ANALYSIS CHALLENGES FOR ASTRONOMY IN THE ERA OF “BIG DATA”

LUCAS MACRI

GEORGE P. & CYNTHIA WOODS MITCHELL INSTITUTE FOR FUNDAMENTAL PHYSICS & ASTRONOMY

DEPARTMENT OF PHYSICS & ASTRONOMY

TEXAS A&M UNIVERSITY
Summary

• The digital revolution has ushered a second era of “big data” in Astronomy

• Our field needs new approaches thanks to
  ◦ Exponential growth in data quantity
  ◦ Ever-increasing data quality

• “New” to us, probably not to most of you!
  ◦ Plenty of opportunities for collaborative research if you are interested!
ONCE UPON A TIME...

- 1609: Galileo Galilei “invents” the telescope
  - Improved by Kepler (1611) & Newton (1671)

- Discoveries limited by poor quality of detector
  - Low angular resolution, limited spectral sensitivity, very short integration time (1/24 second)
    - aka “Mark I eyeball system” 😊

1609
1845
HAND DRAWN BY LORD ROSSE, 1850
**Big data in Astronomy, Part 1**

- The first era of “big data” in Astronomy was ushered by the photographic plate
  - 1880s: first systematic imaging surveys
    - Hour-long exposures possible (increase depth)

- A lot of information per exposure!
  - Plates of 30×25cm @ 11μm/pix = 620 Mpix
    (~80× iPhone 6)
By the 1890s, astronomers were overwhelmed by the amount of photographic data collected.

Solution: parallelize data analysis with high-quality techniques while keeping costs low…
Let’s Get Digital

- 1969: Boyle & Smith (Bell Labs) invent CCDs
- 1974: First astronomical CCD image
  - Pros: digital, linear, low-noise
  - Cons: small (512×384 pix) & expensive

- Today: 112 Mpix CCDs with 9µm pix for $50K
  - Near-perfect detection efficiency ⇒ 100× faster
LET'S GET DIGITAL

- Digital images can be subtracted to reveal transient events (such as supernova explosions)
  - Early applications needed a lot of human vetting

Nobel Prize in Physics 2011: Discovery of “dark energy”
DRINKING FROM THE FIREHOSE...

- 2000: Start of digital survey of 35% of the sky
  - Enabled by “pizza box” camera with 30 CCDs
    - Sloan Foundation, Princeton, Fermilab & many others…

- Huge leap in quantity & uniform quality of data
  - 13 Terapixels over 15 years ⇒ 469M objects
  - Fully-automated pipeline (acquisition to analysis)
  - Not enough astronomers to classify everything…
Citizen Science

- Not enough astronomers? No problem!
  - Plenty of interested and intelligent people out there who will work for free! 😊

- Galaxy Zoo started in July 2007 [galaxyzoo.org]
  - Started with 1 million galaxies to be classified
    - Expected to take a few years…
  - 70K classifications/hr within one day of launch
  - 50M in first year, contributed by 150K people
About to be drowned…

- Large Synoptic Survey Telescope (2022-32)
  - 8-m mirror + 3.2Gpix camera with 6 “colors”
  - “A movie of 61% of the sky” (1 frame/18 sec)
    - ~10 million “alerts” per night ⇒ identify most interesting?
  - 18 Petapixels of science images (0.5 Exabytes total)
  - Final database: 15PB (37 billion unique objects)
  - Peak requirements: 1750 nodes with 1.8 PFLOPS
Texas A&M Astronomy is well poised to capitalize on the data deluge from LSST:
- Endowed Chair in Astro-Statistics
- Founding partner, 24-m Giant Magellan Telescope
  - Last night: A&M Board of Regents approved $50M share
  - First light expected ~2023
MACHINE-ASSISTED DISCOVERY

- Progress in astronomy always data driven, but we are now entering a new era of Knowledge Discovery from Data (KDD) through:
  - Correlations
    - Find patterns & dependencies that reveal new laws
  - Novelty
    - One-in-a-billion/trillion objects/events
  - Classification
    - Learn rules that constrain boundaries
  - Association
    - Unusual/improbable co-occurring events

Djorgovski et al., arXiv:astro-ph/0208246

MATERIAL COURTESY OF PROF. KIRK BORNE (GMU)
LOCAL EXAMPLE

- Take 100s of images of a galaxy over a decade
  - Dense sampling for first half, sparse for the rest
- Automated detection of variable stars (easy)
- Reliable determination of periods (hard)
- Classification into different groups (harder)
LOCAL EXAMPLE

Solution: talk to colleagues in Statistics
  ◦ Use a much better-sampled dataset of same stars in a closer galaxy as training set
  ◦ Derive classification schemes and apply to target
THANKS!

- Dr. Anne Pellerin & Mr. Wenlong Yuan
  ◦ Texas A&M Dept. of Physics & Astronomy
- Prof. Jianhua Huang, Prof. James Long & Mr. Shiyuan He
  ◦ Texas A&M Dept. of Statistics
- Prof. Kirk Borne (George Mason Univ.)