

INTRO TO DATA SCIENCE

LECTURE 2: MACHINE LEARNING

RECAP

LAST TIME:

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

QUESTIONS?

BUZZWORD BREAK

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

AGENDA

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

EXERCISES:

III. J-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:

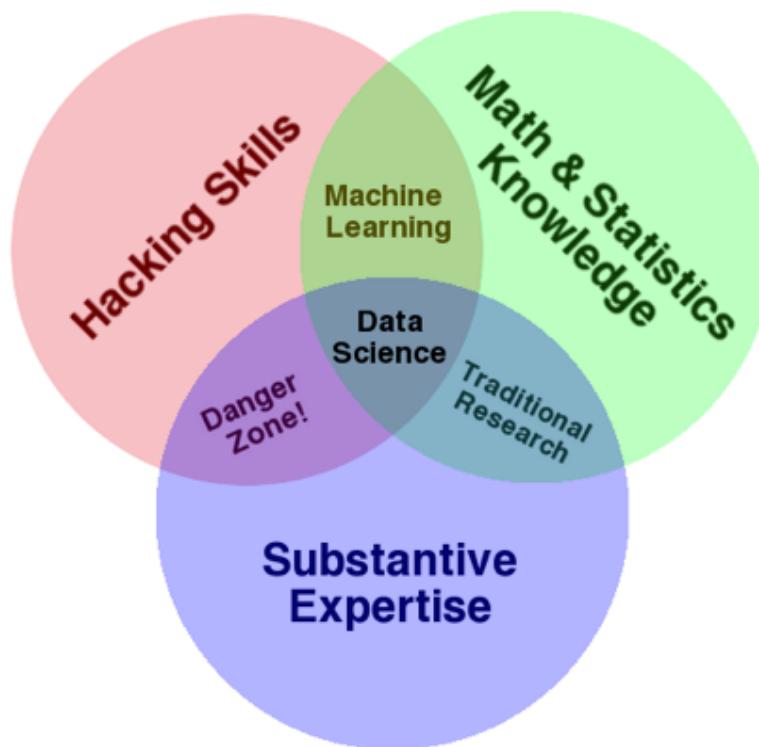
“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

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

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

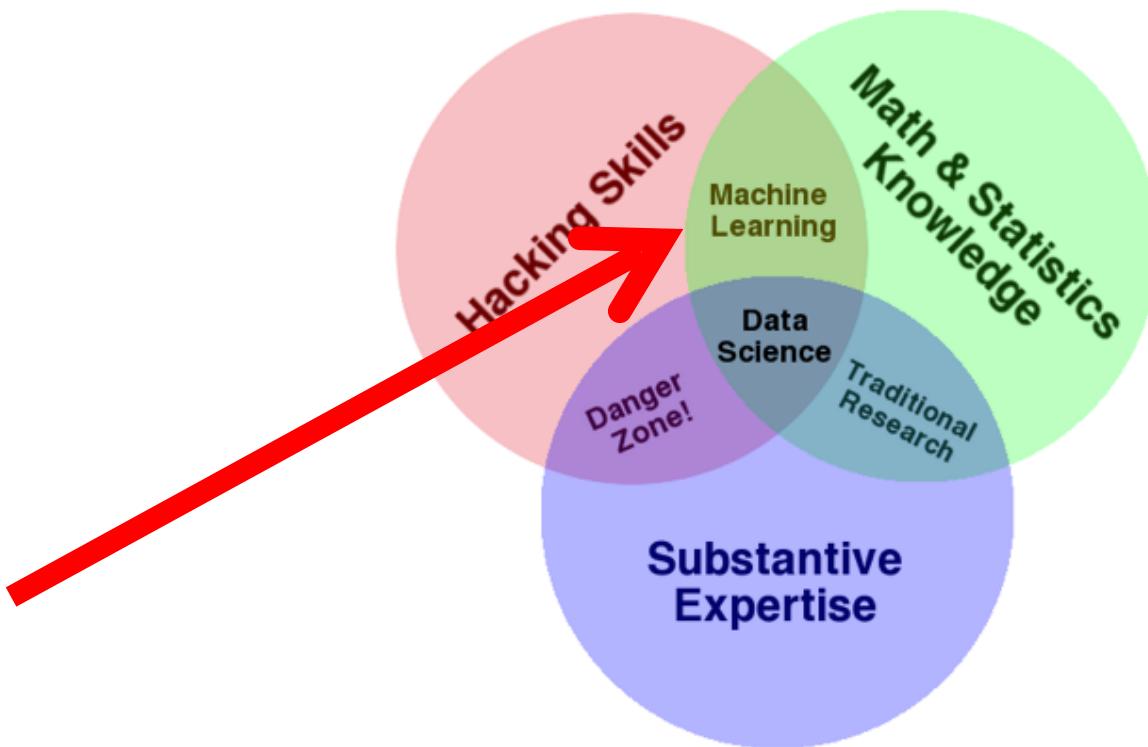
source: http://en.wikipedia.org/wiki/Machine_learning

REMEMBER THIS?

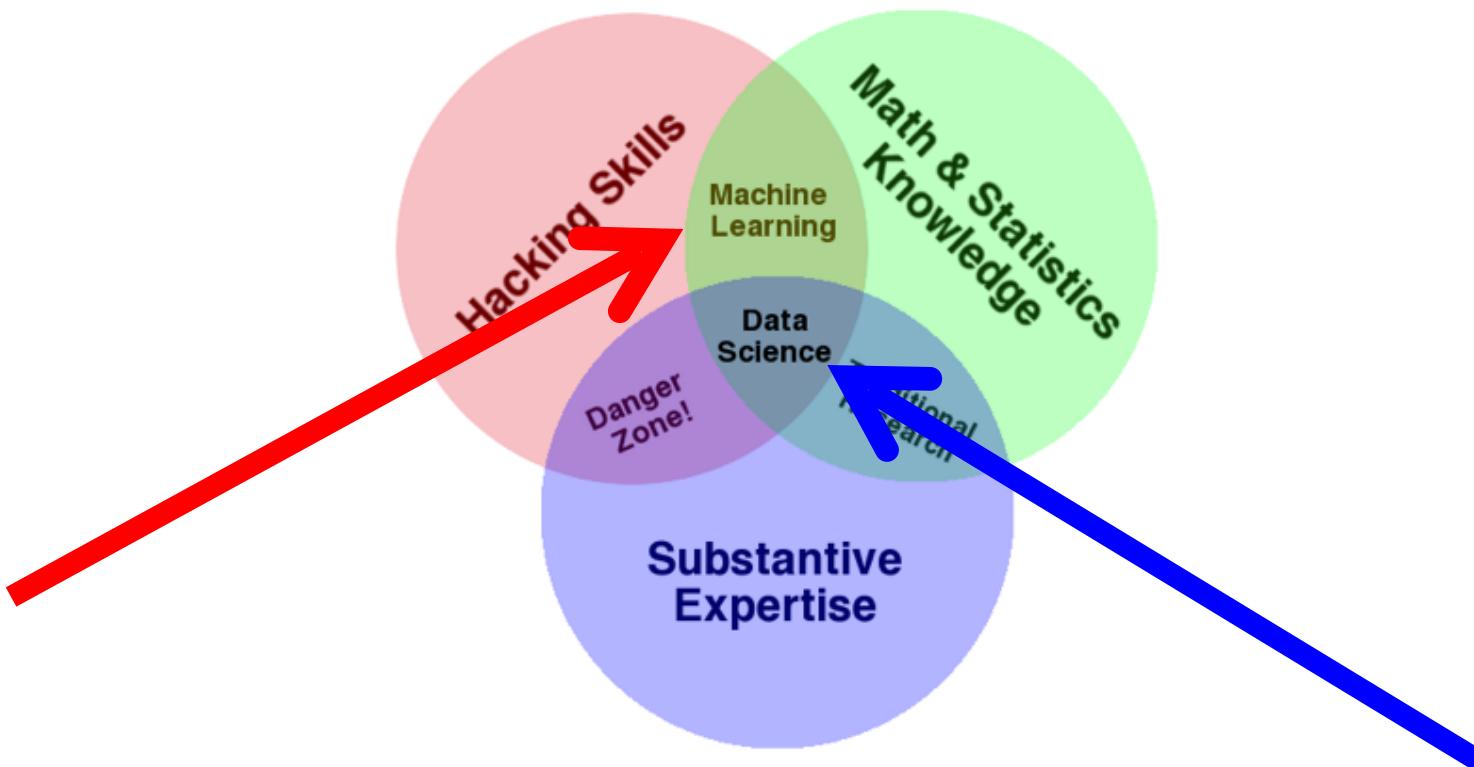


source: <http://www.dataists.com/2010/09/the-data-science-venn-diagram/>

YOU ARE HERE

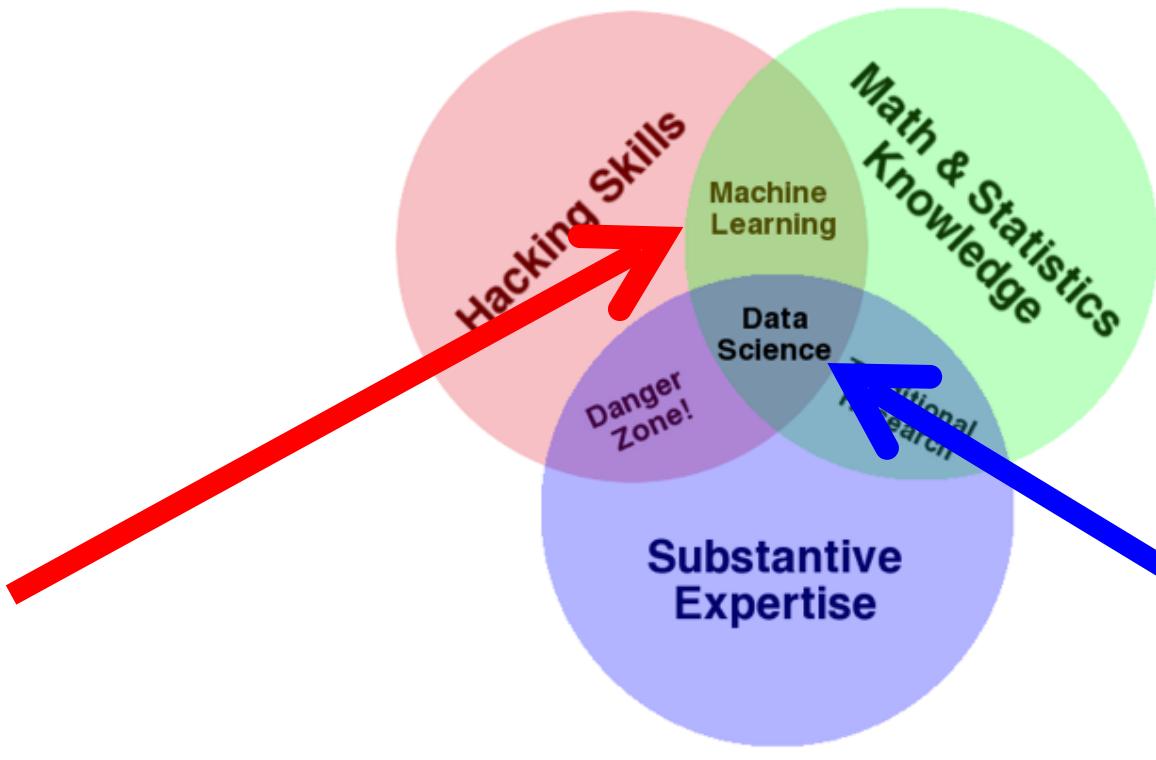


YOU WANT TO GO HERE



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

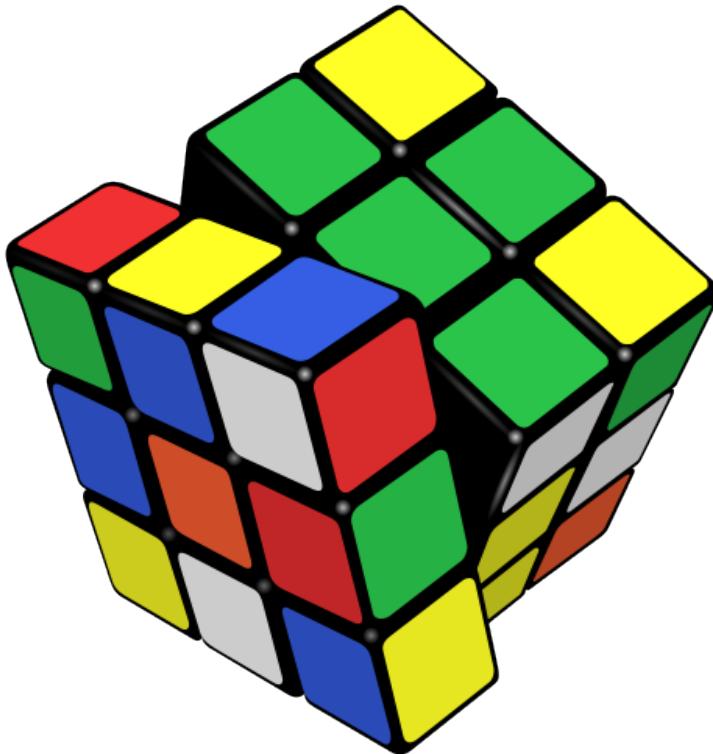
YOU WANT TO GO HERE



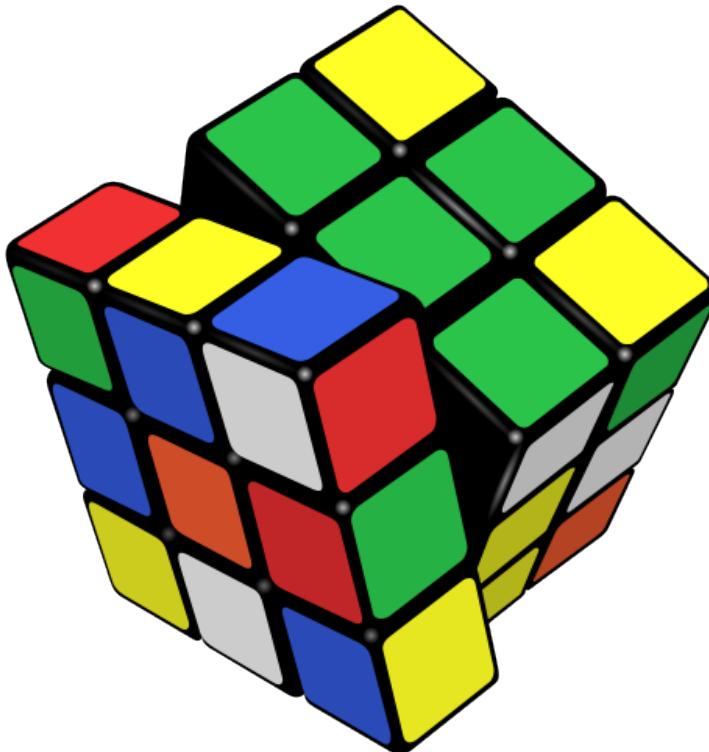
QUESTION

What does it take to make this jump?

ANSWER: PROBLEM SOLVING!



ANSWER: PROBLEM SOLVING!



NOTE

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

II. MACHINE LEARNING PROBLEMS

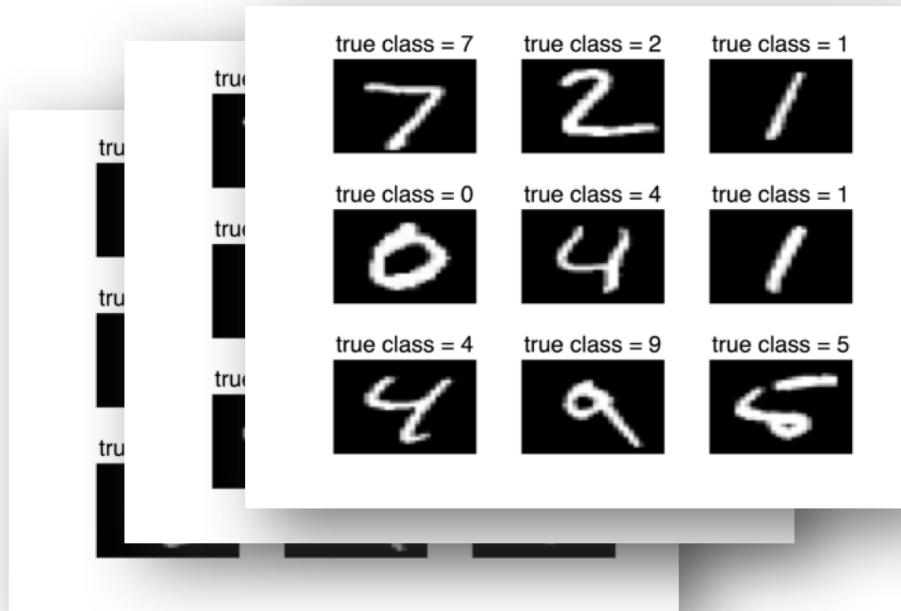
WHAT IS LEARNING?

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



4 1 5 7 1 3 3 6 4 8 1 9 7 6 3 6 9 3 0 6
4 7 7 8 1 3 7 2 4 6 4 3 2 8 0 1 4 3 0 9
1 1 7 6 5 8 6 0 0 3 9 5 4 1 5 7 2 3 2 1
3 5 2 5 2 3 2 9 7 1 6 9 4 6 8 3 2 4 1 9

CREDIT SCORING

[CLICK HERE
TO APPLY TODAY!](#)



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

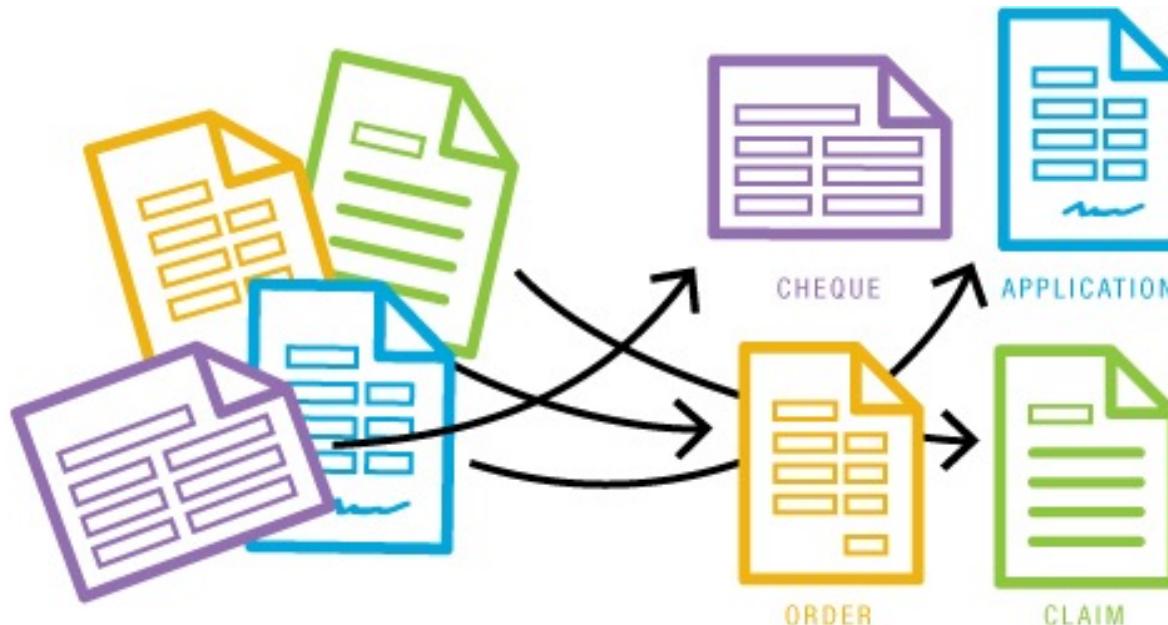
CREDIT SCORING

	<i>Client 1</i>	<i>Client 2</i>	<i>Client 3</i>		<i>Applicant</i>
<i>Age</i>	23	30	19	<i>Age</i>	25
<i>Gender</i>	M	F	M	<i>Gender</i>	M
<i>Annual salary</i>	\$30,000	\$45,000	\$15,000	<i>Annual salary</i>	\$25,000
<i>Years in residence</i>	3 years	1 year	3 month	<i>Years in residence</i>	1 year
<i>Years in job</i>	1 year	1 year	1 month	<i>Years in job</i>	2 years
<i>Current debt</i>	\$5,000	\$1,000	\$10,000	<i>Current debt</i>	\$15,000
<i>Paid off credit</i>	Yes	Yes	No	<i>Credit decision/score</i>	???

EMAIL SPAM DETECTION



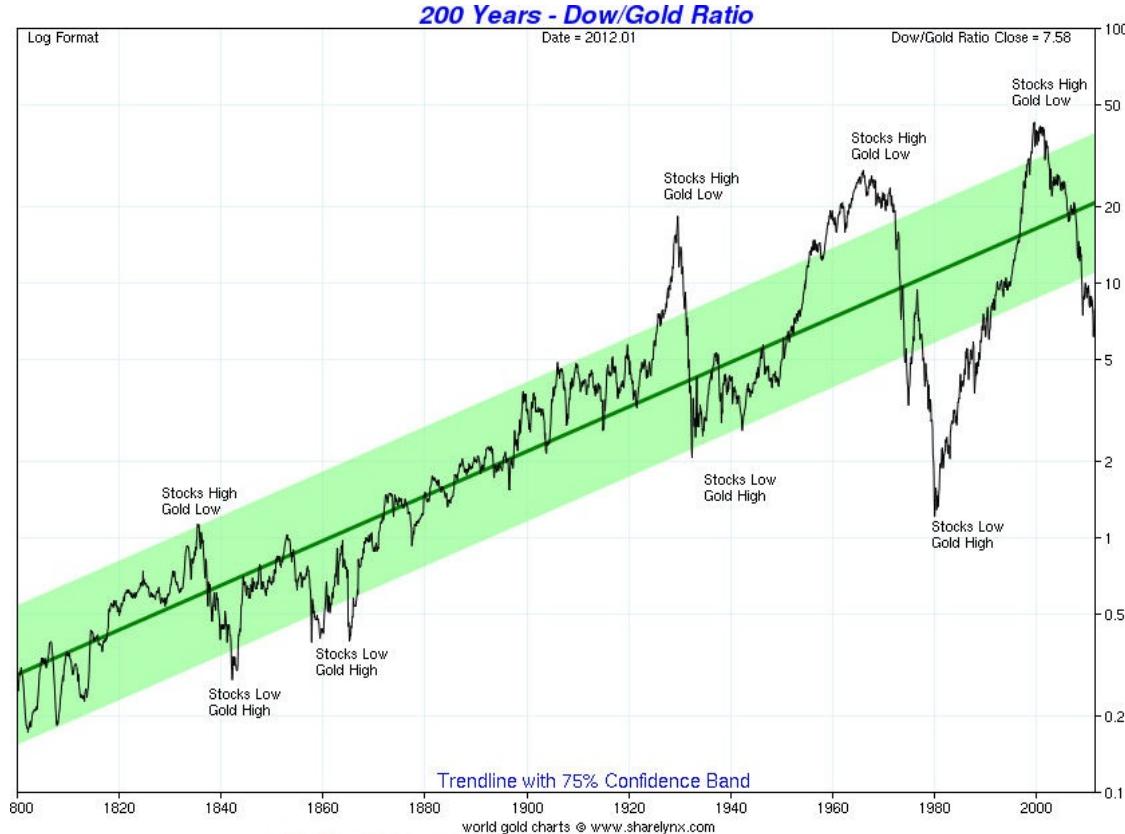
DOCUMENT CLASSIFICATION



REGRESSION - HOUSE PRICE PREDICTION



REGRESSION - STOCK PRICE PREDICTION

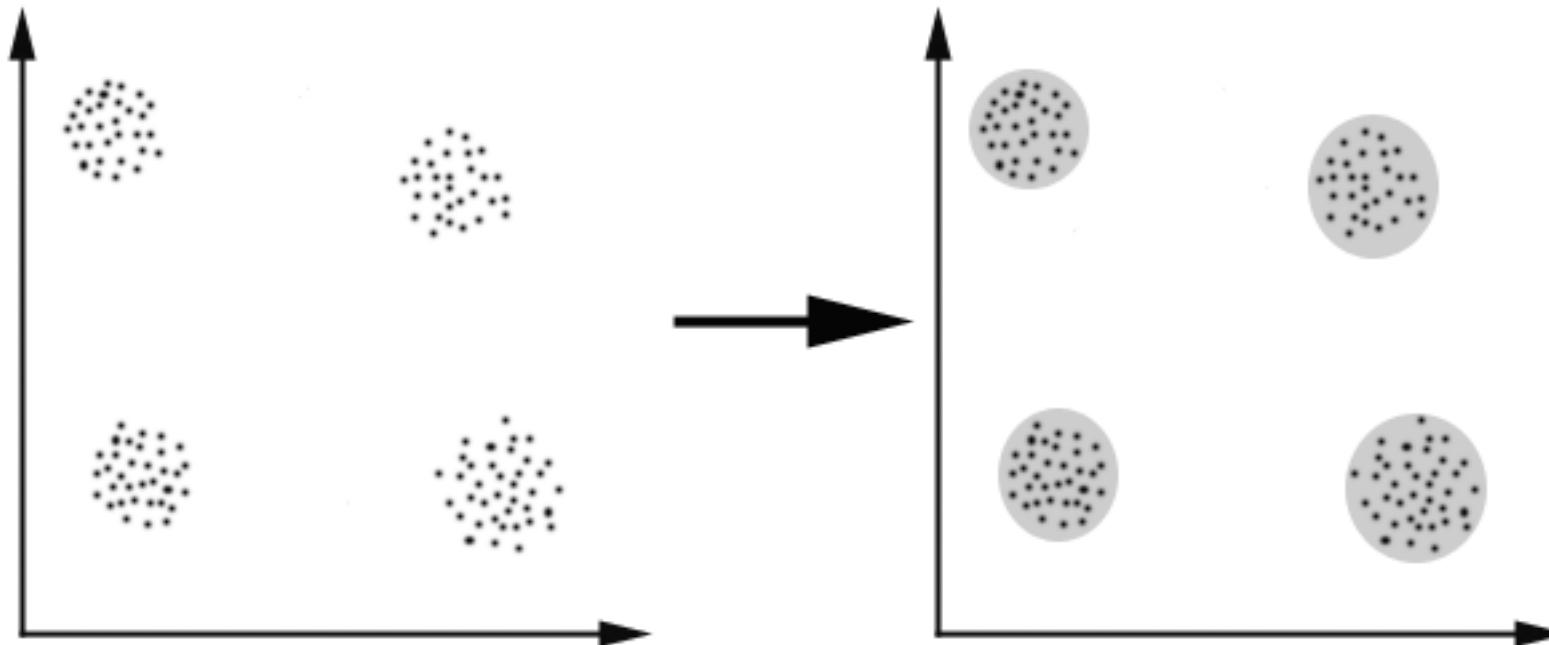


TYPES OF LEARNING?

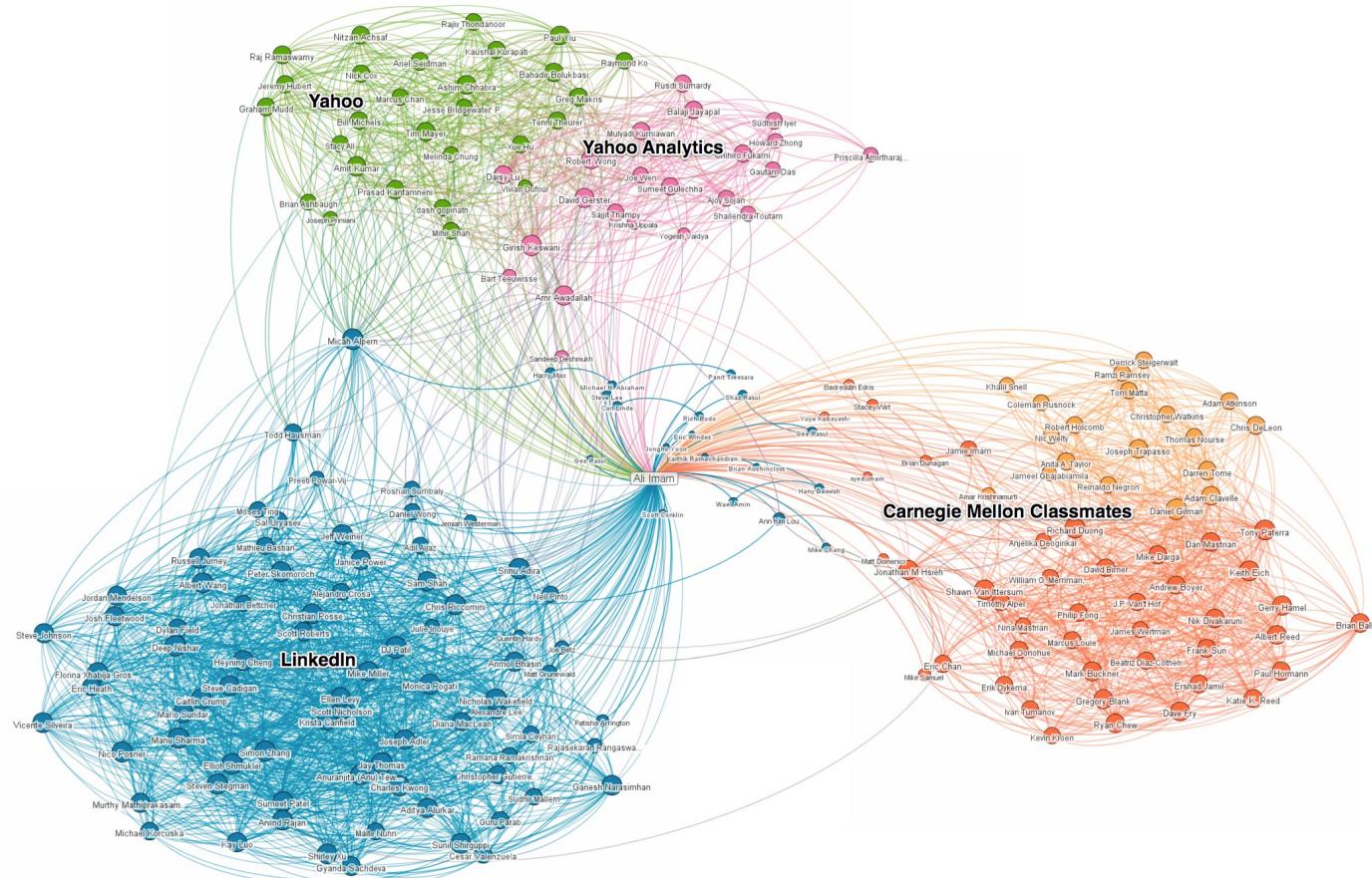
Unsupervised learning: the goal is to learn interesting patterns and **structure** in data given only inputs

no label information given at all

CLUSTERING



COMMUNITY DETECTION IN SOCIAL NETWORKS



TYPES OF LEARNING PROBLEMS

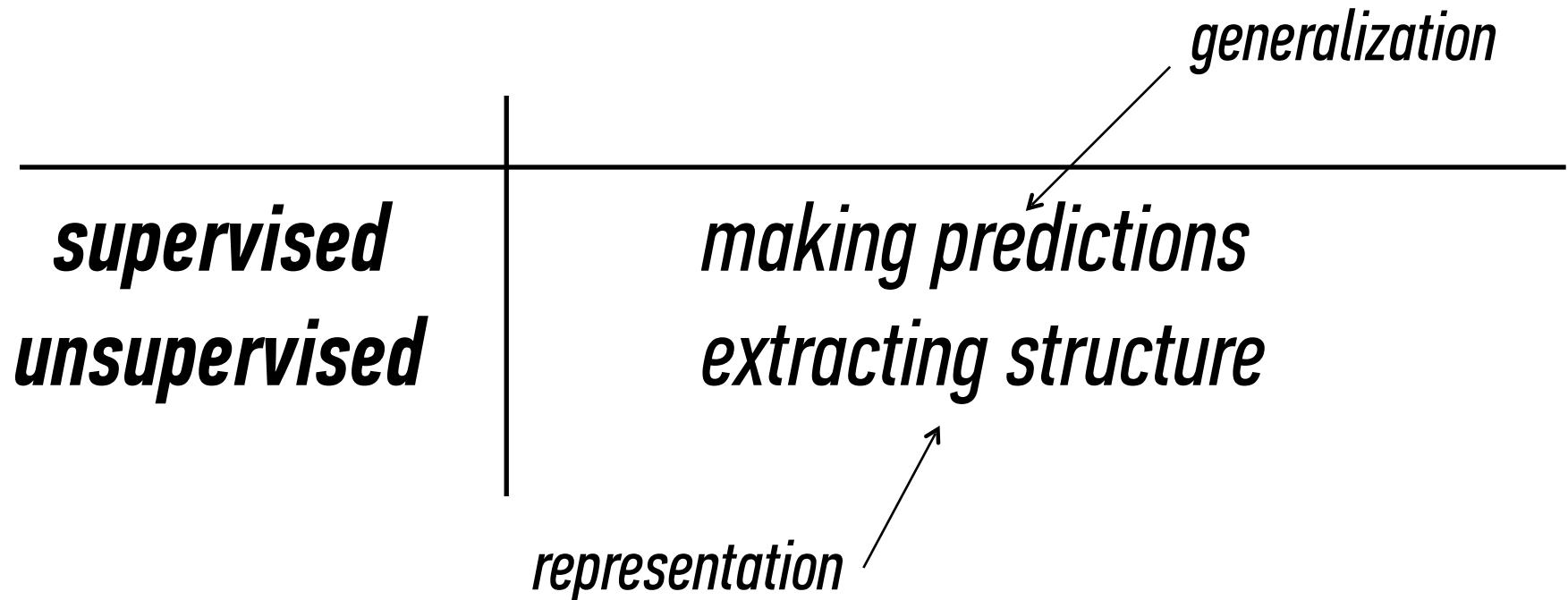
supervised

unsupervised

making predictions

extracting structure

REMEMBER WHAT WE SAID BEFORE?



TYPES OF DATA

continuous

categorical

quantitative

qualitative

TYPES OF DATA***continuous***

Height of children

Weight of cars

Speed of the train

Temperature

Stock price

categorical

Eye colors

Classrooms at GA

Highest degree

Gender

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.

TYPES OF ML SOLUTIONS

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

TYPES OF ML SOLUTIONS

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

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
unsupervised

making predictions
extracting structure

ANSWER

The goal is determined
by the type of problem.

QUESTION

***HOW
DO YOU
DETERMINE
THE RIGHT
APPROACH?***

APPROACHES TO ML PROBLEMS

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

ANSWER

The right approach is determined by the desired solution.

APPROACHES TO ML PROBLEMS

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

ANSWER

NOTE

The

is d

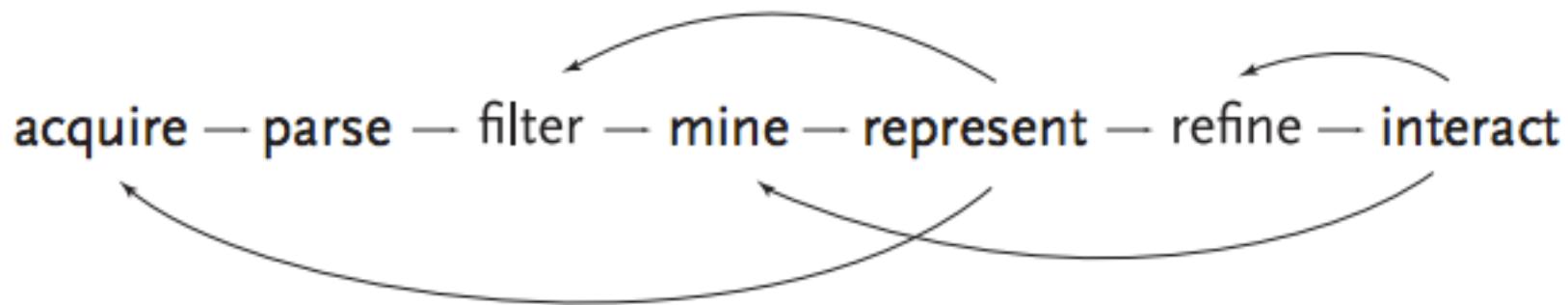
des

All of this depends on
your data!

QUESTION

**WHAT
DO YOU
DO
WITH YOUR
RESULTS?**

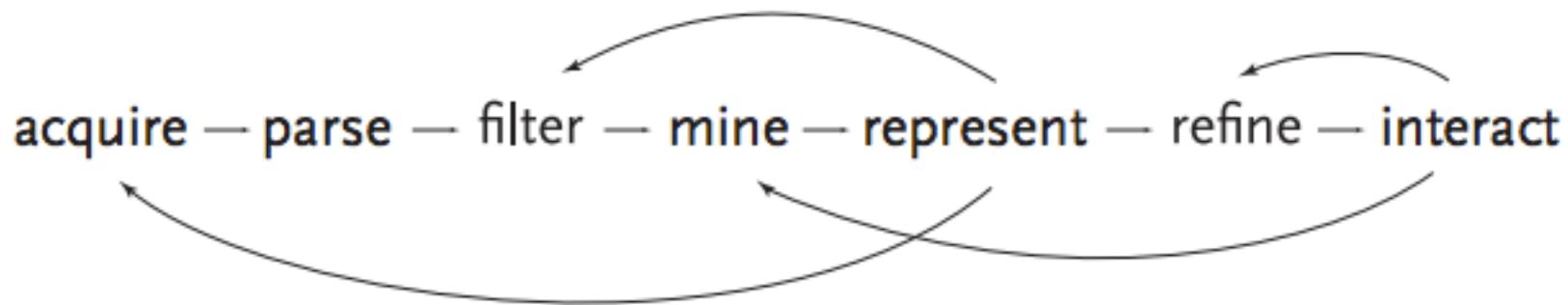
THE DATA SCIENCE WORKFLOW



ANSWER

Interpret them and react accordingly.

THE DATA SCIENCE WORKFLOW



ANSWER

In
re

NOTE
This also relies on
your problem solving
skills!

III. I-PYTHON NOTEBOOK INTRO

INTRO TO DATA SCIENCE

DISCUSSION