# CMPS 142: Machine Learning and Data Mining

Snigdha Chaturvedi (snigdha@ucsc.edu)

The material for slides for this course has been borrowed from several people.

# Welcome

- Instructor: Snigdha Chaturvedi
- TA: Tianyi Luo

- Lectures: TuTh 1:30-3:05 in N. Sci Annex 101
- Sections: W 12-01:05 and F 01:20-02:25 in PhysSciences 140.

- Office hours
- Instructor: Tuesdays 3:15-4:15 PM (343B E2)
- TA: Wednesdays 10:00-11:00 (477 E2)

# Welcome

- Lectures
  - You are expected to attend all lectures
  - Attend first week of lectures if you want to hold your spot
  - Please arrive on time to class
  - There is a no cell-phone rule in the class
  - There is also no need to bring laptops to the class
- Discussion Sections
  - Will be led by the TA
  - Attendance is strongly recommended
  - No discussion sections this week
- If you want to drop out, please do it soon

# Web Info

- Emailing the instructor will not be the fastest way to get a response
- Sign up on Piazza at https://piazza.com/university_of_california_santa_cruz/spring2018/cmps142 (access code=cmps142)
- Webpage: https://ucsc-courses.github.io/CMPS142-Spring2018/
- Check course website and piazza daily or every other day

# Grading

- Grades will be based on
  - Participation (in class and on Piazza) (5%)
  - 4-5 Assignments (40%) → in groups of 2 or 3
  - Midterm (25%) → in class and on May 3, 2018
  - Final Exam (30%)
  - You must pass the exams
  - This allocation can change

- Assignments are due at the beginning of the class on the due date. Late submission is allowed but with a penalty of 10% for every late day, upto 3 days. After 3 days you don't get any points.

# Your responsibilities

- Prepare for the lectures.
- Submit all assignments etc. on time
  - Assignments are challenging and not meant to be completed in a day
  - Don't wait till the last day to start your assignment
- Ensure that you understand the material by attending lectures and sections and asking questions (piazza, class, or office hours)
- If you have special needs, contact the Disability Resource Center and bring your Accommodation Authorization form to me after class or during office hours during the first two weeks

# Notes

- You must be comfortable with probability theory, vector algebra, and of course, programming

- Homework 0 is out
  - Will not be graded, but there are points for submitting it
  - Due at the beginning of the class on April 10, 2018
  - Has to be done individually but discussion is allowed

# Today

- Who are you?
- What is ML and what is this course about?

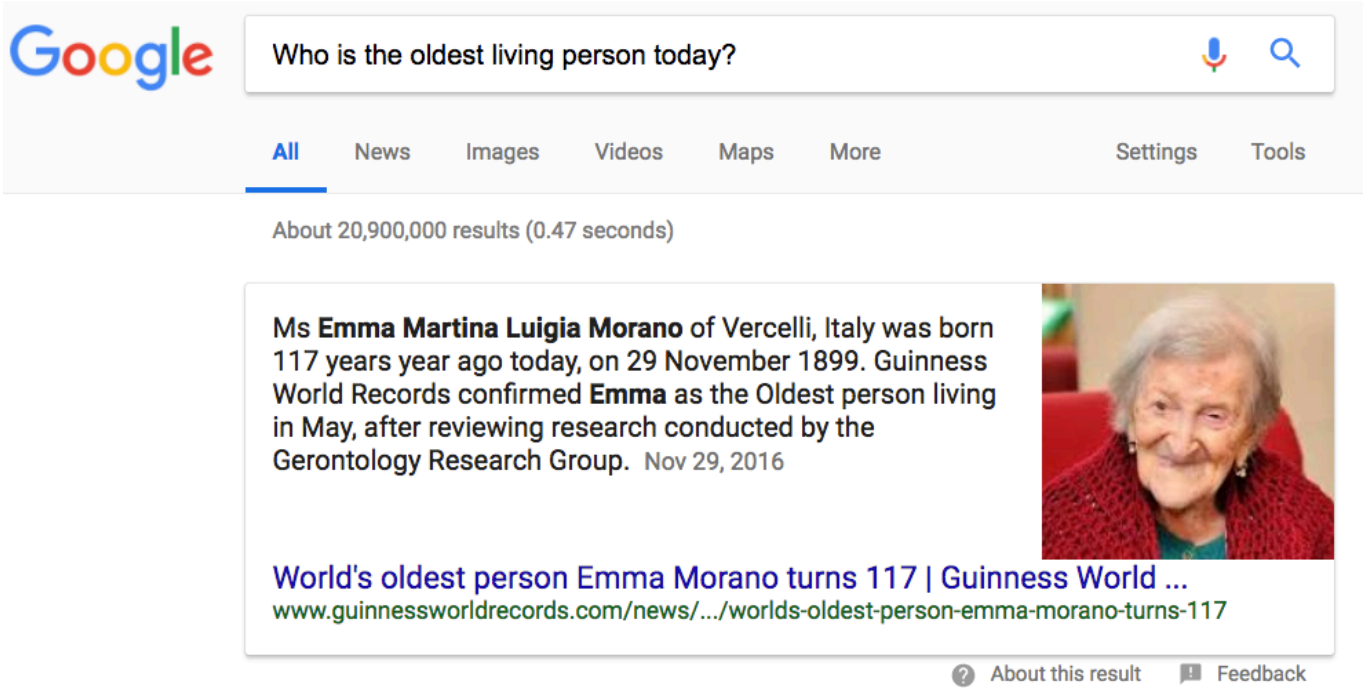
# Syllabus

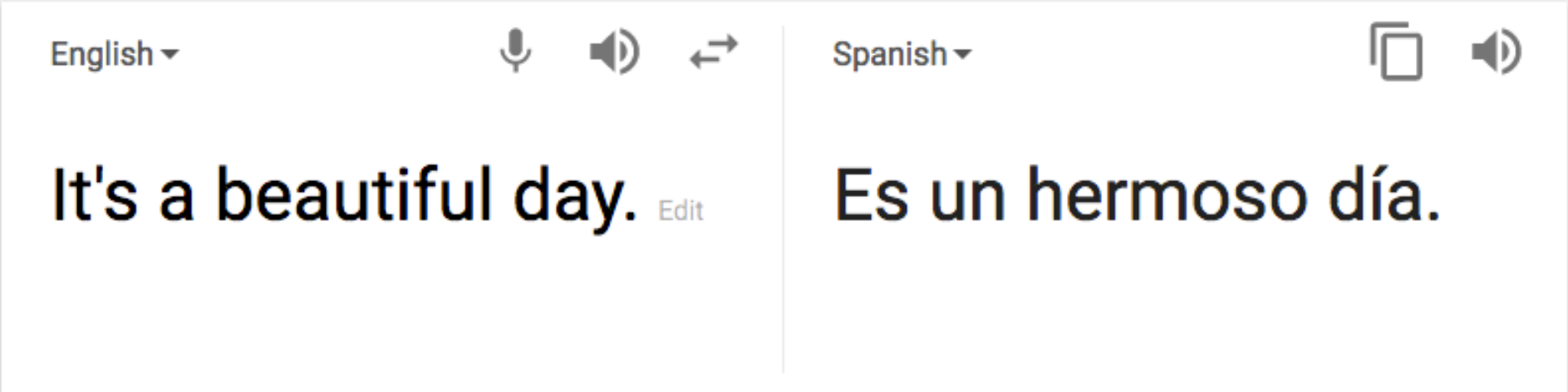
- Introduction and key concepts
- Supervised Learning
  - Linear regression
  - Regularization and Bias-Variance tradeoff
  - Logistic regression
  - Probability review
  - Generative learning models, Naive Bayes
  - Perceptron Algorithm
  - Support Vector Machines
  - Decision Trees
  - Neural networks
  - Model selection and feature selection
  - Ensemble Methods
  - Multi-class classification
- Unsupervised Learning
- On-line Learning

# Introduction to Machine Learning

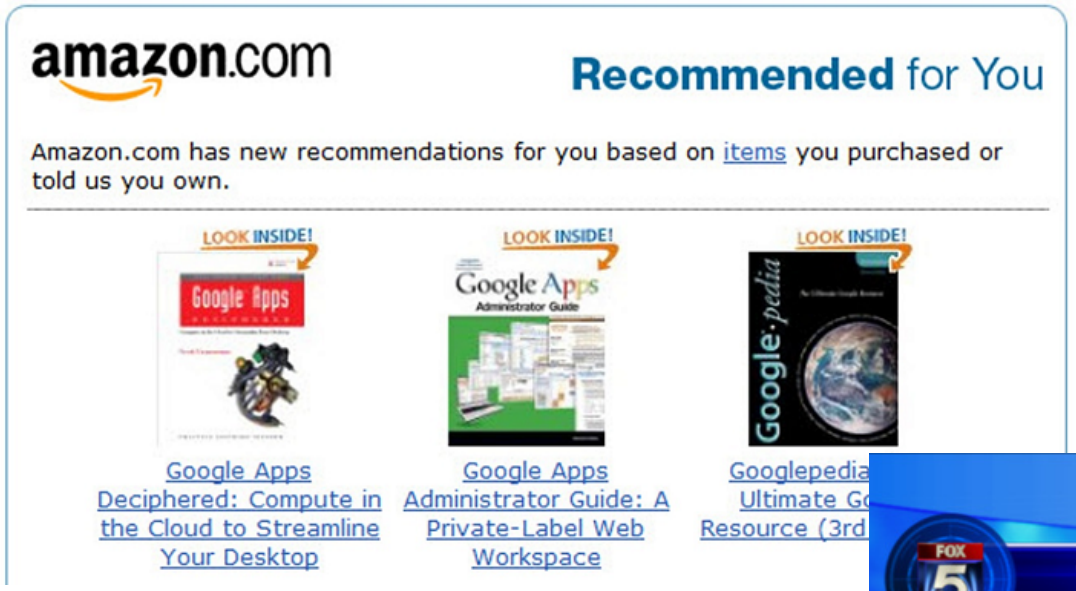
# Machine Learning is everywhere


# Machine Learning is everywhere

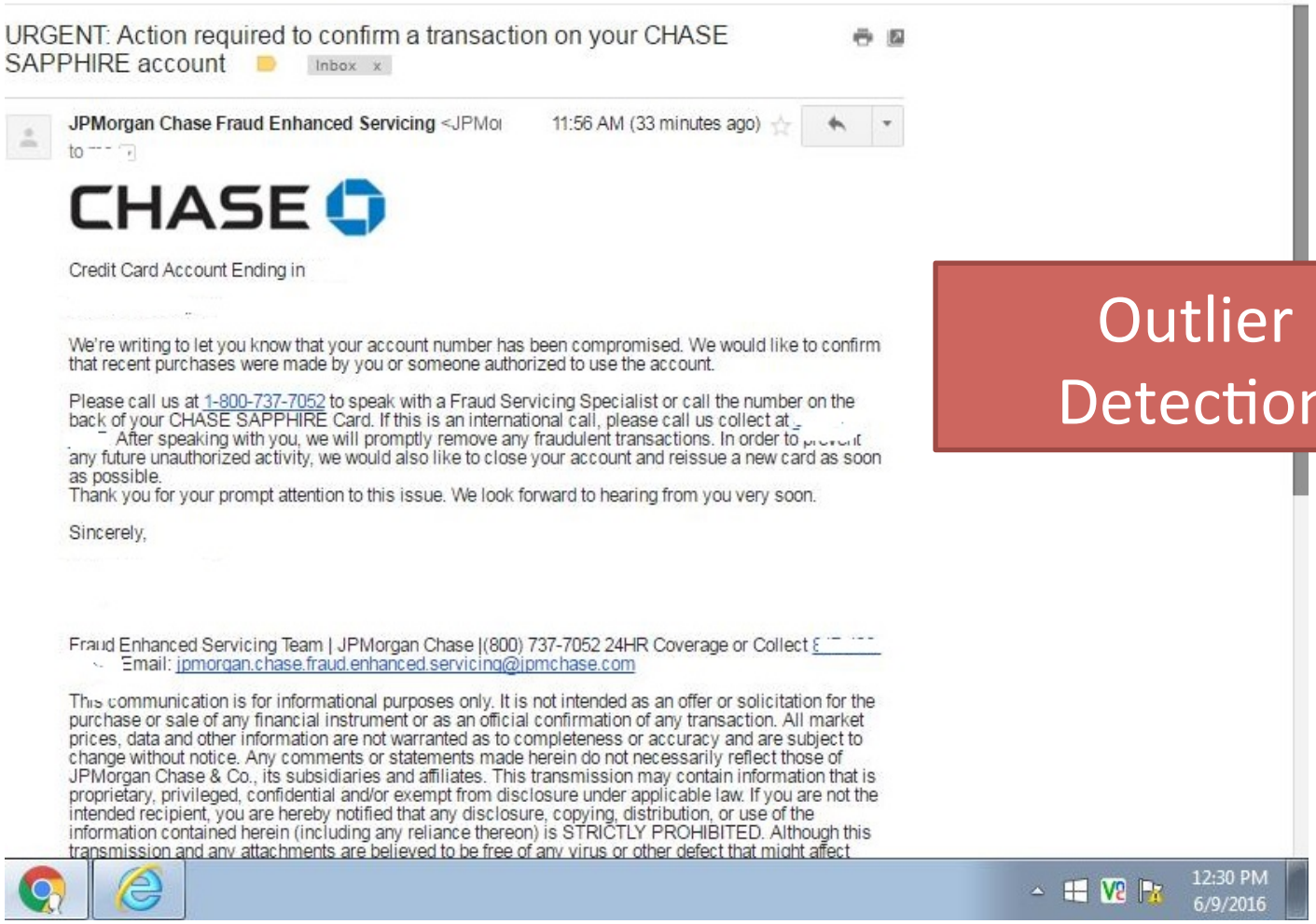


# Machine Learning is everywhere




# Machine Learning is everywhere



URGENT: Action required to confirm a transaction on your CHASE SAPPHIRE account

Inbox x

**JPMorgan Chase Fraud Enhanced Servicing** <JPMor 11:56 AM (33 minutes ago)

to

CHASE

Credit Card Account Ending in

We're writing to let you know that your account number has been compromised. We would like to confirm that recent purchases were made by you or someone authorized to use the account.

Please call us at 1-800-737-7052 to speak with a Fraud Servicing Specialist or call the number on the back of your CHASE SAPPHIRE Card. If this is an international call, please call us collect at After speaking with you, we will promptly remove any fraudulent transactions. In order to prevent any future unauthorized activity, we would also like to close your account and reissue a new card as soon as possible.
Thank you for your prompt attention to this issue. We look forward to hearing from you very soon.

Sincerely,

Fraud Enhanced Servicing Team | JPMorgan Chase |(800) 737-7052 24HR Coverage or Collect
Email: jpmorgan.chase.fraud.enhanced.servicing@jpmchase.com

This communication is for informational purposes only. It is not intended as an offer or solicitation for the purchase or sale of any financial instrument or as an official confirmation of any transaction. All market prices, data and other information are not warranted as to completeness or accuracy and are subject to change without notice. Any comments or statements made herein do not necessarily reflect those of JPMorgan Chase & Co., its subsidiaries and affiliates. This transmission may contain information that is proprietary, privileged, confidential and/or exempt from disclosure under applicable law. If you are not the intended recipient, you are hereby notified that any disclosure, copying, distribution, or use of the information contained herein (including any reliance thereon) is STRICTLY PROHIBITED. Although this transmission and any attachments are believed to be free of any virus or other defect that might affect

12:30 PM
6/9/2016

Outlier Detection

# When to use Machine Learning

- We believe that there is a process that explains the data
- Describing (writing a program for) the process is difficult, but possible to look at examples and 'label' them
- E.g. face recognition, consumer behavior

- Data is cheap and abundant (data warehouses); *knowledge* is expensive and scarce
- Goal is to Build a model that is a good and useful approximation to the data

# Data Mining

- Data Mining: application of ML to large databases
  - Finance (stock price prediction)
  - Telecommunication (optimizing service quality by studying call patterns)
  - Manufacturing (optimization, control and troubleshooting)
  - Medicine (medical diagnosis)

# Why Study Machine Learning?

- "A breakthrough in machine learning would be worth ten Microsofts"
  -Bill Gates, Chairman, Microsoft
- "Machine learning is the next Internet"
  -Tony Tether, Former Director, DARPA
- Machine learning is the hot new thing"
  -John Hennessy, President, Stanford
- "Web rankings today are mostly a matter of machine learning"
  -Prabhakar Raghavan, Dir. Research, Yahoo
- "Machine learning is going to result in a real revolution"
  -Greg Papadopoulos, CTO, Sun
- "Machine learning is today's discontinuity"
  -Jerry Yang, CEO, Yahoo

# Why Study Machine Learning?

- Computer systems with new capabilities.
- Time is right
  - Initial algorithms and theory in place.
  - Growing amounts of on-line data
  - Computational power available.

  - Necessity: many things we want to do cannot be done by "programming".
  - (Think about all the examples given earlier)

# Work in Machine Learning

- Makes Use of:
  - Probability and Statistics; Linear Algebra
- Related to:
  - Philosophy, Psychology (cognitive, developmental), Neurobiology, Linguistics, Vision, Robotics,….
- Has applications in:
  - AI (Natural Language; Vision; Planning; Robotics; HCI)
  - Engineering (Agriculture; Civil; …)
  - Computer Science (Compilers; Architecture; Systems; data bases)
  - The real world…
    - From Internet companies to Finance, Legal, Retail,….

- This class: basic Machine Learning algorithms.

# ML: learning concepts through examples

**SPAM**

WINNER!! As a valued network customer you have been selected to receive a $900 prize REWARD! To claim call 09061701461. Claim code KL341. Valid 12 hours ONLY.

SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info

FREE ENTRY in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's

**Not SPAM**

Nah I don't think he goes to USF, he lives around here though

WHAT DID HE SAY??

Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ?

Examples from UCI SMS SPAM dataset https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

# ML: learning concepts through examples

**SPAM**

WINNER!! As a valued network customer you have been selected to receive a $900 prize REWARD! To claim call 09061701461. Claim code KL341. Valid 12 hours ONLY.

URGENT! Your Mobile No 07808726822 has won a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

SIX chances to wi
100 to 20,000 pounds txt>
CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info

FREE ENTRY in 2 a wkly comp to
win FA Cup final tkts 21st May
2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's

I HAVE A DATE ON SUNDAY WITH WILL!!

**Not SPAM**

h I don't think he
es to USF, he lives
ound here though

WHAT DID HE SAY??

Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your
o m ' s l e ft o v e r
nner ? Do you feel my Love ?

Examples from UCI SMS SPAM dataset https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

# ML: learning concepts through examples

**SPAM**

WINNER!! As a valued network customer you have been selected to receive a $900 prize REWARD! To claim call 09061701461. Claim code KL341. Valid 12 hours ONLY.

SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info

FREE ENTRY in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's

**Not SPAM**

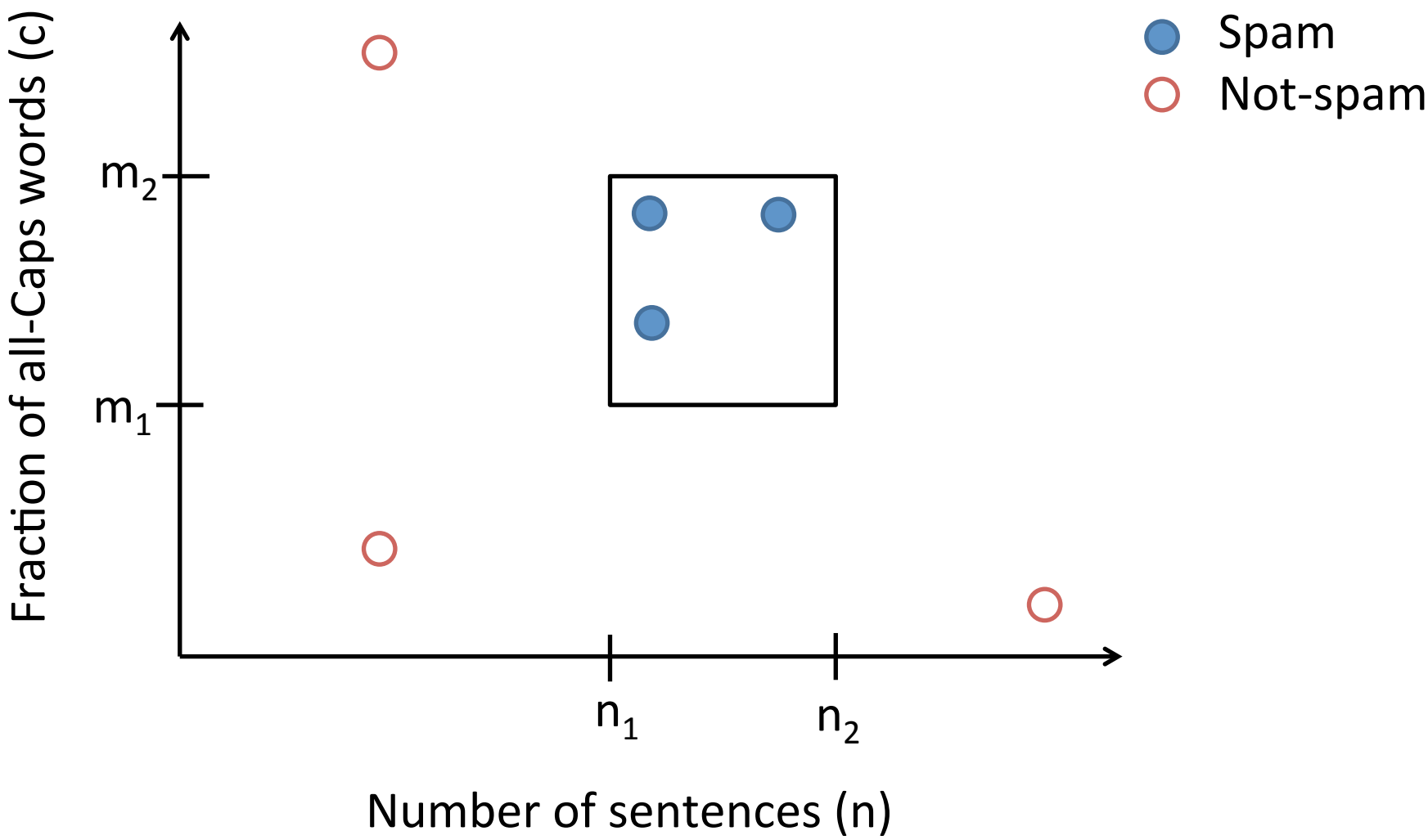
Nah I don't think he goes to USF, he lives around here though

WHAT DID HE SAY??

Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ?

Examples from UCI SMS SPAM dataset https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

If ( ($c>m_1$ & $c<m_2$) & ($n>n_1$ & $n<n_2$) ) then SPAM o/w NOT-SPAM



# Key Terms

- Supervised classification problem
- Labeled training set
  - instances and labels
  - Positive examples and negative examples

# Key Terms

- Step 1: Identify useful properties of instances – Representation/Feature Extraction
- Step 2: Learn a hypothesis (*rules* or *functions* that define each class/label) – Training/Learning
- Step 3: Given a new instance, use the above classifier to predict a label – Testing/Prediction

# Step 1: Feature Extraction

- Identify useful properties of instances – Representation/ Feature extraction
- Most common representation is representing an instance as a *vector* of features (**x**)

WINNER!! As a valued network customer you have been selected to receive a $900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.

Fair enough, anything going on?

**x** **y**

| All-caps words | Has numbers | Call/ Text | Label |
|---|---|---|---|
| yes | yes | yes | spam |
| no | no | no | not-spam |

# Step 1: Feature Extraction

- Identify useful properties of instances – Representation/ Feature extraction
- Most common representation is representing an instance as a *vector* of features (**x**)

Bold fonts to represent vectors

**Designing good features is VERY important in Machine Learning**

WINNER!! As a valued network customer you have been selected to receive a $900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.

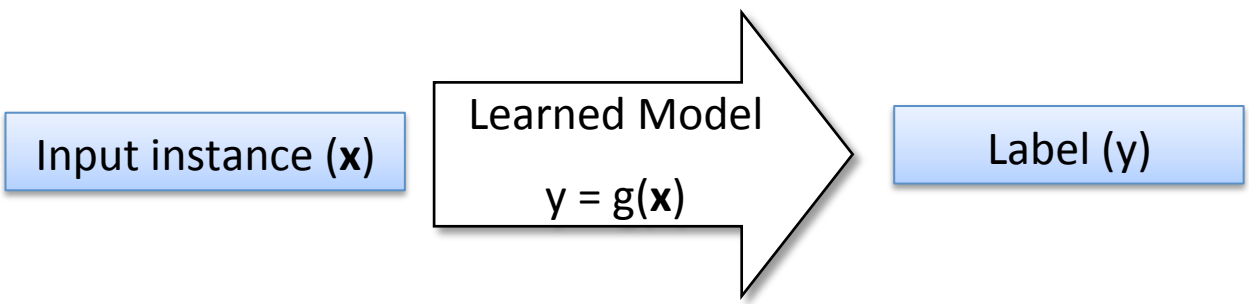
Fair enough, anything going on?

**x** y

| **All-caps words** | **Has numbers** | **Call/ Text** | **Label** |
|---|---|---|---|
| yes | yes | yes | spam |
| no | no | no | not-spam |

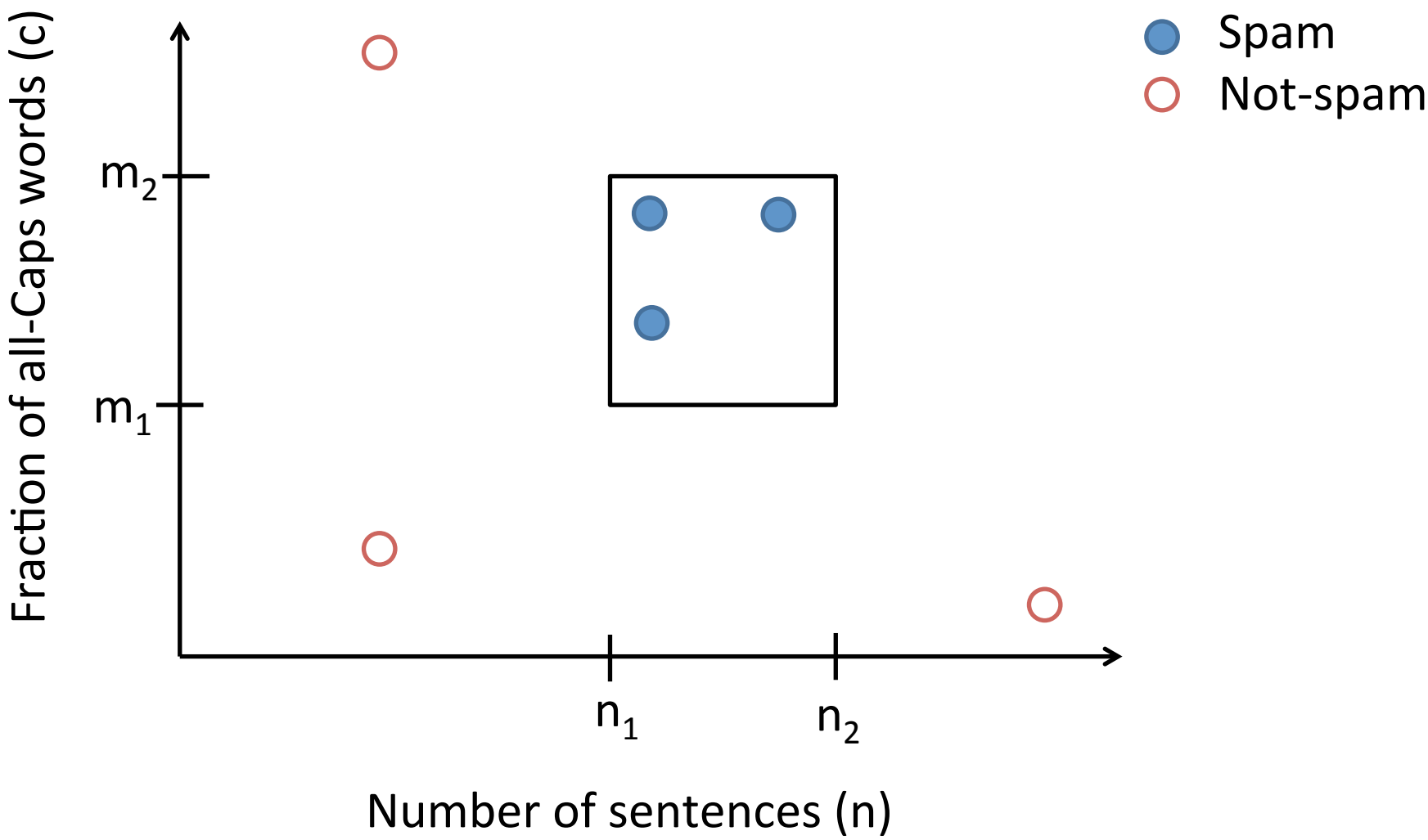
# Step 2: Training

- Choose a hypothesis class
- Learn a hypothesis (*rules* or *functions* that define each class/label) – Training/Learning
- The hypothesis is defined by parameters and learned by a classifier/model




- How to train a classifier to learn g(x)? (later)

If ( ($c>m_1$ & $c<m_2$) & ($n>n_1$ & $n<n_2$) ) then SPAM o/w NOT-SPAM



# Step 3: Testing

- Given a **new** instance, use the above classifier to predict a label – Testing/Prediction/Inference

$y' = g(\mathbf{x})$

- Generalization: The ability to perform a task in a situation which has never been encountered before

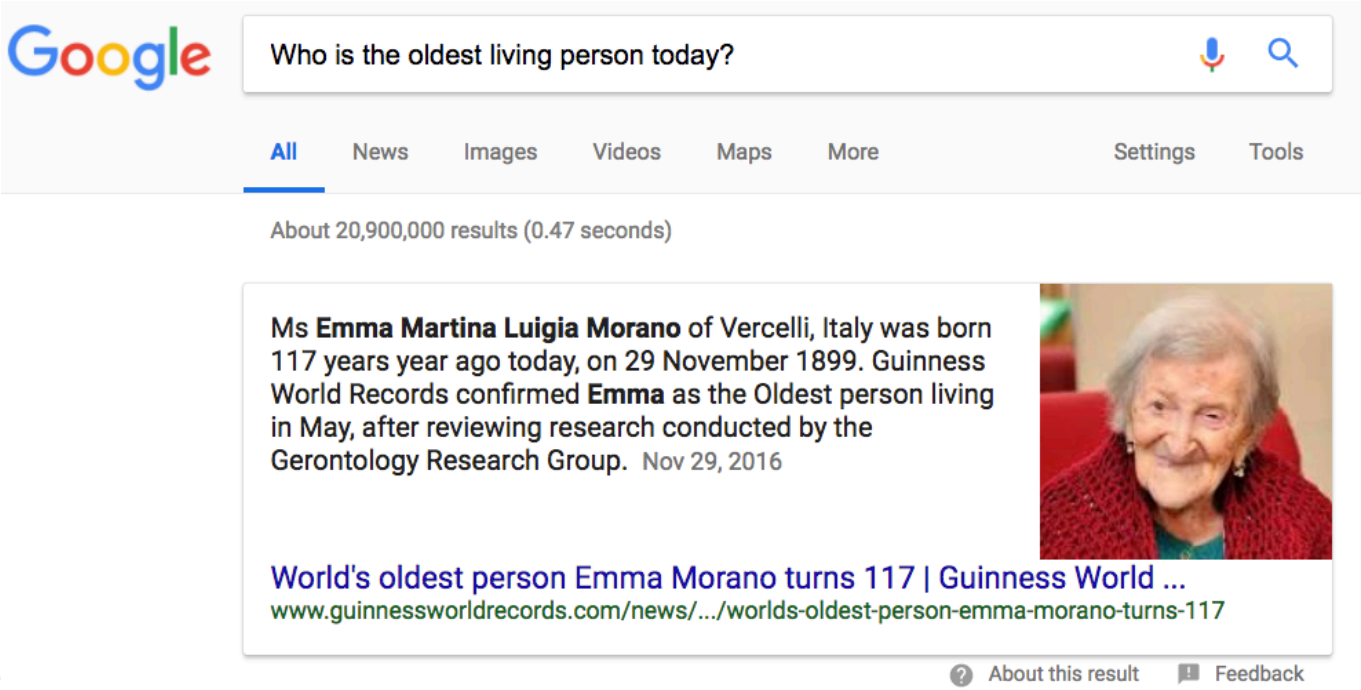
I HAVE A DATE ON SUNDAY WITH WILL!!

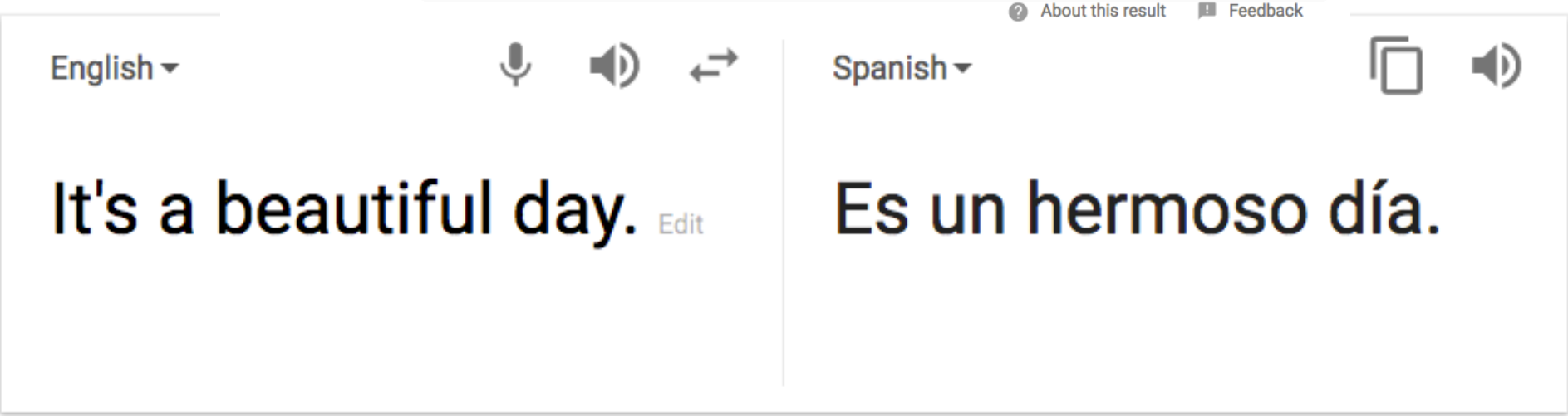
What does the algorithm get as input? (features)

Generalization depends on the Representation as much as it depends on the Classification Algorithm used.

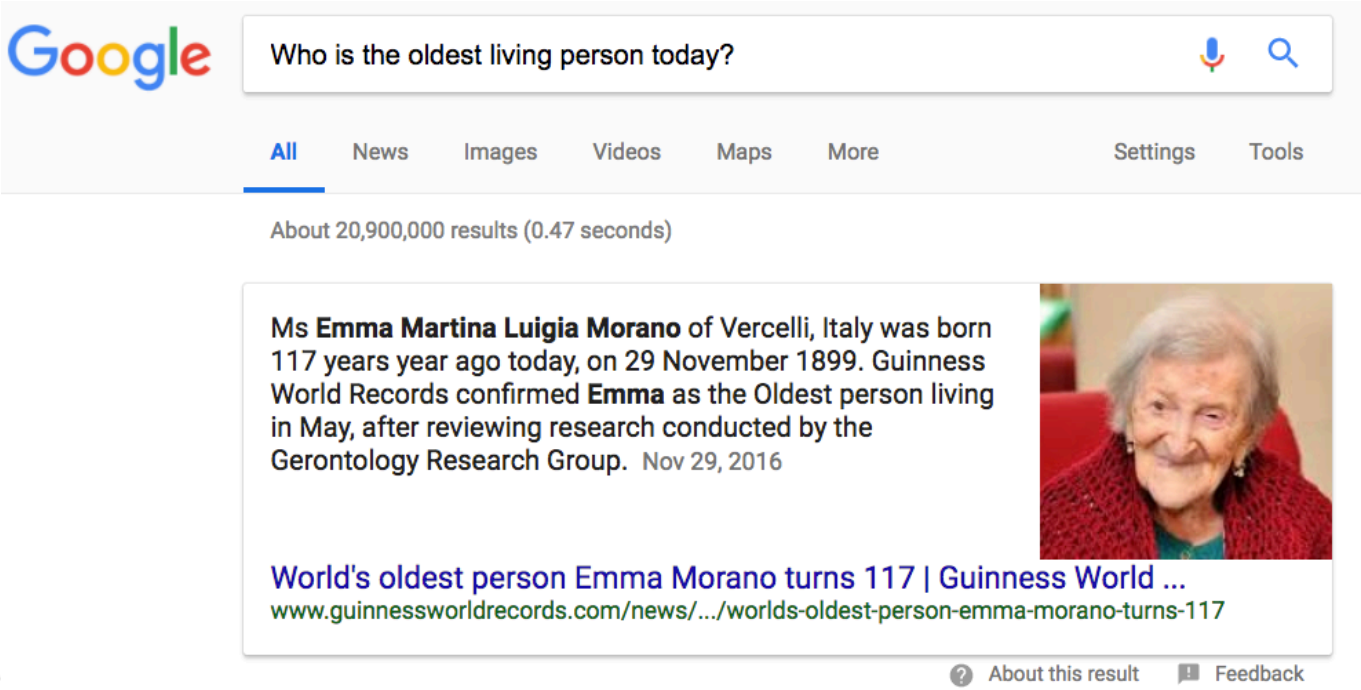
- Test instances should come from the same population

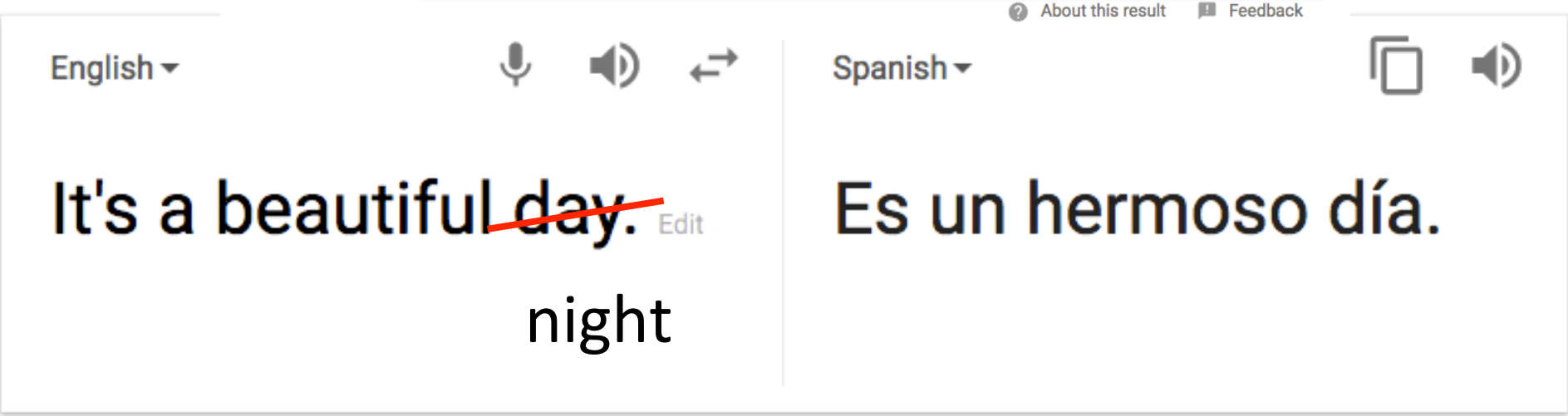
# What does same population mean?



# What does same population mean?



# Same population -- iid assumption

- Distribution of instances (**x**) and labels (y) defines some unknown (but fixed) P(**x**,y)
- Assumption in ML: all training as well as test instances (and their labels) are independent and identically distributed (iid)
- Independent and identically distributed (iid): A collection of random variables is iid if they all have the same probability distribution, and all are mutually independent
  - All training and test instances are sampled from the same P(**x**,y)