SEMI-SUPERVISED LEARNING (SSL)

Bridging the Gap between Labeled and Unlabeled Data

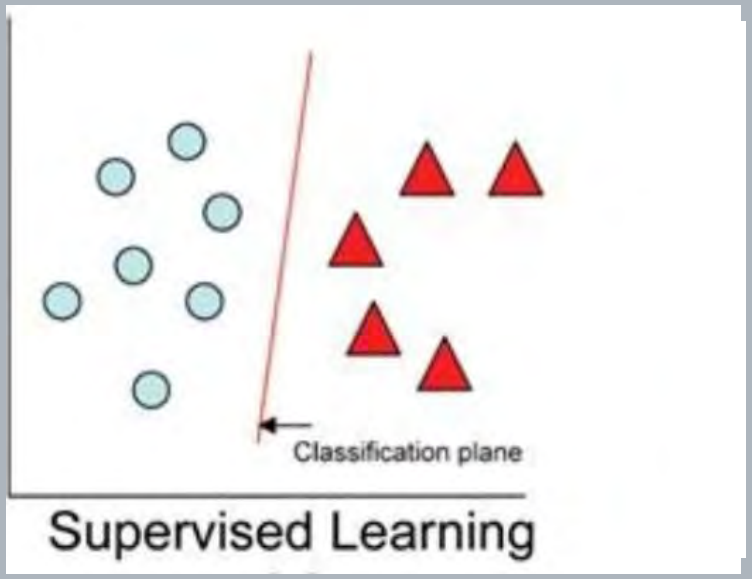
BY:
OLOLADE AKINSANOLA
TIGIST TEFERA

COMP 479: Machine Learning

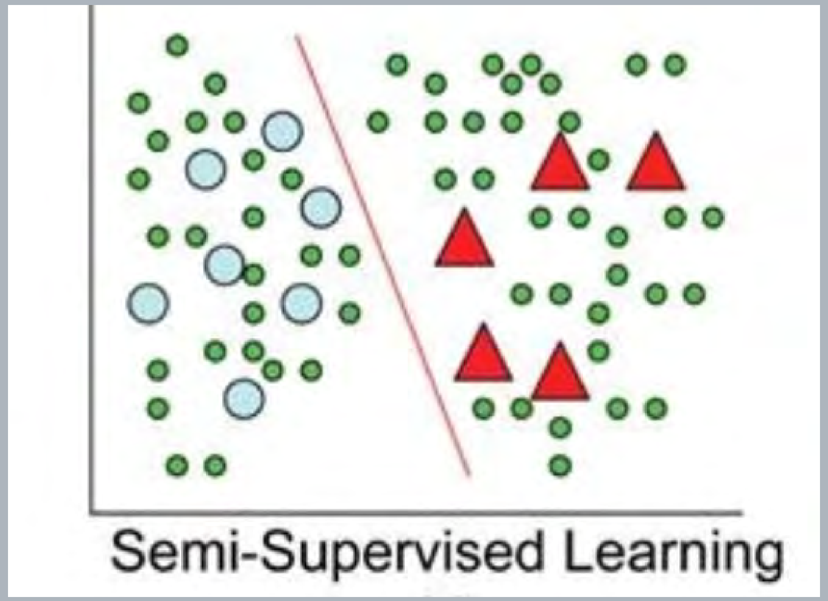
TABLE OF CONTENT

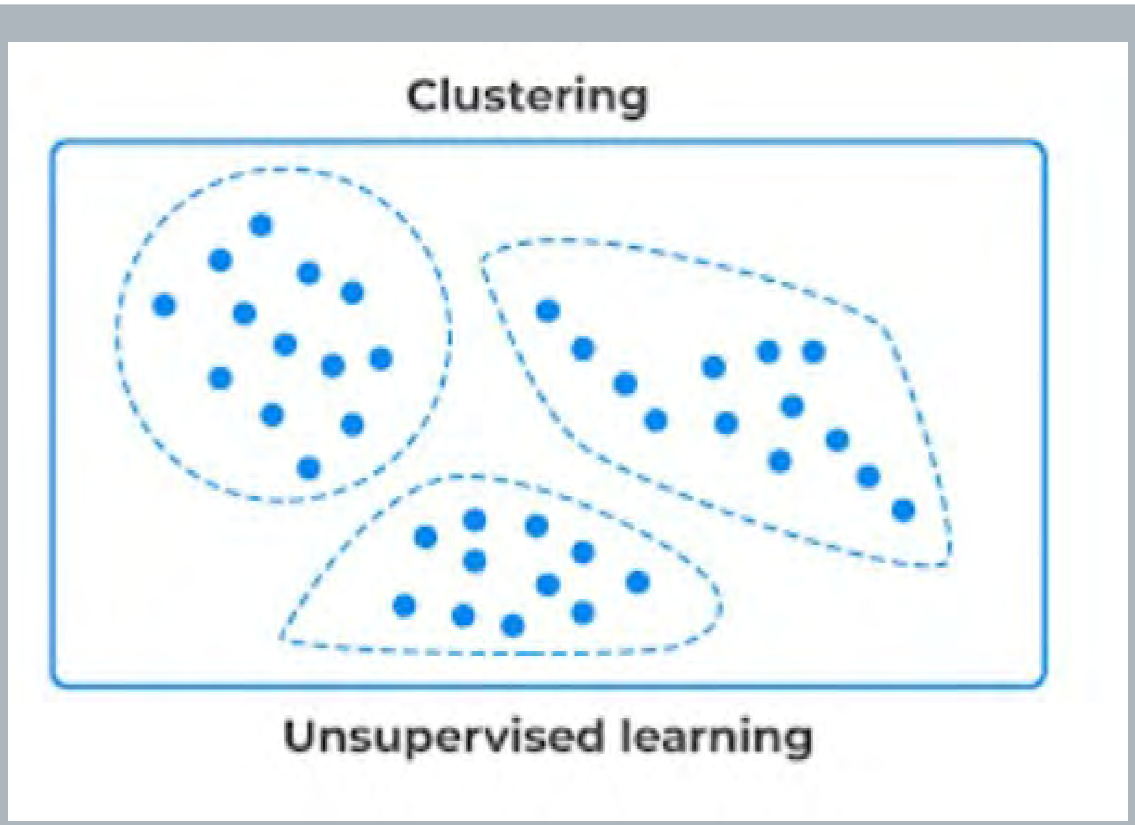
1. Introduction to SSL
2. Core Assumptions in SSL
3. General SSL Pipeline (Example Workflow)
4. SSL Algorithms
    - Self Training
    - Generative Models
    - Graph-Based Algorithms
5. Real-World Applications
6. Pros and Cons
7. Challenges and Research Directions
8. Conclusions + Questions

INTRODUCTION

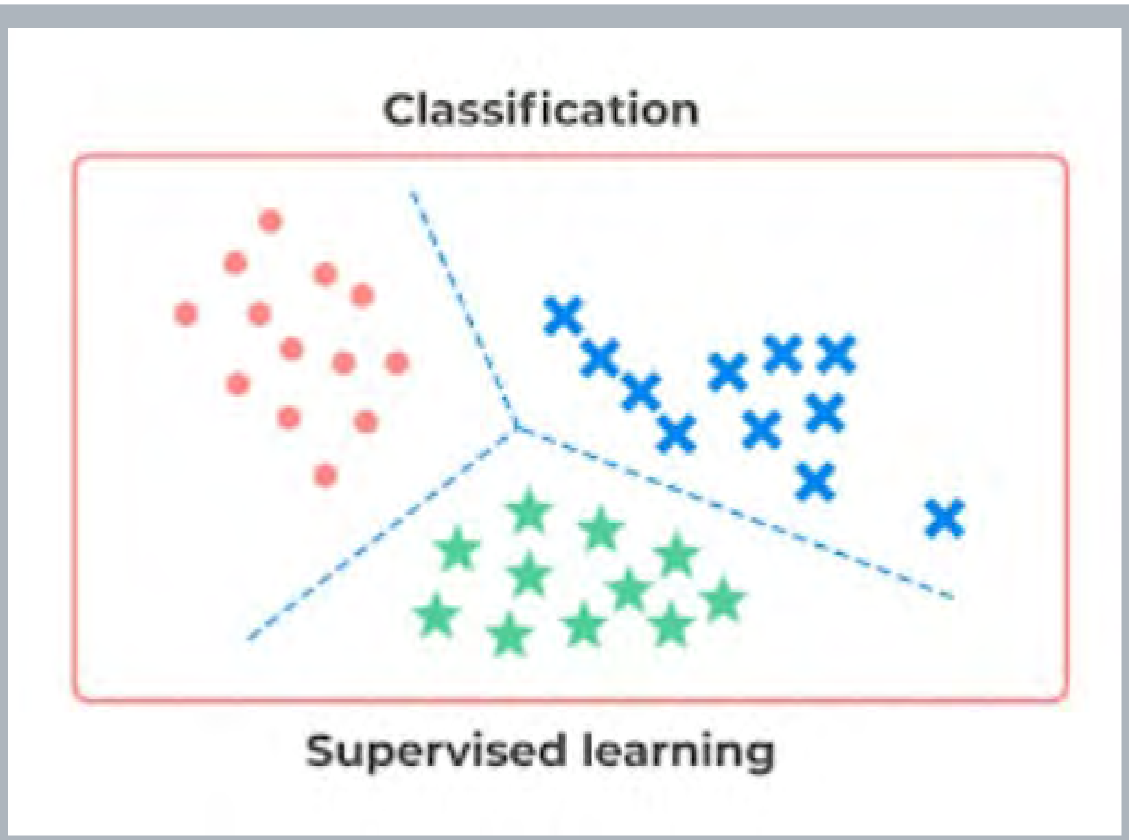


Supervised Learning

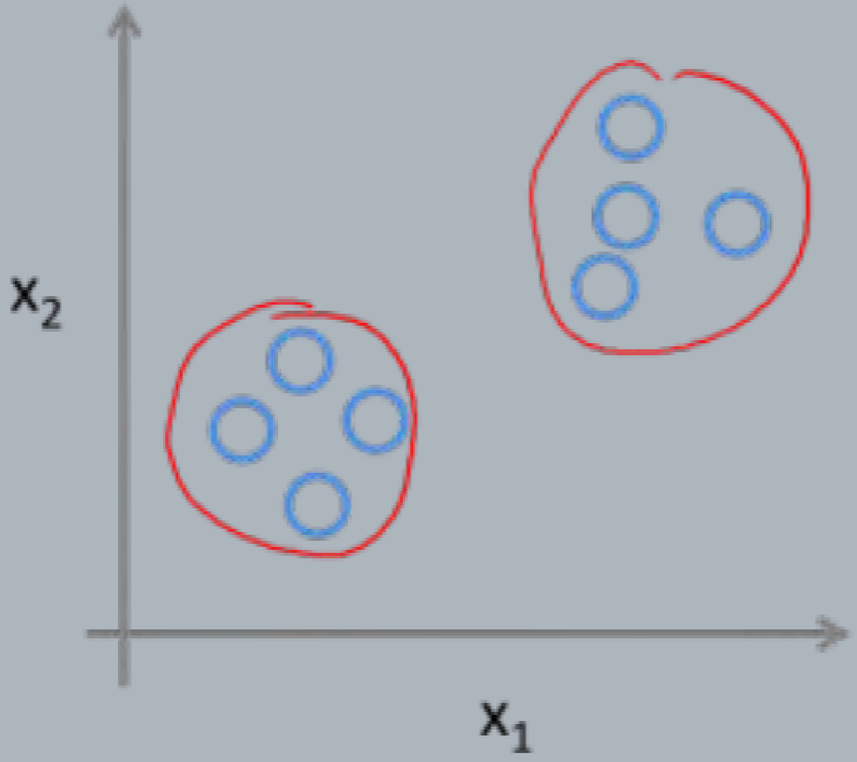

Semi-Supervised Learning

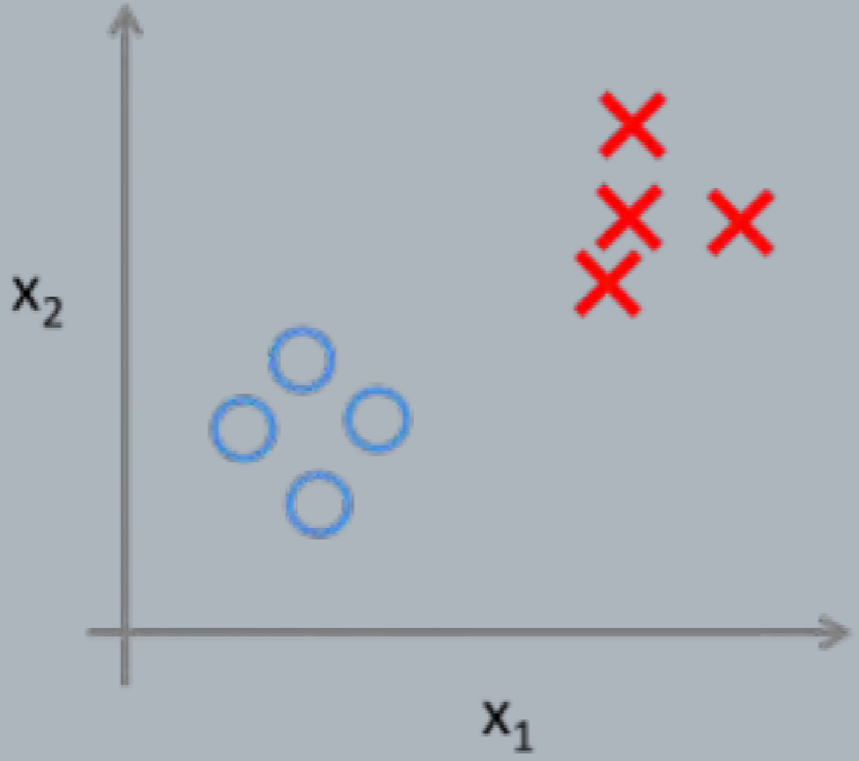

Clustering

Unsupervised learning

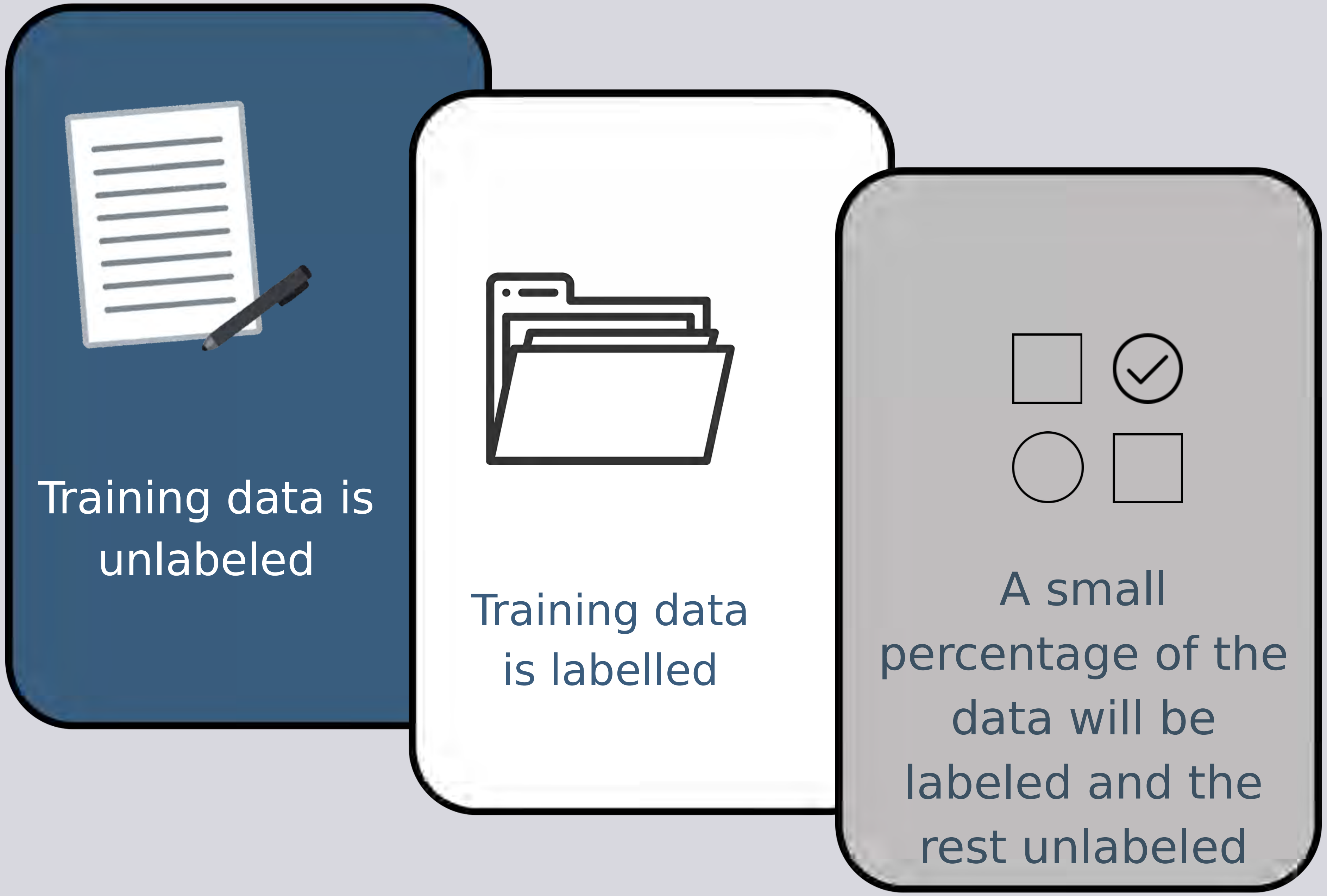

Classification

Supervised learning





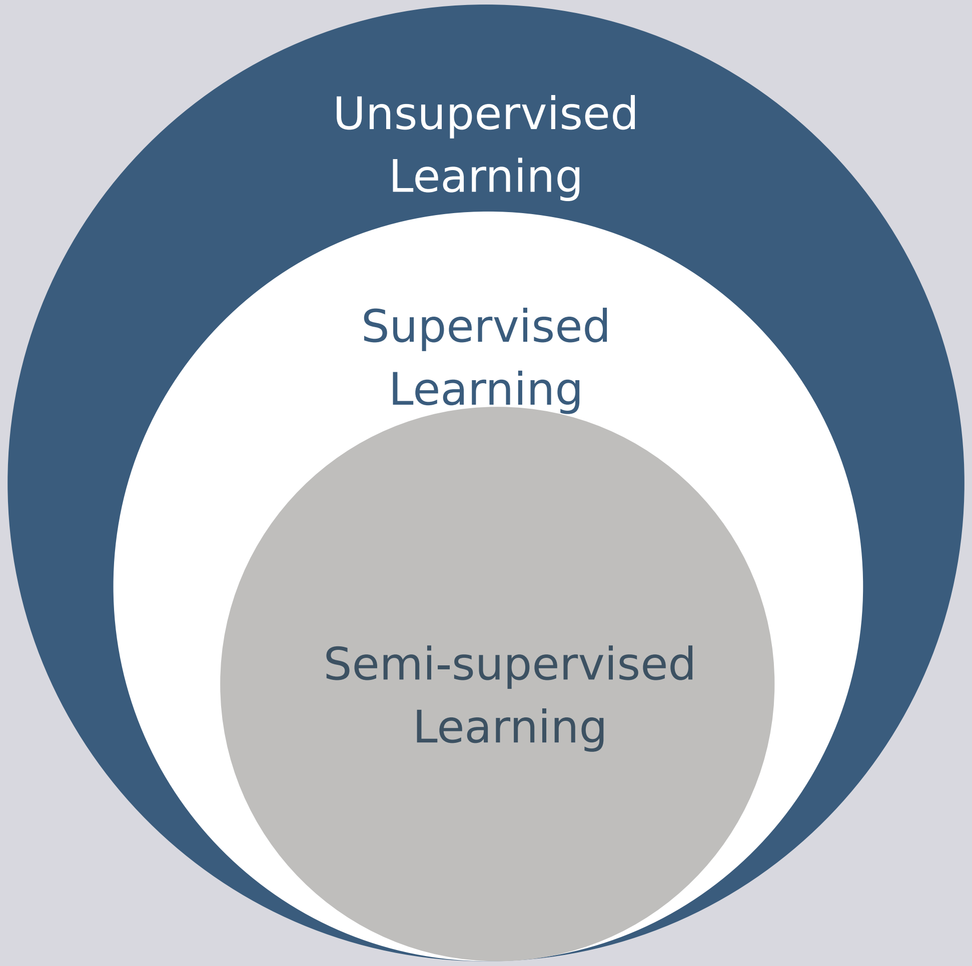
WHY SEMI-SUPERVISED LEARNING?



Training data is unlabeled

Training data is labelled

A small percentage of the data will be labeled and the rest unlabeled



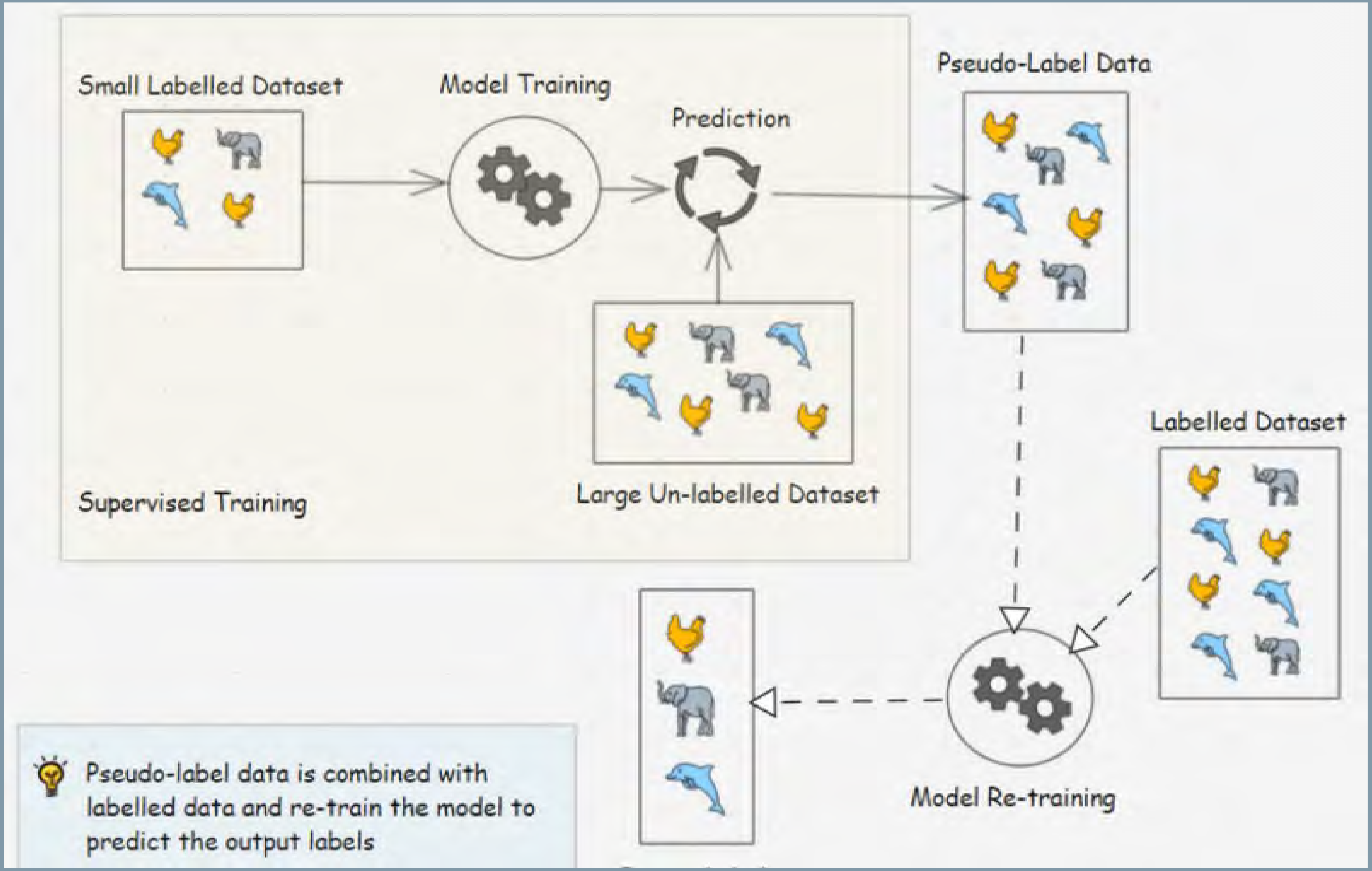
Unsupervised Learning

Supervised Learning

Semi-supervised Learning

- Unlabeled data is cheap and everywhere.
- Labeled data is expensive to get:
  - human annotation is boring
  - labels may require expert or special devices which might not be unique

GENERAL SSL PIPELINE: EXAMPLE WORKFLOW



Pseudo-label data is combined with labelled data and re-train the model to predict the output labels

- Train on small labeled set
- Predict on unlabeled data
- Keep confident Predictions
- Retrain on combined data

EXAMPLES

HARD-TO-GET LABELS

Task: speech analysis

- Switchboard dataset
- Telephone conversation transcription
- 400 hours annotation time for each hour of speech

film ⇒ f ihn uhgln m

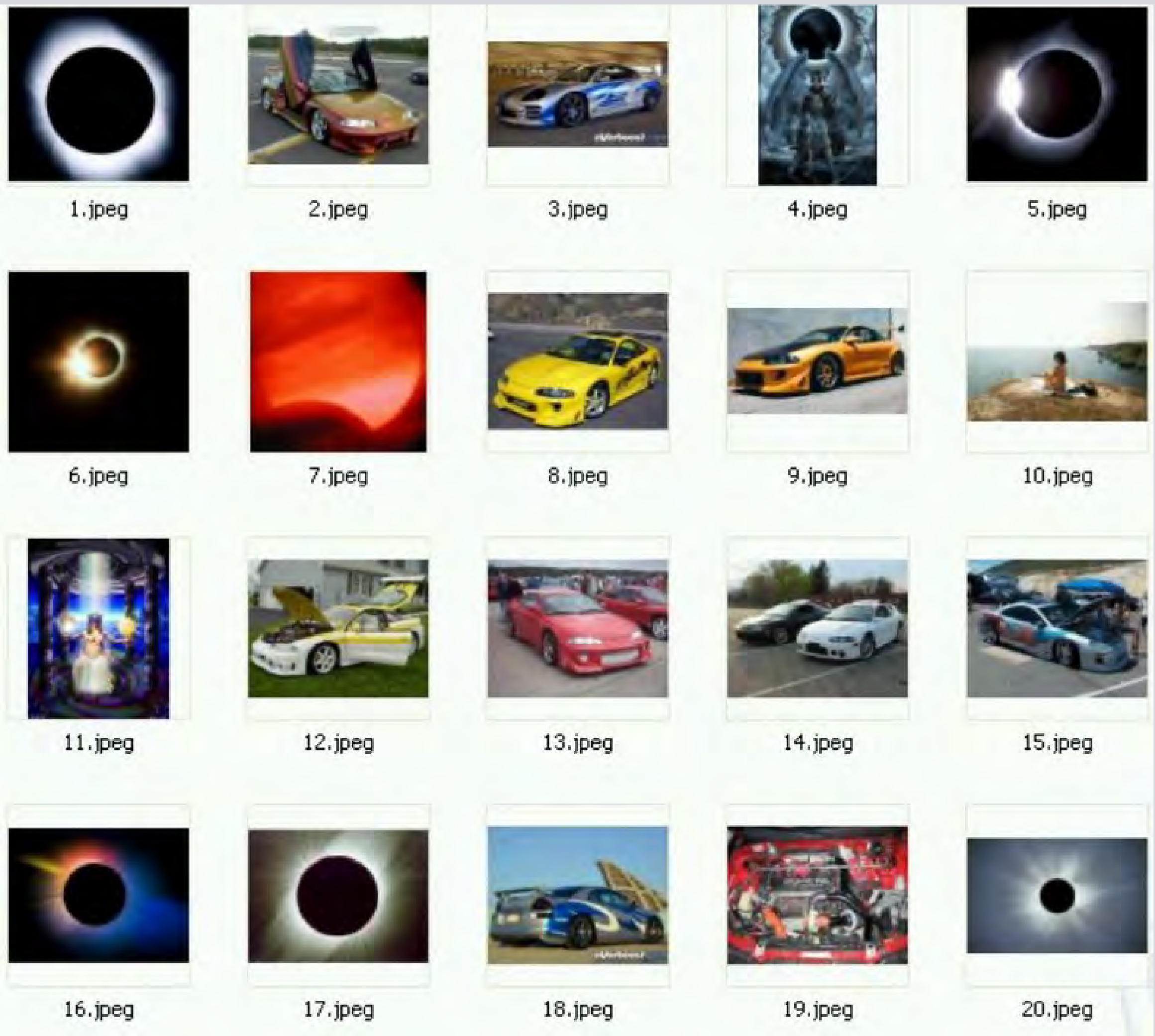
be all ⇒ bcl b iy iy_tr ao_tr ao l_dl

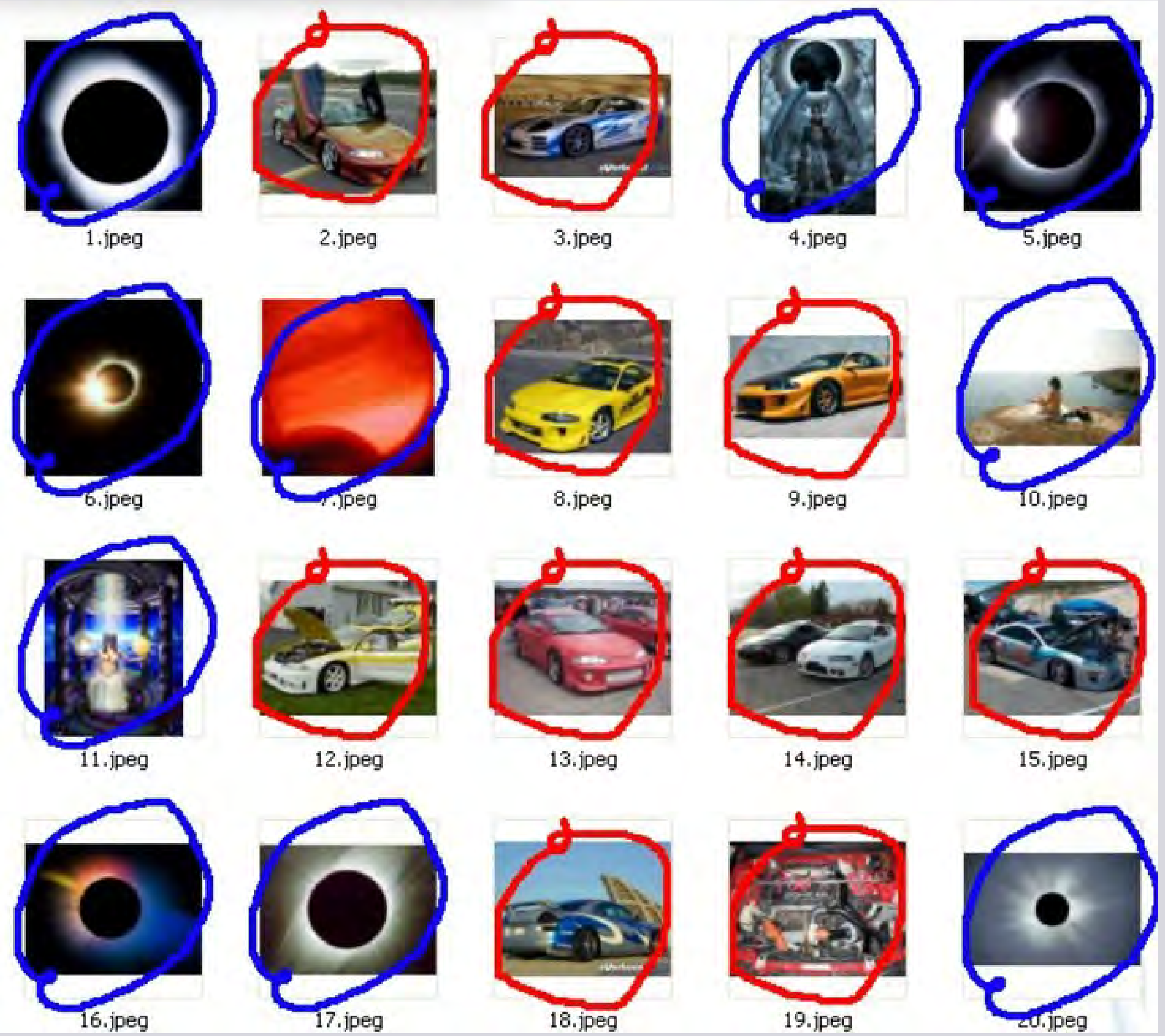
Task: natural language parsing

- Penn Chinese Treebank
- 2 years for 4000 sentences

NOT-SO-HARD-TO-GET LABELS

Task: Image Categorization of eclipse





THE LEARNING PROBLEM

Goal

Use both labeled and unlabeled data to build better models, than using each one alone.

Notations

- input instance *x,* label *y*
- learner $f : X \rightarrow y$
- labeled data $(X_i, Y_i) = \{(x_{1:i.}y_{1:i})\}$
- unlabeled data $X_u = \{x_{i+1:n}\}$, available during training
- usually $i \ll n$
- test data $X_{test} = \{x_{n+1:}\}$, not available during training

TYPES OF SSL

TRANSDUCTIVE LEARNING

- Does not generalize to unseen data (fits only your current dataset)
- Only concerned with unlabeled data
- Produces labels only for the data at training time
  - Assumes labels
  - Train classifier on assumed labels

Real- life application

Medical Imaging: Labeling all unlabeled MRI scans in a specific hospital dataset to help radiologists diagnose tumors, without needing to generalize to new scans.

INDUCTIVE LEARNING

- Does generalize to unseen data (generalize to new data)
- Not only produces labels, but also the final classifier
- Manifold Assumption
- Ultimately applied to the test data

Real- life application

Spam Detection: Training on a small set of labeled emails + large unlabeled corpus to classify future emails, adapting to new spam patterns.

WHEN CAN SSL WORK?

Smoothness Assumption

- 2 points x1, x2 are close, then the outputs y1, y2 must be close too.
- Density is considered:
  - label function is smoother in high-density than in low-density regions.
- By transitivity if 2 points are:
  - Linked by a path of high density then their outputs are close.
  - Linked by a path of low density then their outputs need not be close.
- Applicable to both classification and regression.

Cluster Assumption

- Points in same cluster are in the same class.
- Sets of points are connected by short curves which transverse only high-density regions.
- Decision boundary lies in a low-density region (*low-density separation*).
- Low density vs high density separation gives assumptions that are more sensible in many real-worlds problem.
- Different algorithms for both.
- E.g. Distinguish a handwritten digit “0” and “1”.

Manifold Assumption

- High-dimensional data lies on low dimensional manifold.
- Useful for curse of dimensionality.
- Learning algorithm (data in low-dimensional manifold)operates in a space of corresponding dimension (avoids curse of dimensionality).

Transduction

- Follows *Vapnik’s principle:* Do not solve a more difficult problem as an intermediate step.
- Estimates finite set of test labels (*f: Xu* *y*).⟶
- Takes advantage of unlabeled data.

SSL ALGORITHMS

1.SELF TRAINING

Idea: If I am highly confidence in a label of examples, I am correct.

Algorithm: Given a training set *T = {Xi}*, and unlabeled set *U={Uj}*

- Train *f* from $(X_{i,}Y_i)$
- Predict on $x \in X_u$
- Add $(x, f(x))$ to labeled data
- Repeat

Variations in Self Training

- Add a few most confident $(x, f(x))$ to labeled data
- Add all $(x, f(x))$ to labeled data
- Add all $(x, f(x))$ to labeled data, weigh each by confidence

E.g.: image categorization

Works based on smoothness and cluster assumptions

ADVANTAGES

- The simplest and fast SSL method
- Often used in real tasks like natural language processing
- Applies to existing (complex) classifiers

DISADVANTAGES

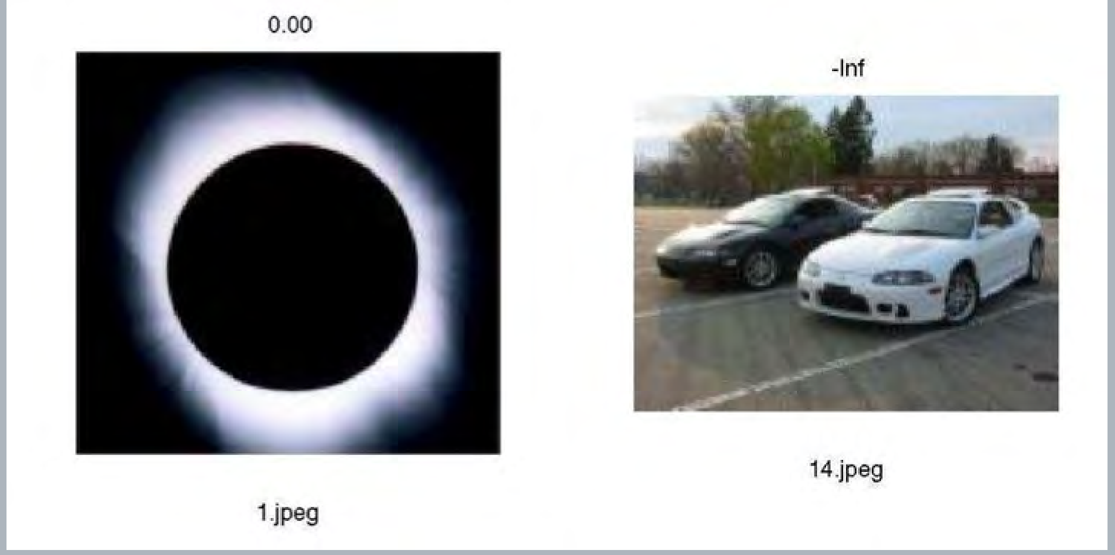
- Early mistakes could reinforce themselves
- Amplifies noise in data
- Requires explicit definition of *P(y|x)*
- Hard to implement for discriminative classifiers (SVM)

[Initial Model] → [Predict on Unlabeled] → [Add Confident Predictions] → [Retrain Model]

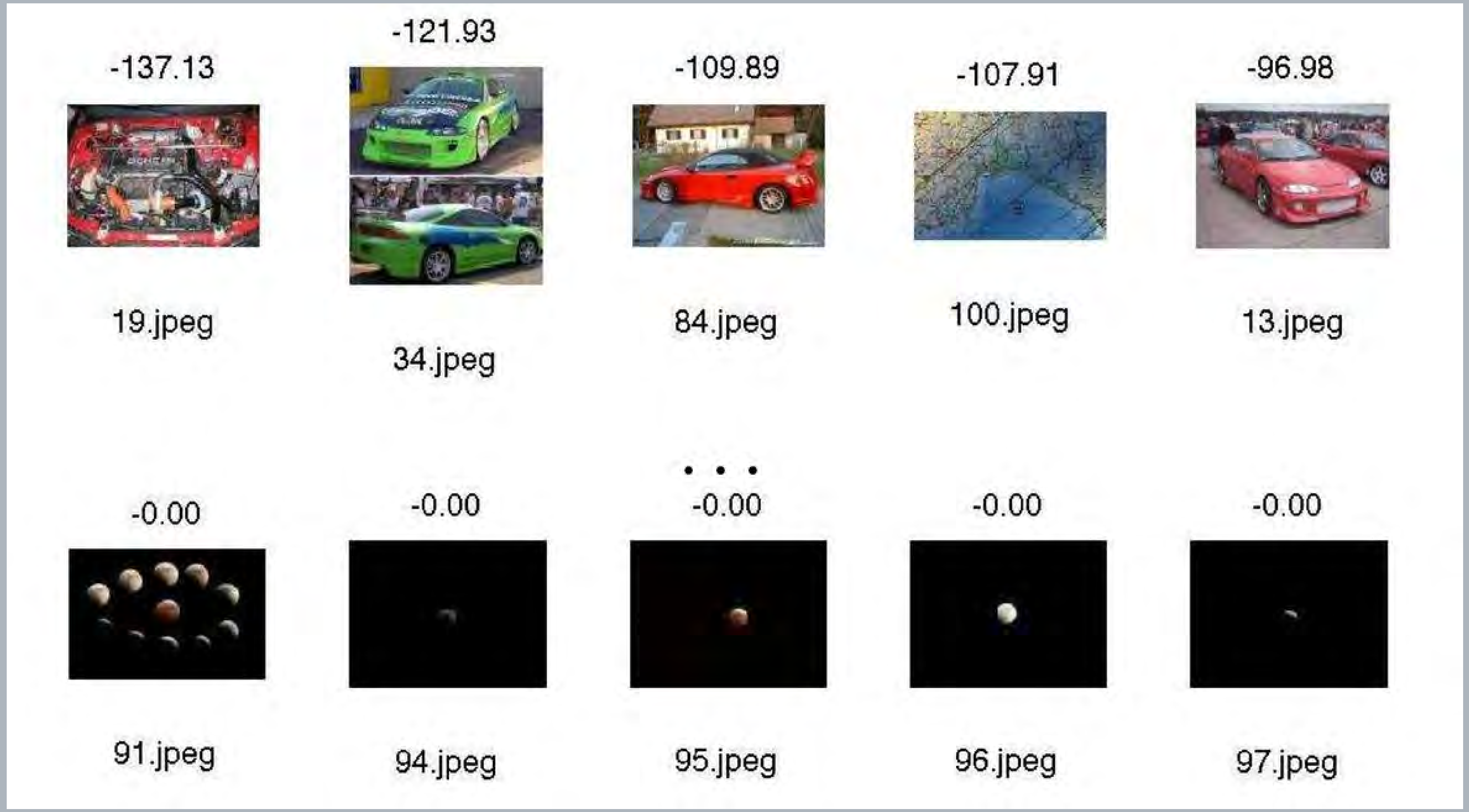
↑____|

SELF TRAINING EXAMPLE: IMAGE CATEGORIZATION

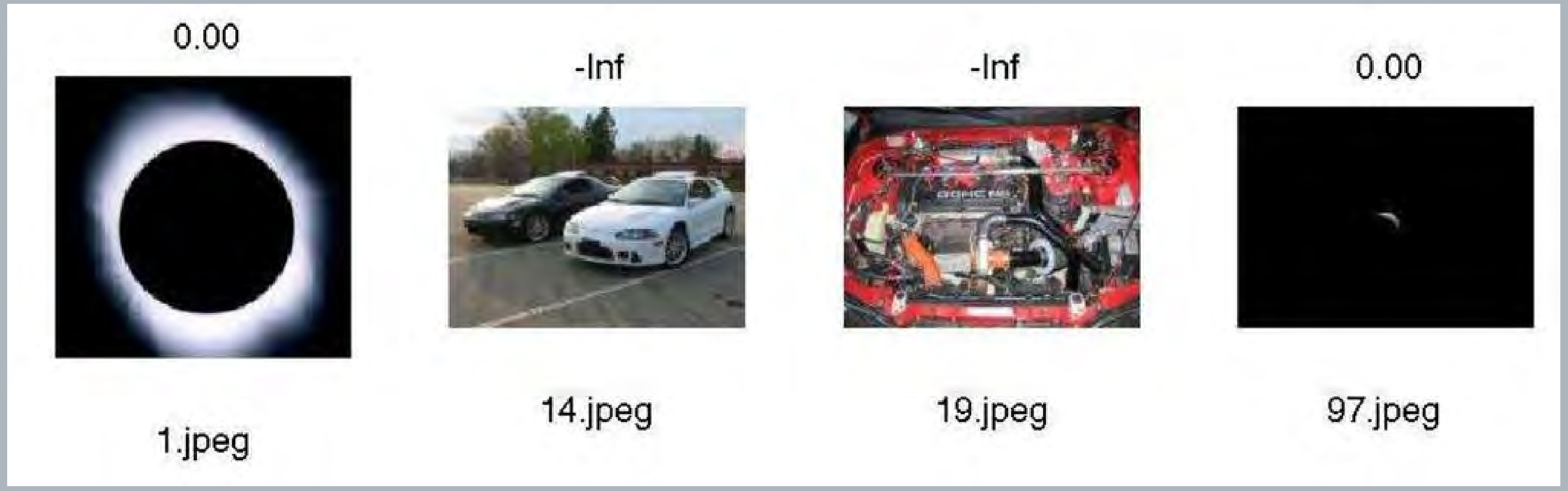
1. Train a naive Bayes classifier on the two initial labeled images



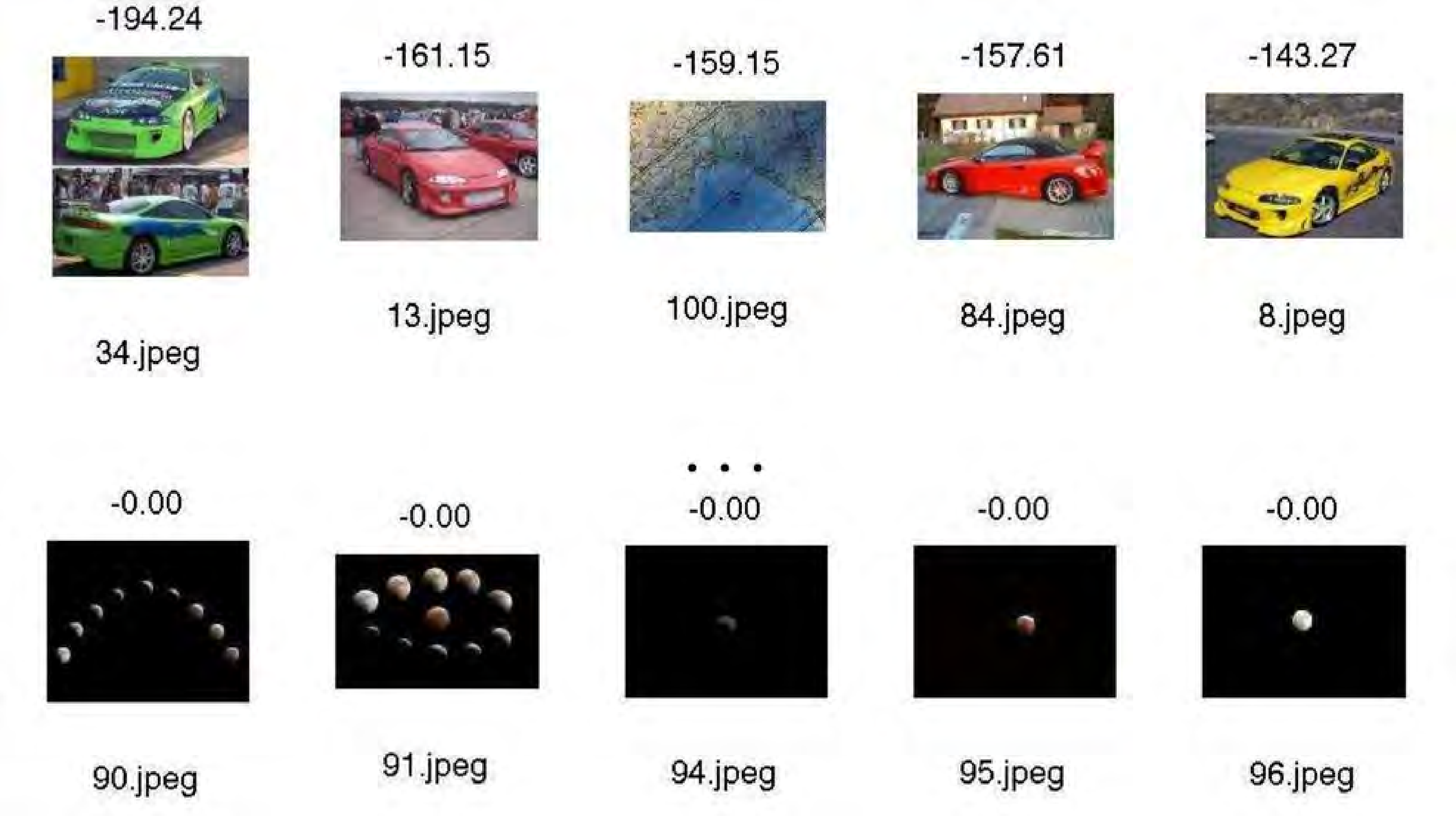
2. Classify unlabeled data, sort by confidence *log p(y = astronomy| x)*



3. Add the most confident images and predicted labels to labeled data



4. Re-train the classifier and repeat

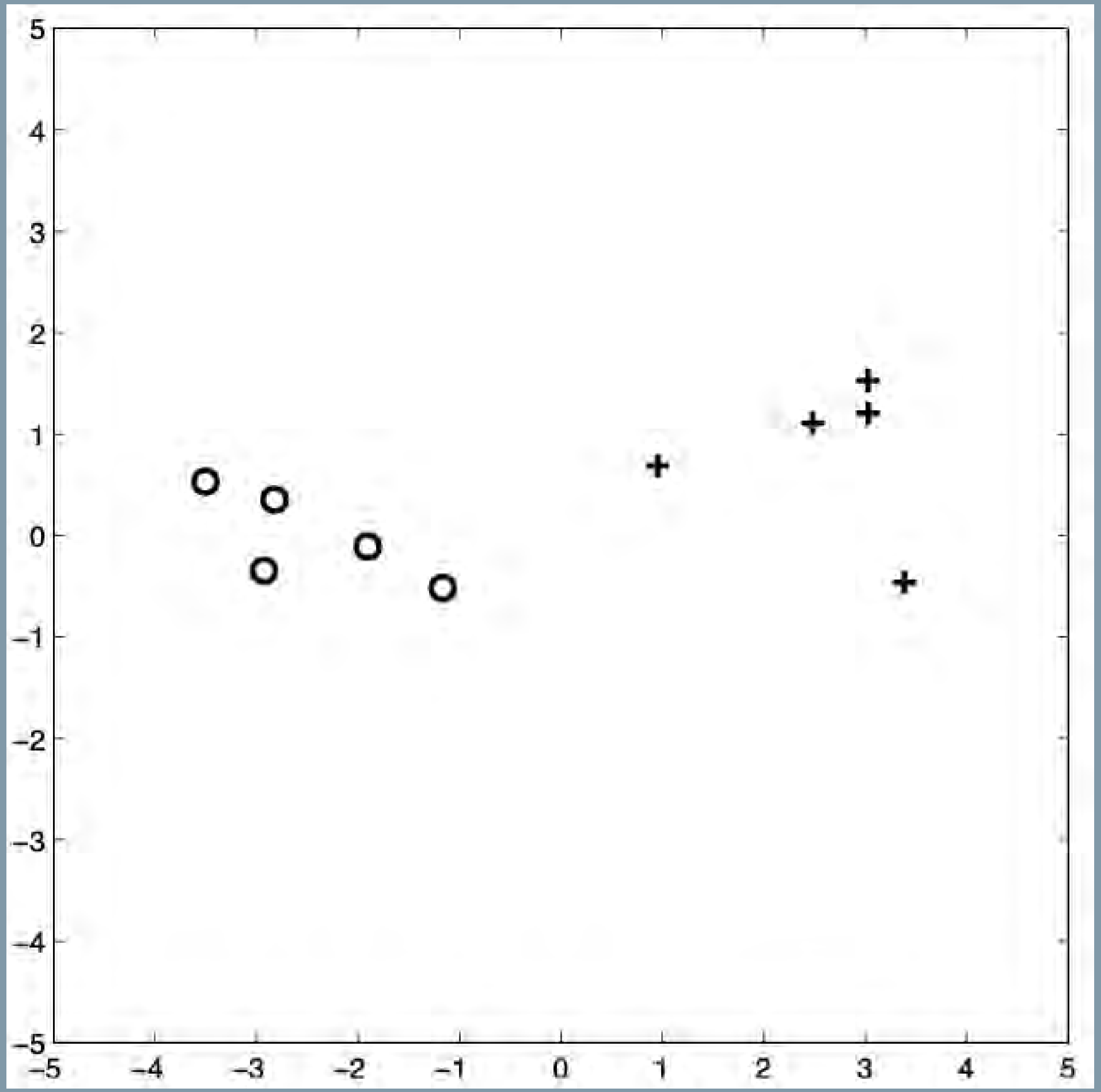


SSL ALGORITHMS

2. GENERATIVE MODELS

Idea: Assumes distribution using labeled data, update using unlabeled data

Labeled data *(Xi,Yi)* and the boundary decision:



Assuming each class has a Gaussian distribution,
what is the decision boundary?

Model parameters: $\theta = \{w_{1,} w_{2,} \mu_{1,} \mu_{2,} \Sigma_{1,} \Sigma_2\}$

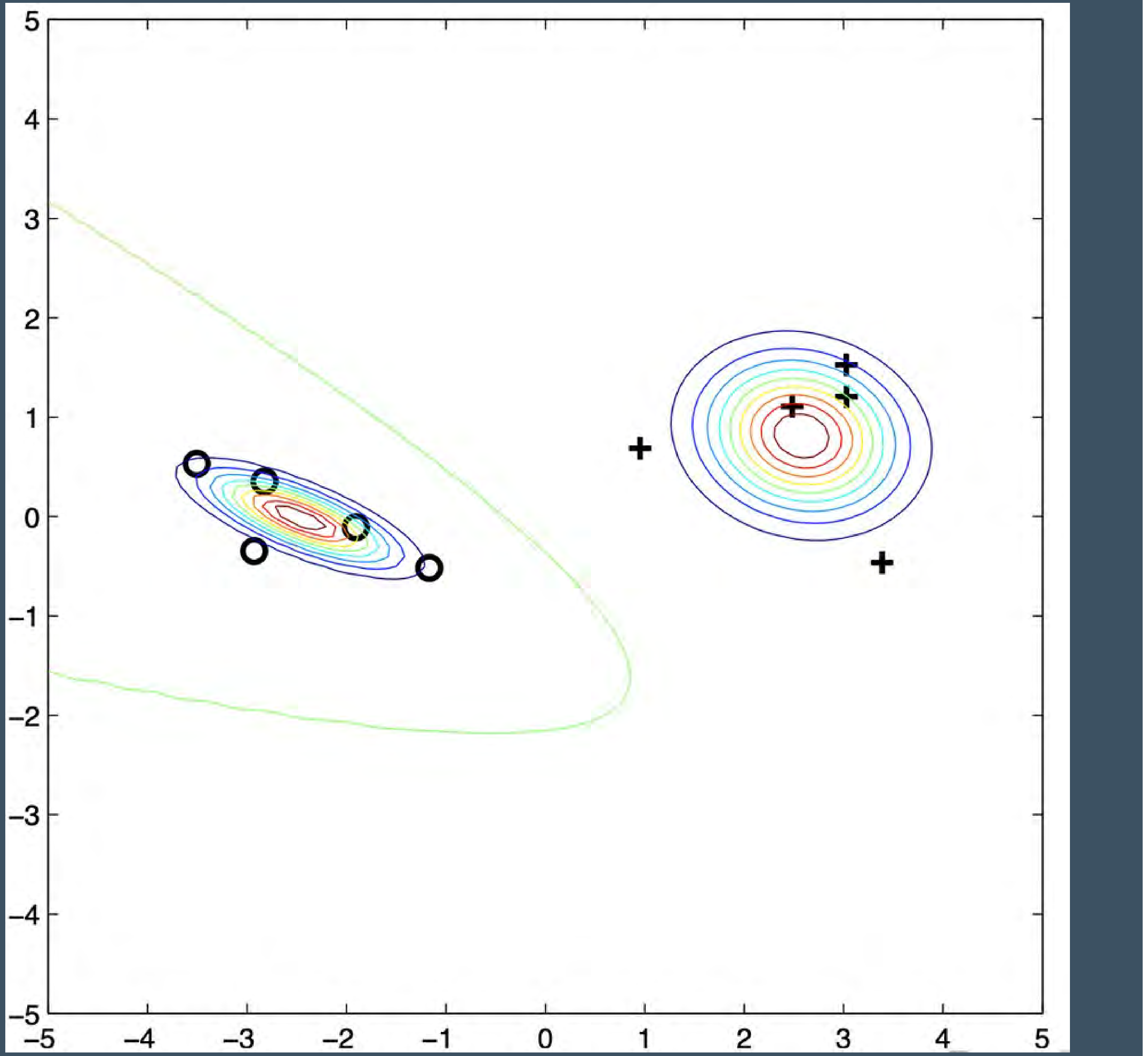
The GMM:
$$p\left(x, y \left|\theta\right|\right) = p\left(y \left|\theta\right.\right) p\left(x \left|y, \theta\right.\right)$$
$$= w_y N\left(x; \mu_{y,} \Sigma_y\right)$$

Classification: $p\left(y \left|x, \theta\right.\right) = \dfrac{p\left(x, y \left|\theta\right.\right)}{\Sigma_{y'} p\left(x, y' \left|\theta\right.\right)}$

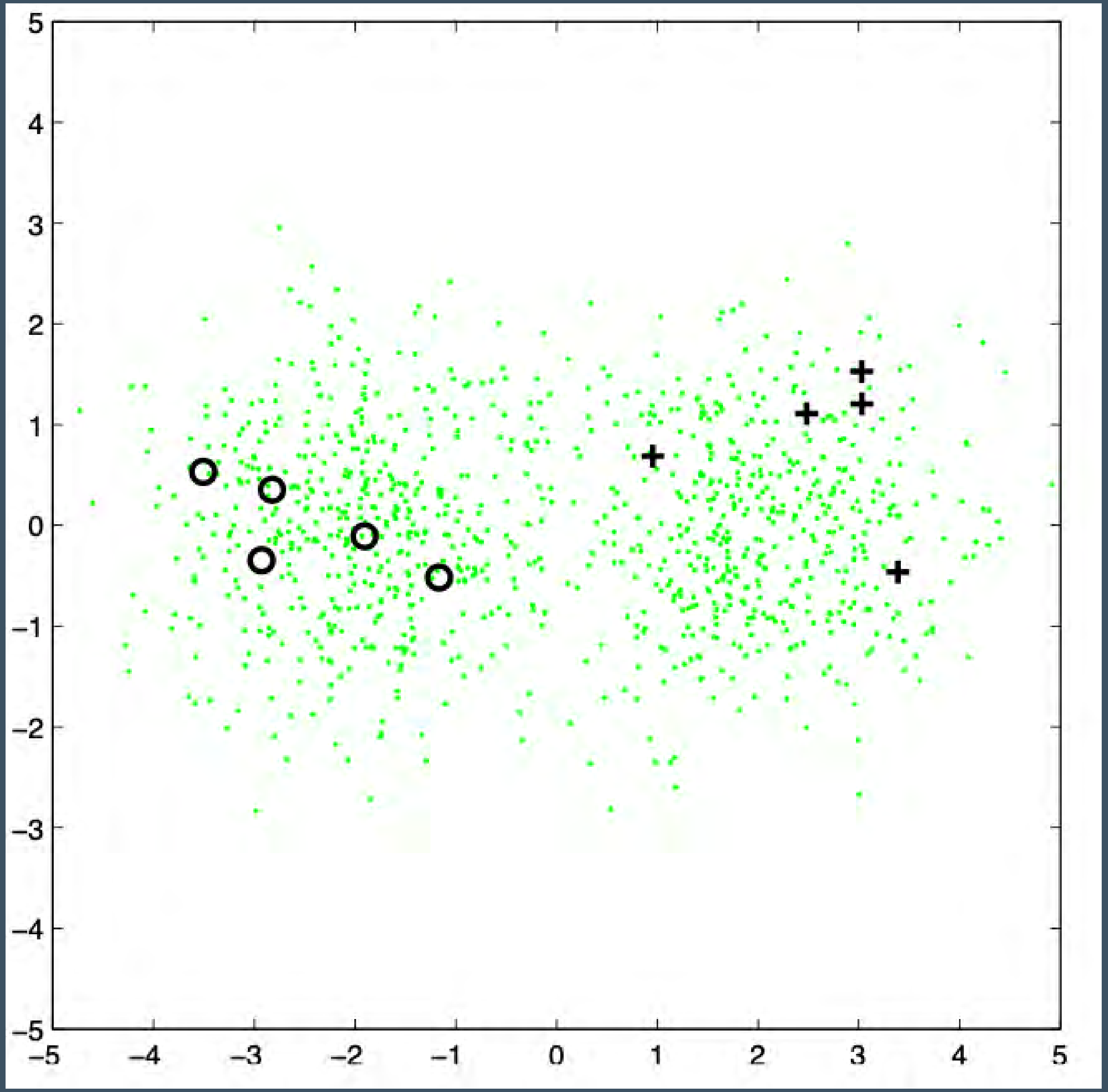
Works based on manifold and cluster assumptions

GENERATIVE MODELS: A SIMPLE EXAMPLE

The most likely model, and its decision boundary:

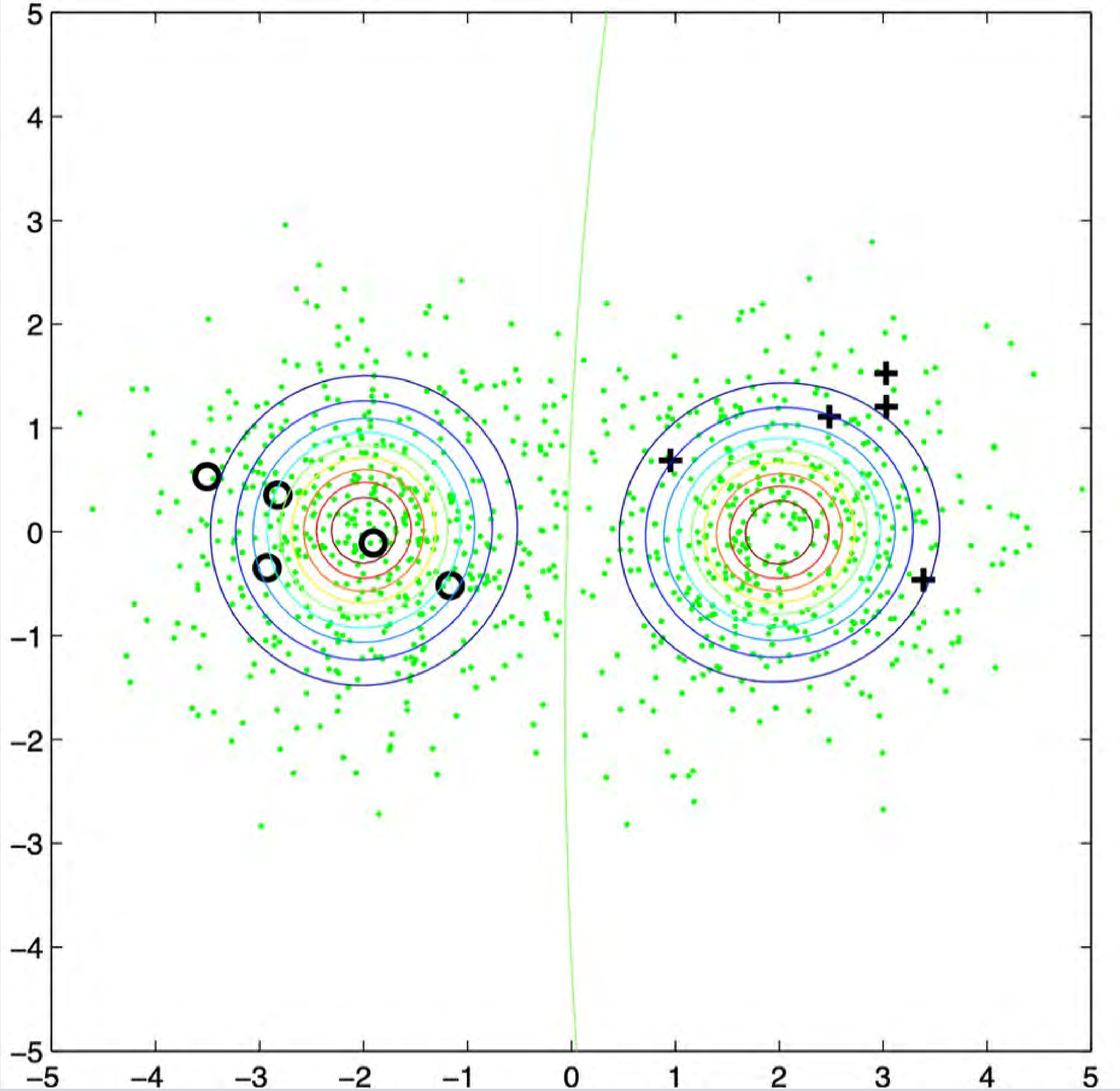


Adding Unlabeled data *(Xu)*, then boundary decision:

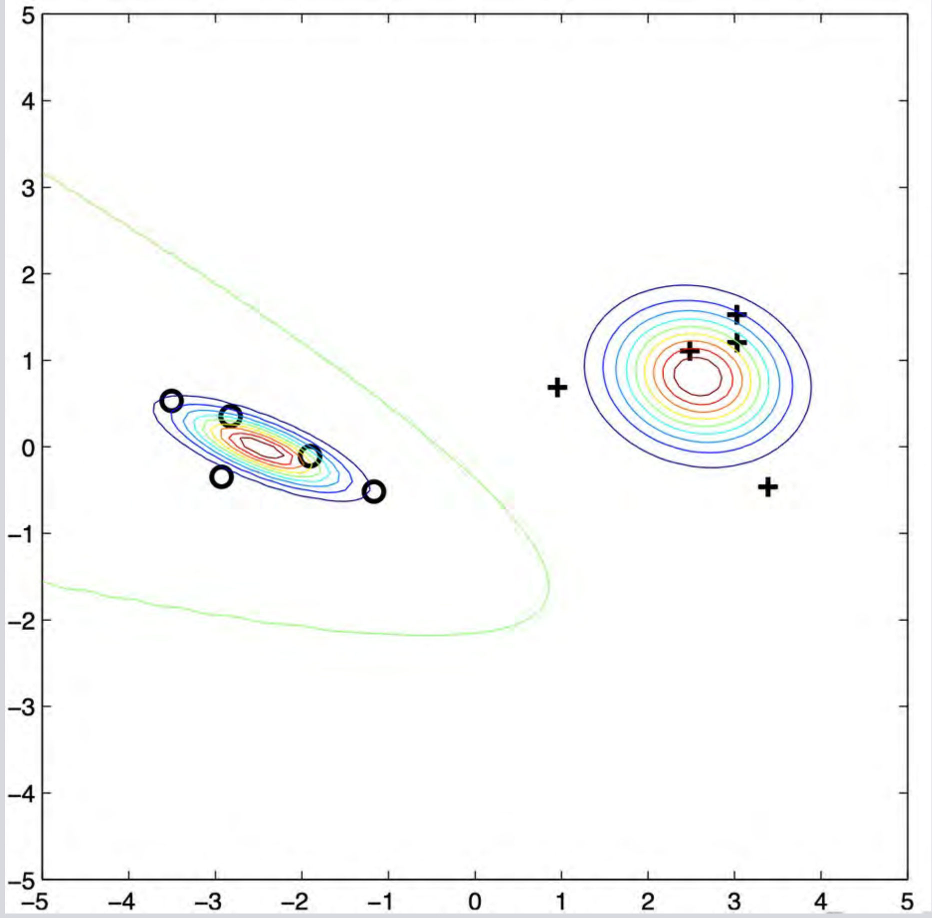


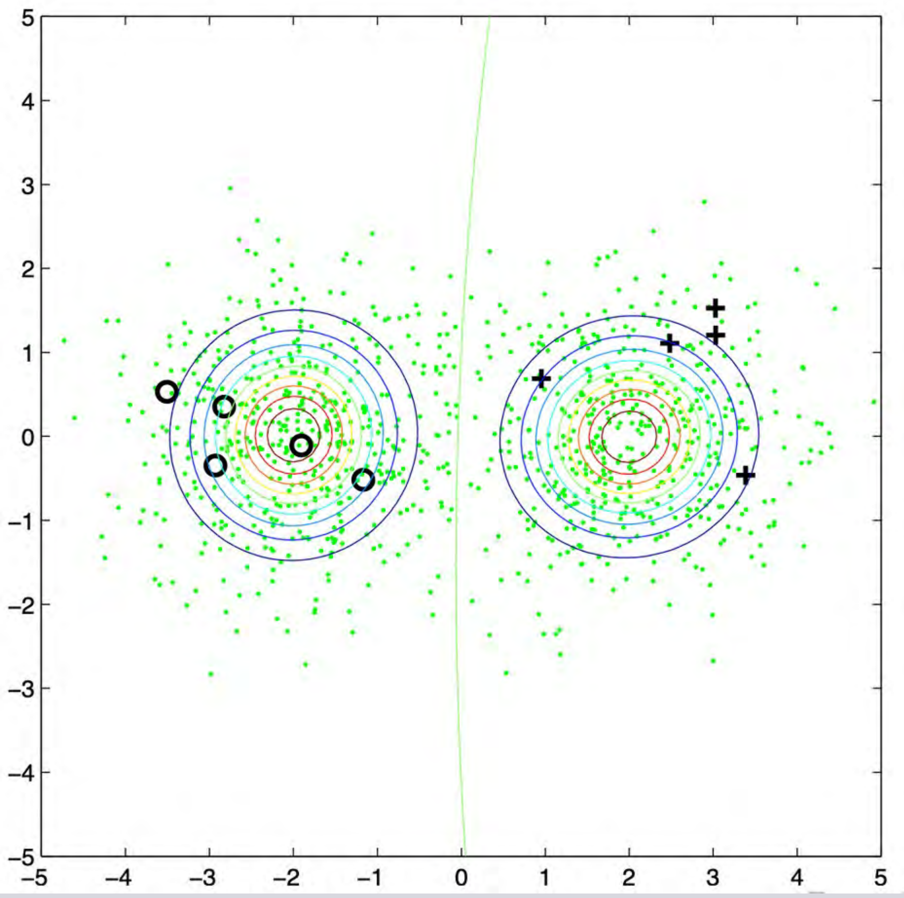
GENERATIVE MODELS: A SIMPLE EXAMPLE

With unlabeled data, the most likely model and its decision boundary:



They are different because they maximize different quantities:





SSL ALGORITHMS: GENERATIVE MODELS

- Full generative model: $p\left(X, Y \mid \theta\right)$
- Quantity of interest: $p\left(X_{i,} Y_{i,} X_{u} \mid \theta\right) = \Sigma_{Yu} p\left(X_{i,} Y_{i,} X_{u,} Y_{u} \mid \theta\right)$
- Find the maximum likelihood estimate (MLE) of θ, the maximum a posteriori (MAP) estimate or Bayesian.
- Often used in:
  - Mixture of Gaussian distributions (GMM): image classification
  - Mixture of multinomial distributions (Naive Bayes): text categorization.
  - Hidden Maekov Models (HMM): Speech recognition

ADVANTAGES

- Clear, Well-studies probabilistic framework
- Instead of EM you can go full Bayesian and include prior with MAP

DISADVANTAGES

- Often difficult to verify the correctness of the model
- EM (Expectation Maximization) local optima
- Makes strong asssumptions about class distribution
- Unlabeled data may hurt if generative model is wrong

To Lessen the danger:

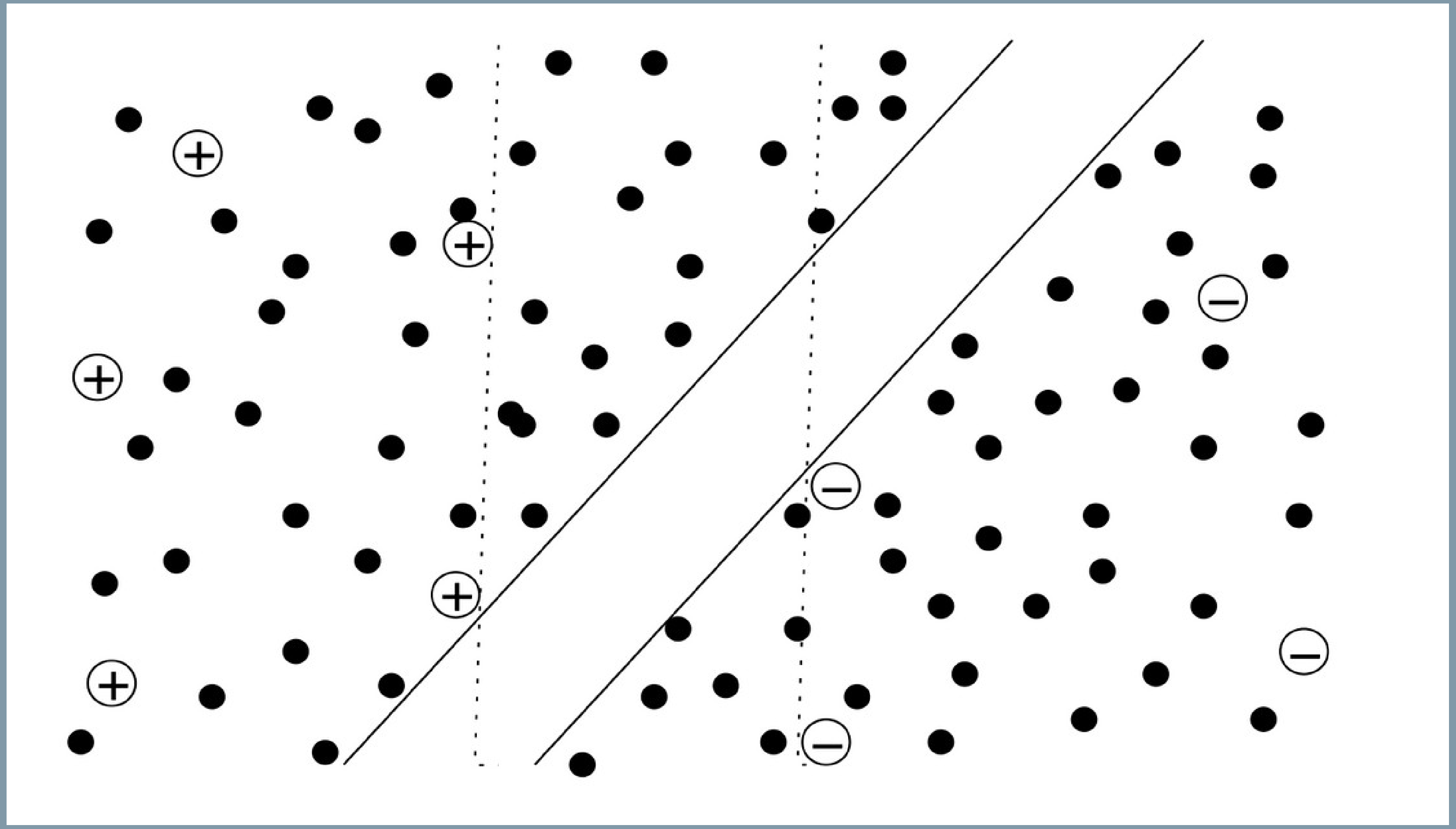
- Carefully construct the generative model to reflect the task e.g multiple Gaussian distributions per class, instead of one
- Down-weight the unlabeled data $(\lambda < 1)$

SSL ALGORITHMS

3. SEMI-SUPERVISED SUPPORT VECTOR MACHINES

Semi-supervised SVMs (S3VMs) = Transductive SVMs (TSVMs)

- It maximizes “unlabeled data margin”



Works based on smoothness assumption

Idea:

Find largest margin classifier, such that, unlabeled data are outside of the margin as much as possible, use regularization over the unlabeled data.

Given the training set $T = \{X_i\}$, and unlabeled set $U = \{u_j\}$

- Find all possible labeling $U_1 .. U_n$ on U
- For each $T_k^{\circ} = T \cup U_k$, train a standard SVM
- Choose SVM with largest margins

SSL ALGORITHMS: TSVM

Methods:

- Local Combinatorial search
- Standard unconstrained optimization solvers (CG,BFGS..)
- Continuation Methods
- Concave-Convex procedure (CCCP)
- Branch and Bound

ADVANTAGES

- Can be used with any SVM
- Clear optimisation criterion, mathematically well formulated

DISADVANTAGES

- Hard to optimize
- Prone to local optima -non convex
- Only small gain given modest assumption

SSL ALGORITHMS

4. MULTIVIEW ALGORITHMS

View: a different set of features that describe the same data point.

Idea: Train 2 classifiers on 2 disjoint sets of features then let each classifier label unlabeled examples and teach the other classifier.

Given Training set T ={Xi}, and unlabeled set U = {uj}

1. Split T into T1 and T2 on the feature dimension
2. Train f1 on T1 and f1 on T2
3. Get predictions P1= f1(U) and P2 =f2(U)
4. Add: top k from P1 to T2; top k from P1 to T2
5. Repeat until |U| =0

Works based on smoothness or cluster assumption

Strategy:

Co-Training (Classic Multiview)

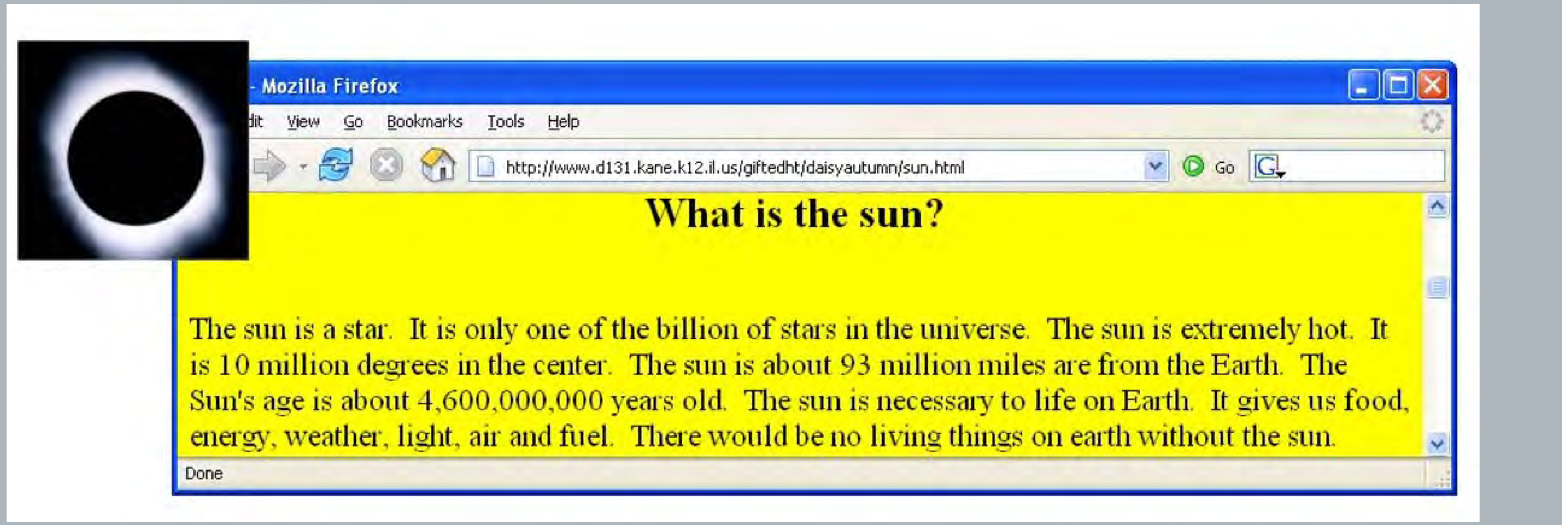
- Train 2 (or more) models on different views.
- Each model predicts labels for the unlabeled data.
- Predictions are used to augment training set of the other model.
- Helps teach each other by labeling new examples.

Consensus-Based Learning

- Multiple models try to agree on labels for unlabeled data.
- Confidence is increased when all models agree.
- Final prediction is made by majority vote or average confidence.

MULTIVIEW ALGORITHM: A SIMPLE EXAMPLE (C0-TRAINING)

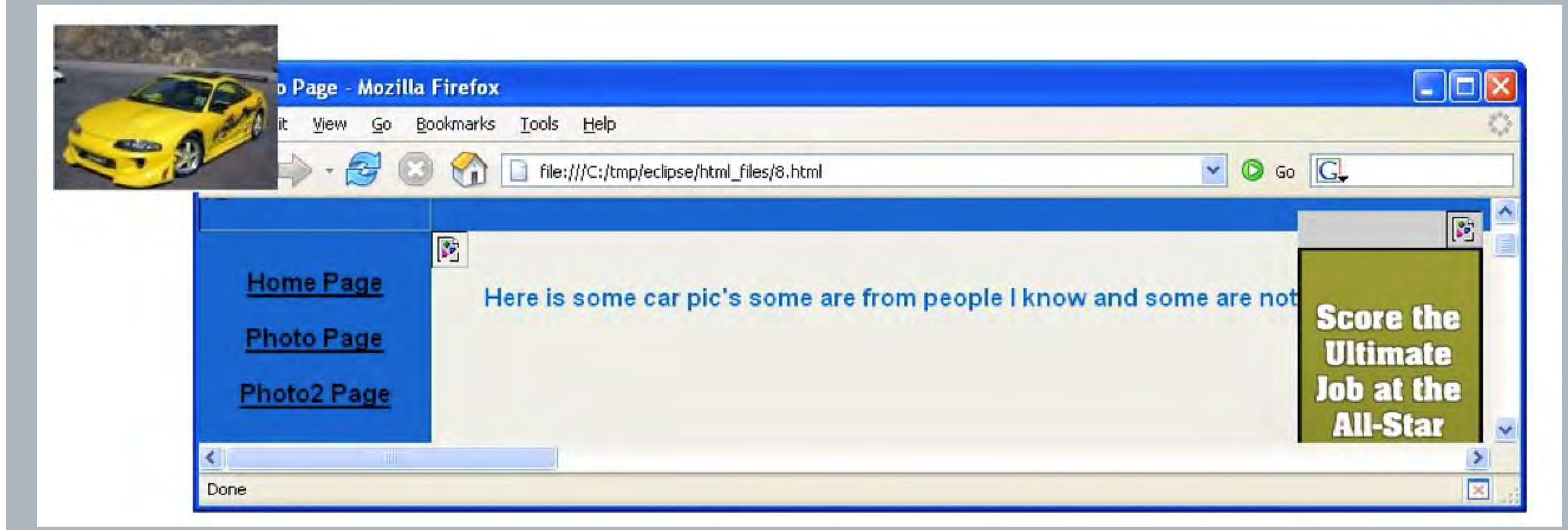
Two views of an item: image and HTML text



http://www.d131.kane.k12.il.us/giftedht/daisyautumn/sun.html

What is the sun?

The sun is a star. It is only one of the billion of stars in the universe. The sun is extremely hot. It is 10 million degrees in the center. The sun is about 93 million miles are from the Earth. The Sun's age is about 4,600,000,000 years old. The sun is necessary to life on Earth. It gives us food, energy, weather, light, air and fuel. There would be no living things on earth without the sun.


o Page - Mozilla Firefox
it View Go Bookmarks Tools Help
Go
Done
Score the Ultimate Job at the All-Star

file:///C:/tmp/eclipse/html_files/8.html

Home Page

Photo Page

Photo2 Page

Here is some car pic's some are from people I know and some are not

Feature Split

Each instance is represented by two sets of features x = [x(1);x(2)]

- x(1) = image feature
- x(2)= web page text
- This is a natural feature split (or multiple views)

Co-training Idea:

- Train an image classifier and a text classifier
- The two classifiers teach each other

SSL ALGORITHMS: MULTIVIEW

ADVANTAGES

- Simple Method applicable to almost all classifiers
- Can correct mistakes in classification between the 2 classifiers
- Less sensitive to mistakes than in self-training
- This makes it useful in domains like multimedia, web mining, and healthcare.

DISADVANTAGES

- Assumes conditional independence between features
- Natural feature splits may not exist
- Artificial feature splits may be complicated if only few features are present
- Models using BOTH features should do better
- Processing multiple views increases computational cost in terms of time and memory.

SSL ALGORITHMS

5. GRAPH-BASED ALGORITHMS

Idea: A graph is given on the labeled and unlabeled data. Instances connected by heavy edge tend to have the same label.

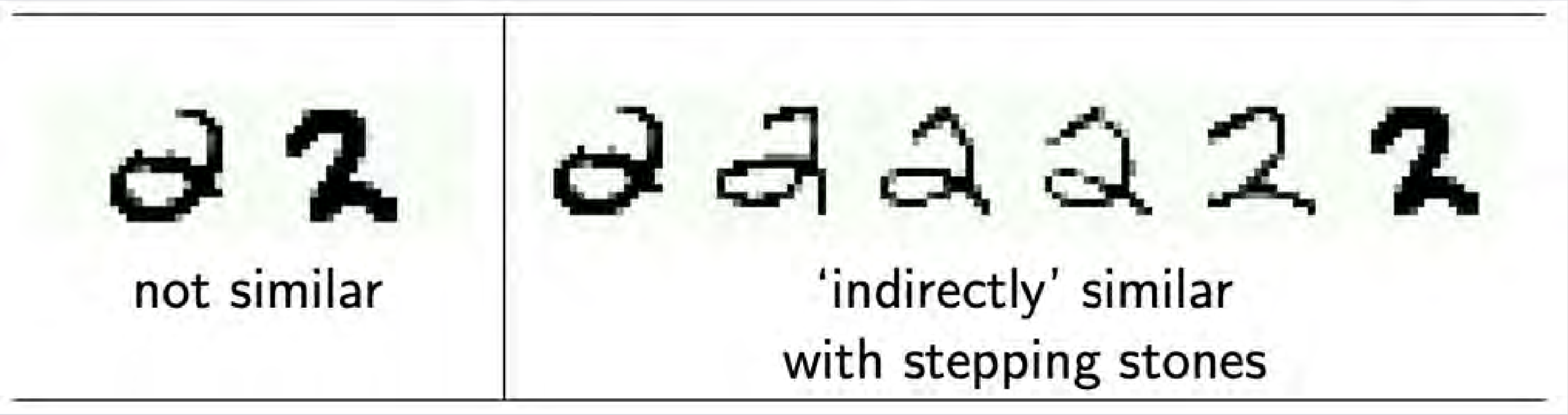
The graph consists of:

- **Nodes:** labeled and unlabeled data points (Xi union Xu).
- **Edges:** similarity weights computed from features (based on distance)
- Works based on the assumption of label smoothness (if two data points are connected, they have same label)

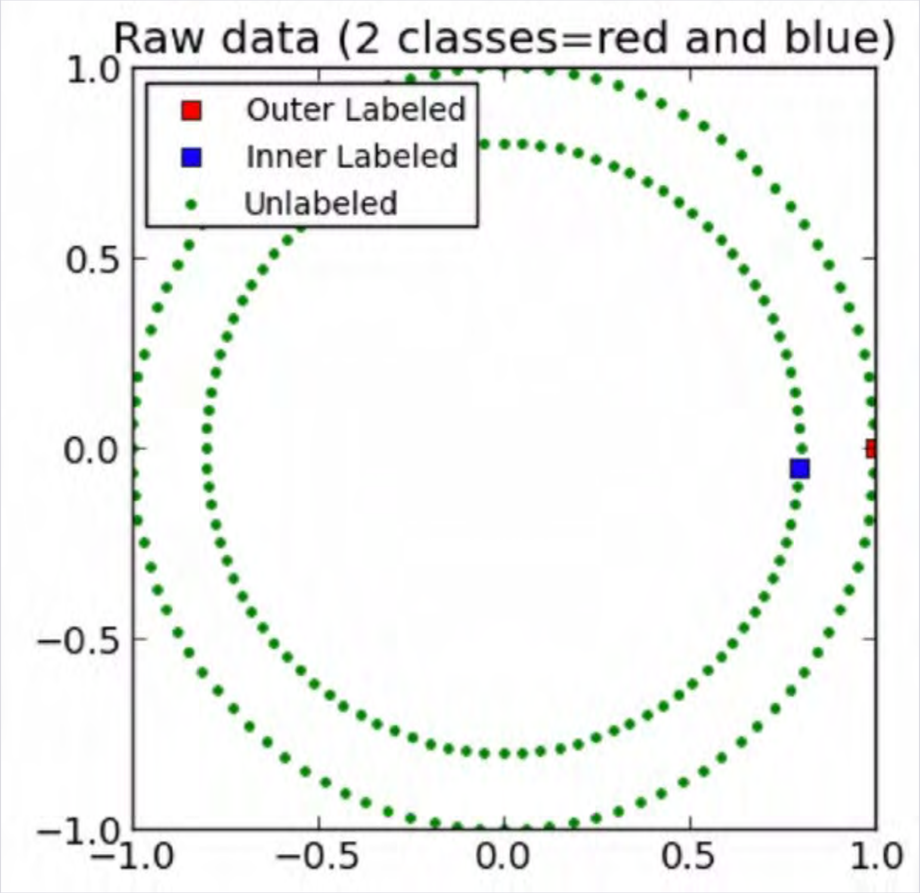
Want: implied similarity via all paths

Works based on smoothness and manifold assumptions

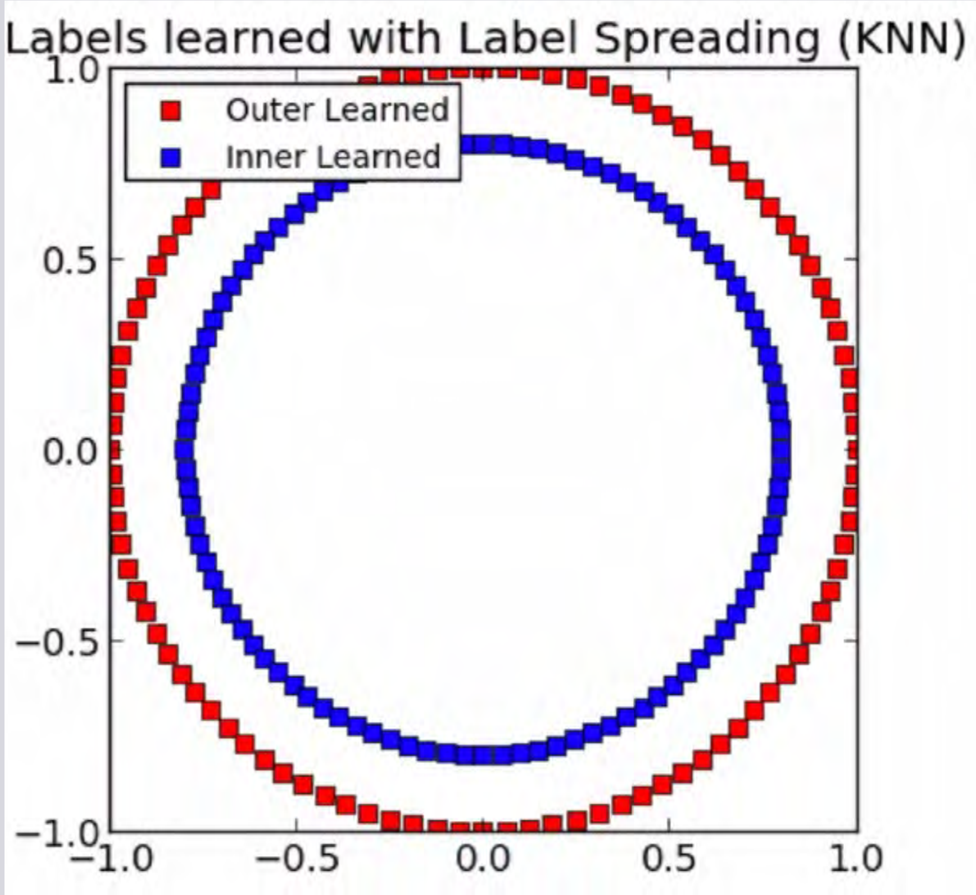
GRAPH-BASED ALGORITHM: A SIMPLE EXAMPLE



Handwritten digits recognition with pixel-wise Euclidean distance



Raw data with two classes



Labels learned with label spreading

SSL ALGORITHMS: GRAPH-BASED ALGORITHMS

Algorithms:

Label Propagation:

- **L**abels from labeled nodes are spread to unlabeled nodes based on graph. structure.
- Use similarity matrix to define how strongly labels should propagate between nodes.

Label Spreading:

- include above, but includes normalization and smoothing, often with a kernel (like RBF).

Graph Convolutional Networks (GCNs):

- Combines graph structure with neural networks.
- Allows learning node representations and classifying nodes.

Others include: mini cut harmonic, manifold regularization, local and global consistency.

ADVANTAGES

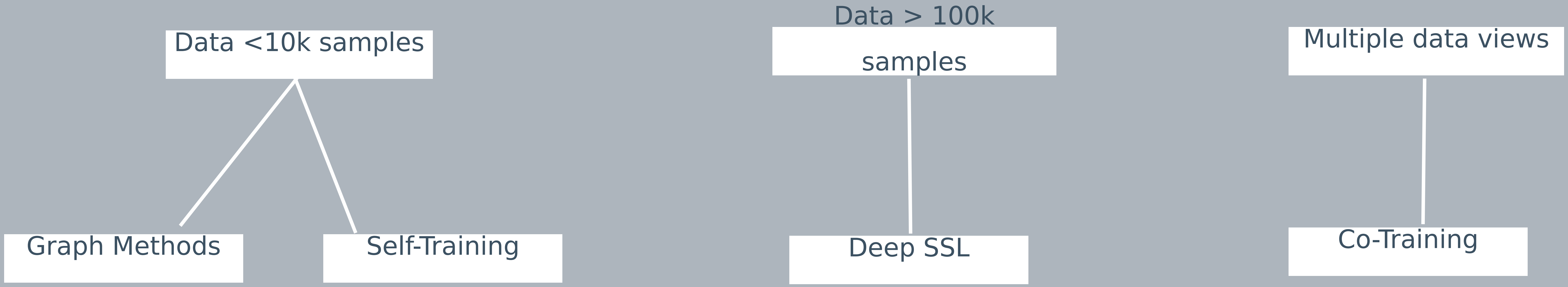
- Clear mathematical framework
- Performance is strong if the graph fits the task
- Can be used in combination with any model
- Excellent use of unlabeled data.

DISADVANTAGES

- Performance is bad if the graph is bad
- Sensitive to graph, construction, structure and edge weights (bad similarity metrics = bad performance).
- Hard to scale to real-time or streaming data.

SSL IMPLEMENTATION ROADMAP

1. Assess Your Data
    a. Labeling cost analysis
    b. Distribution matching check
2. Algorithm selection guide



3. Evaluation Protocol
    a. Always maintain held-out test set
    b. Monitor performance vs labeling budget

REAL WORLD APPLICATION

Medical Imaging and Diagnostics

Task: Label X-rays, MRIs, or pathology

- Train models using a few expert-labeled samples and many unlabeled images.
- E.g. Detecting tumors in MRI scans or classifying diseases from retinal images.

Natural Language Processing (NLP)

Task: Sentiment analysis, named entity recognition (NER), and text classification.

- Labeled text data is limited, but we have huge amounts of raw text (e.g., web pages, forums).
- E.g. Classifying product reviews or detecting spam in emails.

Speech Recognition

Task: Transcribing spoken audio into text

- Helps by learning from unlabeled audio data to improve recognition accuracy.
- E.g. Voice assistants like Siri, Google Assistant.

Fraud Detection

Task: Few known fraud cases vs. large number of transactions.

- Learns transaction patterns and flags potential fraud with limited labeled examples.

Recommendation Systems

Task: Recommend products, music, or videos.

- Learns user preferences even with sparse explicit feedback (e.g., few likes or ratings).

Autonomous Vehicles

Task: Labeling driving scenes or pedestrian actions

- Combines a small labeled dataset with large-scale unlabeled camera
- Sensors data to improve perception models.

Face Recognition & Image Classification

- Uses a few labeled faces and many unlabeled to learn better facial embeddings.
- E.g. Grouping and tagging faces in photo galleries (e.g., Google Photos).

PROS AND CONS

PROS

Less Labeled Data Needed

- It can work well with only a small amount of labeled data: great when labeling is expensive or slow.

Cost-Effective

- Saves time and money by reducing the need for expert-annotated data.

Uses Abundant Unlabeled Data

- Leverages large amounts of available unlabeled data (e.g., images, text, audio).

Better Performance

- Often performs better than purely supervised models when labeled data is limited.

Generalizes Well

- Learns patterns from both labeled and unlabeled data, reduces overfitting.

CONS

Assumption Sensitive

- Assumes that unlabeled data follows the same distribution or structure as labeled data (not always true)

Performance is Hard to Predict

- If unlabeled data is noisy, it can hurt the model more than help.

⚙ Model Complexity

- Some algorithms are more complex to implement, and tune compared to standard supervised models.

Lack of Theoretical Guarantees

- Performance can vary across different datasets (there’s no one-size-fits-all method)

Data Privacy

- Using large amounts of unlabeled personal data (e.g., in healthcare or finance) might raise privacy concerns.

CHALLENGES

Challenges	Description
Label Propagation Risks	If initial labels are noisy or biased, model can propagate incorrect information through unlabeled data.
Distribution Shift	Assumes labeled and unlabeled data comes from the same distribution: may not hold in real-world datasets.
Scalability	Graph-based or complex SSL models may struggle with large-scale data in terms of memory and computation.
Data Quality	Unlabeled data might contain outliers, noise, or irrelevant samples that hurt performance.
Evaluation Difficulties	Without a lot of labeled data, it’s hard to evaluate the model or tune hyperparameters effectively.
Lack of Universality	One SSL algorithm might work great for one task (e.g., image classification) but fail on another (e.g., text or audio).

RESEARCH DIRECTIONS

Area of Research	Description
Self-Supervised Learning Integration	Combining SSL with self-supervised methods to improve learning from unlabeled data.
Robustness to Noisy Labels	Designing SSL models that can resist or correct errors in both labeled and unlabeled data.
Uncertainty Estimation	Using confidence-aware learning to decide when and how to trust the predictions on unlabeled data.
Domain Adaptation	Making SSL effective when labeled and unlabeled data come from different but related domains.
Semi-Supervised Deep Learning	Enhancing deep learning models with SSL capabilities (e.g., consistency regularization, pseudo-labeling).
Theoretical Foundations	Developing better theoretical guarantees for generalization and risk bounds in SSL settings.

REFRENCES

- Olivier Chapelle, Alexander Zien, Bernhard Sch¨olkopf (Eds.). (2006). Semi-supervised learning. MIT Press.

- Xiaojin Zhu (2005). Semi-supervised learning literature survey. TR-1530. University of Wisconsin-Madison Department of Computer Science.

- Matthias Seeger (2001). Learning with labeled and unlabeled data. Technical Report. University of Edinburgh.

- Lukas Tencer (2014). Semi-Supervised Learning

ANY QUESTION?

THANK YOU