Introducing the Whole Tale Project
Merging Science and Cyberinfrastructure Pathways

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

Bertram Ludäscher (UIUC), Kyle Chard (U Chicago), Niall Gaffney (TACC), Matt Jones (UCSB), Jarek Nabrzyski (Notre Dame), Matt Turk (UIUC)

AGU Fall Meeting
December 15, 2016
Convergence of two (ordinarily antagonistic) trends

1. Scientific projects will become massively more computing intensive and,
2. The computing will be become dramatically more transparent.

These must be addressed simultaneously:

Better transparency will allow people to run much more ambitious computational experiments.

And better computational experiment infrastructure will allow researchers to be more transparent.
Ways Forward

• An effort to develop a new infrastructure that promotes good scientific practice downstream like transparency and reproducibility.

• But we plan for people to use it not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work, and enabling efficiency and productivity, and discovery.
Problems Facing Data Researchers

Workflow for data research is **fragmented**:

Data comes from many sources and is “**integrated the old fashioned way**” (email, Excel, …)

Use cloud services **copying data** from (Drop)Box, Google-Drive, … to local storage with a distributed directory structures to organize (and provide discovery) to data

Data provenance is **not captured** (custom scripts, some version of a community developed and supported codebase)

Publication of data with link to publication (never mind DOIs, DMP) is **not sufficient for reproducibility**
Some Bold Assertions

• Software is used as a tool of discovery in nearly all research today.
• When software is a key part of the discovery process, it should be subject to the same philosophy of transparency as any method.
• Software is an integral and inseparable component of the computational infrastructure in which most research takes place.
• Computational research is embedded in a social structure which includes many stakeholders.
Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer\textsuperscript{a,1}, Jamie E. Guillory\textsuperscript{b,2}, and Jeffrey T. Hancock\textsuperscript{b,c}

\textsuperscript{a}Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of \textsuperscript{b}Communication and \textsuperscript{c}Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

whereas people in the negativity-reduced condition should express increased positivity. Emotional expression was modeled, on a per-person basis, as the percentage of words produced by that person during the experimental period that were either positive or negative. Positivity and negativity were evaluated separately given evidence that they are not simply opposite ends of the same spectrum (8, 10). Indeed, negative and positive word use scarcely correlated \([r = -0.04, t(620,587) = -38.01, P < 0.001]\).

We examined these data by comparing each emotion condition to its control. After establishing that our experimental groups did not differ in emotional expression during the week before the experiment (all \(t < 1.5; \) all \(P > 0.13\)), we examined overall posting rate via a Poisson regression, using the percent of posts omitted as a regression weight. Omitting emotional content reduced the amount of words the person subsequently produced, both when positivity was reduced \((z = -4.78, P < 0.001)\) and when negativity was reduced \((z = -7.219, P < 0.001)\). This effect occurred both when negative words were omitted (99.7% as many words were produced) and when positive words were omitted (96.7%). An

Fig. 1. Mean number of positive (Upper) and negative (Lower) emotion words (percent) generated people, by condition. Bars represent standard errors.
Whole Tale (WT) Big Picture

- WT will leverage & contribute to existing CI and tools to support the **whole science story** (= run-to-pub-cycle), and providing access to big CI & HPC for **long tail** researchers.

  ➡ **Integrate tools to simplify usage and promote best practices.**

- **NSF CC*DNI DIBBS:**
  - 5 Institutions, 5 Years ($5M total)
  - Cooperative Agreement
Specific Goals of Whole Tale

Expose existing CI
   ... through popular frontends (Jupyter, RStudio, ..)

Develop necessary “software glue”
   ... for seamless access to different CI-backend capabilities

Enhance data-to publication lifecycle
   ... by empowering scientists to create computational narratives in their usual programming environments
The Approach

WT will integrate established CI components, creating a simple, unified environment to use, share, and publish data and workflows:

1. Unified Authentication via Globus Auth
2. Abstracted Storage Layer with a unified namespace
3. Integrated Python and R APIs with Jupyter Notebook Environments
4. Ingest and publication service linking data, computations, and scholarly articles
5. NextCloud integration for “Dropbox like interface”
6. Event System to react to changes (e.g. new data published)
7. Data Dashboard to ease data management and service interactions

→ Capture full workflow via notebooks, scripts, and applications to be published along with data and research publications
Iterative Design via Working Groups

Whole Tale: Merging Science & CI Pathways ... through Working Groups!

Working Groups (Science Drivers)
- Astronomy and Astrophysics
- Earth & Env. Sciences, Archaeology
- Bioinformatics & Genomics
- Materials Sciences
- Social Sciences

Working Groups (CI Providers)
- Tools Development
- Reproducibility
- Information Science
- Education and Training

Working Groups Driving Use Cases and Adoption

Working Groups to Provide Key Components
• Ingest data from HTTP, Globus, and DataONE
• Store data in a private cloud based home directory
• Move and manage data in iRODS
• Interact with data using Jupyter

• Manage data across ownCloud & iRODS
• Authenticate using ORCID
• Interact with data through a suite of frontends
• Automatically extract key metadata

• Search and manage distributed data from within frontends
• Operate on remote data as if it were local (including using OAI-ORE)
• Utilize a single identity across services

• Discover and share frontends through global repository
• Integrate data and workflows with publications
• Issue, resolve, and track identifiers for distributed data

• Discover data using federated and distributed queries
• Track provenance across services
• Organize data collections via user-defined namespaces
So it looks pretty simple..

- What about big data?
- Complex codes?
- Reuse and bug fixes?
- Meta-analysis?
- Working with external groups, such as publishers?
- Incentives? What if they don’t come?
- Allocating resources? Sustainability models?
- What does citation mean and how are contributions to be rewarded?
Open Issues!

• “I ran all the stuff and it’s still the wrong answer!”
• “I got a different result, using your code and data!”
• “Your code doesn’t work!”
• “Where’s all the documentation? I can’t figure this thing out.”