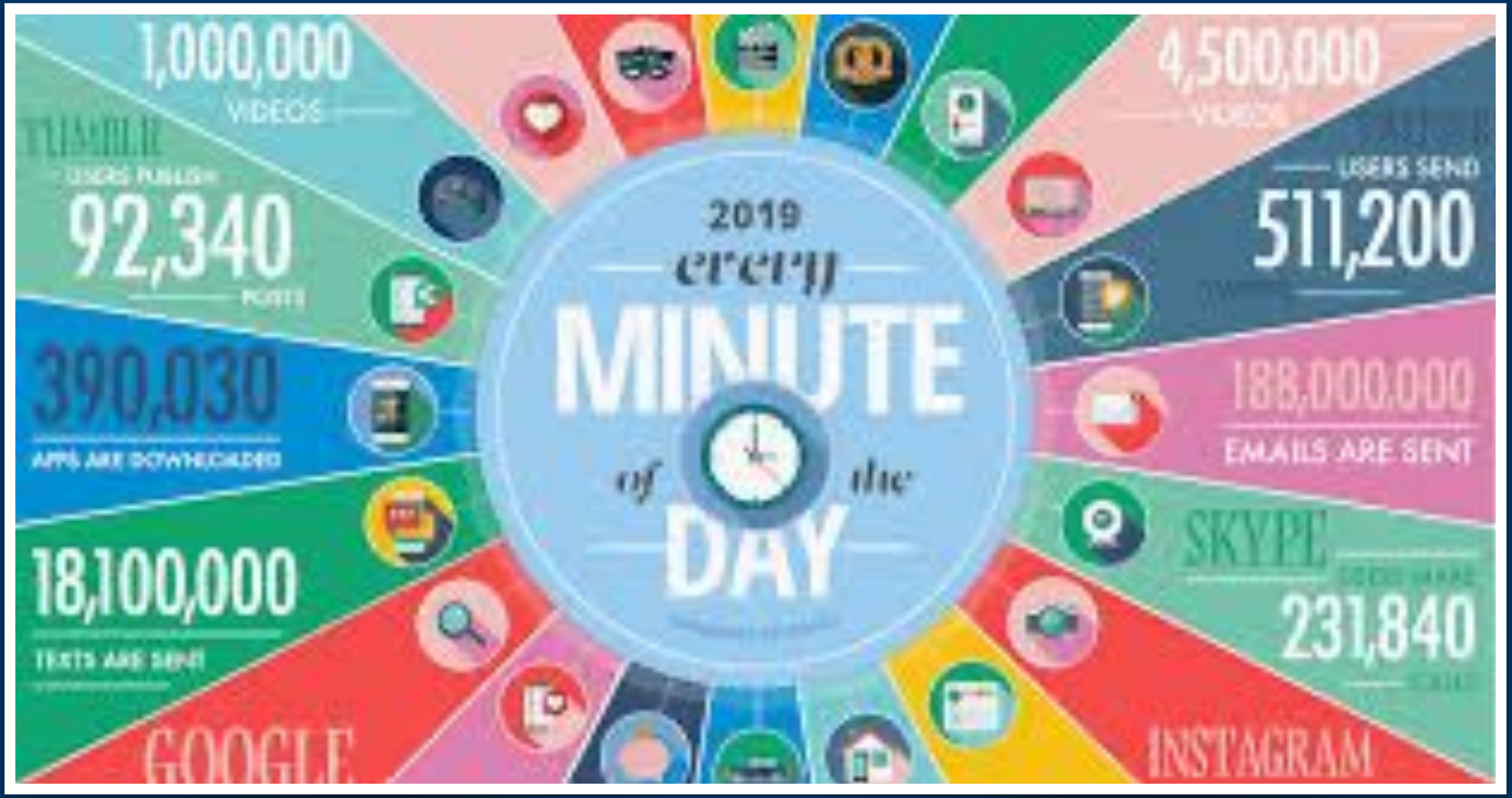


Data: Things to Consider

CSCI-P556 Applied Machine Learning
Lecture 3

D.S. Williamson

Agenda and Learning Outcomes

Today's Topics

- **Topics**:
  - Data and Why is it needed?
  - What challenges do data present?
  - Where to get data? Gather or Collect?
- **Learning Outcomes**: At the end of today's class, students should:
  - Describe the bad data problem
  - Understand how to define needed data for an experiment

TopHat Questions

What is Data?

Anything can be considered as data

- Anything that can be observed and digitized, can function as data for Machine Learning

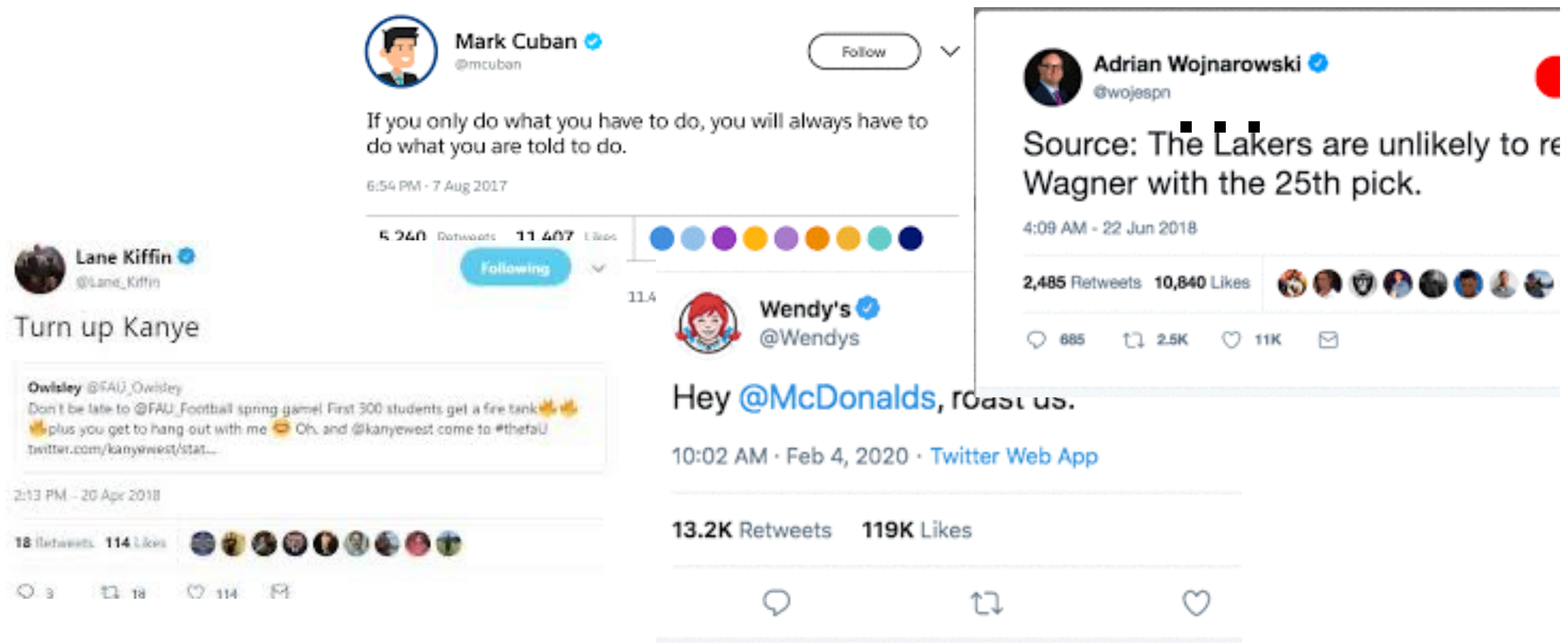
- Depending on the learning algorithm (supervised or unsupervised), additional meta data may be needed (e.g. attributes/features/inputs and labels)

Health Records

Images

Check on
Medical Inform
Names, a
ent I
Electroencephalogram (EEG)
Normal brain activity
Sharp waves show seizure activity
Seizure onset zone (hot spot)
Brain scan (SISCOM)

Tweets



Mark Cuban @mcuban

If you only do what you have to do, you will always have to do what you are told to do.

6:54 PM - 7 Aug 2017

5,240 Retweets 11,407 Likes

Lane Kiffin @Lane_Kiffin

Turn up Kanye

Owlsley @FAU_Owlsley
Don't be late to @FAU_Football spring game! First 300 students get a fire tank🔥🔥 🔥plus you get to hang out with me 😊 Oh, and @kanyewest come to #thefaU
twitter.com/kanyewest/stat...

2:13 PM - 20 Apr 2018

18 Retweets 114 Likes

3 18 114

Wendy's @Wendys

Hey @McDonalds, roast us.

10:02 AM · Feb 4, 2020 · Twitter Web App

13.2K Retweets 119K Likes

Adrian Wojnarowski @wojespn

Source: The Lakers are unlikely to re
Wagner with the 25th pick.

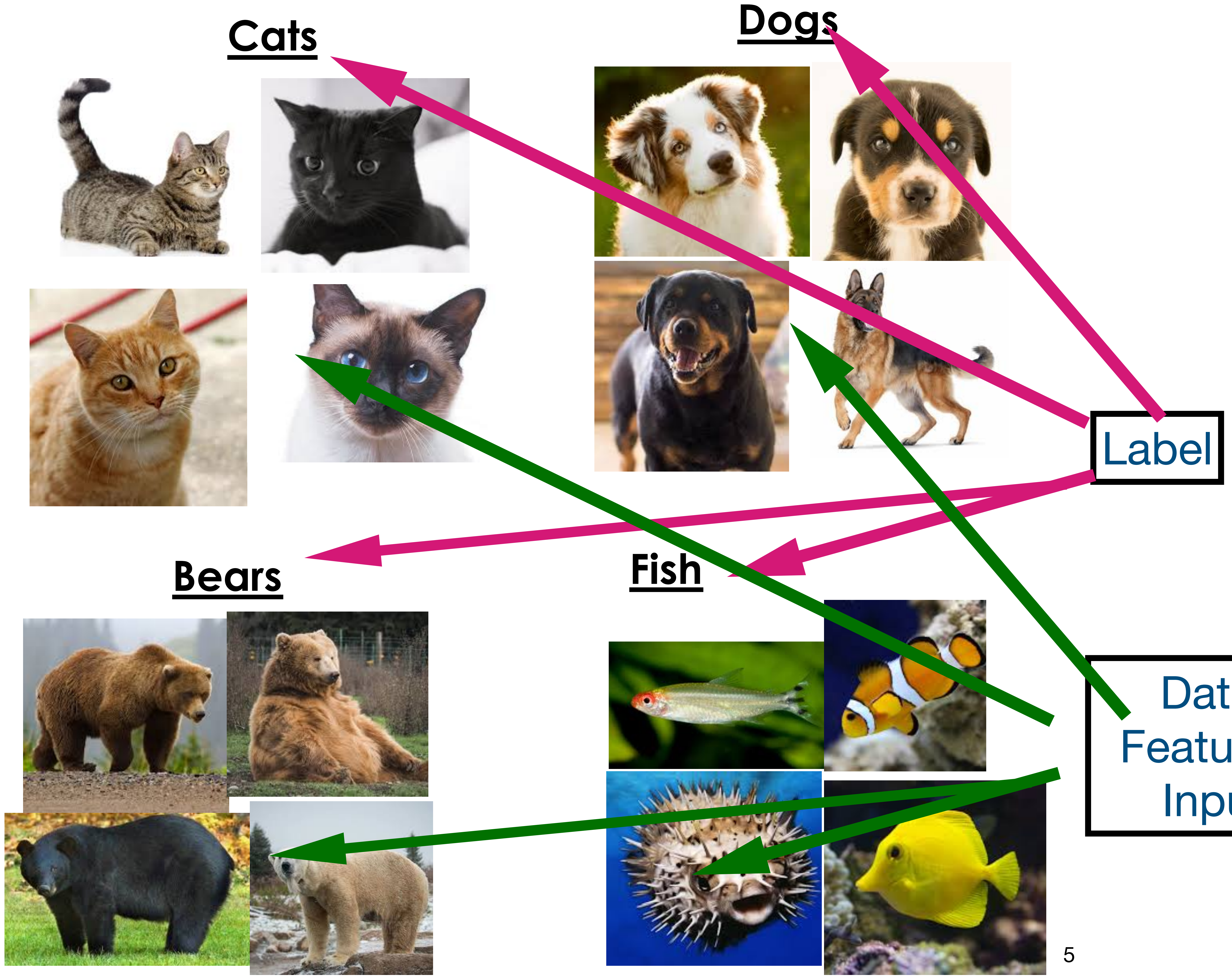
4:09 AM - 22 Jun 2018

2,485 Retweets 10,840 Likes

685 2.5K 11K

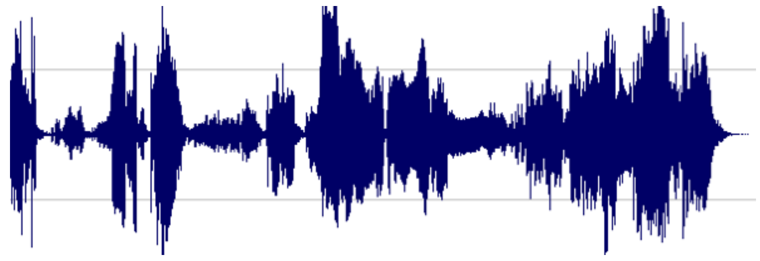
Recall: Attributes and Labels of Data

For Supervised Learning



Data, Features, Input

Data	Label
	Cat
	Food

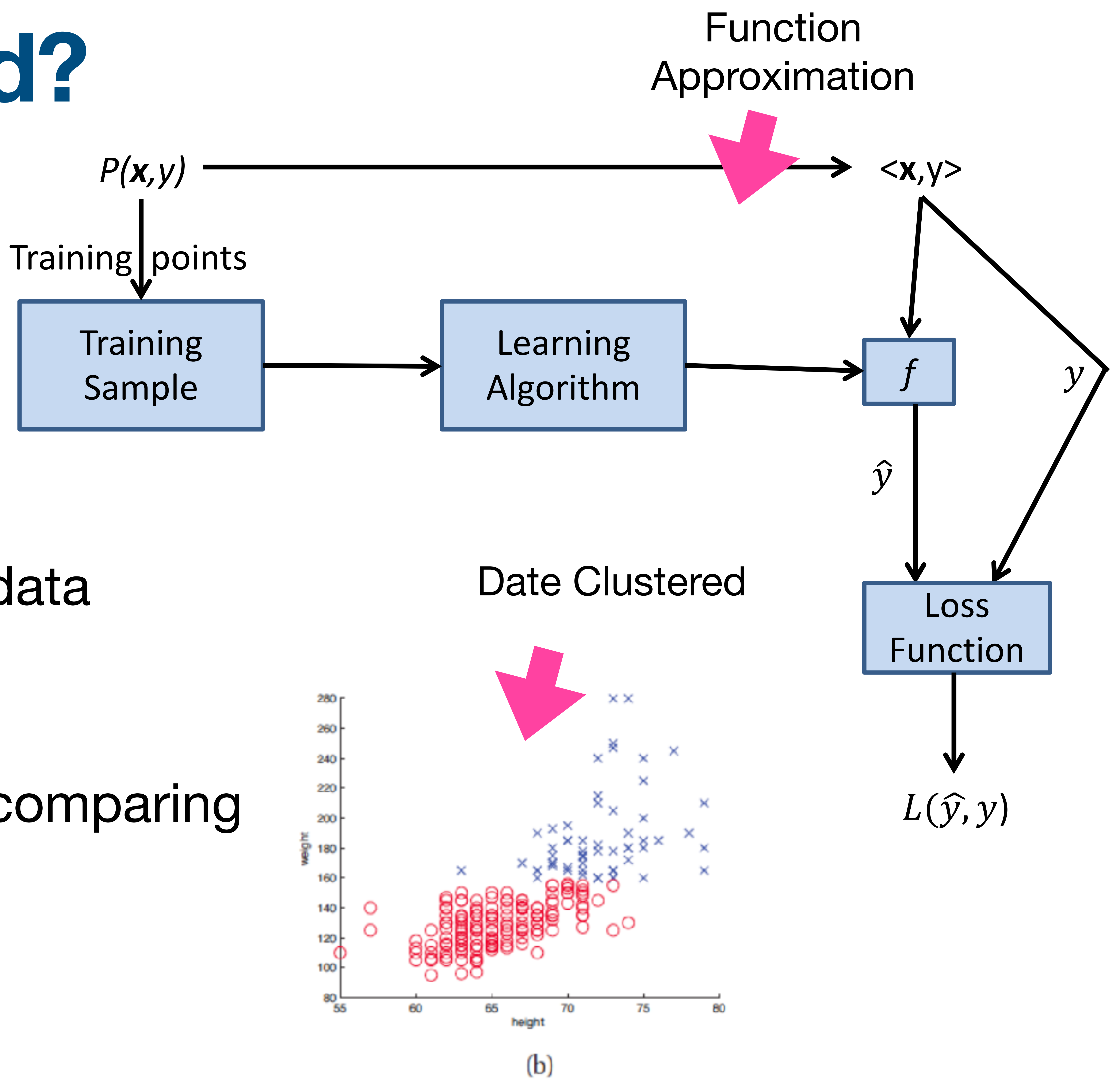
Data: 

Label: “Will you be …”

Why is Data Needed?

No data no learning!

- Machine Learning “learns” from data
    - Either a function for prediction
    - Or patterns for clustering and comparing



Function Approximation

$P(\mathbf{x},y)$ → $<\mathbf{x},y>$

Training points

Training Sample → Learning Algorithm → f

y

$\hat{y}$

Loss Function

$L(\hat{y},y)$

Date Clustered

(b)

Main Challenges of Machine Learning

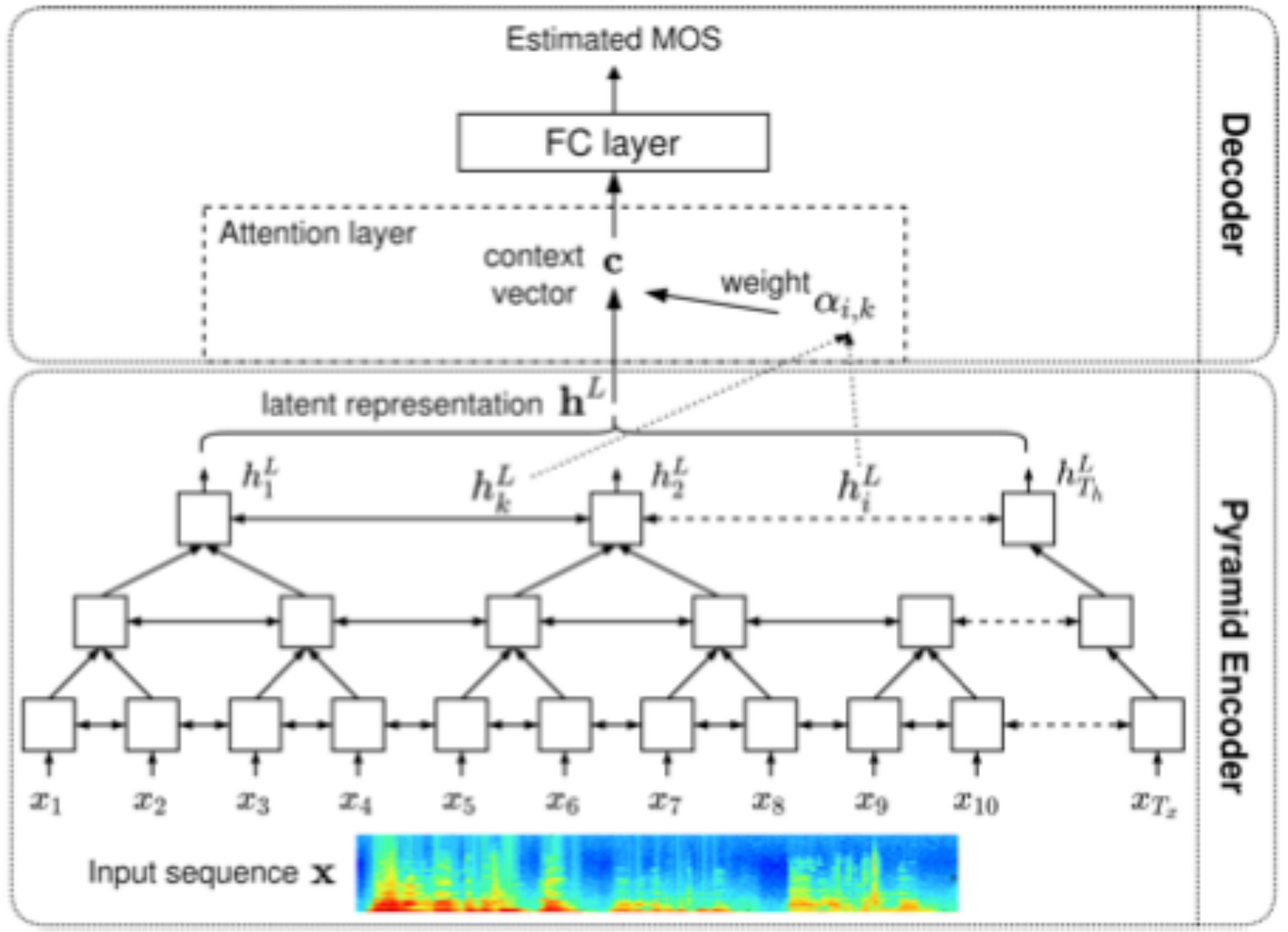
Nothing is perfect!

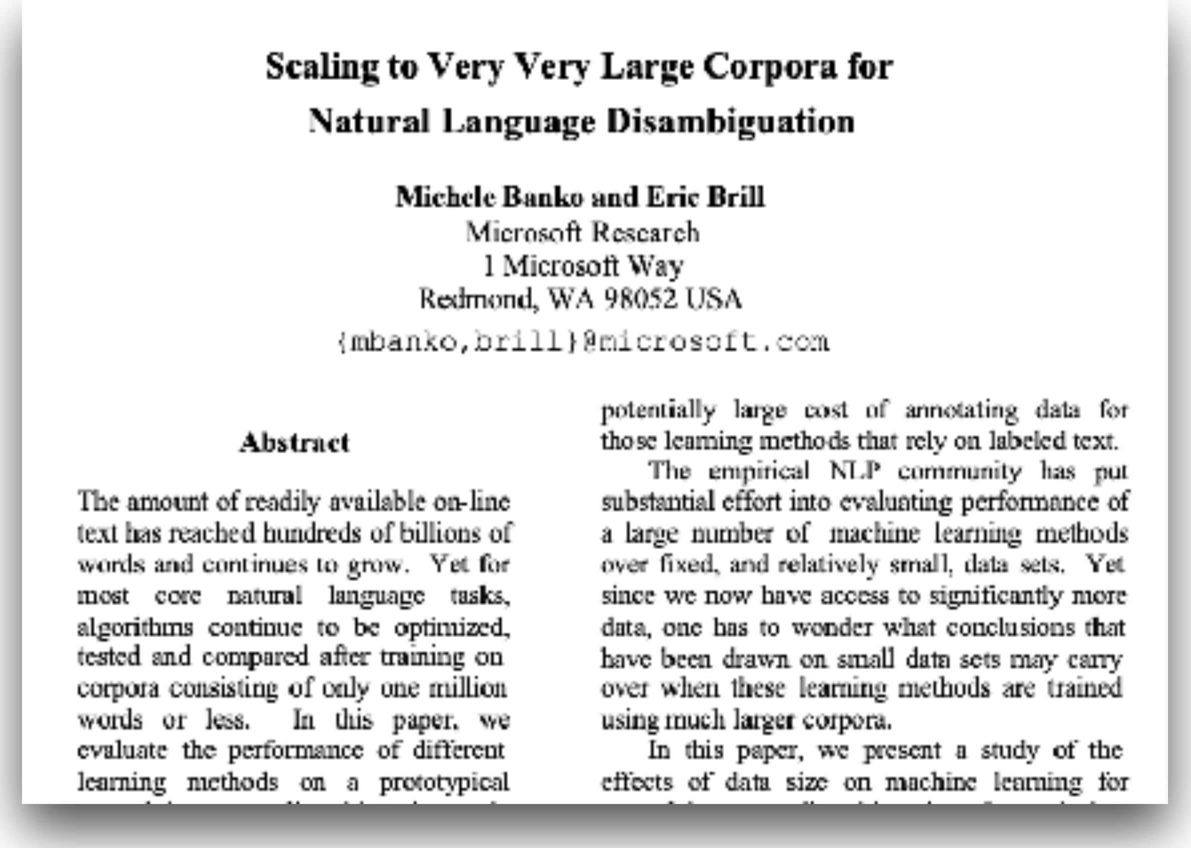
- Our main task, as machine learning scientists, is to select a learning algorithm and train it on some data
- Three potential challenges:
  - **Bad algorithm**: The algorithm is not good enough for this particular problem
  - **Bad data**: The data is not good enough
  - **Bad evaluation**: Metrics aren’t indicative of future usage
- Today, we’ll talk about the ***Bad data problem***

The Unreasonable Effectiveness of Data

Focus at School vs. Work

- In school, algorithms are king!
- In work, data often matters most!





Scaling to Very Very Large Corpora for Natural Language Disambiguation

Michele Banko and Eric Brill
Microsoft Research
1 Microsoft Way
Redmond, WA 98052 USA
{mbanko,brill}@microsoft.com

Abstract

The amount of readily available on-line text has reached hundreds of billions of words and continues to grow. Yet for most core natural language tasks, algorithms continue to be optimized, tested and compared after training on corpora consisting of only one million words or less. In this paper, we evaluate the performance of different learning methods on a prototypical

potentially large cost of annotating data for those learning methods that rely on labeled text.

The empirical NLP community has put substantial effort into evaluating performance of a large number of machine learning methods over fixed, and relatively small, data sets. Yet since we now have access to significantly more data, one has to wonder what conclusions that have been drawn on small data sets may carry over when these learning methods are trained using much larger corpora.

In this paper, we present a study of the effects of data size on machine learning for

EXPERT OPINION

Contact Editor: Brian Brannon, bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

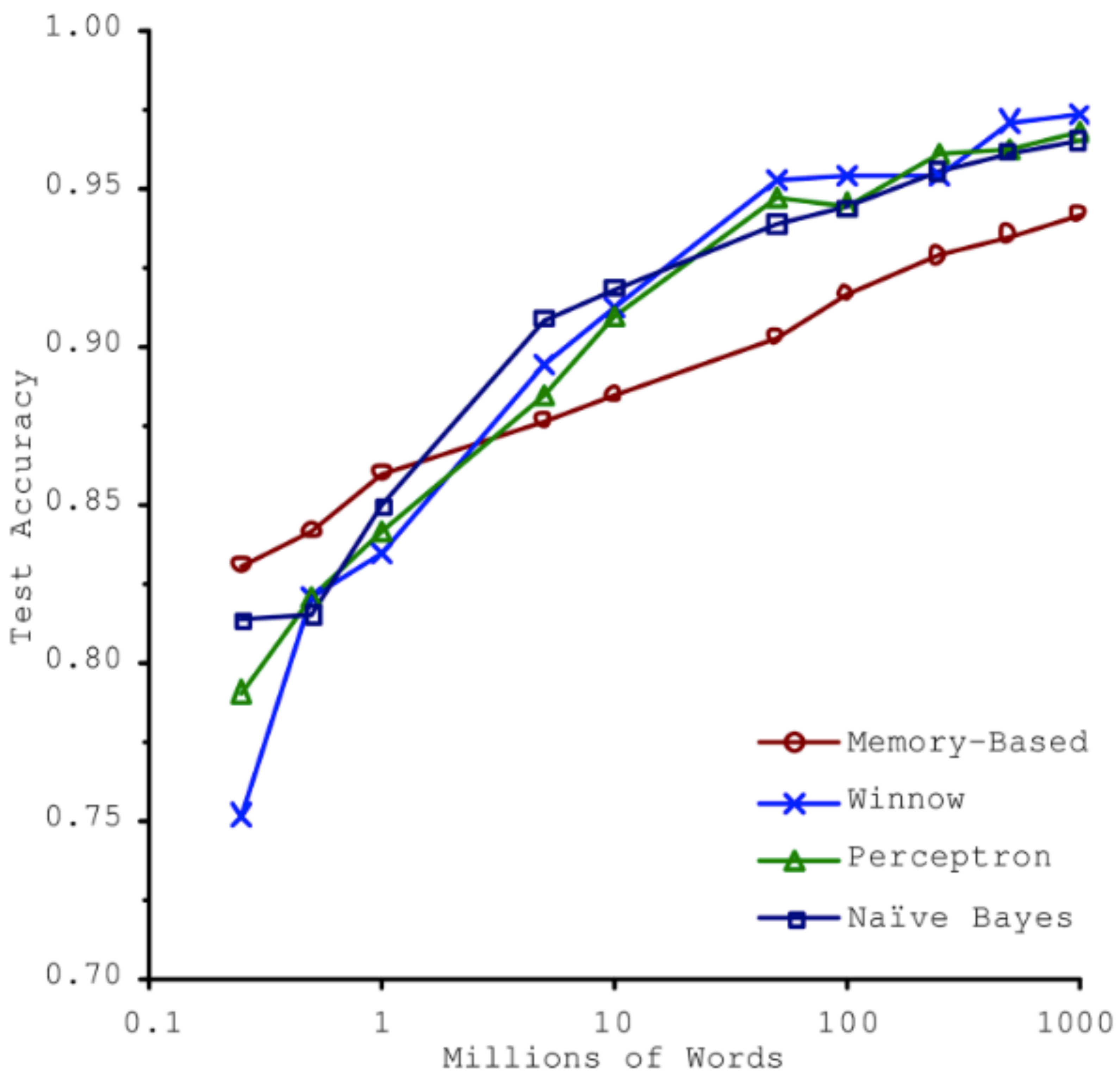
Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The

- Microsoft researchers in 2001 showed that different algorithms performed almost identically all on a complex NLP problem once enough data was given.
- "These results suggest that we may want to reconsider the trade-off between spending time and money on algorithm development versus spending it on corpus development."
- This idea was further popularized by Google researchers.

The Importance of Data Vs. Algorithm

Data Matters

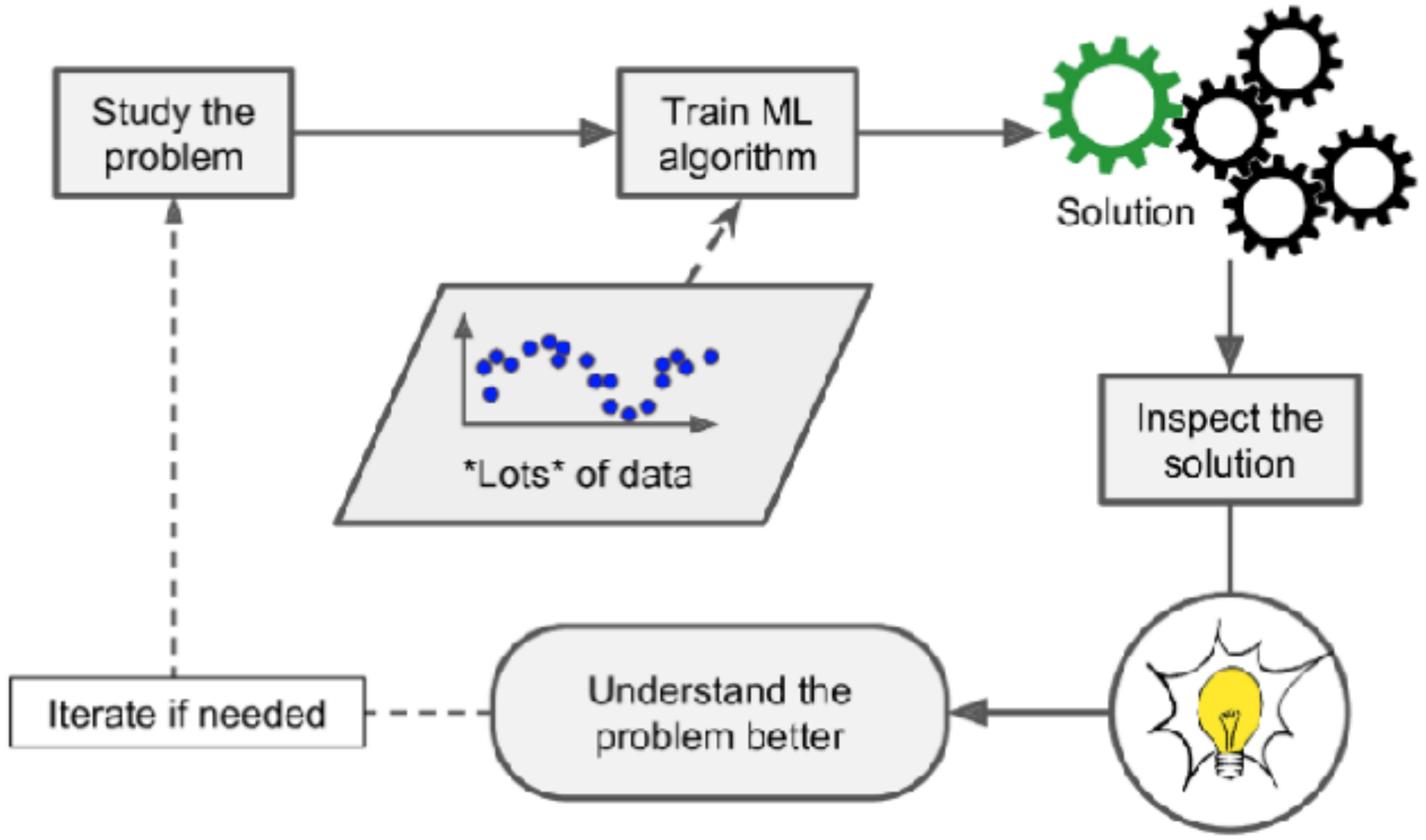


- The performance of different learning algorithms as data varies
- Higher test accuracy indicates better performance
- Notice the main trends,
  - All algorithms perform **“bad” with less data**.
  - All algorithms perform **“well” with large quantities of data.**
- Data is too often overlooked by computer and data scientists!

Main Challenges of Machine Learning

#1 Insufficient Quantity of Training Data

- Babies and Toddlers can learn from few examples
  - Instantly learn how to eat, who are their parents.
  - Easily learn objects (e.g. spoon, bottle, etc.)
- **Machines are not as good!**
  - A **lot** of data is needed for algorithms to work



Main Challenges of Machine Learning

#2 Nonrepresentative Training Data

- Data you ***have*** must be representative of data you ***expect*** to see.

- Suppose you want to know if money makes people happy. You have data that shows the GDP per capita for multiple countries and user rated Life Satisfaction

The current data implies “happiness” increases linearly with money!

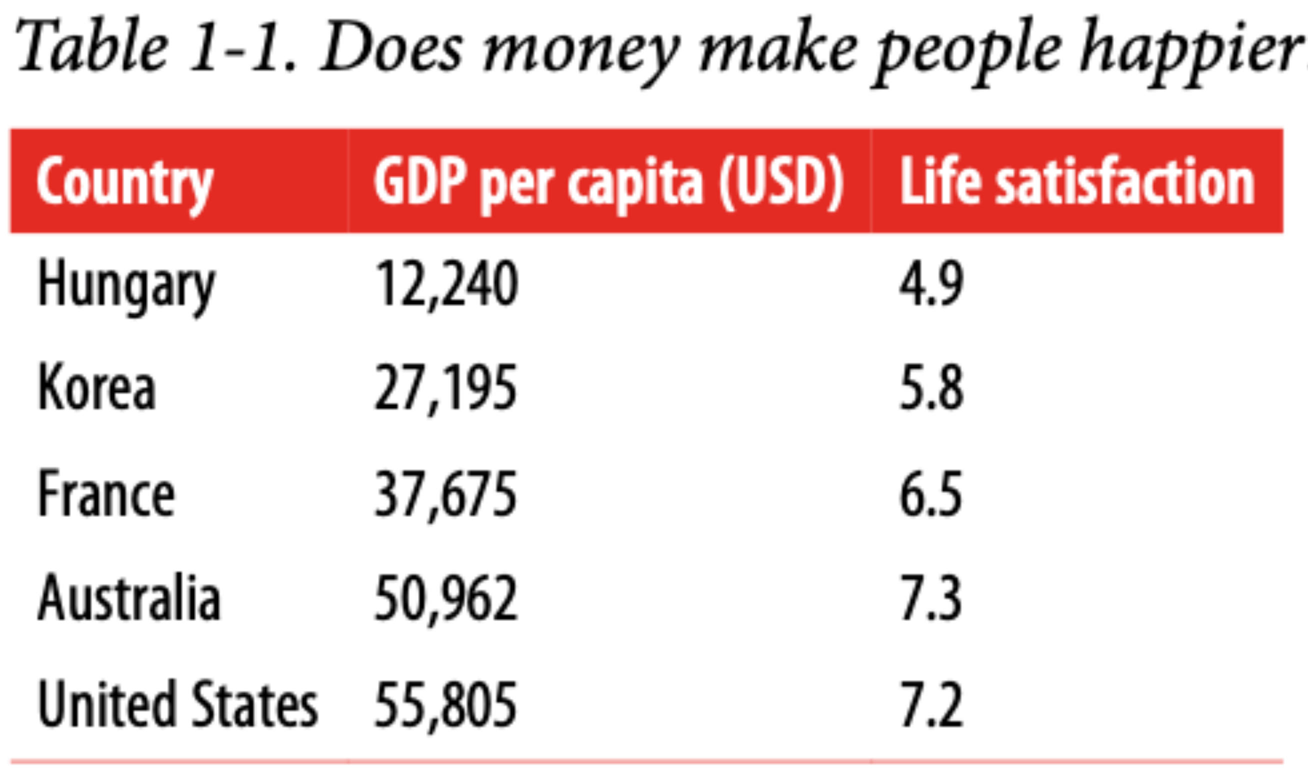
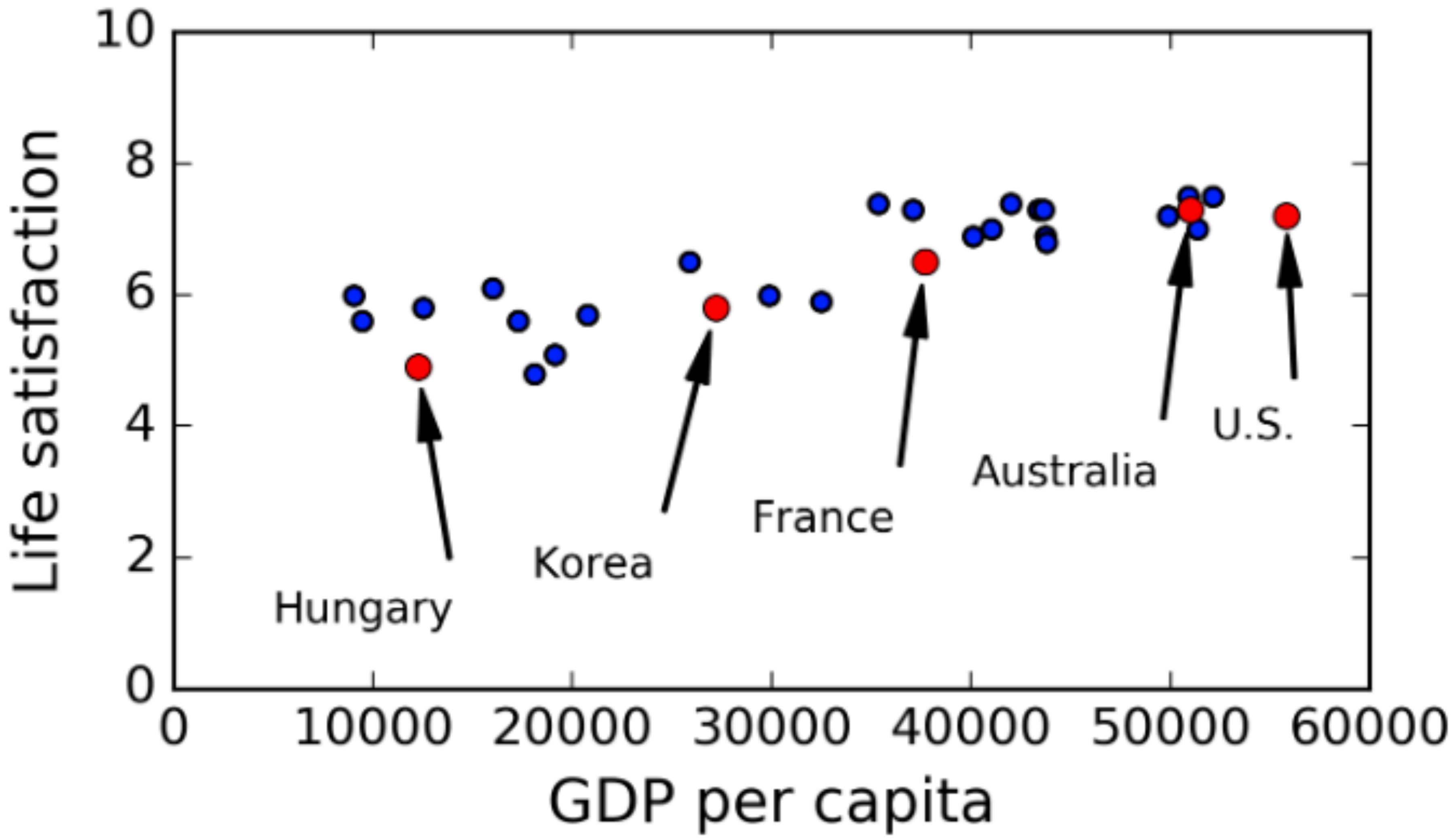


Table 1-1. Does money make people happier?

Country	GDP per capita (USD)	Life satisfaction
Hungary	12,240	4.9
Korea	27,195	5.8
France	37,675	6.5
Australia	50,962	7.3
United States	55,805	7.2



Main Challenges of Machine Learning

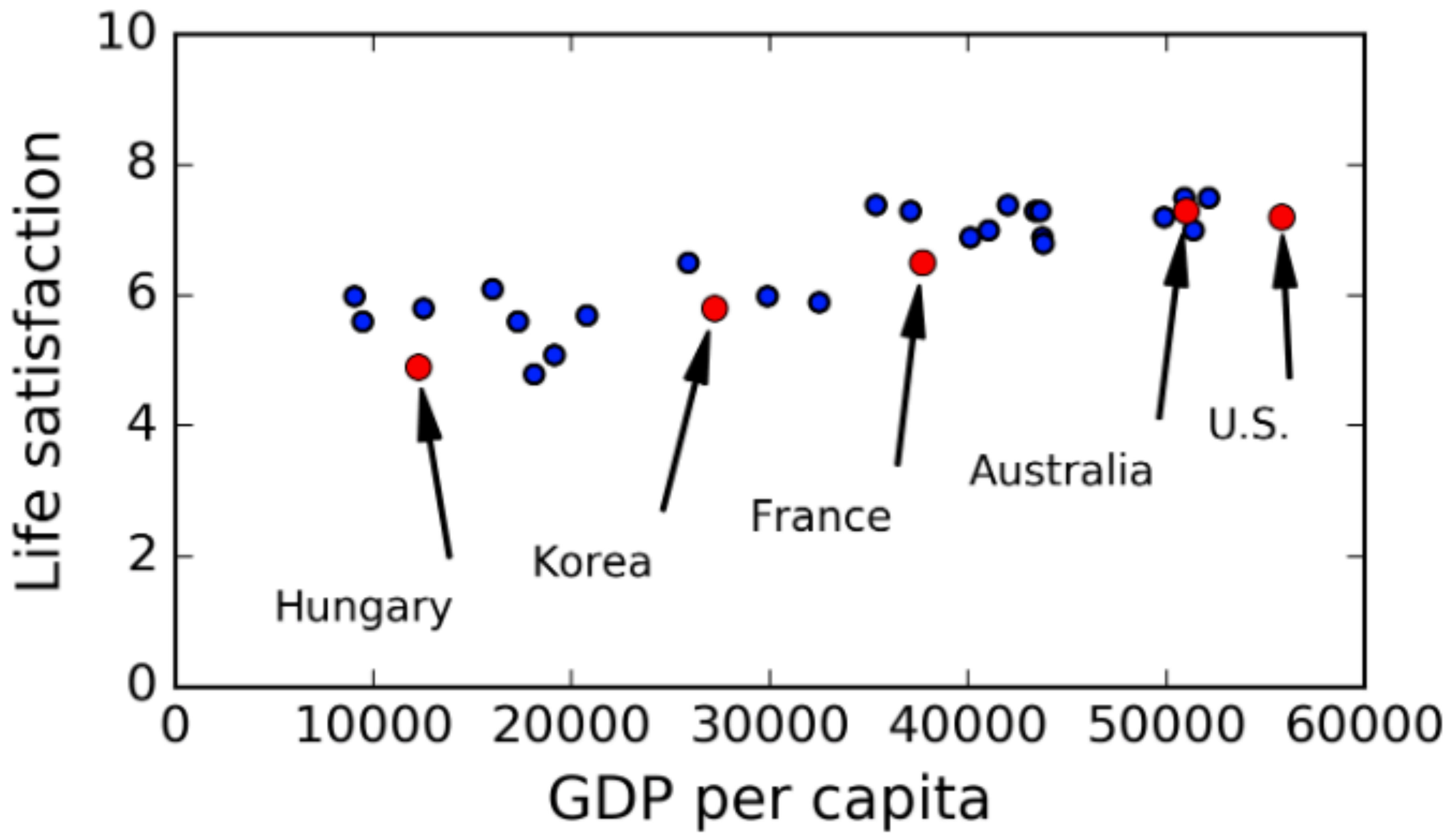
#2 Nonrepresentative Training Data

- Data you ***<u>have</u>*** must be representative of data you ***<u>expect</u>*** to see.
- Suppose you want to know if money makes people happy. You have data that shows the GDP per capita for multiple countries and user rated Life Satisfaction

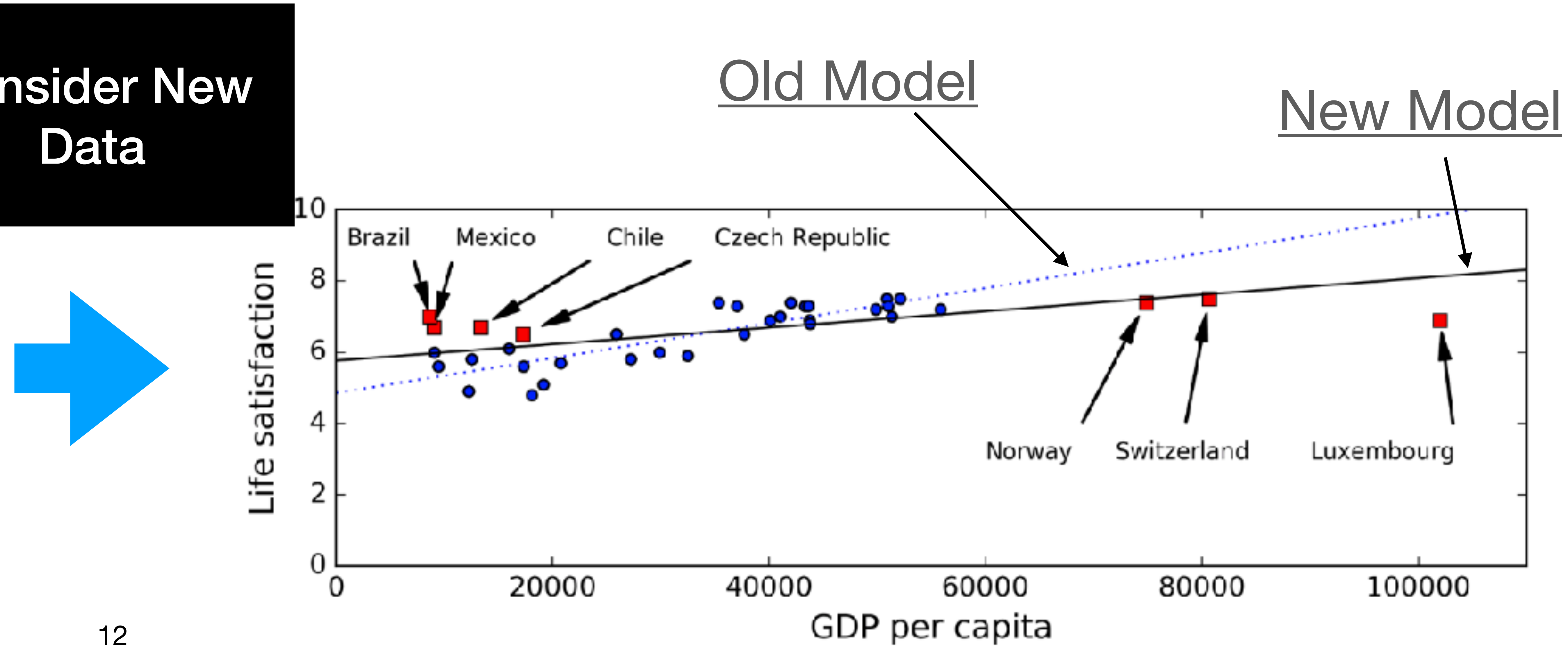
New data doesn't follow the old trends!

Very rich countries are not happier than moderately rich countries (may be unhappy)

Some poor countries seem happier than rich



Consider New Data



Main Challenges of Machine Learning

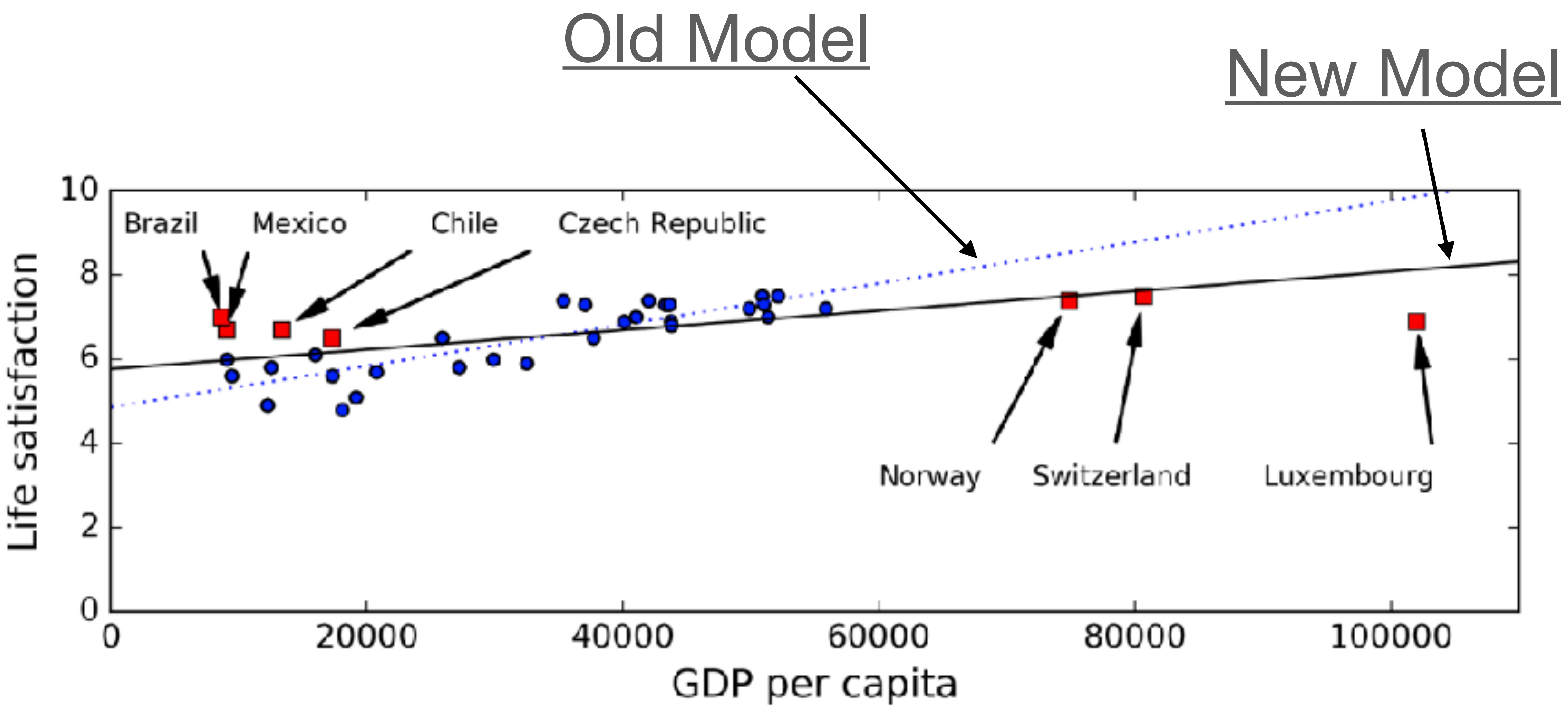
#2 Nonrepresentative Training Data

- Data you ***have*** must be representative of data you ***expect*** to see.
- Suppose you want to know if money makes people happy. You have data that shows the GDP per capita for multiple countries and user rated Life Satisfaction
- Your training data set ***MUST*** be representative of the cases you want to ***generalize*** to!
  - In this case, what type of data do you need?

New data doesn't follow the old trends!

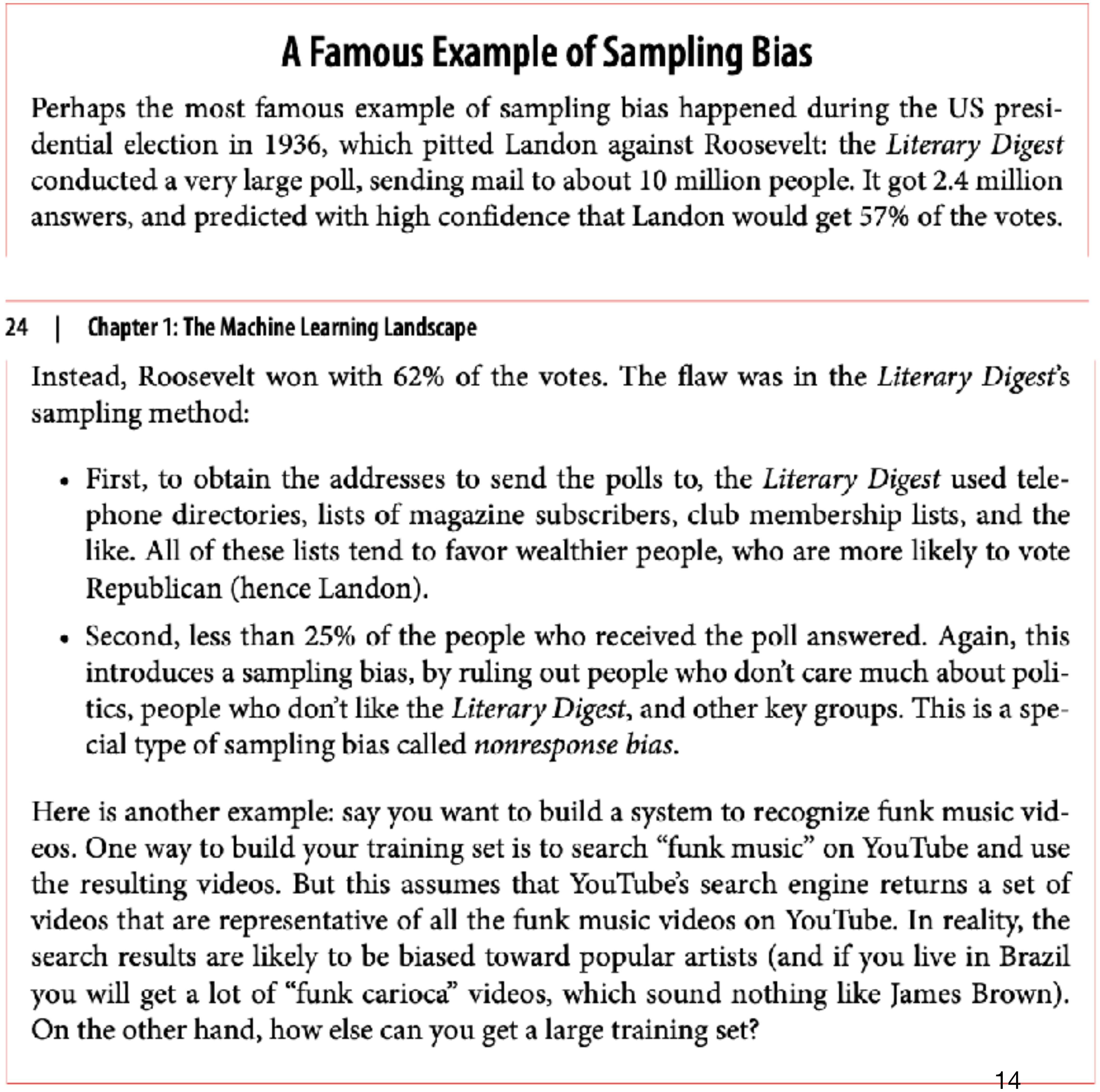
Very rich countries are not happier than moderately rich countries (may be unhappy)

Some poor countries seem happier than rich



Main Challenges of Machine Learning

#2 Nonrepresentative Training Data - Sampling Bias



A Famous Example of Sampling Bias

Perhaps the most famous example of sampling bias happened during the US presidential election in 1936, which pitted Landon against Roosevelt: the *Literary Digest* conducted a very large poll, sending mail to about 10 million people. It got 2.4 million answers, and predicted with high confidence that Landon would get 57% of the votes.

24 | Chapter 1: The Machine Learning Landscape

Instead, Roosevelt won with 62% of the votes. The flaw was in the *Literary Digest*'s sampling method:

- First, to obtain the addresses to send the polls to, the *Literary Digest* used telephone directories, lists of magazine subscribers, club membership lists, and the like. All of these lists tend to favor wealthier people, who are more likely to vote Republican (hence Landon).
- Second, less than 25% of the people who received the poll answered. Again, this introduces a sampling bias, by ruling out people who don't care much about politics, people who don't like the *Literary Digest*, and other key groups. This is a special type of sampling bias called *nonresponse bias*.

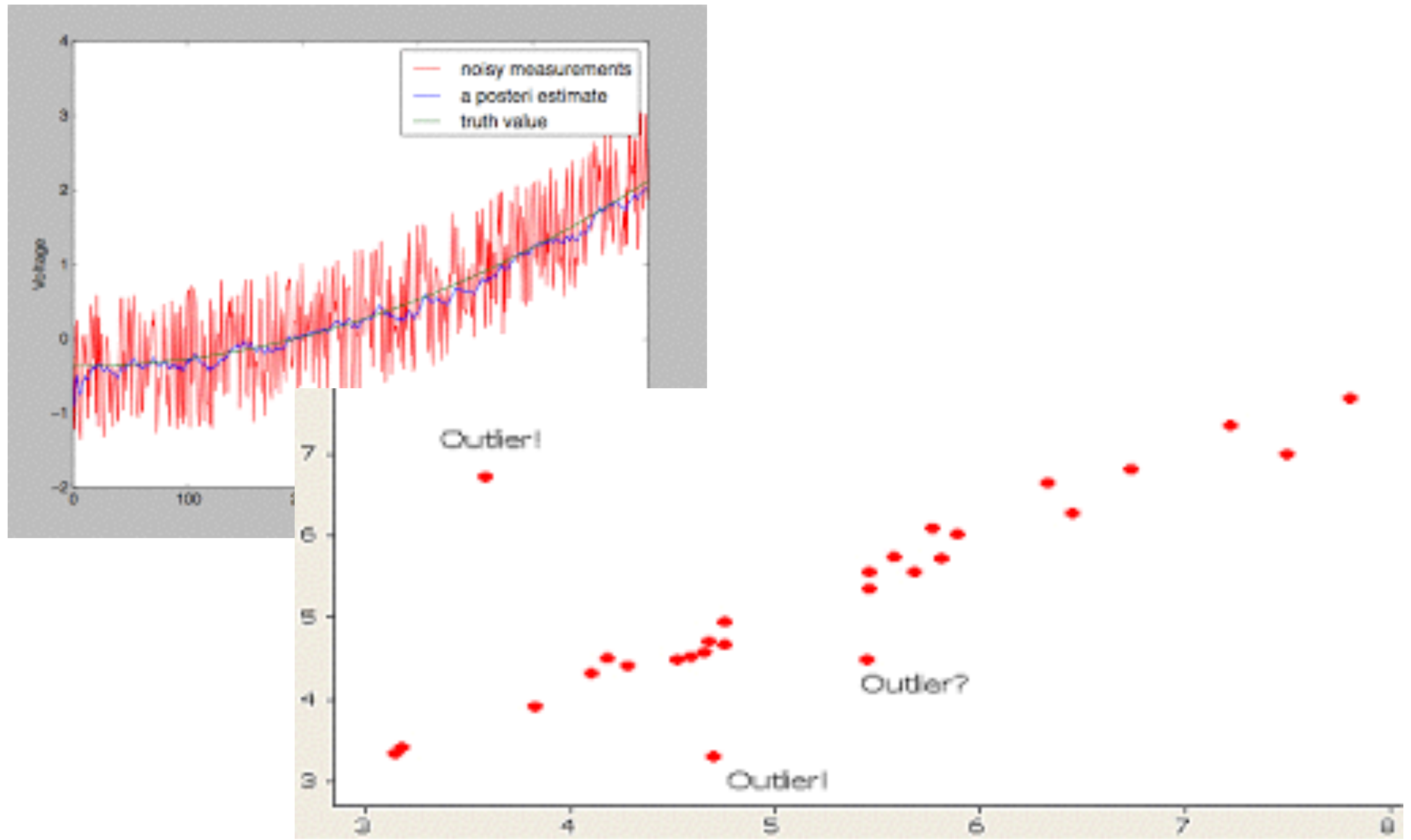
Here is another example: say you want to build a system to recognize funk music videos. One way to build your training set is to search "funk music" on YouTube and use the resulting videos. But this assumes that YouTube's search engine returns a set of videos that are representative of all the funk music videos on YouTube. In reality, the search results are likely to be biased toward popular artists (and if you live in Brazil you will get a lot of "funk carioca" videos, which sound nothing like James Brown). On the other hand, how else can you get a large training set?

- Can you think of or imagine any other instances of sampling bias?

Main Challenges of Machine Learning

#3 Poor Quality Data and #4 Irrelevant Features

- **Poor Quality Data**
  - Errors, outliers and noise (poor sensors) make machine learning more difficult
  - Often need to clean data in this scenario
    - Discard or remove outliers
    - Ignore missing features/attributes, a particular instance, or fill in values
    - Find/collect “replacement” data
- **Irrelevant Features**
  - Garbage In, Garbage Out
  - Feature engineering may be needed:
    - **Selection**: select most useful features
    - **Extraction**: combine existing feaures
    - **Creation**: create new features from new data



ARE YOU SURE THE DATA YOU GAVE ME IS CORRECT?
I'VE BEEN GIVING YOU INCORRECT DATA FOR YEARS. THIS IS THE FIRST TIME YOU'VE ASKED.
WHAT?
I SAID THE DATA IS TOTALLY ACCURATE.
Dilbert.com DilbertCartoonist@gmail.com
5-7-14 ©2014 Scott Adams, Inc./Dist. by Universal Uclick

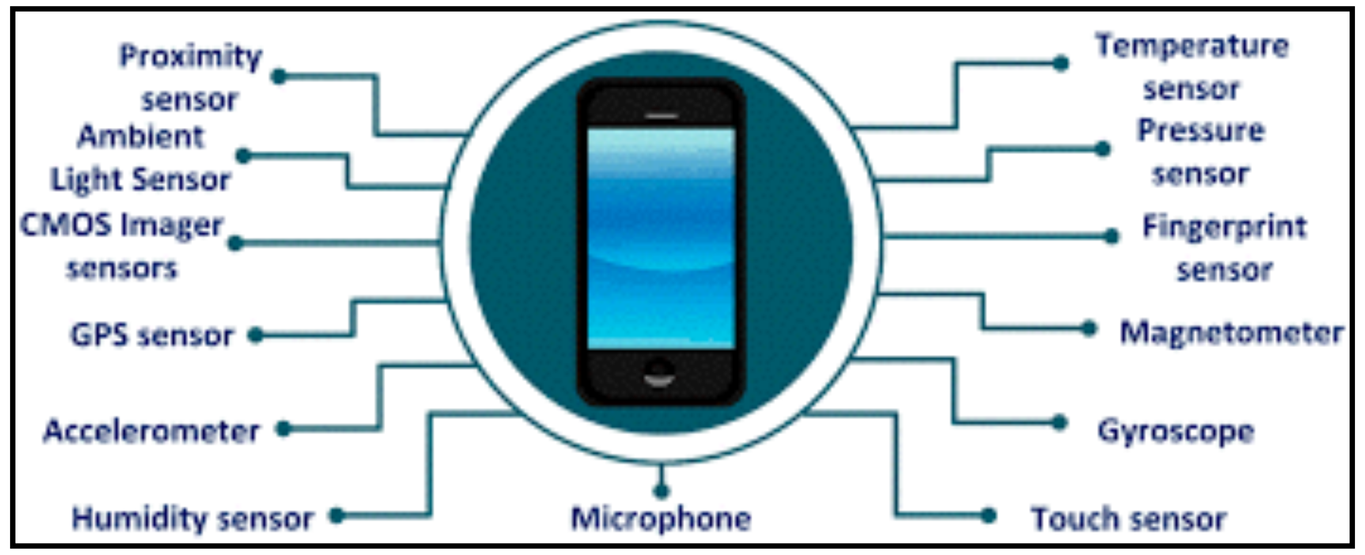
Data Selection for Experimental Design

Group practice - Zoom Break out rooms

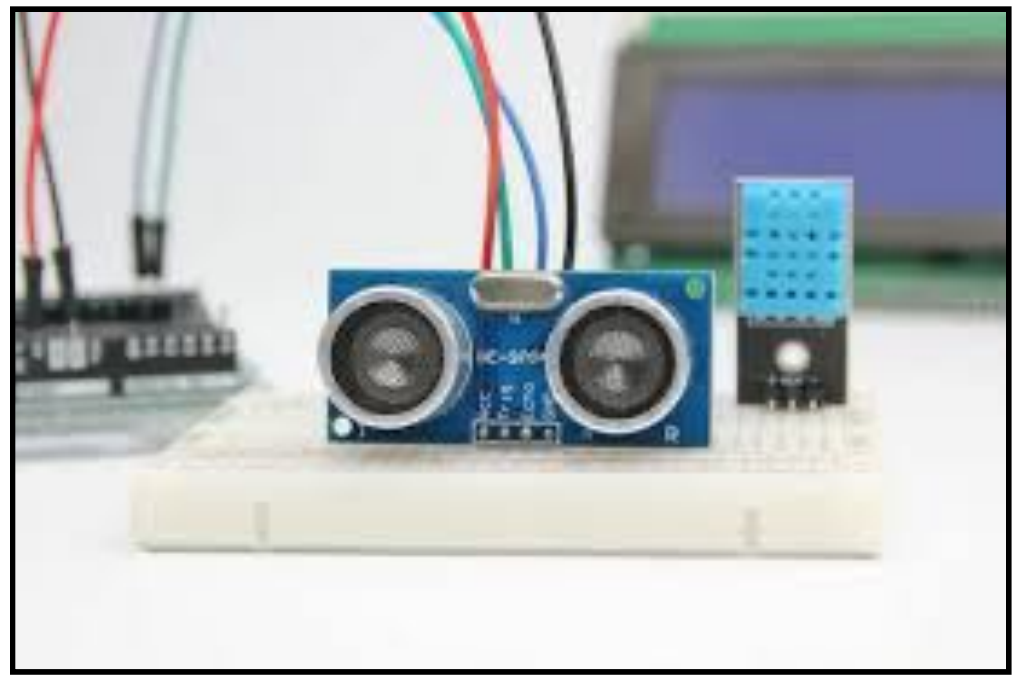
- **Scenario**: You are a machine learning scientists and you want to develop an approach that can classify whether a given fruit on a farm is ripe or not. You want your approach to work for as many fruits and farms as possible.
- **Task**: You need to gather/collect data to train your system. In a group, think of the type of data that you need, including how/where/when it is collected or gathered from. Be sure that you consider the bad data problem.
  - In each group, one person take notes to summarize your approach
  - We'll come back and discuss in 5-10 minutes
- **Assumptions**:
  - You have the perfect classifier to use and metrics for evaluation
  - A drone can be used to fly over the farm(s) to see the fruits
  - Data will be annotated accurately once gathered

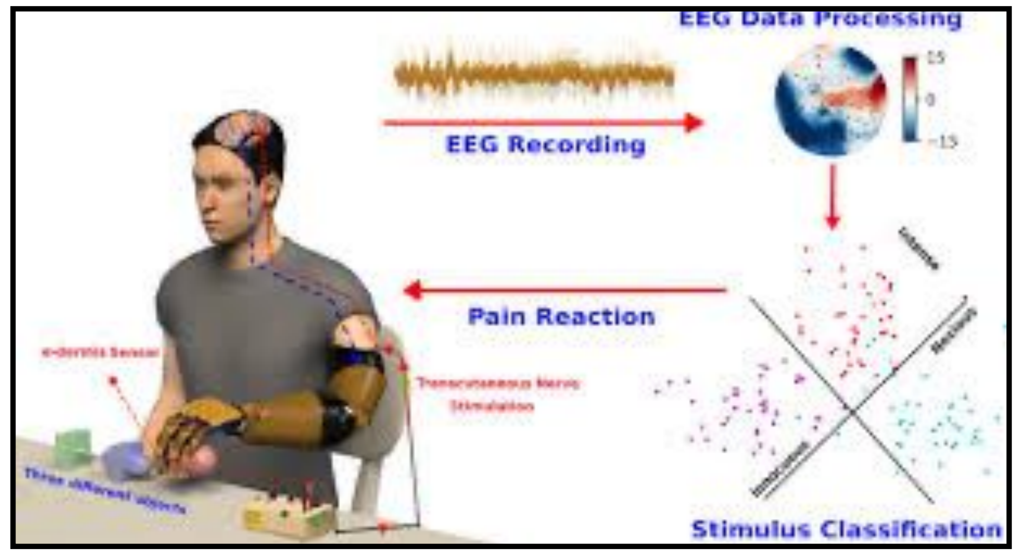
How do we get Data?

Collect it Ourselves



- Sensors for data collection are everywhere!
- Carefully designed experiments must be conducted.
- Must consider data challenges
    - **<u>Inaccurate, noisy and/or missing data</u>**: Over collect when possible
    - **<u>Data imbalance and bias</u>**: must ensure “enough” data from ALL classes, scenarios, and environments are collected

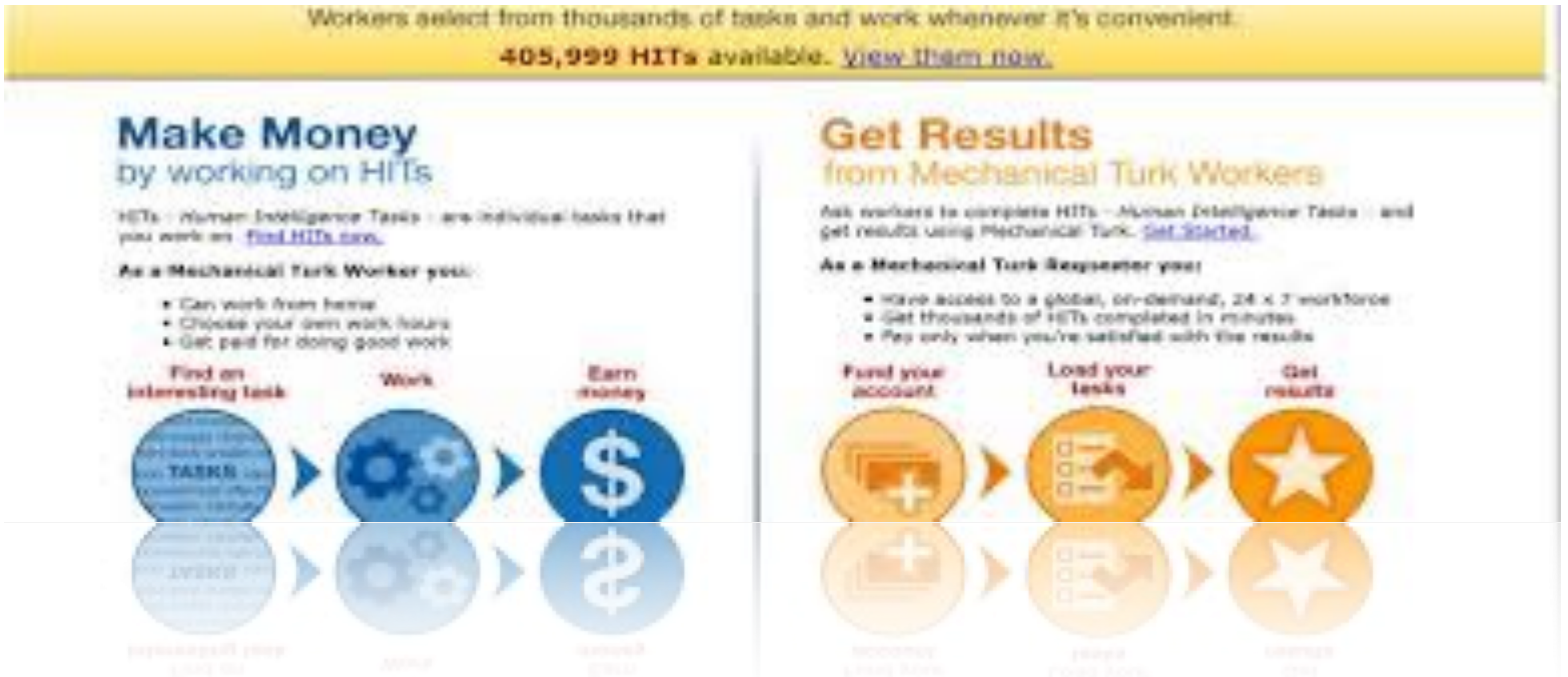




How do we get Data?

Collect it Ourselves

- Now that you have data, it must be annotated!
- This is the hard, costly and time-consuming part!
- For example, objects in pictures must be denoted. Transcriptions of audio must be provided
- You may have to do this yourself! Luckily online crowdsourcing has helped with this.
  - This is perfect, right?



Workers select from thousands of tasks and work whenever it's convenient.

405,999 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

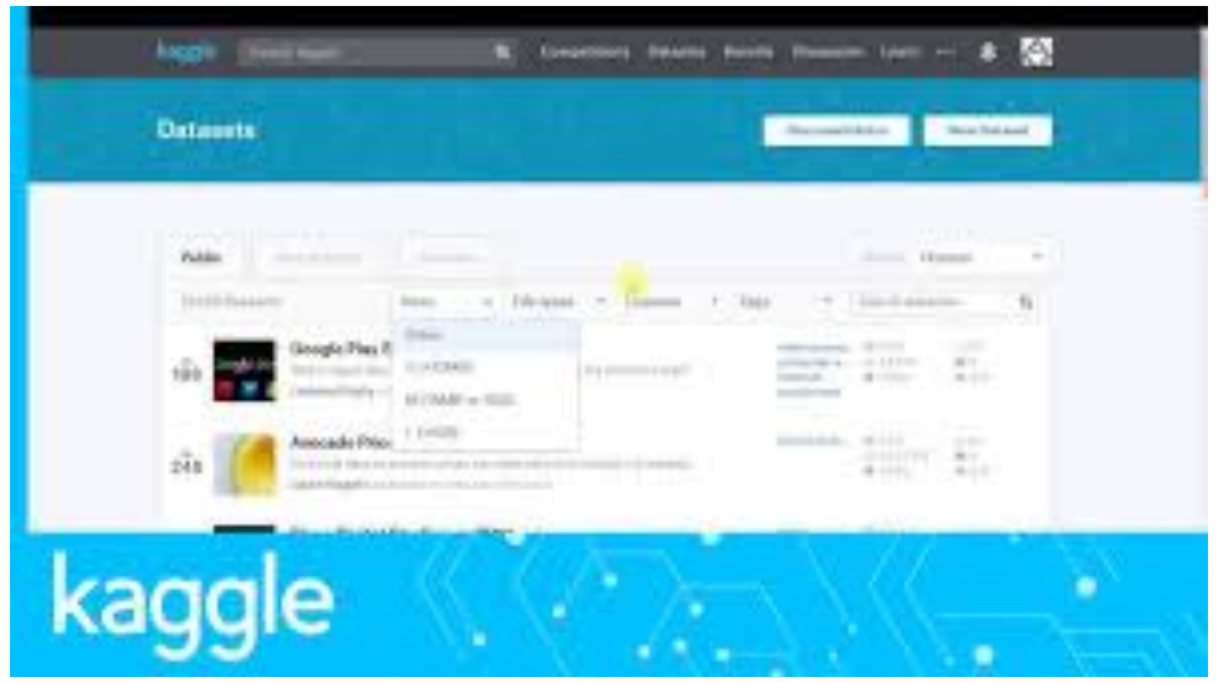
How do we get Data?

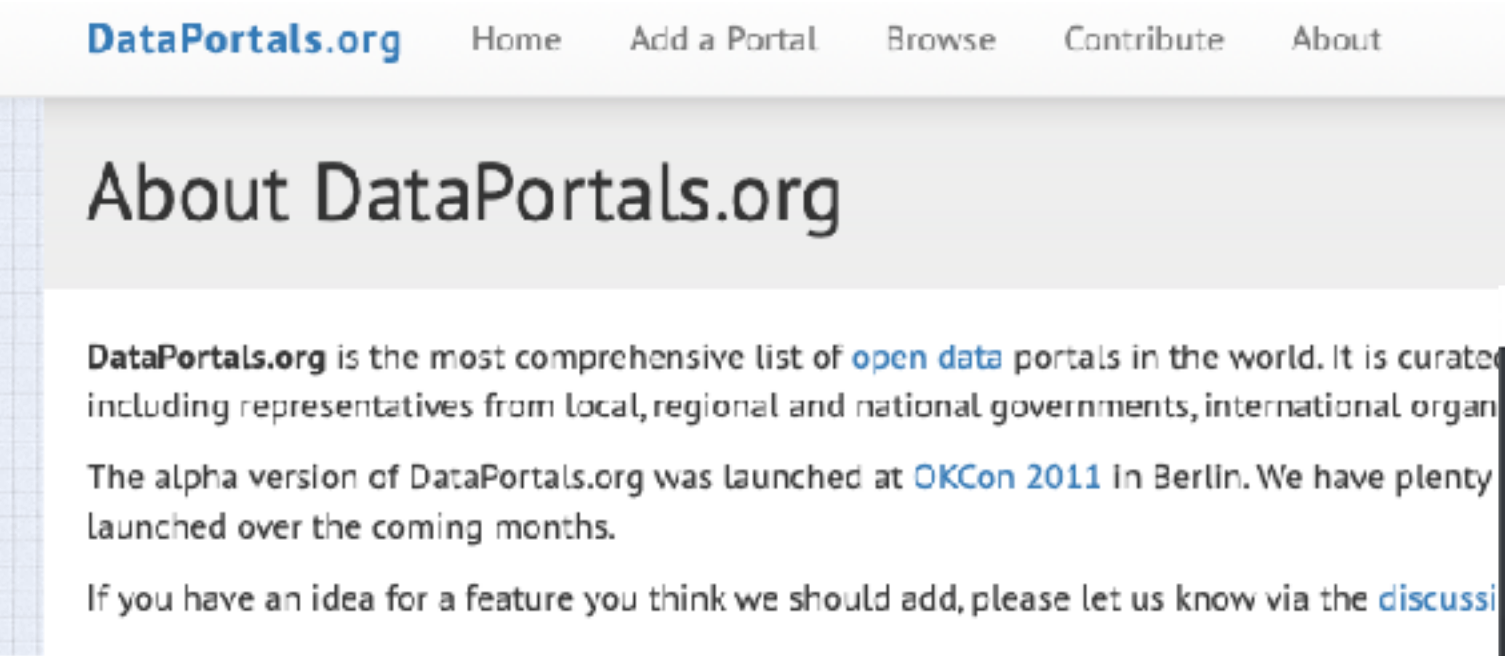
Online Sources

- Thankfully, there are many downloadable datasets that we can start with!

User Tweets Analysis & Data

Export any public Twitter user's recent 3200 tweets in a CSV/Excel datasheet and analyze it with the help of FollowersAnalysis' advanced AI-based analytical tool. The analytics include the best time to tweet, the Most mentioned keywords, Most retweeted, liked tweets and Week-Day Peak Usage Pattern, etc.



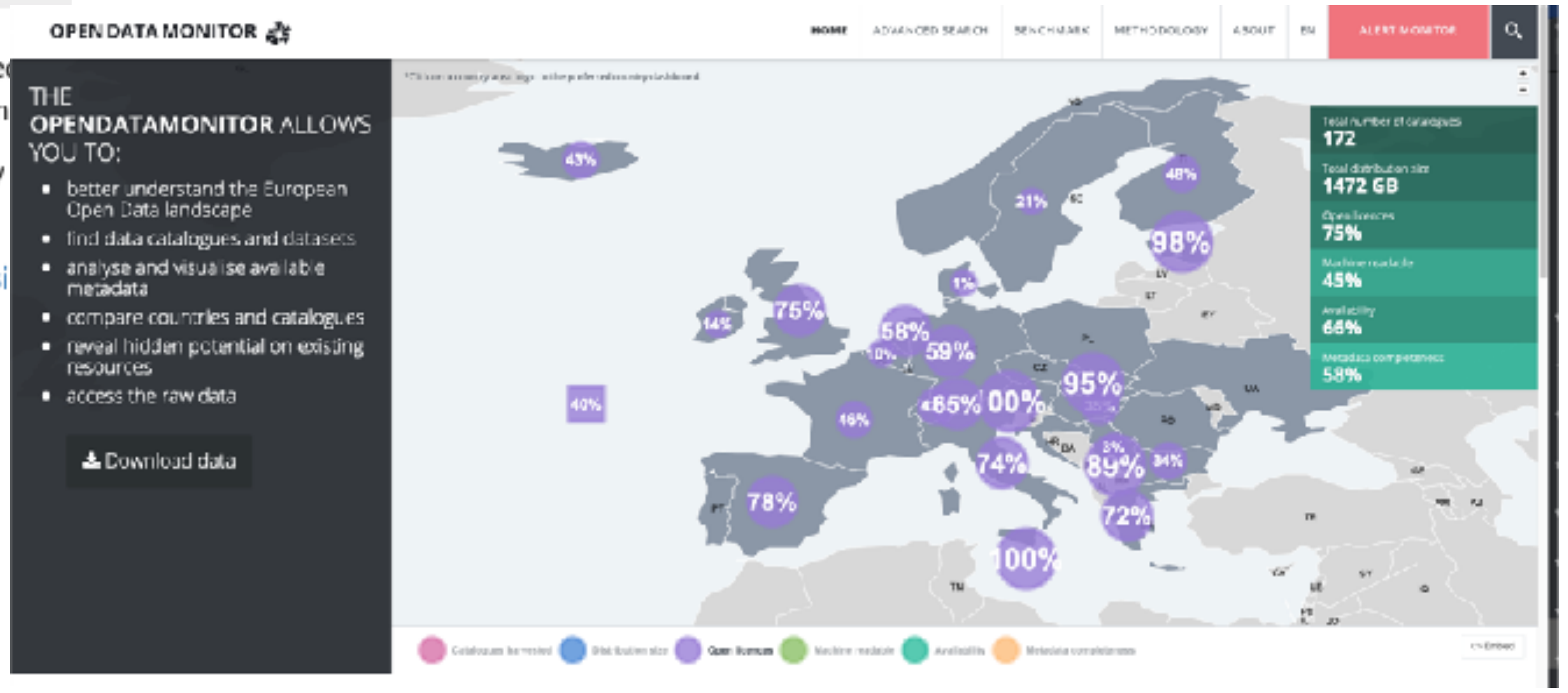


About DataPortals.org

DataPortals.org is the most comprehensive list of open data portals in the world. It is curate
including representatives from local, regional and national governments, international organ

The alpha version of DataPortals.org was launched at OKCon 2011 in Berlin. We have plenty
launched over the coming months.

If you have an idea for a feature you think we should add, please let us know via the discussi



THE **OPENDATAMONITOR** ALLOWS YOU TO:

- better understand the European Open Data landscape
- find data catalogues and datasets
- analyse and visualise available metadata
- compare countries and catalogues
- reveal hidden potential on existing resources
- access the raw data

Download data

Total number of catalogues
172

Total distribution size
1472 GB

Open licences
75%

Machine readable
45%

Availability
66%

Metadata completeness
58%

Data Pre-processing

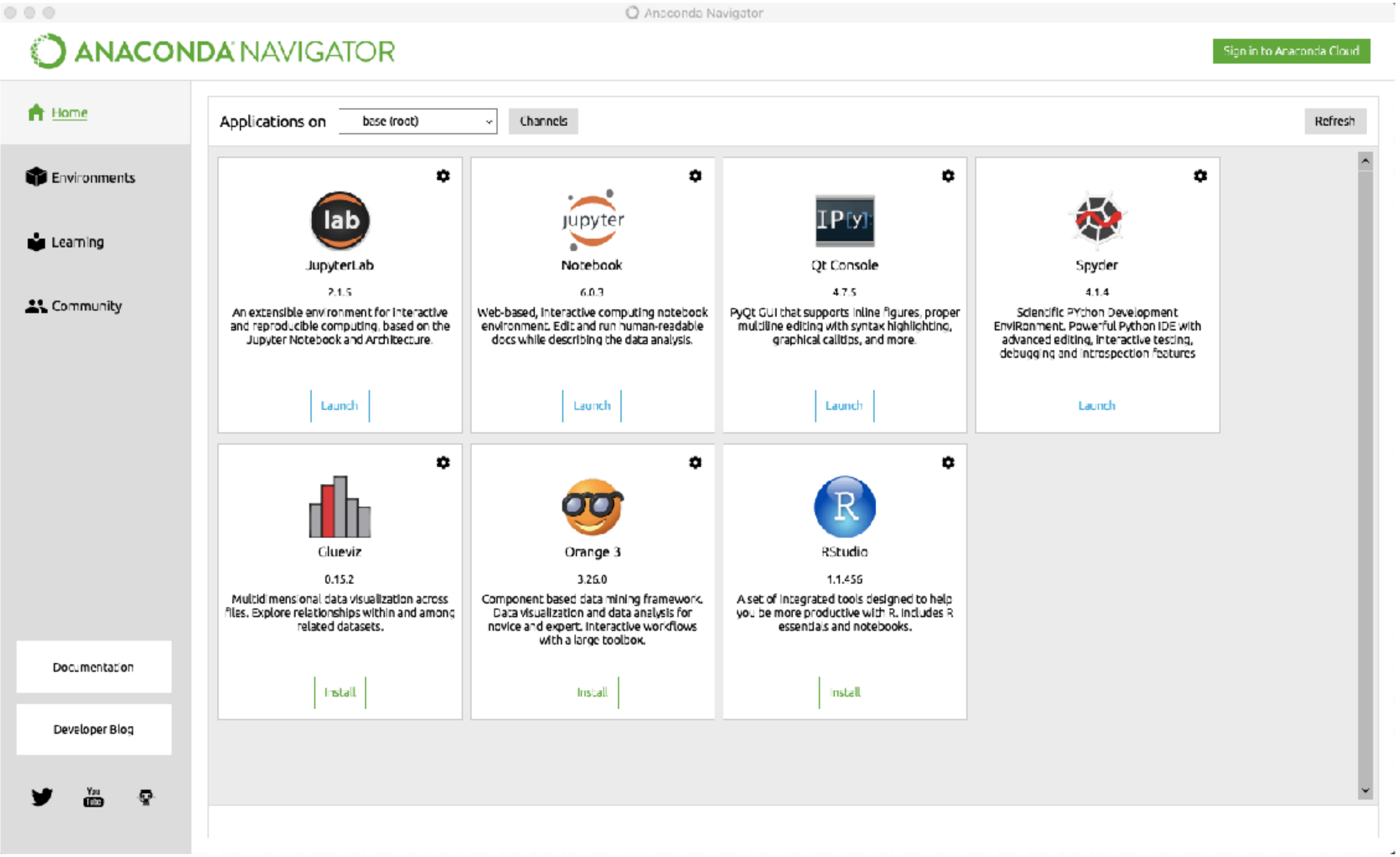
Now that we have data, what’s next? An Example Case

- Suppose you are a Data Scientist at a Housing Corporation. Your boss wants you to build a prediction model of median housing prices in California using their census data
- **Data has info about**: population, median income, median housing prices, … for each block group or district in California.
- **How should this problem be framed**?
  - Supervised Learning, Unsupervised learning, Reinforcement Learning? Why?
  - Classification, Regression, Other? Why?
  - Batch vs. Online?

Python Environment Setup

Option 1. Anaconda Environment

- Use one of the following links for instructions for working with Anaconda
  - https://docs.anaconda.com/anaconda/
  - https://docs.anaconda.com/anaconda/navigator/tutorials/manage-environments/
- Easy and intuitive programming environment



JupyterLab
2.1.5
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Notebook
6.0.3
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Qt Console
4.7.5
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Spyder
4.1.4
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Glueviz
0.15.2
Multidimensional data visualization across files. Explore relationships within and among related datasets.

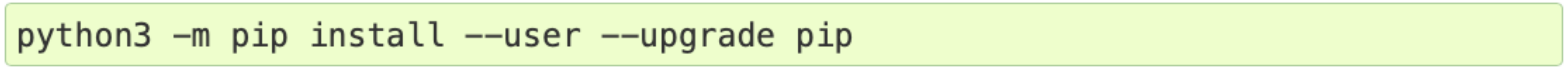
Orange 3
3.26.0
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

RStudio
1.1.456
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Python Environment Setup

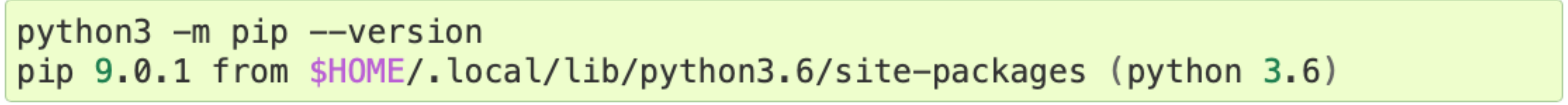
Option 2. Isolated/Local Environment

- Follow specific instructions for your system (Windows, Linux, Apple) from the site: https://packaging.python.org/guides/installing-using-pip-and-virtual-environments/
- Basic steps (Python 3.3 or later on Mac):
  - Install pip:

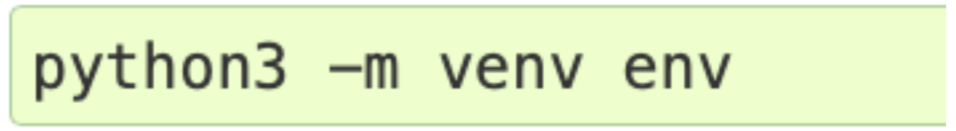
    ```
    python3 -m pip install --user --upgrade pip
    ```

  - Check the installation:

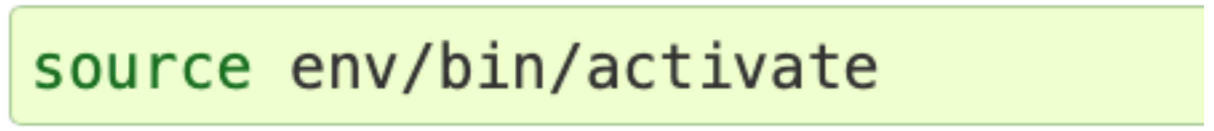
    ```
    python3 -m pip --version
    pip 9.0.1 from $HOME/.local/lib/python3.6/site-packages (python 3.6)
    ```

  - Create Virtual Environment:

    ```
    python3 -m venv env
    ```

  - Activate Virtual Environment:

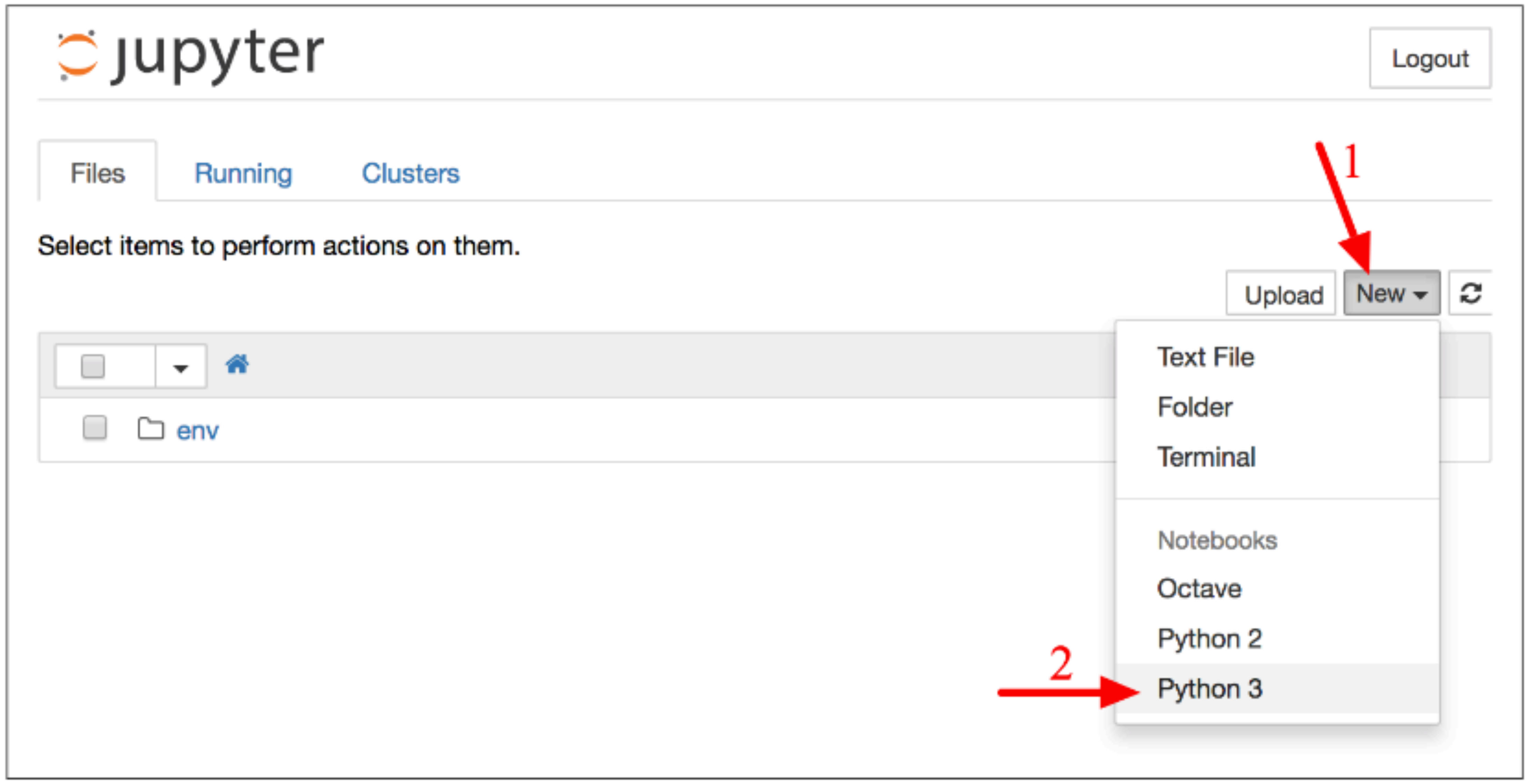
    ```
    source env/bin/activate
    ```

Python Environment Setup

Option 2. Isolated/Local Environment

- Now you an install any/all required modules

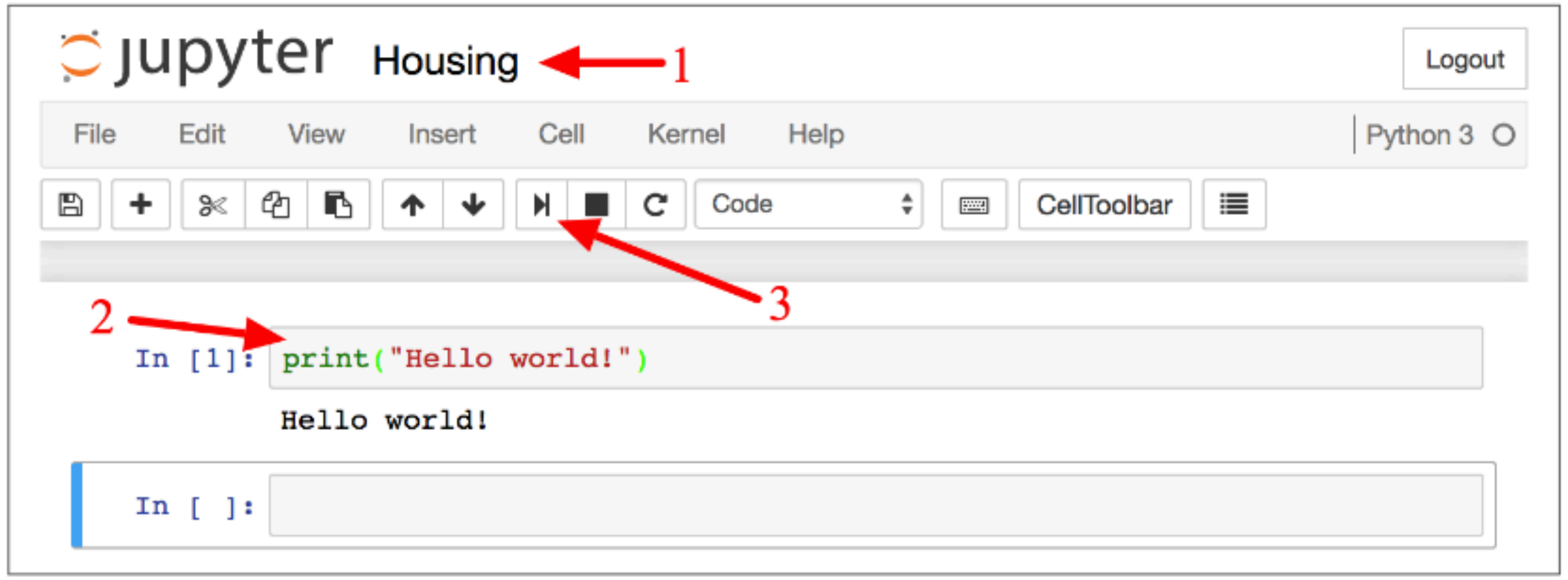
```
python3 –m pip install -U jupyter matplotlib dumpy pandas spicy scikit-learn
```



Select items to perform actions on them.

- You can now work in jupyter notebook (or commandline)

```
jupyter notebook
```



```
In [1]: print("Hello world!")

        Hello world!

In [ ]:
```

Next Class

Be sure your Python environment is setup appropriately

- Post issues to Piazza