Prolegomena unit:

- Course introduction (Monday)
- Basic machine learning terminology (**today**)
- Lab: Python libraries (Friday)
- (Start *The nature of data* on Wed, Jan 22)

Today:

- A few logistical details
- Machine Learning in context
- Fundamental vocabulary
- Demonstration
  - A classification example
  - A regression example

Machine learning. The field of computer science that studies techniques for training algorithms from data.

Artificial intelligence. (Moving target.)

Statistical inference. The area of mathematics that studies building and evaluating statistical models from data.

Data science. An umbrella category for a variety of fields or activities, including data curating, data mining, data analytics, and predictive analytics.

Pattern matching. A field similar to machine learning but from an engineering origin.

There's a joke that says a data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.

Joel Grus, Data Science from Scratch, 2015, pg 1

he problem of searching for patterns in data is a fundamental one and has a long and successful history. For instance, the extensive astronomical observations of Tycho Brahe in the 16th century allowed Johannes Kepler to discover the empirical laws of planetary motion, which in turn provided a springboard for the development of classical mechanics. Similarly, the discovery of regularities in atomic spectra played a key role in the development and verification of quantum physics in the early twentieth century. The field of pattern matching [used as a synonym for machine learning] is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories.

Bishop, Pattern Recognition and Machine Learning, 2006, pg 1

Model The function/program/component/tool we want to make.

Observation. A value of the domain of the model; a data point.

Target. A value of the codomain of the model; an (correct) output, given an observation.

Dataset. A collection of observations collated with targets used for training a model. A model, once trained by a dataset, **predicts** new target values, given new observations.

Feature. A component to an observation/data point. The dimensionality of the observations is the number of features.

Machine learning main tasks:

- Regression, where the target type is $\mathbb{R}$
- Classification, where the target type is a finite set
  - Binary classification, where the target is $\{F, T\}$ (or $\{0, 1\}$ or $\{-1, 1\}$ . . .)
- Density estimation, where the target type is $[0, 1]$.

Other machine learning tasks (see Goodfellow, *Deep Learning*, pg 96–100):

- Transcription, where the observations are unstructured and the targets are text.
- Machine translation, where the observations and targets are text.
- Anomaly detection, where the targets are indicators of whether the observation is atypical.
- Synthesis and sampling, where there are no observations in deployment, but rather the program produces new observations similar to those in training.
- Denoising, where the targets are corrected versions of the observations.

Model family. The general form of a function chosen to train a model from.

Parameters. Constant values used to specify a model from a model family; also called **weights** or **coefficients**.

Hyperparameters. Variables used to specify options within the training algorithm.

Training. Finding parameters to specify a model; also called **learning**, **estimation**, or **fitting**.

Loss function. A function used to evaluate a model (*low* values are good); related terms are **error**, **cost**, and **risk**.

Training set. The portion of an available data set chosen for use in training.

Test set. The portion of an available data set, distinct from the training set, used in testing or evaluating the model trained by the training set; related terms are **held-out set** and **validation set**.

Generalization. The extent to which a trained model performs well on data other than that on which it was trained, such as the test set.

Overfitting. The failure of a model to generalize because training has caused it to reflect the idiosyncrasies of the training set.

Coming up:

Keep learning Python

Take basic-terminology quiz
(due end-of-day Thursday, Jan 16)

Do Python warm-up assignment
(due end-of-day Friday, Jan 17)

Read Hadley Wickham, "Tidy Data", Sections 1–3
(due end-of-day Wednesday, Jan 22)