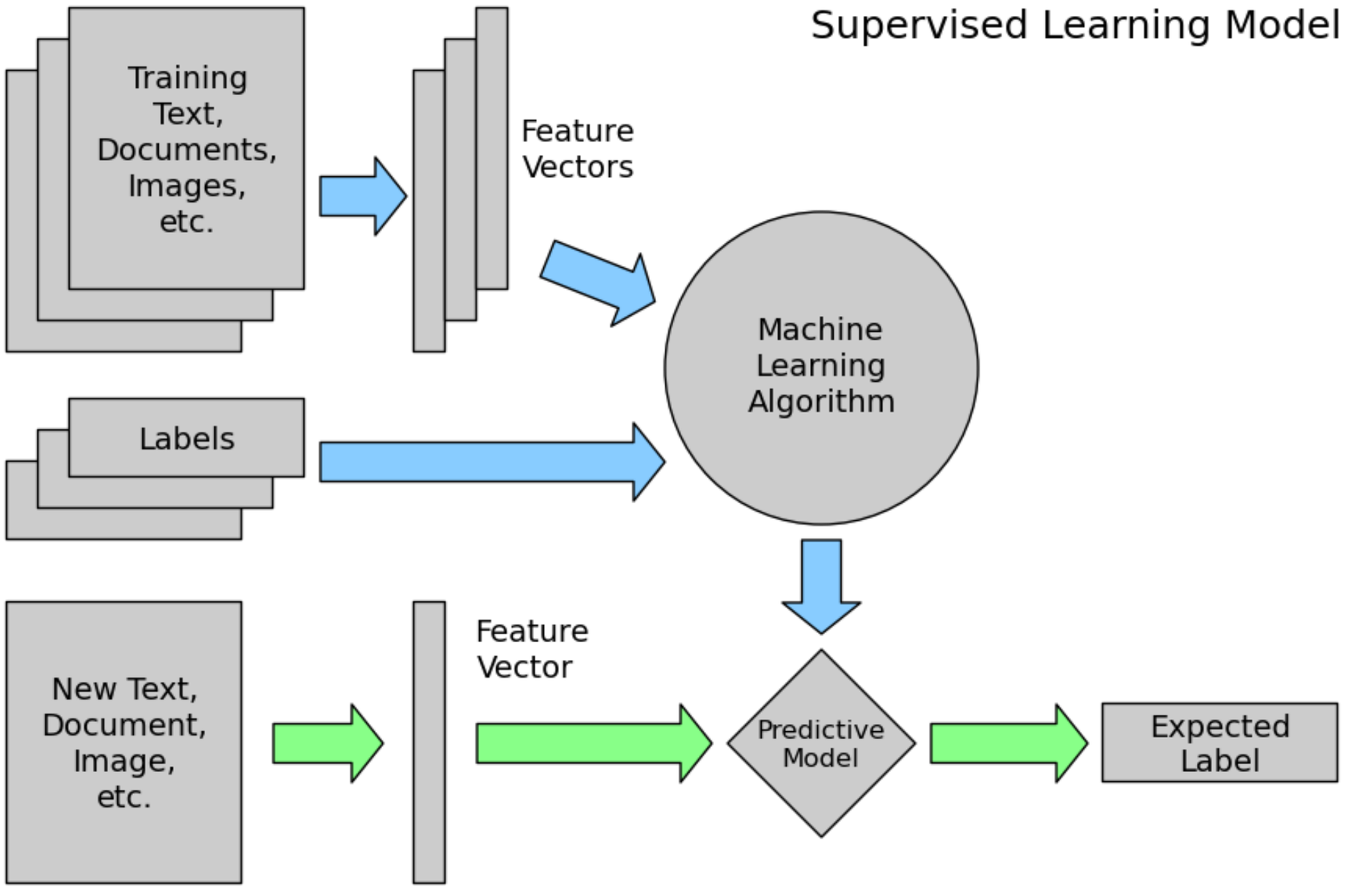
Intro to prediction problems



Reminders

- Thought questions due this Thursday
- Does anyone want a lab this week
  - to ask questions, generally?
  - for help with the assignment?
- All assignments are individual: you must do the assignment yourself
  - UofA takes academic honesty very seriously; if you copy or cheat I will have to report you and you could be kicked out of the program
  - Giving away your solution is still cheating; if someone puts pressure on you to give them your solution, thats a pretty terrible colleague
  - Why cheat? Your personal integrity is important. In the real-world, you won't be able to cheat, so start practicing now (this course is easier than the real-world)

Prediction problem statement

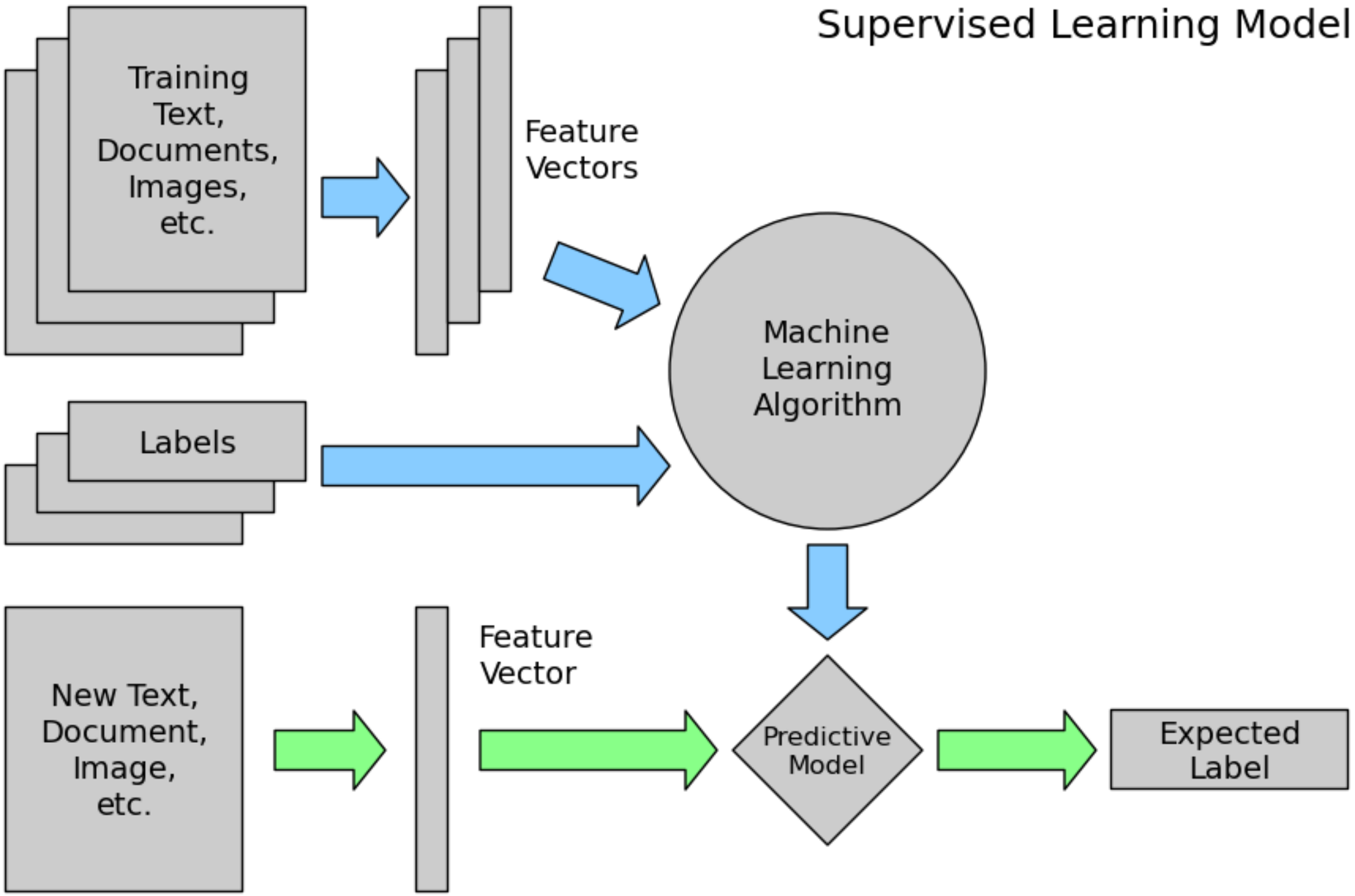
- Data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$

  $\mathbf{x}_i \in \mathcal{X}$ is the i-th object and $y_i \in \mathcal{Y}$ is the corresponding target designation

  We usually assume that $\mathcal{X} = \mathbb{R}^d$, in which case $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$

- Each xi is a data point or sample
- Each dimension of xi is called a feature or attribute
- Underlying assumption: features easy/easier to obtain and targets are difficult to observe or collect

Supervised learning



*from http://www.astroml.org/sklearn_tutorial/general_concepts.html

Data collection setting

- Assume passive setting: data has been collected, and now we must analyze it
  - As opposed to active learning or reinforcement learning, where an important component is deciding where to sample (explore)
- Assume data is i.i.d.
  - e.g., n flips of a coin,
  - e.g., n measurements of height, from a random sample of people
  - As opposed to time series prediction (temporal connections)
- Assume data is complete
  - As opposed to missing feature information

Types of predictions

- The target could be anything; convenient to separate into different types (even though can have related approaches)

- Generally two main types discussed
  - classification, e.g. $\mathcal{Y} = \{\text{sports}, \text{ medicine}, \text{ travel}, \ldots\}$.
  - regression, e.g. $\mathcal{Y} = \mathbb{R}$.

- Structured output often a type of classification problem
  - e.g., trees, e.g., strings

- Unsupervised learning: no labels, just structure of data
  - e.g., can sample be separated into two clusters?
  - e.g., does the data lie on a lower dimensional manifold?

Example: binary classification

	wt [kg]	ht [m]	T [°C]	sbp [mmHg]	dbp [mmHg]	y
$\mathbf{x}_1$	91	1.85	36.6	121	75	-1
$\mathbf{x}_2$	75	1.80	37.4	128	85	$+1$
$\mathbf{x}_3$	54	1.56	36.6	110	62	-1

Table 3.1: An example of a binary classification problem: prediction of a disease state for a patient. Here, features indicate weight (wt), height (ht), temperature (T), systolic blood pressure (sbp), and diastolic blood pressure (dbp). The class labels indicate presence of a particular disease, e.g. diabetes. This data set contains one positive data point ($\mathbf{x}_2$) and two negative data points ($\mathbf{x}_1$, $\mathbf{x}_3$). The class label shows a disease state, i.e. $y_i = +1$ indicates the presence while $y_i = -1$ indicates absence of disease.

Example: regression

	size [sqft]	age [yr]	dist [mi]	inc [$]	dens [ppl/mi^2]	y
$\mathbf{x}_1$	1250	5	2.85	56,650	12.5	2.35
$\mathbf{x}_2$	3200	9	8.21	245,800	3.1	3.95
$\mathbf{x}_3$	825	12	0.34	61,050	112.5	5.10

Table 3.2: An example of a regression problem: prediction of the price of a house in a particular region. Here, features indicate the size of the house (size) in square feet, the age of the house (age) in years, the distance from the city center (dist) in miles, the average income in a one square mile radius (inc), and the population density in the same area (dens). The target indicates the price a house is sold at, e.g. in hundreds of thousands of dollars.

Multi-class versus Multi-label

- Multi-class: must be exactly one class
  - e.g., can only be one of the blood types {A, B, O, AB}
  - Patient with features x (age, height, etc) has target y = A
  - …or represented as indicator vector y = [1 0 0 0]

- Multi-label: can be labeled with multiple categories
  - e.g., categories for articles could be {sports, medicine, politics}
  - an article can be a sports article and a medical article
  - The target y = {sports, medicine}
  - … or again could be represented with the indicator vector y = [1 1 0]

Exercise

- Imagine you have a binary classification problem and someone has given you p(y l x)
- Now you get a new sample, x
- What class might you pick (y = 0 or y = 1)?
- What if you have 4 classes (y = 0, y = 1, y = 2, y = 3)?

Thought exercise: specifying prediction problems

- Imagine someone has given you a database of samples
  - e.g., Netflix data of people's movie rankings

- What do you need to consider before starting to learn a predictor? (without knowing much about algorithms yet)
  - How do I know if my predictions are successful? What is my measure?
  - What simple algorithms can I try first? What are the baselines?
  - How many samples are there? Is it a large database?
  - Is efficiency important?
  - Is the data useful? Could it be significantly improved with different and/or more data collection?
  - How should I represent my prediction problem?

Mini-project

- Pick a dataset
- Pick three algorithms
  - e.g., decision tree, logistic regression and neural network
- Learn three models
- Hypothesis: the decision tree model will make more accurate predictions on new data, for this setting
- Goal: test this hypothesis

Summary so far

- We will learn parameters to distribution models
  - For (joint) distributions p(x | theta) and conditional distributions p(y | x, w)
- We formalize the problem using maximum likelihood and MAP
  - In maximum likelihood, we find optimal parameters theta for p(D | theta) (i.e., theta that makes data most likely)
  - In MAP, also have expert defined prior over parameters p(theta), and optimize p(D | theta) p(theta) (i.e., theta that makes data most likely, but also satisfies prior as much as possible to balance the two)
  - …but how does this relate to prediction? Why MAP?

Optimal prediction

- We want to learn a prediction function f(x) = y
- We will need to define cost/error of a prediction
  - Cost function $c : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$
  - For true target y, get cost $c(\hat{y}, y)$
- We will see that modeling p(y l x) is useful for this task
- Want to find predictor that minimizes the expected cost
  - could choose other metrics, such as minimize number of costs that are very large or minimize the maximum cost

Whiteboard

- Expected cost
- Bayes optimal models
- Next topic: linear regression