GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 2: MACHINE LEARNING

LAST TIME:

- FIRST LOOK AT DATA SCIENCE
- THEN THE DATA MINING WORKFLOW
- GIT INTRO

QUESTIONS?

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

AGENDA

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

EXERCISES:
III. I-PYTHON NOTEBOOK INTRO

I. 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…”

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

WHAT IS MACHINE LEARNING?

from Wikipedia:

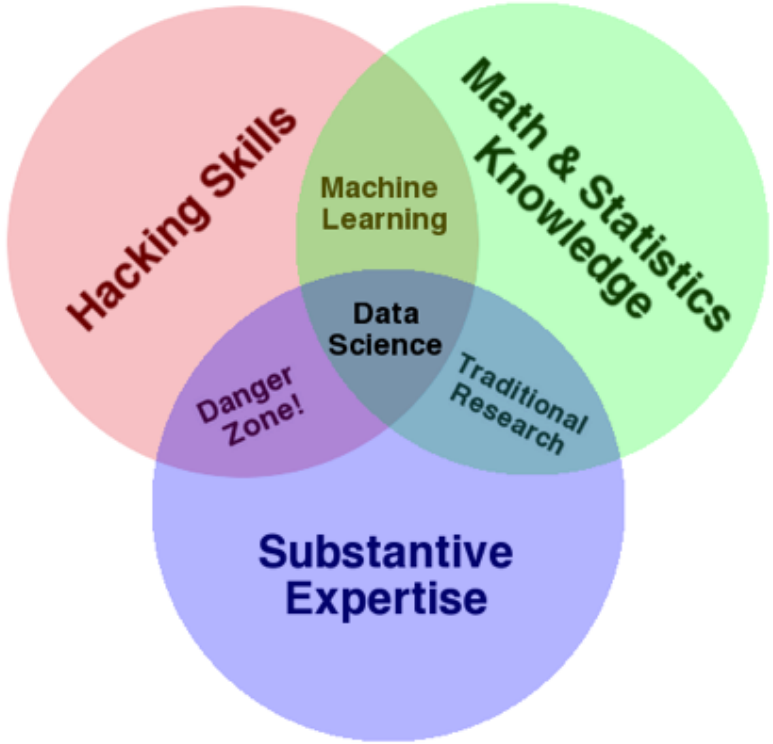
“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

“The core of machine learning deals with representation and generalization…”

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

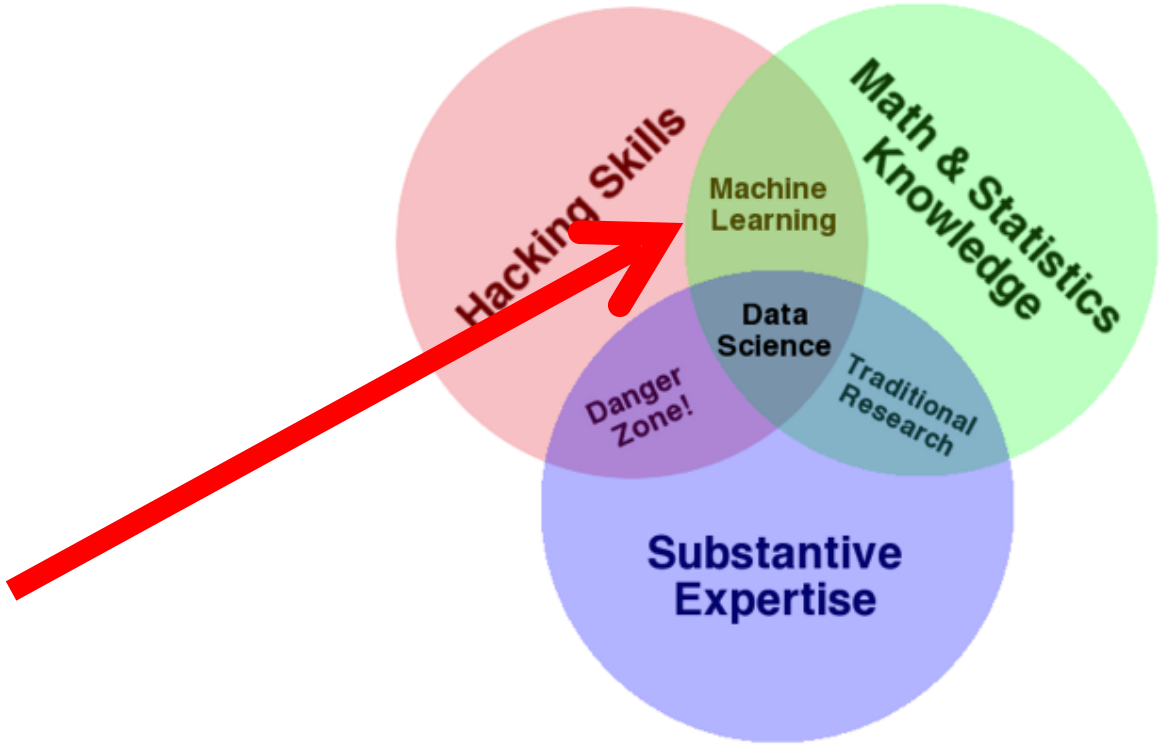
source: http://en.wikipedia.org/wiki/Machine_learning

REMEMBER THIS?



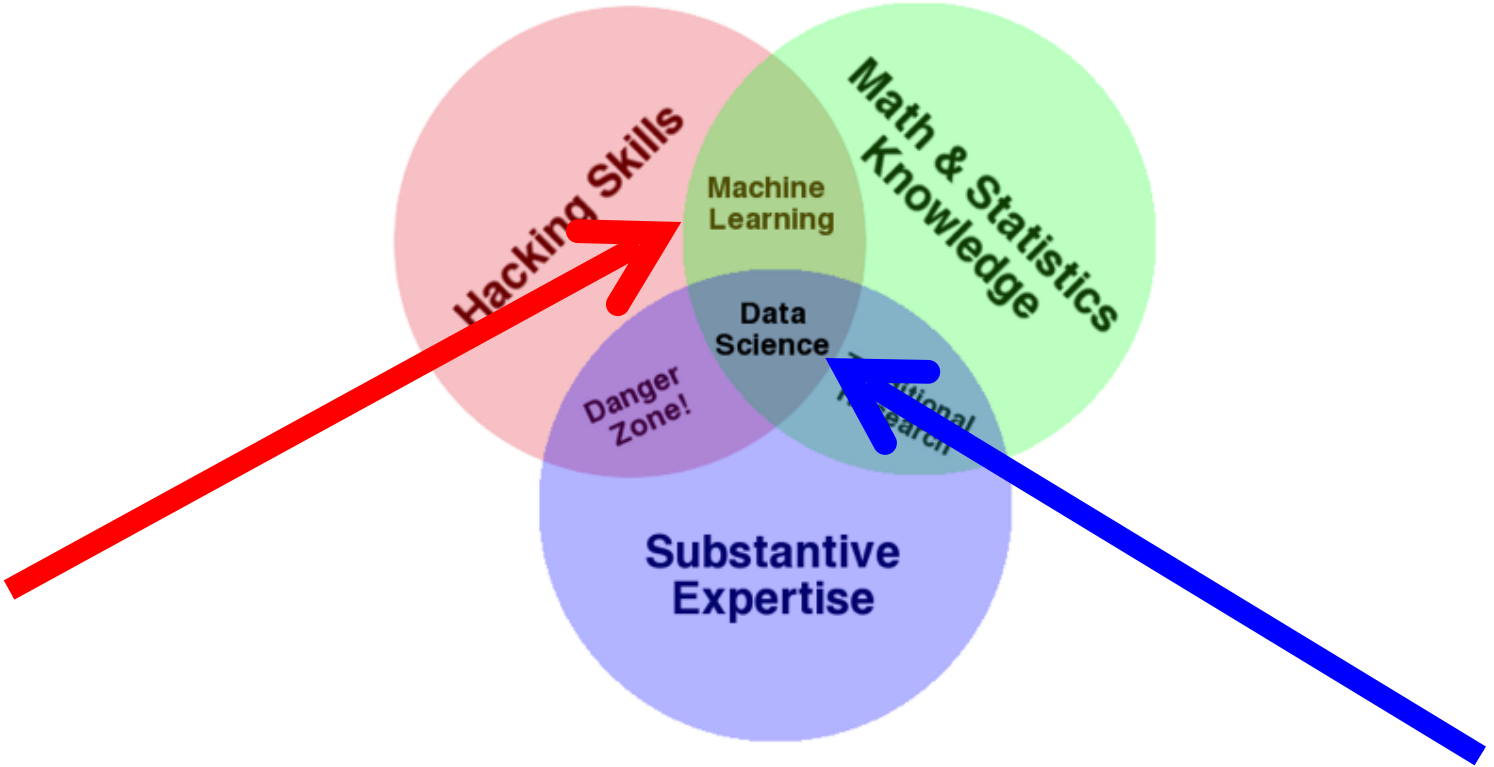
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU ARE HERE



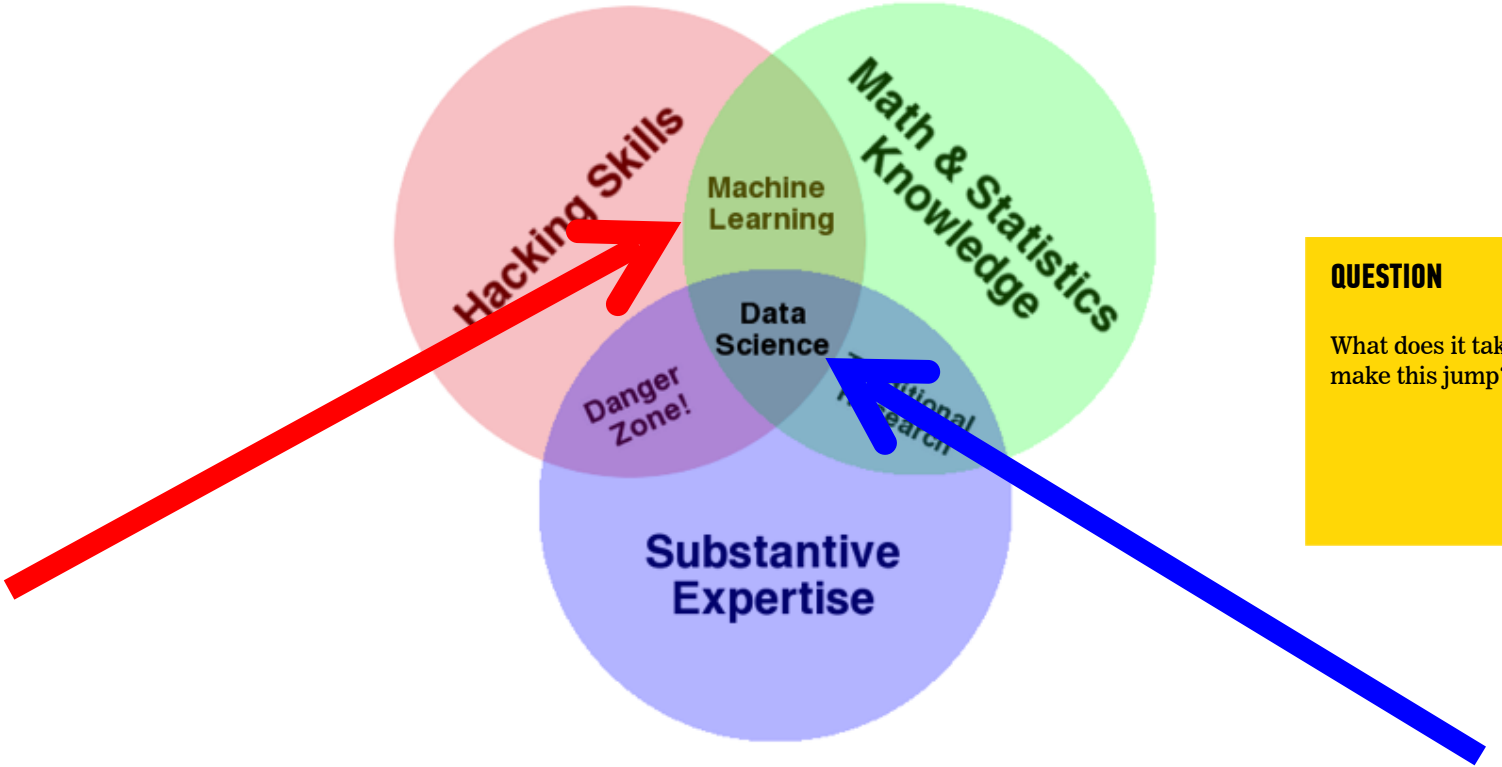
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU WANT TO GO HERE



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

YOU WANT TO GO HERE

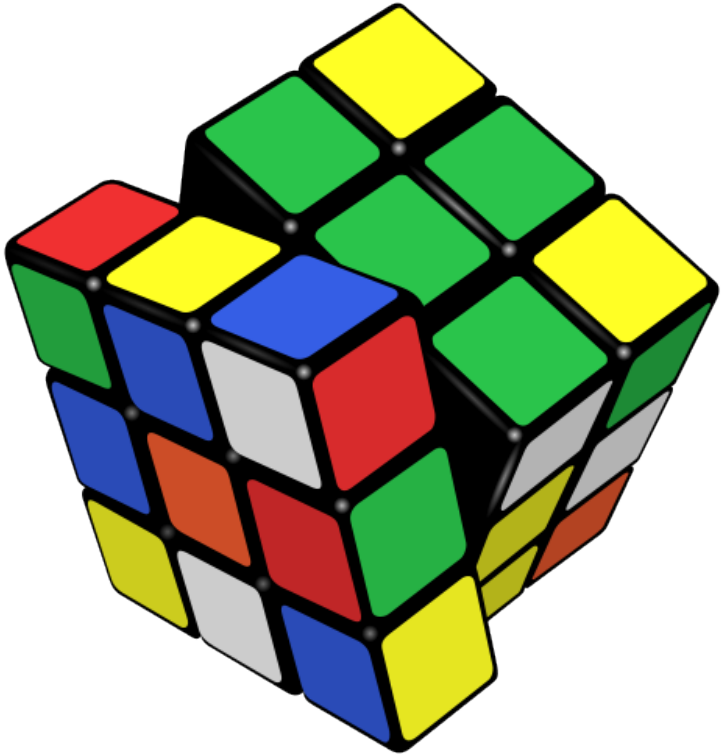


QUESTION

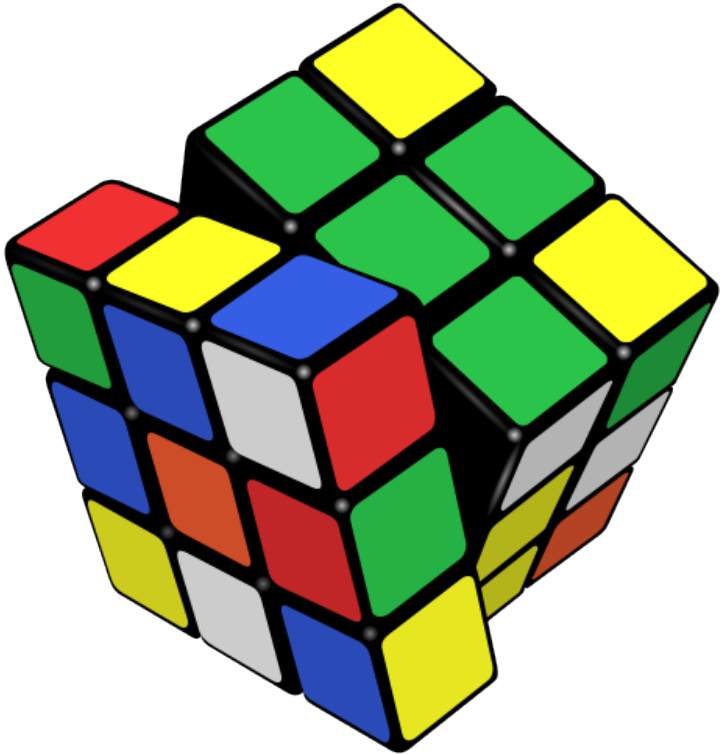
What does it take to make this jump?

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

ANSWER: PROBLEM SOLVING!



ANSWER: PROBLEM SOLVING!



NOTE

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

II. MACHINE LEARNING PROBLEMS

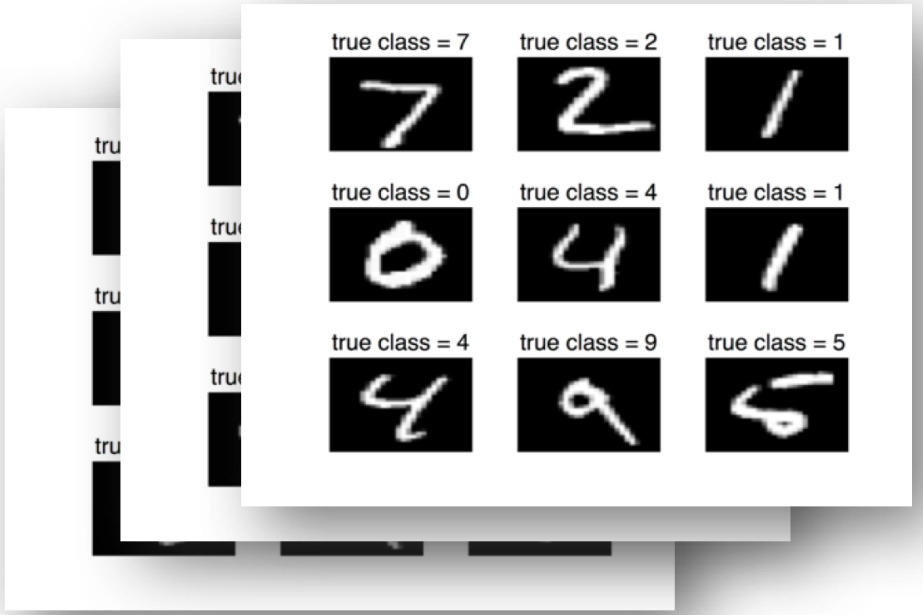
Learning is not about memorizing and being able to recall, it is about generalizing the conclusions to previously unseen examples

TYPES OF LEARNING?

Supervised learning: the goal is to learn mapping from given inputs x to outputs y, given a **labeled** set of input-output pairs

OCR


OCR


true class = 7
true class = 2
true class = 1
true class = 0
true class = 4
true class = 1
true class = 4
true class = 9
true class = 5

CREDIT SCORING

	Client 1	Client 2	Client 3
Age	23	30	19
Gender	M	F	M
Annual salary	$30,000	$45,000	$15,000
Years in residence	3 years	1 year	3 month
Years in job	1 year	1 year	1 month
Current debt	$5,000	$1,000	$10,000
Paid off credit	Yes	Yes	No

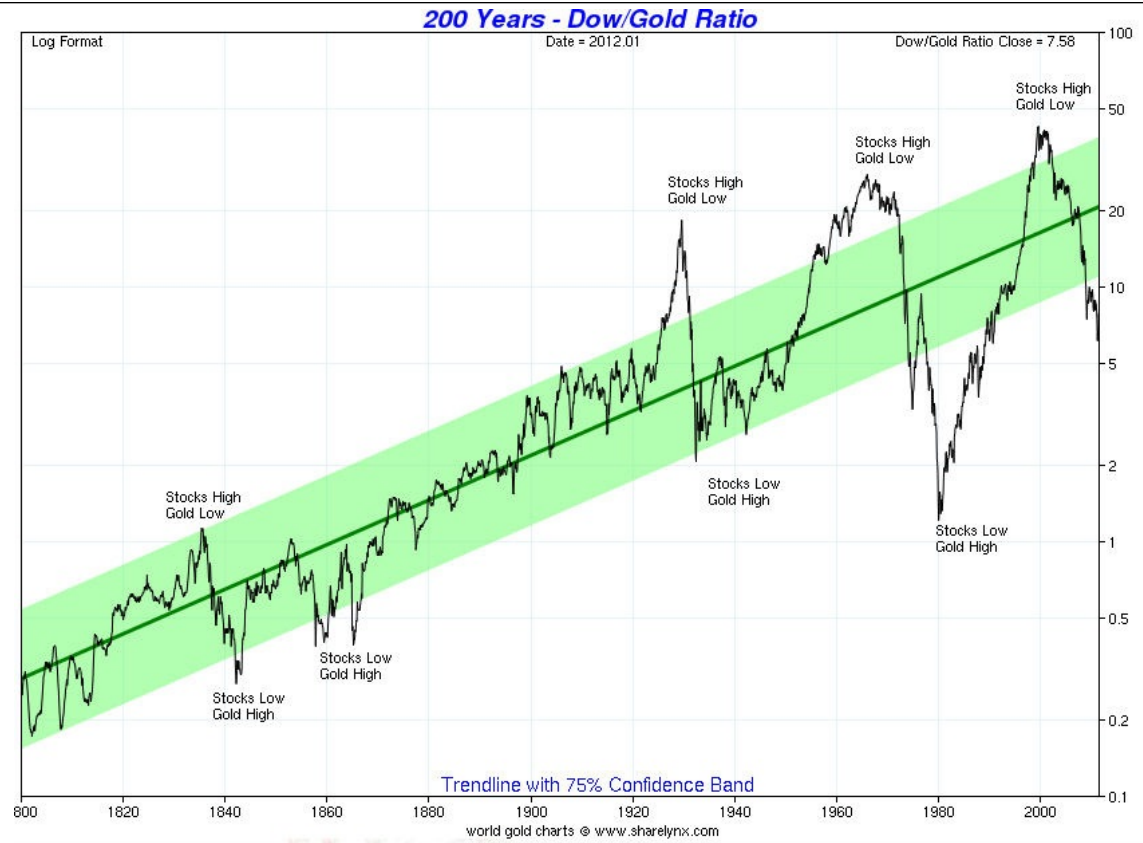
CREDIT SCORING

	Client 1	Client 2	Client 3
Age	23	30	19
Gender	M	F	M
Annual salary	$30,000	$45,000	$15,000
Years in residence	3 years	1 year	3 month
Years in job	1 year	1 year	1 month
Current debt	$5,000	$1,000	$10,000
Paid off credit	Yes	Yes	No

	Applicant
Age	25
Gender	M
Annual salary	$25,000
Years in residence	1 year
Years in job	2 year3
Current debt	$15,000
Credit decision/ score	???

EMAIL SPAM DETECTION

REGRESSION - STOCK PRICE PREDICTION

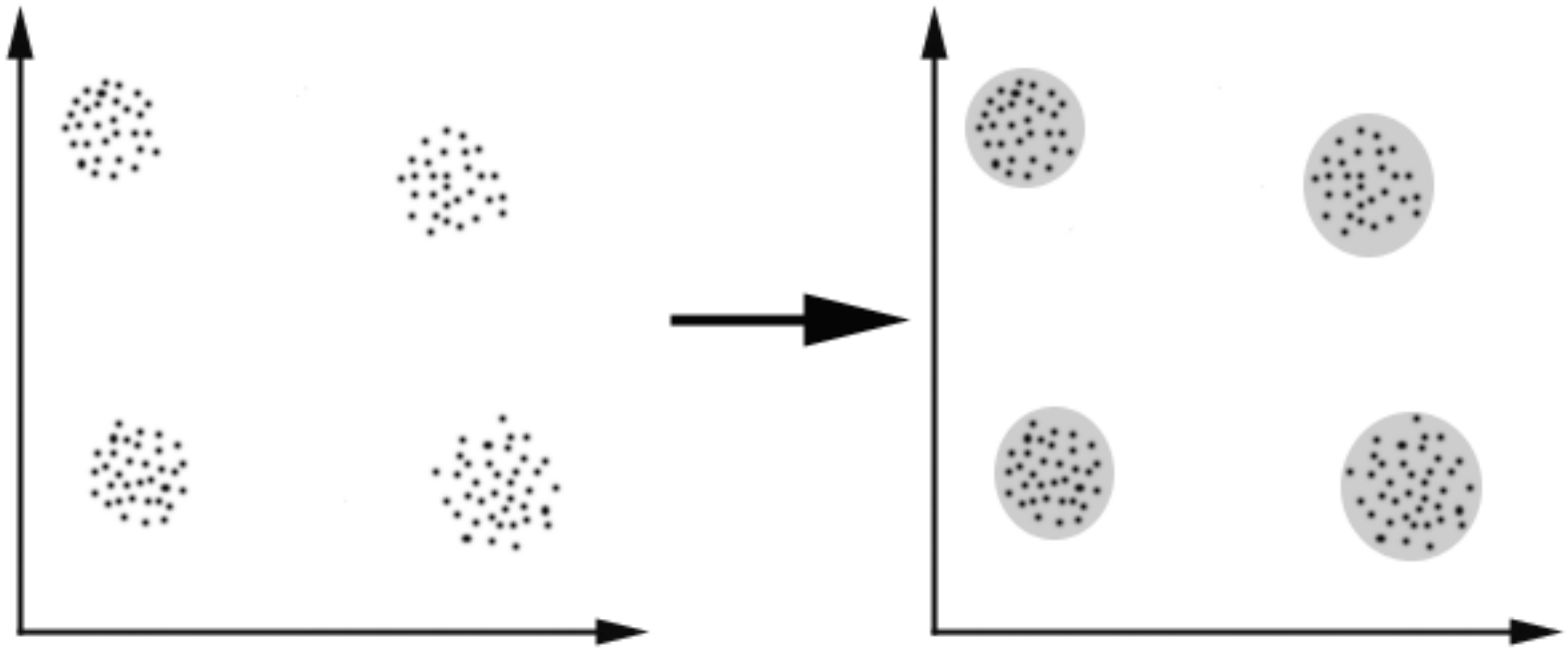


TYPES OF LEARNING?

Unsupervised learning: the goal is to learn interesting patterns and structure in data given only inputs

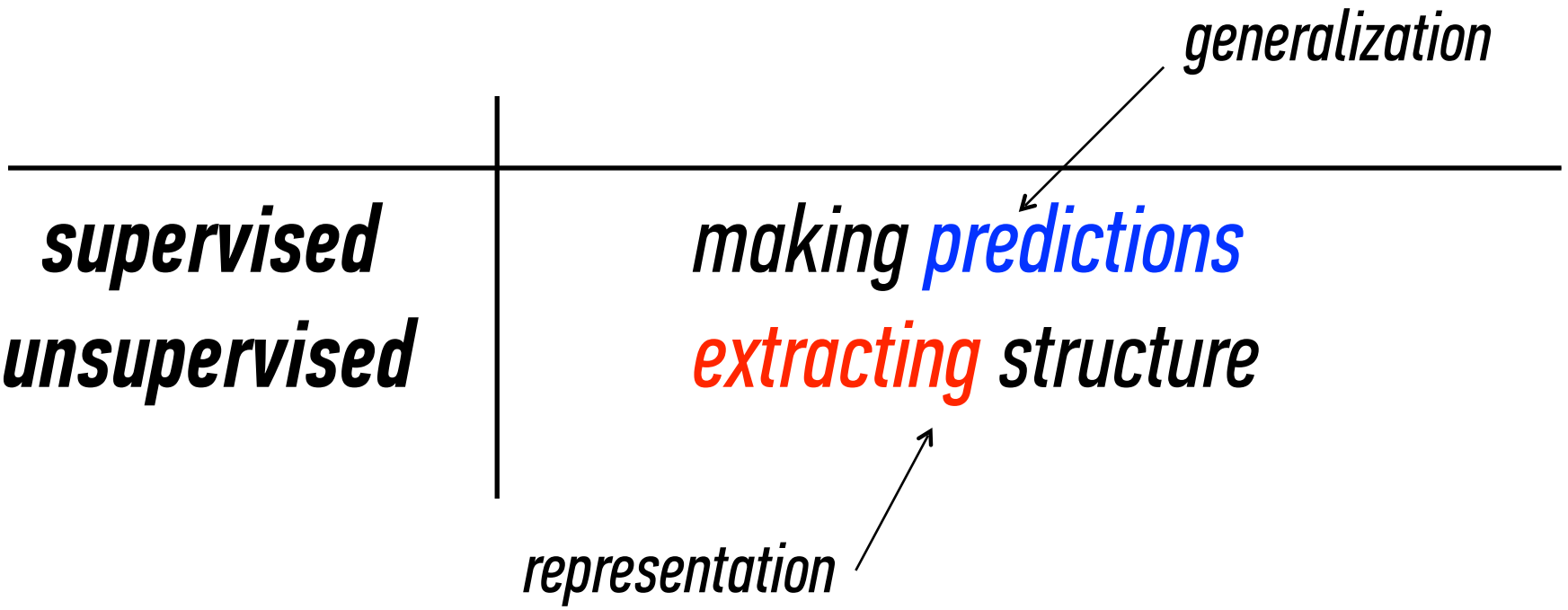
no label information given at all

CLUSTERING



supervised	*making predictions*
unsupervised	*extracting structure*

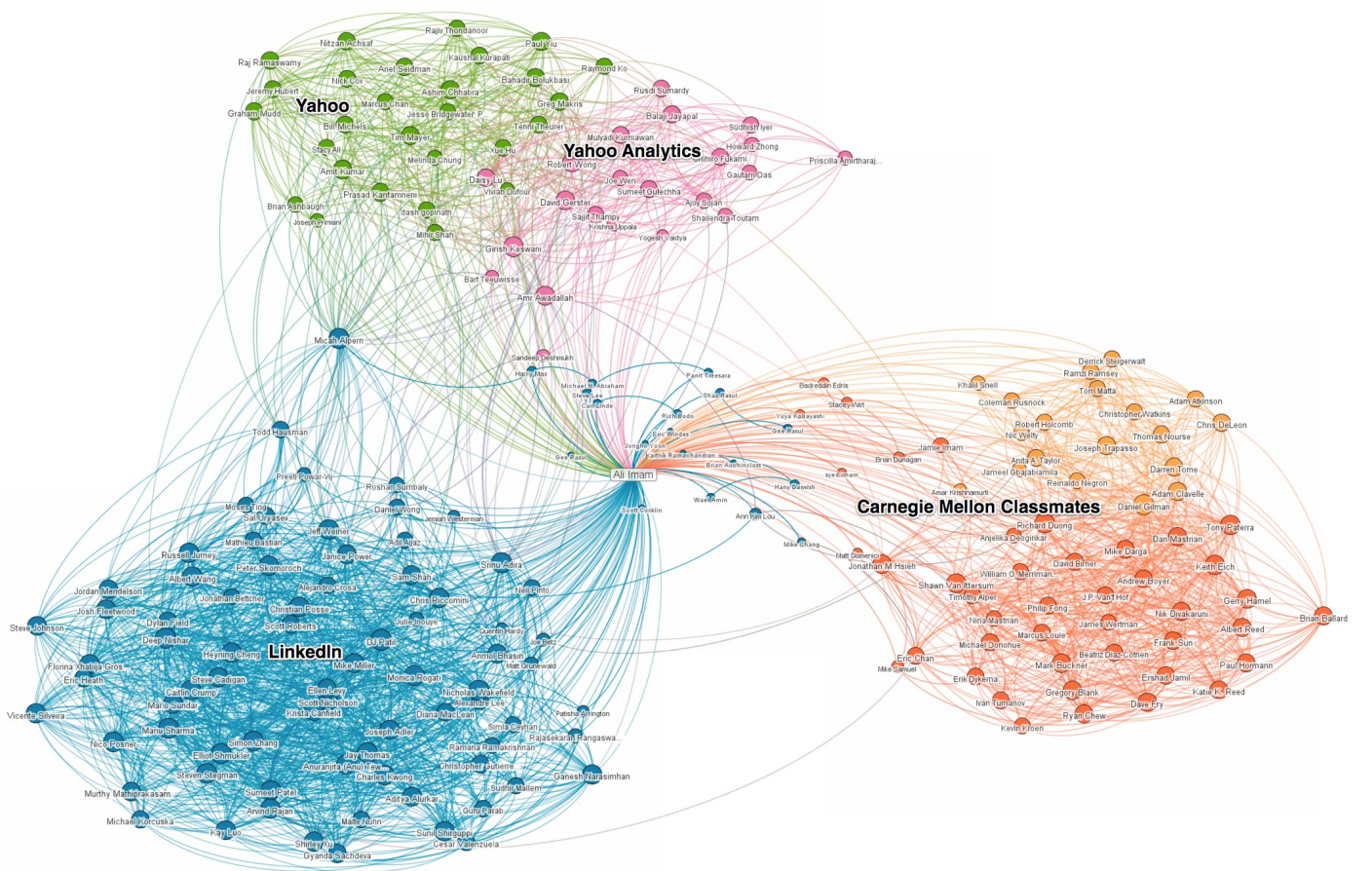
REMEMBER WHAT WE SAID BEFORE?



EXERCISE:

supervised or unsupervised?

COMMUNITY DETECTION IN SOCIAL NETWORKS



REGRESSION - HOUSE PRICE PREDICTION

DOCUMENT CLASSIFICATION

TYPES OF DATA

	continuous	*categorical*
	quantitative	*qualitative*

continuous	*categorical*
Height of children	*Eye colors*
Weight of cars	*Courses at GA*
Speed of the train	*Highest degree*
Temperature	*Gender*
Stock price	*If an email is spam or not*

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.

	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.

QUESTION

WHAT IS THE GOAL OF MACHINE LEARNING?

supervised	*making predictions*
unsupervised	*extracting structure*

ANSWER

The goal is determined by the type of problem.

QUESTION

HOW DO YOU DETERMINE THE RIGHT APPROACH?

APPROACHES TO ML PROBLEMS

	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

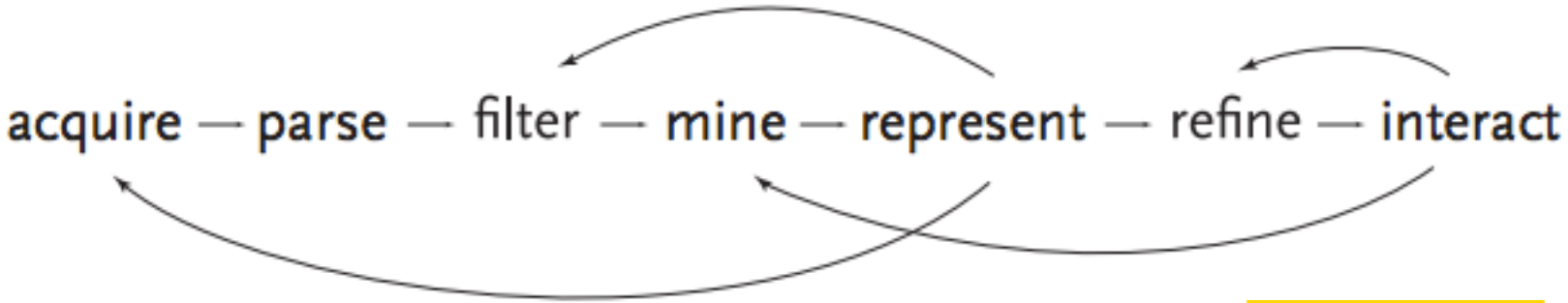
All of this depends on your data!

DO WE HAVE LABELS?

QUESTION

WHAT DO YOU DO WITH YOUR RESULTS?

THE DATA SCIENCE WORKFLOW

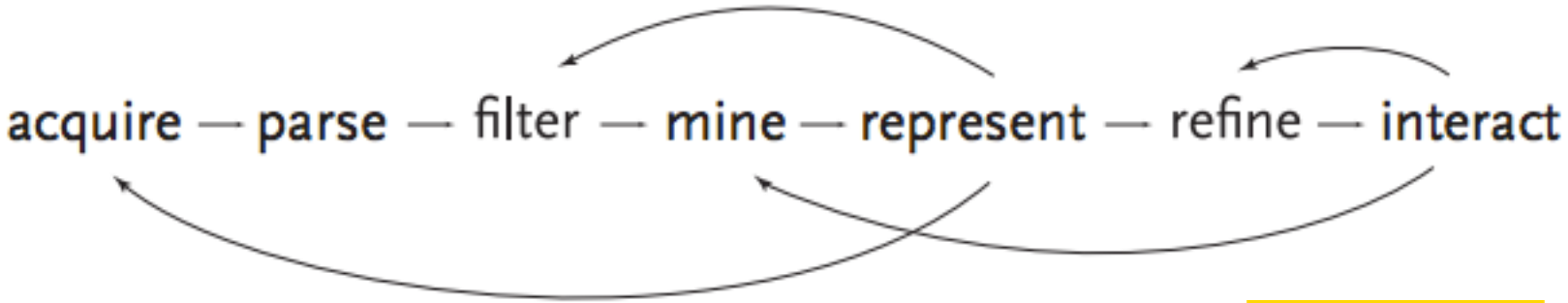


ANSWER

Interpret them and react accordingly.

source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW



source: http://benfry.com/phd/dissertation-110323c.pdf

ANSWER

In re

NOTE

This also relies on your problem solving skills!

III. I-PYTHON NOTEBOOK INTRO

DISCUSSION