Introduction: Reproducibility & Irreproducibility

1. The Duke Scandal
2. Reproducibility & Science
3. Course Structure

Dan Ellis & Brian McFee
Dept. Electrical Engineering, Columbia University
dpwe@ee.columbia.edu brm2132@columbia.edu

1. The Duke Scandal

- **2006: Breakthrough in genomics-based personalized cancer treatment**
  - based on large-scale computational analysis
- **Independent researchers raise questions**
  - unable to duplicate analysis
- **2010: Duke review clears research**
  - based on data provided by researchers
- **2012: Lead researcher agrees data was manipulated**
  - dozens of papers retracted


biggest medical research frauds ever

2. Reproducibility

- **"The Scientific Method"**
  - empirical observation
  - hypothesis
  - tests
    - confirmation or modification
- **Confirmation requires…**
  - effective communication of findings
  - independent reproduction
- **Contemporary Computational Research**
  - "tests" involve highly complex software/hardware

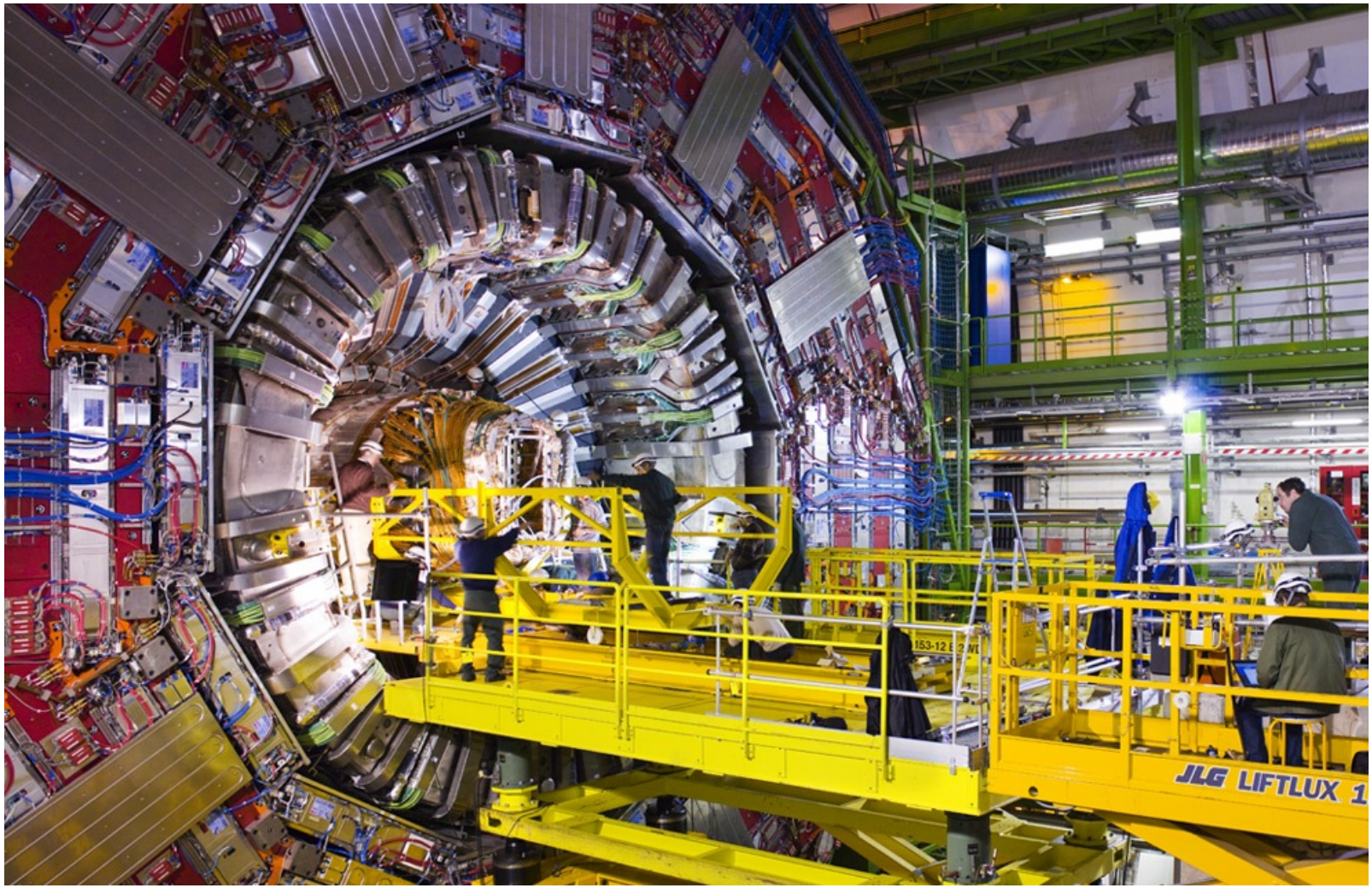
17th Century Science

- **e.g. Astronomy**
  - report observations
  - anyone can repeat
  - .. given the right equipment

Johannes Hevelius

21st Century Science

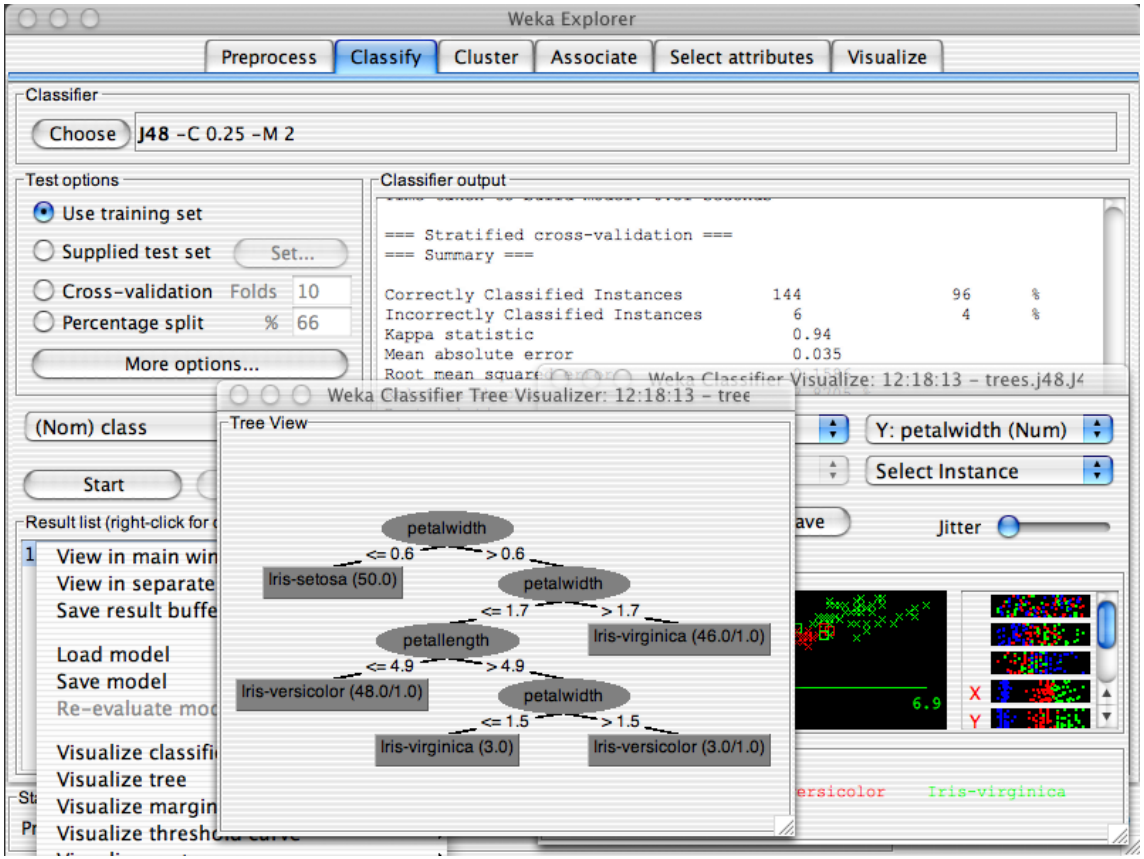
- Large-scale industrialized science



Large Hadron Collider, CERN

Computational Science

- **Software "machinery" can be very complex**
  - far beyond the scope of textual description
- **But: software is easy to duplicate**
  - your own personal LHC



Benefits of Reproduction

- "Ubiquity of error"
  - catching the things you didn't realize you got wrong
- Credibility
  - unbiased confirmation
- Identifying invalid results
  - or mistaken explanations
- Validating advances
  - by direct comparisons between different works

Costs of Reproduction

- **Costs**
  - time, resources, thinking
- **Lowest common denominator**
  - only "reproducible" results count
- **External constraints**
  - e.g. commercially-sensitive or private data
  - Google brain

3. Course Structure

- **Goals**
  - Understand the limitations of current practice
  - Understand the challenges of ideal practice
  - Learn specific tools & techniques
  - Reproduce something you want to understand
- **Methods**
  - Main project: Reproducing a paper of your choice
  - Debugging your "Reproduction package"
  - Training in tools/techniques

Project Schedule

- Feb 05:
  Initial presentations of chosen papers
- Mar 12:
  Mid-semester project updates
  Sharing of Reproduction Packages
- Apr 02:
  Feedback on Reproduction Packages
- Apr 23/30:
  Final presentations
  Final reports

Technical Tools

- Best practices
- How to make good tools
  - programming style
  - testing
  - version control
  - software analysis
  - documentation
- Evaluation Campaigns
- Presenting Statistical Results
- Open code and data distribution

Summary

- **Reproduction is important**
  - for reliable knowledge
- **Reproduction is difficult**
  - to enable
  - to perform
- **Enabling reproduction is worthwhile**
  - impact comes from people using your work
  - helps you sleep at night