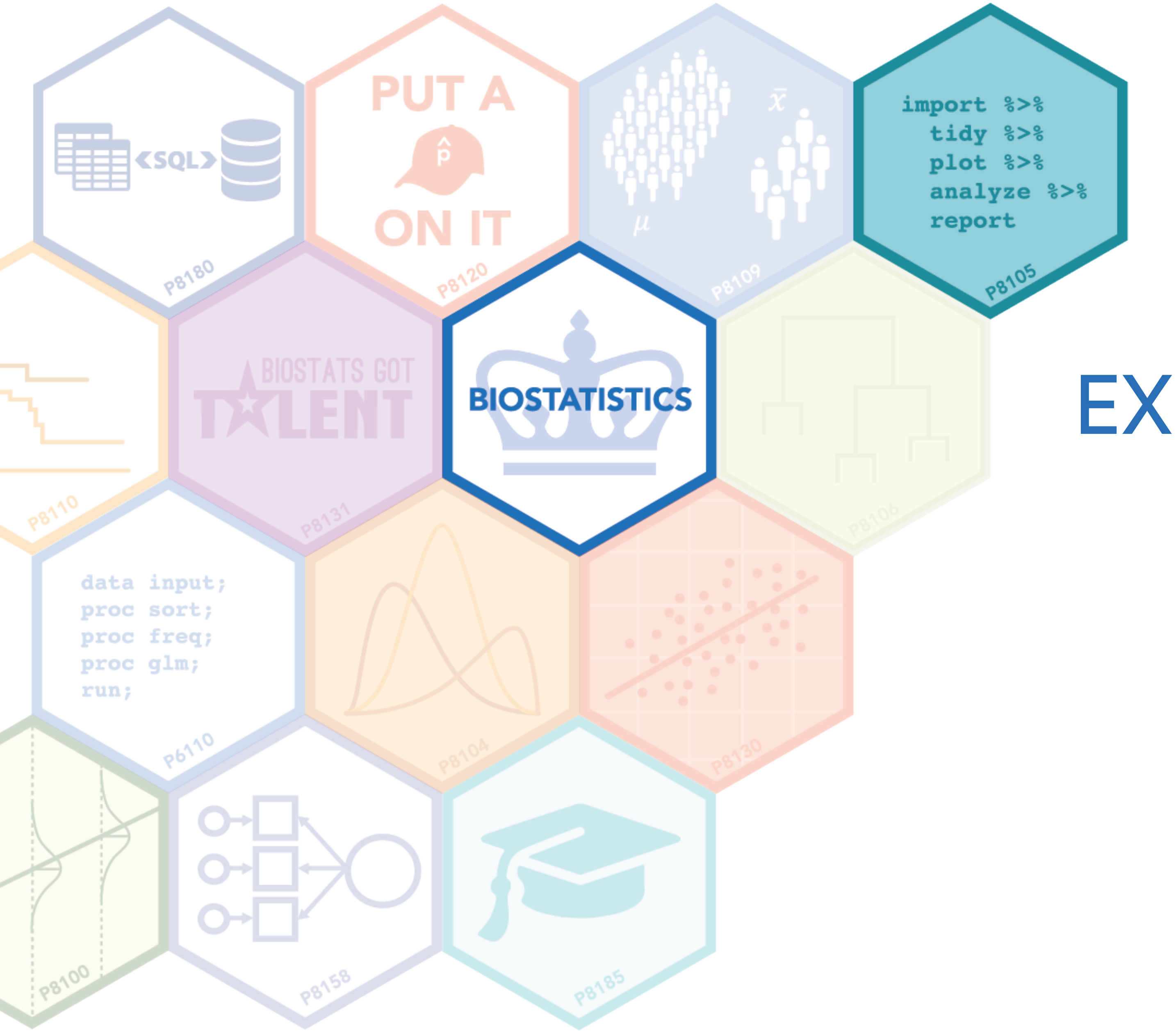

EXPLORATORY NUMERIC ANALYSIS

Jeff Goldsmith, PhD

Department of Biostatistics

Exploratory data analysis

- Exploratory analysis is a loosely-defined process
- Roughly, the stuff between loading data and formal analysis is “exploratory”
- This includes
  - Visualization
  - Checks for data completeness and reliability
  - Quantification of centrality and variability
  - Initial evaluation of hypotheses
  - Hypothesis generation

- Current emphasis is the production of numerical summaries of data, especially within groups

Grouping

- Datasets often consist of groups
  - Sometimes by design
  - Sometimes implied
  - Sometimes nested

- Examples include
  - Treatment groups
  - Age groups
  - Geographic groups
  - Family units

- These are often groups you’ve examined visually

Grouped summaries

- Quantitative comparisons across groups are informative
  - Measures center (mean, median; percent in a category)
  - Measure of variability (standard deviation, variance, IQR)
  - Amount of missingness

- These comparisons should be accompanied by robust visualizations

`group_by() + summarize()`

- `group_by()` makes grouping explicit and adds a layer to yoru data
  - Based on existing variables
  - Changes behavior of some key functions
  - Not exactly invisible, but it's easy to miss …

- `summarize()` allows you to compute one-number summaries
  - Based on existing variables
  - Most useful in conjunction with `group_by()`
  - Produces a dataframe with grouping variables and summaries
  - Easy to integrate into a pipeline

- Sometimes `group_by` and `summarize` are used to make comparisons
- Sometimes they are used to aggregate data before additional analysis

Exploratory data analysis

- A word of caution about exploratory analysis …

- Most statistical tests assume you're only concerned about the current hypothesis, or that you've done appropriate adjustments for multiple comparisons
- The validity of conclusions based on these tests depends on the process that lead you to that hypothesis
  - With any given dataset, you can form a huge number of hypotheses
  - In the end, you will only evaluate a small number of those
  - This can blur the line between "exploratory" and "formal" analysis
  - The problem is sometimes referred to as the "garden of the forking paths"

- Not a problem we'll solve in this class, but you need to be aware of it