Applied Data Science:

Driving the Moonshot of Machine Learning, Deep Learning and Artificial Intelligence

Roger H. French

SDLE Research Center
Materials Science & Engineering Department
Case Western Reserve University, Cleveland OH 44106 USA

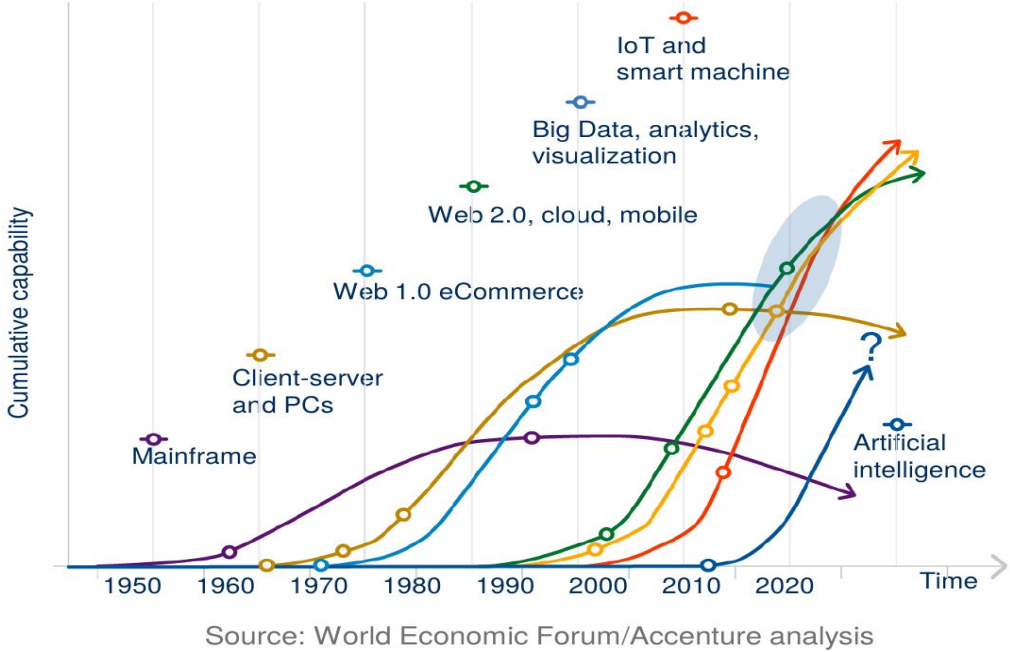
http://sdle.case.edu

Digital Transformation: Combinatorial Effects of Tech. are Accelerating Change

The falling cost of advanced technologies

- A defining characteristic of digital revolution
  - Computing
  - Internet Communications
  - Data Storage

Major role in driving digital transformation



Source: World Economic Forum/Accenture analysis

Examples of the falling cost

- of key technologies

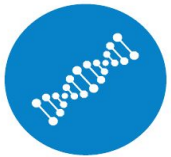
Drones

Cost per unit

2007 **$100,000**

2013 **$700**

DNA Sequencing



Cost per unit

2000 **$2.7 billion**

2007 **$10 million**

2014 **$1,000**

Solar

Cost per kWh*

1984 **$30**

2014 **$0.16**

Industry 4.0: The 4th industrial Revolution

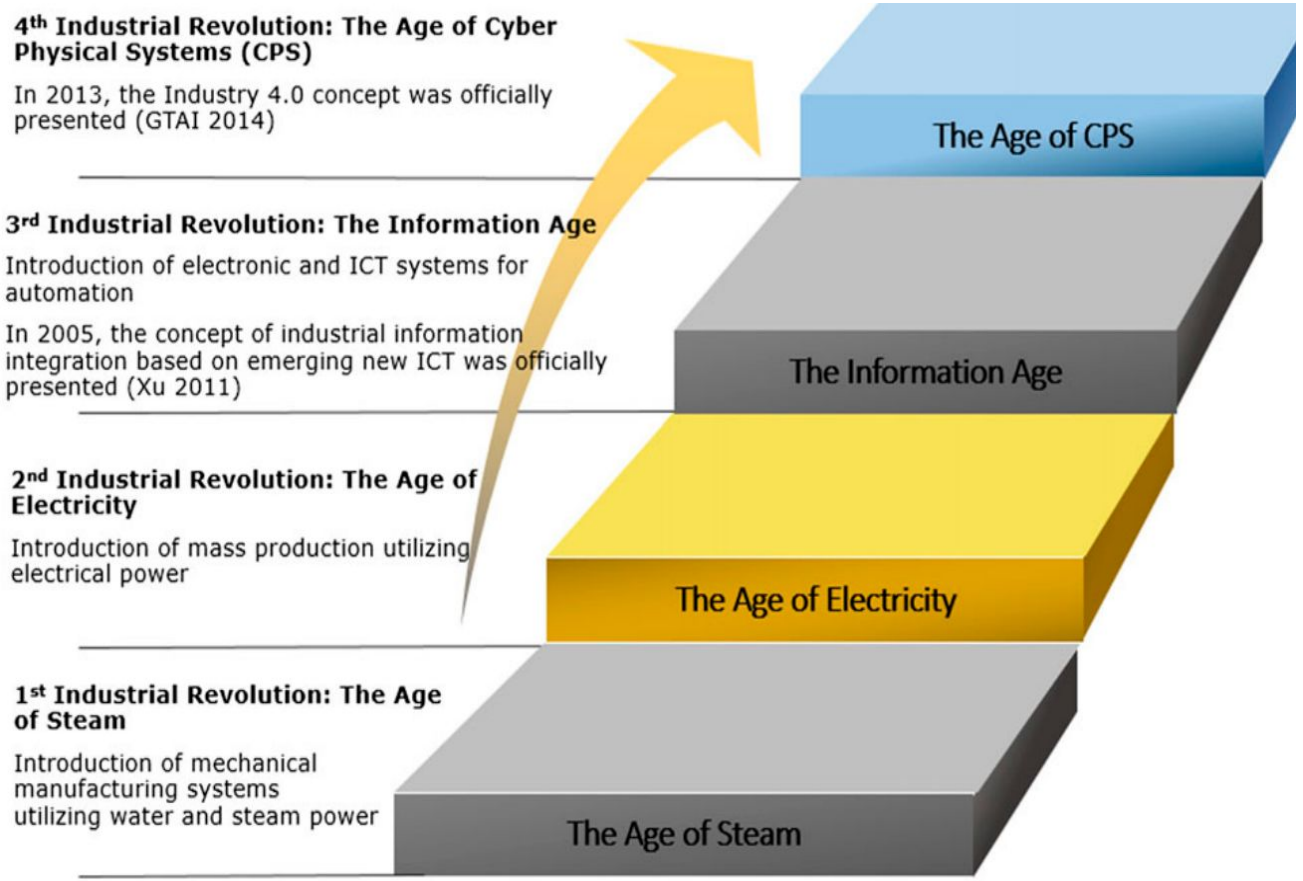
Digital Transformation

- Is transforming Industry
- Into Industry 4.0

Digital Technologies

- Will provide new flexibility
- And Industrial efficiency

Opening new opportunities



Y. Lu, Industry 4.0: A survey on technologies, applications and open research issues, Journal of Industrial Information Integration. 6 (2017) 1–10. https://doi.org/10.1016/j.jii.2017.04.005 .

CWRU Applied Data Science UG/Grad Program

THROUGH THE COLLABORATION of its business and higher education members, the Business-Higher Education Forum (BHEF) launched the National Higher Education and Workforce Initiative (HEWI) to create new undergraduate pathways in high-skill, high-demand fields such as data science and analytics. Data science and analytics must be integrated with T-shaped skills, such as critical thinking, collaboration, and effective communication, which are critical for all graduates entering the 21st century workforce. Knowledge of data science and analytics in recent years has become as fundamental as any other skill for graduates' career readiness. BHEF's Strategic Business Engagement Model with higher education addresses this demand by moving the two sectors from transactional relationships to strategic partnerships through five strategies:

1. **ENGAGE** corporate leadership;
2. **FOCUS** corporate philanthropy on undergraduate education;
3. **IDENTIFY** and tap core competencies and expertise;
4. **FACILITATE** and encourage employee, faculty, and staff engagement;
5. **EXPAND** the focus of funded research to include undergraduate education.

This case study examines how BHEF member Case Western Reserve University (Case Western Reserve) is integrating T-shaped skills into a minor in applied data science.

PROGRAM OVERVIEW

THE APPLIED DATA SCIENCE (ADS) MINOR AT CASE WESTERN RESERVE serves as a national model for undergraduate education in data science. Available to every undergraduate student across all schools at the university, this program of study requires experiential learning opportunities, embeds T-shaped skills, and allows students to master fundamental ADS concepts in their chosen domain area. From strong leadership engagement to funded undergraduate research opportunities, Case Western Reserve applied BHEF's Strategic Business Engagement Model to create a minor that responds to the fundamental need for data science in today's global business community.

Medical Mutual of Ohio
Medtronic
Philips Healthcare
Sherwin-Williams Company
Siemens
Teradata Corporation
Timken Company
University Hospitals

http://www.bhef.com/publications/creating-minor-applied-data-science

CWRU Applied Data Science UG/Grad Program

A BHEF CASE STUDY

CREATING A MINOR IN

APPLIED DATA SCIENCE

Case Western Reserve University Engages B
Leaders to Produce T-Shaped Professionals

The New York Times | https://nyti.ms/2sSkAVI

With Innovation, Colleges Fill the Skills Gap

By JOHN HANC JUNE 7, 2017

How large is the so-called skills gap?

The Manpower Group, a human resources consulting firm, says the gap, which is often defined as the difference in job skills required and the actual skills possessed by employees, is a chasm. Of the more than 42,000 employers the firm surveyed last year, 40 percent said they were having difficulties filling roles, the highest level since 2007.

Case Western Reserve University

Creating 15- or 18-credit minors may be one of the more effective strategies for preparing students to enter high-demand fields. Because a minor requires fewer credits than a major and few, if any, prerequisites, these allow colleges to be more flexible and responsive to changing industries and emerging technologies.

Case Western's minor in applied data science, for example, funnels students into this hot field from other disciplines. The students learn skills like data management, distributed computing, informatics and statistical analytics.

OVERVIEW

) MINOR AT CASE
ional model for
ience. Available to
all schools at the
quires experiential
haped skills, and
ntal ADS concepts
strong leadership
ate research
ve applied BHEF's
del to create a
ental need for data
ommunity.

l Mutual of Ohio
nic
Healthcare
n-Williams
ny
ıs
a Corporation
Company
ity Hospitals

CWRU's Applied Data Science Program: Undergraduate Minor, Graduate Certificate

Applied Data Science program

- Undergraduate Minor
- Graduate Data Science Certificate

Developed in 2014, Courses Started in 2015

- In collaboration with Business Higher Ed. Forum
- UG & Grad Course Section
  - DSCI351/451
- Now have Materials Data Science M sections
  - DSCI351M

Applied/Materials Data Science

- DSCI351M: Exploratory Data Analysis
- DSCI353M: Modeling, Prediction, Machine Learning
- DSCI352M: Materials Data Science Res. Proj., POSEV
- DSCI354: Data Visualization and Analytics

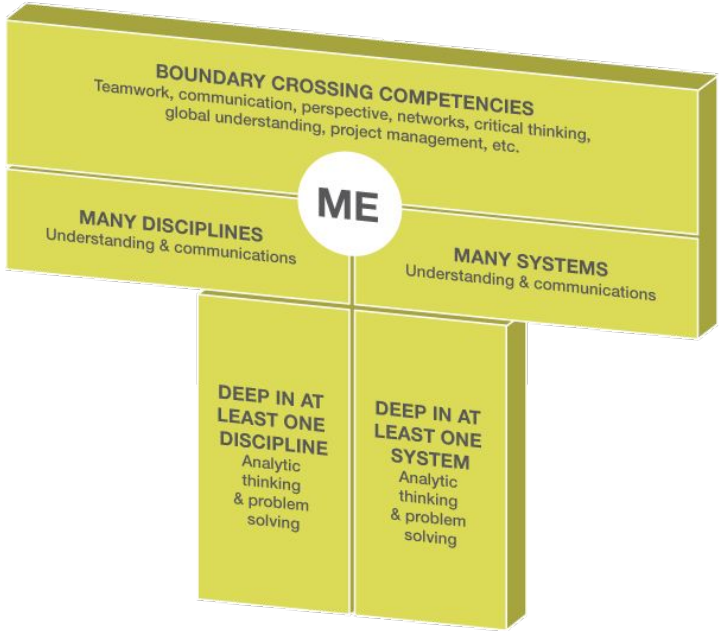
Specialization Courses

- DSCI354M: Data Visualization & Analytics
- DSCI430, Cognition and Computation
- DSCI432, Spatial Statistics for Subsurface Modeling

Now 70 to 80 students per semester

Developing T-shaped <u>Undergraduates</u>

- Deep domain knowledge (in Materials Science)
- Broad knowledge in Applied Data Science



D. Hughes, R. H. French. Crafting a Minor to Produce T-Shaped Graduates. T-Summit 2016, Washington DC, March 21, (2016). at http://tsummit.org/

Components of Applied Data Science Curriculum

Applied/Materials Data Science Core Courses

- DSCI351M: Exploratory Data Analysis
- DSCI353M: Modeling, Prediction, Machine Learning
- DSCI352M: Materials Data Science Res. Project
  - For their GitHub "Portfolio"
- DSCI 354: Data Visualization & Analytics

POSEV Concepts

- Privacy, Openness, Security, Ethics, Value

Taught from "Structure of a Data Analysis" Perspective

Agile Software Development Tools & Approach

Knuth's Literate Programming Perspective

- Integrate Code and Report Writing
- Rmarkdown, Jupyter Notebooks

Textbooks (Open Access)

- Open Intro Statistics
- Introduction to Statistical Learning with R, 2nd Edition

Taught using a Practicum Approach

Each class has two parts

- **Foundation:**
  - Statistics, Regression, ML, Time series ...
- **Practicum**
  - Code Style and commenting
  - Pipelines and Pipe operators
  - Data structures and data frames

Coding/Programming Language

- R with Rstudio IDE
- Python with Spyder (or Jupyter notebooks)

Open Data Science Toolchain

- (cross platform: Linux, Mac, Win)
- R, Python
- Rstudio, Spyder
- Markdown, Rmarkdown
  - Jupyter Notebooks for Python, R
- LaTeX engine, TexStudio
- Chrome, Firefox, html

“Structure of a Data Analysis” Perspective, For SemProj’s & Class Practicum

Part a) Define Question

- Background on the research area & critical issues
- Define the question
- Define the ideal data set
- Determine what data you can access
- Define critical capabilities, identify packages you will draw upon
- Obtain the data, define you target data structure
- Clean and tidy the data

Part b) Cleaning and Exploratory Data Analysis (EDA)

- Write your databook, defining variables, units and data structures
- Data visualization and exploratory data analysis
- Observations of trends and functional forms
- Power transformations
- Validate with reference to domain knowledge
- Evaluate the types of Modeling Approaches to take

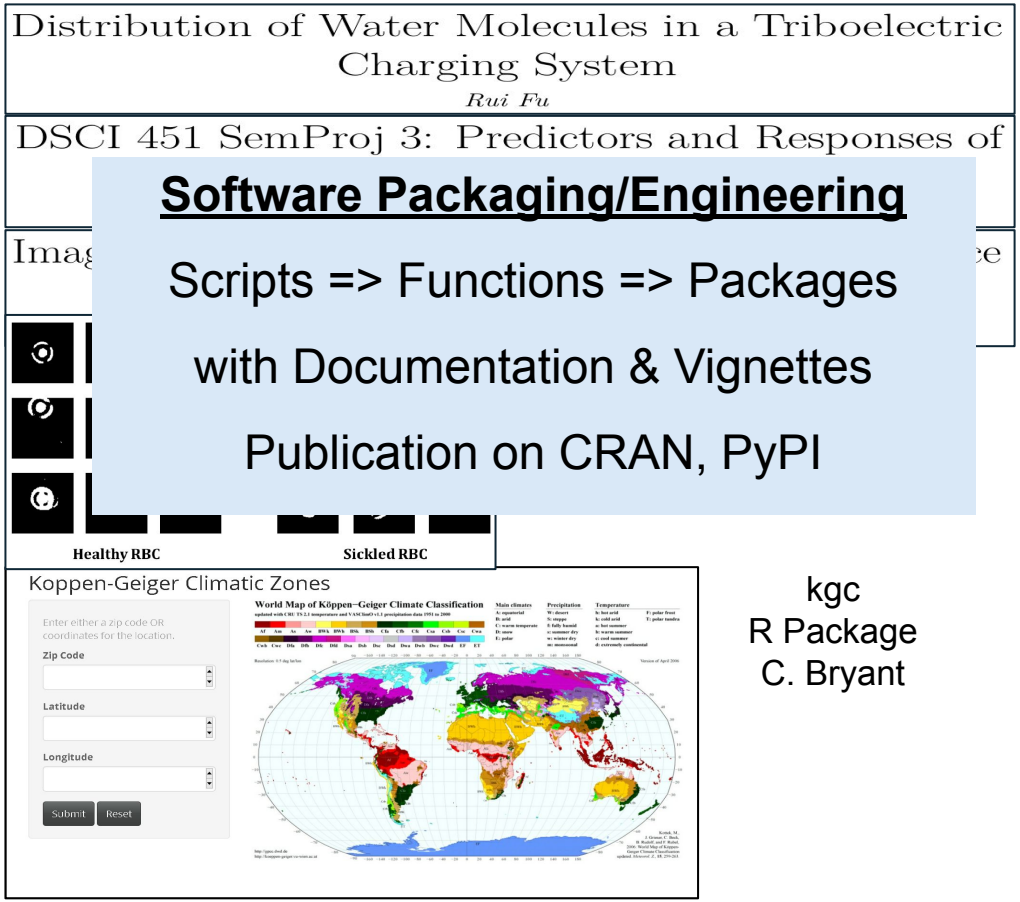
Part c) Modeling, Prediction, Machine Learning

- Types of modeling to try
- Statistical prediction/modeling
- Model selection
- Cross-validation, Predictive R2
- Interpret results
- Challenge results

Part d) Present Your Final Models and Learnings

- Present your results
- Present reproducible code
- Comparison to literature modeling approaches

Jeff Leek, JHU, Data Analytic Style



kgc
R Package
C. Bryant

Open Data Science Tool Chain

Using Open Source, Agile Tools

- Manifesto for Agile Software Development

Reproducible Research

- Using Rmarkdown reports
- Python/R Jupyter Notebooks
- When data updates
- Recompile your report
- All new figures and report!
- Well Documented Codes & Reports

High Level Scripting Languages: R, Python

- Use Machine Learning Frameworks
- Such as Keras/TensorFlow for Deep Neural Networks

Rstudio Integrated Development Environment

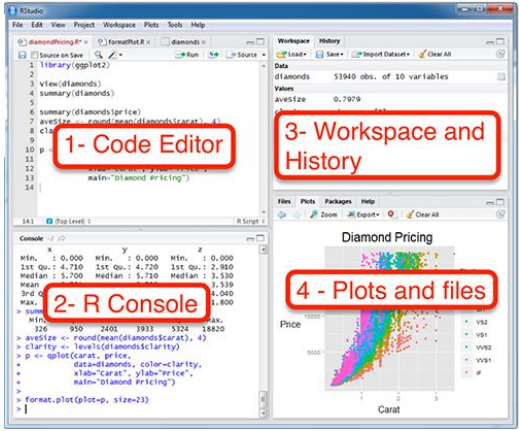
- Spyder IDE for Python

Git Repositories for Code Version Control

- Share code scripts with colleagues
- Share project data and reports with others

Github, BitBucket, GitLab for Collaboration

- Website hosting your Code Repositories



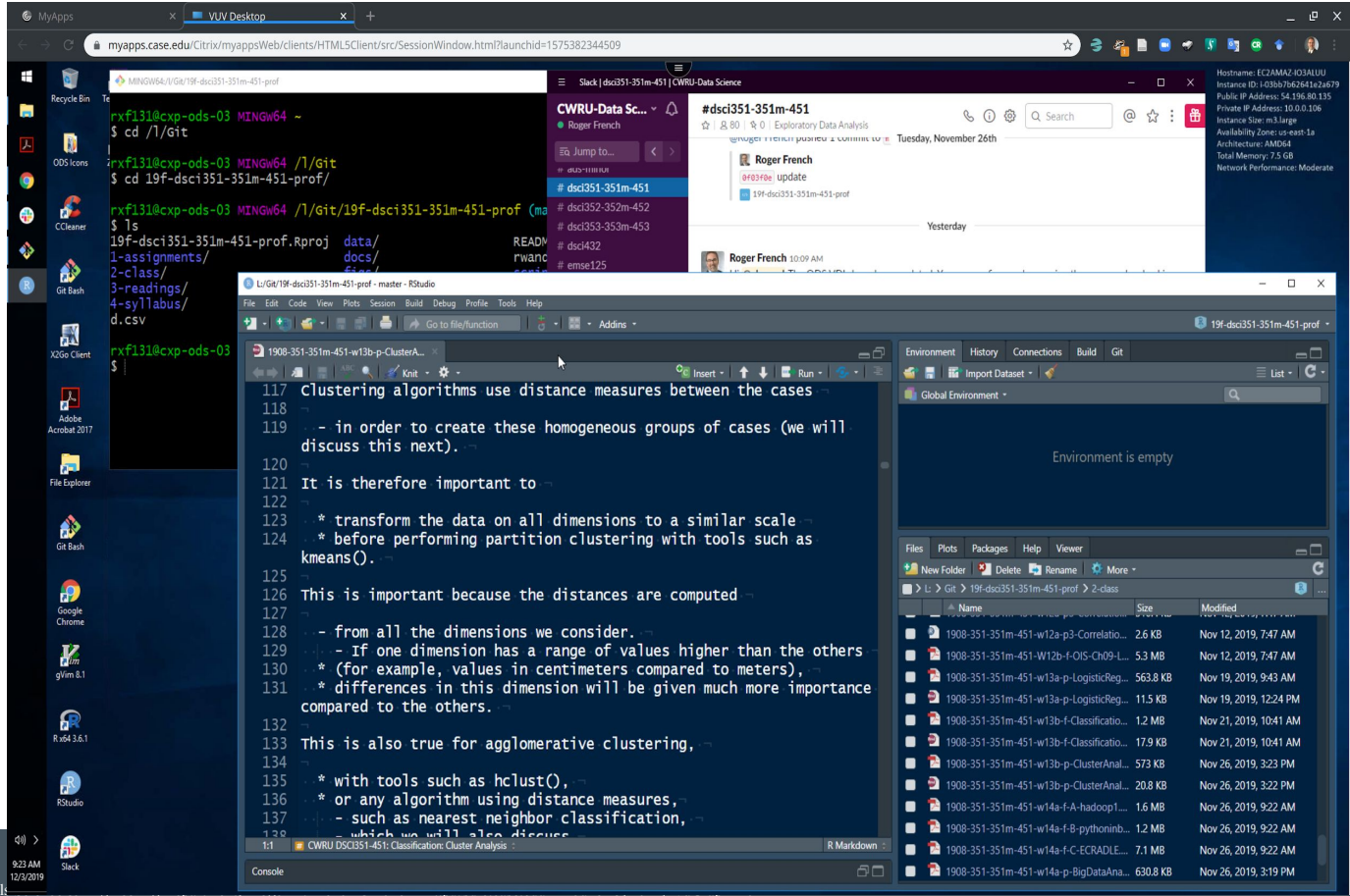
Compute Infrastructure for the ADS program

Provide students Open Data Science Computers

- Win10 Cloud Computers (Citrix)
- Hosted by CWRU
  - Scalable,
  - Good Performance
- With R, with Rstudio IDE
- Python3, with PyCharm IDE
  - “standard” R packages
  - “standard” Python packages
- Git, (git bash)
- Pandoc, LaTeX, html
- Slack
- StackExchange

Standard ODS Env.

- No time lost fixing computers
- Full install instruc. provided



CWRU Markov Data Science Cluster: Hosted by [U]Tech Res. Computing

Markov Total = 1120 CPU cores, 174k GPU cores

- 28 nodes: 2 Xeon CPUs (20 cores/CPU).
- 20 nodes: 2 Xeon CPUs with 2 Nvidia RTX2080Ti
- 1120 CPU cores and 174K GPU cores.

Enables Batch & Interactive GUI sessions

- Using either compute or GPU nodes.

Running R/Rstudio and Python3/PyCharm

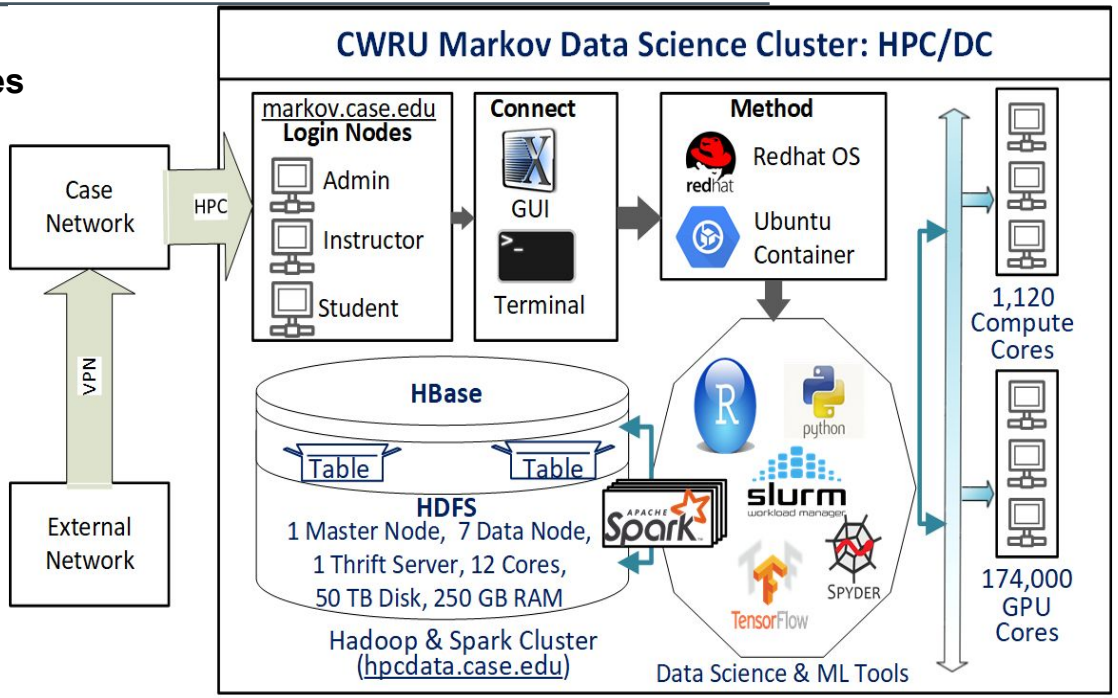
- Along with Keras/TensorFlow2/Cuda/CDNN

Implemented using the Open Data Science (ODS) Ubuntu 20.04 Container

- Based on Singularity.

Markov's Hadoop cluster: hpcdata.case.edu

- Loaded with publically available datasets



CWRU Markov Data Science Cluster: Hosted by [U]Tech Res. Computing

Markov Total = 1120 CPU cores, 174k GPU cores

- 28 nodes: 2 Xeon CPUs (20 cores/CPU).
- 20 nodes: 2 Xeon CPUs with 2 Nvidia RTX2080Ti
- 1120 CPU cores and 174K GPU cores.

Enables Batch & Interactive GUI sessions

- Using either compute or GPU nodes.

Running R/Rstudio and Python3/Spyder,

- Along with Keras/TensorFlow2/Cuda/CDNN

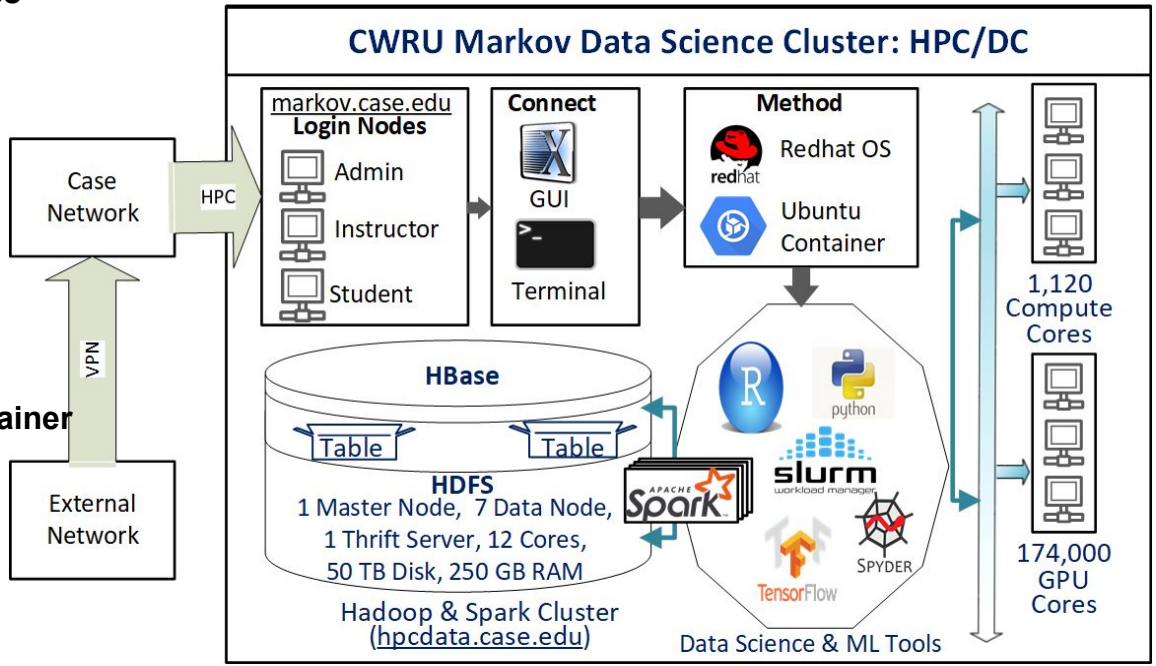
Implemented using the

Open Data Science (ODS) Kubuntu Container

- based on Singularity.

Markov's Hadoop cluster: hpcdata.

- Loaded with publically available datasets



Teaching with Git & The Tools of Agile Software Development

Coursework distributed using Git Repository

- Fork the “Prof” repo
- Students tend their personal repo

Coherent Repo Structure

- Codes using relative pathing
- So codes work cross-platform

Git Sync and Pull

- For each class

Git Add, Commit, Push

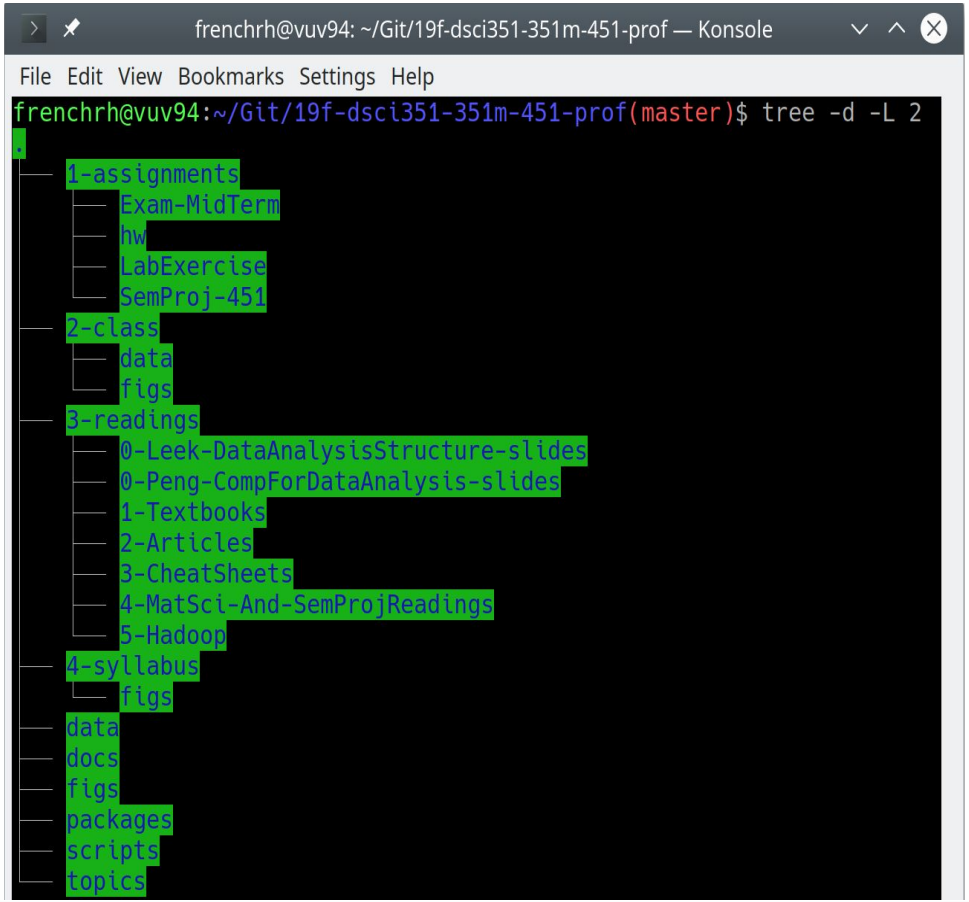
- Students own work

Class Notes in Rmarkdown

- Compiled to pdf
- With Pandoc & LaTeX

Assignments

- Traditional homeworks
- Lab Exercises: Two week assignments



Applied Data Science Research:

In the SDLE Research Center & MDS-Rely Center

Roger H. French

SDLE Research Center
Materials Science & Engineering Department
Case Western Reserve University, Cleveland OH 44106 USA

http://sdle.case.edu

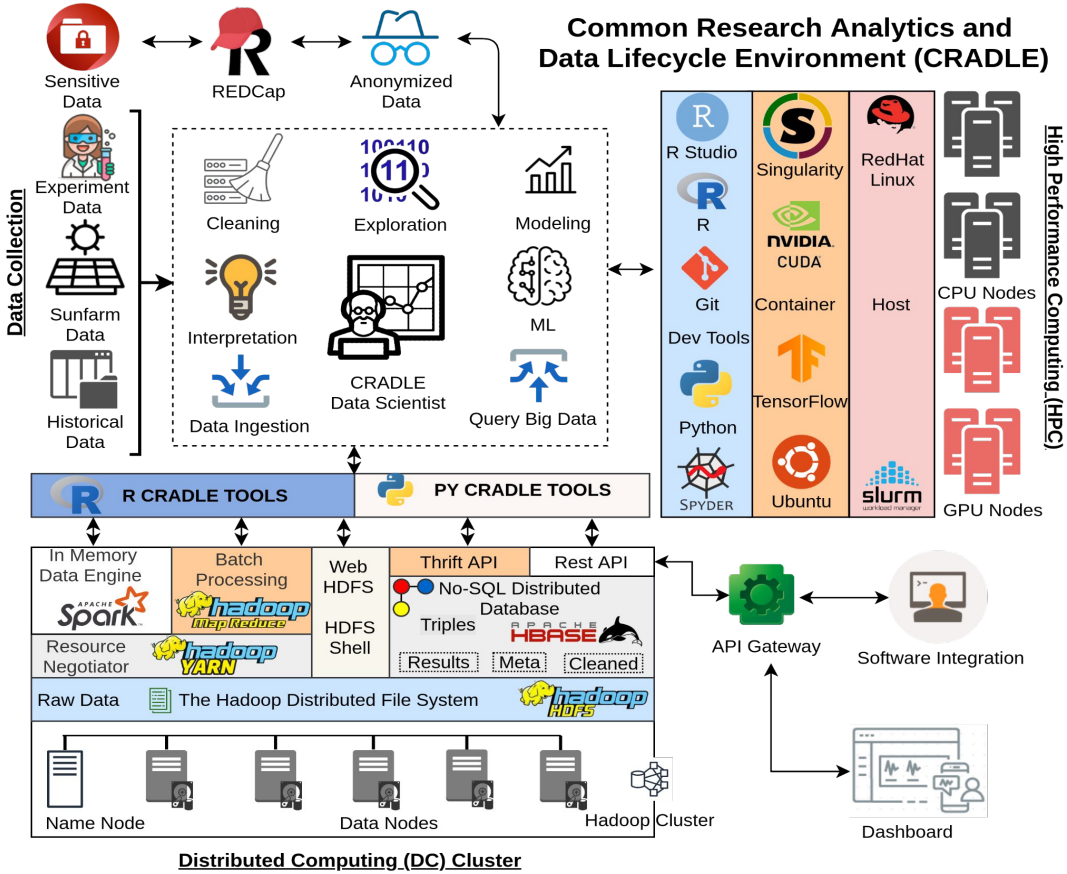
CRADLE Analytics Environment

CRADLE 2.2

- 192 Cores
- 1 Tb RAM
- 75 Tb Storage
- CDH 5.13.0

CRADLE 2.3

- 128 Cores
- 0.5 Tb RAM
- 100 Tb Storage
- CDH 5.16.2
- NSF SP-800-171
- For DOD CUI (Controlled Unclassified Information)



Distributed Computing vs. High Performance Computing

Distributed Computing

- Lots of data: Gigabyte computing
  - e.g. Facebook collecting all user data
  - Google, Amazon, etc
- 1st example: Google indexing the web
  - S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google File System," in Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, New York, NY, USA, 2003, pp. 29–43, doi: 10.1145/945445.945450
  - F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," ACM Transactions on Computer Systems (TOCS), vol. 26, no. 2, p. 4, 2008.
  - J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Commun. ACM, vol. 51, no. 1, pp. 107–113, Jan. 2008, doi: 10.1145/1327452.1327492.

This transformed computer science

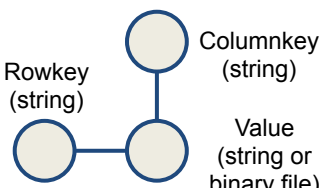
High Performance Computing

- Lots of Flops: Gigaflop computing
- Usually small startup file
- Lots of compute operations
- A small output file

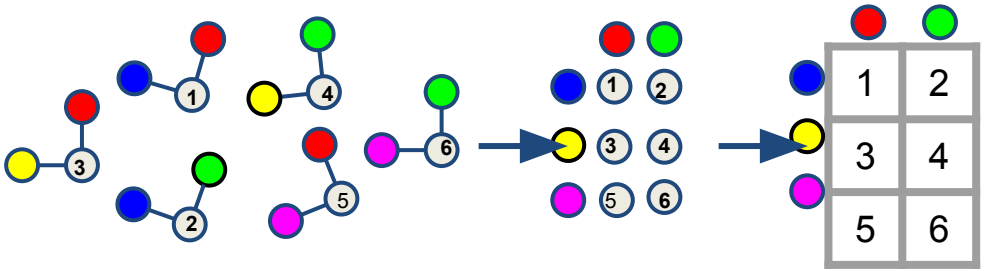
This is HPC Compute

- or HPC Cloud Computing

NoSQL DB Abstraction of Hadoop/Hbase



HBase RDF 'Triple'



Combines Lab data (Spectra, Images etc.) With Time-series Data (PV Power Plant Data)

High Performance PV Data Analytics: Petabyte Data Warehouse In A Petaflop HPC Environment

- In-place Analytics: Distributed Spark Analytics in Hadoop/HDFS
- In-memory Data Extraction: To Separate HPC Compute Nodes

IEEE JPV

A non-relational data warehouse for the analysis of field and laboratory data from multiple heterogeneous photovoltaic test sites

Yang Hu, *Member, IEEE,* Venkat Yashwanth Gunapati, Pei Zhao, Devin Gordon, Nicholas R. Wheeler, Mohammad A. Hossain, *Member, IEEE,* Timothy J. Peshek, *Member, IEEE,* Laura S. Bruckman, Guo-Qiang Zhang, *Member, IEEE,* and Roger H. French, *Member, IEEE*

Hu, Y., V. Y. Gunapati, P. Zhao, D. Gordon, N. R. Wheeler, M. A. Hossain, T. J. Peshek, L. S. Bruckman, G. Q. Zhang, R. H. French. "A Nonrelational Data Warehouse for the Analysis of Field and Laboratory Data From Multiple Heterogeneous Photovoltaic Test Sites." IEEE Journal of Photovoltaics 7, no. 1 (January 2017): 230–36. doi:10.1109/JPHOTOV.2016.2626919.

FAIRification of Datasets and Models, Enables AI learning

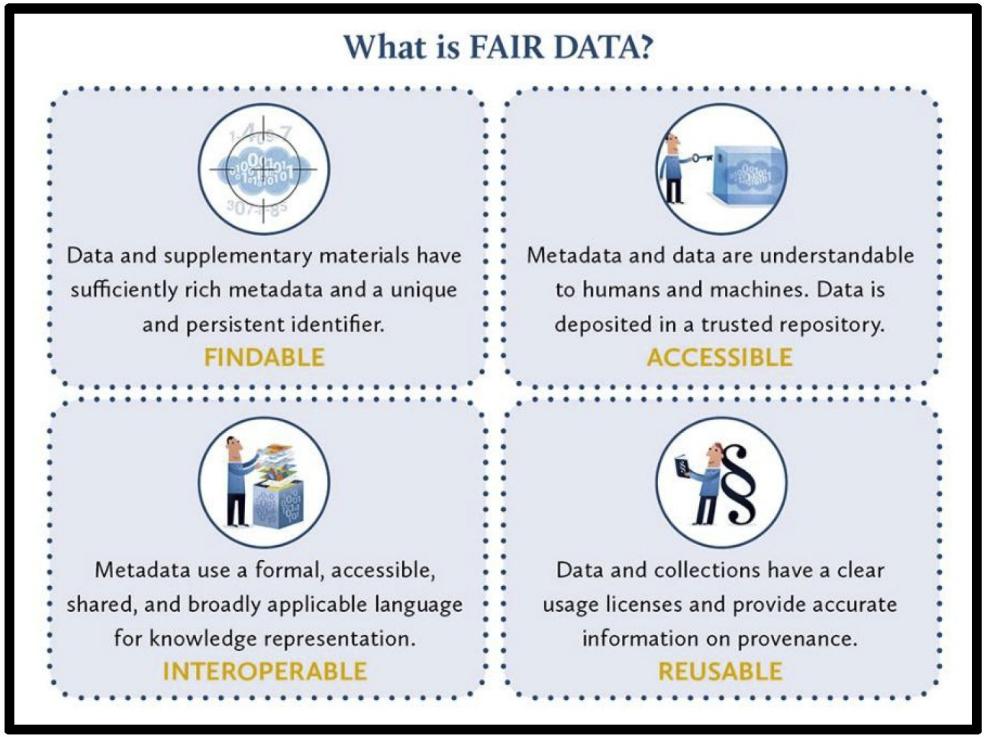
Making Datasets & Models FAIR
- By “FAIRification”

Enables Models to find Data
- And Data to find Models

So that they can advance
- Without human intervention

This is an aspect of the Semantic Web
- And Resource Description Framework
- Hbase triples are an example of RDF

We just received a DOE SETO AI award
- For st-GNN, that includes FAIRification



M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Sci. Data, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.1038/sdata.2016.18.
https://en.wikipedia.org/wiki/Resource_Description_Framework

Open Source, Open Data, Reproducible Research Tools For Science

Using Open Source tools

- R & Python coding
- Git code versioning & collaboration
- Cross-Platform (Linux, Mac, Windows)
- LaTeX & Markdown

Reproducible Research

- Distribute Code & Datasets
- At time of paper publication
- Your research can be reproduced by others
- Others can build on your research and data

Use Agile Development Tools

- Slack team messaging
- Jira Cloud Issue Tracking
- BitBucket/GitHub/GitLab

Build Packages for Science

Use Package-based systems

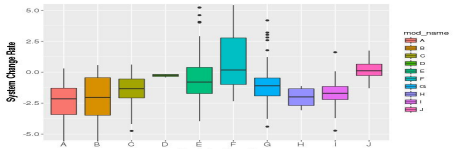
- Rely on well-vetted Open Source Codes

R Packages

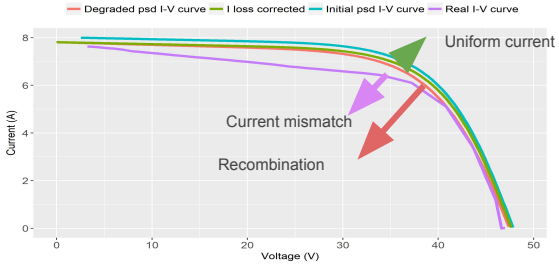
- Well vetted, with know package dependencies
- With Vignettes on Theory, & Use
- With Data Sets and Results for Validation

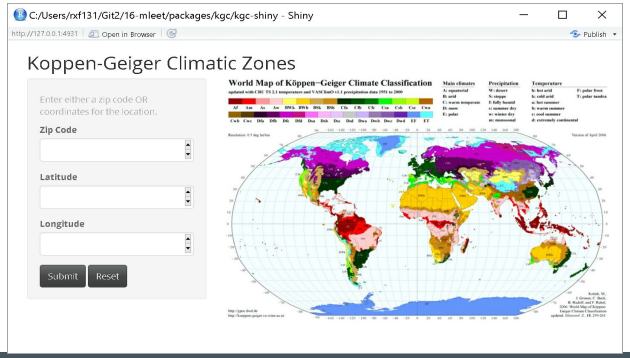
Performance Loss Rate Determination

IEA PVPS Task 13 PV Reliability



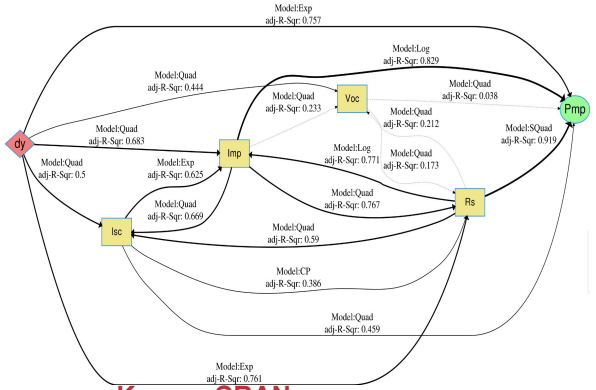
Suns-V_{oc} from Time-Series I-V





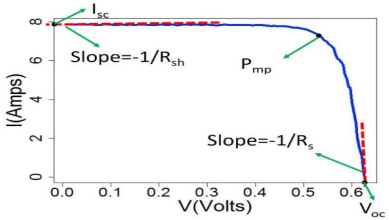
NetSEM on CRAN

Network Structural Equation Modeling



Kgc on CRAN

Köppen-Geiger Climate Zone Package



ddiv on CRAN

Data-driven I-V Feature Extraction