Cyberinfrastructure shapes scientific outcomes in crucial and largely unrecognized ways

Victoria Stodden
School of Information Sciences & National Center for Supercomputing Applications
University of Illinois at Urbana Champaign

Opportunities for Accelerating Scientific Discovery:
Realizing the Potential of Advanced and Automated Workflows
National Academies of Science, Engineering, and Medicine Workshop
March 16, 2020 - Washington DC

This material is based upon work supported by the National Science Foundation under Grant No. 1941443. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Cyberinfrastructure (CI) is a broad term

CI: Computational/digital resources deployed to assist scientific discovery pipelines:

- compute resources (e.g. cloud, Jetstream, Frontera, ...),
- repositories (e.g. Dataverse, zenodo, Dryad, MLOSS, ...),
- scientific and statistical analysis software (e.g. python, R, ...),
- middleware (e.g. jupyter notebooks, containers, ...),
- scalable database provisioning (e.g. MongoDB, SQL, ...),
- scientific workflow software (e.g Pegasus, Galaxy, MyExperiment, ...),
- collaboration tools and platforms (e.g. colab, kaggle, Agave, ...),
- publishing and curation software (e.g. Whole Tale, reana, ...),
- many many other examples..

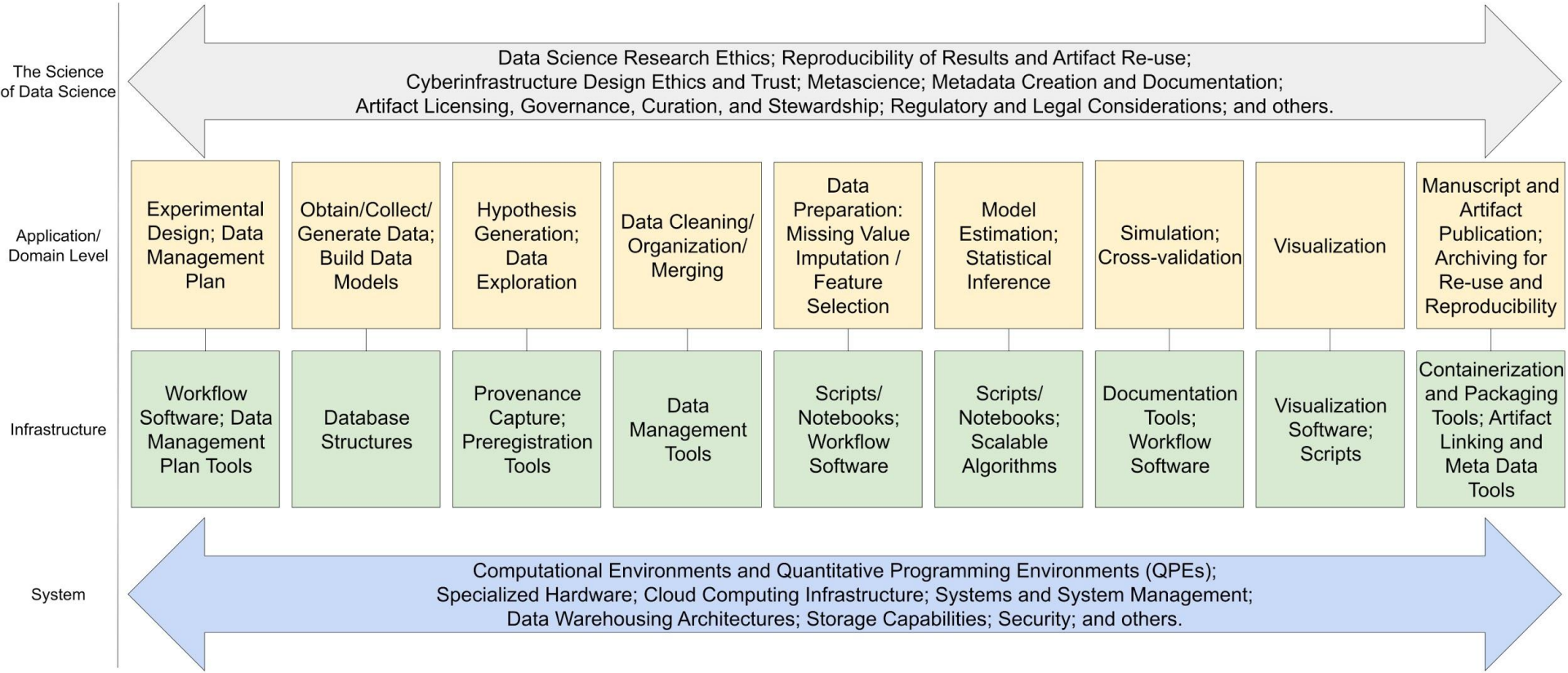
Researchers often use CI without control or even awareness of some elements.

Learning from ML: 5 articles, one dataset

	Preprocessing/Feature Selection Method						
Classifier(Paper)	1	3	6a	6b	9	29	**Average**
WeightedVote(1)	.91	.94	.97	.97	.89	.74	.90
NN(3)	.97	.94	.91	.94	.97	.97	.95
Linear SVM(3)	.97	.97	.94	.97	.97	.77	.93
Quadratic SVM(3)	.97	.88	.97	.97	.97	.91	.95
Adaboost(3)	.91	.91	.97	.97	.91	.91	.93
Logit(6)	.97	.97	.97	.97	.97	.88	**.96**
QDA(6)	.94	.91	.94	.97	.97	.85	.93
NN(9)	.97	.91	.85	.97	.94	.94	.93
Decision Trees(9)	.91	.91	.97	.97	.91	.77	.90
Bagging(9)	.94	.91	.97	.97	.92	.77	.91
Bagging CPD(9)	.74	.85	.82	.91	.77	.68	.79
FLDA(9)	.88	.88	.97	.97	.88	.88	.91
DLDA(9)	.97	.94	.97	.97	.97	.88	.95
DQDA(9)	.97	.94	.97	.97	.97	.88	.95
BayesNetwork(29)	.74	.88	.97	.97	.83	.62	.83
Average	.92	.92	.95	**.97**	.92	.83	

From: Stodden, Wu, & Sochat (2018). AIM: An Abstraction for Improving Machine Learning Prediction. IEEE First Workshop on Data Science.

An Abstraction: The Data Science Life Cycle



From: V. Stodden (2020). The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science. Forthcoming in CACM.

What does this new research/CI conceptualization imply for the research ecosystem?

A greater awareness of:

1. **the extent of and nature of the reliance on CI** across the research landscape e.g. integration of mathematics/algorithms/CI for discovery; representation of abstract reasoning in CI.
2. the importance of and **impact on findings CI can have** (introducing uncertainties, opacity, complexity, p-hacking/poor inference practices, ...).
3. **opportunities to (automatically) integrate and compare research findings**, including algorithms, methods, and data that generated them (why do we believe these findings?).
4. how **research pipelines are shaped by CI,** and how this creates opportunities for accelerated science through pipeline publication.

Some next steps

1. Development of best or better **practices and guidance for CI design and use in research and discovery** (e.g. required CI capabilities, capturing uncertainties, reliable and reproducible inference, transparency, integration and comparison of findings, ...)**.**
2. Increased funding to support the **development of CI for scientific research** (e.g. delivering transparency in computational methods, access to and re-use of digital scholarly objects). Most CI is repurposed e.g. docker, excel, GitHub, GPUs, ...
3. Recognition of the **connection between digital scholarly objects, CI, and scientific *results* and correctness**. Enable “system checks” on knowledge production, coherent benchmarking and comparisons.
4. **Inclusivity** in a CI-dominated research world and recognition of CI contributors.
5. Improved **training** in CI at all levels for all scientific contributors.