Scaling the Teaching of Statistics and Data Science for Varied Learning Modes

Vik Gopal
National University of Singapore (vik.gopal@nus.edu.sg)

*Emerging Challenges And Opportunities
in Statistics for Higher Education and Organizations.*

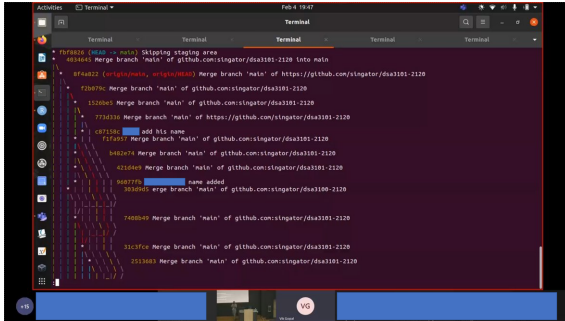This talk is based on my personal experiences teaching stats/data science.

Undergraduate/Master's classes

- Fulltime classes.
- In statistics and data science.
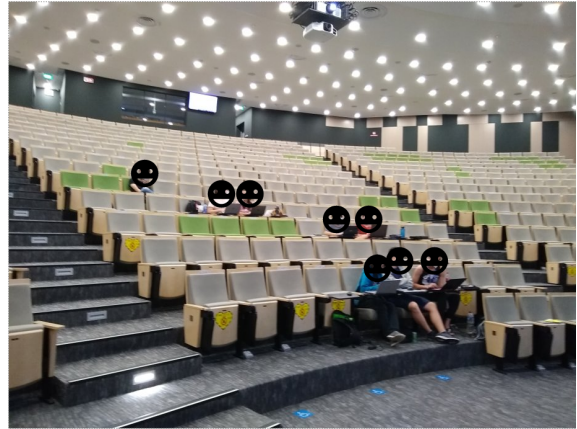
Classes for working adults

- Come back for re-training/switching careers.
- Some are traditional (13 weeks long)
- Some are "short" courses (anywhere between 3 – 8 weeks)
- Most are "blended learning" classes.

These are the various learning modes the courses are delivered in.

These are the main challenges I face.

Time

To create content.

Background of students

Classes for working adults are most challenging.

What is most useful to them?

Engaging students

Preventing collusion.

"No student left behind."

By the way, what does it actually mean to scale?

1. Scaling means we increase the number of students who take our class.
2. Scaling means we can teach *any topic to anyone with any background or interest.*
3. Scaling means we increase the number of students *while maintaining the learning experience and outcomes.*

Creating e-learning content takes a great deal of time.

- Blended learning classes require e-learning content.
    - videos
    - SCORM content
    - quizzes
    - interactive components, e.g. learnr
- It is best to have a team of instructional designers. They can turn this into this:
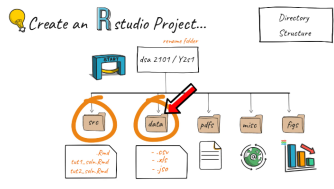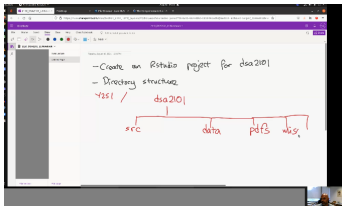
We only have one, so this is the pipeline we use for video-creation.
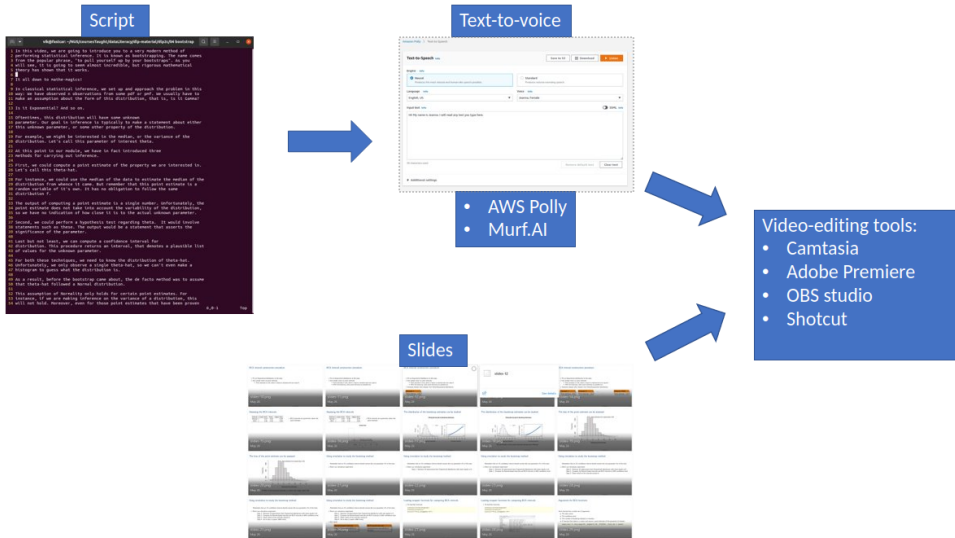
Scaling up Assessments - Quizzes and Worksheets

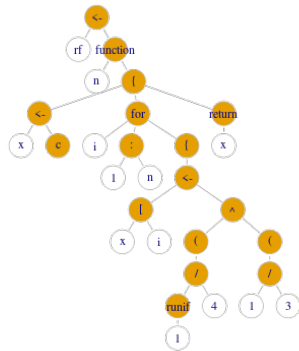- Individualising/auto-generating quiz questions is possible with
  - r/exams package by Achim Zeilis, combined with
  - your learning management's API.
- This is good for introductory classes, but not for coding classes.
- I use a semi-automatic grader:
  - autoharp

R expression as a tree

Scaling up Assessments - Take home exams (1)

Anscombe's 4 Regression data sets
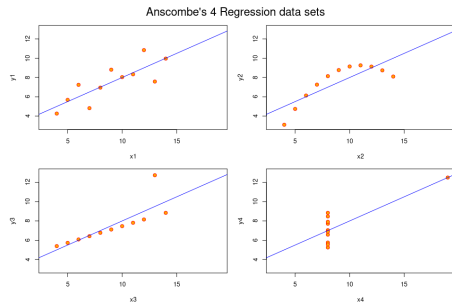
From R:

`? anscombe`

From: Alberto Cairo's blog.

Scaling up Assessments - Take home exams (2)

- Can we create "similar" datasets with "traps" in them?
  - dirtydf
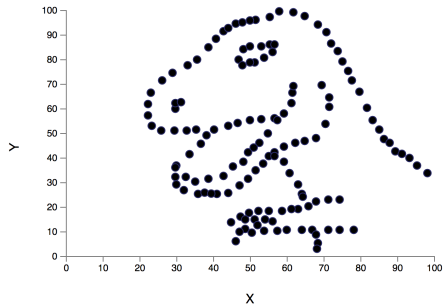 - Able to insert misspellings, missing values, create ragged dataframes.
  - Able to generate bivariate data with similar marginals but chosen correlation (see chapter 9.3 of Luc Devroye's book).
- John Tukey had an idea to identify anomalies in scatterplots, using "scagnostics". It would be nice to do that in reverse.
  - Graph-theoretic scagnostics
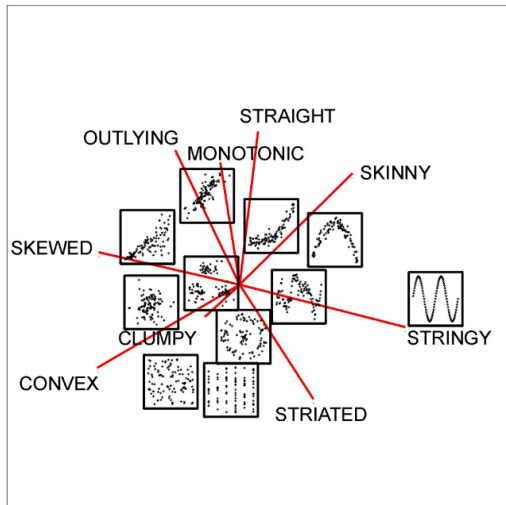
Scaling up Assessments - Projects

- Learning outcomes are not always technical.
- In projects, it is about collaboration and integration.
  - Put more in each group, and make them do more!
  - Split into front-end team, back-end team
- Cannot rely on Kaggle datasets or UCI ML repository any more.
- Need to have a sustainable stream of projects.

Project tracking

- git commits
- technical manual writing
- meeting recordings
- journals

Summary and final thoughts

- Education in our domain is already behind the industry; it is difficult to find instructors with the right qualifications. We need to be able to focus on remaining content experts, not becoming content producers.
- Scaling up to large classes requires **more** manpower; we can't do it with the same manpower, or **less**.
- For blended learning, even more skills are necessary.
- An efficient method of producing and combining e-learning material is important.
- Version control is critical.
- How to measure the effectiveness of what we are doing?