Infusing Data with Compute: Developing and Advancing an Institution-wide Strategy around Research Data Science
Shawna Dark
Chief Academic Technology Officer and Executive Director of Research, Teaching, and Learning

Salwa Ismail
Associate University Librarian for Digital Initiatives and IT; Associate CIO

Anthony Suen
Director of Programs, Division of Computing, Data Science, and Society

Ken Lutz
Interim Director, Research IT and Co-Director Berkeley Wireless Research Center
Agenda

1. Opportunities
2. Problem
3. Key players in Landscape
4. Requirements and Concepts
5. (In-Progress) Solution(s)
The Scientific Paper is Obsolete!
Here’s what’s next...
What is Interactive Computing?
Exploring the Lorenz System

In this Notebook we explore the Lorenz system of differential equations:

\[
\begin{align*}
\dot{x} &= \sigma(y - x) \\
\dot{y} &= \rho x - y - xz \\
\dot{z} &= -\beta z + xy
\end{align*}
\]

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters \(\sigma, \beta, \rho\) are varied, including what are known as chaotic solutions. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

In [7]: interact(Lorenz, N=fixed(10), angle=(0., 360.),
\sigma=(0.5,50.0), \beta=(0.,5), \rho=(0.0,50.0))

angle: 308.2
max_time: 12
\sigma: 10
\beta: 2.6
\rho: 28
A multi-user version of the notebook designed for companies, classrooms and research labs

Pluggable authentication
Manage users and authentication with PAM, OAuth or integrate with your own directory service system.

Centralized deployment
Deploy the Jupyter Notebook to thousands of users in your organization on centralized infrastructure on- or off-site.

Container friendly
Use Docker and Kubernetes to scale your deployment, isolate user processes, and simplify software installation.

Code meets data
Deploy the Notebook next to your data to provide unified software management and data access within your organization.

Learn more about JupyterHub
Interactive Computing

Impacts...

Research

Education

Publishing
The introductory data science course at @Cal is Data 8. The course is so popular that it's in Zellerbach Hall. Fall semester 2018, Day 1.

881 7:33 AM - Oct 5, 2018

309 people are talking about this
Research

Using Python and Jupyter at NERSC
Plot a q-transform of the data

```
In [7]:
dt = 1  #-- Set width of q-transform plot, in seconds
hq = strain.q_transform(outseg=(t0-dt, t0+dt))
fig4 = hq.plot()
ax = fig4.gca()
fig4.colorbar(label="Normalised energy")
ax.grid(False)
ax.set_yscale('log')
```

Pretty interesting! Those two measurements do seem helpful for predicting whether the banknote is counterfeit or not. However, in this example you can now see that there is some overlap between the blue cluster and the gold cluster. This indicates that there will be some images where it's hard to tell whether the
Transdisciplinary transformation underway

Data Science

Bioinformatics

Machine Learning

Digital Humanities

Precision Agriculture

Computational Neuroscience
Problem

Demand for interactive computing is exploding!
Problem

Demand for interactive computing is exploding!

How will we supply interactive computing for the entire campus?
Creating Campus Strategy

- Campus has now made the leap to articulating an institution-wide strategy around data science.

- Various campus units have seen an infusion of new knowledge and participants, discussion forums, communities of practice, and other opportunities to learn from each other.

- Many of our units have new leadership, and change is being spurred by the shared purpose of Reinventing IT (One IT).
Values Campus Data & Compute

across domains  collaborative vision  scalability
reimagining of the research and instructional support experience on campus.
- campus services have had to pivot to support programs not just for instructors, but overlapping needs for researchers
- we also want to continue to support opportunities and areas for innovation
Data Services and Support Landscape - Division of Computing, Data Science and Society

bCourses

DATA SCIENCE CONNECTOR COURSES
Companion Courses for Data 8

LEGALST 88
Taking Measure of the Justice System

STAT 88
Probability & Mathematical Statistics

COMPSCI 88
Computational Structures in DS

COGSCI 88
Data Science & the Mind

PSYCH 88
DS for Cognitive Neuroscience

UGBA 96-4 & 5
Data & Decisions

STAT 89A
Linear Algebra for Data Science

GEOG 88
Data Science Applications in Geography

L&S 88-1
Children in the Developing World

L&S 88-2
Sports Analytics

MCB 88
Immunotherapy of Cancer
So we’re pivoting too --

- We’ve always been the repository for Data Access and Data Deposit -- buying datasets and publishing data.
- The demands while the same have pivoted -- can we provide normalized data? Can we published data that can be normalized?
- What will our data repository be? Published data vs. purchased data? and storage -- part of research data lake with campus?
- Other entities on campus buying data
- Dryad - through CDL play a role (partnership with Zenodo (DJ D-Zed: Mixing Up Repositories)
- Outreach and consulting - RDM partnerships (shared positions)
Data Services and Support: Research Data Management (RDM)

RDM must adapt to the opportunities and challenges presented by Interactive Computing

- Research Data management will change: Individual researchers working with and managing their data sets => Researchers drawing on shared, dynamic data sets that may be ubiquitously available, supporting extensive collaboration

- Just as Jupyter notebooks can be shared so to can the data that underlies the notebook

- Verifying and maintaining Data provenance and integrity will become part of the data management process and lifecycle
It used to be that:

- A data set was tightly associated with a particular computer, data and compute were location bound.
  - Where is the computer (IP Address) that has the data I want to compute on?
  - Go to the Data/Compute or...
  - Explicitly move the data to my computer and then run
    - Data has a location
    - Computers have state
  - Researchers worked in labs with their peers and only their peers. But..... the world is changing
  - becoming much more dynamic and collaborative
The Pivot;  
Factor Data and Computation

- Interactive computing will map to onto a portfolio of computational facilities
  - High Performance Computing Clusters
  - Secure computation for sensitive data
  - Virtual Machine environments
    - Computers are *stateless* and fungible
- Data sets, ubiquitously available map in a similar, agile manner to support Interactive Computing
  - Data becomes *locationless* and readily available
Experimenting with coming together as one

Convenors:
Division/DSEP, Research IT, Library (Digital Strategies)

Participating:
IST, ETS, RTL, EDW, CoE, D-Lab/BIDS, VCRO, IPIRA, Academic Senate CIT chair
Vision for Campus Data & Compute

Universal interactive computing

Hybrid scalable computing infrastructure

Integrated consulting services
Thank you!
Questions?