IRDS: Data Mining Process

Charles Sutton
University of Edinburgh

(many figures used from Murphy. *Machine Learning: A Probabilistic Perspective*.)

“Data Science”

- Our working definition
  - Data science is the study of the computational principles, methods, and systems for extracting knowledge from data.
- A relatively new term. A lot of current hype…
  - “If you have to put ‘science’ in the name…”
- Component areas have a long history
  - machine learning
  - databases
  - statistics
  - optimization
  - natural language processing
  - computer vision
  - speech processing
  - applications to science, business, health….
- Difficult to find another term for this intersection

The term “data mining”

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner.

— Hand, Mannila, Smyth, 2001

The term “data mining”

Data mining is the analysis of (often large) **observational data sets** to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner. — Hand, Mannila, Smyth, 2001

not collected for the purpose of your analysis

The term “data mining”

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data **in novel ways** that are both understandable and useful to the data owner.

— Hand, Mannila, Smyth, 2001

Many “easy” patterns already known

e.g., pregnant example from association rule mining

The term “data mining”

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both **understandable and useful** to the data owner. — Hand, Mannila, Smyth, 2001

Tradeoff between

- predictive performance
- human interpretability
  Ex: neural networks vs decision trees

Before I get too far ahead of myself…

What problem am I trying to solve?

Problem Types

- Visualization
- Prediction: Learn a map **x** —> y
  - Classification: Predict categorical value
  - Regression: Predict a real value
  - Others
    - Collaborative filtering
    - Learning to rank
    - Structured prediction
- Description
  - Clustering
  - Dimensionality reduction
  - Density estimation
  - Finding patterns
    - Association rule mining
    - Detecting anomalies / outliers

supervised learning

unsupervised learning

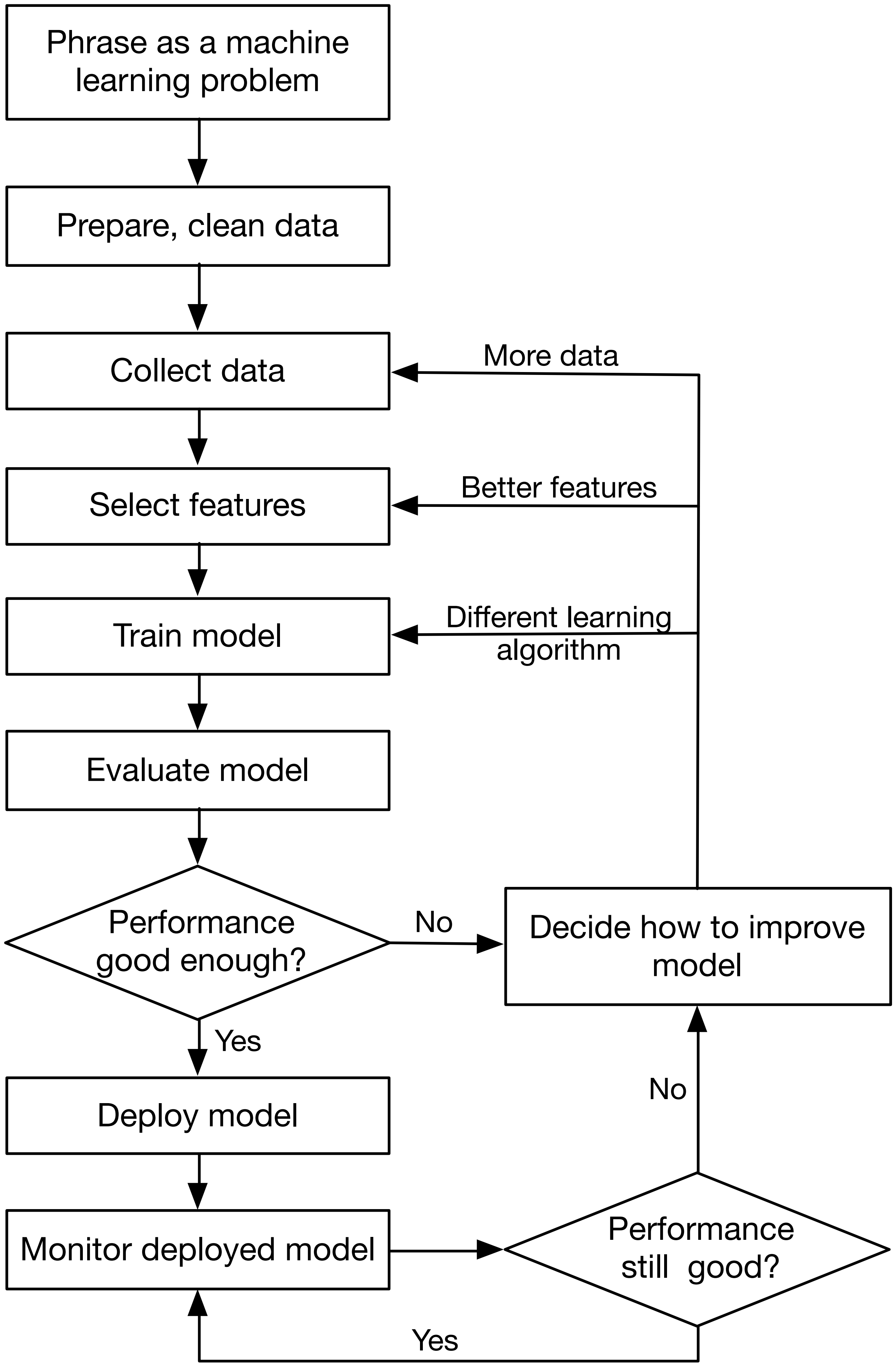
Prediction Examples

- Classification
  - Advertising
    - Ex: Given the text of an online advertisement and a search engine query, predict whether a user will click on the ad
  - Document classification
    - Ex: Spam filtering
  - Object detection
    - Ex: Given an image patch, dose it contain a face?
- Regression
  - Predict the final vote in an election (or referendum) from polls
  - Predict the temperature tomorrow given the previous few days
- Sometimes augmented with other structure / information
  - Structured prediction
    - Spatial data, Time series data
    - Ex: Predicting coding regions in DNA
  - Collaborative filtering (Amazon, Netflix)
  - Semi-supervised learning

Description Examples

- Clustering
  - Assign data into groups with high intra-group similarity
    - (like classification, except without examples of “correct” group assignments)
  - Ex: Cluster users into groups, based on behaviour
    - Social network analysis
  - Autoclass system (Cheeseman et al. 1988) discovered a new type of star,
- Dimensionality reduction
  - Eigenfaces
  - Topic modelling
- Discovering graph structure
  - Ex: Transcription networks
  - Ex: JamBayes for Seattle traffic jams
- Association rule mining
  - Market basket data
  - Computer security

Data Analysis Process



Inspired by Wagstaff, 2012.
"Machine Learning that Matters"

For another more industrial process, see CRISP-DM.

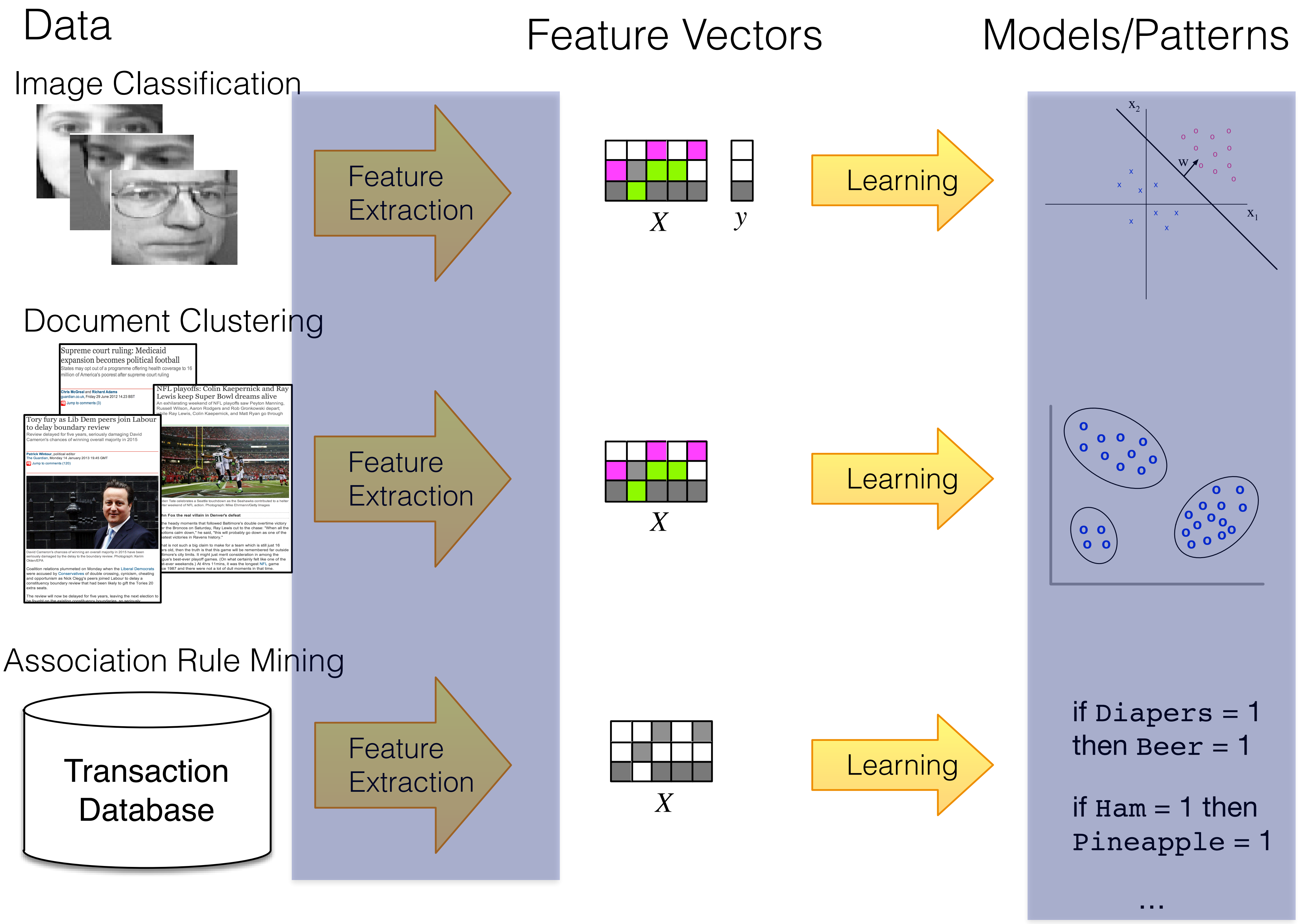
Roadmap

In the next few weeks, we'll talk about

- Visualization
- Feature extraction
- Evaluation and debugging

But to talk about these, we still need to understand **representation** behind the algorithms

Two Representation Problems



Data

Feature Vectors

Models/Patterns

Image Classification

Document Clustering

Association Rule Mining

Transaction Database

Feature Extraction

Feature Extraction

Feature Extraction

Learning

Learning

Learning

if Diapers = 1 then Beer = 1

if Ham = 1 then Pineapple = 1

...

X
y
X
X
x_2
x_1
w
Supreme court ruling: Medicaid expansion becomes political football
NFL playoffs: Colin Kaepernick and Ray Lewis keep Super Bowl dreams alive
Tory fury as Lib Dem peers join Labour to delay boundary review

1. What is the feature vector?

2. Given input, set of possible models?

Two Representation Problems

1. What features to use
2. What is the space of possible models

- In these lectures, we discuss features.
  - For model, see —> IAML, PMR, MLPR
- But: To pick features, must understand model.
- So: Whirlwind tour of models, leaving out learning algorithms

Summary

- Different types of model structures
    1. Linear boundaries (for classification and regression)
    2. Nonlinear boundaries (but linear in a set of features)
    3. “Wavy” boundaries (nonparametric, piecewise linear)
    4. Convex boundaries (with respect to Euclidean distance)
- This will affect feature construction, soon.