GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 2: MACHINE LEARNING

LAST TIME:

- FIRST LOOK AT DATA SCIENCE & THE DATA MINING WORKFLOW
- DATA EXPLORATION WITH UNIX
- DATA VISUALIZATION WITH R & GGPLOT2

TODAY:

- WHAT IS MACHINE LEARNING?
- DATA EXPLORATION WITH UNIX
- DATA VISUALIZATION WITH R & GGPLOT2

What's big data?

The practical viewpoint:

1. $O(n^2)$ algorithm feasible: small data
2. Fits on one machine: medium data
3. Doesn't fit on one machine: big data

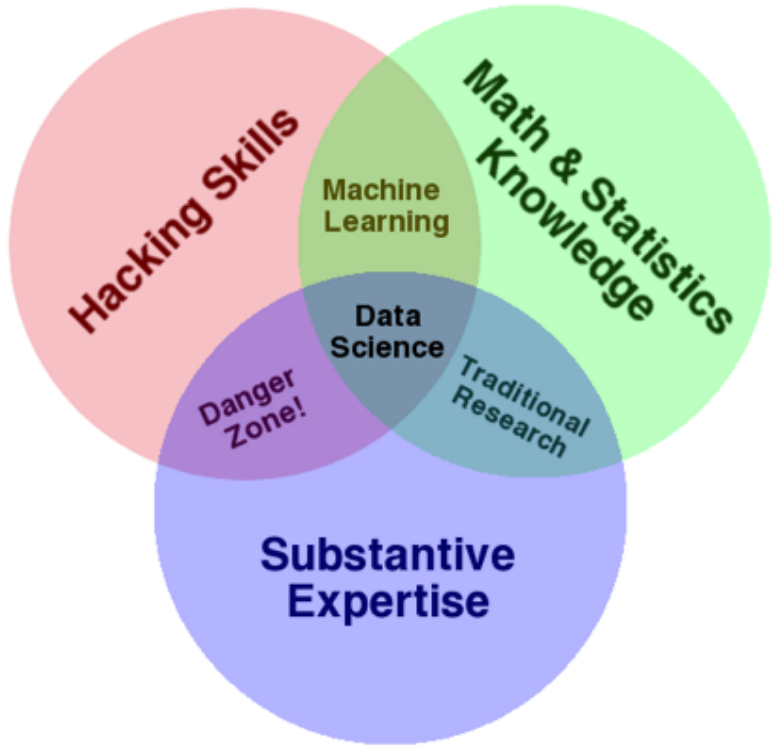
source: http://people.cs.umass.edu/~mcgregor/stocworkshop/langford.pdf

I. WHAT IS MACHINE LEARNING?
II. MACHINE LEARNING PROBLEMS

EXERCISES:
III. MULTIPLE REGRESSION & FEATURE EXTRACTION

I. WHAT IS MACHINE LEARNING?

REMEMBER THIS?



source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

"A field of study that gives computers the ability to learn without being explicitly programmed." (1959)

Arthur Samuel, AI pioneer
Source: Stanford

"A field of study that gives computers the ability to learn without being explicitly programmed."

- What does it mean for a computer to learn?
- Do they think?
- Do they feel?

"A computer program is said to learn from experience E with respect to some set of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". (1989)

Tom Mitchell, Professor, CMU
(Source: CMU)

"A computer program is said to learn from experience *E* with respect to some set of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*".

A person is said to learn from a college course E with respect to some set of readings and midterms T and grades P, if its performance at tasks in T, as measured by P, improves with E.

from Wikipedia:

"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."

source: http://en.wikipedia.org/wiki/Machine_learning

from Wikipedia:

"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."

"The core of machine learning deals with *representation* and *generalization*…"

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."

"The core of machine learning deals with *representation* and *generalization*…"

- *representation* – extracting structure from data

source: http://en.wikipedia.org/wiki/Machine_learning

from Wikipedia:

"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."

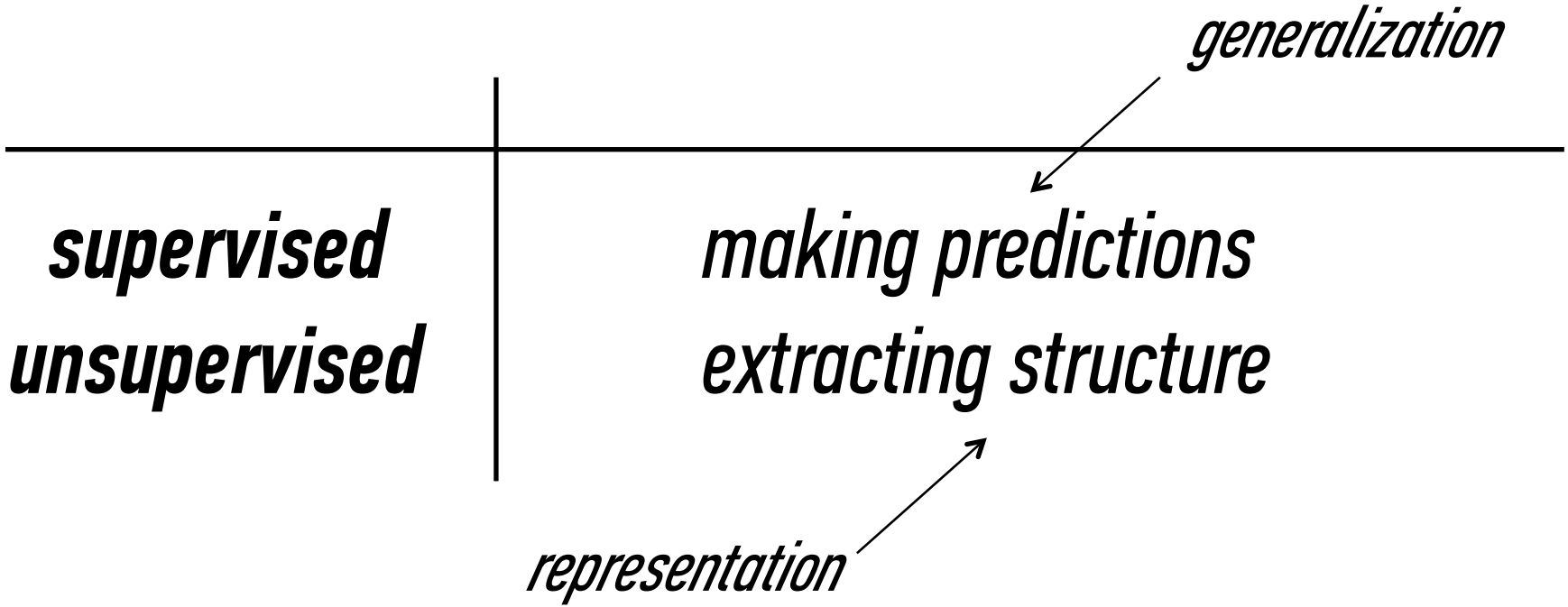
"The core of machine learning deals with *representation* and *generalization*…"

- *representation* – extracting structure from data
- *generalization* – making predictions from data

source: http://en.wikipedia.org/wiki/Machine_learning

II. MACHINE LEARNING PROBLEMS

supervised	*making predictions*
unsupervised	*extracting structure*


generalization
supervised
unsupervised
making predictions
extracting structure
representation

	continuous	*categorical*
	quantitative	*qualitative*

	continuous	*categorical*
	quantitative	*qualitative*

NOTE

The space where data live is called the *feature space*.

Each point in this space is called a *record*.

	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

NOTE

We will implement solutions using *models* and *algorithms*.

Each will fall into one of these four buckets.

supervised	*past predictions/making predictions*
unsupervised	*extracting structure*

ANSWER

The goal is determined by the type of problem.

	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

ANSWER

The right approach is determined by the desired solution.

	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

ANSWER

The
is d
des

NOTE

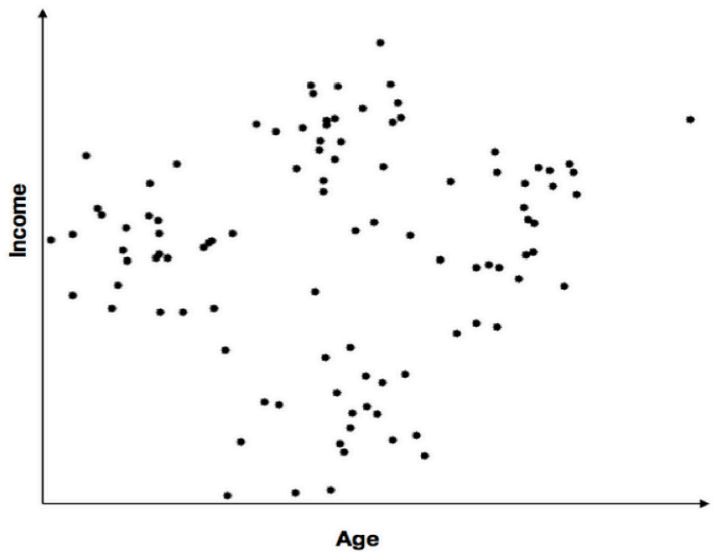
All of this depends on your data!

Supervised Learning - Can we create a function that predicts a value based on labeled training data?

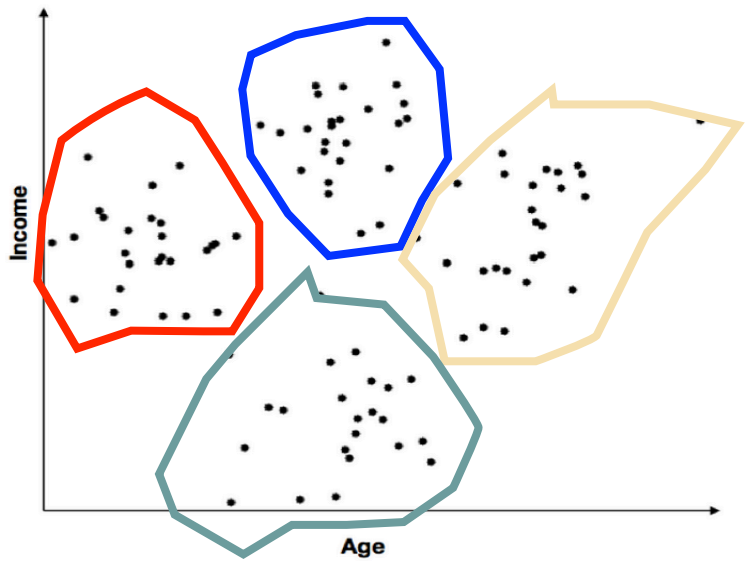
Regression example: Alan is 30 years old and can eat *four donuts an hour*. Betty is 60 years old, and can eat *two donuts an hour.* Cameron is 15 years old--how many donuts an hour eaten would be a good guess? This prediction is a regression model.

Classification example: Let' s use the same data above. What is the probability that Cameron will eat eight donuts? Here, we have an answer and am now calculating the probability that an outcome has occurred.

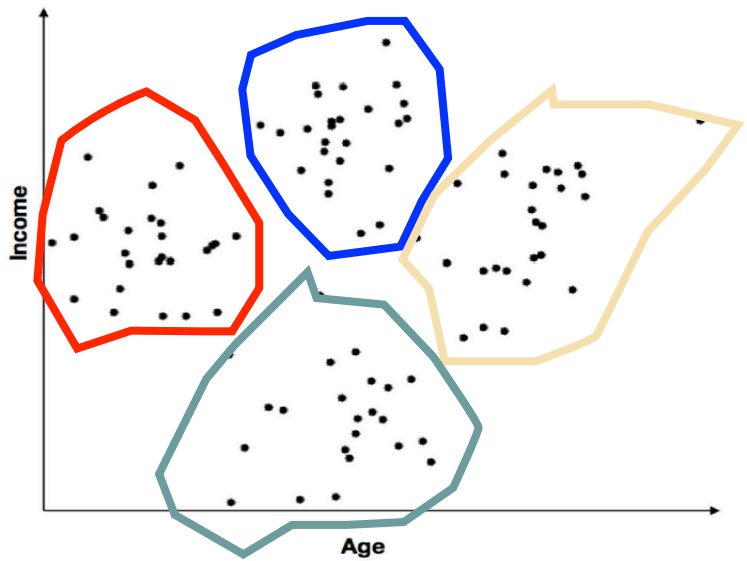
Unsupervised Learning - Can we find structure to unlabeled data?

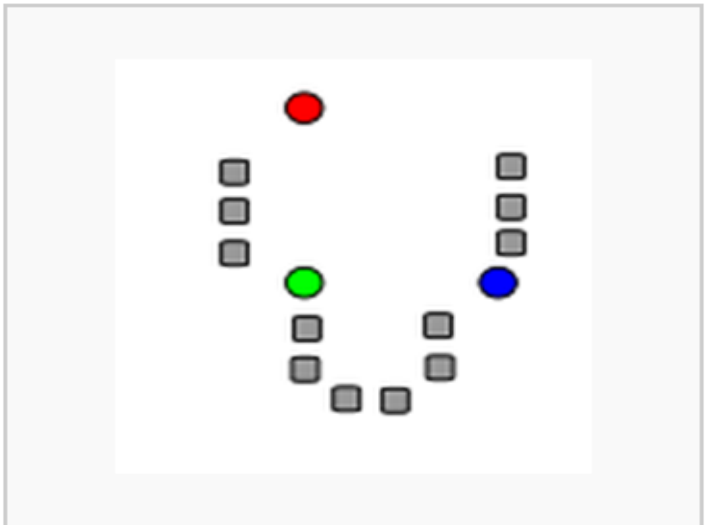


Unsupervised Learning - Can we find structure to unlabeled data?

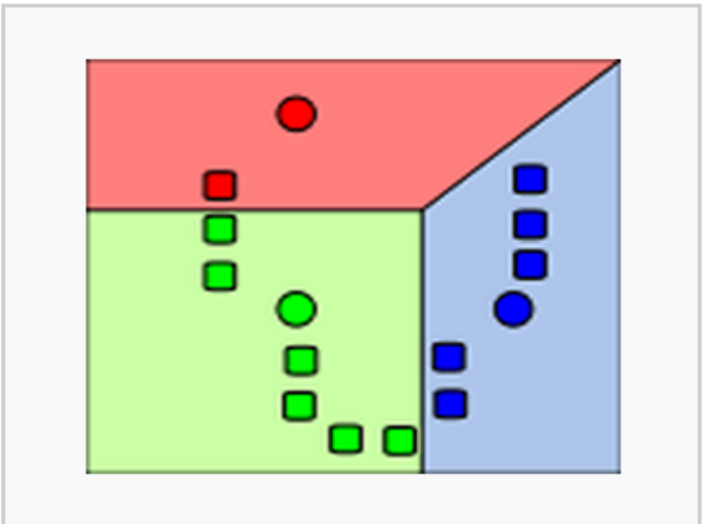


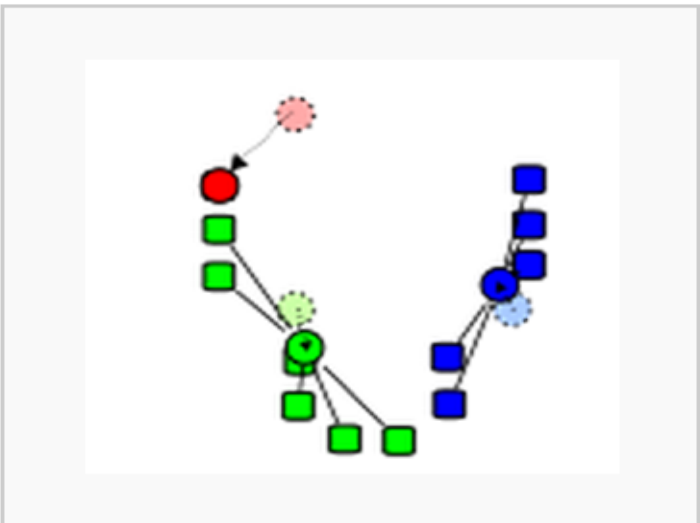
Unsupervised Learning - Can we find structure to unlabeled data?





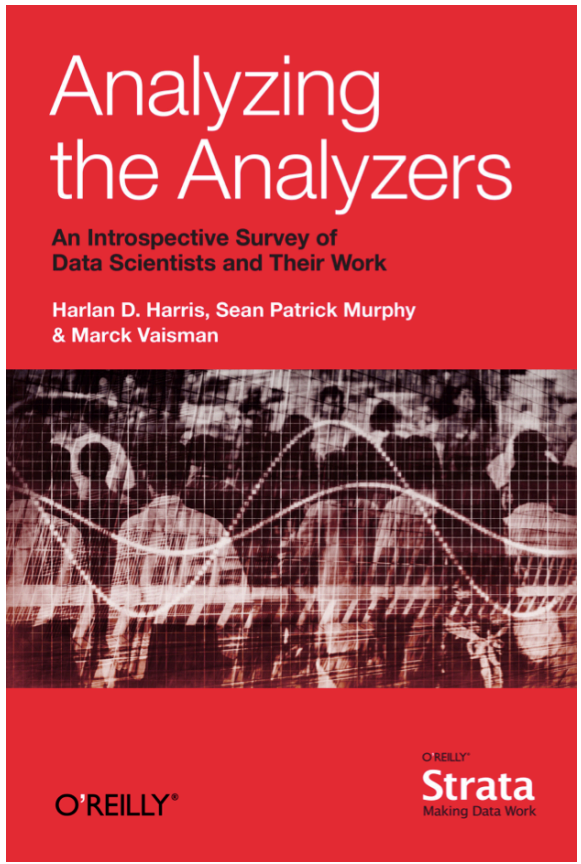
http://shabal.in/visuals/kmeans/2.html





DATA SCIENCE IN THE NEWS:

The four types of Data Scientists (June 2013)


Analyzing
the Analyzers
An Introspective Survey of
Data Scientists and Their Work
Harlan D. Harris, Sean Patrick Murphy
& Marck Vaisman
O'REILLY®
O'REILLY®
Strata
Making Data Work

There has been intense excitement in recent years around activities labeled "data science," "big data," and "analytics." However, the lack of clarity around these terms and, particularly, around the skill sets and capabilities of their practitioners has led to inefficient communication between "data scientists" and the organizations requiring their services.

Problem

To address this issue, we surveyed several hundred practitioners via the Web to explore the varieties of skills, experiences, and viewpoints in the emerging data science community.

Acquire

We created a five-page web survey, taking less than 10 minutes to complete and focusing on five areas: skills, experiences, education, self-identification, and web presence. (Regarding web presence, we asked for those willing to share their LinkedIn, Meetup, and GitHub profiles, so that we could perform additional text analysis. However, due to relatively low response rates and some technical issues, the re- sults were not usable, and will not be reported.)

Acquire (Interact)

After testing the survey on a small group of friends and colleagues, we shared it broadly, evangelized it to professional Meetups, and posted links on every relevant social network we could think of. By the end of the project, we received over 250 completed surveys from around the globe.

Acquire (Interact)

We used dimensionality reduction techniques, non-negative Matrix Factorization and unsupervised learning [clustering] *algorithms to divide potential data scientists into five categories based on their self-ranked skill sets (Statistics, Math/Operations Research, Business, Programming, and Machine Learning/Big Data), and four categories based on their self-identification (Data Researchers, Data Businesspeople, Data Engineers, and Data Creatives).*

Parse, Filter, Mine

Further examining the respondents based on their division into these categories provided additional insights into the types of professional activities, educational background, and even scale of data used by different types of Data Scientists.

Refine

In this report, we combine our results with insights and data from others to provide a better understanding of the diversity of practitioners, and to argue for the value of clearer communication around roles, teams, and careers.

Represent and interact

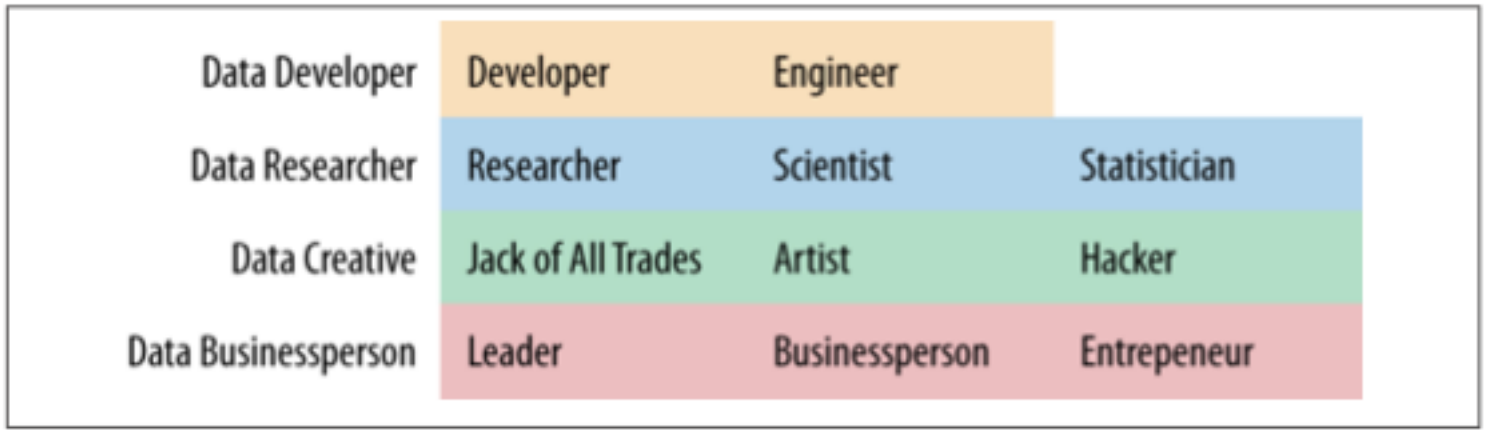


Figure 3-1. Respondents tended to agree or disagree consistently to questions such as "I think of myself as a/an X" in each Self-ID Group. Our suggested Self-ID Group names are shown, along with the self-ID categories most strongly associated with each Group.

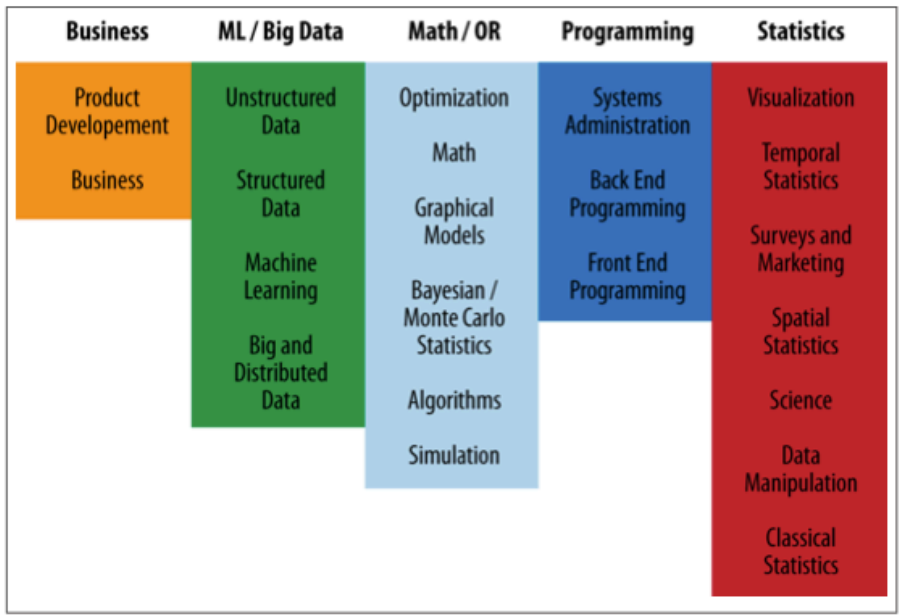


Figure 3-2. Respondents tended to rank similarly skills in each Skills Group. Our suggested Skills Goup names are shown, along with the skills most strongly associated with each Group. ML = "Machine Learning" and OR = "Operations Research."

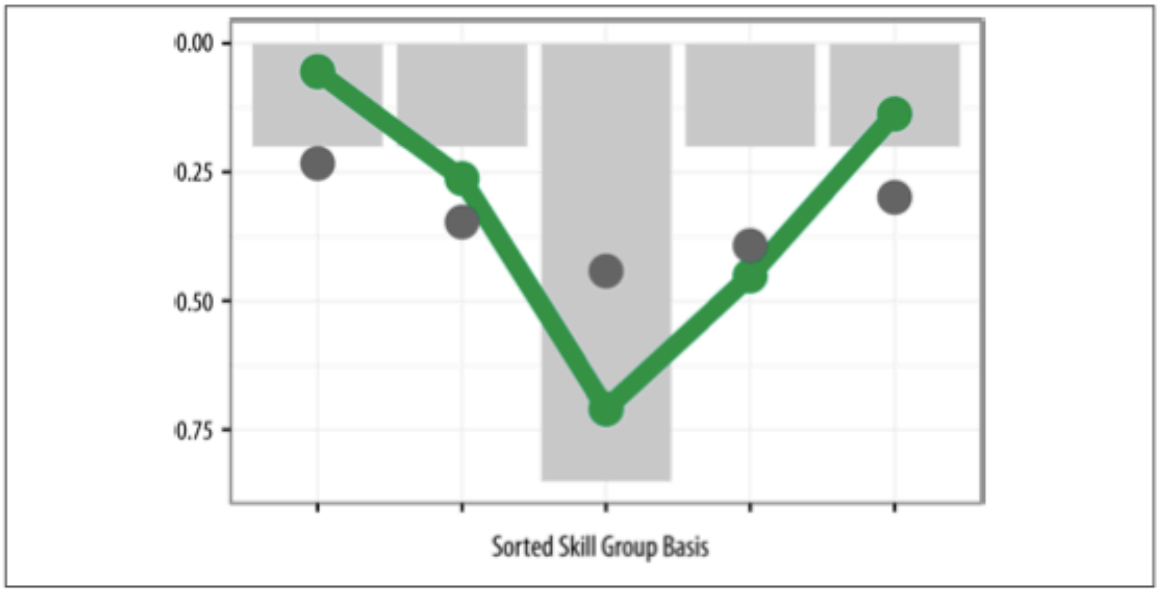


Figure 4-1. Skill Group strength for "ideal" professionals (grey bar), simulated controls (grey dots), and mean of surveyed respondents (green). Loadings are sorted from center out on a per-respondent basis.

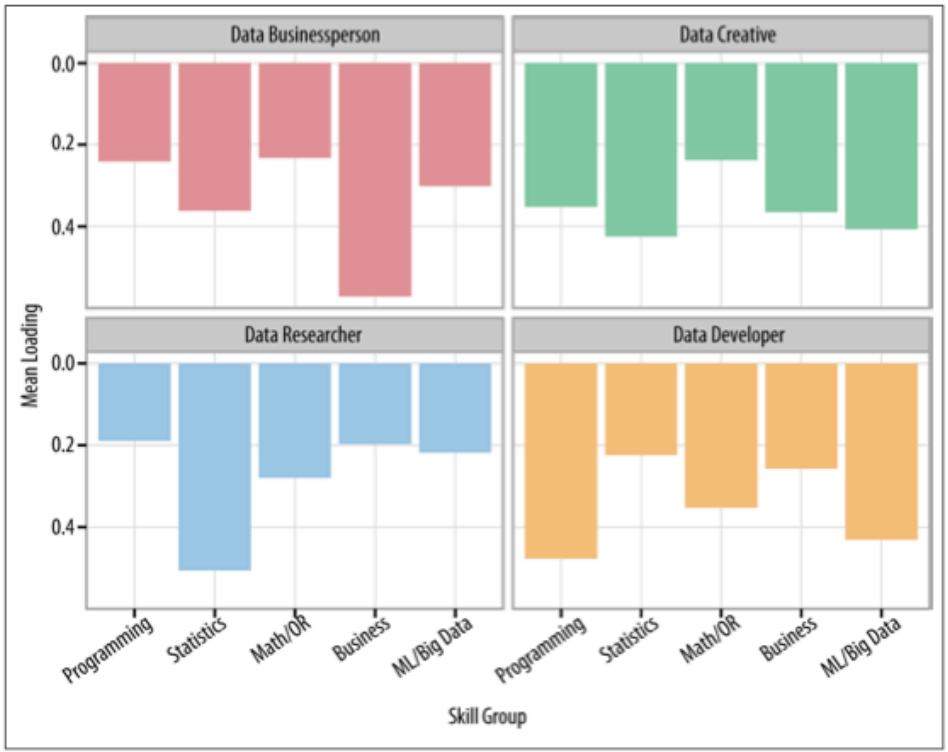


Figure 4-2. Mean Skill Group loadings for survey participants categorized into four Self-ID Groups.

III. RELATIONSHIPS AMONG SEVERAL VARIABLES

EXERCISE – MULTIPLE REGRESSION (BACKWARD ELIMINATION)

KEY OBJECTIVES

- Create a regression model using several independent variables

- Extract meaningful features

TOOLS

- R (plot, lm, update)

HOMEWORK

There has been intense excitement in recent years around activities labeled "data science," "big data," and "analytics." However, the lack of clarity around these terms and, particularly, around the skill sets and capabilities of their practitioners has led to inefficient communication between "data scientists" and the organizations requiring their services.

Problem