DS 4400:
Machine Learning and Data Mining I

Ehsan Elhamifar

eelhami@ccs.neu.edu

Slides adapted from a variety of sources including A. Ng, P. Domingos, K. Murphy…

Logistics

- **Instructor: Ehsan Elhamifar**
  - Email: eelhami@ccs.neu.edu
  - Office hours: Fridays, 4:30pm—5:15pm, 310E WVH

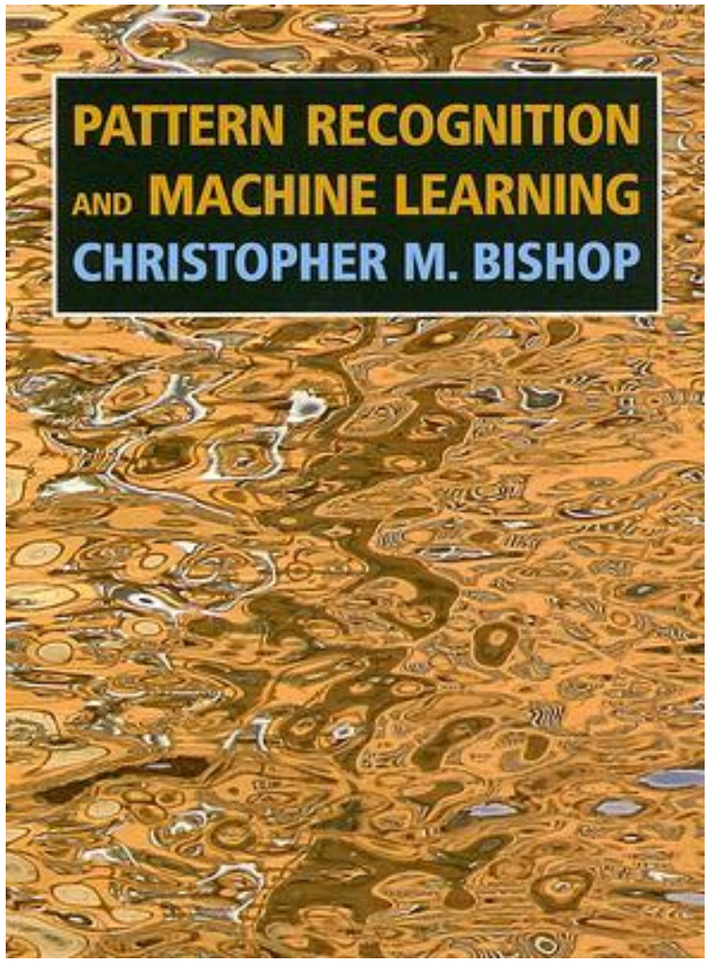
- **Teaching Assistant:**
  - Rebecca McBrayer: mcbrayer.r@husky.neu.edu
  - Likith Ponnanna patrapandabelliapp.l@husky.neu.edu
  - Time/location: TBA

- **Course page:** http://khoury.neu.edu/home/eelhami/courses_nu_ds4400_spring20.htm

- **Discussions, HWs, Solutions, Announcements: Piazza**

Optional Textbooks

- **Pattern recognition and machine learning, Christopher Bishop, Springer 2007**



- **Machine Learning: A Probabilistic Perspective, Kevin Murphy, MIT Press 2013**

- **Stanford CS229**: http://cs229.stanford.edu/syllabus.html

Grading

- **Homeworks** (40% of the grade)
  - ~4 HWs: analytical + programming assignments (Python)
  - **No late** HWs, submit at the **beginning** of class, done **individually**

Grading

- **Homeworks** (40% of the grade)
  - ~4 HWs: analytical + programming assignments (Python)
  - **No late** HWs, submit at the **beginning** of class, done **individually**
- **Project** (20% of the grade)
  - **Project teams and proposal**: **March 10, 11:59pm**; teams of <= 2
  - **Project report**: **April 25, 11:59pm**

Grading

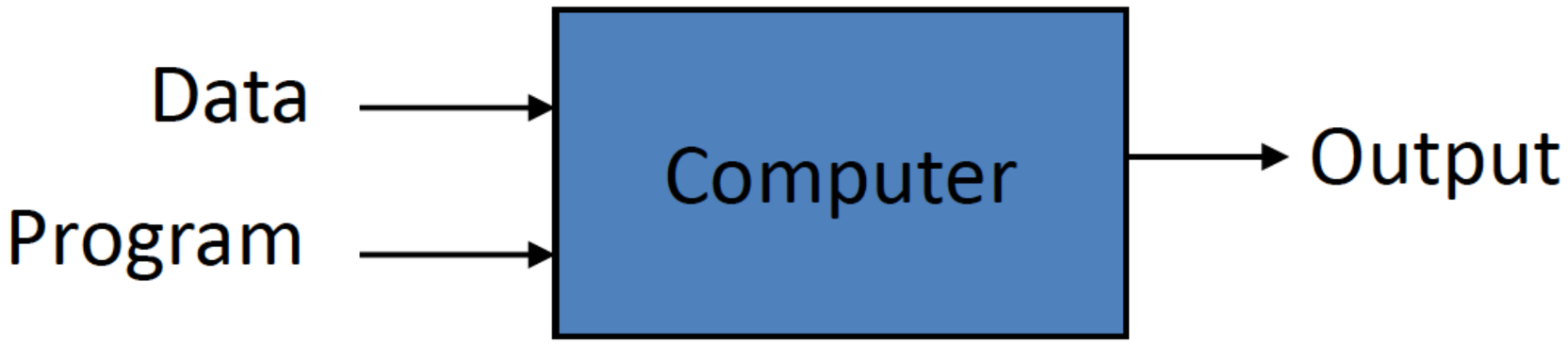
- **Homeworks** (40% of the grade)
  - ~4 HWs: analytical + programming assignments (Python)
  - **No late** HWs, submit at the **beginning** of class, done **individually**
- **Project** (20% of the grade)
  - **Project teams and proposal**: **March 10, 11:59pm**; teams of <= 2
  - **Project report**: **April 25, 11:59pm**
- **Midterm 1 and 2** (40% of the grade)
  - **Friday, February 28**, 1 cheat sheet
  - **Tuesday, April 14**, 1 cheat sheet

Grading

- **Homeworks** (40% of the grade)
  - ~4 HWs: analytical + programming assignments (Python)
  - **No late** HWs, submit at the **beginning** of class, done **individually**
- **Project** (20% of the grade)
  - **Project teams and proposal**: **March 10, 11:59pm**; teams of <= 2
  - **Project report**: **April 25, 11:59pm**
- **Midterm 1 and 2** (40% of the grade)
  - **Friday, February 28**, 1 cheat sheet
  - **Tuesday, April 14**, 1 cheat sheet
- **Class participation** (-5% to +5%)

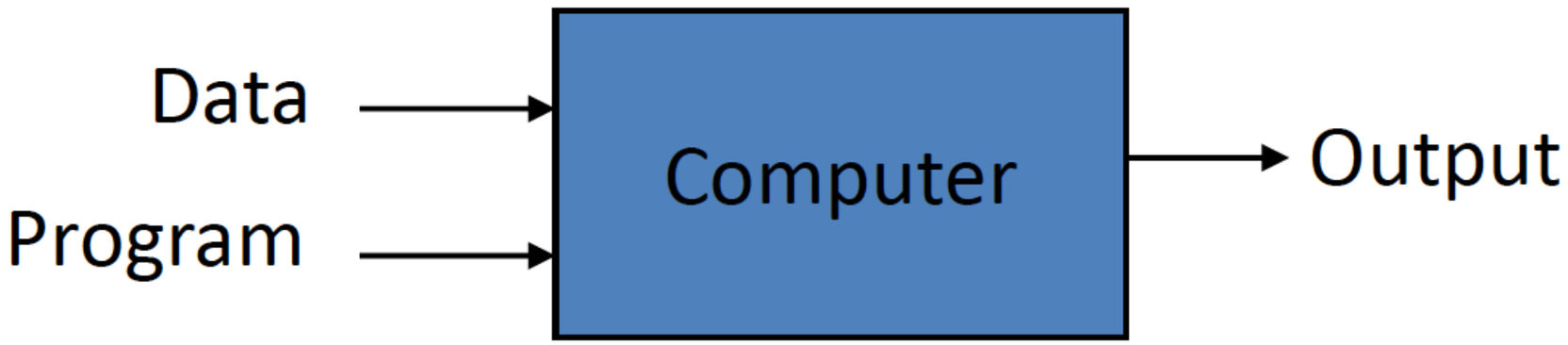
What is machine learning?

- Traditional Programming

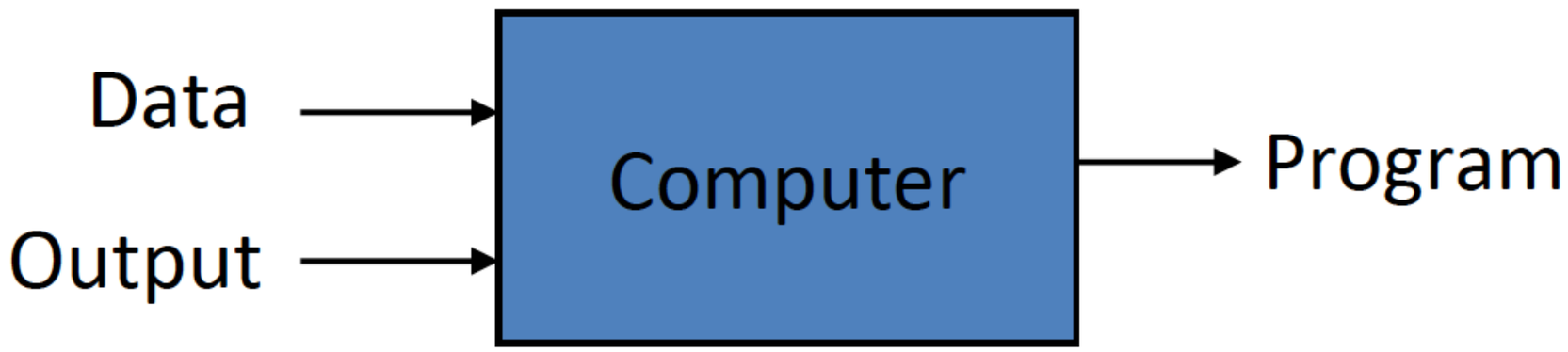


What is machine learning?

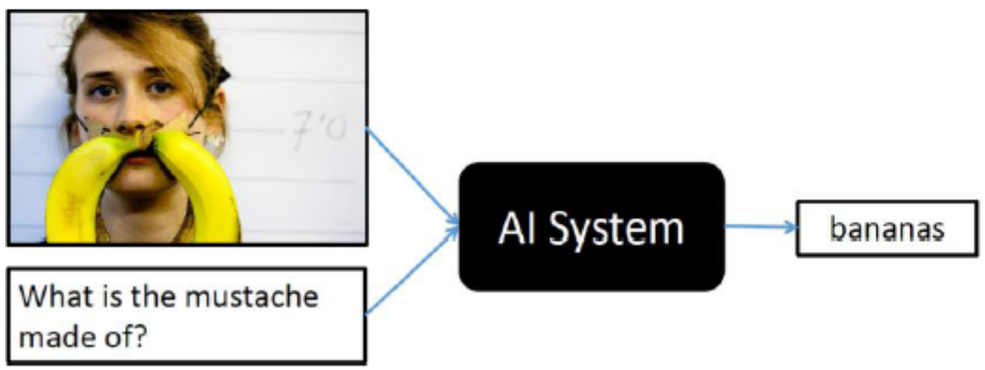
- Traditional Programming

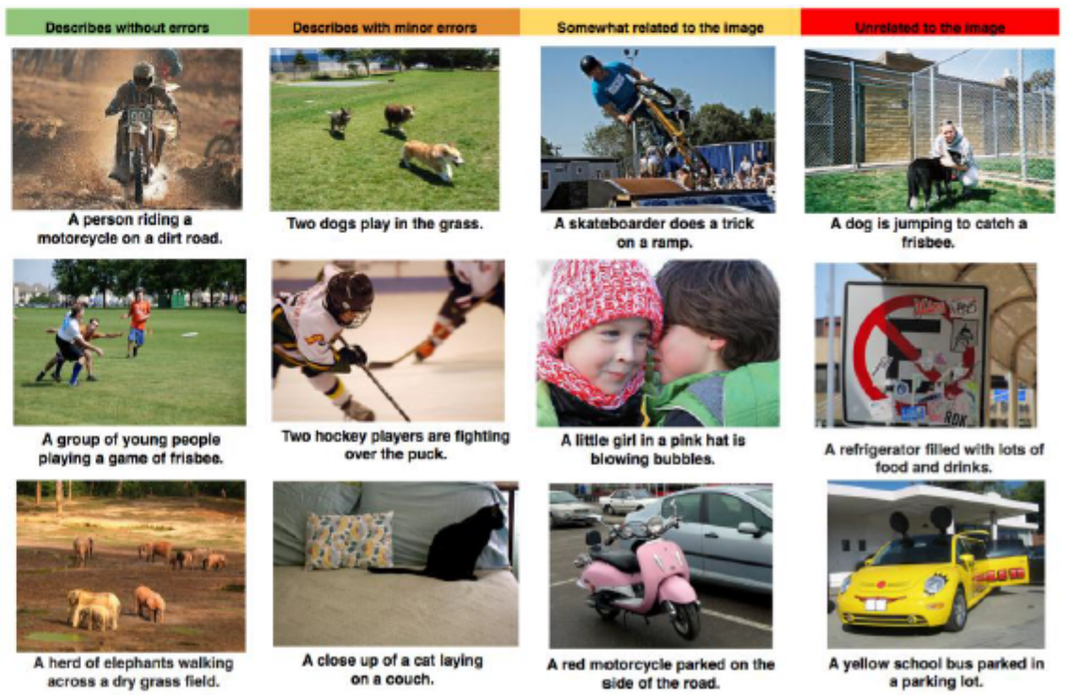


- **Machine Learning**



What can ML do?





What is Machine Learning ?
(by examples)

Classification

(from data to discrete classes)

Spam filtering

data

prediction

Osman Khan to Carlos show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
> Let's try to chat on Friday a little to coordinate and more on Sunday in person?
>
> Carlos

Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Michel show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Spam

vs

Not Spam

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | x

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: anç show details 9:52 PM (1 hour ago) Reply

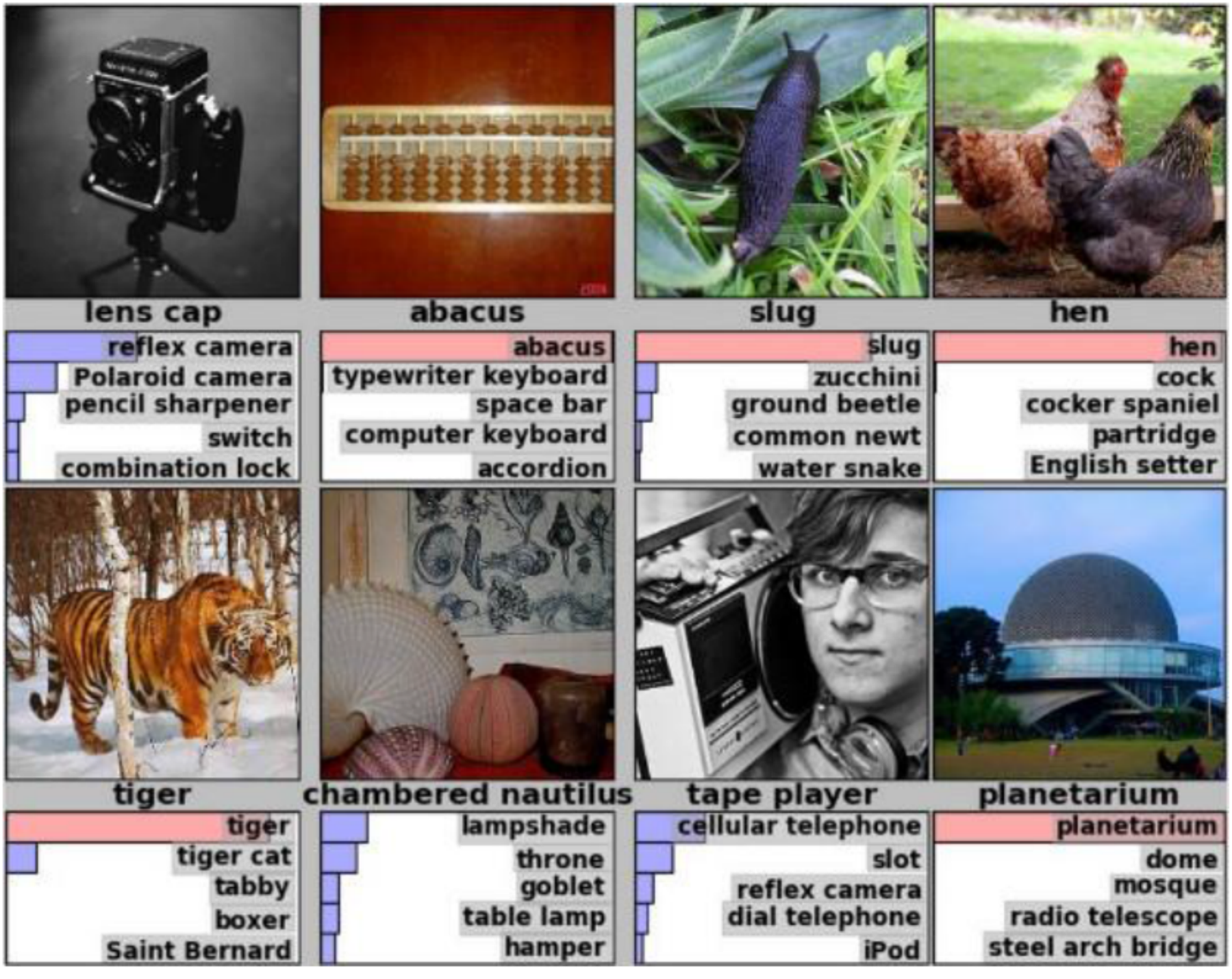
=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

Object classification



Weather prediction

Regression

(predicting numeric values)

Stock market prediction

Weather prediction (revisited)

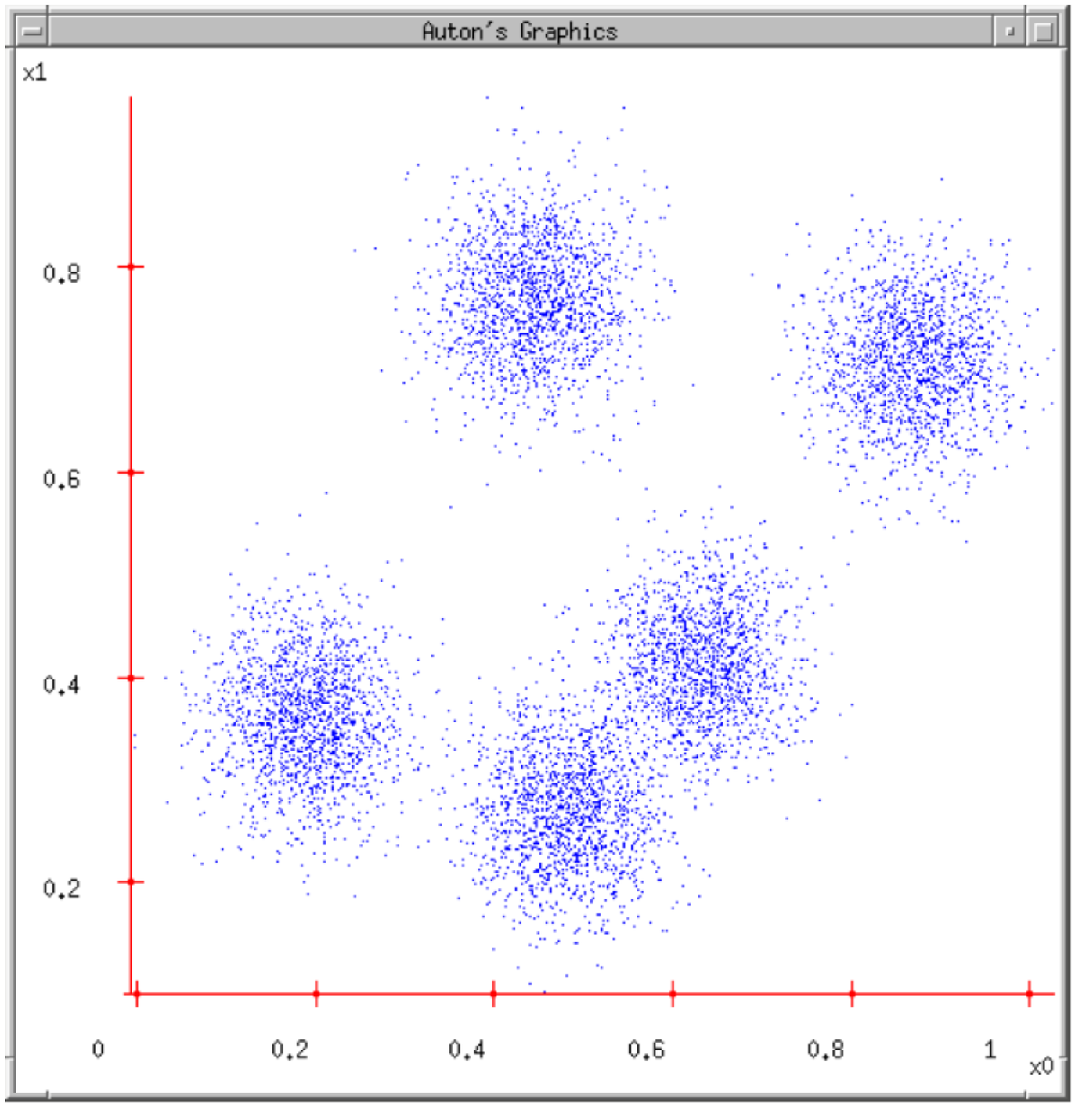
Temperature

72° F

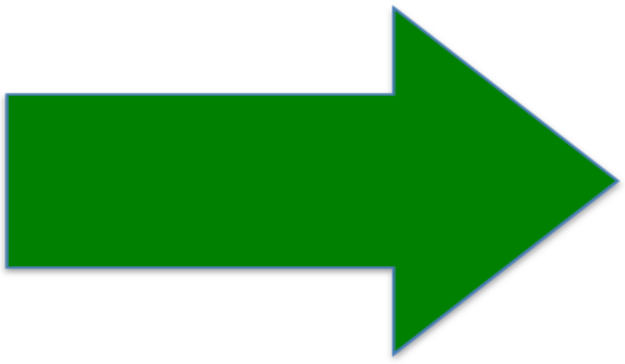
Clustering

(discovering structure in data)

Clustering: group similar items



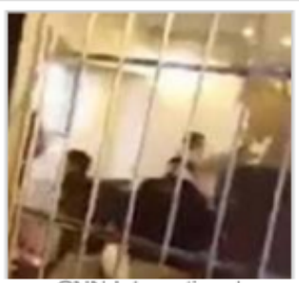
Clustering images



Clustering news

U.S. edition | Modern

Top Stories



CNN International

See realtime coverage

Saudi execution of Shia cleric threatens to deepen regional sectarian crisis

CNN International - 3 hours ago

(CNN) Sheikh Nimr al-Nimr was not among the "A-list" of Shia clerics in Saudi Arabia. But his execution has provoked a regional crisis, sparking condemnation from Iraq, Iran and even senior U.N.

Oil Rises in Asia Due to Iran-Saudi Arabia Tensions Wall Street Journal

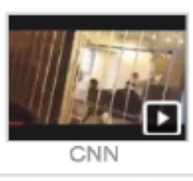
A reckless regime Washington Post

Highly Cited: Iranian Protesters Ransack Saudi Embassy After Execution of Shiite Cleric New York Times

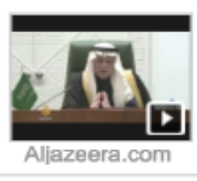
From Saudi Arabia: Saudi Arabia severs Iran ties Arab News

Wikipedia: Nimr al-Nimr

Related
Saudi Arabia »
Sheikh Nimr »
Iran »



CNN



Aljazeera.com

YouTube

Washingt...

Armed activists in Oregon touch off unpredictable chapter in land-use feud

Washington Post - 2 hours ago

BURNS, Ore. - An unpredictable new chapter in the wars over federal land use in the West unfolded Sunday after a group of armed activists split off from an earlier protest march and occupied a national wildlife refuge in remote southeastern Oregon.

Firstpost

One dead as 6.8 magnitude quake strikes eastern India - police

Reuters - 1 hour ago

GUWAHATI, India At least one person was killed and a dozen injured when an earthquake measuring 6.8 struck near Imphal in eastern India on Monday, sending people running from their homes and knocking out power to the city near the Myanmar border.

CBS News

ISIS threatens UK in new execution video

CBS News - 5 hours ago

BEIRUT -- A video circulated online Sunday purported to show the Islamic State of Iraq and Syria (ISIS) killing five men accused of spying for Britain in Syria.

Press He...

NTSB releases haunting video of El Faro wreckage on ocean floor

Press Herald - 23 minutes ago

The merchant ship carrying 33 crew members, including four from Maine, sank off the Bahamas last fall. By Dennis Hoey Staff Writer.

The Bost...

In NH, Clinton hits on opioid abuse as a top concern

The Boston Globe - 2 hours ago

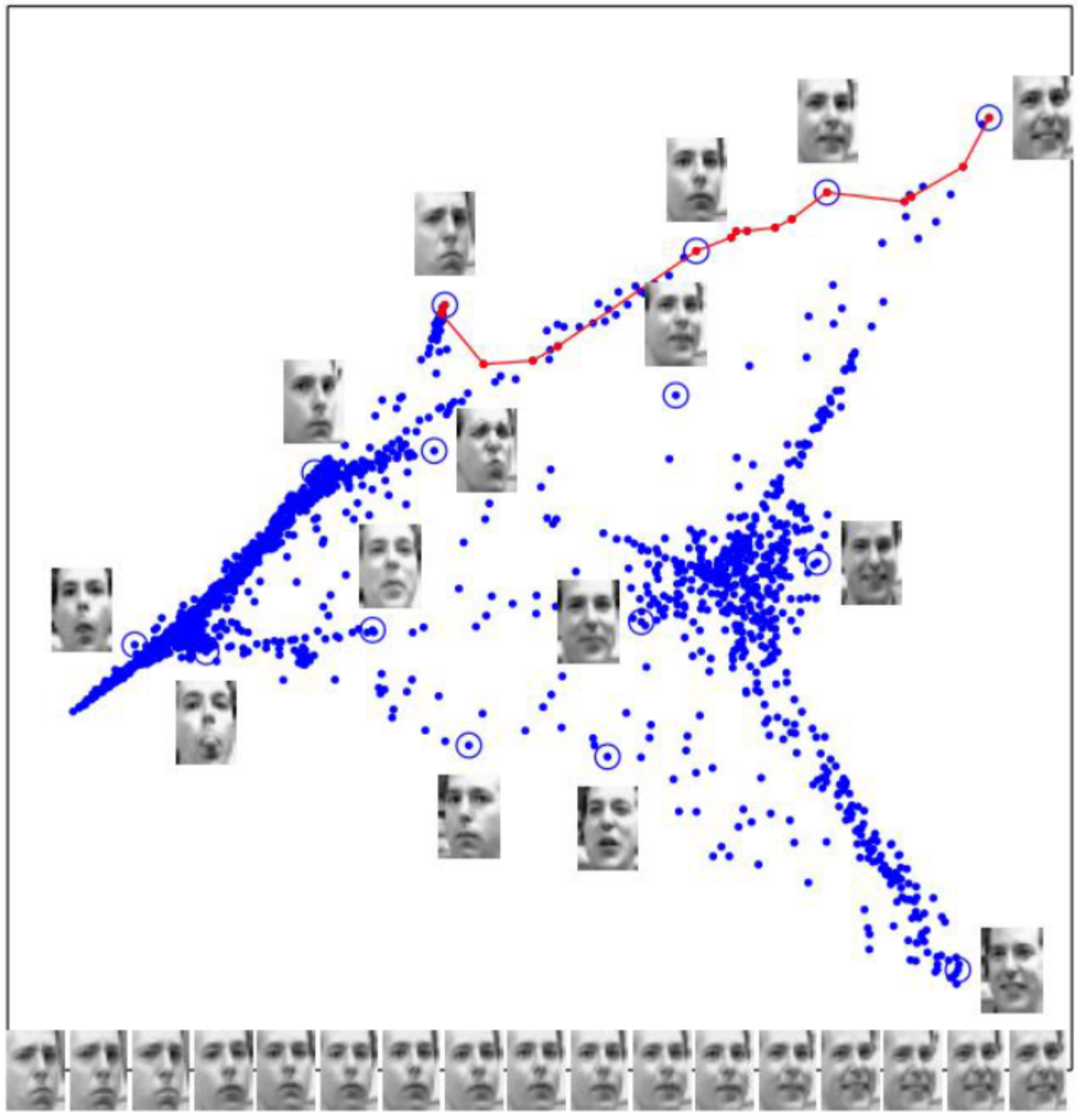
DERRY, N.H. - Hillary Clinton, who arrived to loud applause here at one of three New Hampshire campaign stops Sunday, said prohibitively expensive education, lack of support for families coping with Alzheimer's disease, and the rising tide of opioid ...

Dimensionality reduction

(visualizing data)

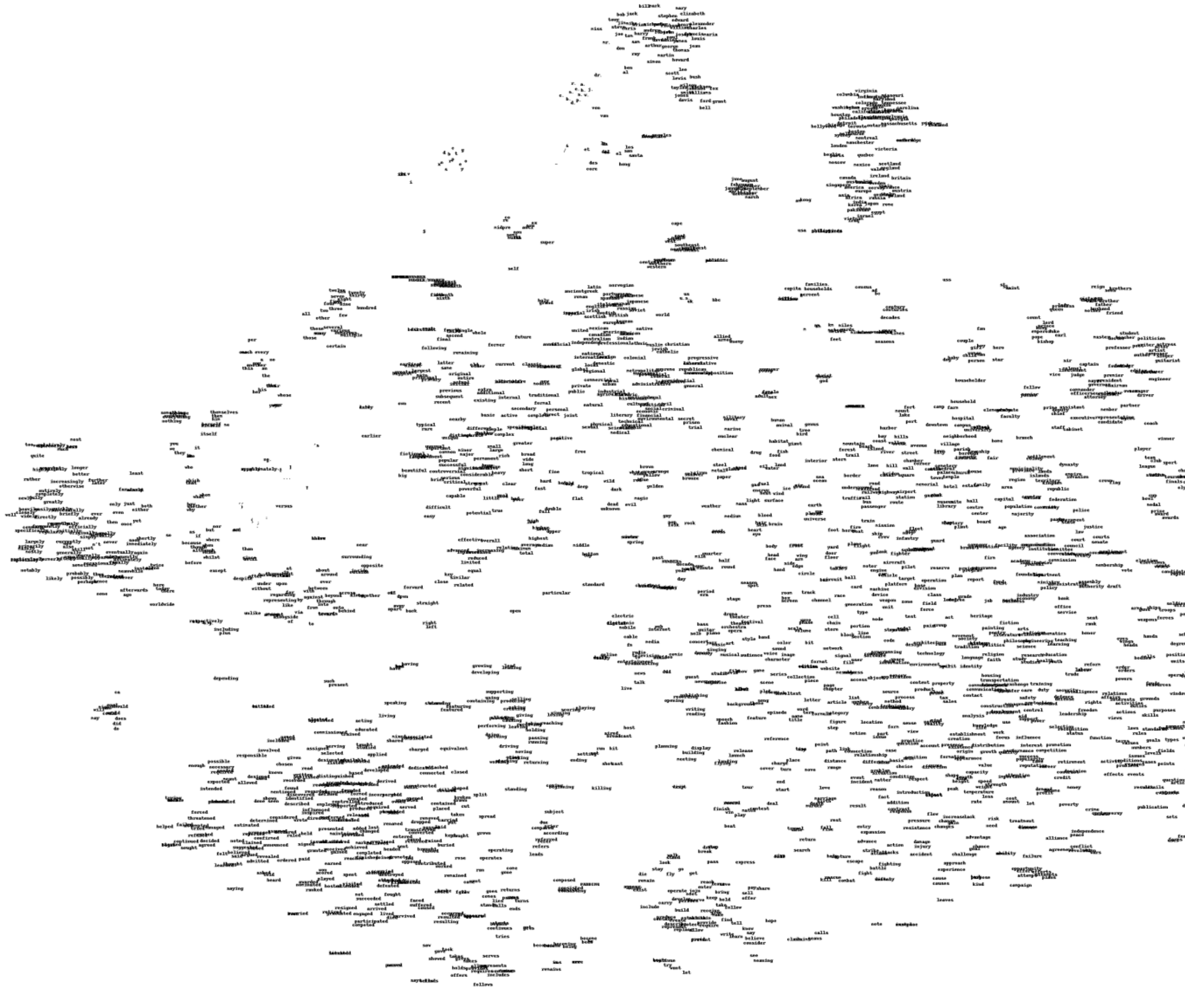
Image embedding

- Images have millions/ billions of pixels
- Can we give each image a few coordinates so that similar images are near each other?

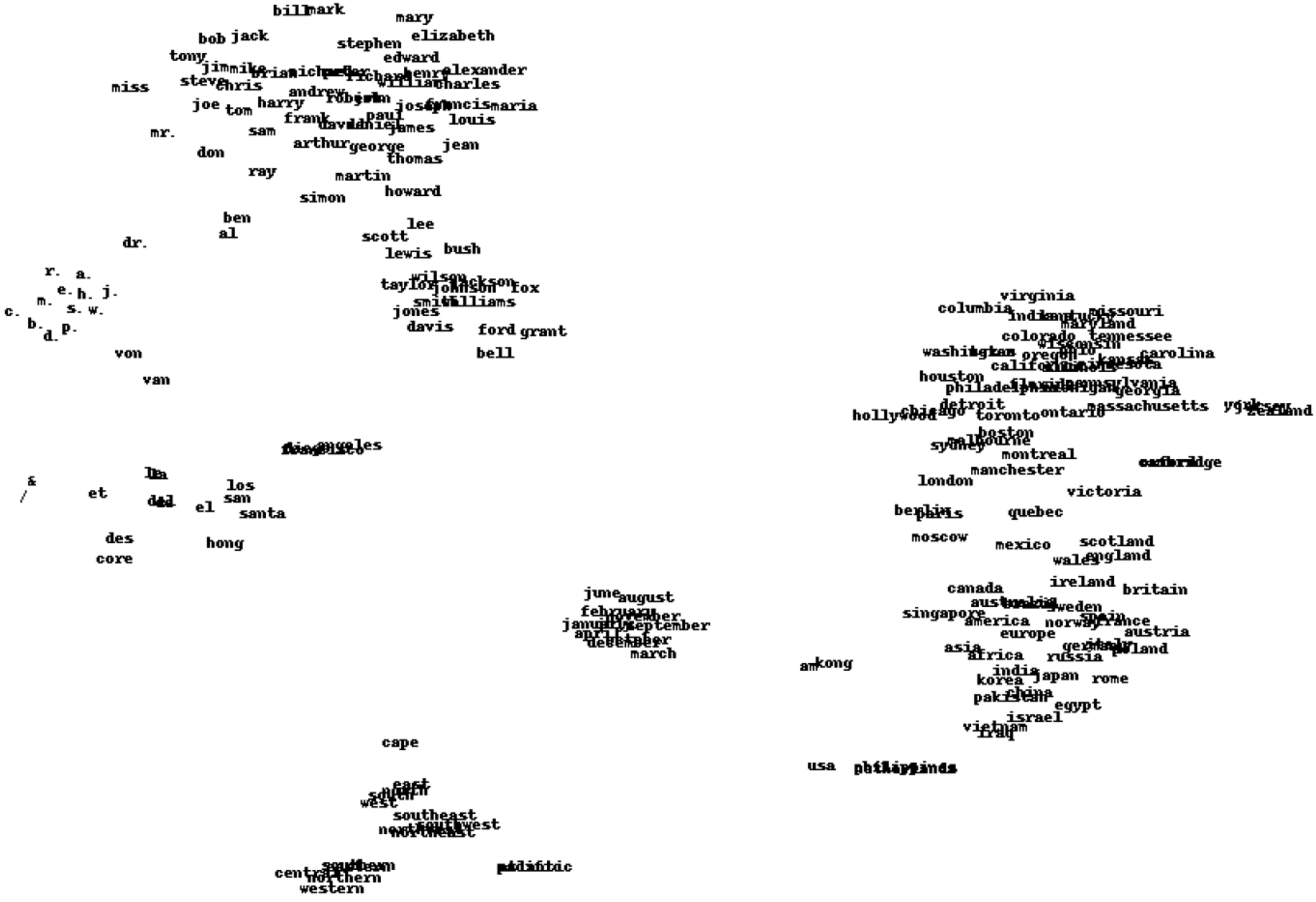


Roweis & Saul ‘00

Word embedding



Word embedding (zoom in)



Machine learning types

- **Supervised** learning
  - **Regression**
  - **Classification**
- **Unsupervised** learning
  - **Clustering**
  - **Dimensionality reduction**
  - **Matrix completion**
- **Reinforcement** learning
  - Learning by **weak supervision**: reward function

Supervised learning

- **Given**: training data $\{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, N\}$
- **Find**: a good approximation to $f : \ \mathcal{X} \to \mathcal{Y}$

Model

- Examples:
  - Spam detection: $\mathcal{X}$ = email, $\mathcal{Y}$ = spam/ham
  - Digit recognition: $\mathcal{X}$ = digit images, $\mathcal{Y}$ = {0,1,2, …, 9}
  - Stock prediction: $\mathcal{X}$ = new/historic prices, etc, $\mathcal{Y}$ = real numbers

Example: spam filter

- Input: email
- Output: spam/ham
- Setup:
  - Get a large collection of example emails, each labeled “spam” or “ham”
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future emails

- Features: The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: CAPS
  - Non-text: SenderInContacts
  - …

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Example: digit classification

- Input: images / pixel grids
- Output: a digit 0-9
- Setup:
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new digit images

0

1

2

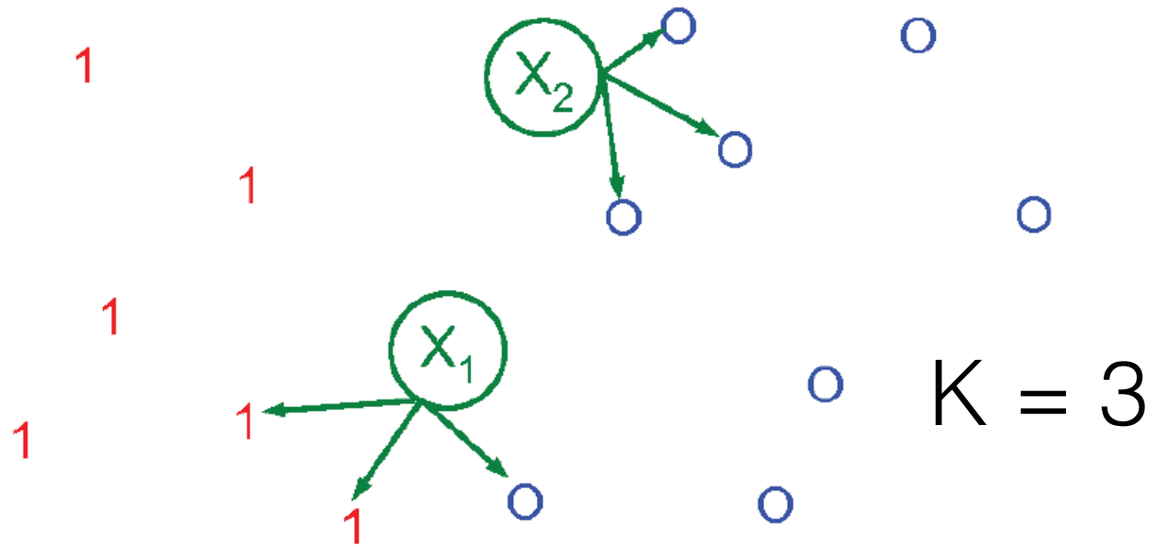
1

??

- Features: The attributes used to make the digit decision
  - Pixels: (6,8)=ON
  - Shape Patterns: NumComponents, AspectRatio, NumLoops
  - …

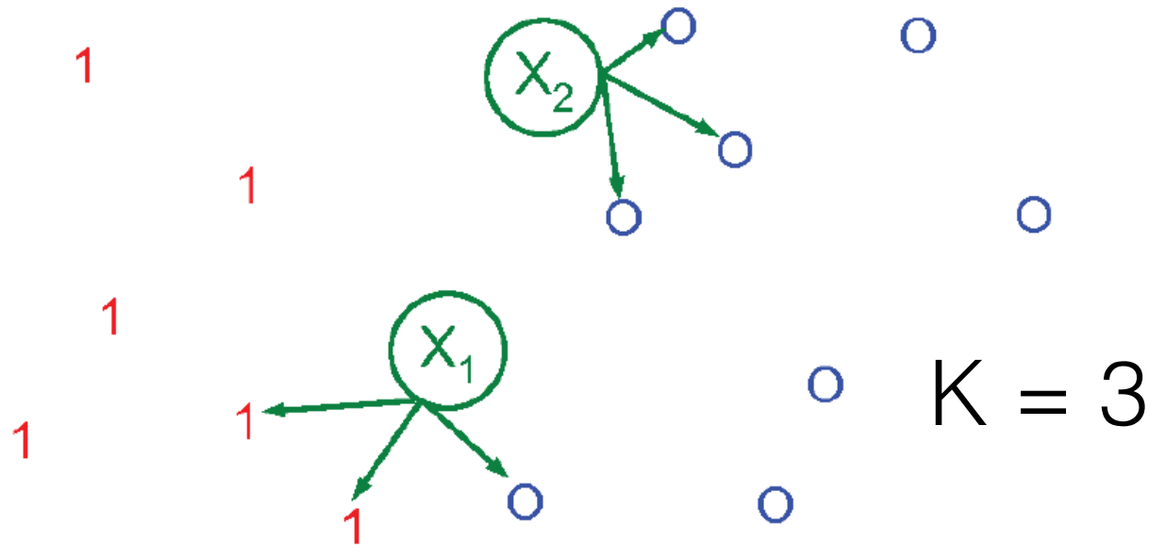
KNN classifier

- Assign the test sample to the class with largest #votes

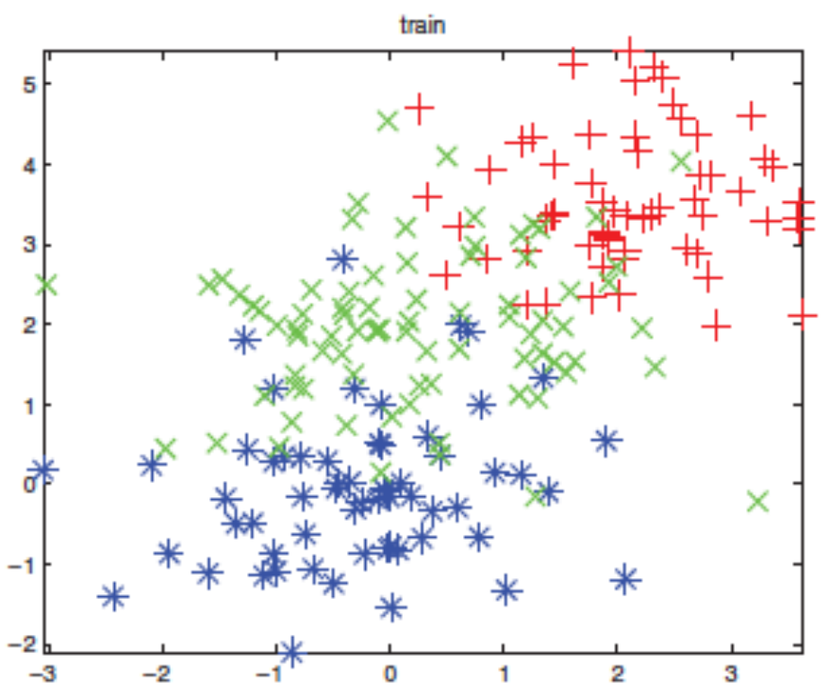


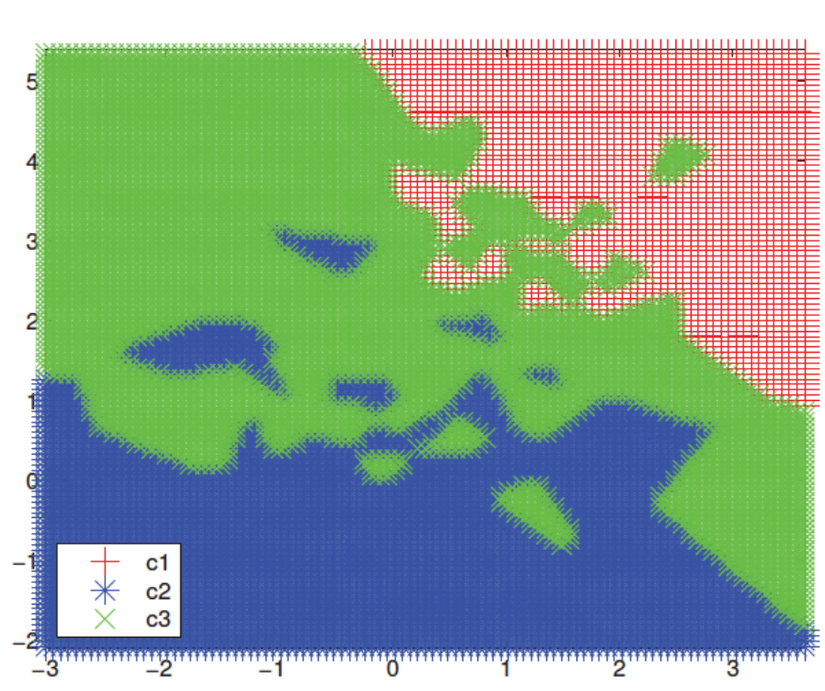
KNN classifier

- Assign the test sample to the class with largest #votes

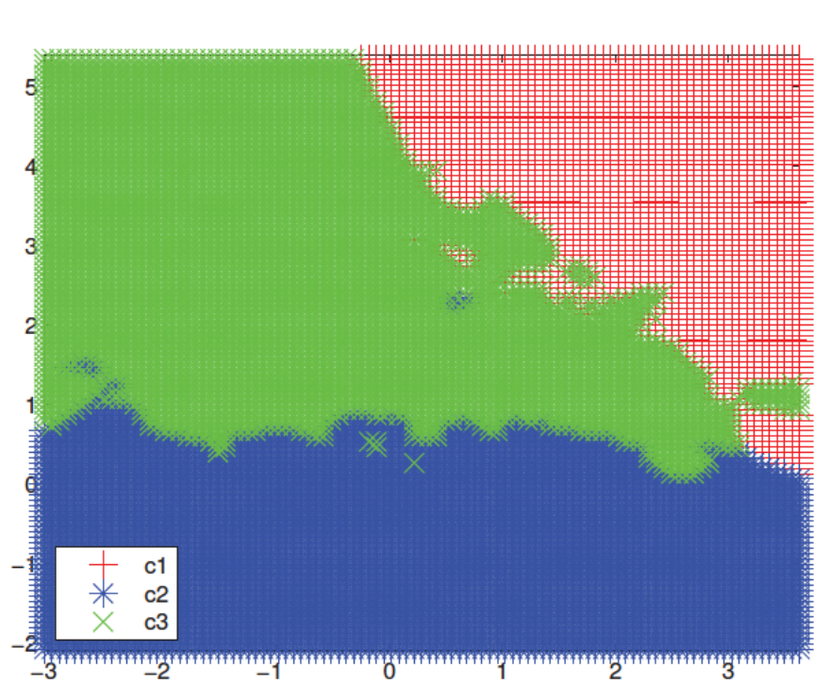


- How to choose K?





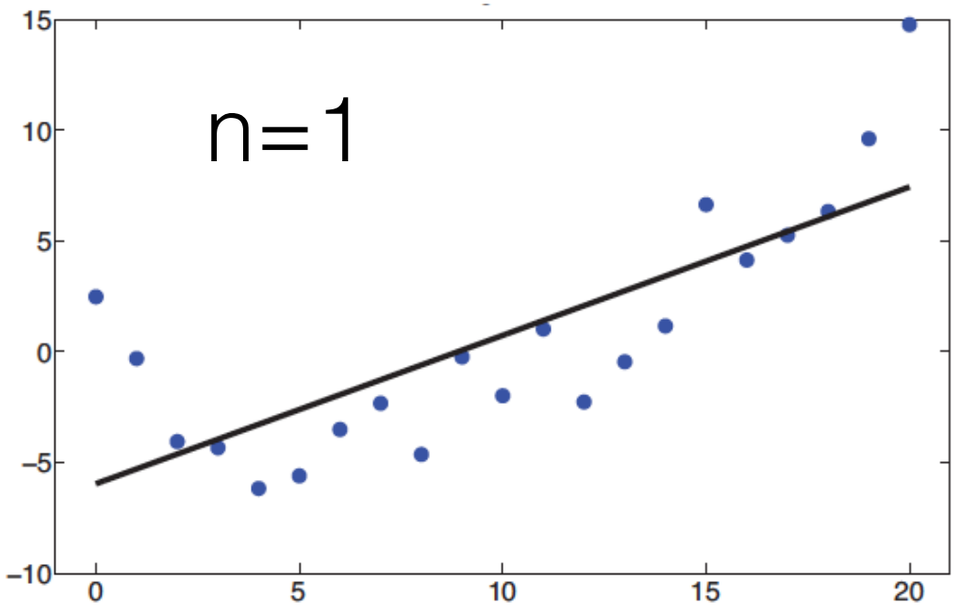
K = 1



K = 5

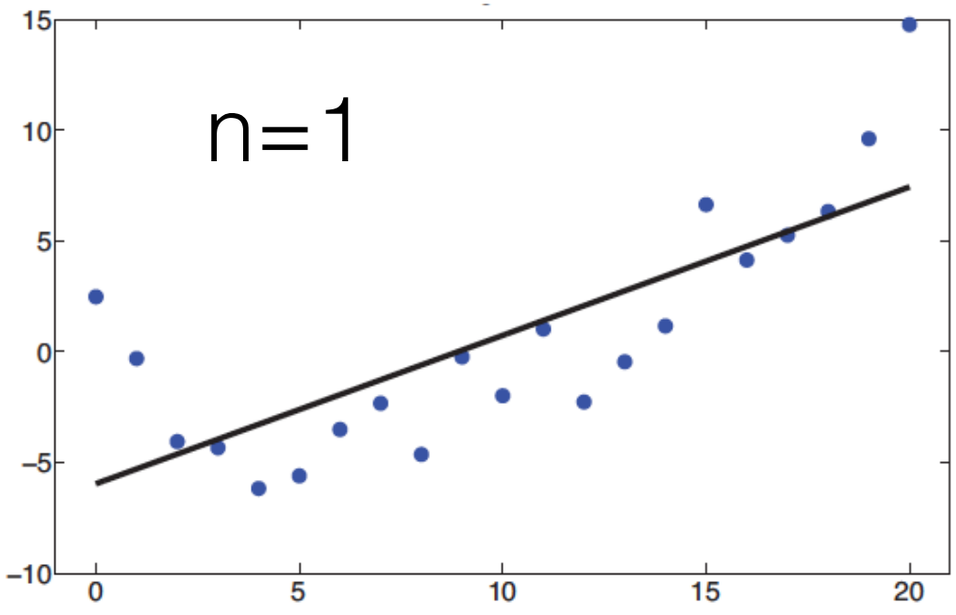
Regression

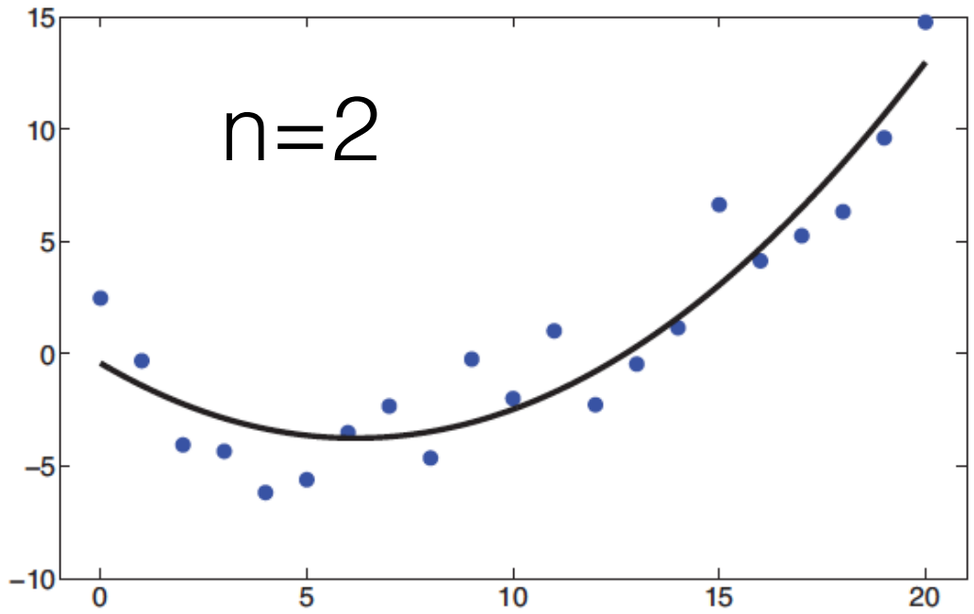
- **What degree** polynomial (n) to fit to data?



Regression

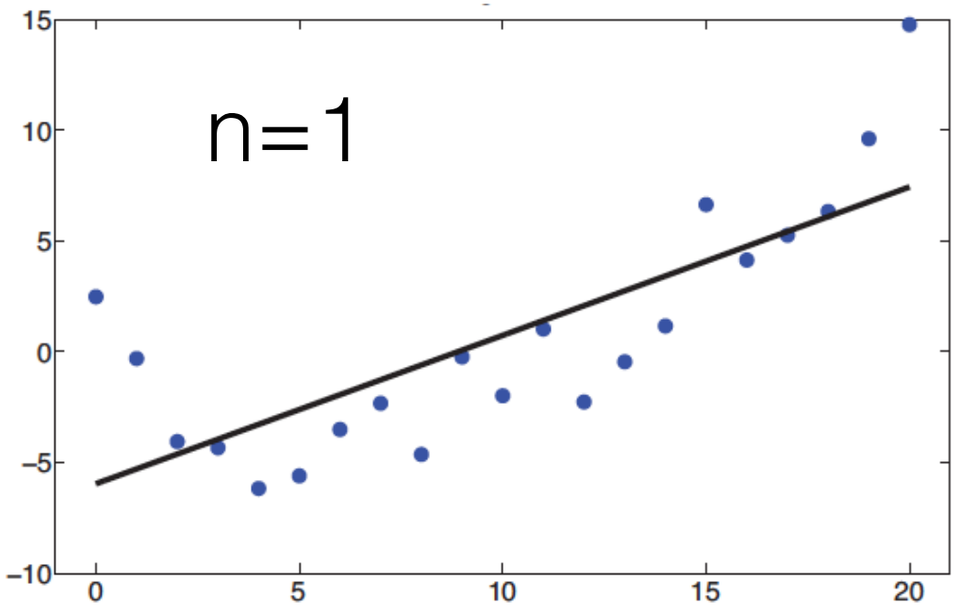
- **What degree** polynomial (n) to fit to data?

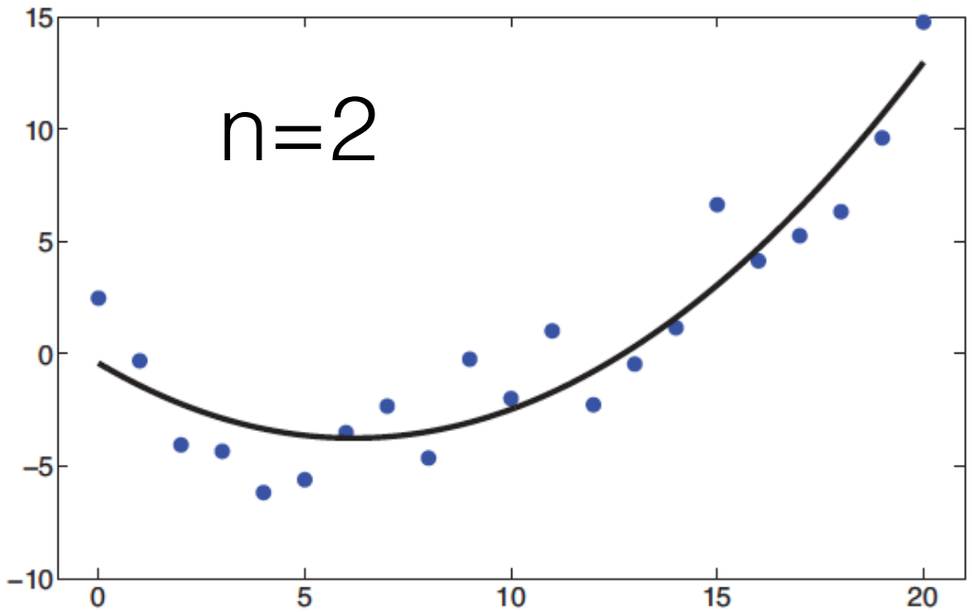


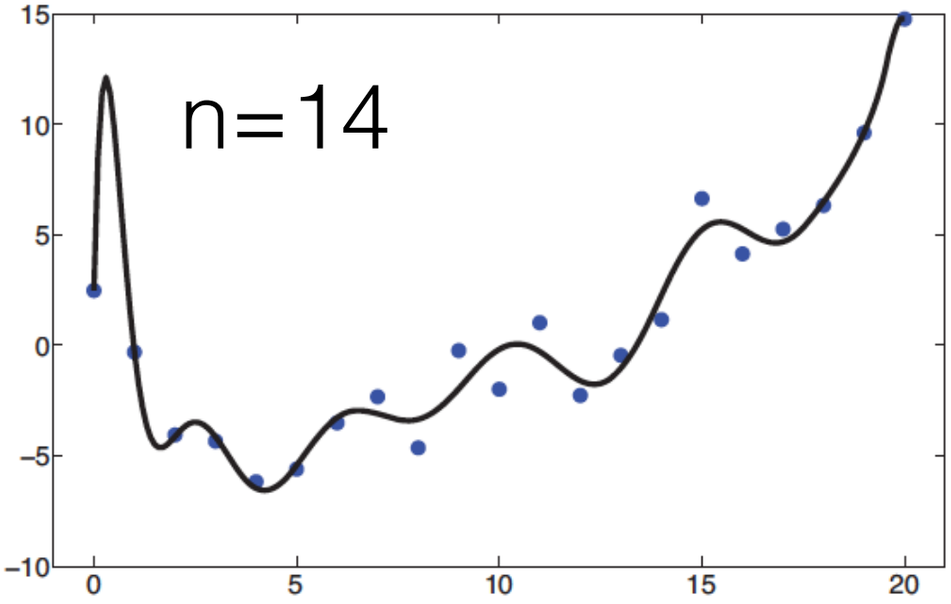


Regression

- **What degree** polynomial (n) to fit to data?

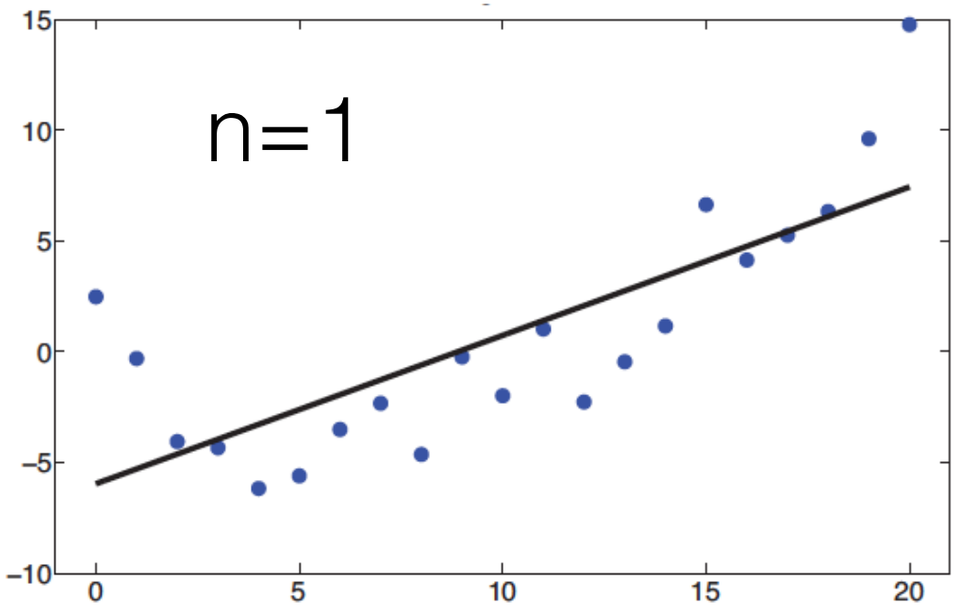


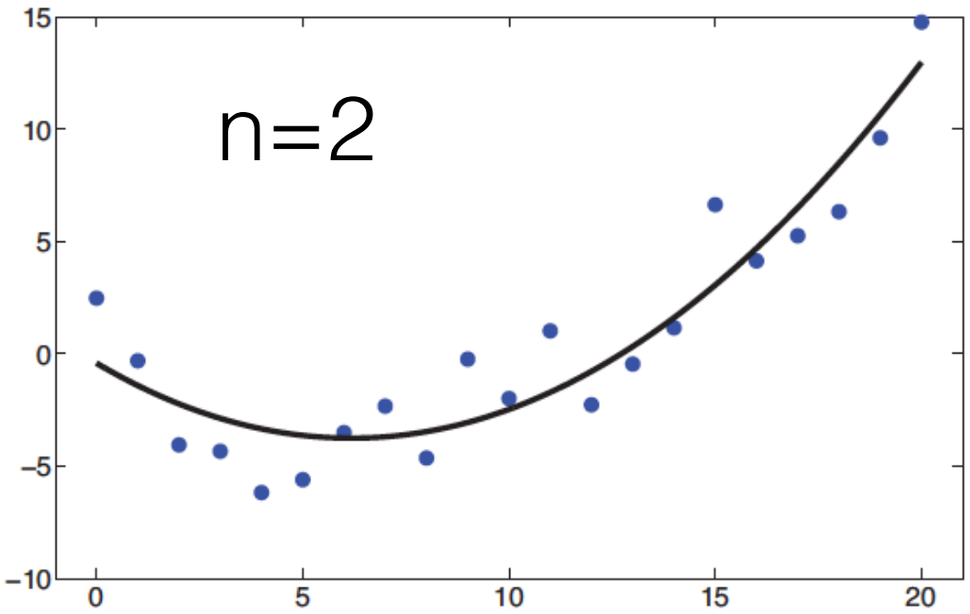


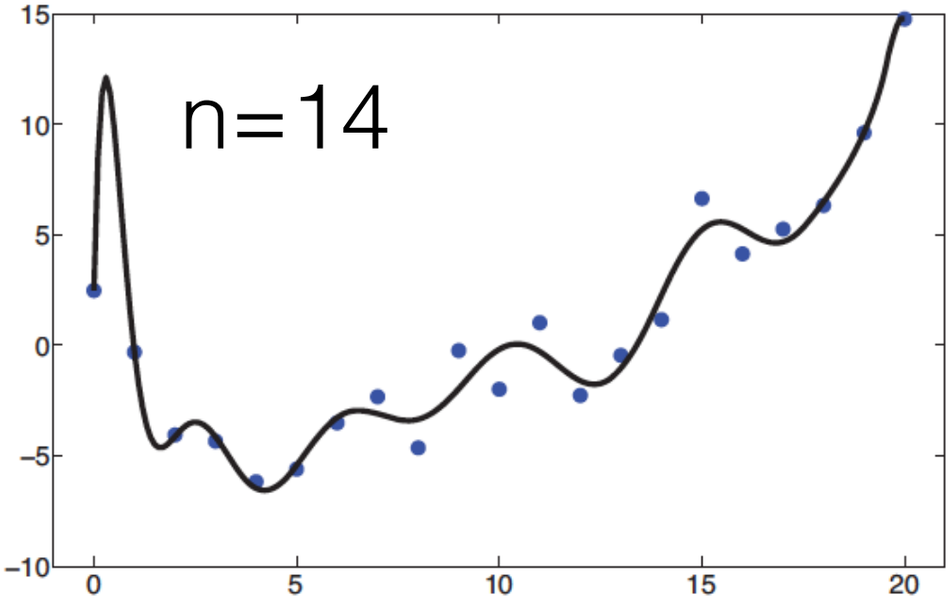


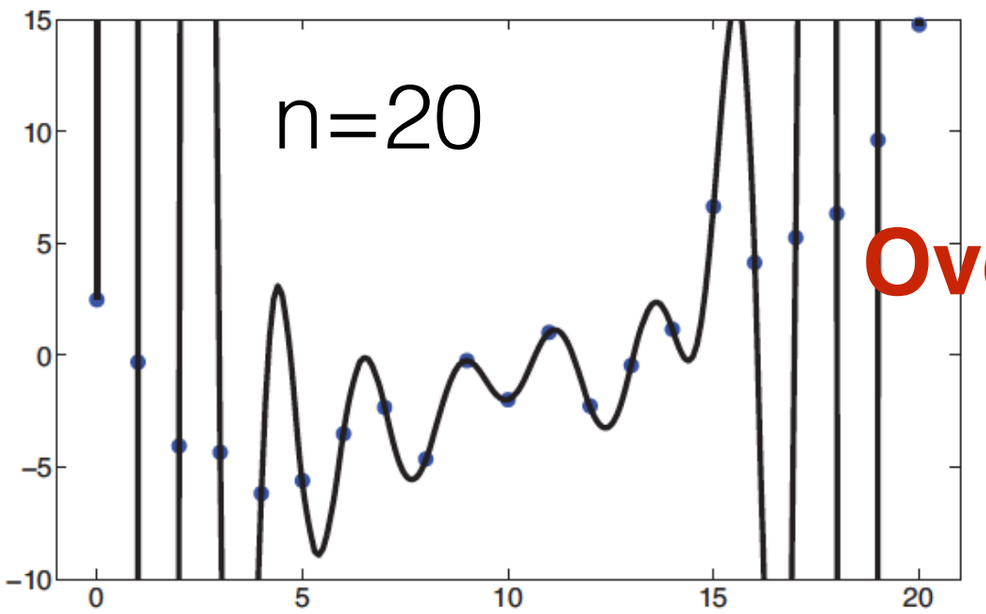
Regression

- **What degree** polynomial (n) to fit to data?









Overfitting

Key issues in ML

- What are good **models**? finding best model?
- How to optimize for accuracy of **unseen data** (avoid overfitting)?
- How much data is needed to be **confident** in results?
- How to deal with **challenging data**
  - **Missing** data, **errors** and **outliers**
  - **High-dimensional** data
  - **Large** datasets

- How to **model applications** as machine learning problems?

Summary

- **Components of machine learning** systems
  - **Data** (input: x, output: y)
  - **Model** (e.g. polynomial degree)
  - **Objective/cost functions**: what makes one ‘incorrect answer’ better/ worse than another ‘incorrect answer’

- Important concepts
  - Want high accuracy on unseen test data, but have only see training data
  - Central challenge in ML: **generalization**

Course materials

- Introduction
- Linear Algebra, Probability Review
- Regression
- Maximum Likelihood
- Logistic Regression
- Naive Bayes
- Convex optimization, SGD
- Support vector machines
- Kernel SVM
- Neural networks: Deep NNs, CNNs
- Bayesian Learning
- Ethics in ML

What you need to review

- Concepts in **probability**:
  - Conditioning, marginals, expectations
  - Gaussian/Laplace/Bernoulli/Multinomial/… distributions
- Concepts from **linear algebra** and matrix analysis
  - Eigenvalues, eigenvectors, SVD
  - Cholesky/QR decomposition, projections, subspace

Read tutorials on the course webpage