101C

ADMIN STUFF

IMPORTANT STUFF

- Read the syllabus (on CCLE)
- Office Hours: Thursday 5:00-6:00pm
- Midterm in class: November 12
- Final: TBD
- Send all academic questions to Piazza, not email. Feel free to answer/participate in discussions.

IMPORTANT STUFF

- Please turn homework in on time. Late homework will not be accepted. And "late" means any moment after the deadline.
- We will do lots of working in teams/pairs.
- No enrollment after start of second lecture.

101C

PREAMBLES

STRATEGY VS TACTICS

Strategy without tactics is the slowest route to victory.
Tactics without strategy is the noise before defeat.

STRATEGY VS TACTICS

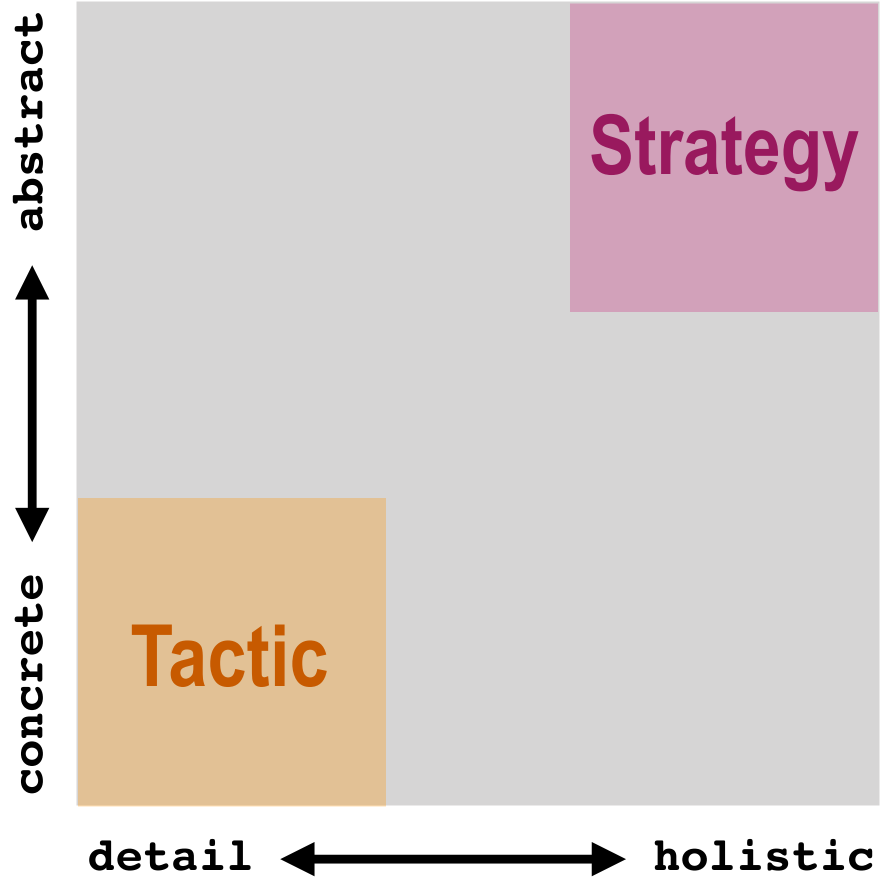
Strategy without tactics is the slowest route to victory.
Tactics without strategy is the noise before defeat.

– Sun Tzu

STRATEGY VS TACTICS

- Someone who understands fundamental concepts but who is weak with tactics will be wildly more successful than someone who has mastered the tactics but doesn’t know why they’re doing what they’re doing.
- Tactics : Strategy :: Knowledge : Wisdom

STRATEGY VS TACTICS



THE IMPORTANCE OF DATA

- Suppose we have a data set of $n = 1000$ observations of (x, y) pairs that have been realized (randomly) from a known joint density function $\phi(x, y)$.

- **Our objective** is to create a function, $y_0 = f(x_0)$, that will map new realizations (from ϕ) of x_0 to unobserved y_0 as precisely as possible (say, minimizes the RMSE if y is quantitative).

- **Question:** What is a good first step with our data?

THE IMPORTANCE OF DATA

- Ignore it.

THE IMPORTANCE OF DATA

$$\phi(Y\,|\,X) = \frac{\phi(X, Y)}{\phi(X)}$$

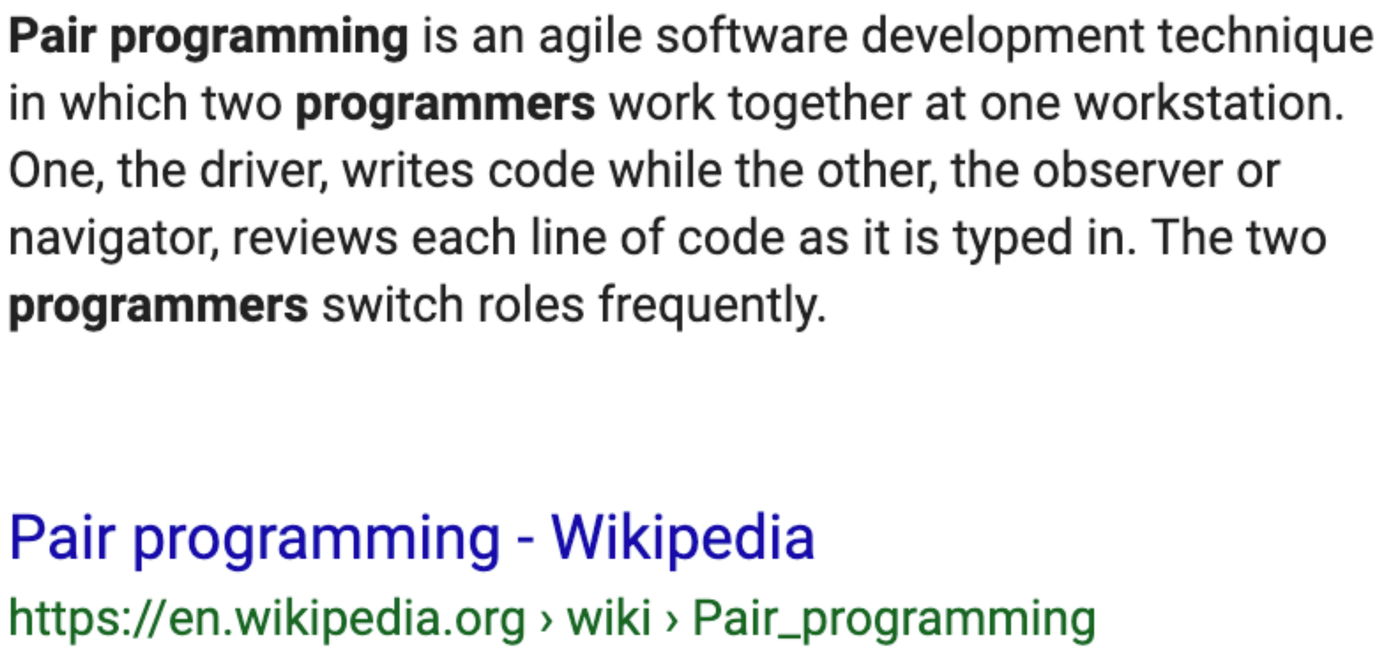
- E.g.,

$$E[Y\,|\,X] = \int_Y Y\,\phi(Y\,|\,X)\,dY$$

THE IMPORTANCE OF DATA

- The entire reason this course exists is because in real world settings, we don't know $\phi(x, y)$.

PAIR PROGRAMMING



Pair programming is an agile software development technique in which two **programmers** work together at one workstation. One, the driver, writes code while the other, the observer or navigator, reviews each line of code as it is typed in. The two **programmers** switch roles frequently.

Pair programming - Wikipedia

https://en.wikipedia.org › wiki › Pair_programming

101C

0

COURSE GOALS

- Understand the purposes and uses of statistical modeling.
- Learn a number of modeling approaches.
- Understand some of the strengths and weaknesses of these approaches.

RUN R CODE IN LECTURE

- We will be running R code in these lectures to illustrate various concepts.
- Please bring your laptop and be prepared to participate results of these R code examples/results.

HEADS UP

- Prefer using simulated data in lecture
    - Point of lectures is to focus on concepts
    - Don’t get caught up in variable definitions and units
    - Some data isn’t interesting to some

TERMINOLOGY

- x often represents an explanatory variable.
- $\mathbf{X}$ can be used to represent a matrix of explanatory variables.
- y commonly represents the response variable; the value we're seeking to predict/explain.

TERMINOLOGY

- $f(x)$ often represents the (true) function that maps the explanatory input to our response value.
- $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ is a linear function w.r.t. $\mathbf{x}$ – commonly and conveniently assumed in multivariate regression analysis.
- $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ is matrix shorthand notation for the above linear function.

PARAMETERS

- All functions use constants, either implicitly or explicitly.
- When some function is assumed to relate $\mathbf{x}$ to y, like $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, the constants that we believe describe important features of our population, like $\boldsymbol{\beta}$, are called parameters.
- Parameters are constant across the population.
- Our choice of $f(\mathbf{x})$ implies which parameters we feel are important.

PARAMETRIC

- Situations like linear regression, in which we assume the functional form of $f(\mathbf{x})$ (for example, we know that $f(\mathbf{x})$ is linear) are called parametric problems.
- Once we assume that, e.g., $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, it becomes a "simple" matter of estimating the parameters $\boldsymbol{\beta}$ (and sigma, the standard deviation of the errors).
- See pg 21

NON-PARAMETRIC

- In contrast, when it may be unclear how to explicitly describe $f(\mathbf{x};\boldsymbol{\theta})$ in terms of $\boldsymbol{\theta}$, OR, even if we did, it would be difficult to ascribe meaning to $\boldsymbol{\theta}$, we may turn to Non-parametric models.
- See pg 23

CROSS-VALIDATION

- One technique we'll make frequent use of is cross validation.
- In cross validation, we set aside some data and use it to test our model.
- Why? **Inference!**

TRAINING VS TESTING

- “Training” data are data used to fit a model.
- “Test” data are data that were NOT used in the fitting process, but are used to test how well your model performs on “unseen” data.

TRAINING VS TESTING

- For example, for your final exam, we’ve set aside some testing data that you’ll never see.
- You’ll use the data we give you to fit a model, and make predictions for a given set of covariates. Only we will know what the correct response values are, and will use this to test your model.

ADVICE

- **Read the Book!**

MISC

- Finals
- Before our next get together:
  - Read Ch 1
  - Read Ch 2.1, 2.2, 2.3
  - Skim Ch 5