
# CSC380: Principles of Data Science

## Introduction to Machine Learning

Prof. Jason Pacheco

TA: Enfa Rose George

TA: Saiful Islam Salim

With content from Prof. Kwang Sung-Jun

# What is machine learning?

- **Tom Mitchell** established Machine Learning Department at CMU (2006).

**Machine Learning, Tom Mitchell, McGraw Hill, 1997.**


*Machine Learning is the study of computer algorithms that improve automatically through experience*. Applications range from datamining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests.

*This book provides a single source introduction to the field*. It is written for advanced undergraduate and graduate students, and for developers and researchers in the field. No prior background in artificial intelligence or statistics is assumed.

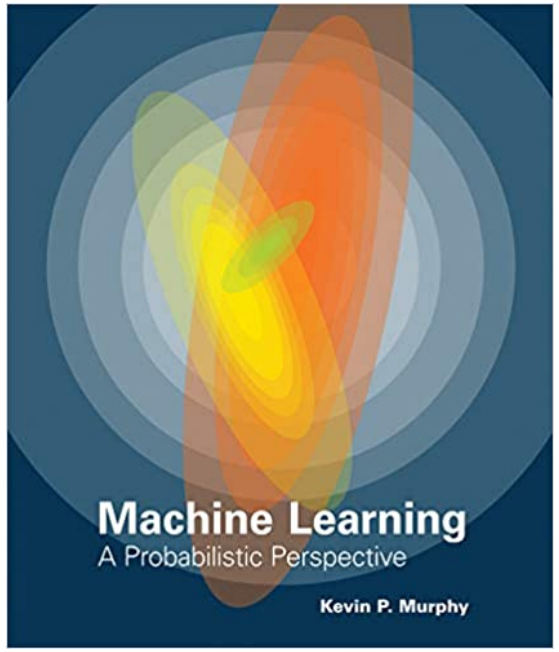
- A bit outdated with recent trends, but still has interesting discussion (and easy to read).
- A subfield of **Artificial Intelligence** – you want to perform nontrivial, smart tasks. The difference from the traditional AI is "**how**" you build a computer program to do it.

# Textbooks

*We will use a more recent textbook for readings*

*Takes a probabilistic approach to machine learning*

*Consistent with the goals of data science in this class*




Murphy, K. "Machine Learning: A Probabilistic Perspective." MIT press, 2012

( UA Library )

# AI Task 1: Image classification

- Predefined categories: $C$ = {cat, dog, lion, …}
- Given an image, classify it as one of the set $C$ with the highest accuracy as possible.
- Use: sorting/searching images by category.
- Also: categorize types of stars/events in the Universe (images taken from large surveying telescopes)

# AI Task 2: Recommender systems

- Predict how user would rate a movie
- **Use**: For each user, pick an unwatched movie with the high predicted ratings.
- **Idea**: compute user-user similarity or movie-movie similarity, then compute a weighted average.

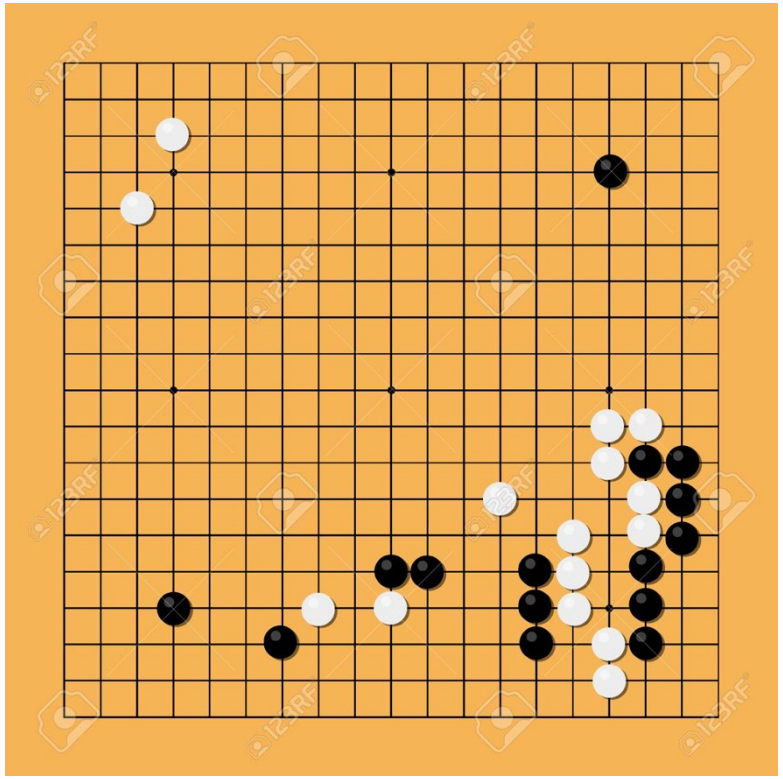
| | User 1 | User 2 | User 3 |
|---|---|---|---|
| **Movie 1** | 1 | 2 | 1 |
| **Movie 2** | ? | 3 | 1 |
| **Movie 3** | 2 | 5 | 2 |
| **Movie 4** | 4 | ? | 5 |
| **Movie 5** | ? | 4 | 2 |

# AI Task 3: Machine translation

- No need to explain how useful it is.

# AI Task 4: Board game

- Predict win probability of a move in a given game state (e.g., AlgphaGo)
- Traditionally considered as a “very smart” task to perform.
- **Use**: From the AI Go player, you can do practice play or even learn from it.



# Traditional AI vs Machine Learning (ML)

- **Traditional AI**: *you* encode the knowledge (e.g., logic statements), and the *machi* executes it, with some more '**inference**' like if a -> b and b-> c, then a-> c.
  - e.g., if you see some feather texture with two eyes and a beak, classify it as a

- **ML**: I give you a number of input and output observations (e.g., animal picture + label), and you give me a **function (can be a set of logical statements or a neural network)** that maps the input to the output accurately.
  - As the "big data" era comes, data is abundant => far better to learn from data than to encode domain knowledge manually.
  - "statistical" approach // data-driven approach
  - *"Every time I fire a linguist, the performance of the speech recognizer goes up." – 1988,* Frederick Jelinek, a Czech-American researcher in information theory & speech recognition.

# Overview of ML Methods

**Supervised Learning**

- Provide *training* data consisting of input-output pairs and learn mapping
- E.g. Spam prediction, object detection or image classification, machine translation, etc.

**Unsupervised learning**

- Finds patterns in the data without the help of labels (outputs)
- E.g. clustering, dimensionality reduction, target tracking, image segmentation, etc.

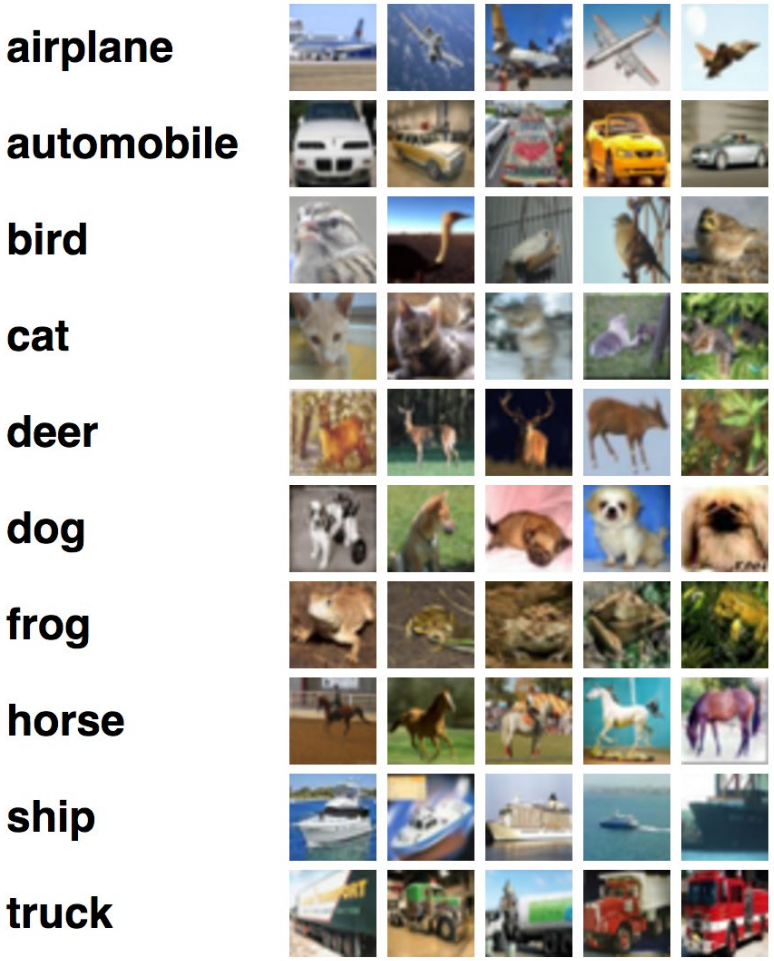
**Reinforcement learning** ⇐ **We won't cover this**

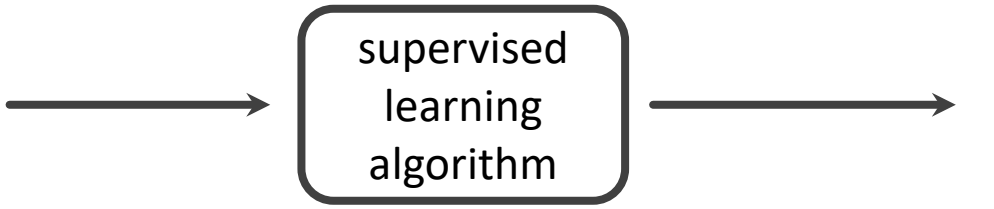
- The environment interacts with your action, transferring you to different states.
- When there are no states: "**bandit**" feedback.
  - E.g., Amazon recommends you a pair of shoes. You did not click it. Amazon don't know if you would've clicked had it recommended speakers or cookware.
  - The dataset is now dependent on the recommendation algorithm => biased data.

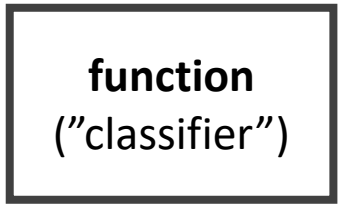
# Supervised Learning

# Basic setting: Supervised learning

- Training data: dataset comprised of *labeled examples*: a pair of (input, label)




cat!

training

testing

# Examples function 1: Decision tree

```
Task: predict the rating of a movie by a user
If age >= 60 then
   if genre = western then
      return 4.3
   else if release date > 1998 then
      return 2.5
   else ...
   ...
   end if
else if age < 60 then
...
end if
```

# Example function 2: Linear

Task: Image classification

Let $x$ be a set of pixel values of a picture (30x30) => 900-dimensional vector $x$.

If $0.124 \cdot x_1 - 2.5 \cdot x_2 + \cdots + 2.31 \cdot x_{900} > 2.12$ then

    return cat

else

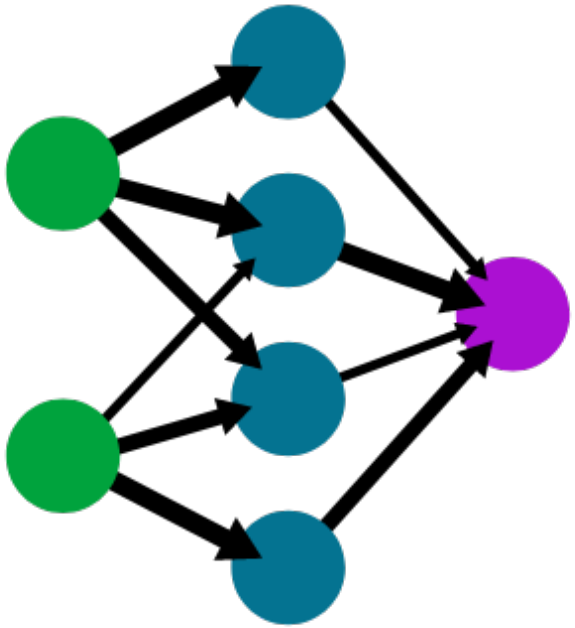
    return dog

end

"linear combination"

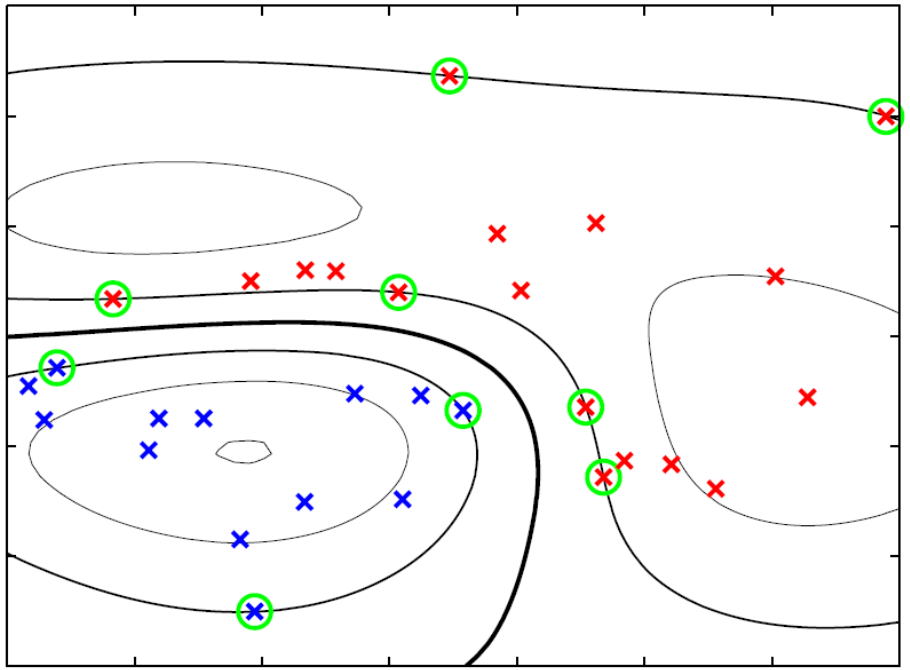
# Example function 3: Nonlinear

Neural network



(stacked **linear** models with nonlinear **activation functions**)

Support Vector Machine

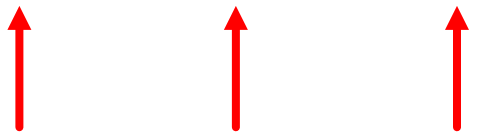


(**linear** in the induced feature space)

# Example: Naïve Bayes Classification

**Training Data:**

| Person | height (feet) | weight (lbs) | foot size(inches) |
|---|---|---|---|
| male | 6 | 180 | 12 |
| male | 5.92 (5'11") | 190 | 11 |
| male | 5.58 (5'7") | 170 | 12 |
| male | 5.92 (5'11") | 165 | 10 |
| female | 5 | 100 | 6 |
| female | 5.5 (5'6") | 150 | 8 |
| female | 5.42 (5'5") | 130 | 7 |
| female | 5.75 (5'9") | 150 | 9 |



**Features**

**Task:** Observe features $x_1, \ldots, x_n$ and predict class label $C_k$

**Model:** Treat features as *conditionally independent*, given class label,

$$p(x, C) = p(C) \prod_{i=1}^{n} p(x_i \mid C)$$

Doesn't capture correlation among features, but is easier to learn.

**Classification:** Bayesian model so classify by posterior,

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

# Supervised learning: Types of prediction problems

**Binary classification:** Choose between 2 classes

- Given an email, is it spam or not (ham)? (or the probability of it being spam)

**Multi-class classification:** Multiple discrete outputs

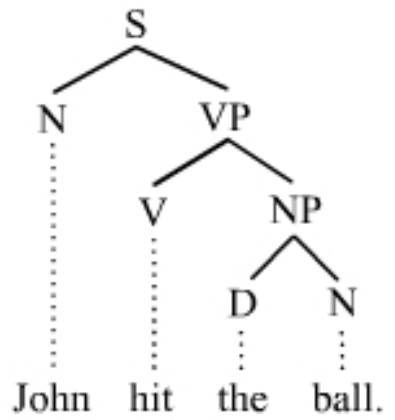
- Image classification with 1000 categories.

**Regression:** the label is real-valued (e.g., price)

- Say I am going to visit Italy next month. Given the price trends in the past, what would be the price given (the # of days before the departure, day of week)?
- Pricing: predict the lowest price

**Structured output prediction:** more than just a number

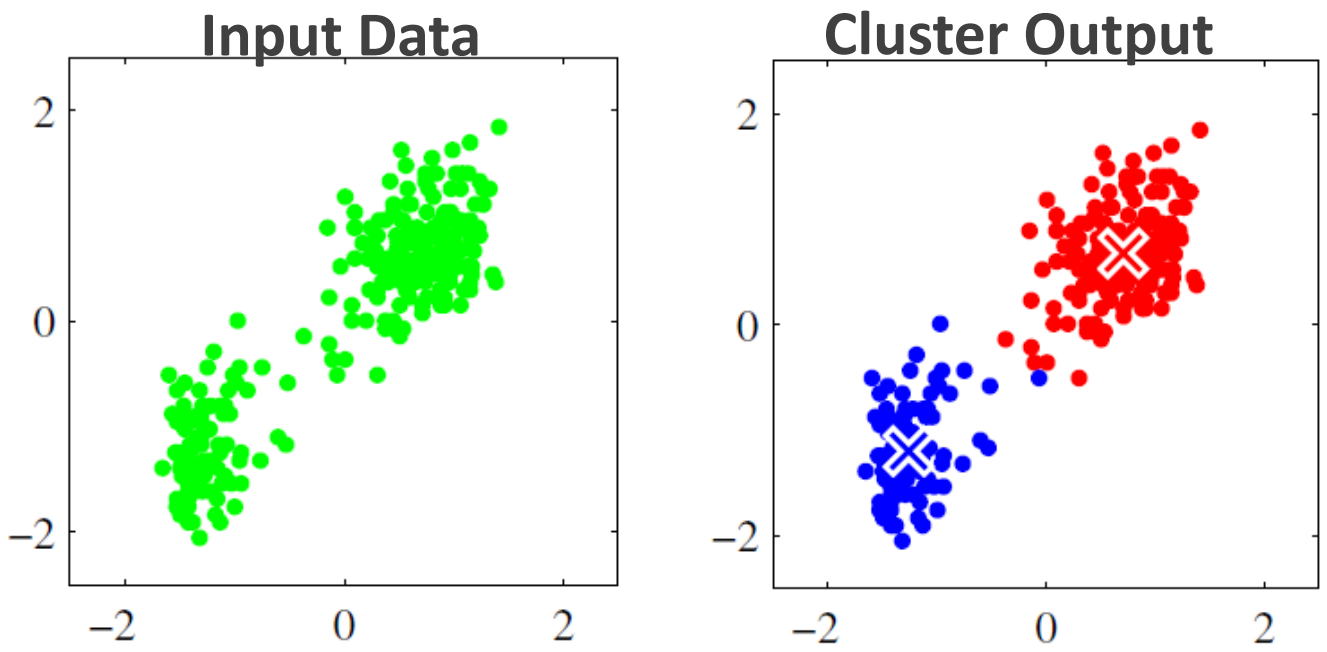
- Given a sentence, what is its grammatical parse tree?



# Unsupervised Learning

# Example: Clustering

Identify groups (clusters) of similar data



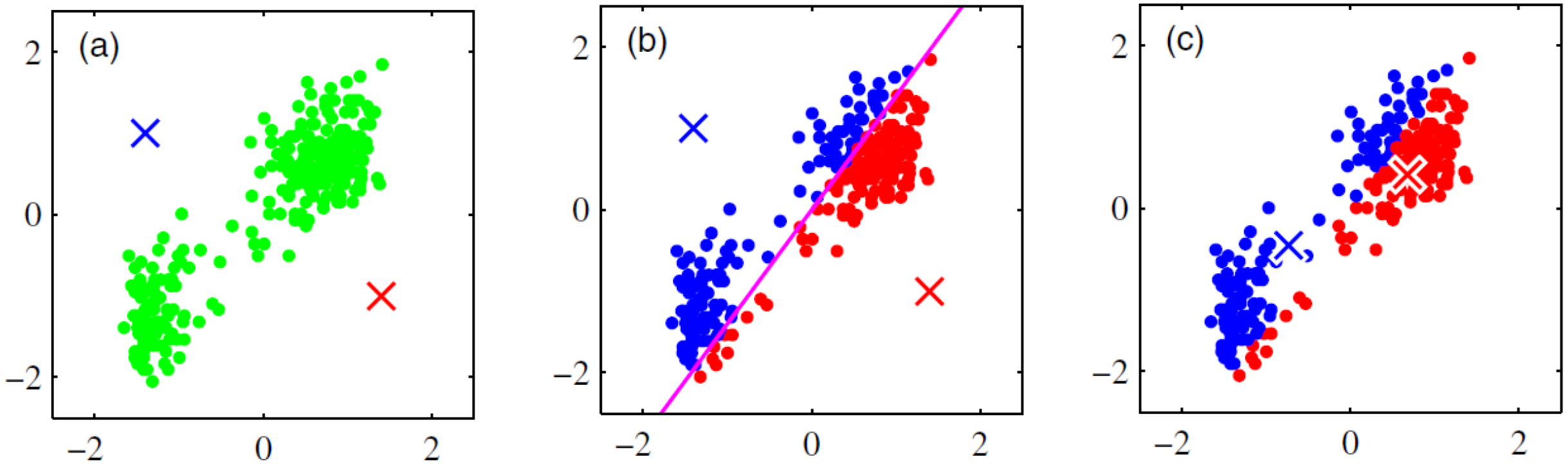

Useful for interpreting large datasets

Clusters are assigned arbitrary labels (e.g. 1, 2, …, K); if we provide meaning then:

clustering → classification

Common clustering algorithms: K-means, Expectation Maximization (EM)

# K-Means Clustering

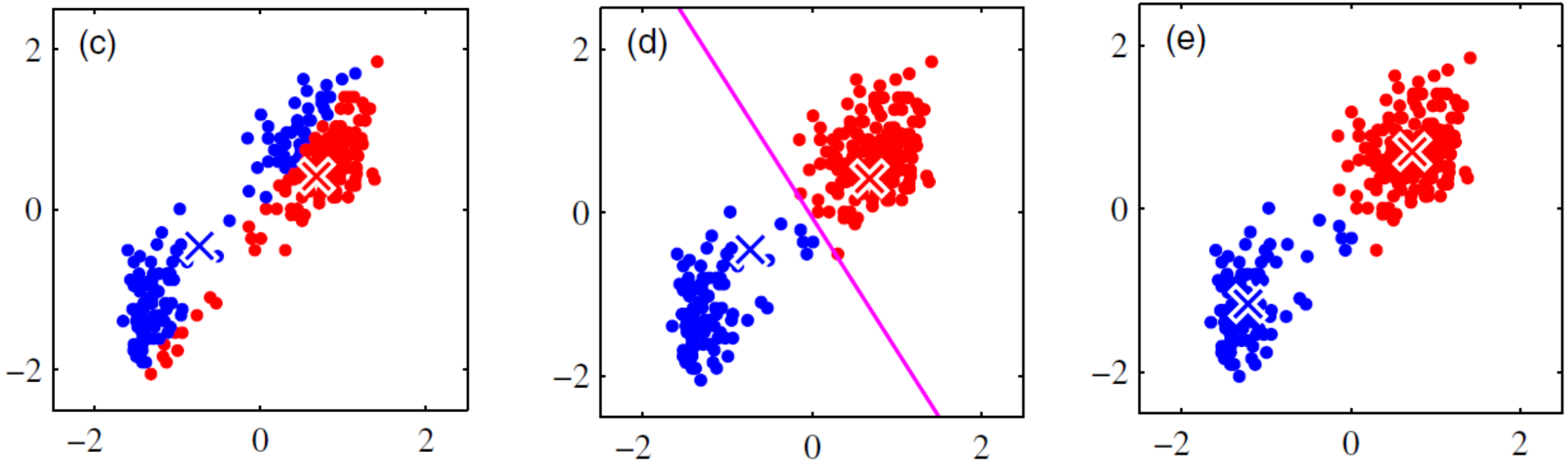



**K-Means Algorithm**

1. Update cluster centers
2. Assign data to closest cluster center
3. Repeat

**Source: Bishop, C. PRML**

# K-Means Clustering



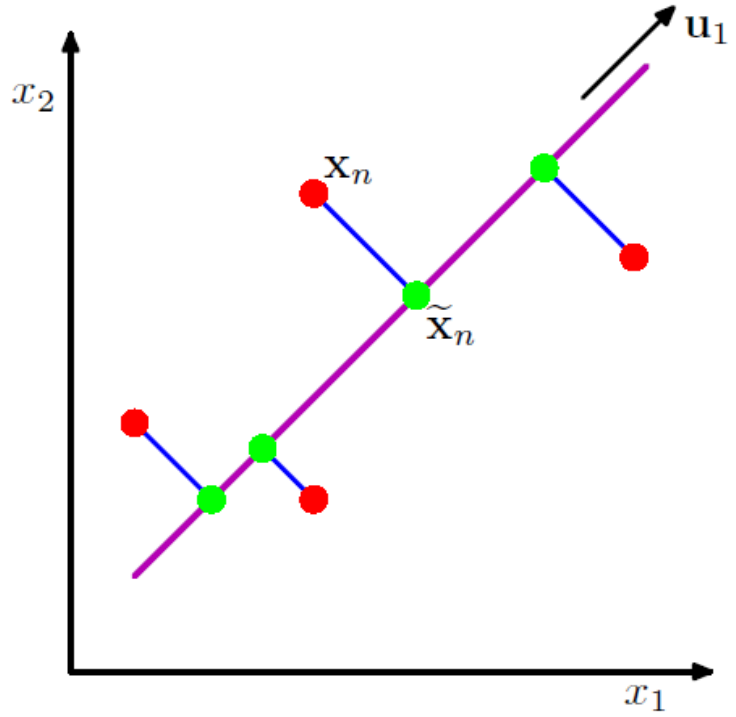

**K-Means Algorithm**

1. Update cluster centers
2. Assign data to closest cluster center
3. Repeat

**Source: Bishop, C. PRML**

# Example: Principal Component Analysis (PCA)

Reduce dimension of high-dimensional data using linear projection




Identify directions of maximum variation in the data by computing *eigenvectors*

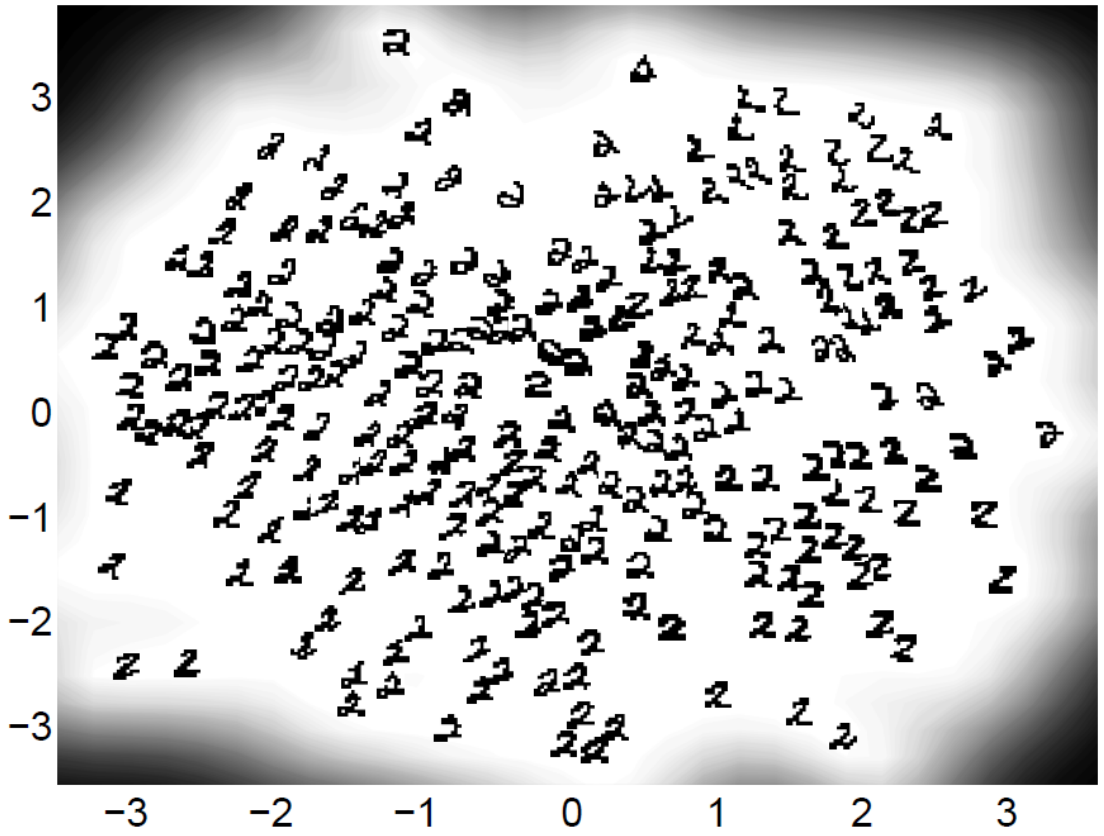
Linear projection onto K-dimensional subspace spanned by top K eigenvalues

Can be used for visualization (project to 2D) or for modeling

**Source: Bishop, C. PRML**

# Example: Principal Component Analysis (PCA)

Reduce dimension of high-dimensional data using linear projection

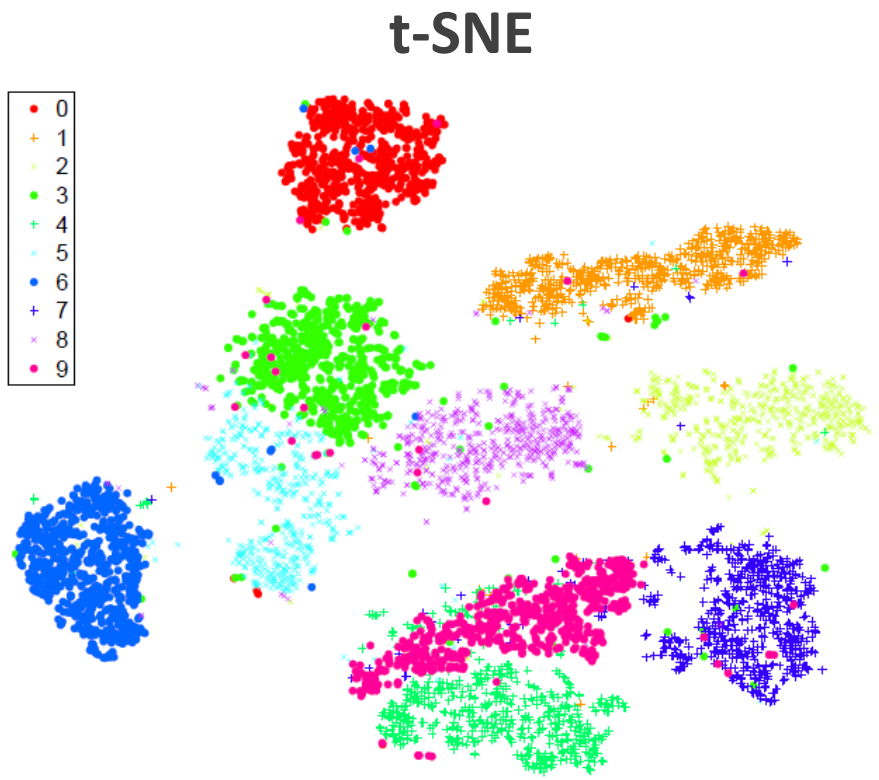



**Source: Lawrence, N. (2005)**

Example for modeling / visualizing handwritten digits

Each digit is a black/white image with 28x28 pixels ($28^2$ dimensions) projected down to 2D

# Example: Nonlinear Dimensionality Reduction

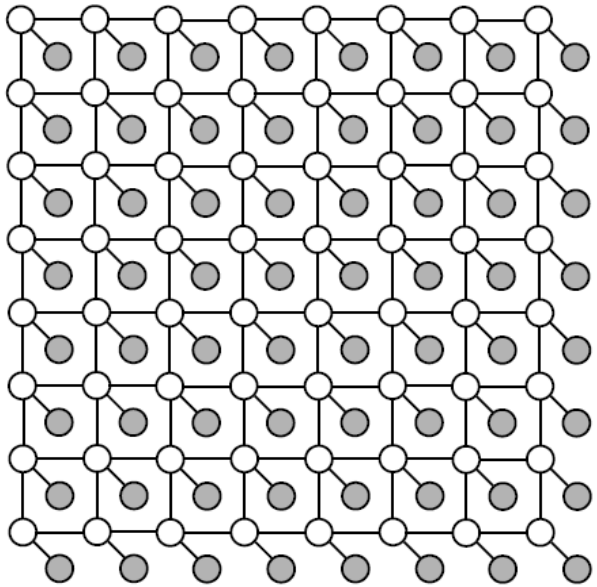



Nonlinear reduction can (potentially) amplify clustering properties

**t-Distributed Stochastic Neighbor Embedding (t-SNE)** Models similarity between data as a Student's-t distribution in high / low dimensions and optimizes reduction to preserve similarity

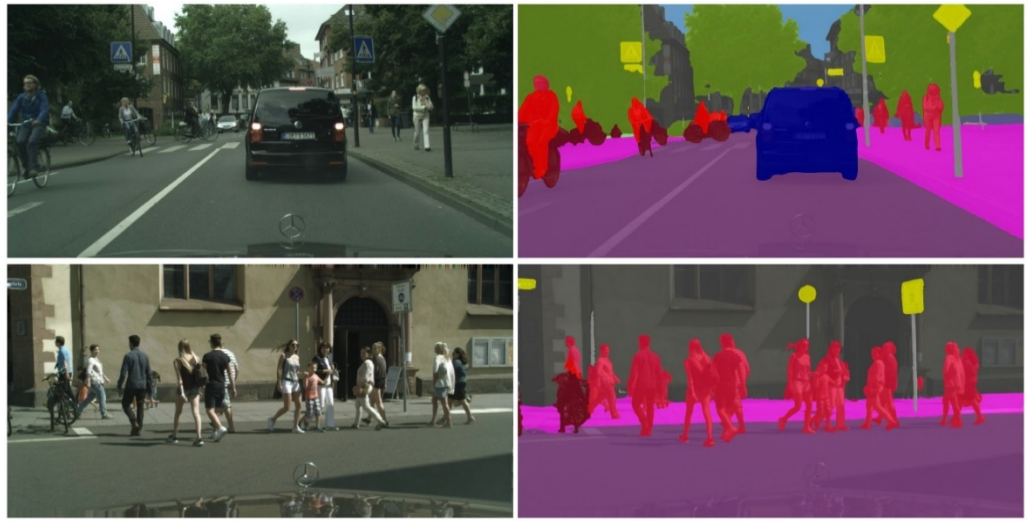
Visualization shows same handwritten digits from previous slide, projected to 2D and clustered

# Example: Image Segmentation



Don't need to know log-normalizer to specify model

[Source: Kundu, A. et al., CVPR16]



**Pairwise MRF *energy*:** $$-\log p(x,y) = \log Z + \sum_i \psi_i(x_i, y_i) + \sum_{(i,j)} \psi_{i,j}(x_i, x_j)$$

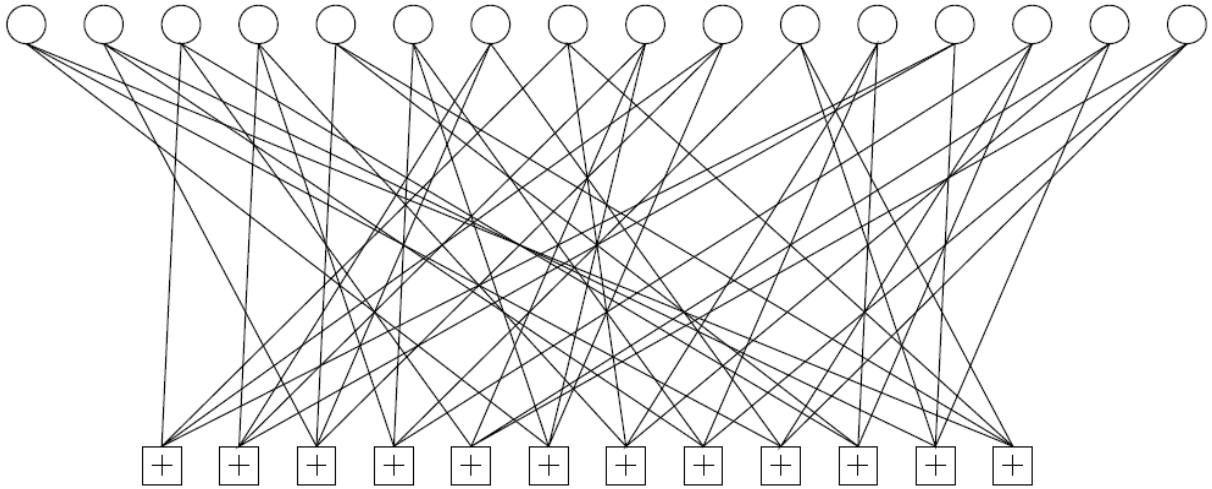
*Low energy configurations = High probability*

**L2 Likelihood:** $\psi_i(x_i, y_i) = \|x_i - y_i\|^2$ **Potts model:** $\psi_{i,j}(x_i, x_j) = \mathbb{I}(x_i = x_j)$

*MAP (minimum energy) configuration = Piecewise constant regions*

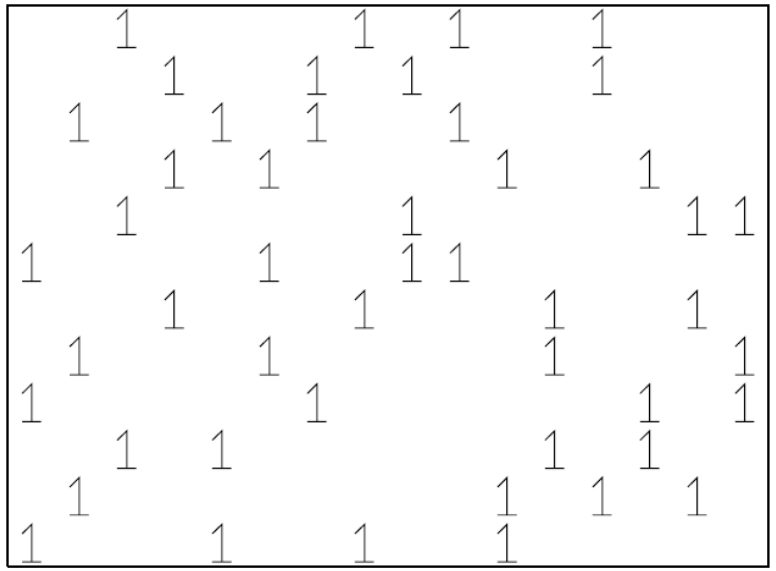
# Example: Low Density Parity Check Codes

Factor Graph Representation



Sparse Parity Check Matrix

$\mathbf{H} =$



Transmitted Code

$t \sim p(t)$

Noisy Channel

Received Code

$r \mid t \sim p(r \mid t)$

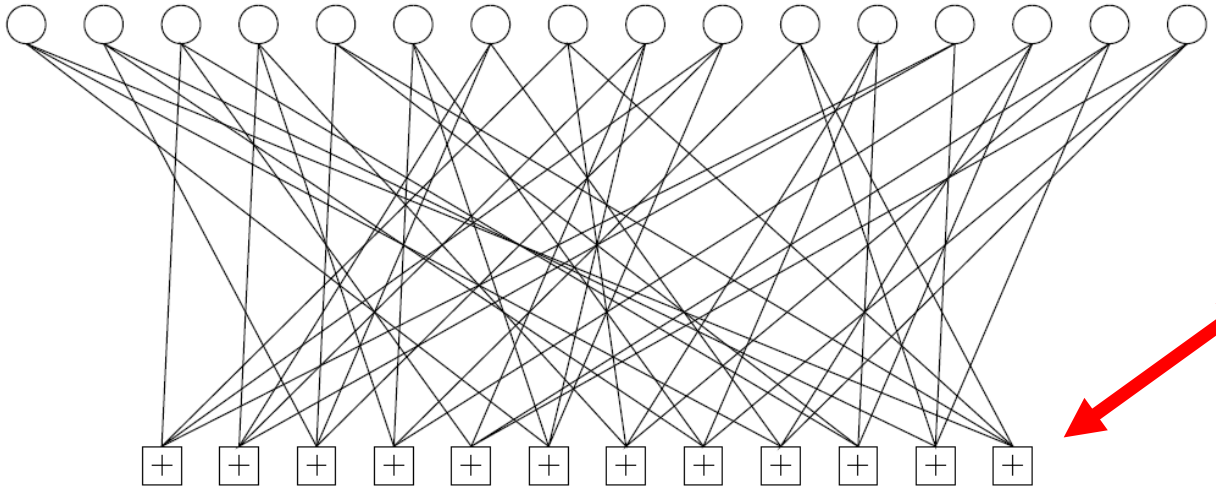
Decoder

$t^* = \arg\max_t p(t \mid r)$

[Source: David MacKay]

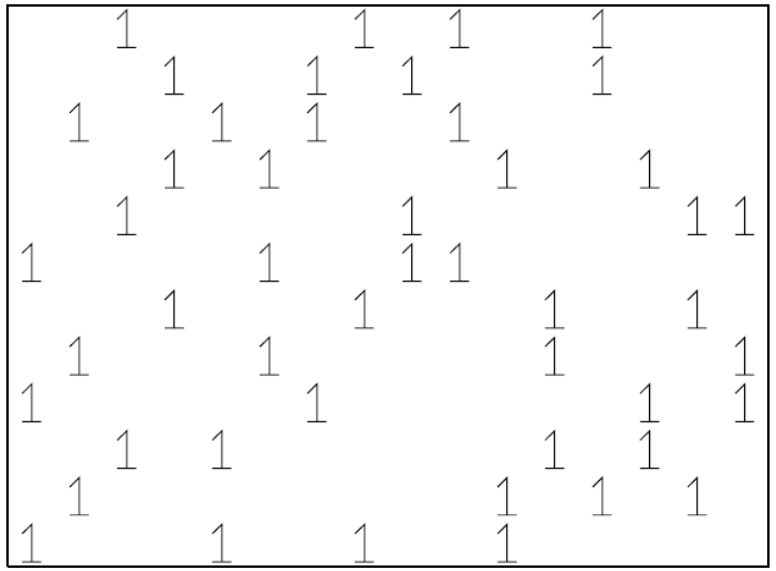
# Example: Low Density Parity Check Codes

Factor Graph Representation



$\mathbf{H} =$

Sparse Parity Check Matrix



- Valid codes have zero parity: $p(t) \propto \mathbb{I}(Ht = 0 \bmod 2)$
- Chanel noise model arbitrary, e.g. flip bits w/ probability: $\epsilon$

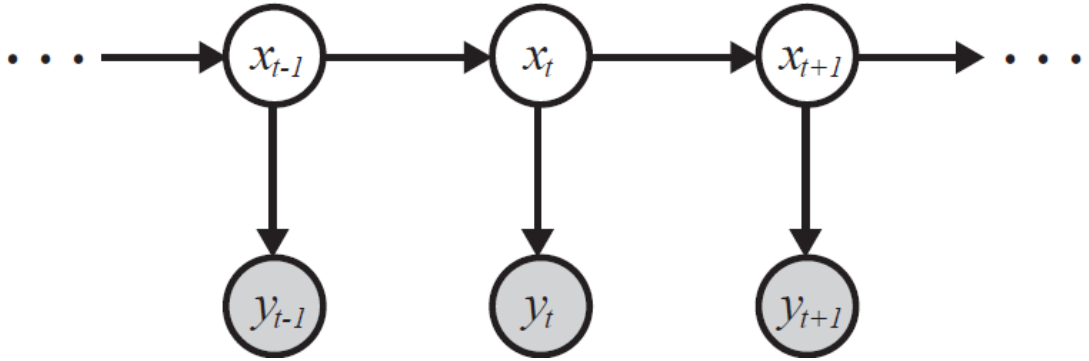
$$p(r \mid t) = \prod_n p(r_n \mid t_n) = \prod_n (1-\epsilon)^{\mathbb{I}(r_n = t_n)} \epsilon^{\mathbb{I}(r_n \neq t_n)}$$

n-th bit →

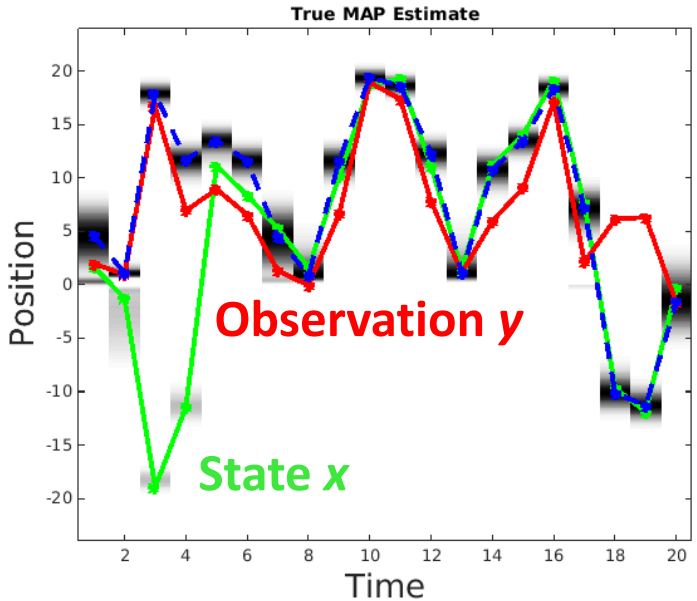
[Source: David MacKay]

# Example: Time Series and Target Tracking

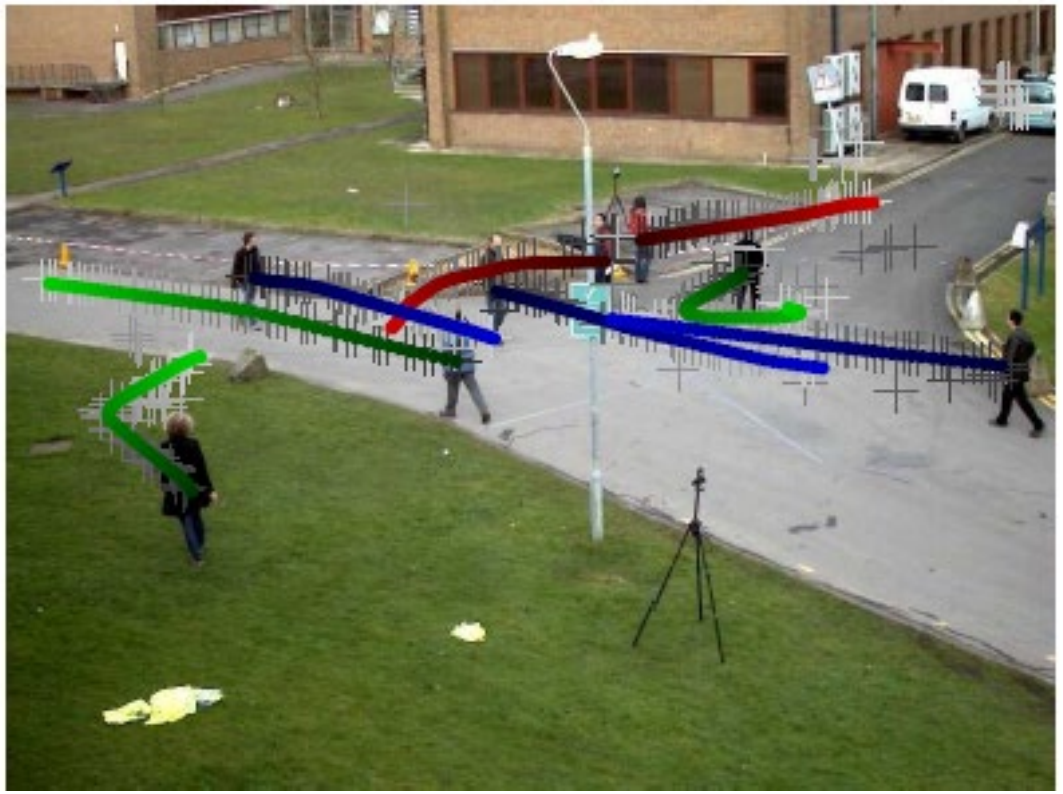
*Sequential* mod*els of continuous quantities of interest*




**Example: Nonlinear Time Series**

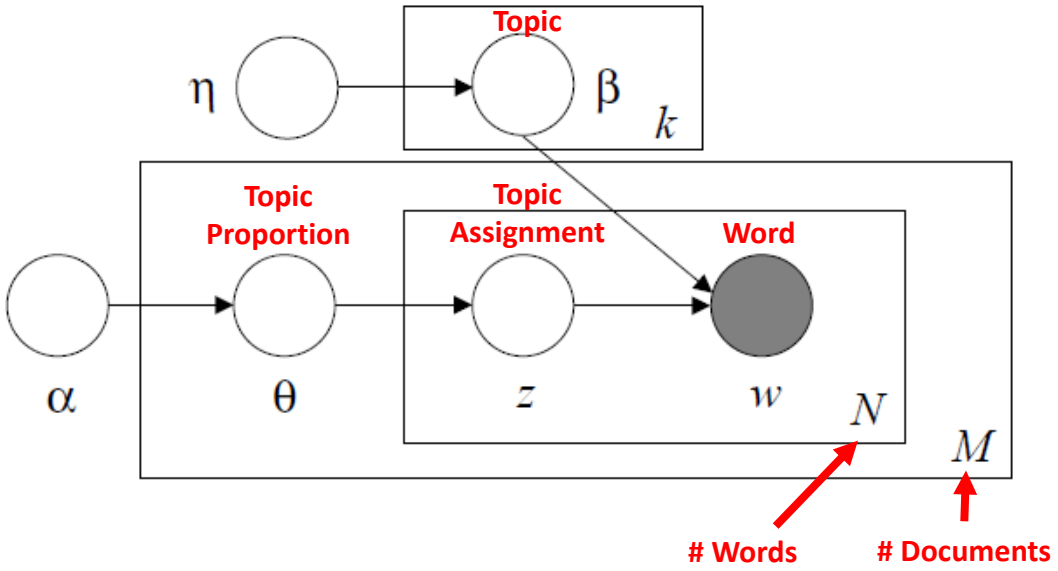



**Example: Multitarget Tracking**



# Example: Topic Modeling

**Latent Dirichlet Allocation (LDA)**




| "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 |

*This is really just a clustering model, but clusters words into topics*

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.

*Blei, D. et al., JMLR 2003*