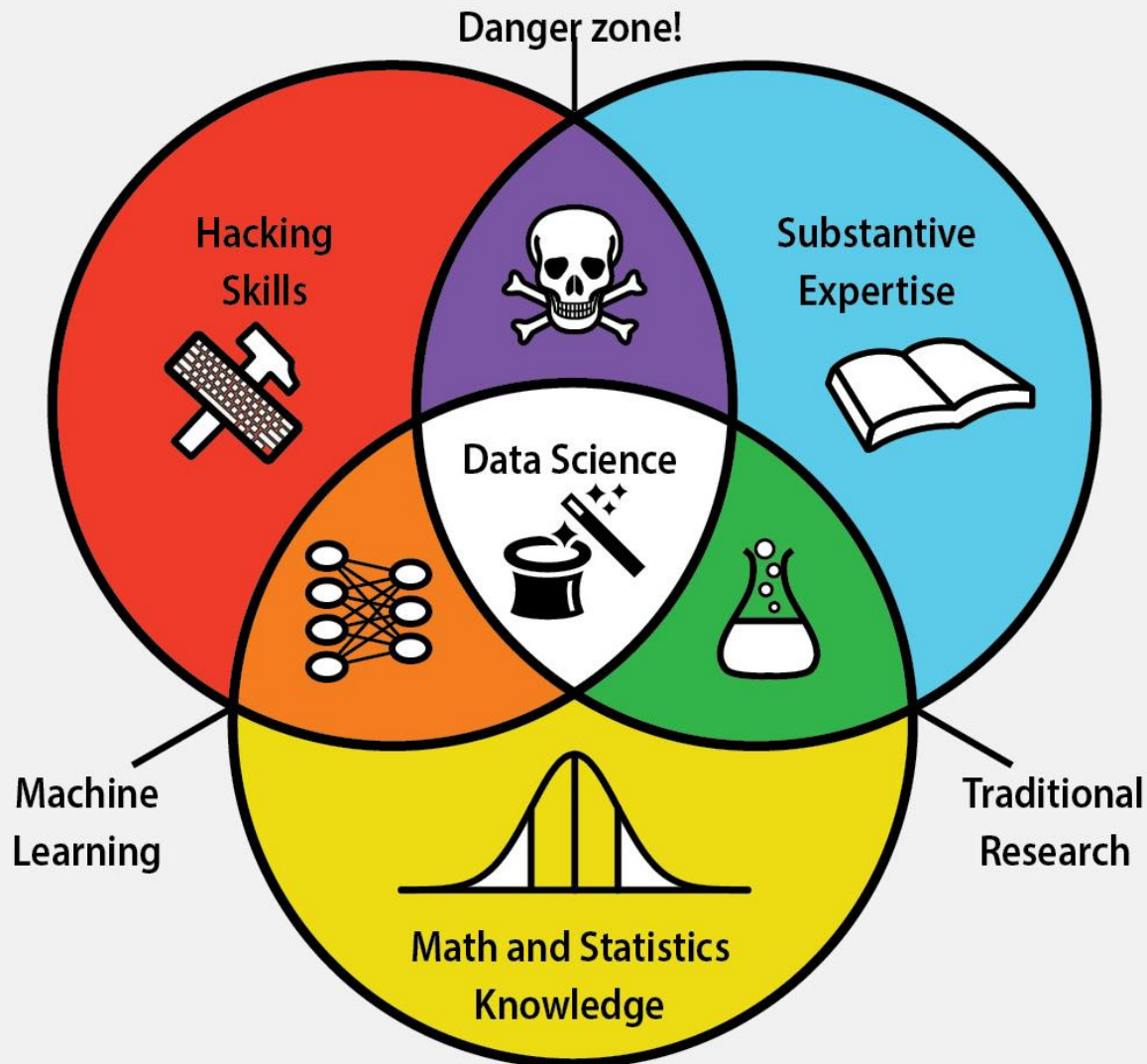
“Only **27%** of the big data projects are regarded as successful”

Only **13%** of organizations have achieved full-scale production for their Big Data implementations

“Only **8%** of the big data projects are regarded as VERY successful”

“Only **17%** of survey respondents said they had a well-developed Predictive/Prescriptive Analytics program in place, while 80% said they planned on implementing such a program within five years” – Dataversity 2015 Survey

# DATA SCIENCE SKILLSET



Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills**, **math and statistics knowledge**, and **substantive expertise** in a field of science.



**Hacking skills** are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.



**Math and statistics knowledge** allows a data scientist to choose appropriate methods and tools in order to extract insight from data.



**Substantive expertise** in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.



**Traditional research** lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.



**Machine learning** stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



**Danger zone!** Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

# How we work: The unspoken challenges of doing data science

*Yael Garten (LinkedIn)*

11:50am-12:30pm Wednesday, March 15, 2017

Data-driven business management, Strata Business Summit

Location: 210 D/H

Level: Non-technical

## Description

Beyond being dubbed "sexiest job in the 21st century," data science is a rewarding career. It's also really hard—not just the technical work itself but also "how to do the work well" in an organization. The term "data scientist" covers a broad range of specific roles ranging from data engineer to data analyst to machine learning expert. Yael Garten explores what data scientists do, how they fit into the broader company organization, and how they can excel at their trade and shares the hard and soft skills required, tips and tricks for success, and challenges to watch out for.

Walking through specific examples, Yael outlines the tips and tactics that she has employed to enable herself and her data science team to thrive, feel empowered, and have impact. She'll leave time for some role playing of challenging scenarios to provide guidance on how to effectively approach and solve some of the difficult issues you might encounter.

## Yael Garten

LinkedIn

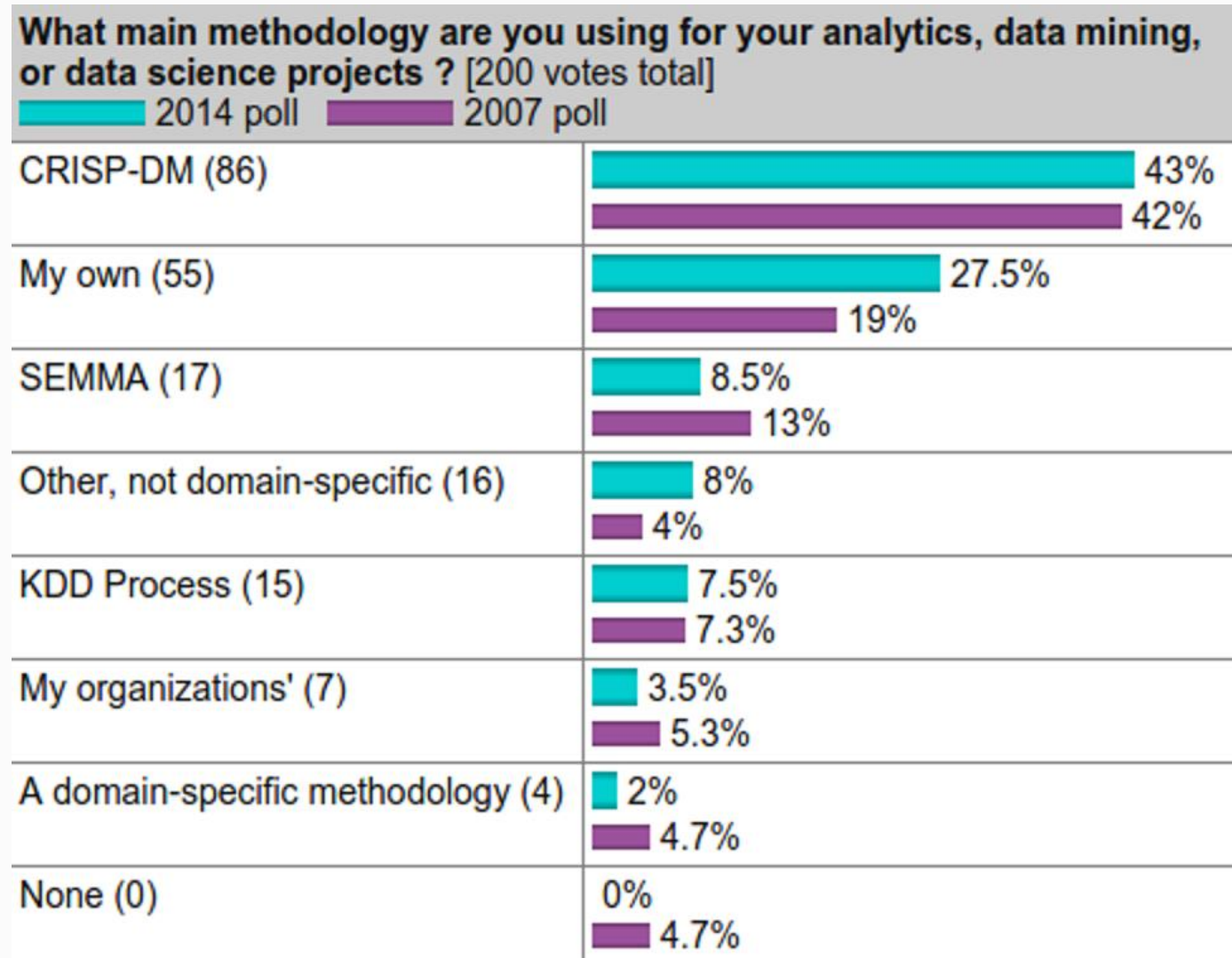
Yael Garten leads a team of data scientists at LinkedIn that focuses on understanding and increasing growth and engagement of LinkedIn's 400 million members across mobile and desktop consumer products. Yael is an expert at converting data into actionable product and business insights that impact strategy. Her team partners with



# How to Do the Work Well in an Organization?

- We need A Standardized data science process
  - For data scientists, a standardized data science process can:
    - Support collaboration
    - Support quality assurance in the entire lifecycle of a project
    - Support security control of project assets
    - Support better planning and tracking
  - For organization, a standardized data science process can:
    - Provide convenience for management to plan and track
    - Accumulate knowledge and expertise over time to continuously improve productivity

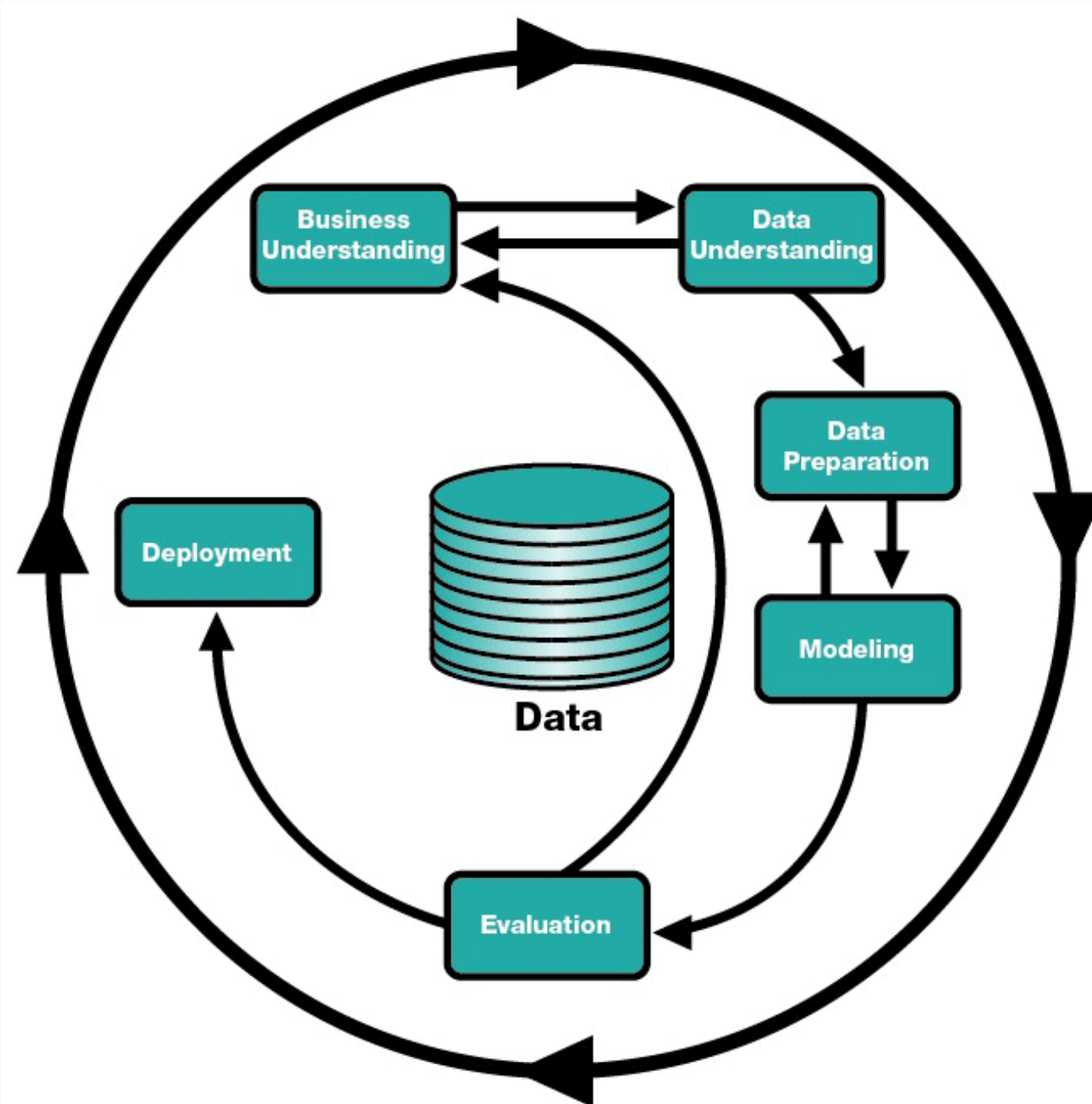
# Who uses a Process for Data Science?



Source: [KDD Nuggets](#), October 2014



# CRISP-DM



# Business Understanding

## Business background

---

- Who is the client, what business domain the client is in.
- What business problems are we trying to address?

## Scope

---

- What data science solutions are we trying to build?
- What will we do?
- How is it going to be consumed by the customer?



# Examples of Scoping Down A Data Science Project (1)

## Example 1: An F1 Racing Car Team

“I want to optimize my racing strategy” – Customer’s request

Questions to ask:

- What is the current racing strategy?
- How do you determine whether a racing car needs to change tires or refuel?
- What are the decision factors that are qualitative or experience-based?
- Whether there is any decision factors that can be data-driven?

Project #1:

Build a machine learning model to predict the tire surface temp

# Examples of Scoping Down A Data Science Project (2)

## Example 2: An Speaker Manufacturer

“I want to improve the efficiency of my customer service” –  
Customer’s request

Questions to ask:

- What is the major pain point of you customer service department or your customers
- What are you doing now when you are tackling this major pain point

Project #1:

Build a machine learning model to do automatic fault diagnosis and propose solutions to customers

# Success Criteria and Deployment Plan

- What is the performance of the current system?
- What is the expected performance of the data science solution
  - Be optimistic
  - But never over-commit
- If the expected performance is achieved, what is the plan of deployment?
  - It might take 2-3 months to complete a data science project
  - Management priority might have changed
  - But you should still be able to claim your project a successful as long as you reach the expected performance

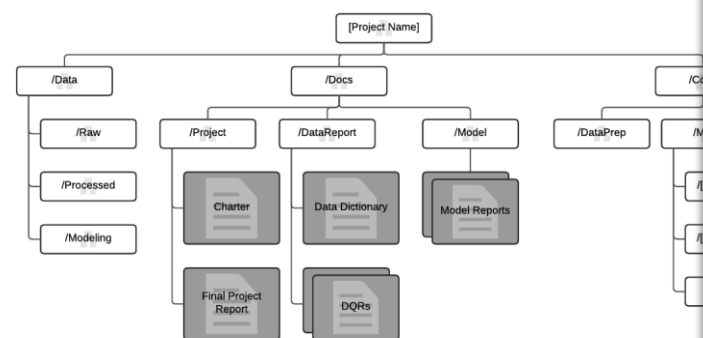
# Data Understanding

- What possibly relevant data are there?
- Which is the most relevant set of data?
- Are they relevant to the business problem we want to address?
- With basic feature engineering on the most relevant set of data, how good the machine learning model can perform? (baseline model)
- If no relevant data available, any other business problem can be addressed by this data? Alternatively, any other data source might be relevant to the selected business problem?

# Standardized Git Repository and Docs, Shared Productivity Utilities

- One git repository per project
- Standardized git repository directory structure
- A set of standardized document templates
- A shared data science utility repository, and a process to enrich it over time

## Templates



### Project Charter

#### Business background

- Who is the client, what business domain the client is in.
- What business problems are we trying to address?

#### Scope

- What data science solutions are we trying to build?
- What will we do?
- How is it going to be consumed by the customer?

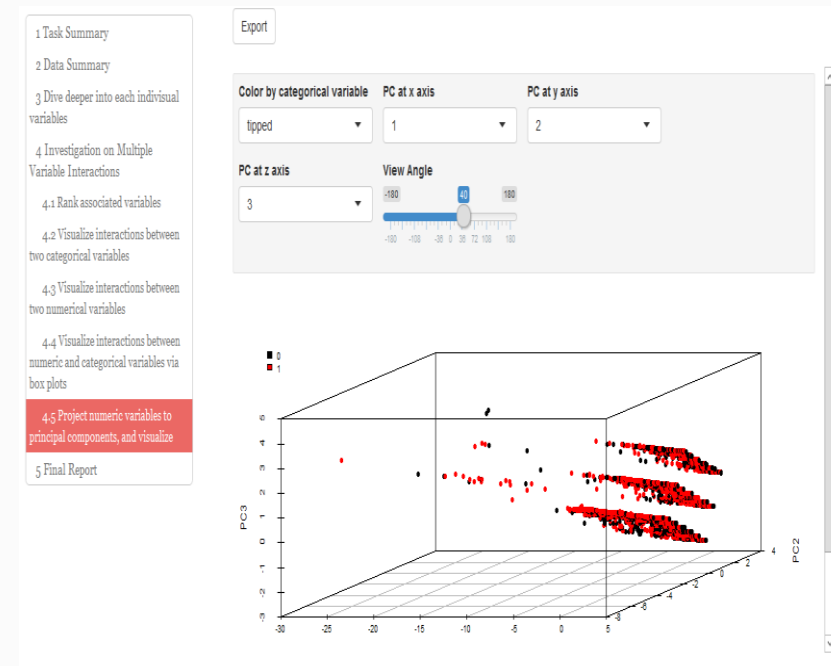
#### Personnel

- Who are on this project:
  - Microsoft:
    - Project lead
    - PM
    - Data scientist(s)
    - Account manager
  - Client:
    - Data administrator
    - Business contact

#### Metrics

- What are the qualitative objectives? (e.g. reduce user churn)
- What is a quantifiable metric (e.g. reduce the fraction of users with 4-week inactivity)
- Quantify what improvement in the values of the metrics are useful for the customer scenario (e.g. reduce the fraction of users with 4-week inactivity by 20%)
- What is the baseline (current) value of the metric? (e.g. current fraction of users with 4-week inactivity = 60%)
- How will we measure the metric? (e.g. A/B test on a specified subset for a specified period; or comparison of performance after implementation to baseline)

## Utilities



# Why we recommend separate git repositories for different data science projects?

- Track all changes in your code
- Rollback bad code
- Git operations are reversible
- Facilitate collaboration
  - Work on the same code simultaneously
  - Code reviews
  - Identify and resolve conflicts
- Recover lost repos
- Operations are local
  - No need to repeatedly download from server for every operation
  - Can work offline

# Three Things to Keep in Mind when Doing Data Science

- Keep the data/model SECURED as required by your customer
- Keep the data/model SECURED as required by your customer
- Keep the data/model SECURED as required by your customer

# AOL search data leak

From Wikipedia, the free encyclopedia

The **AOL search data leak** was the release, in August 2006, of detailed search logs by AOL of a large number of AOL users. The release was intentional and intended for research purposes; however, the public release meant that the entire Internet could see the results rather than a select number of academics. AOL did not redact any information, which caused privacy concerns since users could potentially be identified from their searches.

Contents

[hide]

1

Overview

2

Lawsuits

3

Notable users

3.1

Thelma Arnold

3.2

User 927

4

See also

5

References

6

External links

## Overview

On August 4, 2006, AOL Research, headed by Dr. Abdur Chowdhury, released a compressed text file on one of its websites containing twenty million search keywords for over 650,000 users over a 3-month period intended for research purposes. AOL deleted the search data on their site by August 7th, but not before it had been mirrored and distributed on the Internet.

AOL did not identify users in the report; however, personally identifiable information was present in many of the queries. As the queries were attributed by AOL to particular user numerically identified accounts, an individual could be identified and matched to their account and search history by such information.<sup>[1]</sup> *The New York Times* was able to locate an individual from the released and anonymized search records by cross referencing them with phonebook listings.<sup>[2]</sup> Consequently, the ethical implications of using this data for research are under debate.<sup>[3][4]</sup>

AOL acknowledged it was a mistake and removed the data; however, the removal was too late. The data was redistributed by others and can still be downloaded from mirror sites.<sup>[5][6]</sup>

In January 2007, Business 2.0 Magazine on CNNMoney ranked the release of the search data #57 in a segment called "101 Dumbest Moments in Business."<sup>[7]</sup>