Design the Right Experiments

- Flawed experiment design is a common cause of rejection of an IR paper (e.g., a poorly chosen baseline)
- The data should match the hypothesis
  - A general claim like "method A is better than B" would need a variety of representative data sets to prove
- The measure should match the hypothesis
  - Multiple measures are often needed (e.g., both precision and recall)
- The experiment procedure shouldn't be biased
  - Comparing A with B requires using identical procedure for both
  - Common mistake: baseline method not tuned or not tuned seriously
- Test multiple hypotheses simultaneously if possible (for the sake of efficiency)