Data Science for Smart Cities

CE88

Prof: Alexei Pozdnukhov
GSI: Madeleine Sheehan

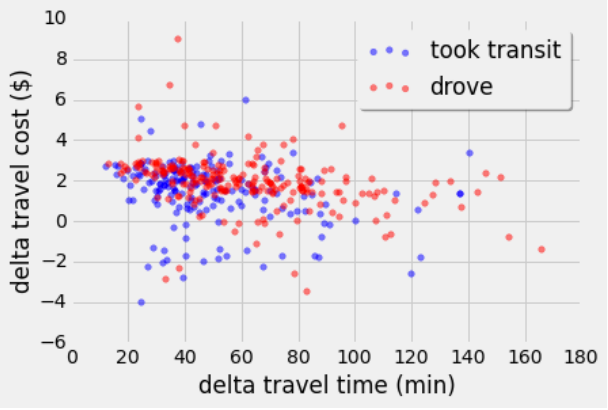
115 McLaughlin Hall

alexeip@berkeley.edu
m.sheehan@berkeley.edu
CE88 in title

Today

Reminder on the goals of Exploratory Data Analysis:
- suggest a hypothesis about the causes of the phenomena
- state the problem of data analysis
- assess the validity of the assumptions
- select an algorithm to approach the stated problem

Taxonomy of algorithms
Approaches to model specification



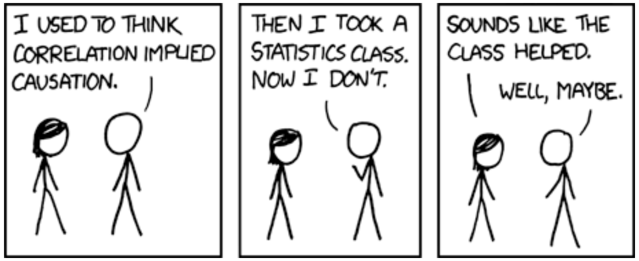
Mini Lab 7: specify and apply a simple predictor algorithm

“Statistics starts with data”

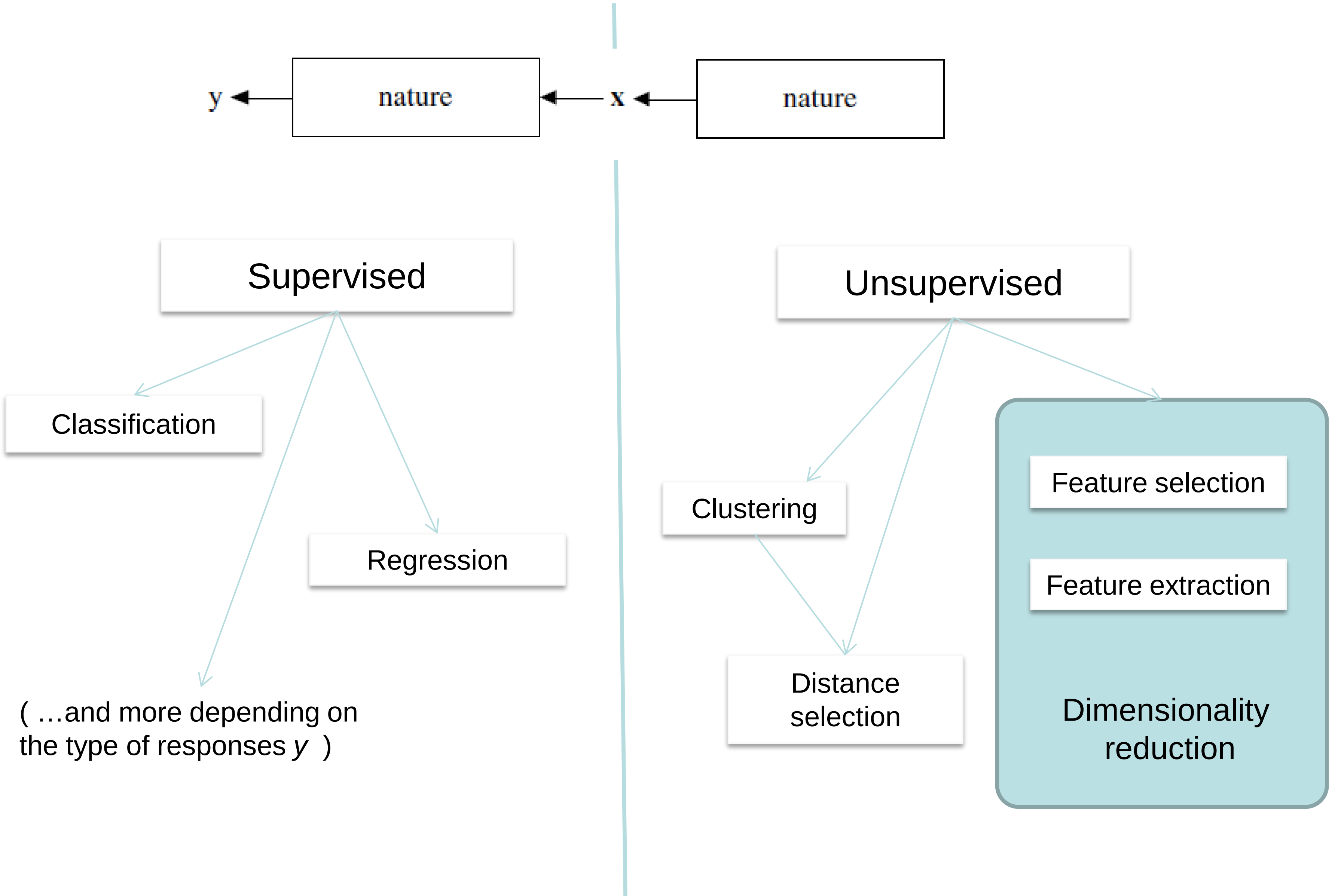
Think of the data as being generated by a black box in which input variables *x* (predictor or inputs, independent or explanatory variables) go in one side, and on the other side the response variables *y* come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

There are at least **two** goals in analyzing the data:

- To be able to predict what the responses are going to be to future input variables;
- To extract some information about *how* nature is associating the response variables to the input variables.



Taxonomy of problems and algorithms



Supervised modeling

Nowhere it is set in stone how nature black box should be modelled when the only thing you got is empirical data. It is **our** choice to specify an appropriate model. How the translation from subject-matter problem to statistical model is done is often the most critical part of an analysis.

Best thing one can do is to specify a model in a way that achieves both goals we just stated.

One can assume a stochastic data model for the inside of the black box, i.e. assume that data are generated by independent draws from:

response variables = *f* (predictor variables, random noise, parameters).

Then the values of the parameters can be estimated from the **sample** and the model then used for information (inference) and/or prediction for the **population**.

Statistical modelling

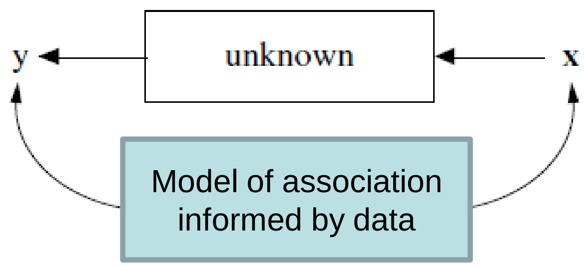
If one starts with “assume that the data are generated by the following model: … “, this implies that by imagination and by looking at the data, one can invent a reasonably good parametric class of models for a complex mechanism devised by nature.

The conclusions made with such approach could be true about the model’s mechanism, but not necessarily about the nature’s mechanism. If the model is a poor emulation of nature, the conclusions maybe wrong.

This approach is most solid when indeed the nature of the observed association between x and y can be identified, for example, when a known physical process is observed under noise (noise in the process parameters or as additive random factors to the observable).

Statistical modelling

Most often, the inside of the nature box is complex and unknown.



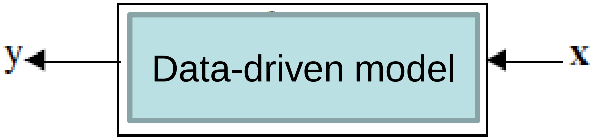
The approach to find an algorithm that operates on **x** to predict the responses y must be well thought through. Particularly, to make statistical inference about the nature of the x to y association, one has to:

- control for confounding factors
- use randomized controlled trials
- validate and test models on previously unobserved (out-of-sample) data

Analysis of models

When the model is built, we are comparing:

vs



The great advantage of a model is if it produces a simple and understandable picture of the relationship between the input variables and responses.

Examples are:

- decision rule based on thresholds

```
if  (travel distance X > 10 miles):
          person Y travels by car
else:
          person Y travels by bike
```

- model based on empirical evidence

```
if (most similar person to person Y
                    travels by car):
          person Y travels by car
```

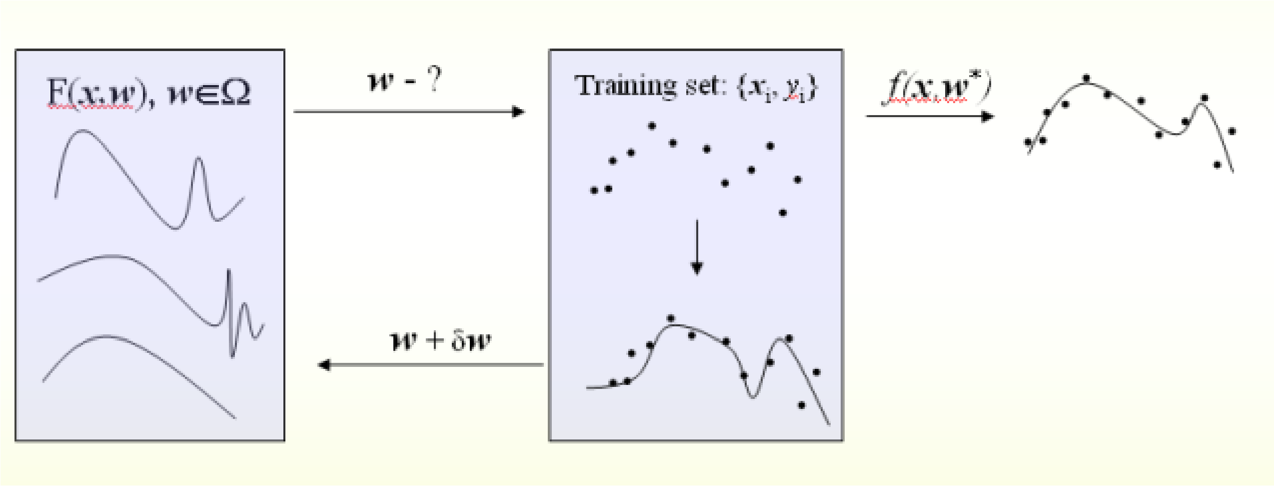
- model linear in parameters

$$y = b_0 + \sum_{1}^{M} b_m x_m + \varepsilon$$

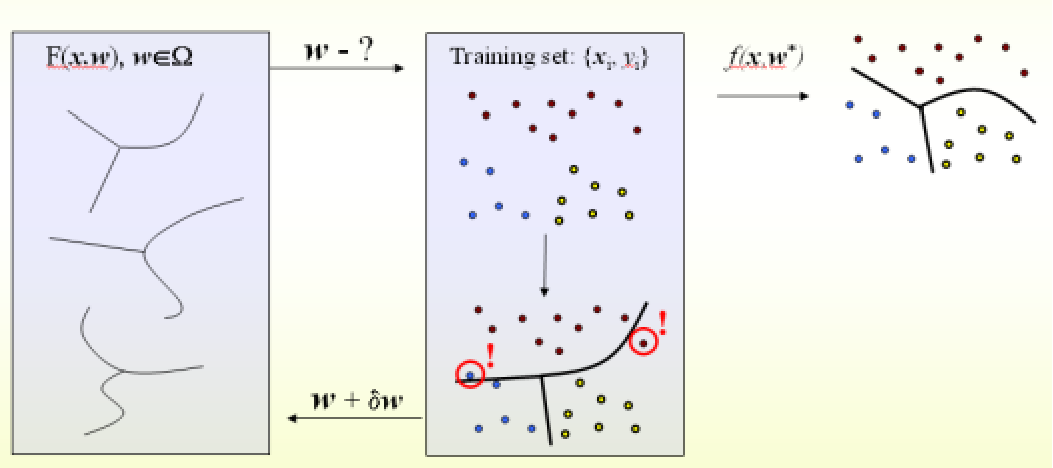
Note: such models may seem oversimplified, and not fitting the data perfectly. The choice between accuracy and interpretability is a difficult one. In a choice between accuracy and interpretability, practitioners often go for interpretability. However, a model does not have to be simple to provide reliable information about the relation between predictor and response variables.

Algorithmic framework

Regression



Classification



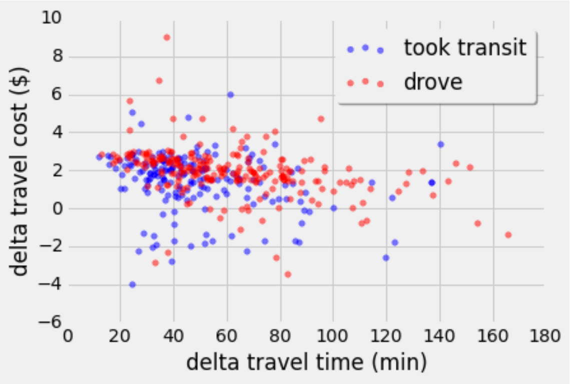
Geometry of the input space

A simple prediction method is a ‘nearest neighbor’:

Given an unseen situation $\mathbf{x_0}$, provide a likely y_0 that can be associated to it:

- find an example $\mathbf{x}$ in the available dataset that is most similar to $\mathbf{x_0}$
- assign an observed response variable y as your best guess for y_0

What do we mean by ‘nearest’ or ‘most similar’?



It has to be an informed decision made by us!

Ideas:

- use Euclidean distances,
- normalize the input variables,

or

- apply scaling to input variables justified by domain knowledge.

Take away ideas

- The are multiple tasks one can think of when dealing with data,
so keep in mind a taxonomy of algorithms and methods

- Methods deal with a sample and are used to get information (infer) and/or predict for the entire population

- Methods can be conceptually represented as an algorithmic model of nature

- Algorithms can have parameters that one can 'tune' to achieve better performance

In the Mini-lab:

1) we will learn a simple programming concept to implement algorithms
2) we will solve a simple prediction problem