THE UNIVERSITY OF CHICAGO

STAT 37710 / CAAM 37710 / CMSC 35400
Machine Learning

Introduction

Cong Ma

What is machine learning?

- Wiki's definition of machine learning (ML):
  - Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks ---adapted from Tom Mitchell

Examples of machine learning

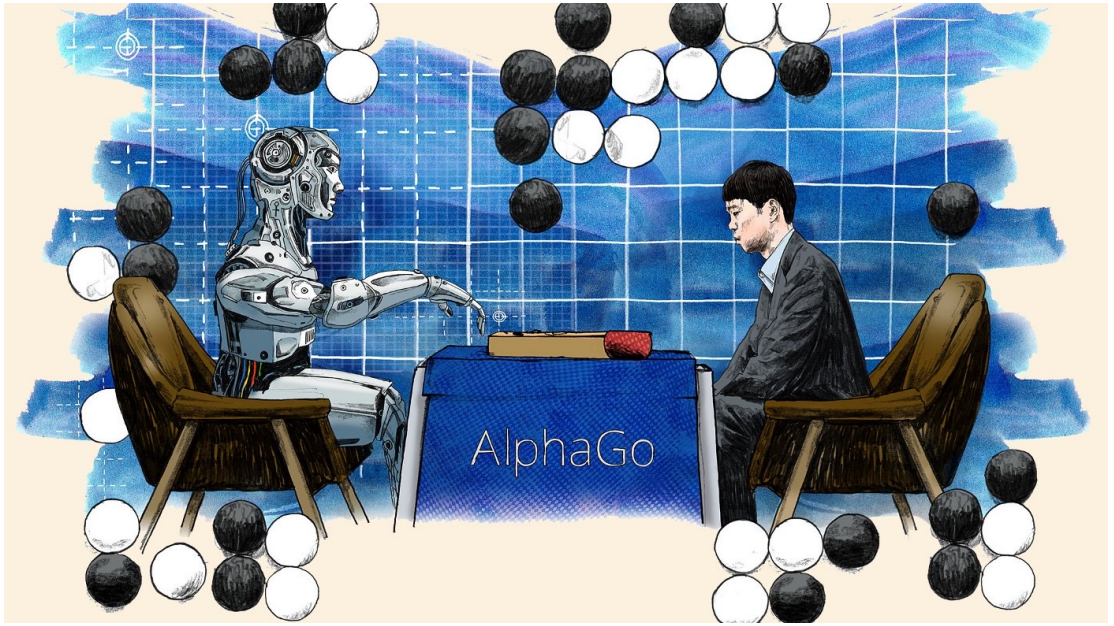
- spam filter
  - Data: emails together with labels
  - Task: label emails to either spam or non-spam
  - Performance: accuracy in labelling emails

Examples of machine learning

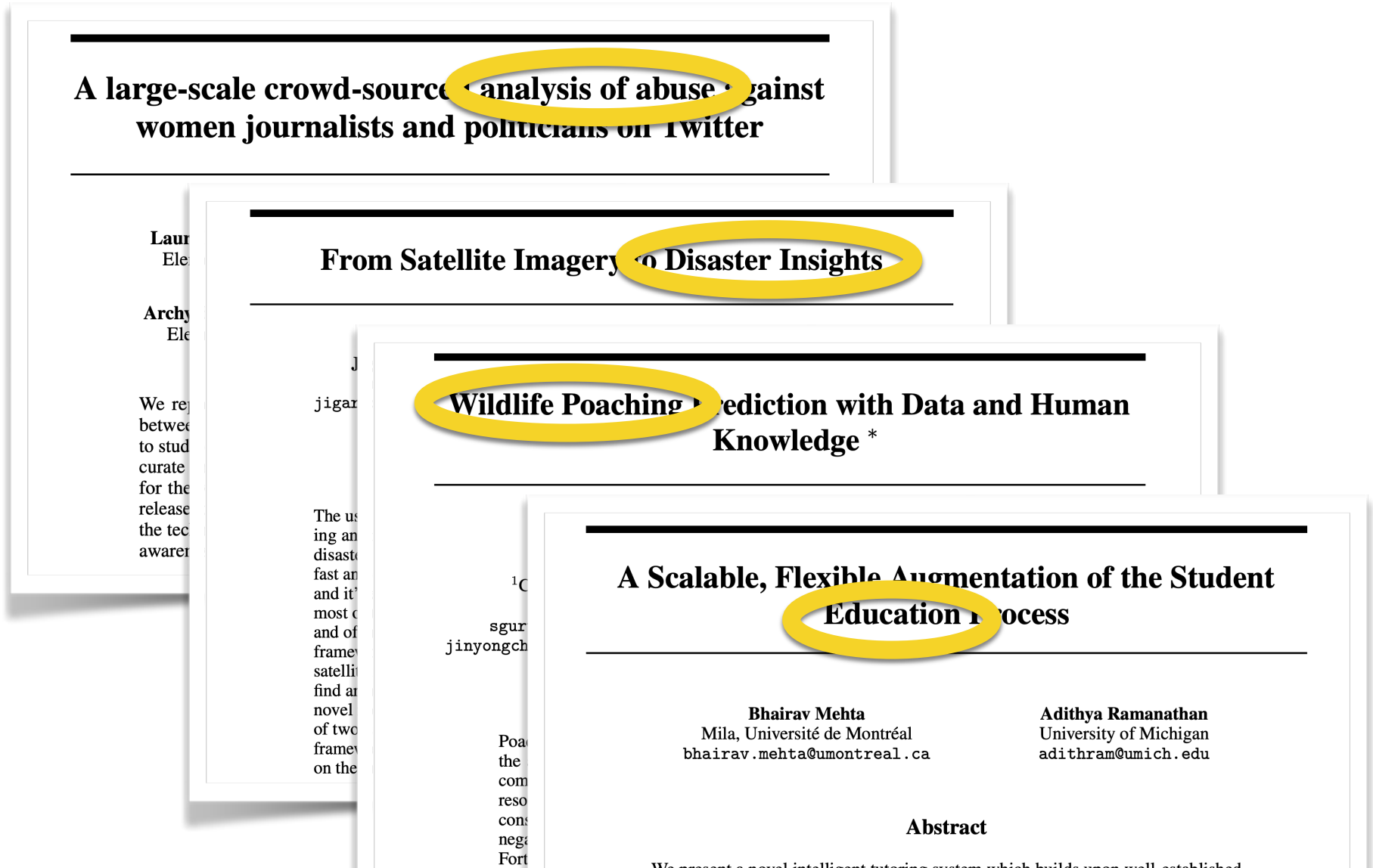
- image classification
  - Data: images together with labels (ImageNet)
  - Task: label images to categories (e.g., cat, dog)
  - Performance: accuracy in labelling images

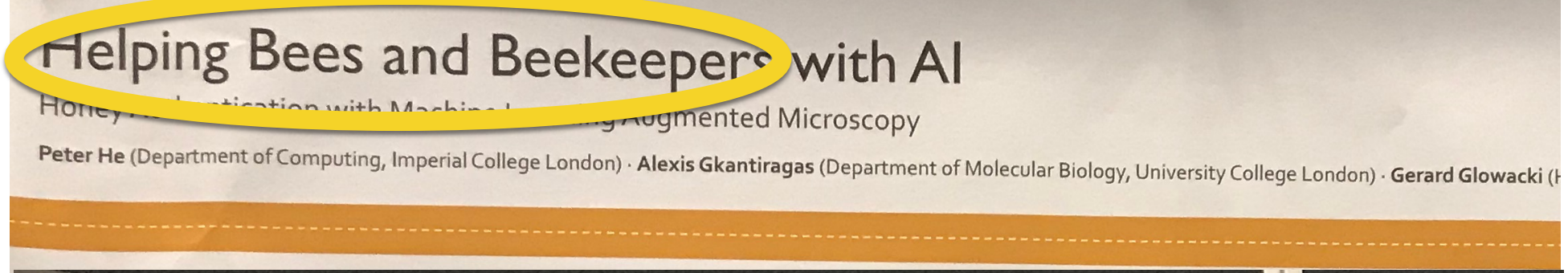
Examples of machine learning

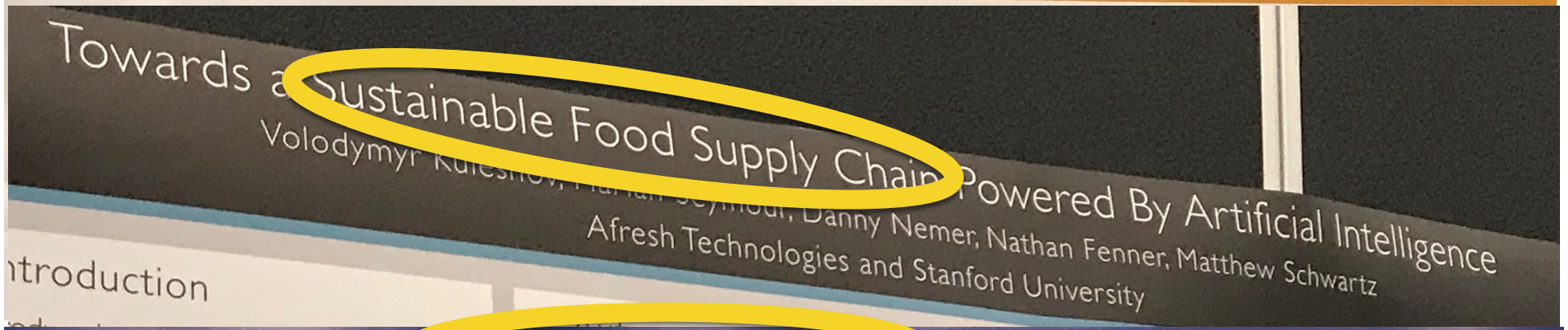
- Playing Go
  - Data: history of game playing
  - Task: playing Go well
  - Performance: winning rate against world champion

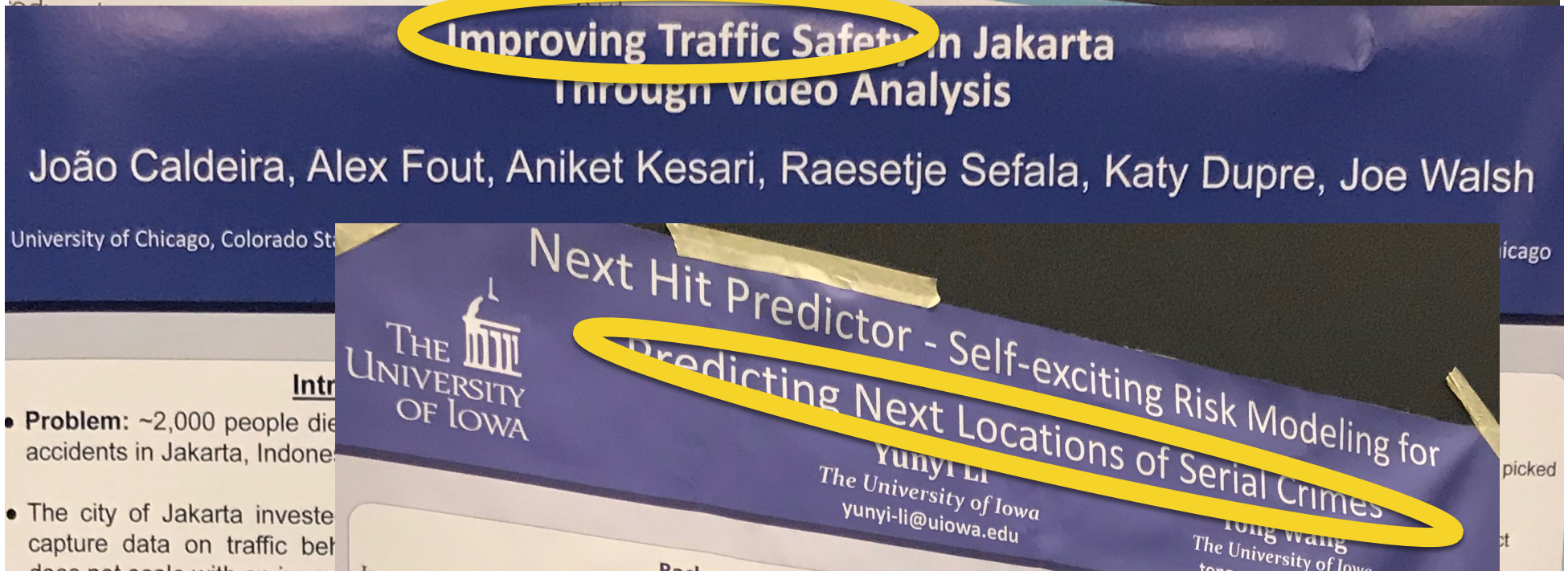


Machine learning is pervasive...




Helping Bees and Beekeepers with AI
Augmented Microscopy
Peter He (Department of Computing, Imperial College London) · Alexis Gkantiragas (Department of Molecular Biology, University College London) · Gerard Glowacki


Towards a Sustainable Food Supply Chain Powered By Artificial Intelligence
Danny Nemer, Nathan Fenner, Matthew Schwartz
Afresh Technologies and Stanford University
ntroduction


Improving Traffic Safety in Jakarta
Through Video Analysis
João Caldeira, Alex Fout, Aniket Kesari, Raesetje Sefala, Katy Dupre, Joe Walsh
University of Chicago, Colorado St
Intr
Problem: ~2,000 people die
accidents in Jakarta, Indone
The city of Jakarta investe
capture data on traffic beh
Next Hit Predictor - Self-exciting Risk Modeling for
Predicting Next Locations of Serial Crimes
THE UNIVERSITY OF IOWA
Yunyi Li
The University of Iowa
yunyi-li@uiowa.edu
The University of Iowa
picked

ML is interdisciplinary

statistics

information theory

Machine Learning

optimization

algorithms

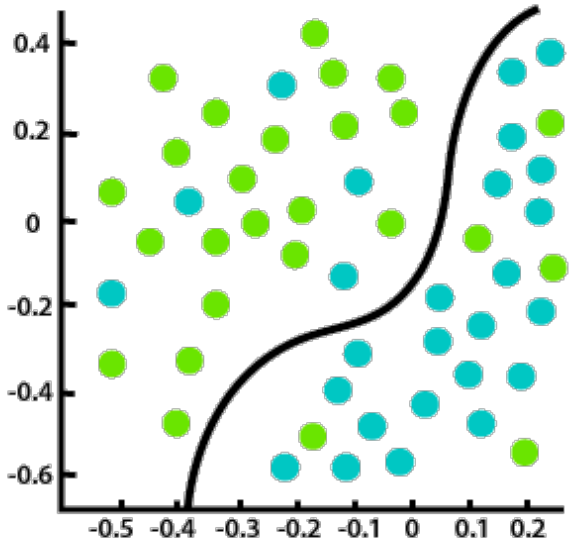
Machine learning tasks

- *Depending on the feedbacks, machine learning can be decomposed into*
  - *Supervised Learning*
    - Regression
    - Classification
  - *Unsupervised Learning*
    - Clustering
    - Dimension reduction Anomaly detection, …
- *Many other specialized tasks*

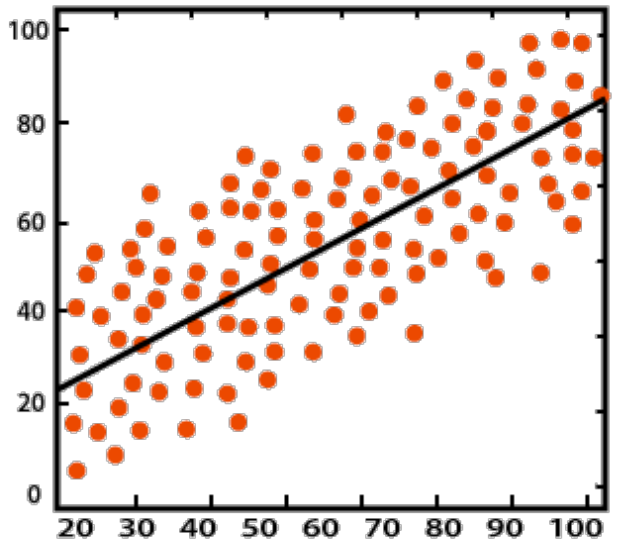
Supervised learning

$$f : X \rightarrow Y$$

Supervised learning

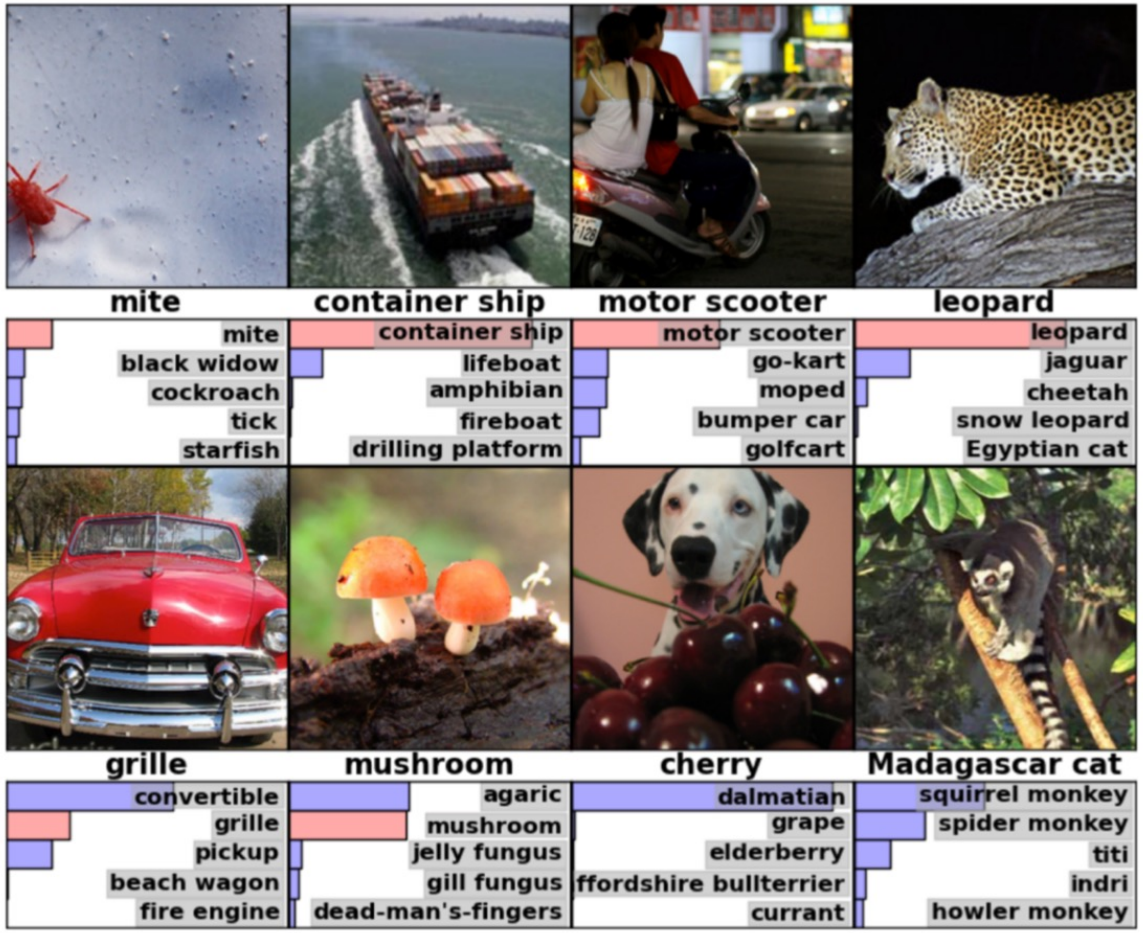


Classification



Regression

Example: Image classification



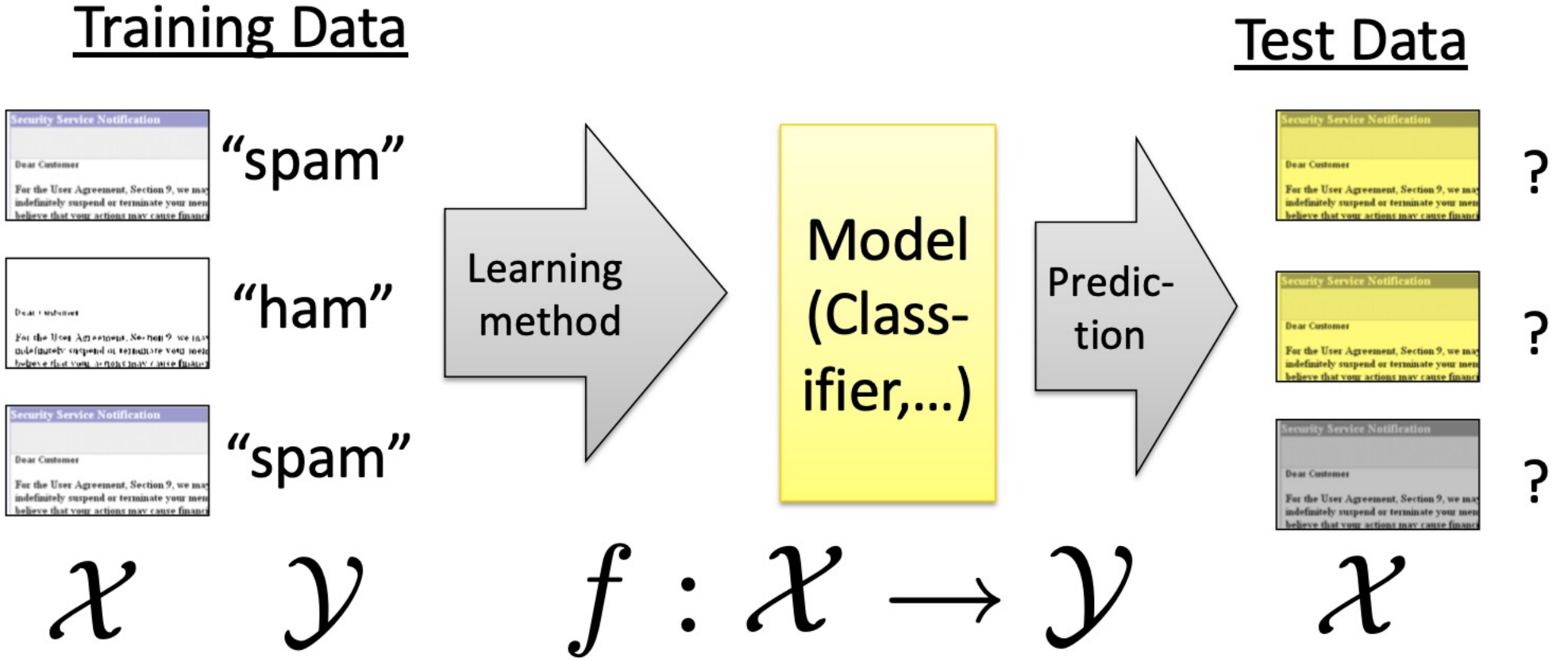
Krizhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks ‘12

Regression

- **Goal**: Predict real valued labels (possibly vectors)
- Examples:

X	Y
Flight route	Delay (minutes)
Real estate objects	Price
Patient & drug	Treatment effectiveness ...

Basic supervised learning pipeline

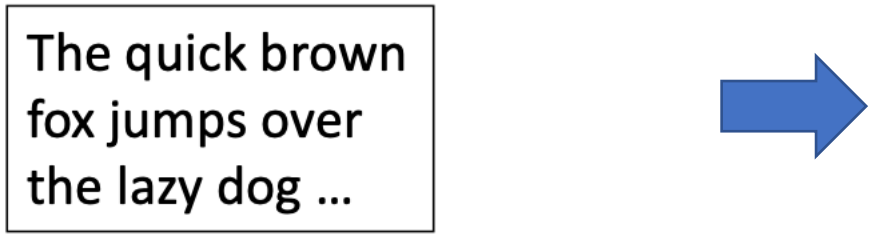


Example: Classifying documents

- **Input**:
  - Documents with labels, but how to represent documents
- **Goal**:

Representing data

- Learning methods expect standardized representation of data
  - (e.g., Points in vector spaces, nodes in a graph, similarity matrices ...)



- Concrete choice of representation ("features") is crucial for successful learning
- This class (typically): feature vectors in R^d

Example: Bag-of-words

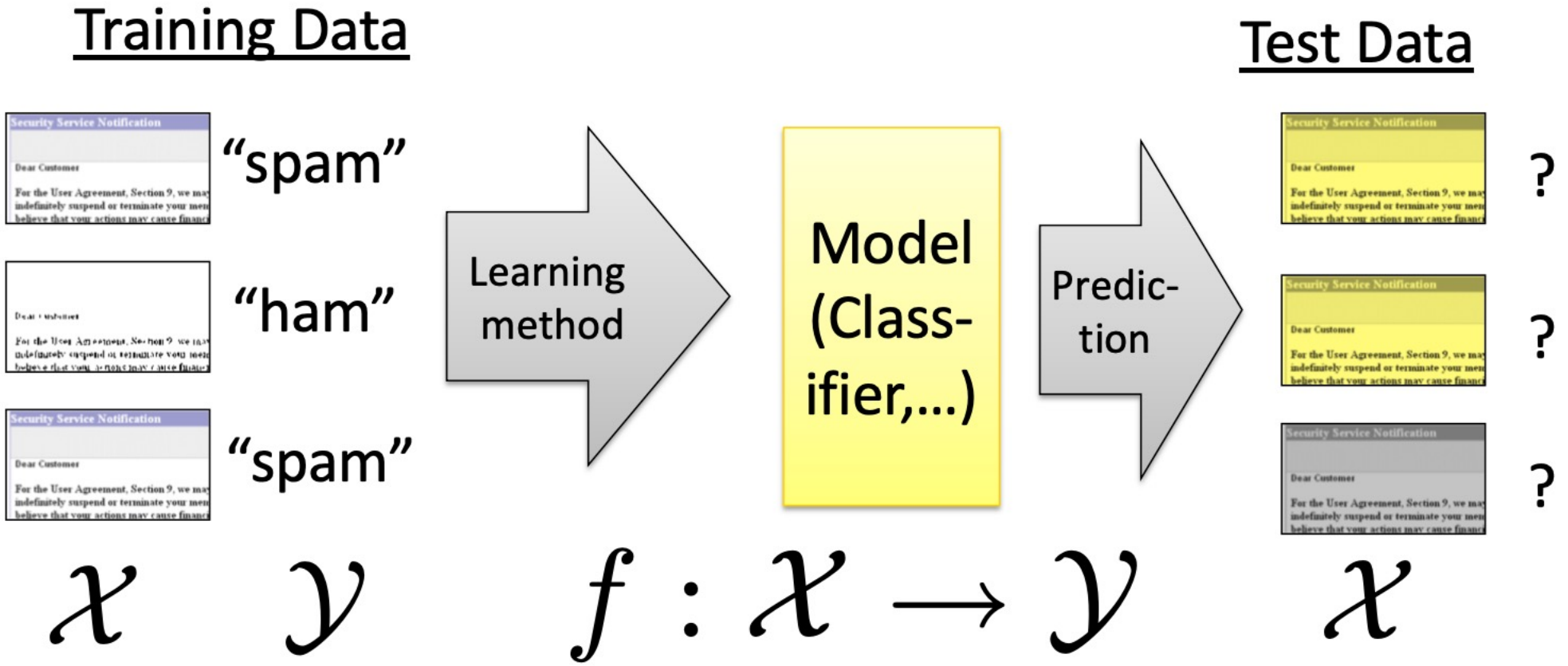
- Suppose language contains at most *d=100000* words
- Represent each document as a vector $\boldsymbol{x}$ in R^d
- *i*-th component x_i counts occurrence of *i*-th word

Word	Index
a	1
abandon	2
ability	3
...	
is	578
...	
test	2512
...	
this	2809
....	

Bag-of-words: Improvements

- Length of the document should not matter
  - Replace counts by binary indicator (yes/no) Normalize to unit length
- Some words more “important” than others
  - Remove “stopwords” (the, a, is, ...)
  - Stemming (learning, learner, learns -> learn)
  - Discount frequent words (tf-idf)
- Bag-of-words ignores order
  - Consider pairs (n-grams) of consecutive words
- Does not differentiate between similar and dissimilar words (ignores semantics)
  - Word embeddings (e.g., word2vec, GloVe)

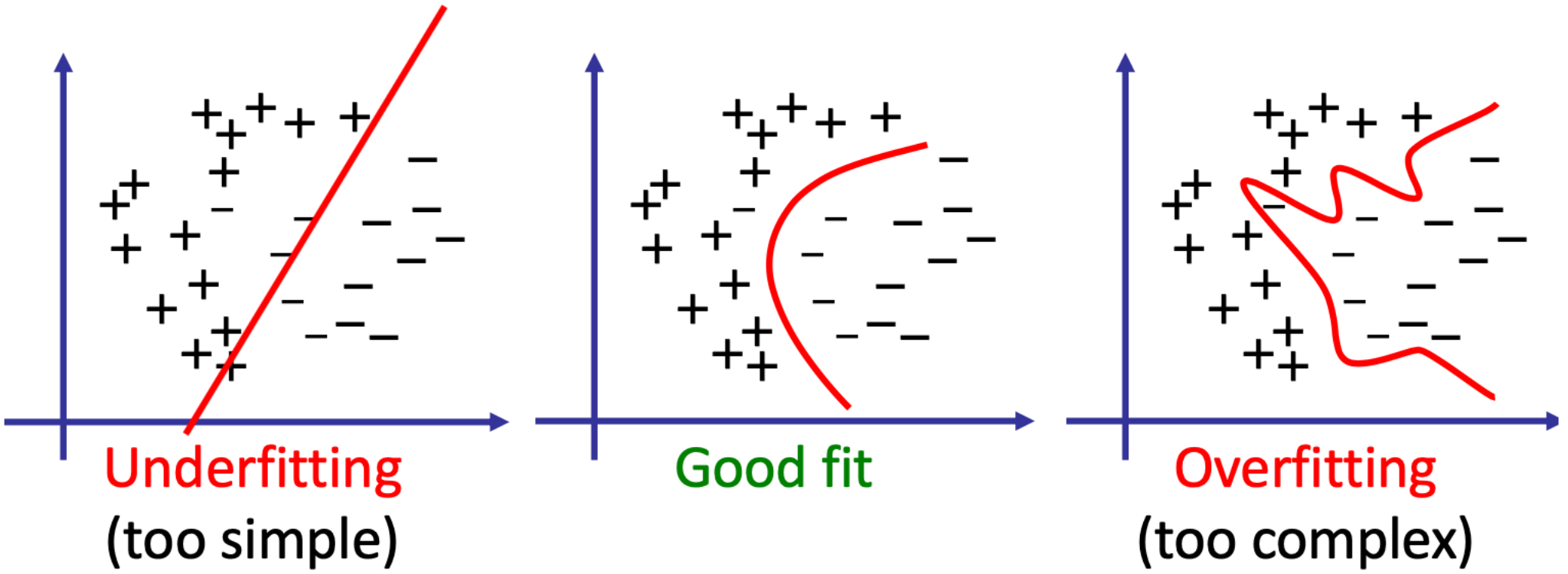
Basic supervised learning pipeline



Model class

- Linear
- decision tree
- random forests
- graphical models
- deep neural networks

Model selection and validation



Other models of learning

- Unsupervised learning
  - Learning without labels
- Semi-supervised learning
  - Learning from both labeled and unlabeled data
- Transfer learning
  - Learn on one domain and test on another
- Active learning
  - Acquiring most informative data for learning
- Online learning
  - Learning from examples as they arrive over time
- Reinforcement learning
  - Learning by interacting with an unknown environment

Summary

- Where we are
  - Basic forms of learning:
    - Supervised learning and other modes of machine learning
  - Key challenge in ML
    - Trading goodness of fit and model complexity
  - Representation of data is of key importance

- What's next
  - Formally state machine learning problems
  - Estimation theory and bias-variance tradeoff