DCF Week 7 Warm-Up

Data and Computing Fundamentals

```
require(rpart)
```

```
## Loading required package: rpart
```

Tonight is about review and adding an important new approach to interpreting data.

We have spent most of our time on two subjects:

1. making data-oriented graphics
2. transfiguring data: getting from the data we are given to the "glyph-ready" data that you need to make a graphic or some other mode to guide our interpretation of the data.

- Problems from Stanford/CMC intro computer science course.
- A new data set to explore with graphics
- If time, a clustering approach to genetics.

Then we'll go on to machine learning. Think of it as a way of presenting data that allows us to handle multiple variables.

What is learning? Recognizing patterns knowing facts. Gutenberg Bible, 1492, Reformation

Perhaps "understanding" is broader: knowing where to look for patterns and how to apply a kind of pattern to a new situation.

Machines can learn: they can spot patterns.

Supervised and Unsupervised.

1. Two problems from Stan

Unsupervised Learning

Cluster the countries

NCI 60

Show a glaucoma prediction

```
data("GlaucomaM", package = "TH.data")
glaucoma_rpart <- rpart(Class ~ ., data = GlaucomaM,
                        control = rpart.control(xval = 100)
```

Calculate a diabetes prediction

What are the highest risk groups?

Supervised Learning

Baby Names

Look at all the letters in girls' names, compare to boy's names.

As features, use length, last letter,

Predictors of Immigration

- Build a decision-tree model of the factors that determine strong one-way emigration. (Ratio of GDP, life expectancy, . . .)