GENERAL ASSEMBLY

# DATA SCIENCE

## MACHINE LEARNING

# AGENDA

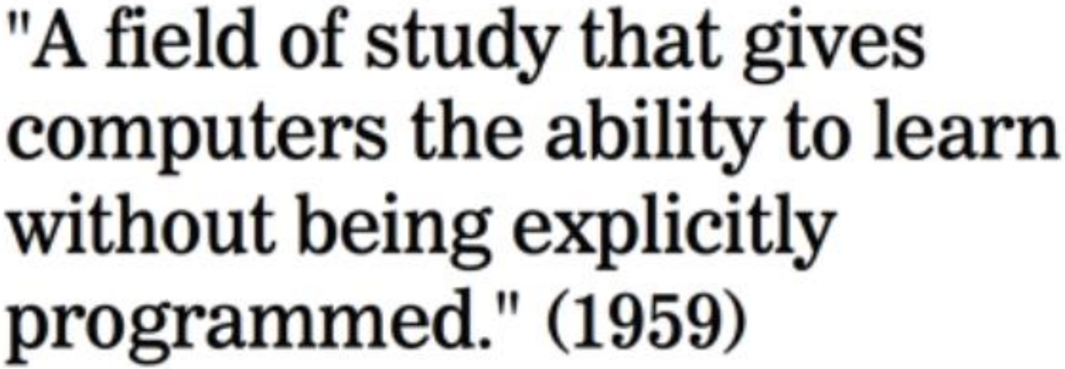
I. WHAT IS MACHINE LEARNING?

II. SUPERVISED LEARNING

III. UNSUPERVISED LEARNING

# I. WHAT IS MACHINE LEARNING?

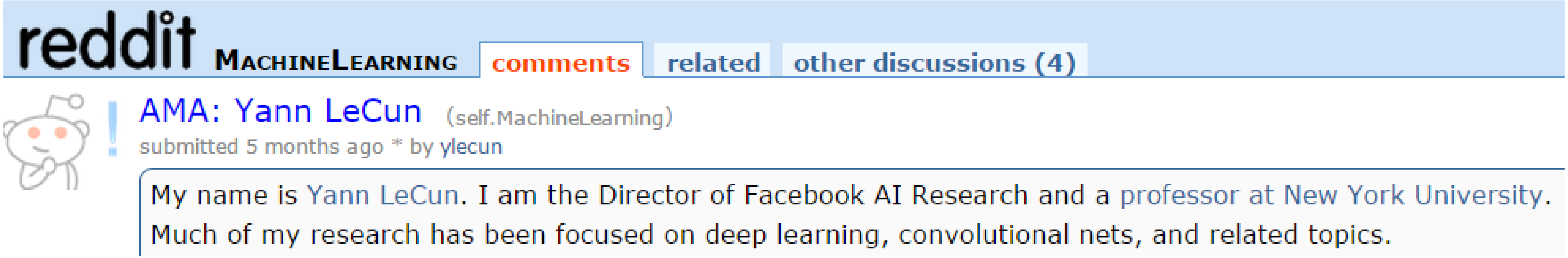
# WHAT IS MACHINE LEARNING?



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

Arthur Samuel, AI pioneer
Source: Stanford

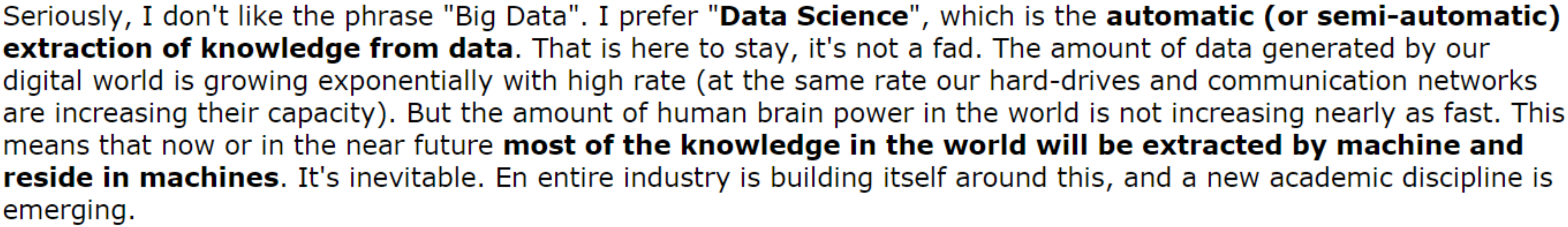
# WHAT IS MACHINE LEARNING?



AMA: Yann LeCun (self.MachineLearning)

submitted 5 months ago * by ylecun

My name is Yann LeCun. I am the Director of Facebook AI Research and a professor at New York University. Much of my research has been focused on deep learning, convolutional nets, and related topics.



Seriously, I don't like the phrase "Big Data". I prefer "**Data Science**", which is the **automatic (or semi-automatic) extraction of knowledge from data**. That is here to stay, it's not a fad. The amount of data generated by our digital world is growing exponentially with high rate (at the same rate our hard-drives and communication networks are increasing their capacity). But the amount of human brain power in the world is not increasing nearly as fast. This means that now or in the near future **most of the knowledge in the world will be extracted by machine and reside in machines**. It's inevitable. En entire industry is building itself around this, and a new academic discipline is emerging.

Source: http://www.reddit.com/r/MachineLearning/comments/25lnbt/ama_yann_lecun

# WHAT IS MACHINE LEARNING?

One definition: “Machine learning is the semi-automatic extraction of knowledge from data.”

- **Knowledge from data:** Starts with a question that might be answerable using data
- **Automatic extraction:** A computer provides the insight
- **Semi-automatic:** Requires many smart decisions by a human

# II. SUPERVISED LEARNING

# TYPES OF MACHINE LEARNING: SUPERVISED LEARNING

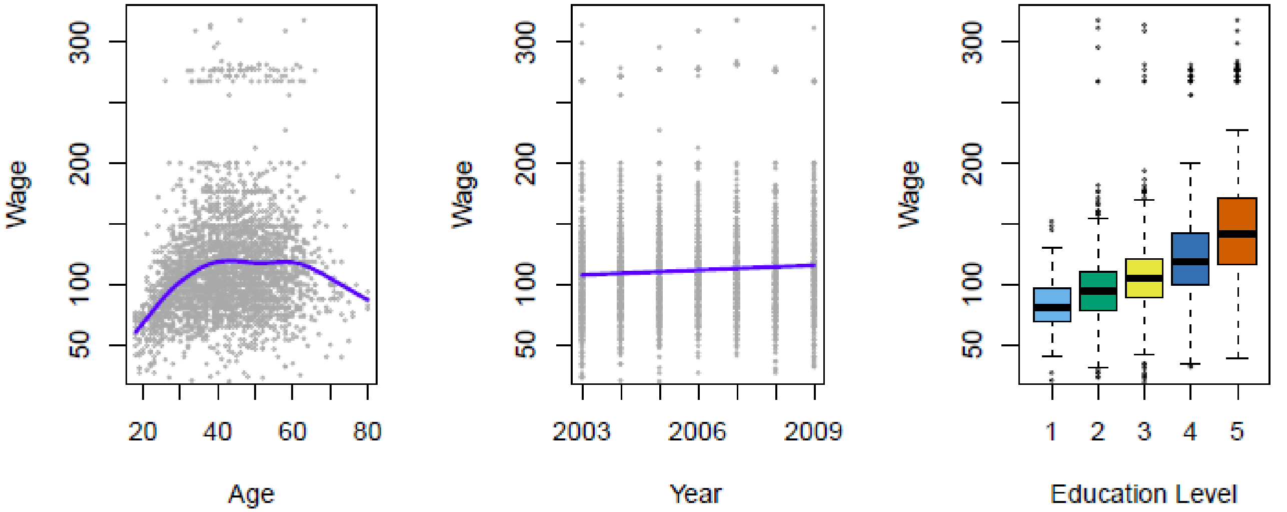
There are two main categories of machine learning: **supervised learning** and **unsupervised learning**.

**Supervised learning** (aka “predictive modeling”):

- Predict an outcome based on input data
- Example: predict whether an email is spam or ham
- Goal is “generalization”

# SUPERVISED LEARNING EXAMPLE

Predict salary using demographic data

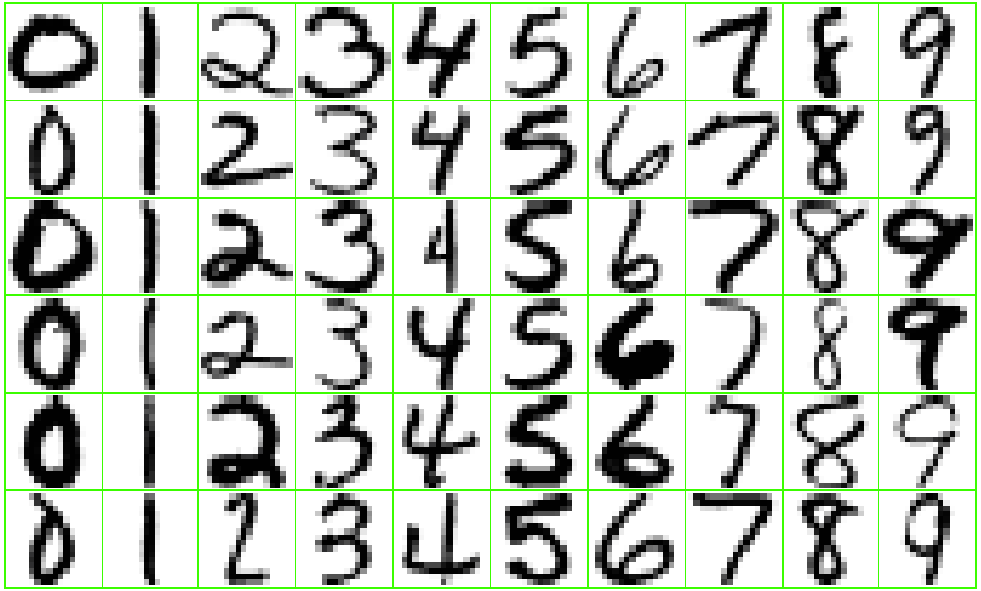


Income survey data for males from the central Atlantic region of the USA in 2009

Source: https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf

# SUPERVISED LEARNING EXAMPLE

Identify the numbers in a handwritten zip code



Source: https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf

# CATEGORIES OF SUPERVISED LEARNING

There are two categories of supervised learning:

**Regression**

- Outcome we are trying to predict is continuous
- Examples: price, blood pressure

**Classification**

- Outcome we are trying to predict is categorical (values in a finite, unordered set)
- Examples: spam/ham, cancer class of tissue sample

# REGRESSION OR CLASSIFICATION?

**Problem:** Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick

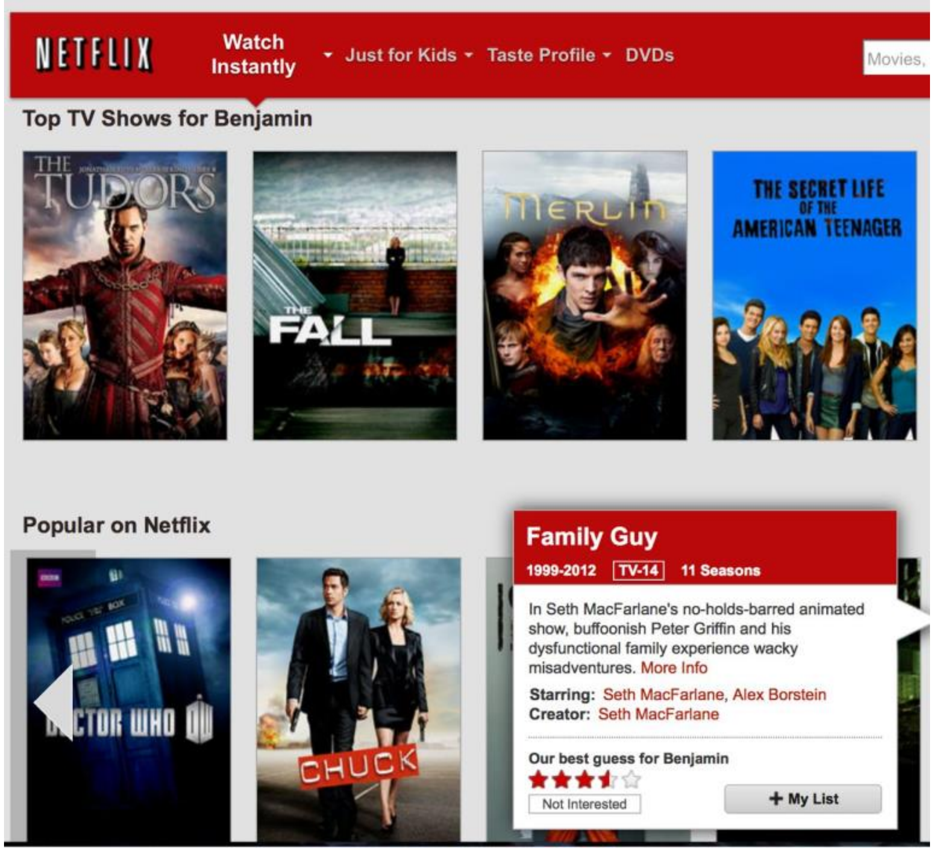
**Goal:** Detect subtle patterns in the data that predicts infection before it occurs

**Data:** 16 vital signs such as heart rate, respiration rate, blood pressure, etc…

**Impact:** Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

**Image**: http://www.babycaretips4u.com/wp-content/uploads/2014/03/premature-baby.jpg
**Case Study**: http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695

# REGRESSION OR CLASSIFICATION?



**Family Guy**

1999-2012 TV-14 11 Seasons

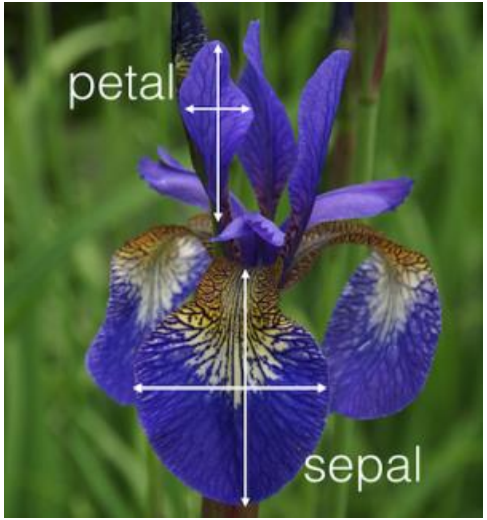
In Seth MacFarlane's no-holds-barred animated show, buffoonish Peter Griffin and his dysfunctional family experience wacky misadventures. More Info

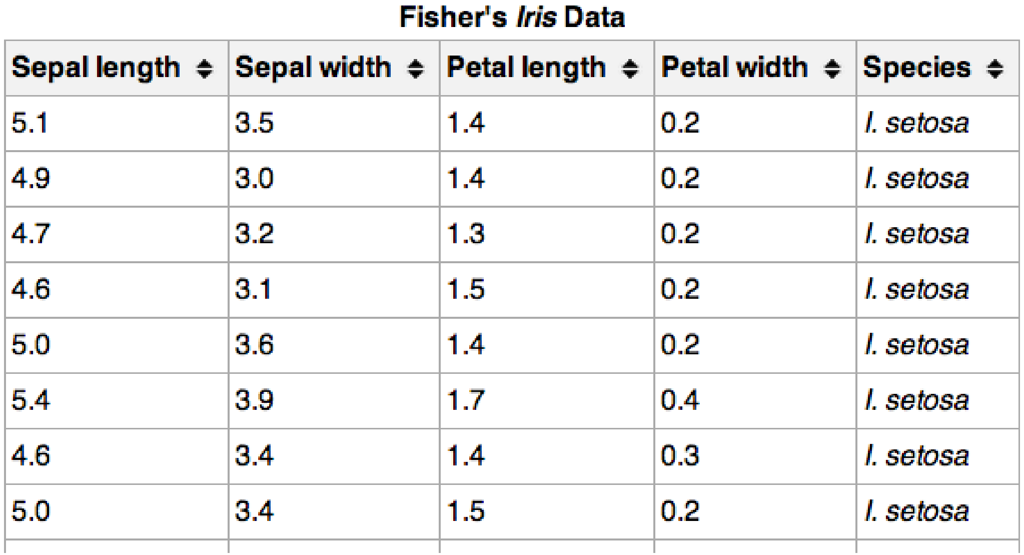
**Starring:** Seth MacFarlane, Alex Borstein
**Creator:** Seth MacFarlane

**Our best guess for Benjamin**

★★★½☆

# REGRESSION OR CLASSIFICATION?

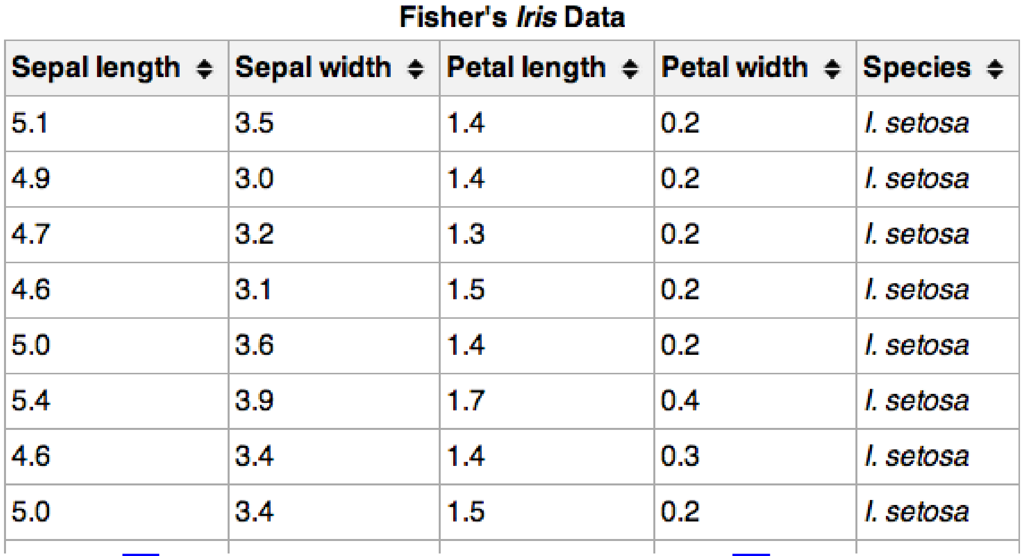




Fisher's *Iris* Data

| Sepal length | Sepal width | Petal length | Petal width | Species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | *I. setosa* |
| 4.9 | 3.0 | 1.4 | 0.2 | *I. setosa* |
| 4.7 | 3.2 | 1.3 | 0.2 | *I. setosa* |
| 4.6 | 3.1 | 1.5 | 0.2 | *I. setosa* |
| 5.0 | 3.6 | 1.4 | 0.2 | *I. setosa* |
| 5.4 | 3.9 | 1.7 | 0.4 | *I. setosa* |
| 4.6 | 3.4 | 1.4 | 0.3 | *I. setosa* |
| 5.0 | 3.4 | 1.5 | 0.2 | *I. setosa* |

# MACHINE LEARNING TERMINOLOGY



Fisher's *Iris* Data

| Sepal length ⬍ | Sepal width ⬍ | Petal length ⬍ | Petal width ⬍ | Species ⬍ |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | *I. setosa* |
| 4.9 | 3.0 | 1.4 | 0.2 | *I. setosa* |
| 4.7 | 3.2 | 1.3 | 0.2 | *I. setosa* |
| 4.6 | 3.1 | 1.5 | 0.2 | *I. setosa* |
| 5.0 | 3.6 | 1.4 | 0.2 | *I. setosa* |
| 5.4 | 3.9 | 1.7 | 0.4 | *I. setosa* |
| 4.6 | 3.4 | 1.4 | 0.3 | *I. setosa* |
| 5.0 | 3.4 | 1.5 | 0.2 | *I. setosa* |

*150 observations*

*(n = 150)*

*4 features (p = 4)*

*response*

Feature matrix “X” has n rows and p columns

Response “y” is a vector with length n

# MACHINE LEARNING TERMINOLOGY

**Observations** are also known as: samples, examples, instances, records

**Features** are also known as: predictors, independent variables, inputs, regressors, covariates, attributes

**Response** is also known as: outcome, label, target, dependent variable

**Regression problems** have a continuous response. **Classification problems** have a categorical response. The type of supervised learning problem has nothing to do with the features!

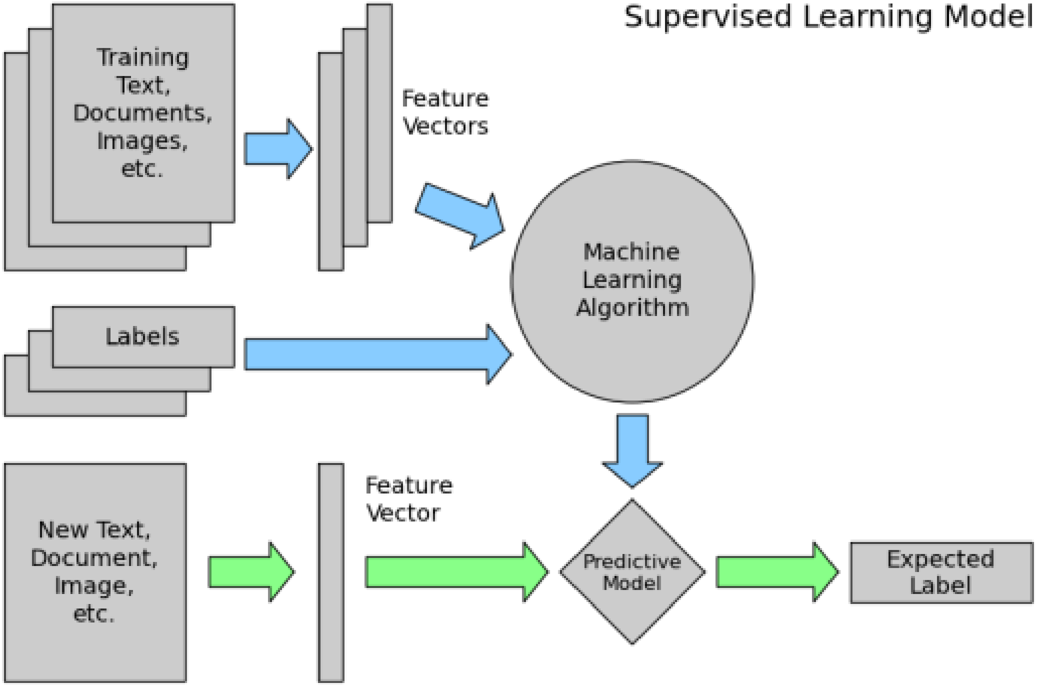
**How does supervised learning “work”?**

1. Train a **machine learning model** using **labeled data**
    - “Labeled data” is data with a response variable
    - “Machine learning model” learns the relationship between the features and the response

2. Make predictions on **new data** for which the response is unknown

The primary goal of supervised learning is to build a model that “generalizes”: It accurately predicts the **future** rather than the **past**!

## How does supervised learning “work”?


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

# SUPERVISED LEARNING EXAMPLE

**Supervised learning example: Dog detector**

- Input data: Images from Google
- Features: Numerical representations of the images
- Response: Dog (yes or no), hand-labeled

1. Train a **machine learning model** using **labeled data**
    - Model learns the relationship between the image data and the "dog status"

2. Make predictions on **new data** for which the response is unknown
    - Give it a new image, predicts the "dog status" automatically

# III. UNSUPERVISED LEARNING

# TYPES OF MACHINE LEARNING: UNSUPERVISED LEARNING

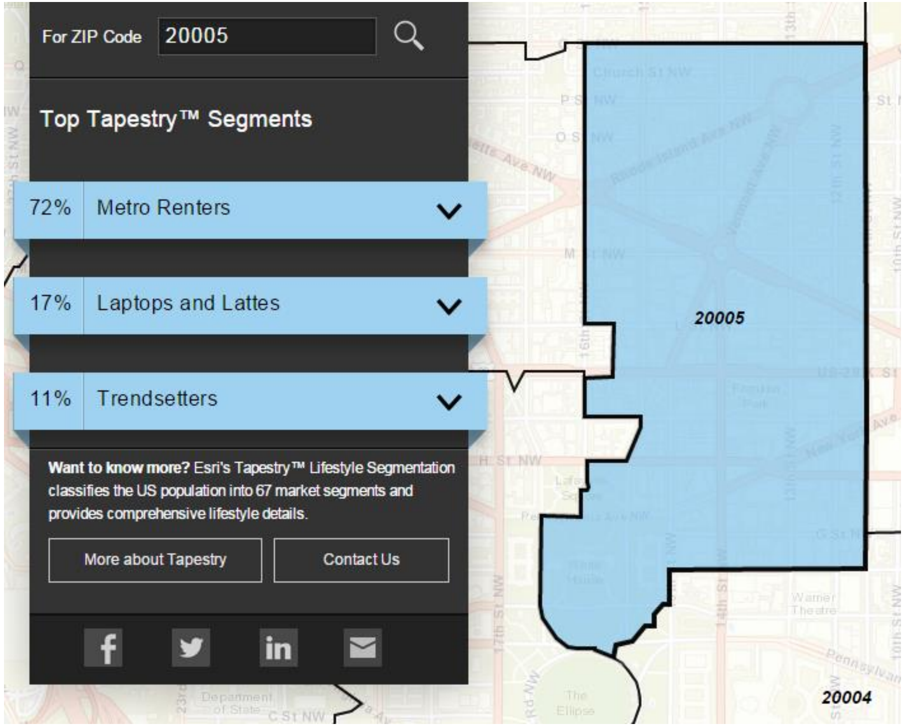
There are two main categories of machine learning: **supervised learning** and **unsupervised learning**.

**Unsupervised learning:**

- Extracting structure from data
- Example: segment grocery store shoppers into “clusters” that exhibit similar behaviors
- Goal is “representation”

# UNSUPERVISED LEARNING EXAMPLE

Classify US residential neighborhoods into 67 unique segments based on demographic and socioeconomic characteristics



For ZIP Code 20005

Top Tapestry™ Segments

- 72% Metro Renters
- 17% Laptops and Lattes
- 11% Trendsetters

**Want to know more?** Esri's Tapestry™ Lifestyle Segmentation classifies the US population into 67 market segments and provides comprehensive lifestyle details.

Metro Renters:

Young, mobile, educated, or still in school, we live alone or with a roommate in rented apartments or condos in the center of the city. Long hours and hard work don't deter us; we're willing to take risks to get to the top of our professions… We buy groceries at Whole Foods and Trader Joe's and shop for clothes at Banana Republic, Nordstrom, and Gap. We practice yoga, go skiing, and attend Pilates sessions.

Source: http://www.esri.com/landing-pages/tapestry/

# UNSUPERVISED LEARNING

Unsupervised learning has some clear differences from supervised learning. With **unsupervised learning:**

- There is no clear objective
- There is no “right anwser” (hard to tell how well you are doing)
- There is no response variable, just observations with features
- Labeled data is not required

## Unsupervised learning example: Image clustering

- Input data: Images from Google
- Features: Numerical representations of the images
- Response: There isn't one (no hand-labeling required!)

1. Perform **unsupervised learning**
    - Cluster the images based on "similarity"
    - Might find a "dog cluster", might not
    - You're done!

Sometimes, unsupervised learning is used as a "preprocessing" step for supervised learning.