GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 2: MACHINE LEARNING

LAST TIME:

- FIRST LOOK AT DATA SCIENCE & THE DATA MINING WORKFLOW
- DATA VISUALIZATION WITH R & GGPLOT2
- FIRST LINEAR MODEL

QUESTIONS?

EXERCISES:

I. MULTIPLE REGRESSION & FEATURE EXTRACTION

II. WHAT IS MACHINE LEARNING?

III. MACHINE LEARNING PROBLEMS

I. 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)

ASSUMPTIONS FOR LINEAR REGRESSION

1) **Linearity** of the relationship between dependent and independent variables (doesn't mean the relation between y and x has to be linear since we can use transformations if y and x as well)
2) **Independence** of the errors
3) **Homoscedasticity** (constant variance of the errors)
   1) versus time
   2) Versus the predictions or any independent variables
4) **Normality** of the error distribution

II. WHAT IS 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*.”

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
- *generalization* – making predictions from data

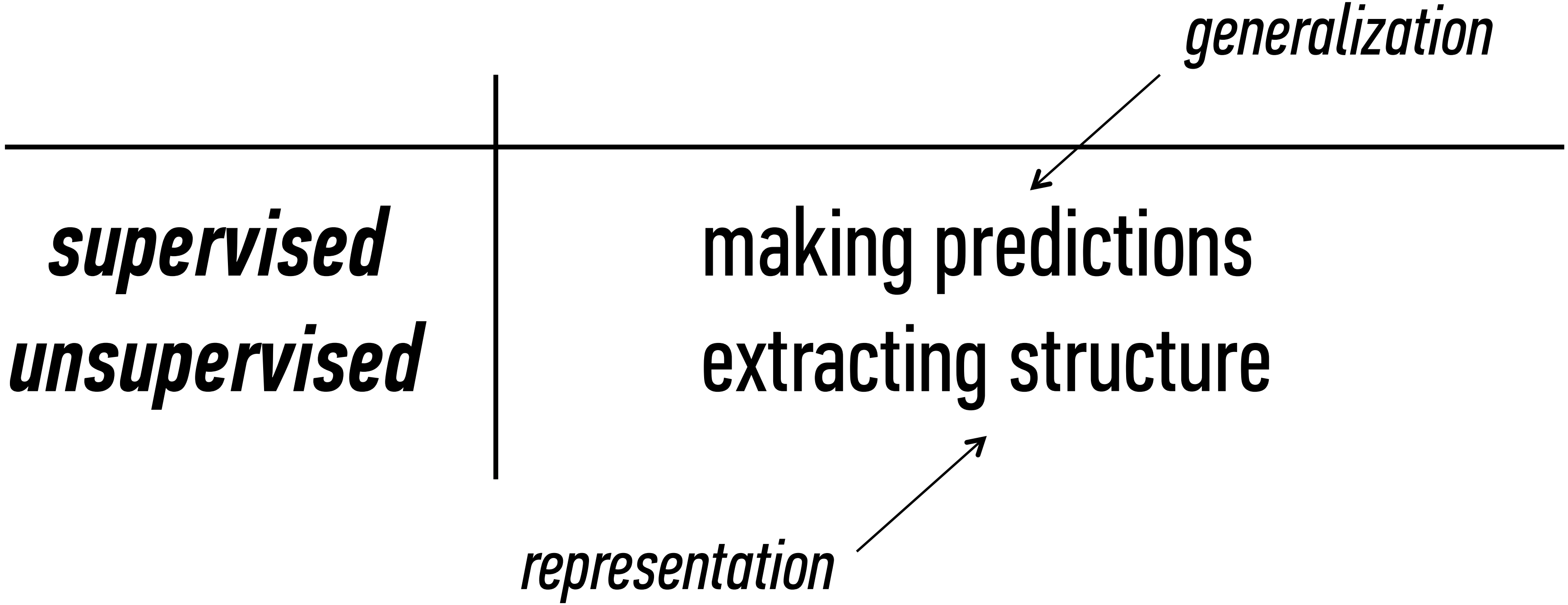
source: http://en.wikipedia.org/wiki/Machine_learning

III. MACHINE LEARNING PROBLEMS

TYPES OF LEARNING PROBLEMS

supervised	making predictions
unsupervised	extracting structure

REMEMBER WHAT WE SAID BEFORE?



generalization

supervised — making predictions

unsupervised — extracting structure

representation

TYPES OF DATA

	continuous	*categorical*
	quantitative	qualitative

TYPES OF DATA

	continuous	*categorical*
	quantitative	qualitative

NOTE

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

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

TYPES OF ML SOLUTIONS

	continuous	*categorical*
supervised	regression	classification
unsupervised	dimension reduction	clustering

TYPES OF ML SOLUTIONS

	continuous	*categorical*
supervised	regression	classification
unsupervised	dimension reduction	clustering

NOTE

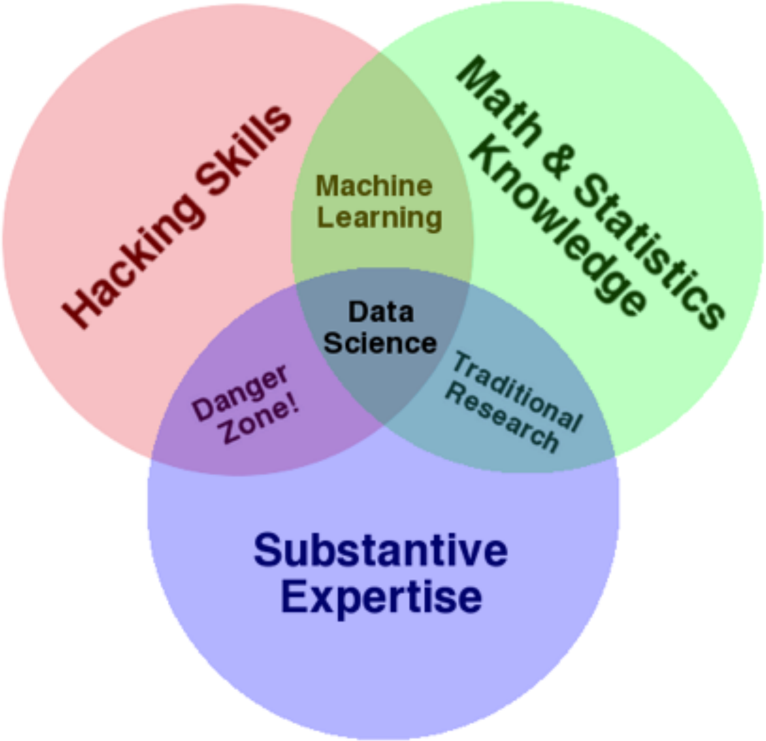
We will implement solutions using *models* and *algorithms.*

Each will fall into one of these four buckets depending on the type of problem and type of data.

GENERAL ASSEMBLY

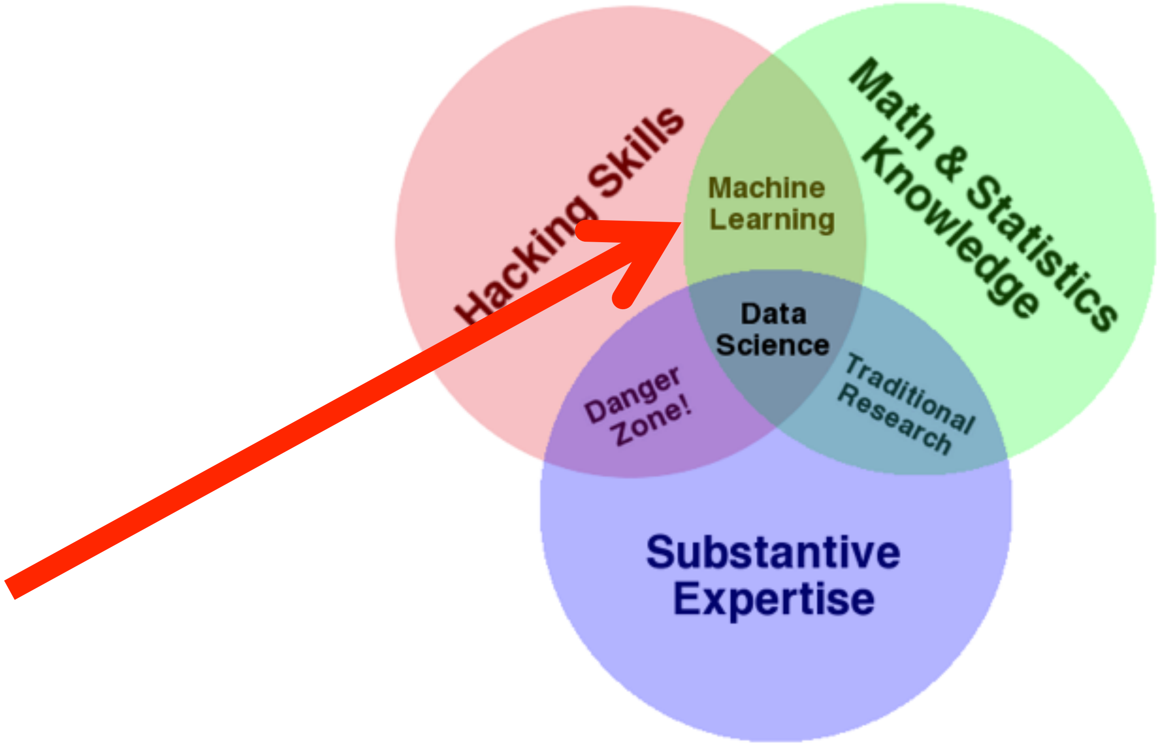
DATA SCIENCE
AND
MACHINE LEARNING

REMEMBER THIS?



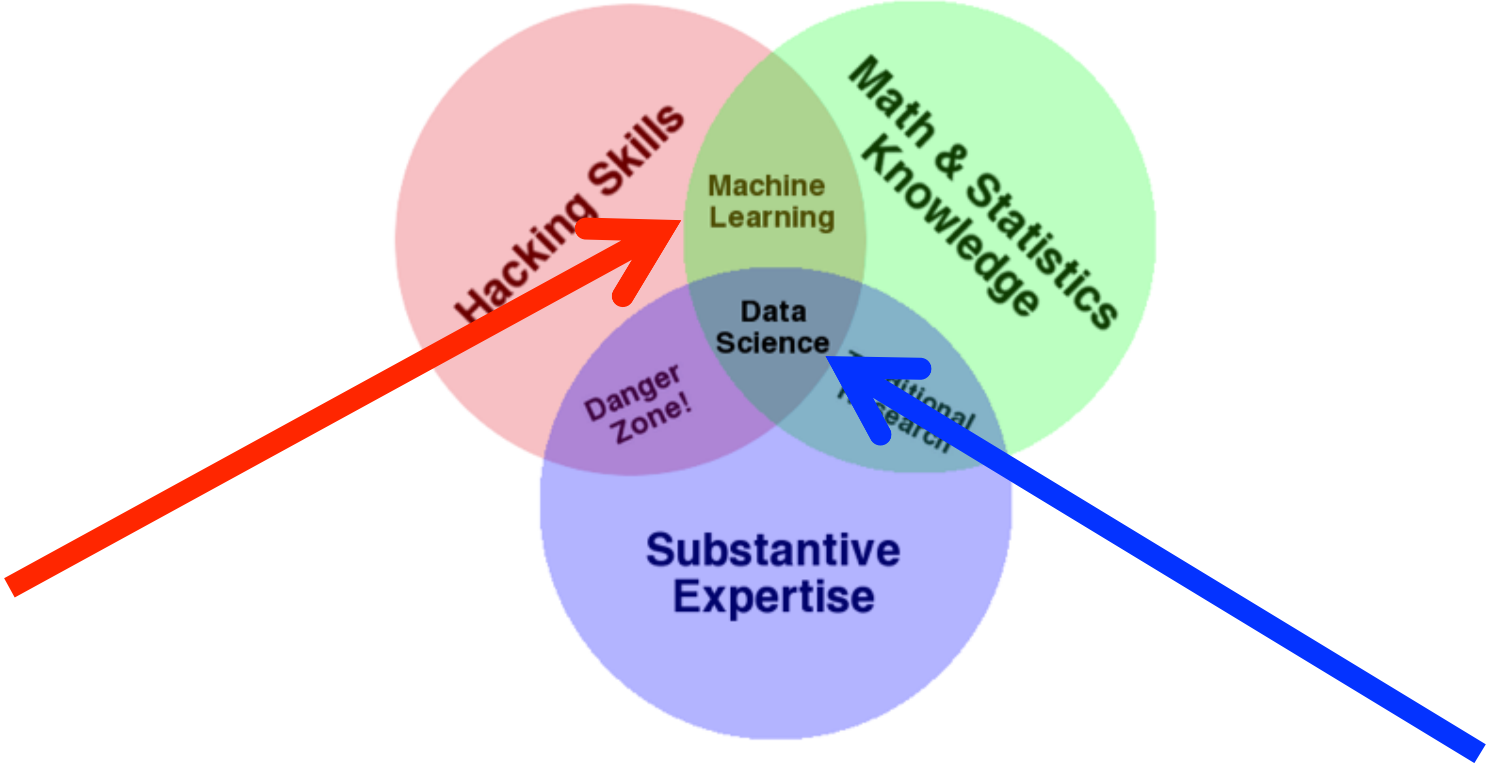
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU ARE HERE



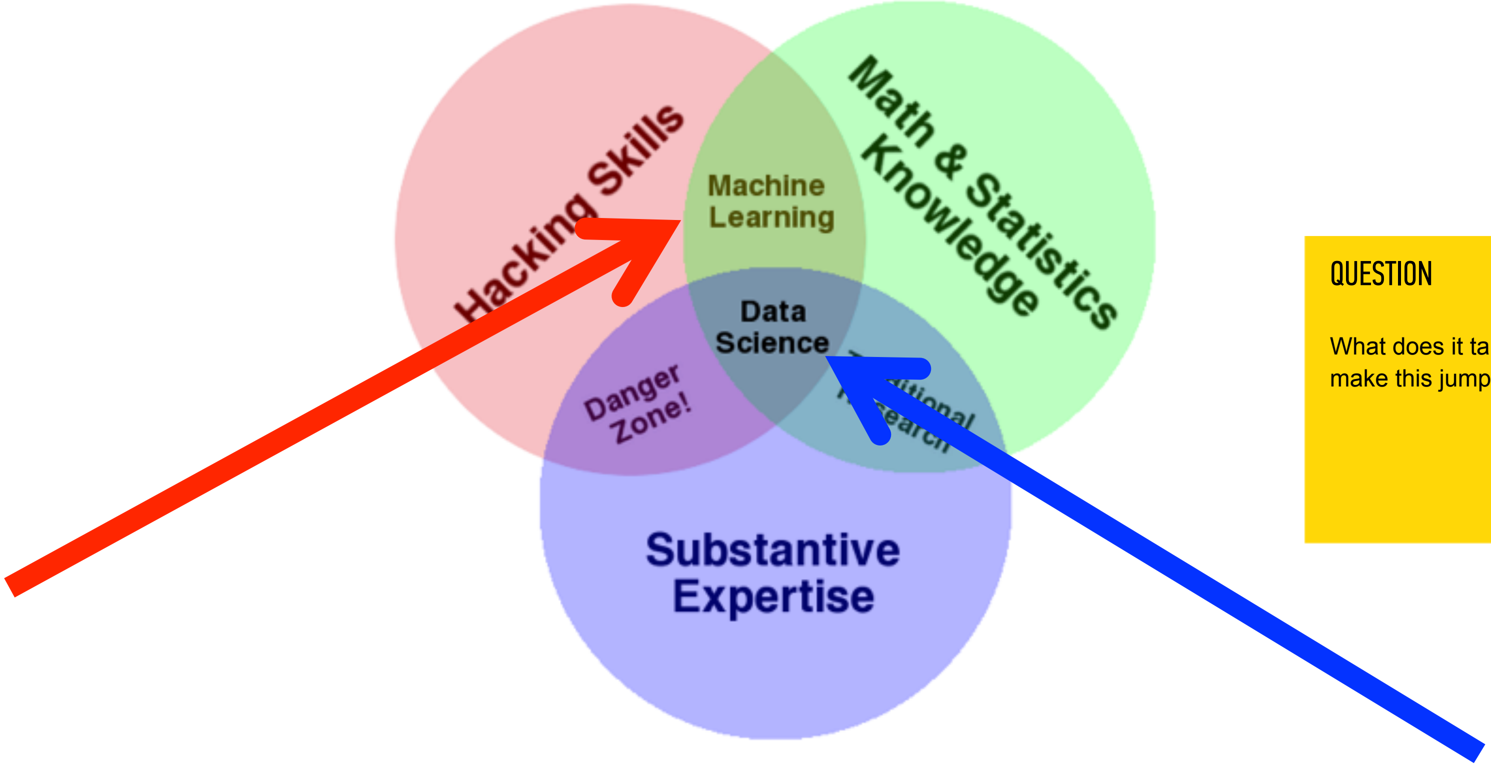
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU WANT TO GO HERE



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

YOU WANT TO GO HERE

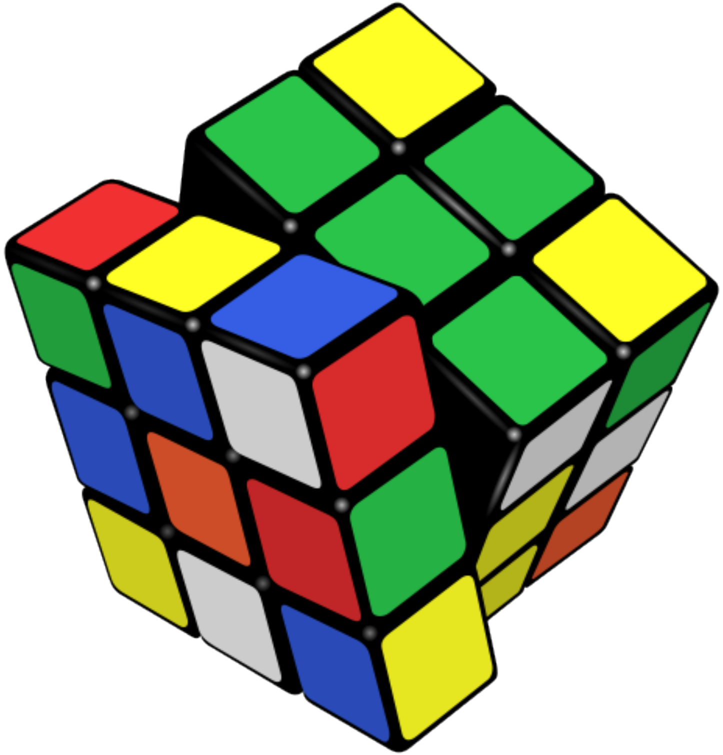


QUESTION

What does it take to make this jump?

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

ANSWER: PROBLEM SOLVING!



NOTE

Implementing solutions to ML problems is the focus of this course!

DISCUSSION