CSE 6740: Computational Data Analysis

Spring 2026

Overview

Anqi Wu

01/13

Based on materials by B. Aditya Prakash, Le Song, Mahdi Roozbahani, Carlos Guestrin

Course Information

Instructor

• Anqi Wu

Assistant Professor, CSE, College of Computing

• Email: anqiwu@gatech.edu

  • Include string “CSE 6740” in the subject

• Research Interests: Machine Learning, Computational Neuroscience, Deep Learning, Probabilisitic Generative Models, Computer Vision, Reinforcement Learning

Course Information

TAs

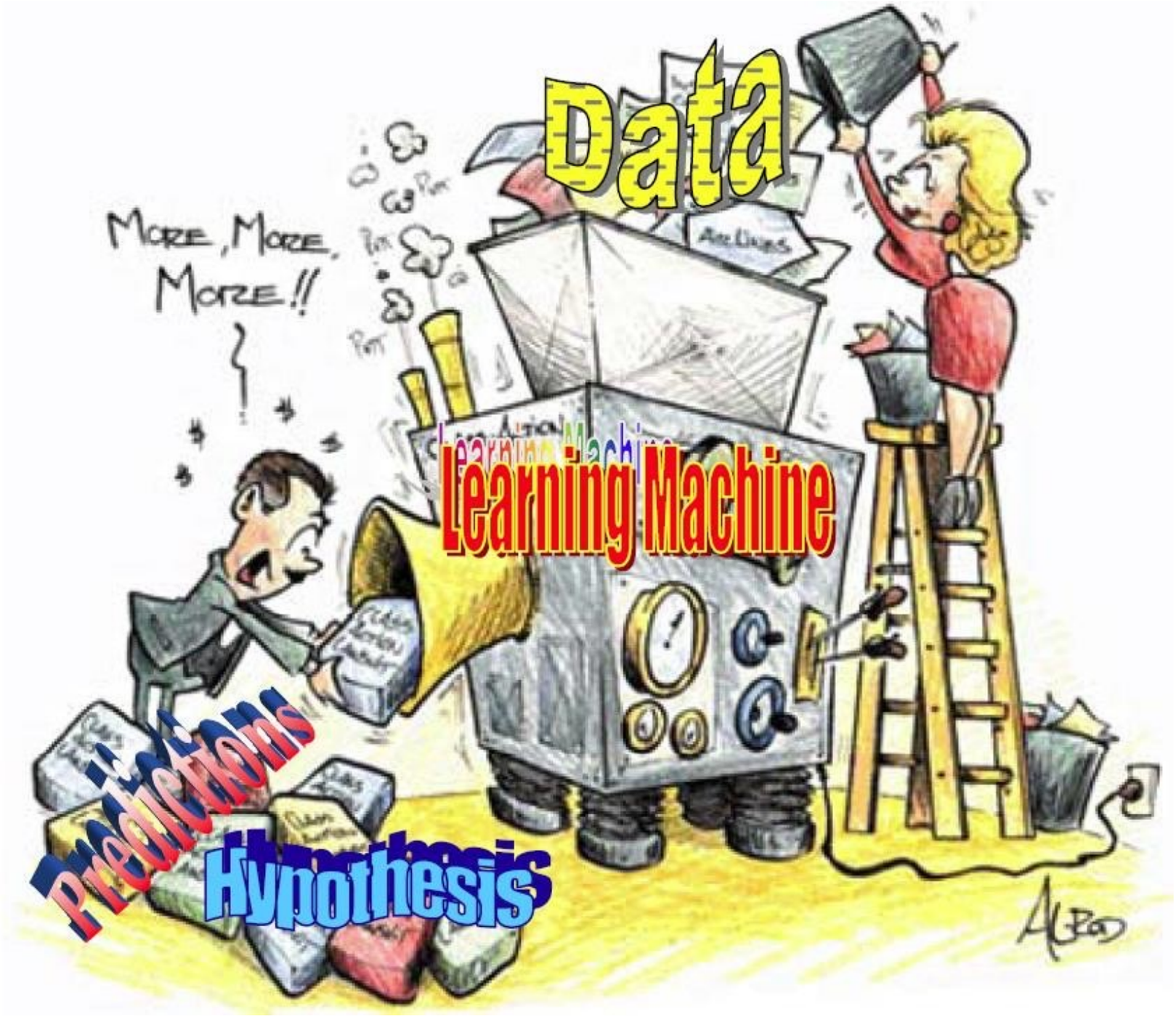
Amelie Minji Kim

Aman Garg

They will hold office hours and answer questions on Piazza and Gradescope for you. Both have been activated on Canvas.

What is machine learning (ML)

- Study of algorithms that improve their <u>performance</u> at some <u>task</u> with <u>experience</u>



Common to industrial scale problems

13 million wikipedia pages

Meta

800 million users

flickr

6 billion photos

340 million tweets per day

YouTube

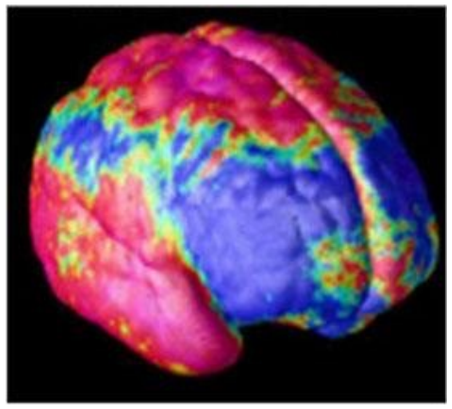
24 hours video uploaded per minutes

Google

> 1 trillion webpages

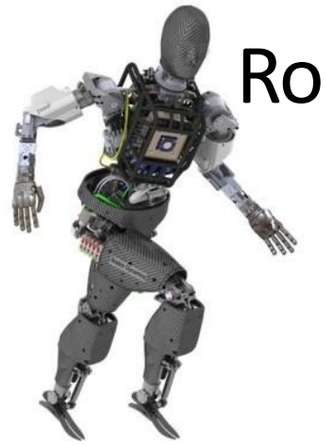
Increasingly relevant to science problems

Brain



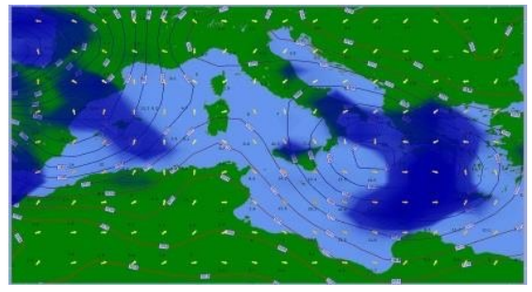
Galaxy

Self-driving car



Robots

Genome

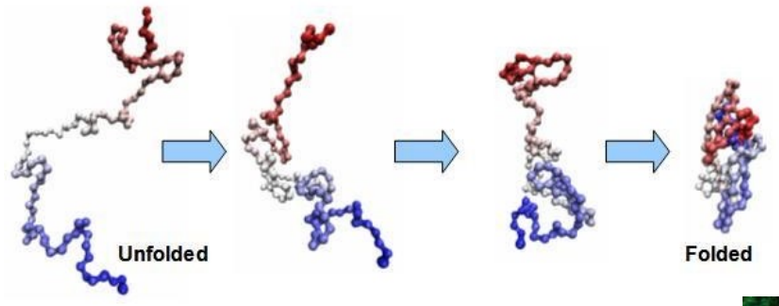


Weather

Finance

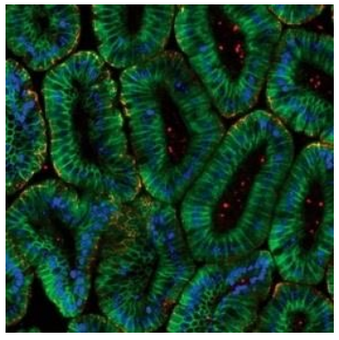
Music

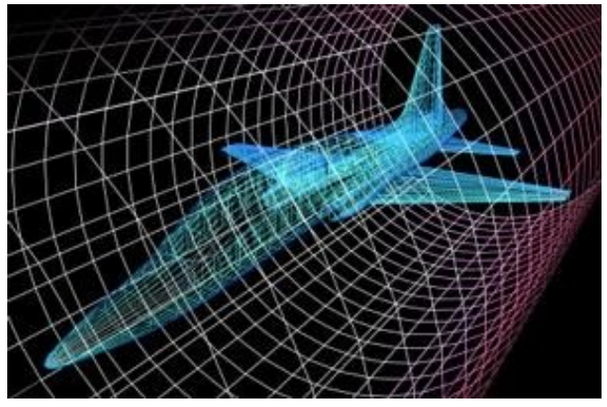
Sustainability



Medicine

Cell





Design

Syllabus

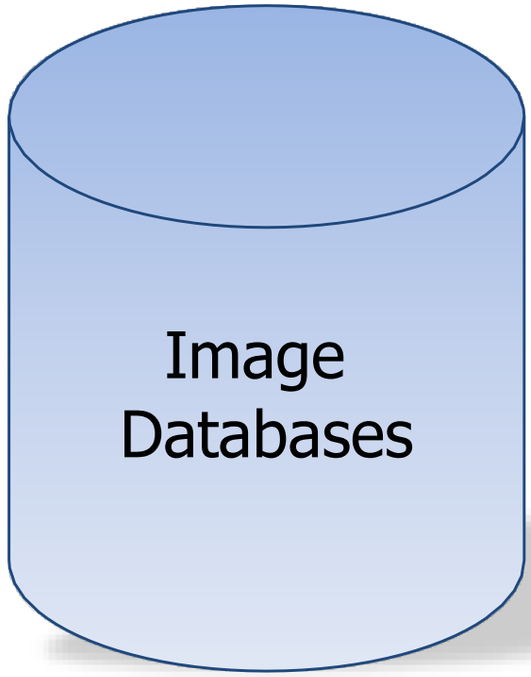
- Cover a number of most commonly used machine learning algorithms in sufficient amount of details on their mechanisms.

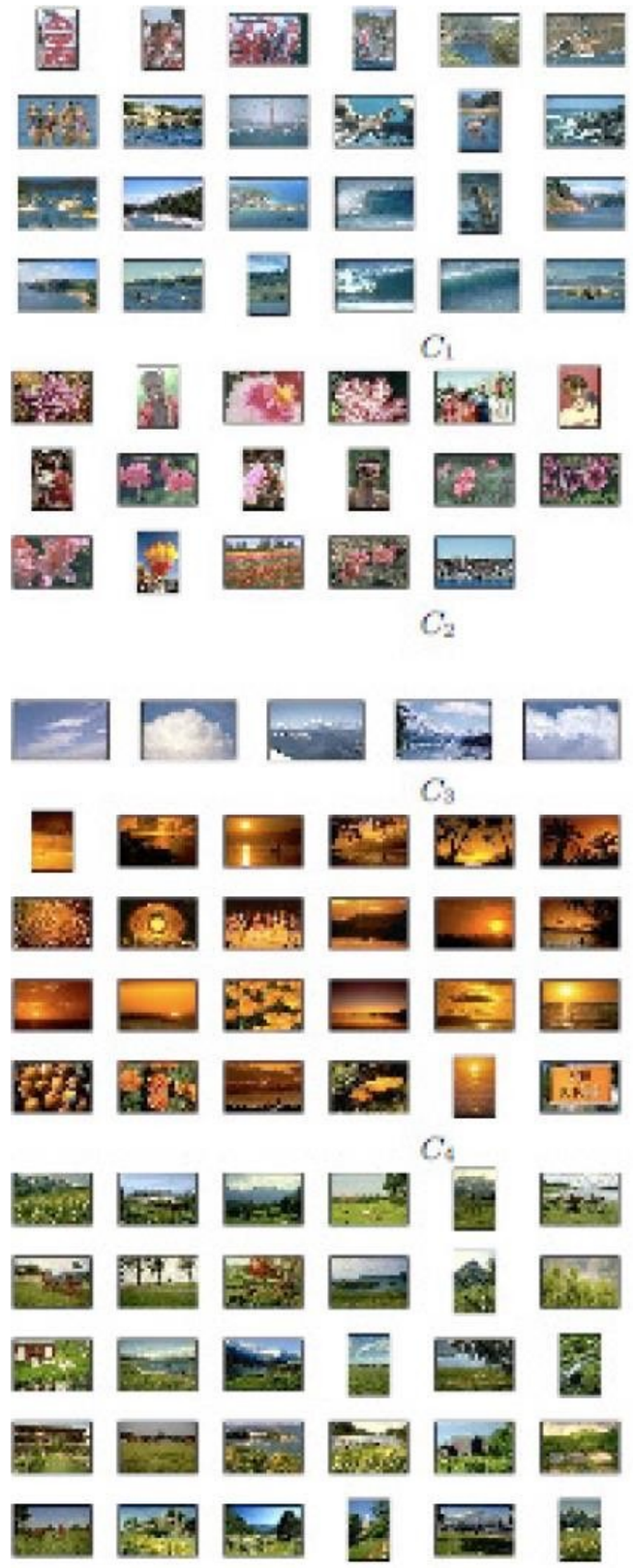
- Organization
  - Unsupervised learning (data exploration)
    - Learning without labels or without optimizing for predictive task
  - Supervised learning (predictive models)
    - Learning with labels, focusing on predictive performance
  - Advanced models (dealing with nonlinearity, combine models etc)
    - Nonlinearity, complex dependency, real world applications
  - Basics

Syllabus: unsupervised learning

- Learning without labels or without optimizing for predictive task
  - Clustering vectorial data
    - Kmeans
    - Hierarchical clustering
  - Clustering networks
    - Spectral algorithm
  - Dimensionality reduction,
    - Principal component analysis
  - Dimensionality for manifold
    - Locally linear embedding
  - Density estimation
    - Feature selection
    - Novelty/abnormality detection

Organizing Images

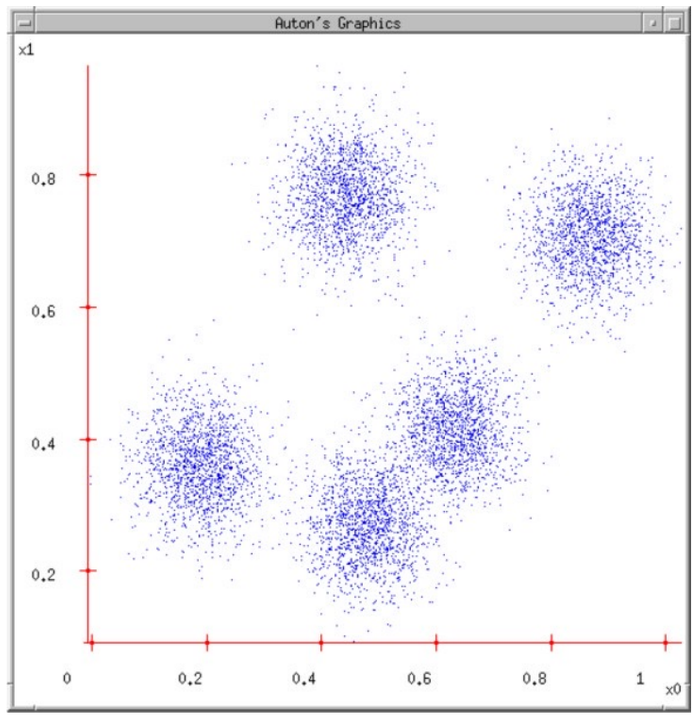




What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

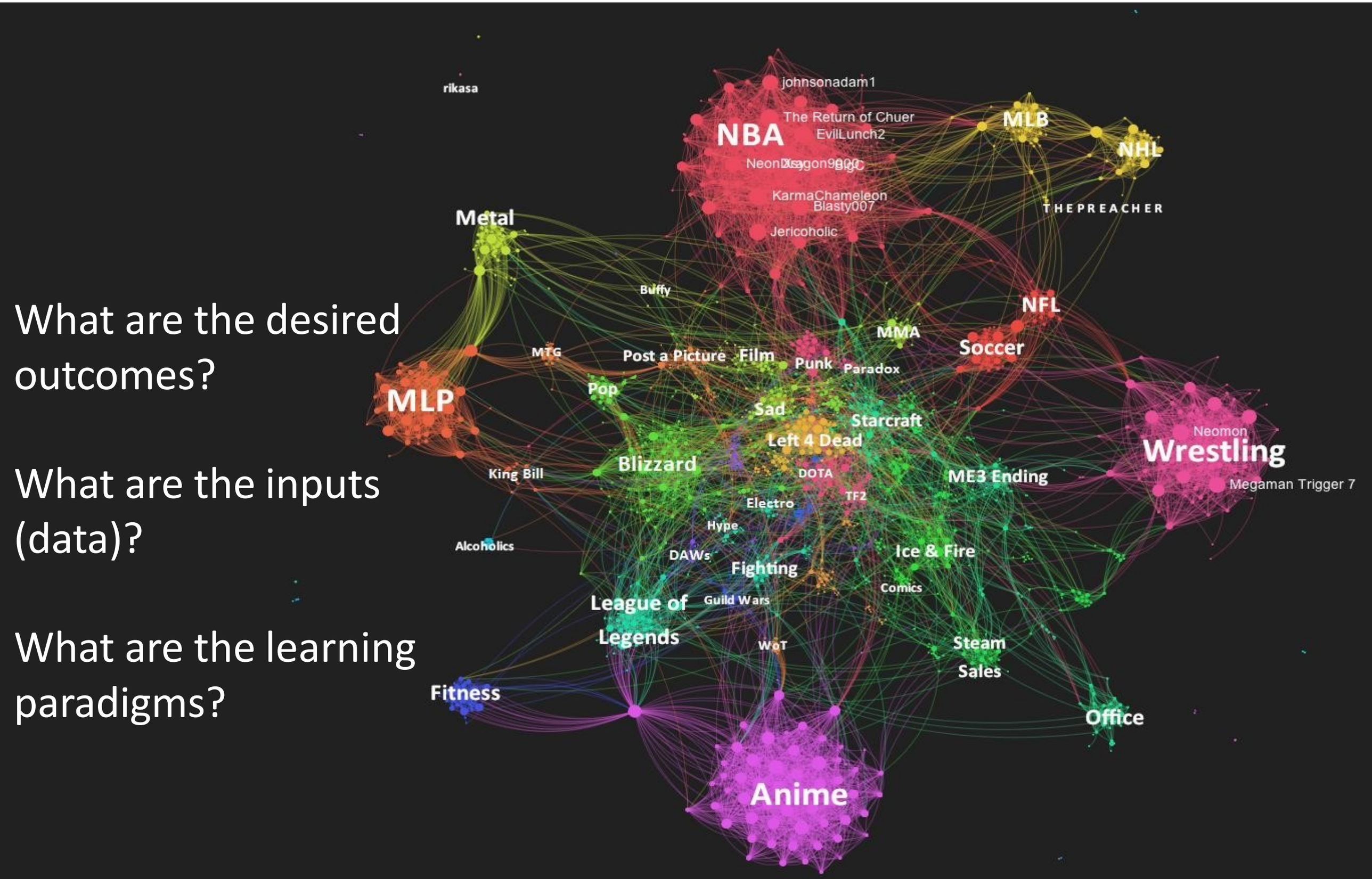


Find communities in social networks

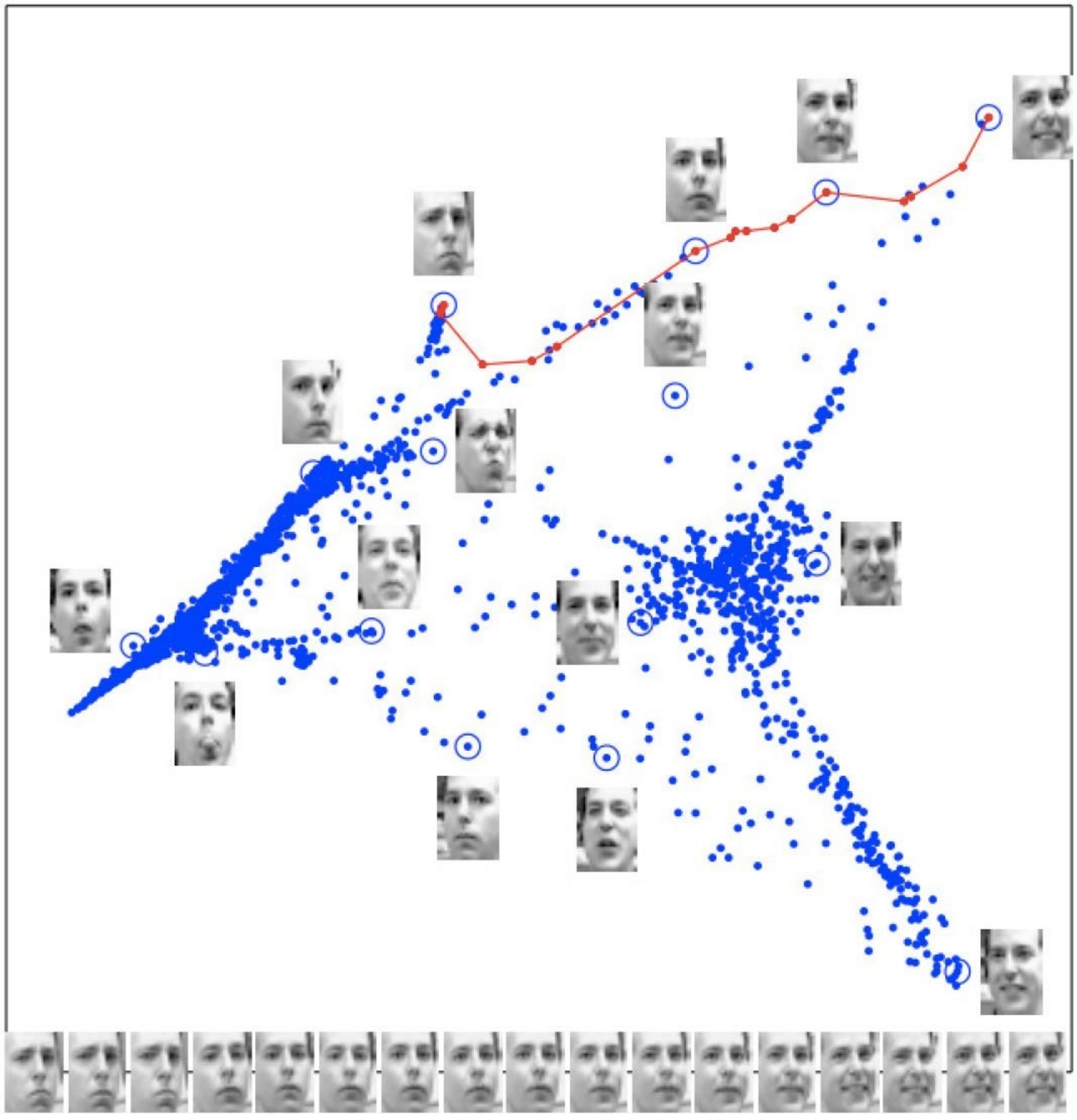
What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?



Visualize Image Relations



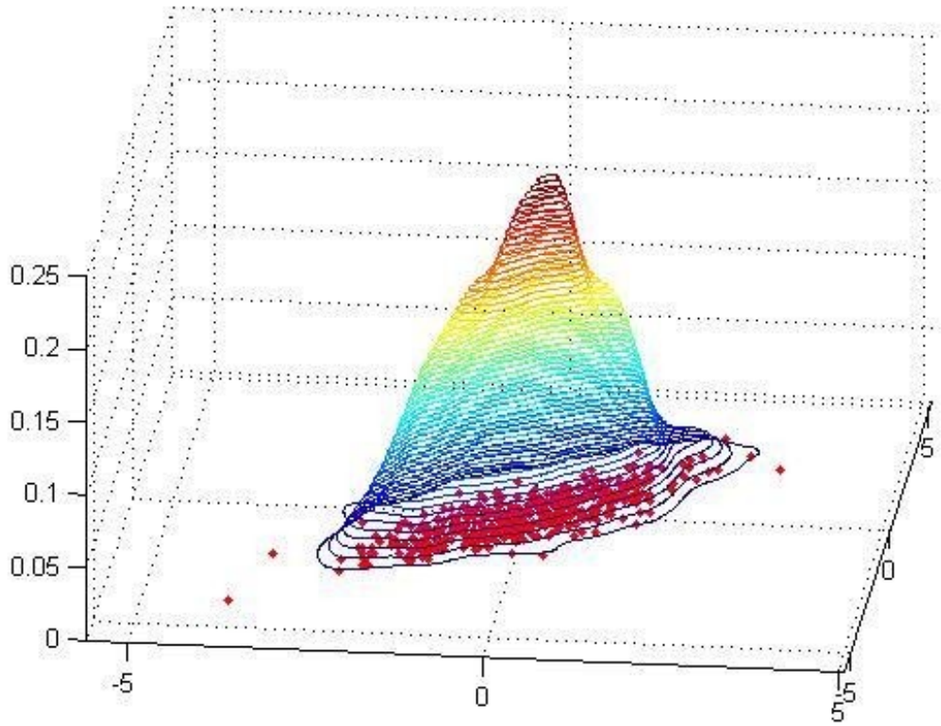
Each image has thousands or millions of pixels.

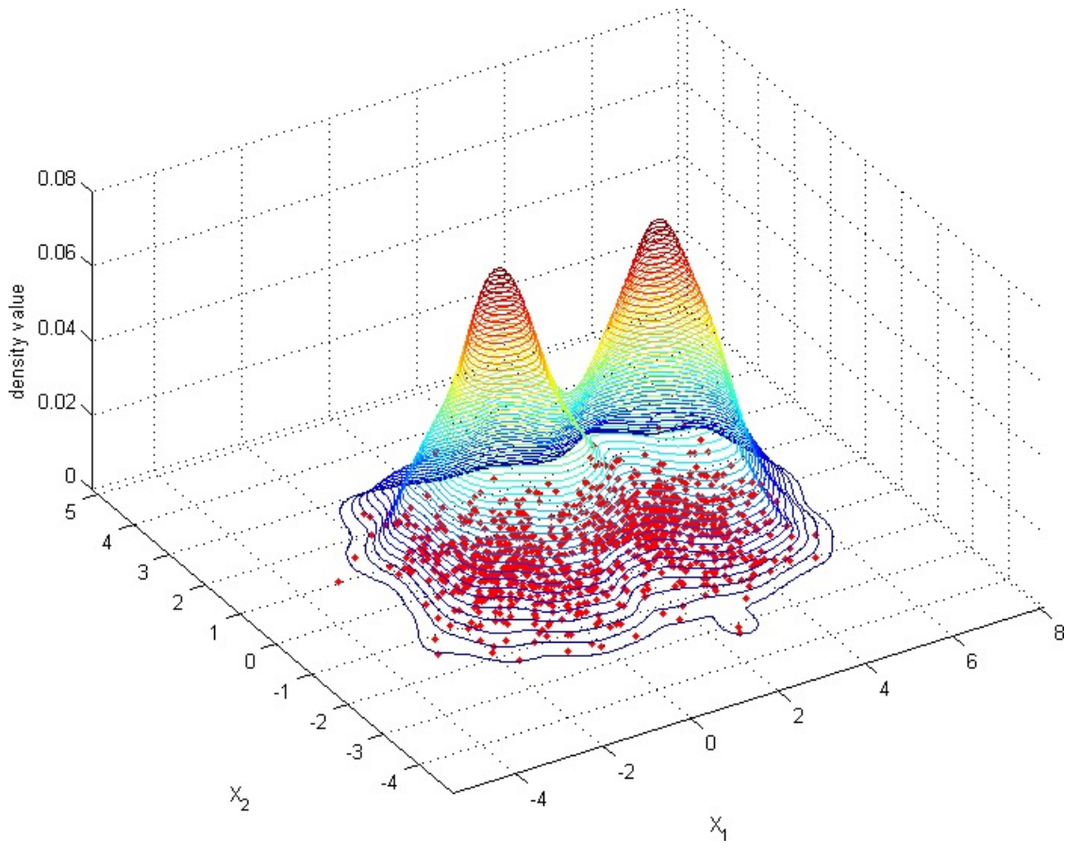
What are the desired outcomes?

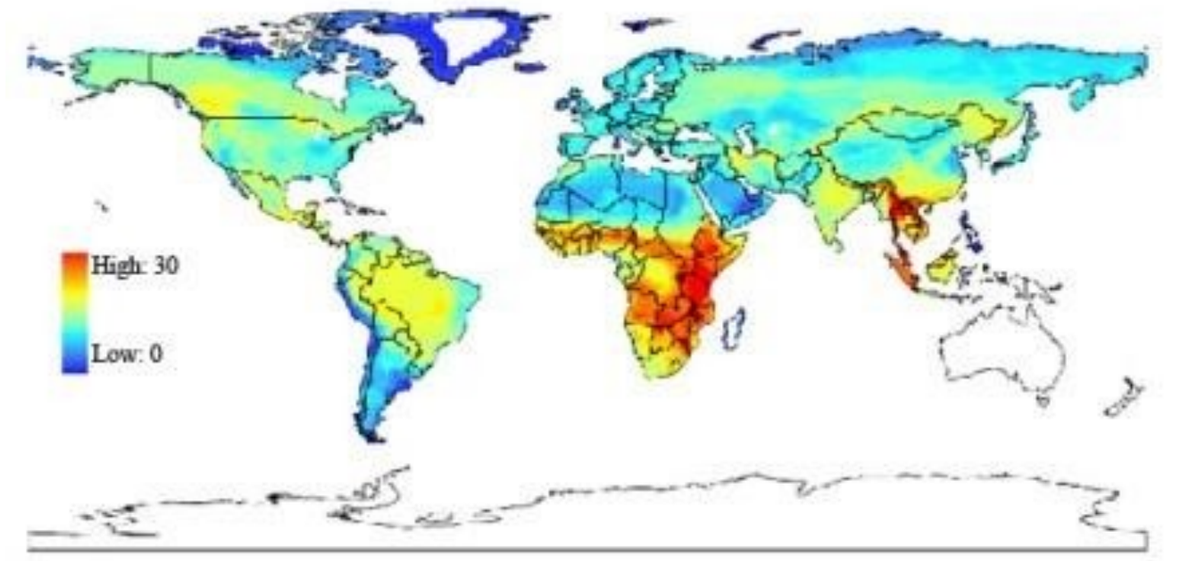
What are the inputs (data)?

What are the learning paradigms?

Shape of data







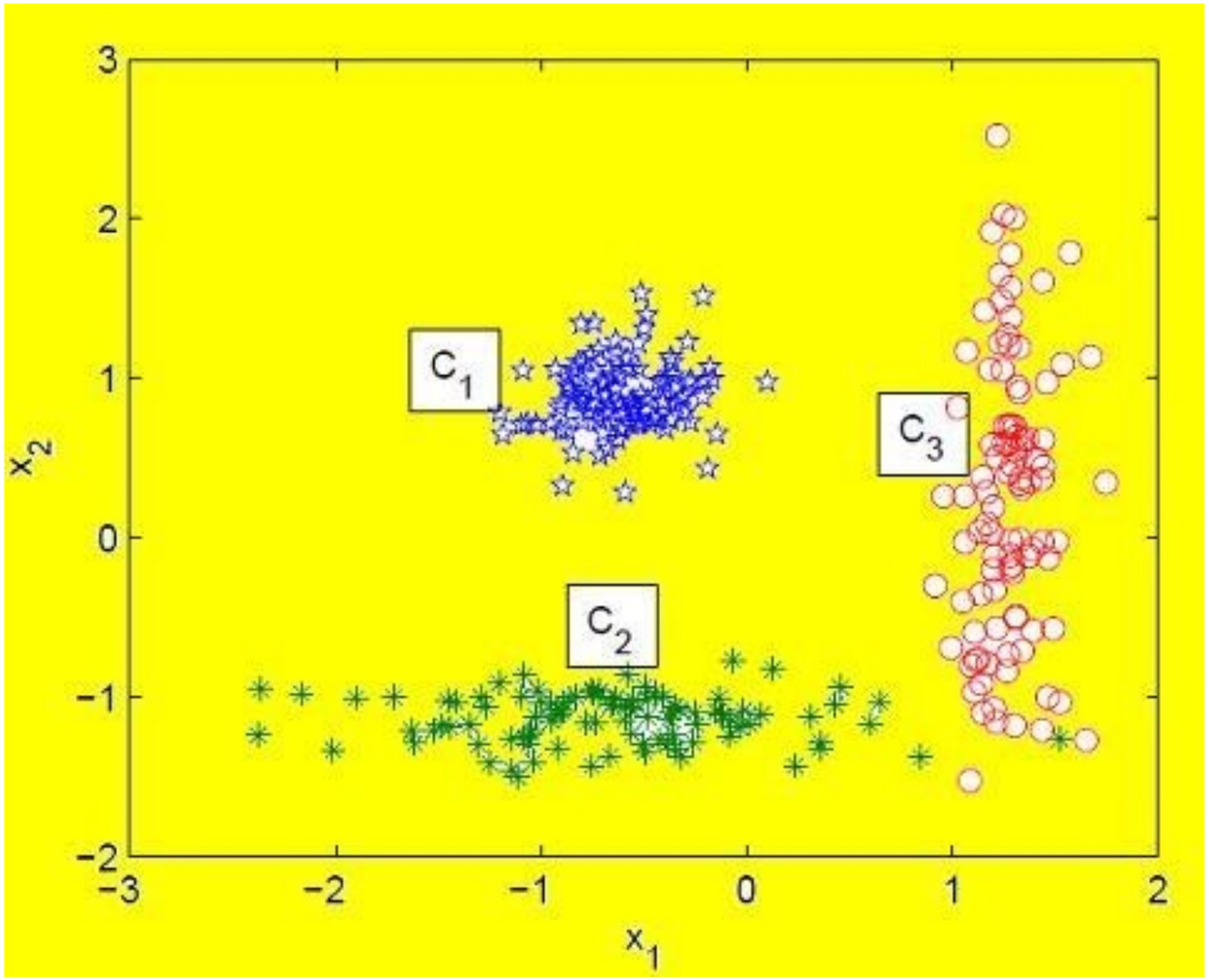
Number of carnivore species found in different regions across the globe

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Feature selection



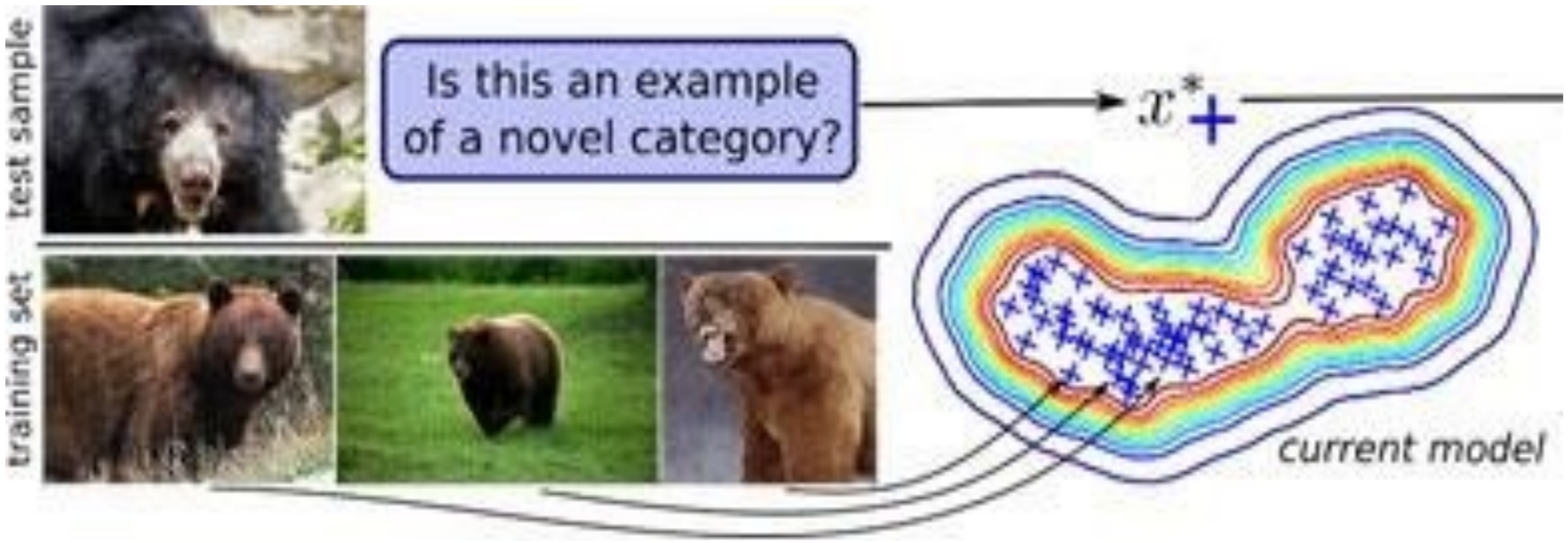
http://splinedoctors.com/2009/02/hurry-up-and-choose/

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Novelty/abnormality detection



Find abnormal object

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Syllabus: supervised learning

- Learning with labels, focusing on predictive performance
  - Classifications
    - Nearest neighbor classifier
    - Naïve Bayes classifier
    - Logistic regression
  - Combined classifiers
    - Boosting
  - Regressions
    - Ridge regression
    - Cross-validation

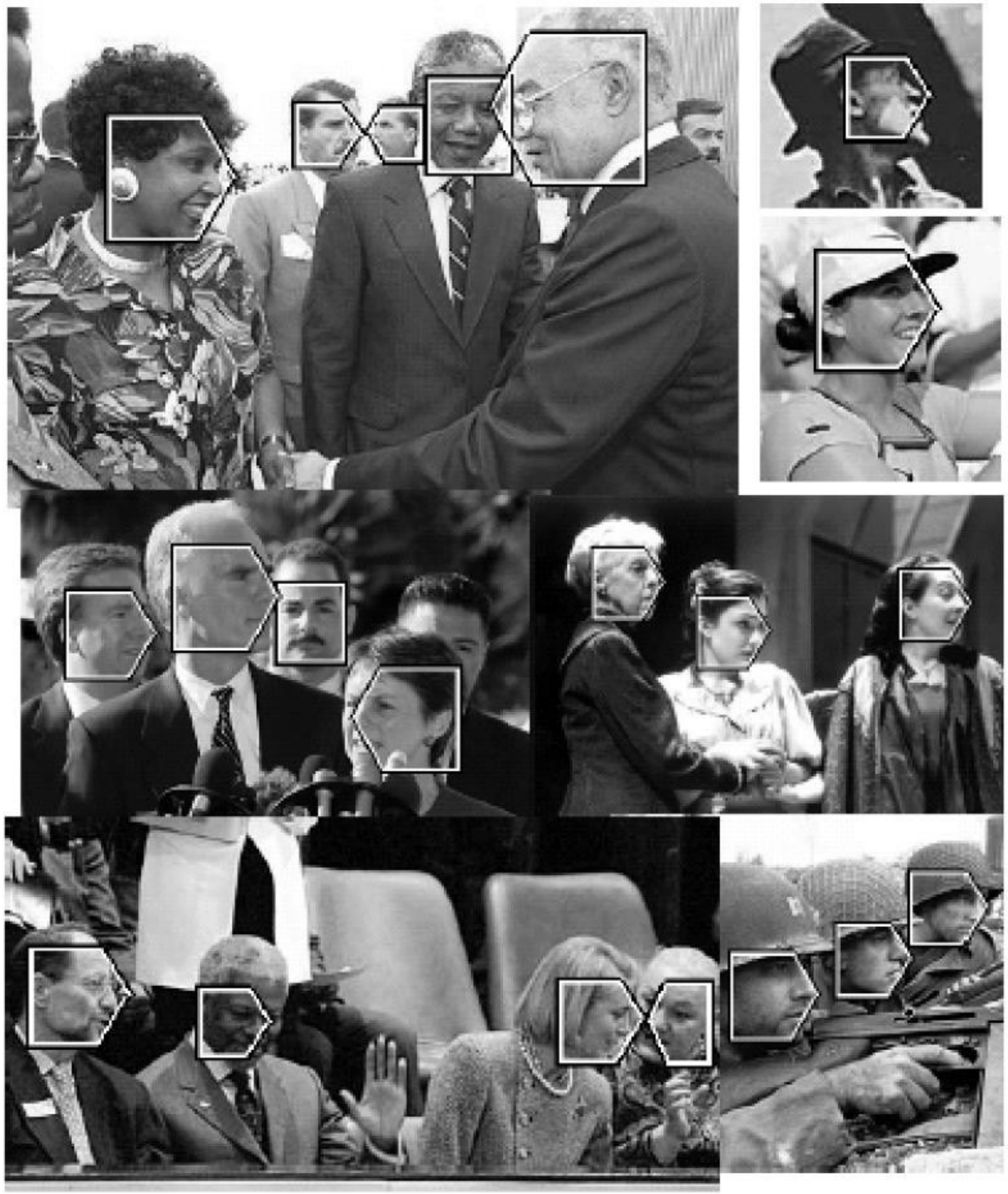
Image classification

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Face Detection

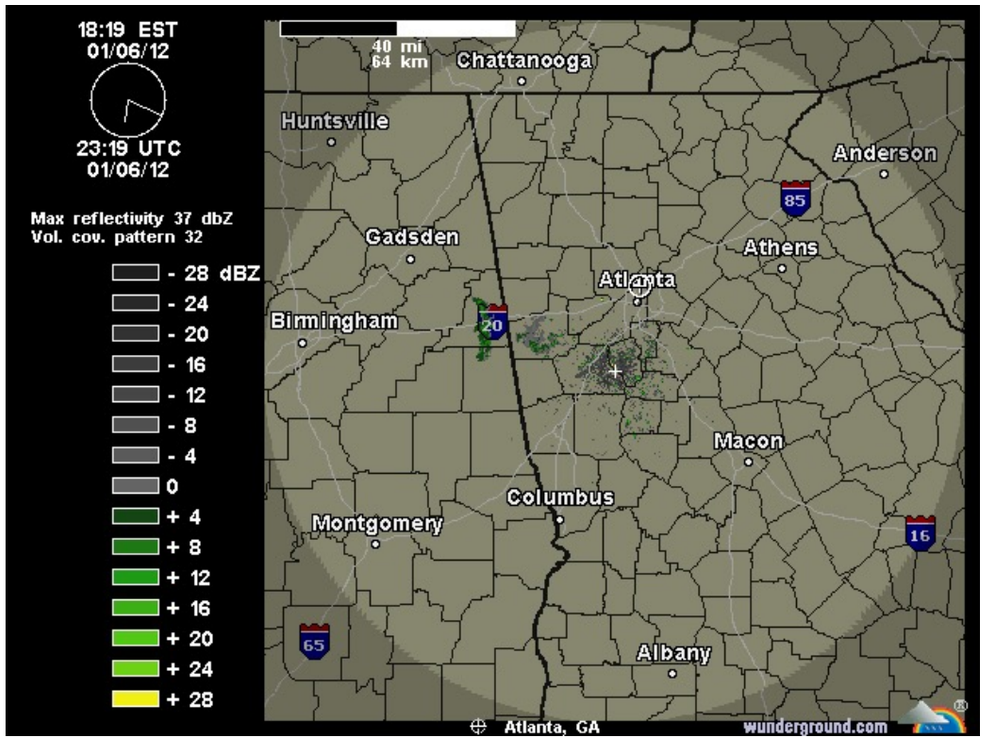


What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Weather Prediction



Predict

Numeric values:
40 F
Wind: NE at 14 km/h
Humidity: 83%

Predict

What are the desired outcomes?

What are the inputs (data)?

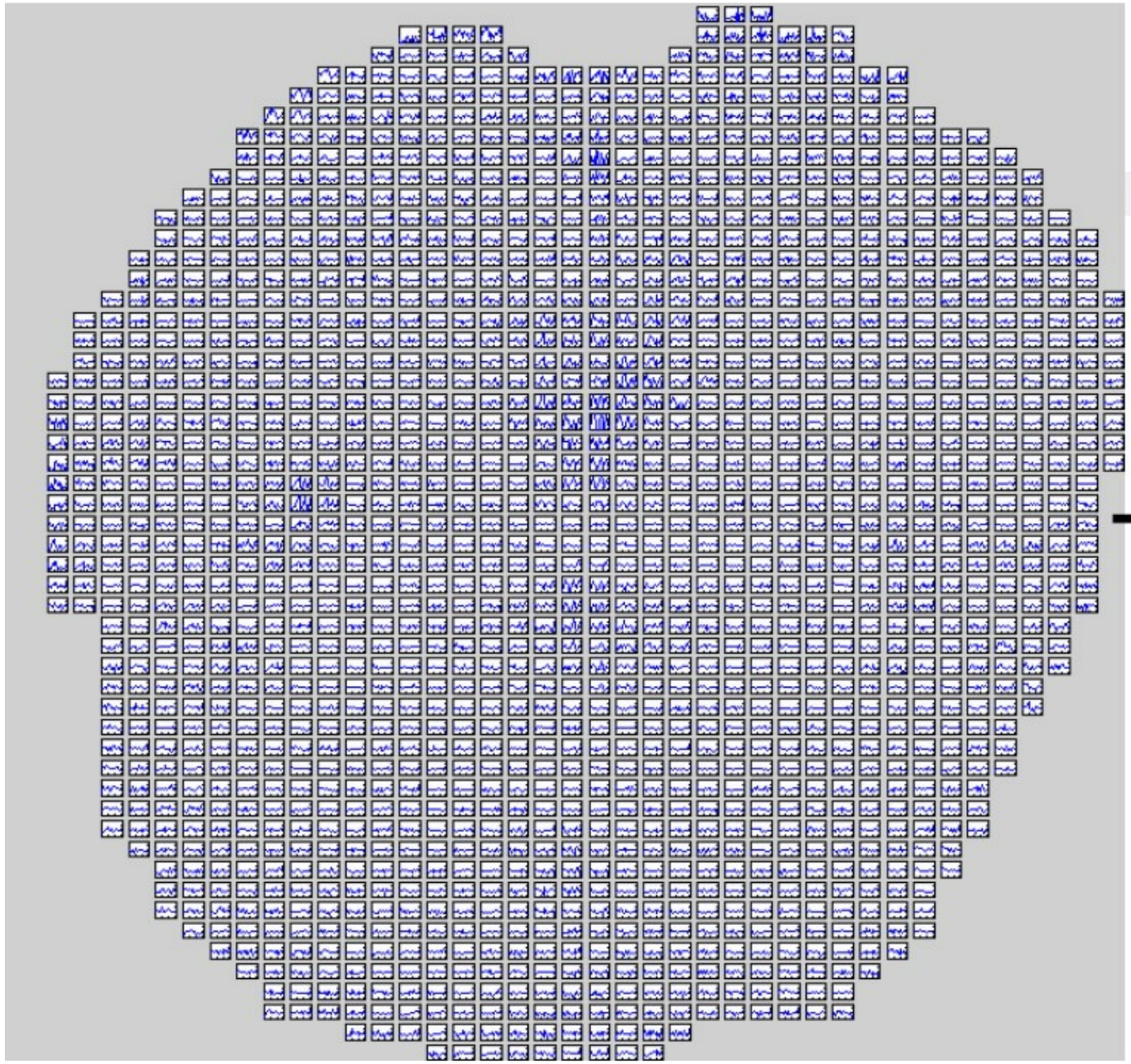
What are the learning paradigms?

Understanding brain activity

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

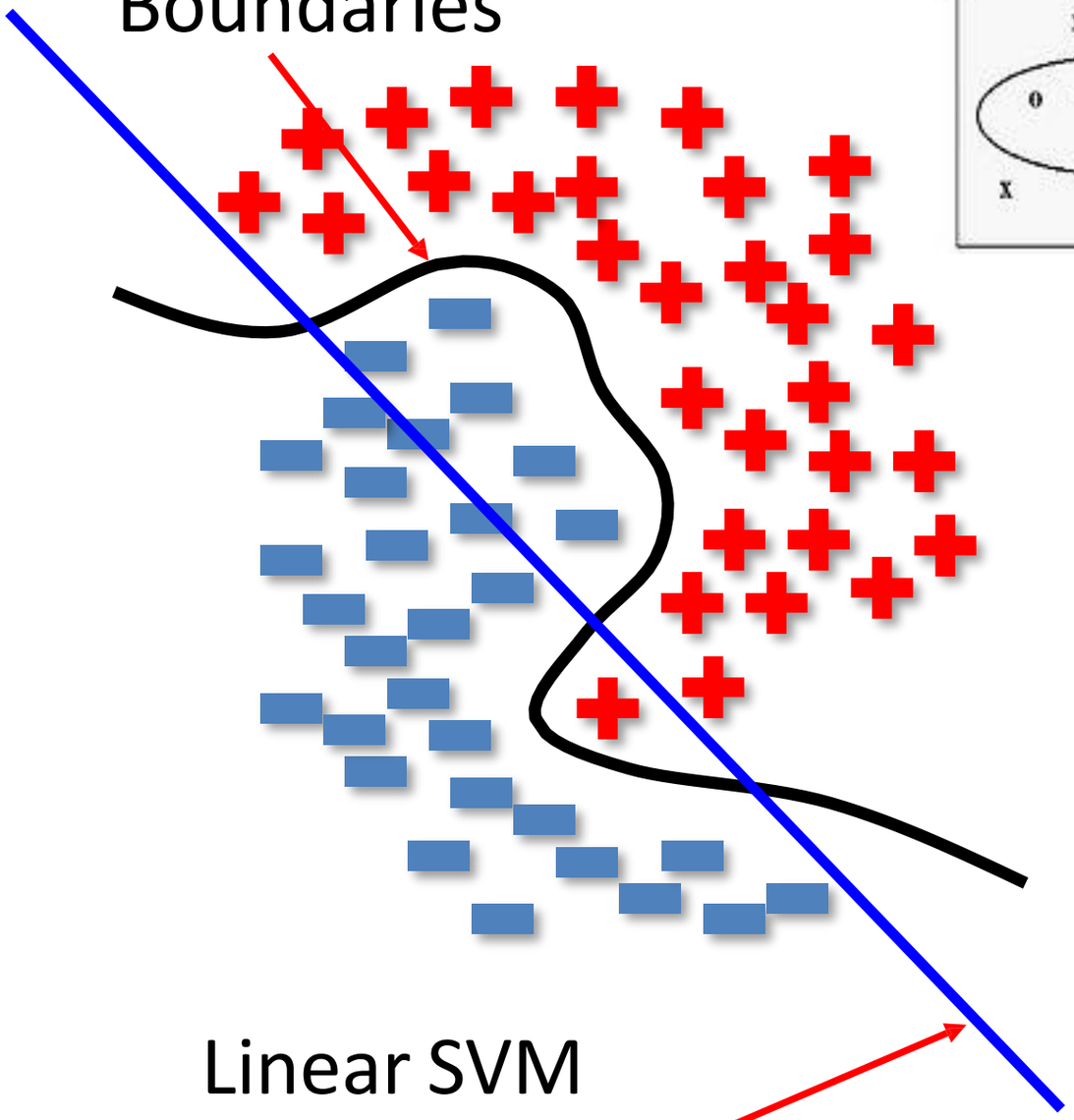


Reading a noun (vs verb)

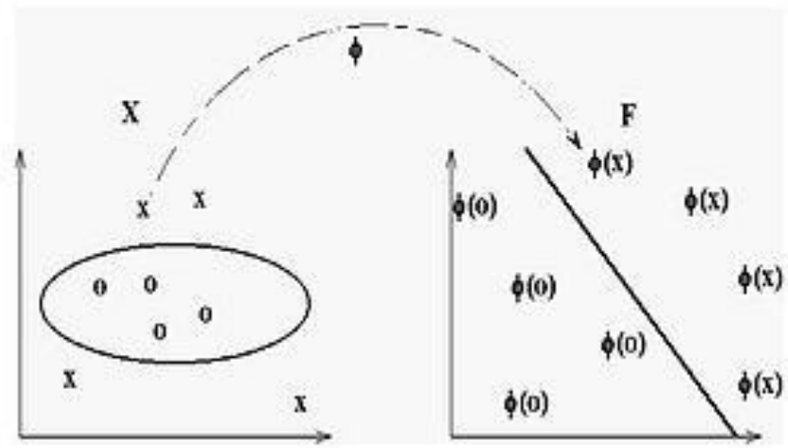
[Rustandi et al., 2005]

Nonlinear classifier

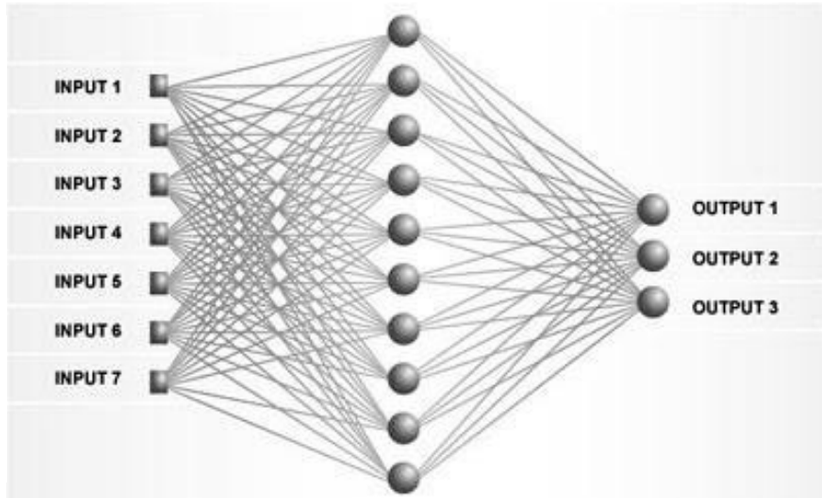
Nonlinear Decision Boundaries



Linear SVM Decision Boundaries



Kernel methods



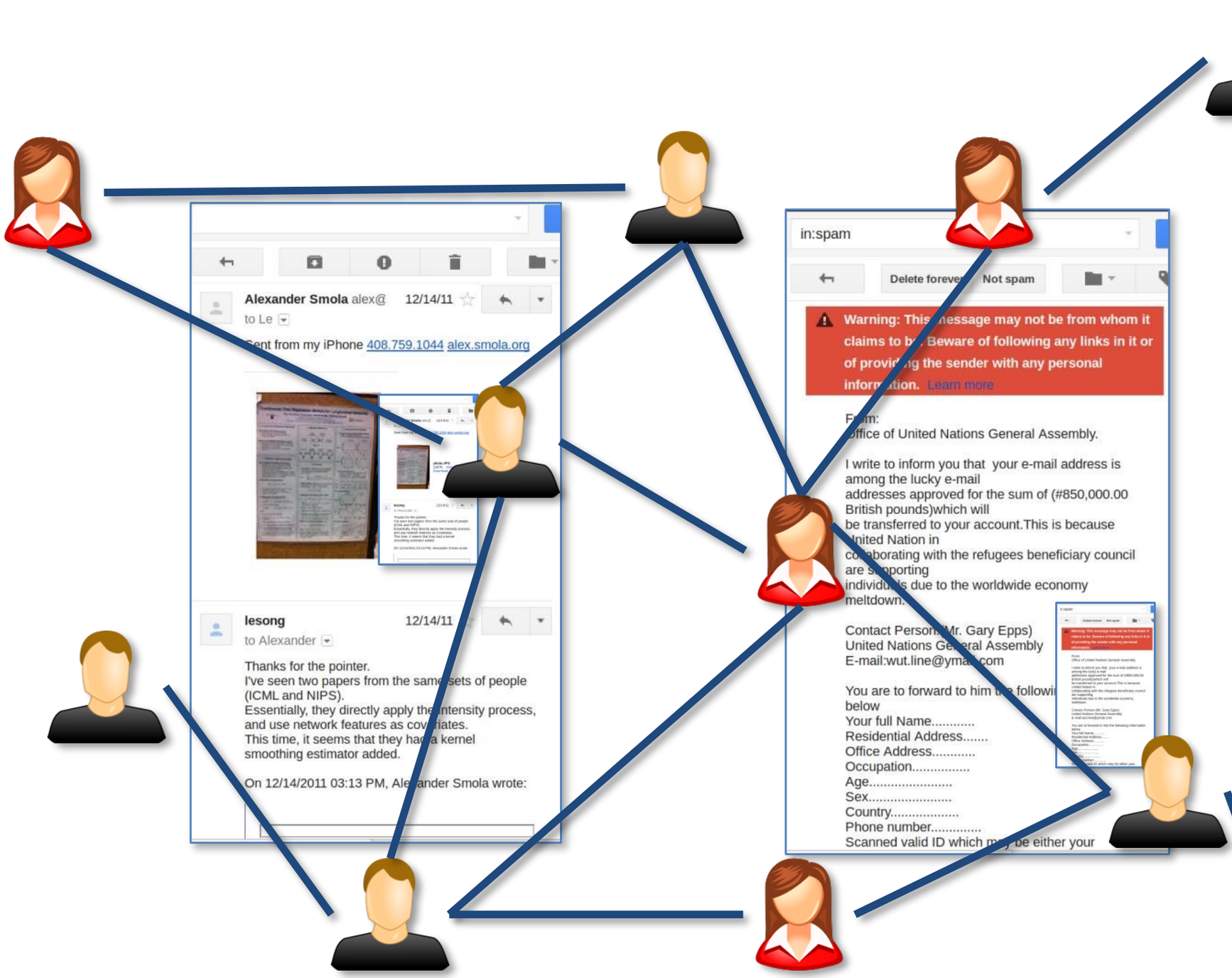
Neural Networks

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Spam Filtering



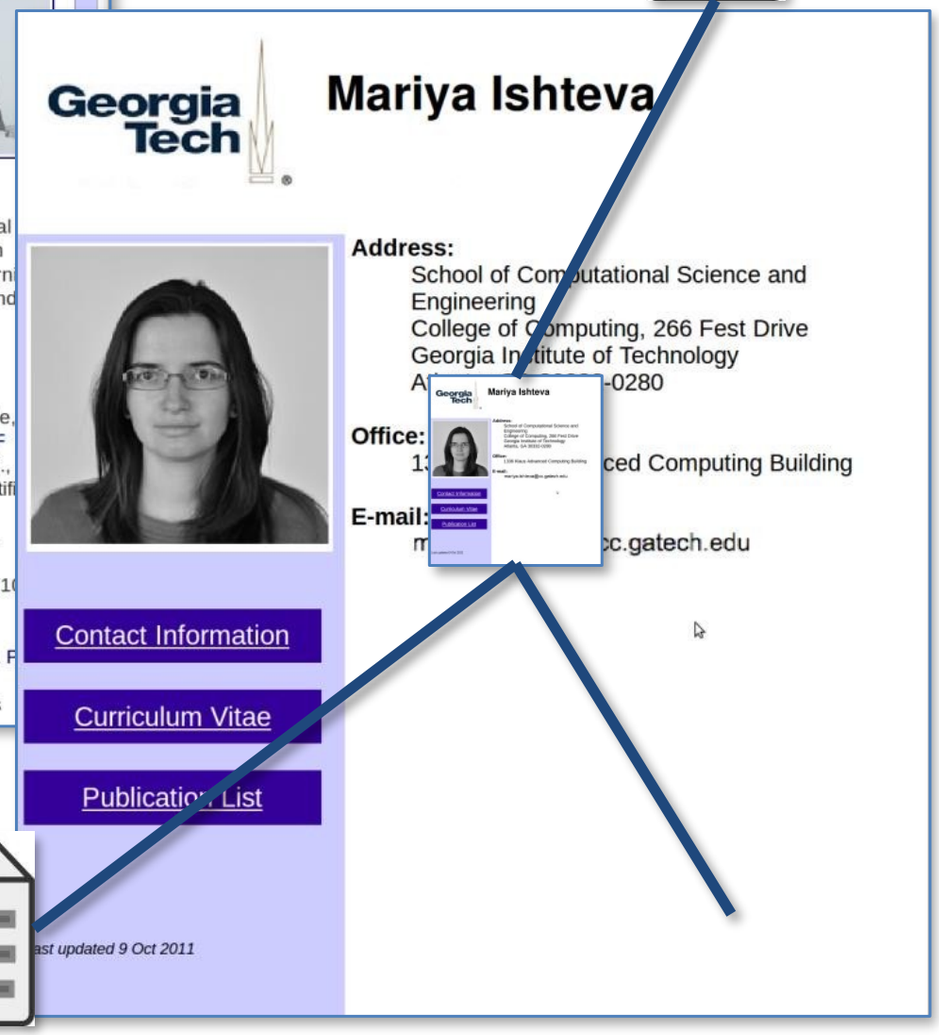
What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Webpage classification

Company homepage vs. University homepage

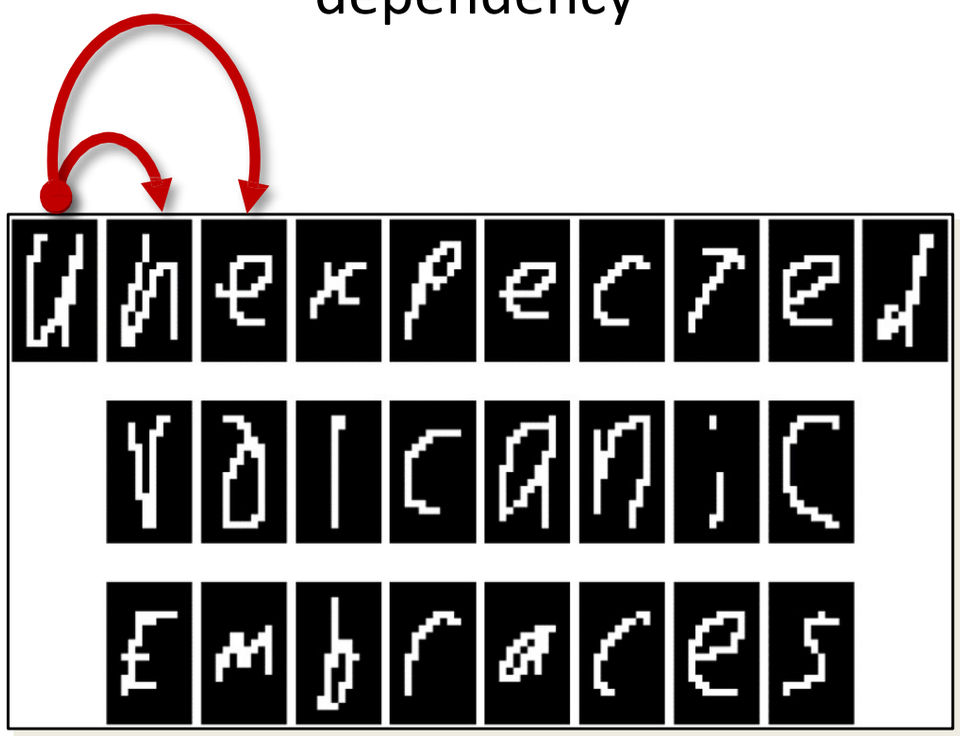


Syllabus: advanced topics, complex models

- Nonlinearity, complex dependency, real world applications
  - Kernel methods
  - Hidden Markov models
  - Graphical models
  - Transformers
  - Generative models
  - Reinforcement learning

Handwritten digit recognition/text annotation

Inter-character dependency



Inter-word dependency

Aoccdrnig to a sudty at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a ttoal mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

What are the desired outcomes?

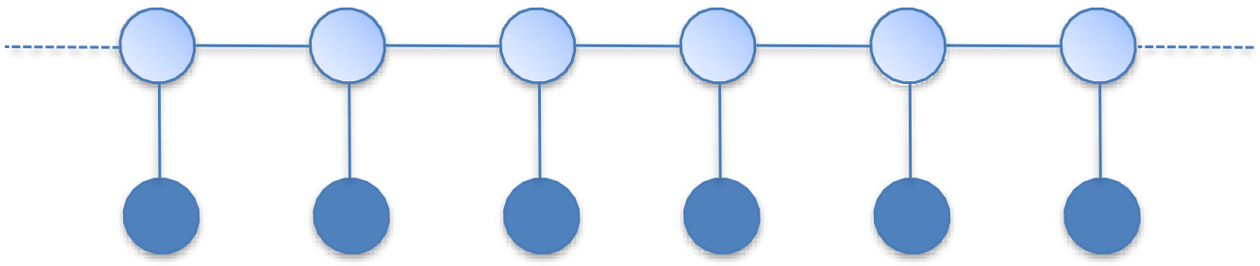
What are the inputs (data)?

What are the learning paradigms?

Speech recognition

Models

Hidden Markov Models



Text

"Machine Learning is the preferred method for speech recognition …"

Audio signals

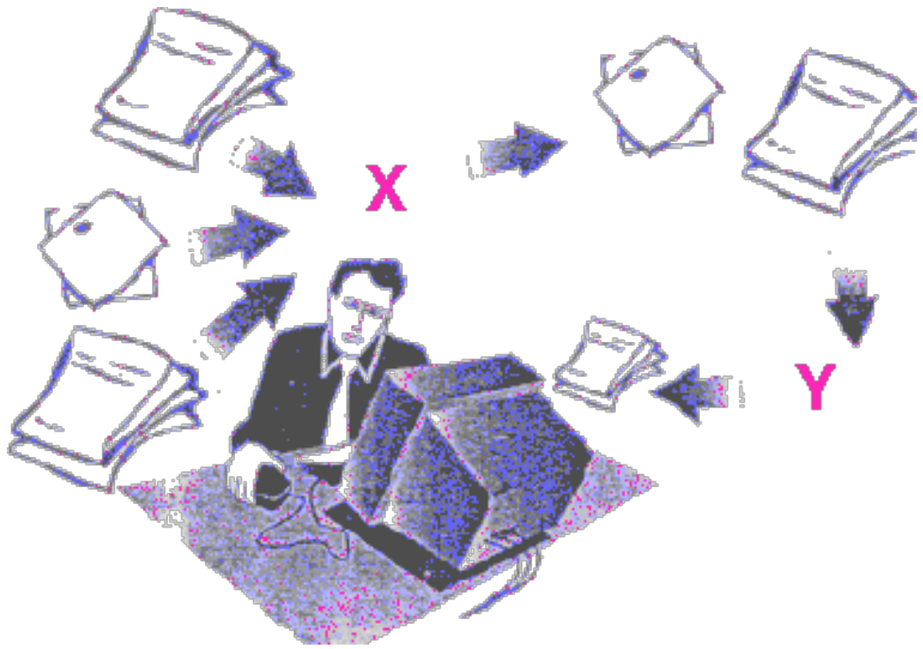
Bioinformatics

Georgia Tech

Where is the gene?

Organizing documents

- Reading, digesting, and categorizing a vast text database is too much for human!



What are the desired outcomes?

What are the inputs (data)?

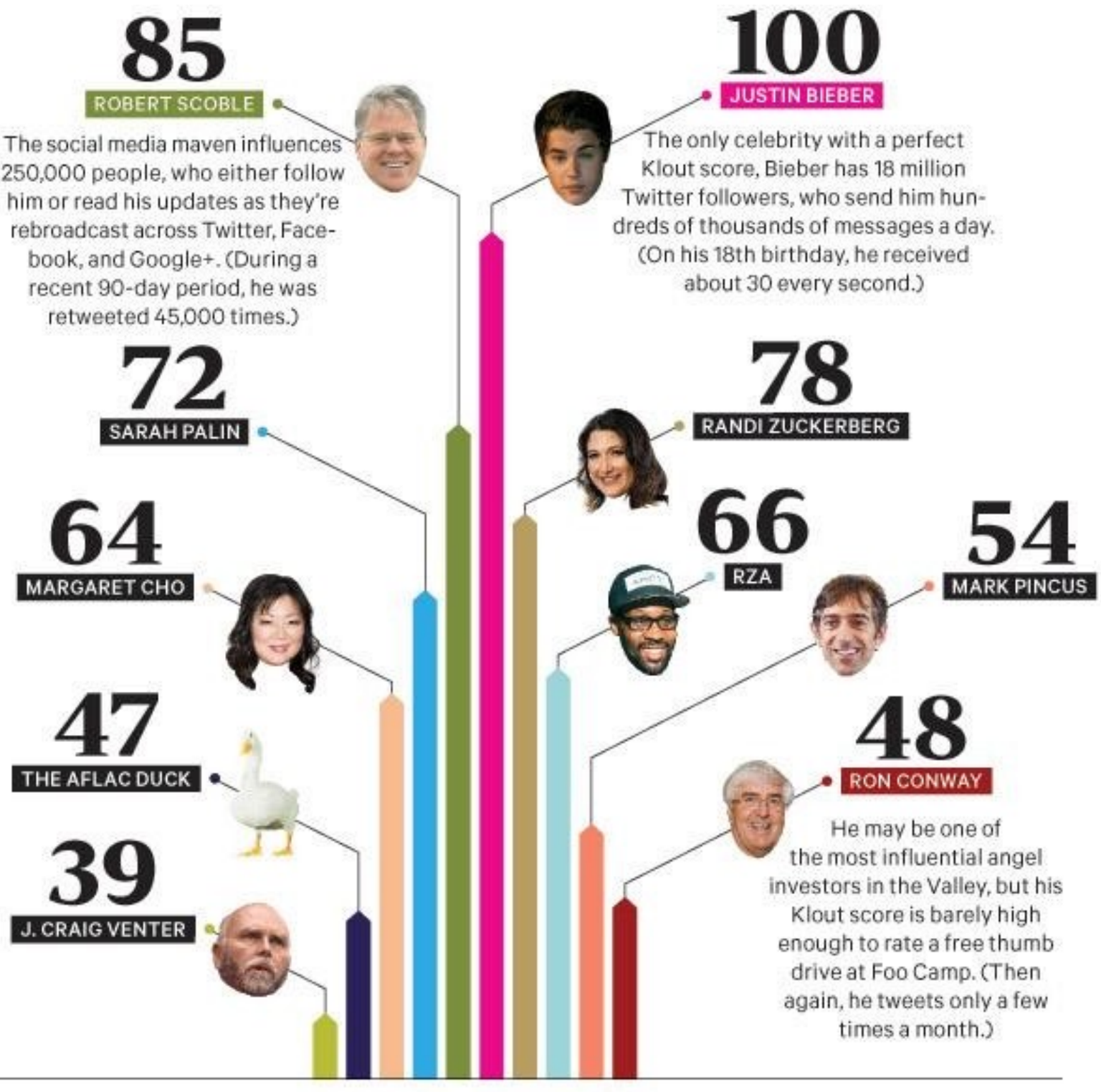
What are the learning paradigms?

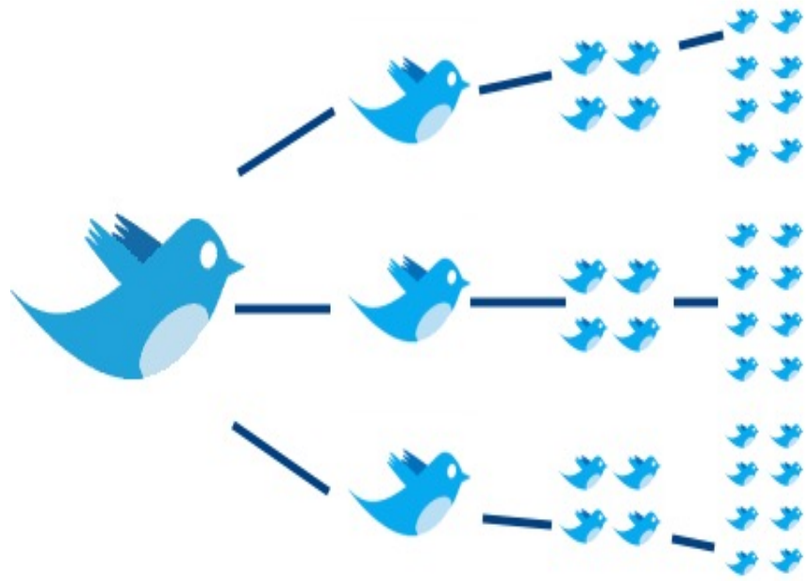
- **We want:**

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Quantifying social influence



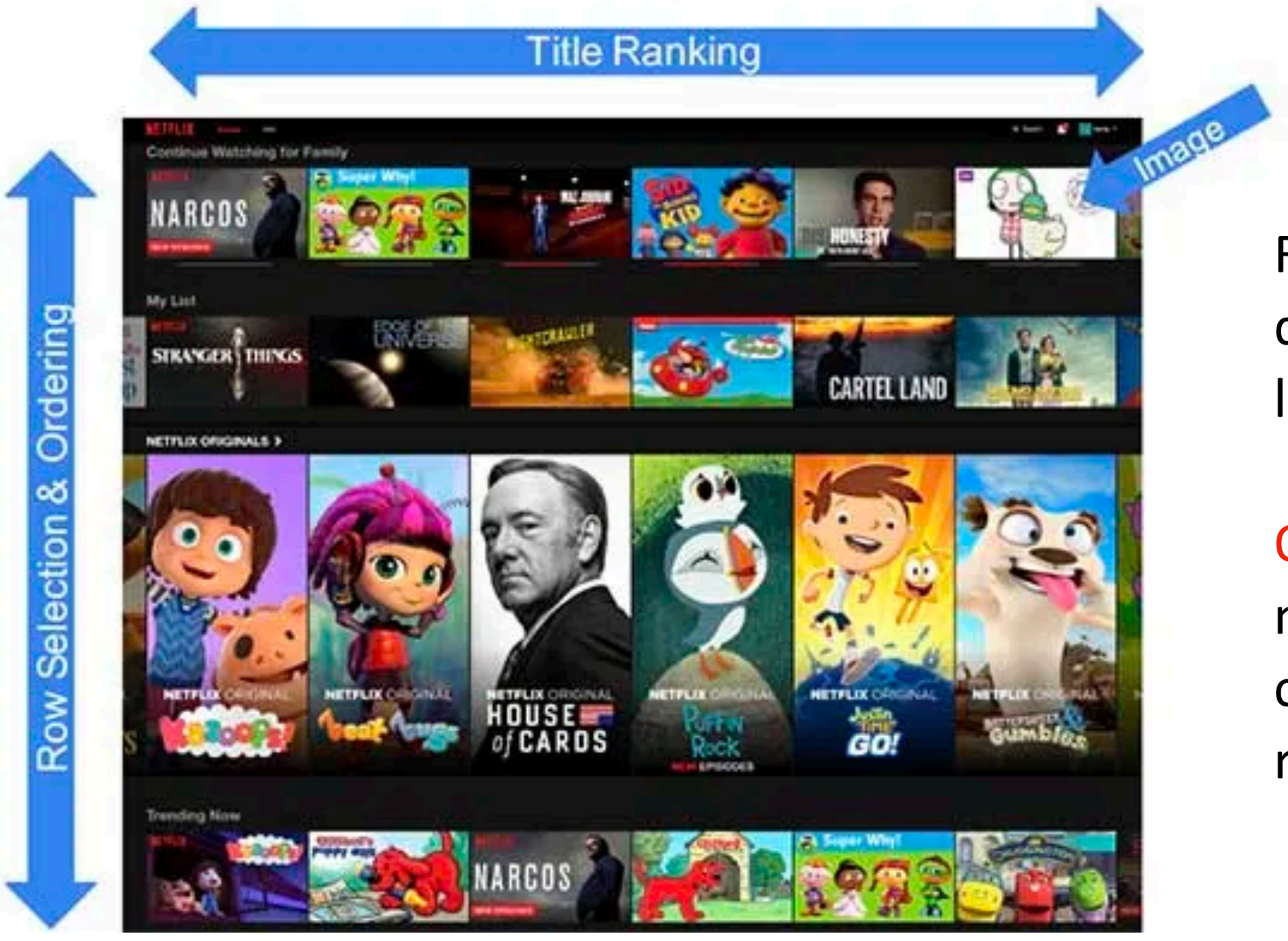


What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

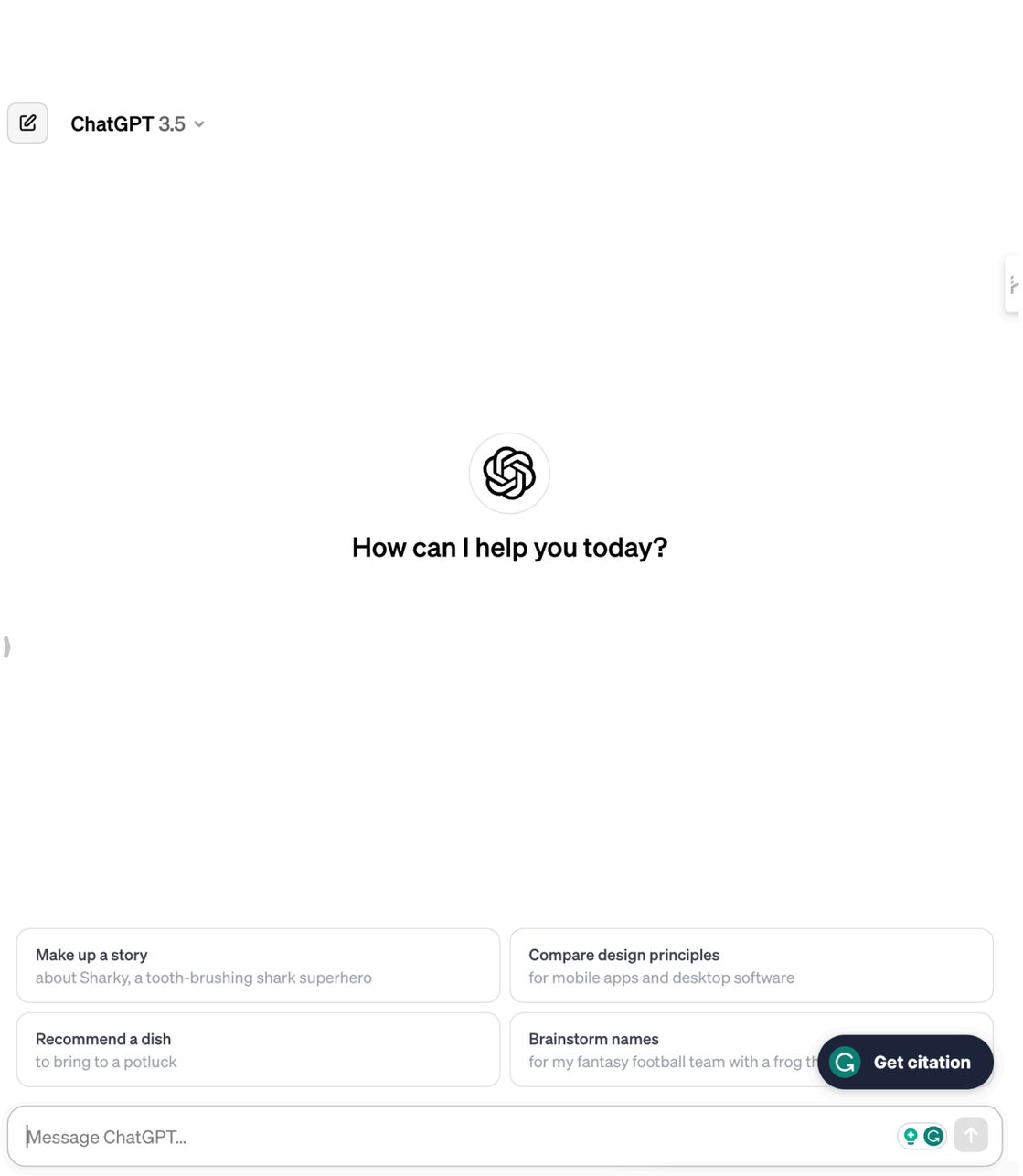
Recommendation with human feedback



Recommendations are driven by machine learning algorithms

Over 80% of what members watch comes from Netflix's recommendations

Language models



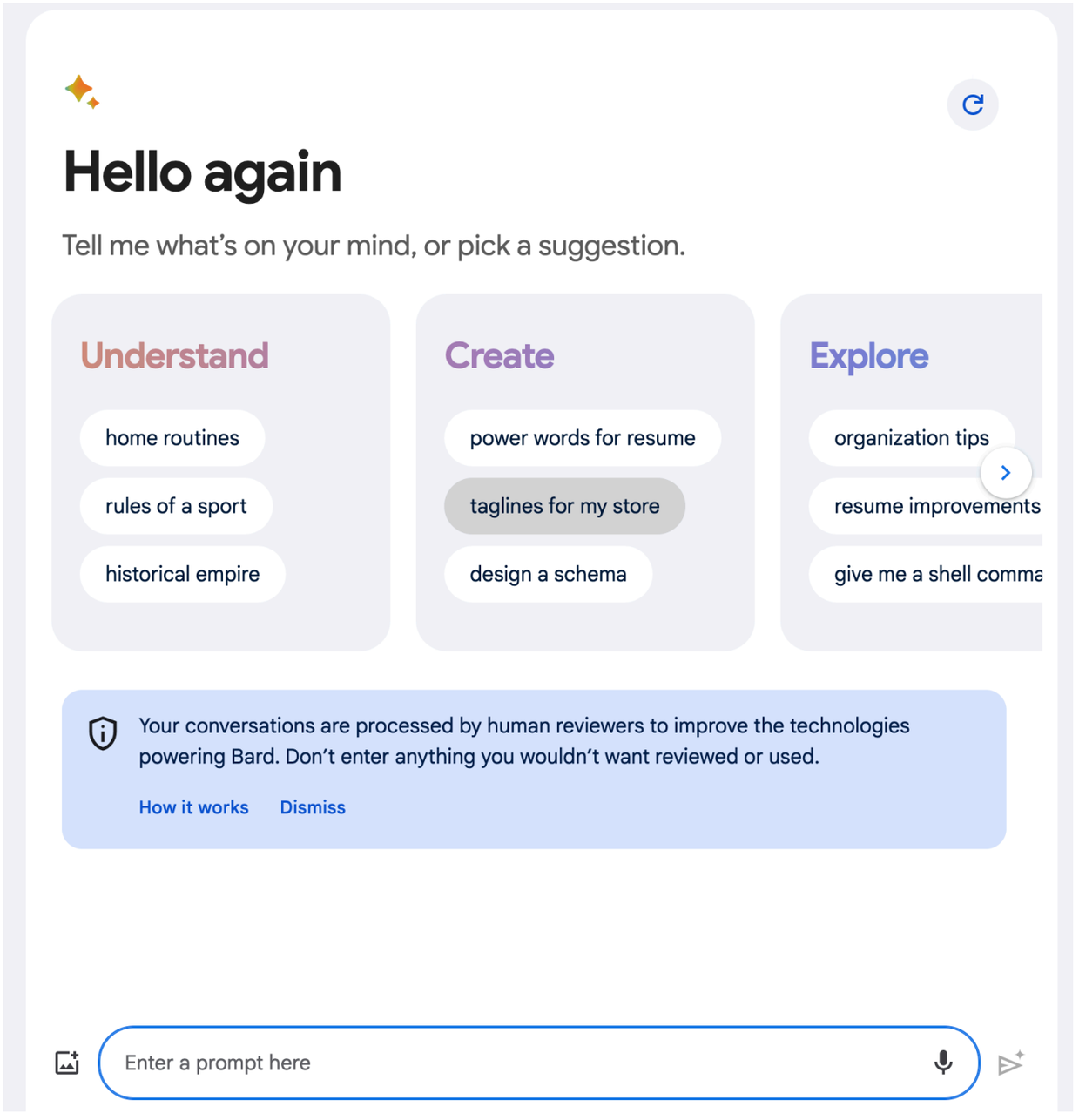
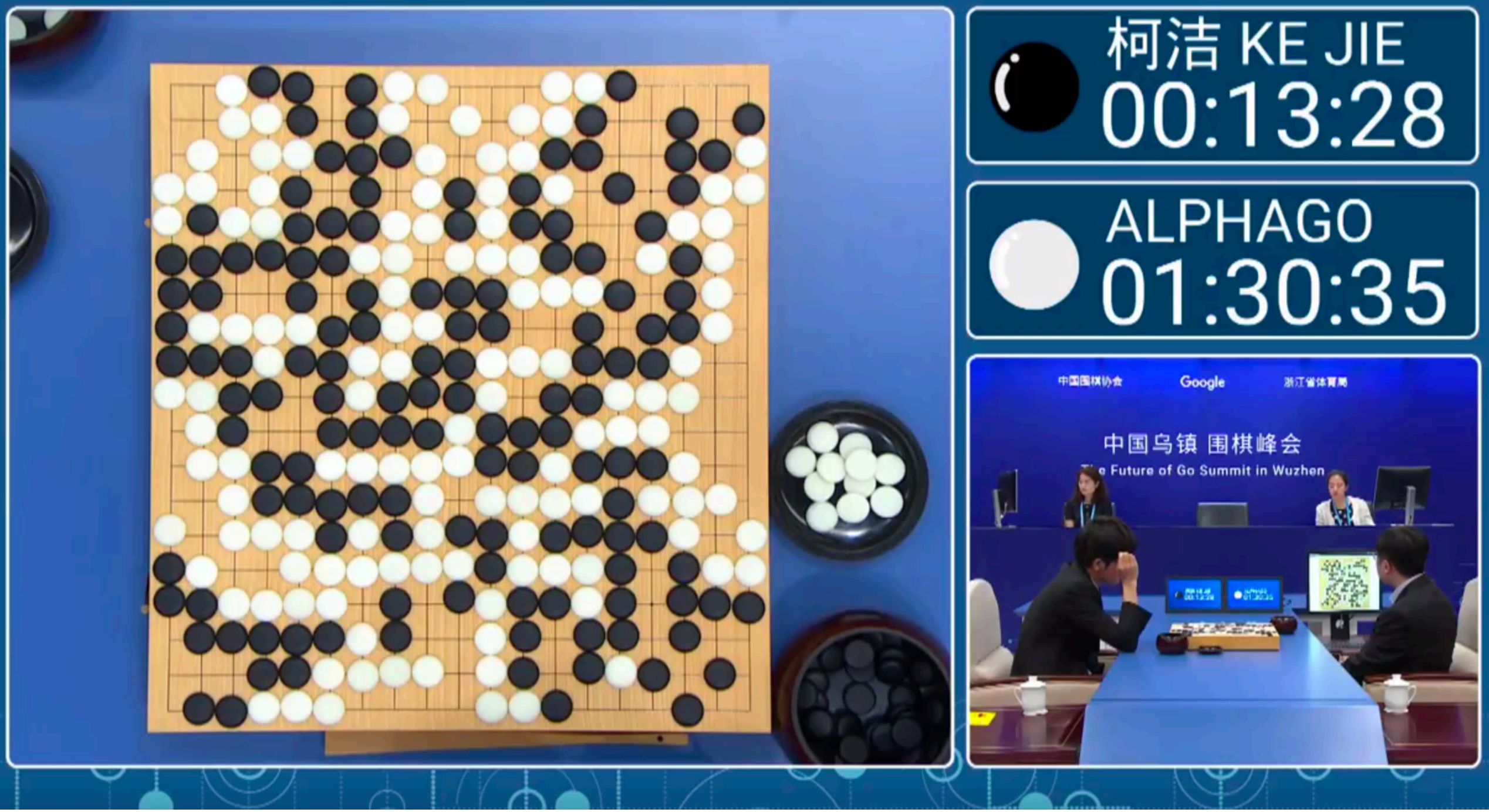


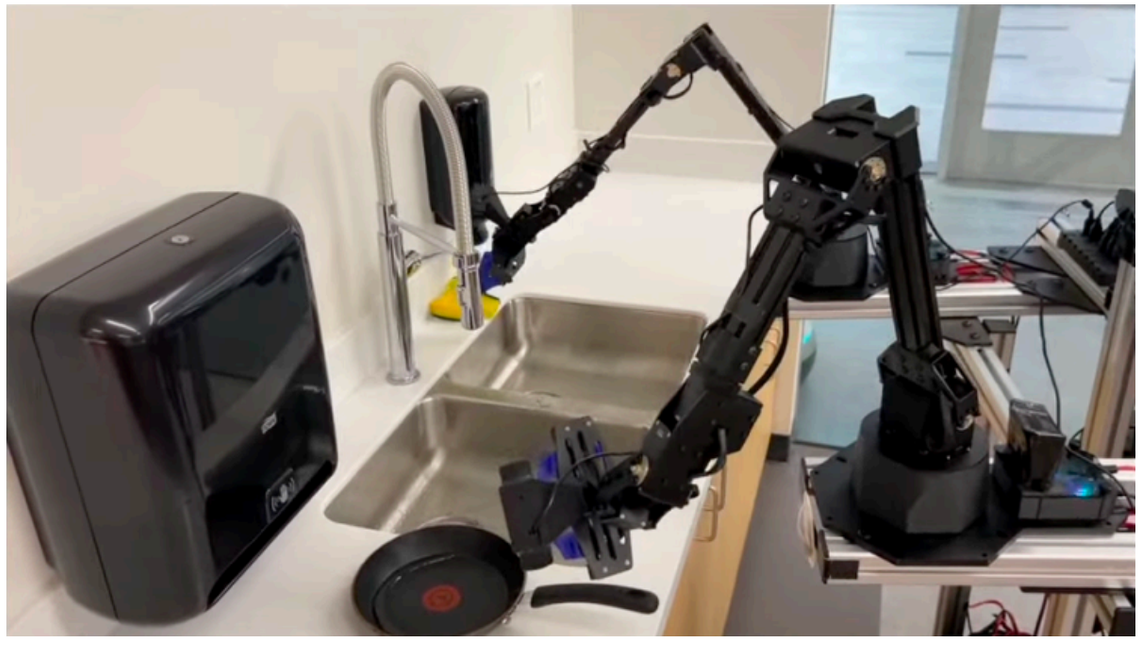
Image generation

AlphaGo


柯洁 KE JIE
00:13:28
ALPHAGO
01:30:35
中国围棋协会
Google
浙江省体育局
中国乌镇 围棋峰会
The Future of Go Summit in Wuzhen

Robot Control

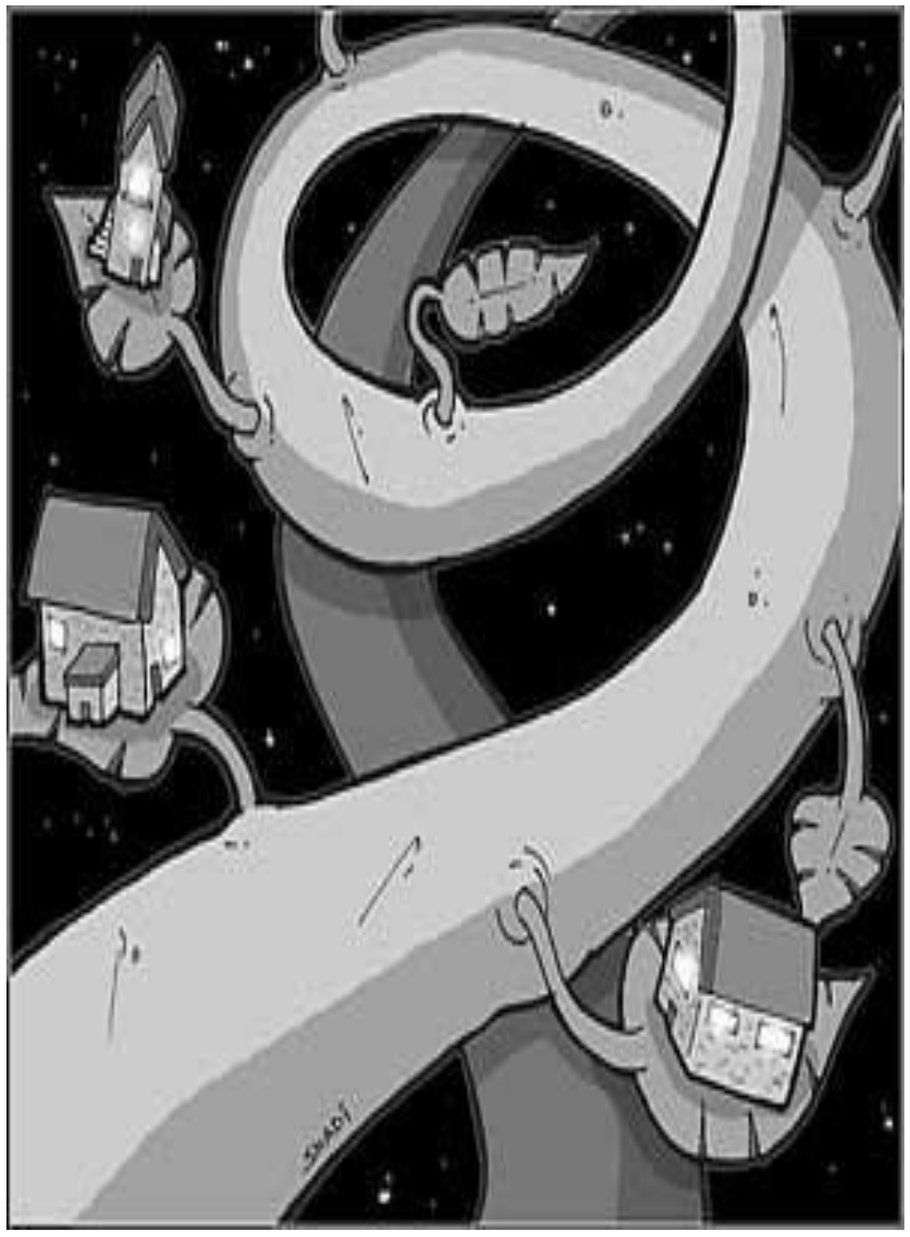
- Now cars can find their own ways!



Basics/Prerequisites

- Probability
  - Distributions, densities, marginalization, conditioning
- Statistics
  - Mean, variance, maximum likelihood estimation
- Linear algebra
  - Vector, matrix, multiplication, inversion, eigen-decomposition
- Algorithms and Programming
  - Python, Basic data structures, computational complexity
- Convex optimization
  - Basics will be covered during lecture

Machine learning for apartment hunting



- Suppose you are to move to Atlanta
- And you want to find the **most reasonably priced** apartment satisfying your **needs: (I know it is hard)**

square-ft., # of bedroom, distance to campus …

Living area (ft^2)	# bedroom	Rent ($)
230	1	600
506	2	1000
433	2	1100
109	1	500
…		
150	1	?
270	1.5	?

Linear Regression Model

- Assume y is a linear function of x (features) plus noise ϵ

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n + \epsilon$$

where ϵ is an error model as Gaussian $N(0, \sigma^2)$ ← Probability

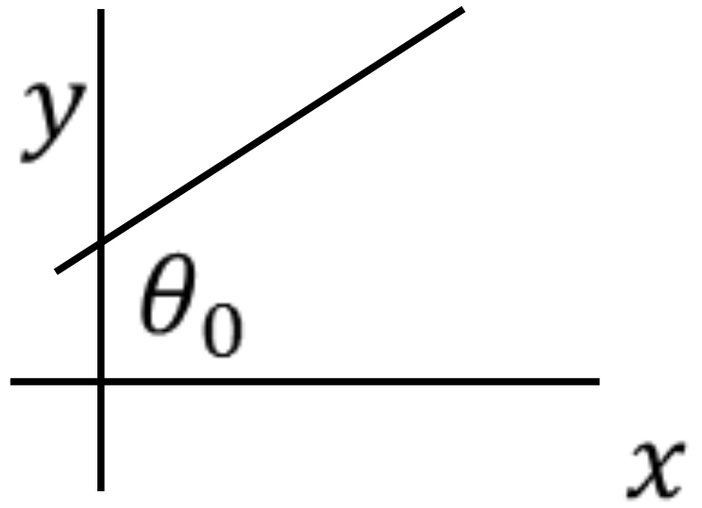
- Let $\theta = (\theta_0, \theta_1, \ldots, \theta_n)^\top$, and augment data by one dimension

Linear algebra

$$x \leftarrow (1, x)^\top$$

Then $y = \theta^\top x + \epsilon$

Linear algebra



Least mean square method

- Given m data points, find θ that minimizes the mean square error

$$\hat{\theta} = argmin_{\theta}\, L(\theta) = \frac{1}{m}\sum_{i}^{m}\left(y^i - \theta^\top x^i\right)^2$$

Optimization

Statistics

- Set gradient to 0 and find parameter

Optimization

Linear algebra

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m}\sum_{i}^{m}\left(y^i - \theta^\top x^i\right)x^i = 0$$

$$\Leftrightarrow -\frac{2}{m}\sum_{i}^{m} y^i x^i + \frac{2}{m}\sum_{i}^{m} x^i x^{i\top}\theta = 0$$

Statistics

Statistics

Matrix version of the gradient

- Define $X = (x^1, x^2, \ldots x^m), y = (y^1, y^2, \ldots, y^m)^\top$, gradient becomes

Linear algebra →

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m} Xy + \frac{2}{m} XX^\top \theta$$

Linear algebra →

$$\Rightarrow \hat{\theta} = (XX^\top)^{-1} Xy$$

Algorithms Programming →

- Matrix inversion in $\hat{\theta} = (XX^\top)^{-1} Xy$ expensive to compute
  - Gradient descent

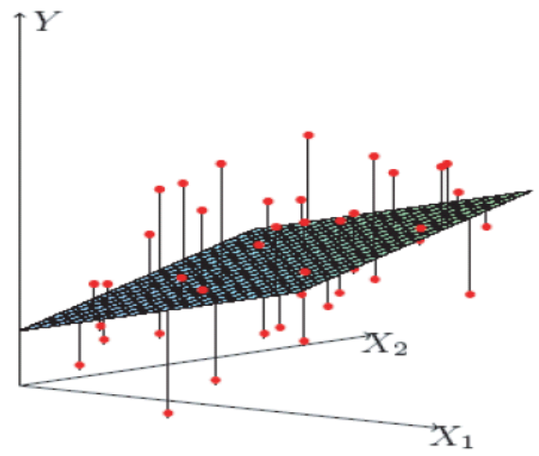
$$\hat{\theta}^{t+1} \leftarrow \hat{\theta}^t + \frac{\alpha}{m} \sum_i^m \left(y^i - \hat{\theta}^{t\top} x^i \right) x^i$$

Optimization ←

Probabilistic Interpretation of LMS

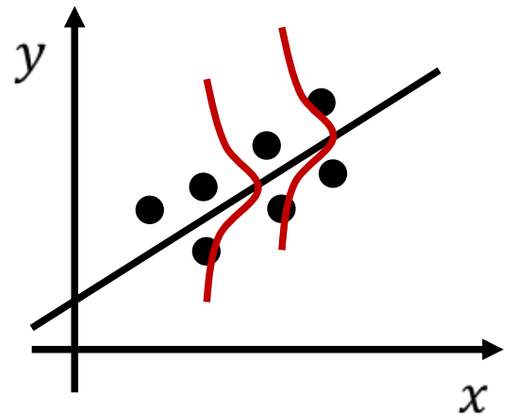
- Assume y is a linear in x plus noise ϵ

$$y = \theta^\top x + \epsilon$$



- Assume ϵ follows a Gaussian $N(0,\sigma)$

$$p\left(y^i \middle| x^i; \theta\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^i - \theta^\top x^i\right)^2}{2\sigma^2}\right)$$



- By independence assumption, likelihood is

$$L(\theta) = \prod_i^m p\left(y^i \middle| x^i; \theta\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(-\frac{\sum_i^m \left(y^i - \theta^\top x^i\right)^2}{2\sigma^2}\right)$$

Probability

Probabilistic Interpretation of LMS, cont.

- Hence the log-likelihood is:

$$\log L(\theta) = m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_i^m \left(y^i - \theta^\top x^i\right)^2$$

Statistics

- LMS is equivalent to MLE of θ !

$$LMS: \quad \frac{1}{m} \sum_i^m \left(y^i - \theta^\top x^i\right)^2$$

- How to make it work in real data?

Algorithms
Programming

Textbooks

- Machine Learning: a Probabilistic Perspective, Kevin Murphy algebra, and matlab.
  - Sufficient details for self-study

- Pattern Recognition and Machine Learning, Chris Bishop
  - Presented from probabilistic and graphical model view

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - Presented from statistical view

- These books offer complementary views

Grading

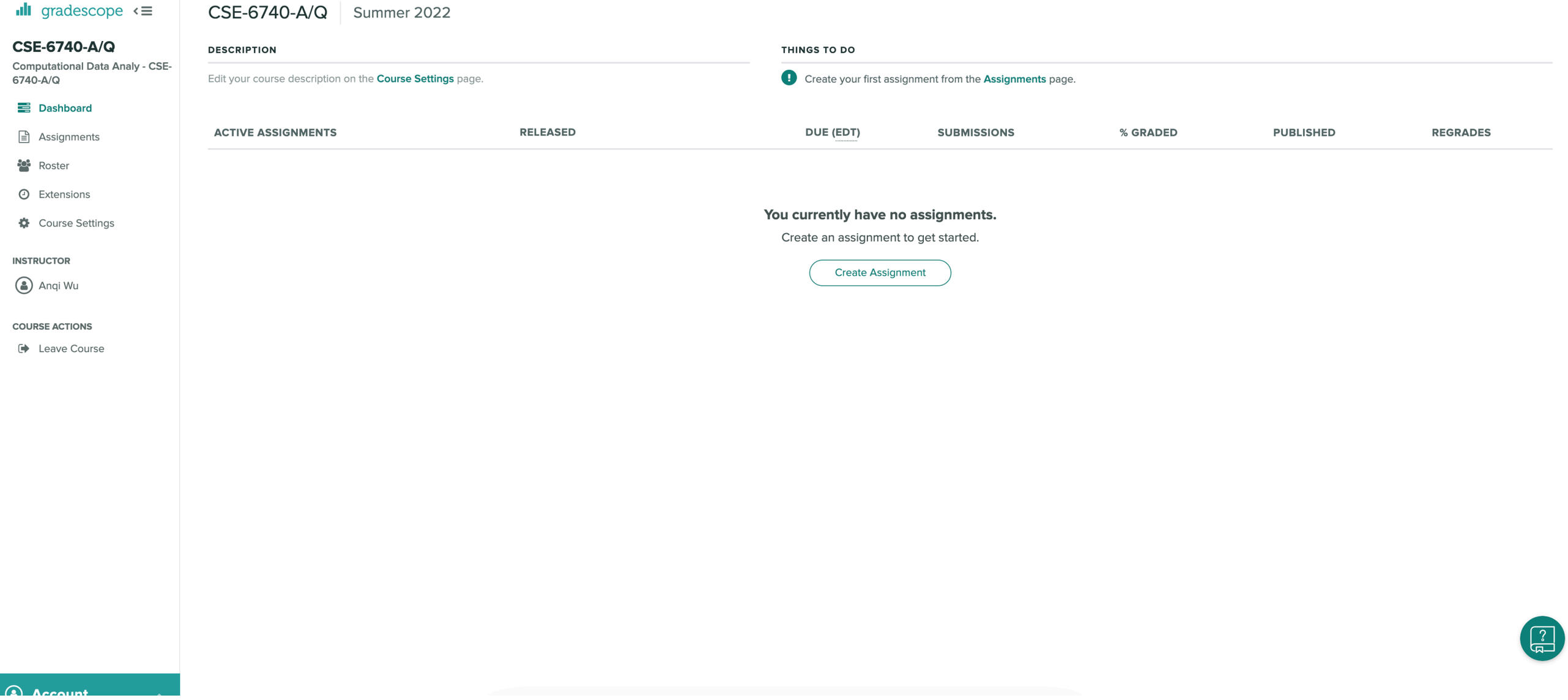
- 4 assignments, 70%
  - Approximately 1 assignment every 5 lectures
  - More next slide

- Midterm exam (10%)
  - 02/26 in class

- Final exam (15%)
  - 05/04 8:00 AM - 10:50 AM

- In-person attendance (5%)

Grading


gradescope
CSE-6740-A/Q
Computational Data Analy - CSE-6740-A/Q
Dashboard
Assignments
Roster
Extensions
Course Settings
INSTRUCTOR
Anqi Wu
COURSE ACTIONS
Leave Course
Account
CSE-6740-A/Q
Summer 2022
DESCRIPTION
Edit your course description on the Course Settings page.
THINGS TO DO
Create your first assignment from the Assignments page.
ACTIVE ASSIGNMENTS
RELEASED
DUE (EDT)
SUBMISSIONS
% GRADED
PUBLISHED
REGRADES
You currently have no assignments.
Create an assignment to get started.
Create Assignment

Assignments

- All homeworks are due by the beginning of class.
- Homework is penalized by 20% for each day that it is late (this applies additively, meaning that no credit is gained after 5 late days).
- We strongly encourage the use of LaTeX for your submission. Unreadable handwriting is subject to zero credit.
- We encourage you to discuss course content, homework problems, and project ideas with your classmates. However, all answers and codes should be prepared independently.
- If you refer to any material, it should be properly cited.
- If you discussed homework problems with your classmates, indicate which problems you discussed with whom. Any kind of academic misconduct is subject to F grade will be reported to the Dean of Students.

The Principle of Independent Thinking

In this course, we prioritize student thinking and analysis independent of AI-generated content.

- **The Nature of Learning:** Real learning involves recognizing what you don't know and connecting new knowledge personally.
- **Critical Usage:** You must move beyond being a consumer of AI to being a critical user who understands tool limitations.
- **Responsibility:** You are 100% responsible for the correctness and quality of your submitted work.

What is (and is NOT) Allowed

AI tools in this class are restricted to **technical support only**.

Permissible Uses

- **LaTeX Formatting:** Assistance with mathematical typesetting and document structure.
- **Programming Syntax:** Debugging or fixing syntax in `Matlab`, `Numpy`, or `Python`.
- **Polishing:** Improving the clarity and flow of *your own* written explanations.

Strict Prohibition

AI is **strictly prohibited** from generating mathematical derivations, probabilistic proofs, or any core conceptual content.

Enforcement and Point Loss

To protect the integrity of your learning, we enforce the following:

- **Content Generation Penalty:** If the key answer or derivation of a question is generated by AI, **you will lose all points for that question**.
- **Citations Do Not Exempt:** This penalty applies even if you provide a citation or reference.
- **Academic Credit:** Points are awarded only for your independent logic and unique contributions.

How We Monitor Academic Integrity

We treat AI-generated content as work produced by someone other than the student.

- **Detection Tools:** We use specialized software to audit submissions for AI patterns in both text and math.
- **Refined Assessment:** Grading will focus on your unique contributions and the alignment of research questions and methods.

The Requirement for Every Submission

For every assignment, you must include a 150-300 word acknowledgment:

1. **Identification:** Which tool(s) did you use?
2. **Explanation:** Why did you decide to use it (e.g., LaTeX fixing)?
3. **Documentation:** Describe exactly how the tool was used to manage assignment requirements.
4. **Reflection:** Explore what worked, what didn't, and acknowledge tool limitations.

**If you used no AI, use this space to highlight your non-AI approach.*

Proper Documentation Practices

Situate all AI use within established citation practices:

- **As Third-Party:** AI content must be documented and, if relevant, quoted or paraphrased.
- **Format:** Title of AI Tool. Prompt or brief description of topic. Date of creation.
- **Official Guides:** Follow the provided APA and MLA links for citing generative tools.
  - APA provides guidance here: https://apastyle.apa.org/blog/how-to-cite-chatgpt
  - MLA provides guidance here: https://style.mla.org/citing-generative-ai/

Closing Guidance

- Use AI as a tool for **expression**, not for **thinking**.
- Ensure your work demonstrates knowledge in spontaneous and interpersonal ways.
- **When in doubt:** Ask before you submit.

Questions?