HPC IN NON-TRADITIONAL DISCIPLINES

JASON L. SIMMS, PH.D., M.P.H.

ABOUT ME

- B.A. in Classics (Latin and Greek languages)
- M.A. in Cultural Anthropology
- M.P.H. in Environmental Health (water and wastewater treatment)
- Ph.D. in Applied Anthropology (environmental anthropology)
- Graduate Certificate in GIS (spatial statistics and surface generation)

WHAT'S THE PROBLEM?

- "Standard" or "introductory" HPC use cases typically arise from STEM fields (the so-called "traditional" disciplines)
- How do we best support users in non-traditional disciplines?
  - Use cases rising in *every* discipline, including Social Sciences and Humanities
- Challenges
  - Less familiar with concepts, workflows, systems, interfaces, etc.
  - Often don't know what's even possible
  - Disciplinary acceptance varies widely!

WHAT DO WE MEAN BY "HPC"?

- *Our* answer may depend on scale
  - Especially important at smaller institutions, where we are lucky to have *maybe* 1 or 2 FTE to support the entire R/HPC ecosystem
  - At *my* institution, this includes: traditional computational resources, services (VMs, JupyterHub, RStudio Server, Omeka, Scalar, etc.), and storage and data management workflows
  - Answering this will clearly influence who is best positioned to work with users

WHAT DO WE MEAN BY "HPC"?

- *Our* answer may depend on scale
  - Especially important at smaller institutions, where we are lucky to have *maybe* 1 or 2 FTE to support the entire R/HPC ecosystem
  - At *my* institution, this includes: traditional computational resources, services (VMs, JupyterHub, RStudio Server, Omeka, Scalar, etc.), and storage and data management workflows
  - Answering this will clearly influence who is best positioned to work with users
- *Their* answer often is: anything more powerful than my desktop (or phone!)
  - And perception is reality!

IDENTIFYING NON-TRADITIONAL USERS

- Again, scale-dependent!
- Unlike with STEM, “If you build it, they won’t come.”
- So, requires evangelization, and often a good amount of (re)training
- IMPORTANT: do we care?

IDENTIFYING NON-TRADITIONAL USERS

- New faculty orientation
  - Sign-up sheet for in-person meetings
- Insert ourselves into the hiring process
- Insert ourselves into the grant process
  - Data Services Team in conjunction with Office of Sponsored Research
- Insert ourselves into the purchasing process
- Darken the doors of offices! Attend departmental meetings! Presentations!

REFRAME THE QUESTION

- Typically ask: “What are you trying to accomplish?”
  - Often incomplete! Works generally ok with STEM users, but less so with non-STEM

REFRAME THE QUESTION

- Typically ask: “What are you trying to accomplish?”
  - Often incomplete! Works generally ok with STEM users, but less so with non-STEM
- With non-traditional users, it is often necessary and helpful to reframe the question: “What would be possible…”
  - …if you can analyze 10 million rows rather than 10,000?
  - …if you can quantitatively analyze qualitative information, such as images?
  - …if you can ask your questions across ALL related corpora rather than a small subset?
  - Might even open up possibilities for grants or other funding

COMMUNICATION AND COMFORT

- Your role as an interlocutor
  - I would argue that this is *the essential* role of a research facilitator
- You do need a bit of "linguistic legerdemain" and facility with discipline-specific lingo, *and more importantly* with translating and conveying technical information back to them

COMMUNICATION AND COMFORT

- Your role as an interlocutor
  - I would argue that this is *the essential* role of a research facilitator
- You do need a bit of "linguistic legerdemain" and facility with discipline-specific lingo, *and more importantly* with translating and conveying technical information back to them
- Your role as ombudsperson
  - Making connections among people, near and far (6-step letter experiment)

COMMUNICATION AND COMFORT

- Your role as an interlocutor
  - I would argue that this is *the essential* role of a research facilitator
- You do need a bit of "linguistic legerdemain" and facility with discipline-specific lingo, *and more importantly* with translating and conveying technical information back to them
- Your role as ombudsperson
  - Making connections among people, near and far (6-step letter experiment)
- Their comfort with tools and interfaces is critical!
  - My advice: be willing to do it for or with them at first, since results will often hook them
  - Give them a VM! Useful sandbox for their own training, courses, etc.

SOCIAL SCIENCES

- All over the map!
  - (Elements of) Economics, Sociology, Geography, sometimes Psychology
  - Less so PolySci, Anthropology
- Often have challenging data issues
  - STEM data are the easy case! Often standardized formats.
  - Commonly a mix of analog and digital, and qualitative and quantitative data (field notes, images, interviews, surveys, external data sets (e.g., census, economic, etc.), and so forth)
  - Higher incidence of PII or other sensitive info in data (SSNs, minors, etc.)
  - In my experience, one result is that some are not used to strict research workflows, or may require hyper-strict workflows
- HPC versions of existing software (Stata, MATLAB, whatever)

HUMANITIES

- Historically, little need for powerful computation or analytical tools
- BUT, interesting exceptions
  - Text and image analysis, but often not yet at scale
- “Digital humanities”
  - Embraced by a range of disciplines, but especially History; often a mashup of text, images, videos, audio recordings, etc.
  - Still some resistance to quantitative research approaches, but perhaps there is space for a nascent “computational humanities”?

MIXED APPROACHES

- GIS and Spatial Analyses
  - Point pattern analyses (again, e.g., 5,000 points vs. 5,000,000) and surface generation
  - Tools beyond the desktop, such as R and Python; model builder
- Data Visualization
  - Increasingly important to represent data visually, including for lay audiences and clients
  - Often computationally-intensive, and “interestingly” challenging
- Team-teaching
  - Say, CompSci and Graphic Design, but differing levels of faculty and institutional support
- Interdisciplinary Programs
  - Data Science minor, Environmental Studies programs, etc.

CODE REUSE

- Regardless of discipline, reusing publicly-available code can reduce barriers to entry
  - R and Python for: Twitter analysis (e.g., for sentiment analysis), point-pattern analysis, text analysis
  - Springer books
- When speed and automation are demonstrated, this can also hook them
- BUT, research question often doesn't fall within the scope of the generic code (the proverbial "last mile" often requires 99% of the effort)

DATA SERVICES

- Data lifecycle as entry point
  - Data access and longevity increasingly mandated by funding agencies, and NSF and NIH regularly fund Social Science research
  - As noted, data needs are often complex, and many proposals require data management plans (DMPs), on which the Data Services Team offers consultation
  - Use data conversations as an opportunity to discuss scale and alternate analyses
- Data Associates to support courses and one-off requests for, e.g., data analysis, visualization, etc. (Data Science minor and DSDS thrust)
- Aggregating (large) datasets for teaching purposes
- Data Services Team as a way to standardize response regardless of entry point

TOOLKITS

- https://programminghistorian.org/en/lessons/
- http://cyberdh.github.io/Text-Analysis/
- http://tapor.ca/home
- https://www.datacamp.com/
  - https://www.datacamp.com/groups/education