CSE 5449: Intermediate Studies in Scientific Data Management

Lecture 1: Introductions and Scientific Data Life Cycles

Dr. Suren Byna

The Ohio State University

E-mail: byna.1@osu.edu

https://sbyna.github.io

01/10/2023

Outline of this lecture

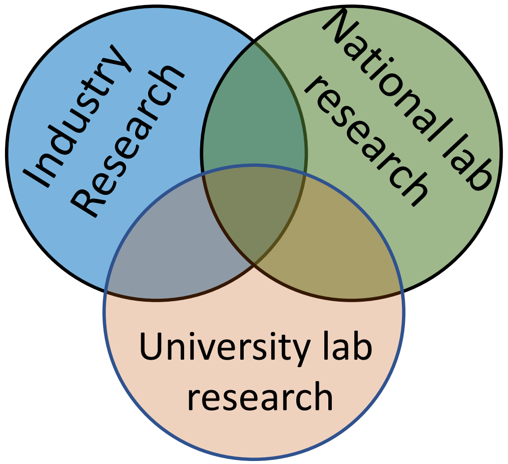
- Introductions
- Class logistics
- Basics: Scientific data and data life cycles
- Data life cycle examples
  - 1. Simulation use case
  - 2. Experimental use case
  - 3. Observation use case

Introductions - Who am I?

- Newly joined Professor at OSU (1 week)
- Senior Computer Scientist at Lawrence Berkeley National Laboratory
  - Joined as a research scientist in 2010
  - Led projects related to data management for science, data analysis, next generation storage systems
    - ExaIO, Proactive Data Containers, etc.
- A researcher at NEC Labs Inc, Princeton, NJ
  - Worked on a new programming model for ML applications - best effort parallelism
- Research assistant professor at Illinois Tech and Guest Researcher at Argonne National Laboratory
  - Developed a prefetching-based data management framework

Introductions - Who am I?

- Have been exposed to research at national labs, industry, and universities
- I'm also a Visiting Faculty at Lawrence Berkeley National Laboratory
- Collaborators at various Department of Energy national laboratories, universities, and industry (HPE, Intel, HDF Group, etc.)
- Current projects
  - ExaIO - Part of USA's Exascale Computing Project ($13M), ending in 2023
  - End-to-end object-focused software-defined data management ($2.7M), ending in 2025
  - Experimental and observational data management ($2M), ending in 2023



Funded research assistant positions available

Summer intern positions at LBNL are available

Introductions

- Student intros
  - Basics, current research interests, etc.

- What are your expectations from this course?

- What excites you about research?

Class logistics - Time, location, contact

- Tuesdays and Thursdays
  - 11:15 AM-12:10 PM
- Baker Systems 180
- Office hours
  - Right after the class
  - By appointment -- Planning on being in my office (Dreese Labs, Room # 791) most of the semester except when I'm on travel
- E-mail: byna.1@osu.edu

sbyna.github.io/teaching/5449-sdm.html

Class logistics - What would you learn by the end of the course?

- Data life cycles in science (applies to areas beyond science)
  - Simulations, experiments / observations, and analytics (visualizations, ML/AI, etc.)
- Data management software and hardware stacks
  - HPC, cloud, edge systems
- I/O libraries
  - HDF5, NetCDF, ADIOS, ROOT, etc.
- File and data management systems
  - Lustre, Spectrum Scale, BeeGFS, Ceph, DAOS, etc.
- Performance understanding, common causes of bottlenecks, and tuning
  - I/O performance characterization, performance issues, visualization of I/O logs, auto-tuning, …
- Knowledge management - metadata and provenance
- Designing next-gen data management systems for science
  - Object-centric systems, software-defined data management systems, …
- Research gaps and challenges

Class logistics - Grading plan

Grading component	%
Attendance, participation in class discussions, and evaluation of class presentations	15%
Class presentation	20%
Final exam (comprehensive, open book)	25%
Class project	40%

Class logistics - Class schedule (tentative)

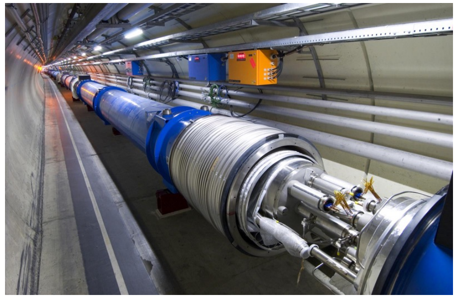
Week	Topic	Presenter	Notes / Due dates
1 (1/10 & 1/12)	Intros & Data life cycles	Suren Byna	
2 (1/17 & 1/19)	Software and hardware stacks of storage and data management	Suren Byna	Project topics - Provided by Prof
3 (1/24 & 1/26)	I/O libraries	Suren Byna	Select project topics and presentations
4 (1/31 & 2/02)	File and data management systems	Suren Byna	Discuss a project initial plan w/ Prof
5 (2/07 & 2/09)	Parallel I/O Stack performance tuning	Suren Byna	
6 (2/14 & 2/16)	Performance understanding, bottlenecks, and tuning	Suren Byna & Guest	
7 (2/21 & 2/23)	Knowledge management - metadata and provenance	Suren Byna	Discuss project progress w/ Prof
8 (2/28 & 3/02)	Student presentations - related research, gaps, proposal	Students	Discuss project progress w/ Prof
9 (3/07 & 3/09)	Student presentations - related research, gaps, proposal	Students	
10	**Spring Break - No class**		
11 (3/21 & 3/23)	Designing next-gen data management systems for science	Suren Byna	
12 (3/28 & 3/30)	SDM - Research gaps and challenges	Suren Byna	Discuss project progress w/ Prof
13 (4/04 & 4/06)	Student project presentations - Progress reports	Students	
14 (4/11 & 4/13)	Scientific data discovery, data quality, etc.	Guest	Discuss project progress w/ Prof
15 (4/18 & 4/20)	Student project presentations - Final report outs	Students	
16 (4/25 & 4/27)	Final Exam & Recap / Guest lecture	Suren Byna / Guest	Final Exam on 4/25 (to be confirmed)

Class logistics - Etiquette and reading materials

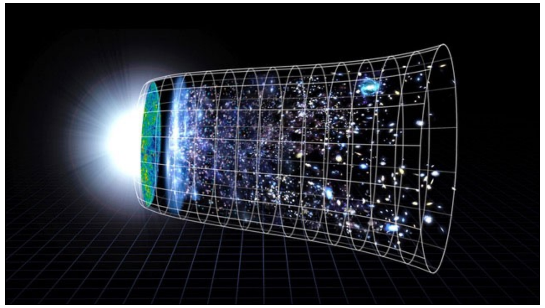
- Feel free to ask questions, discuss in class (15% for participation)
- Phones on mute, laptops closed (unless instructed by me), and be respectful to everyone in the class
- Learn about my approach to R&D
  - General research resources (on the class page)
  - Not universal or a single formula
  - Some guidelines that could be used based on what's applicable to you
- Reading materials
  - Will post a set of papers on the class page →
    - sbyna.github.io/teaching/5449-sdm.html
    - sbyna.github.io
  - Reference books (not required for the class)
    - "High Performance Parallel I/O", edited by Prabhat and Quincey Koziol
    - "Scientific Data Management: Challenges, Technology, and Deployment", edited By Arie Shoshani and Doron Rotem
  - Project ideas → (Will be provided on the class page by next week)

- Any questions?

Data driven science - Datasets are large, diverse, and complex

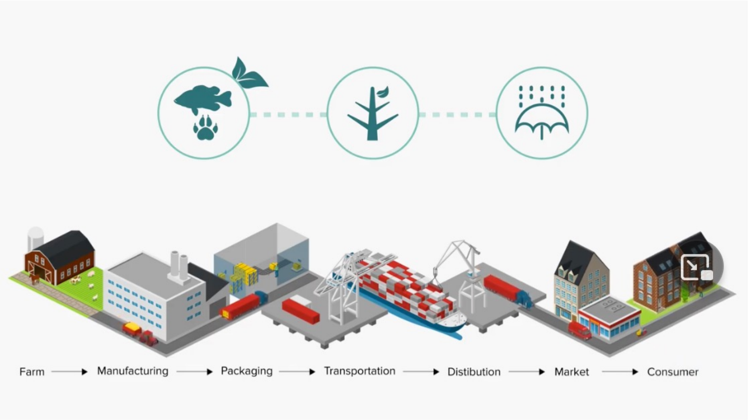


CERN Data Centre stores more than 30 petabytes of data per year from the LHC experiments - CERN Facts



Vera Rubin Observatory: the 20 terabytes of data collected per night is as much as the entire 10 years of data collected by the Sloan Digital Sky Survey. The ten years of the LSST survey will contribute to building a 500 petabyte database of images and a 15 petabyte catalog of text data describing properties of nearly 40 billion individual stars and galaxies.

AmeriFlux: Data contributed to the AmeriFlux Network are complex and diverse. Ecosystem-level field sites acquire continuous measurements from a large number of sensors at high temporal resolution, which can result in large quantities of data.

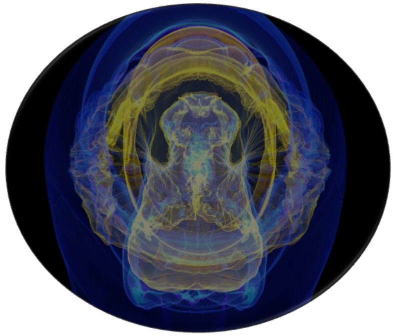
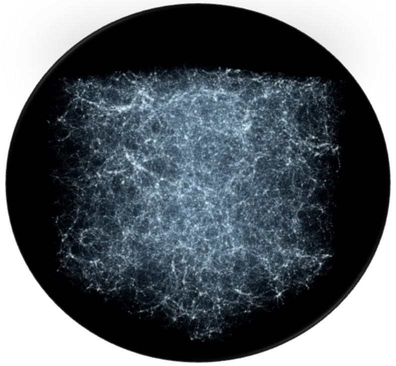
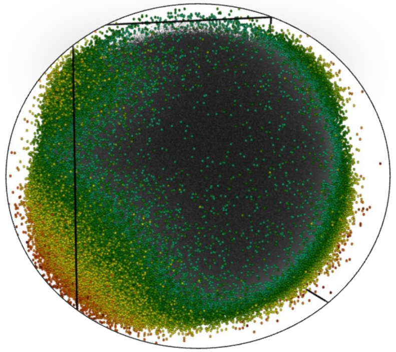
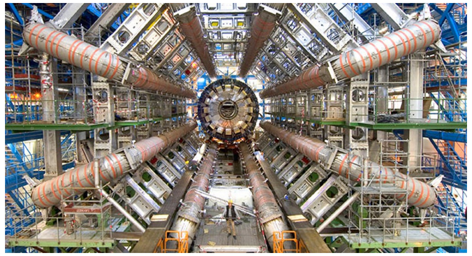


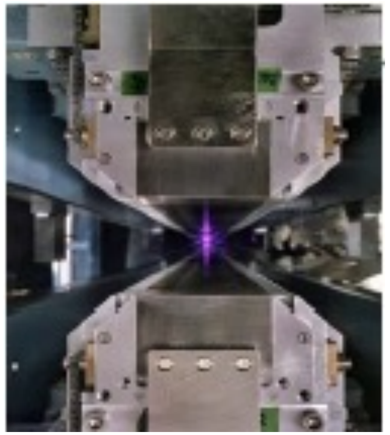
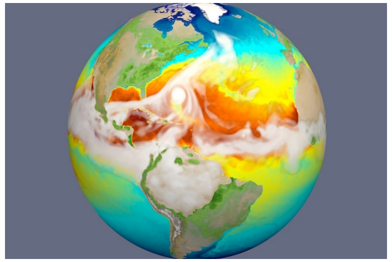
ICICLE: data collected from a wide range of sources.

Sources of figures:
CERN Facts: https://home.cern/resources/faqs/facts-and-figures-about-lhc
HEP Cosmic frontier: https://science.osti.gov/hep/Research/Cosmic-Frontier
AmeriFlux: https://ameriflux.lbl.gov/data/aboutdata/
ICICLE: https://icicle.osu.edu/

Scientific data generators

- Simulations
- Experiments
- Observations

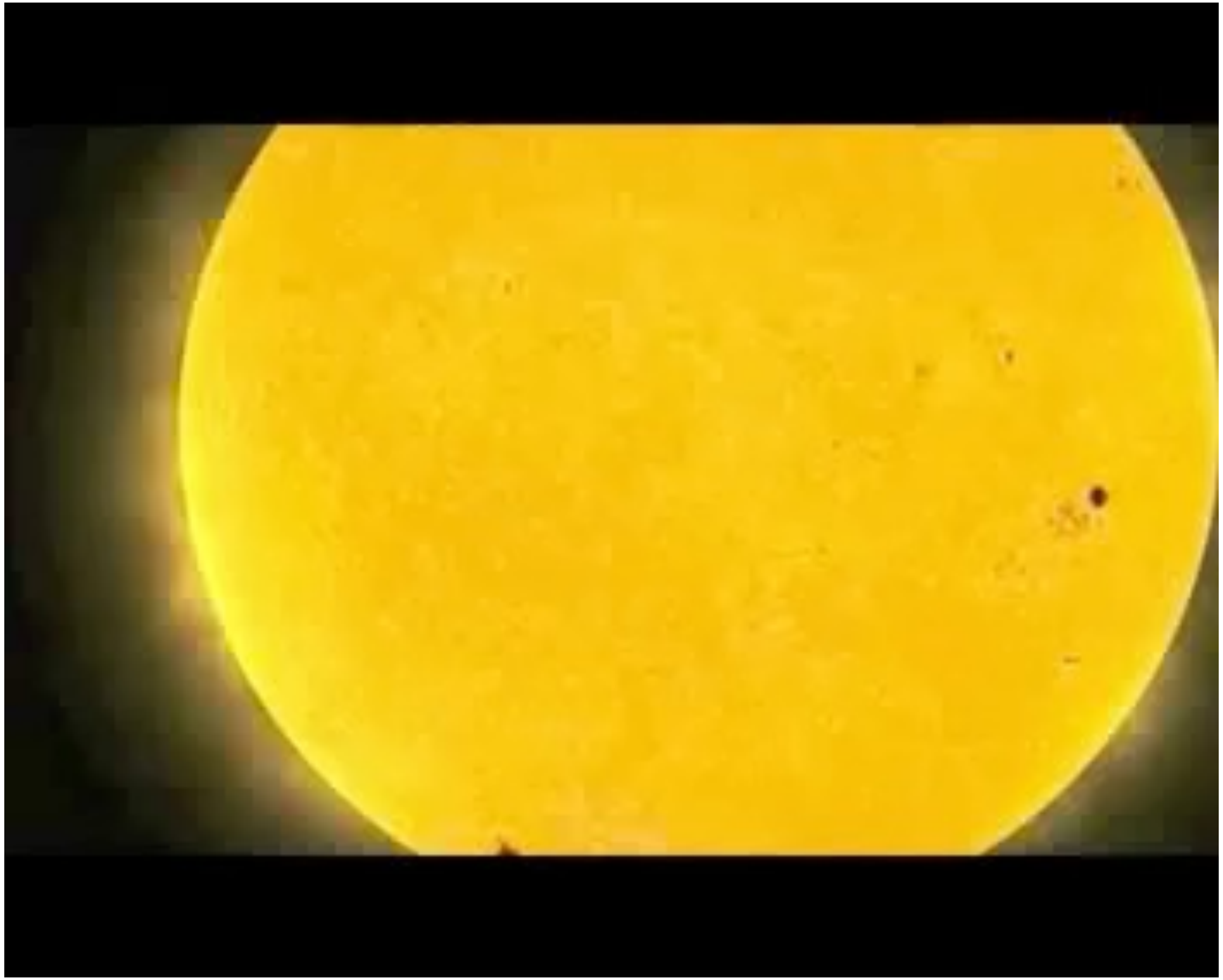






SLAC
NATIONAL ACCELERATOR LABORATORY


SQUARE KILOMETER ARRAY

Simulation use cases – Plasma physics

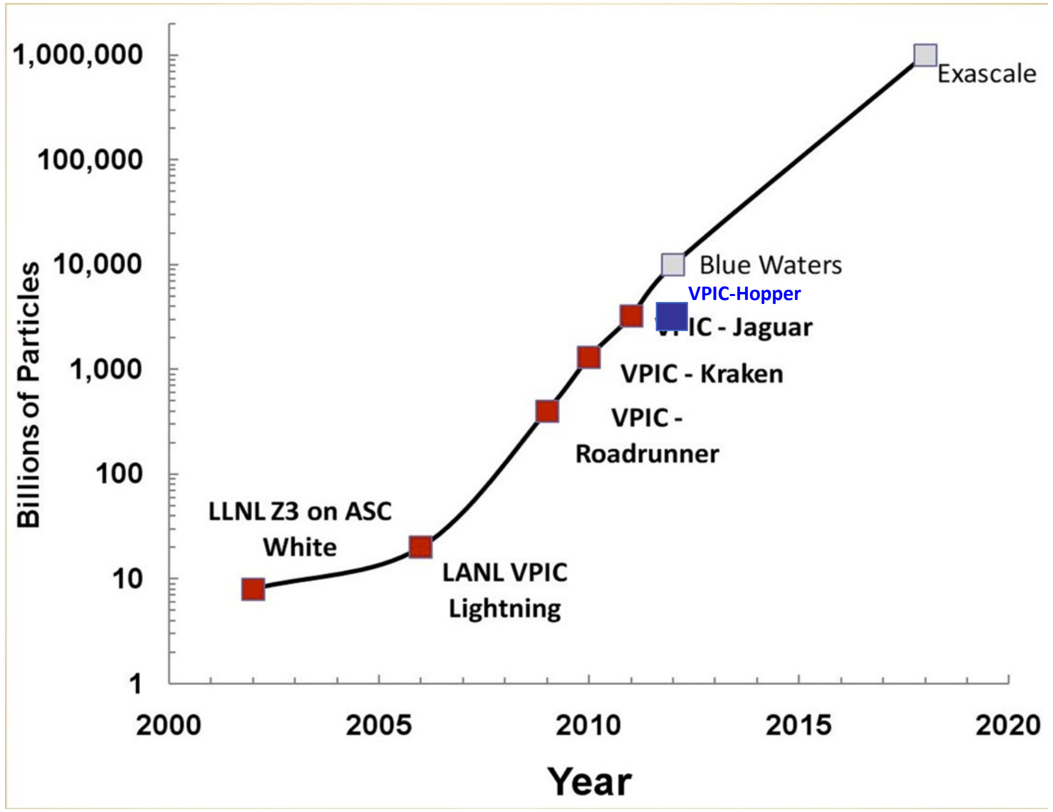


- Understanding physical mechanisms responsible for producing magnetic reconnection in a collision-less plasma
- Are the highly energetic particles preferentially accelerated along the magnetic field?
- What is the spatial distribution of highly energetic particles?
- What are the properties of particles near the reconnection hot-spot (the so-called X-line)?

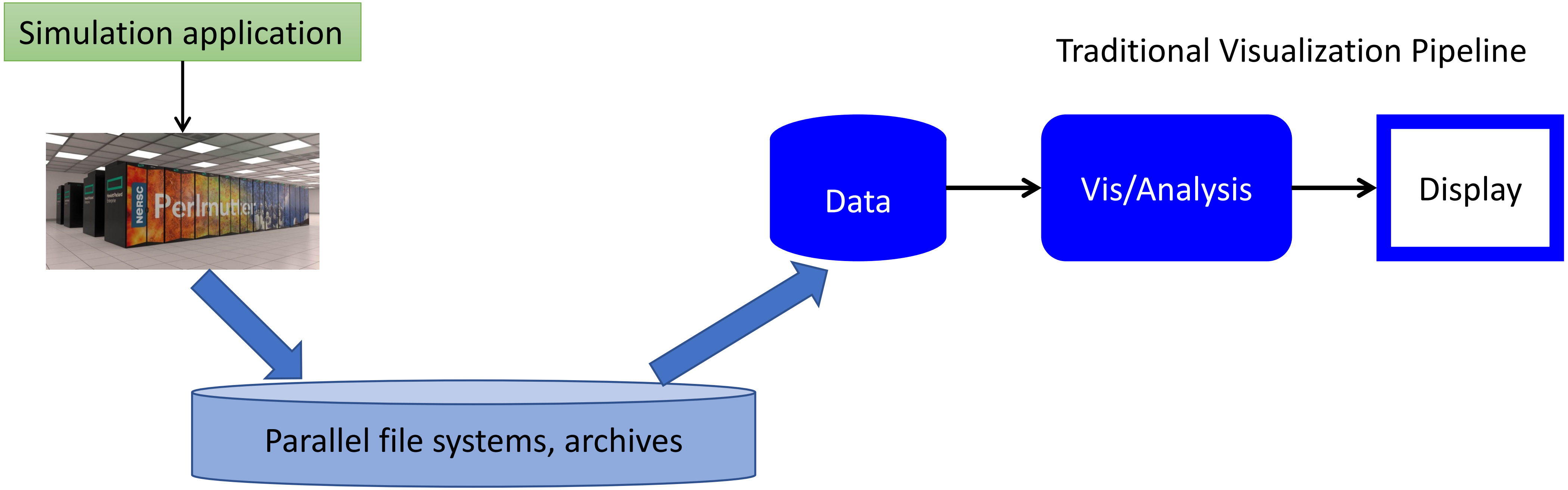
https://www.youtube.com/watch?v=mgUZwoR0gcE

Vector Particle-in-Cell (VPIC) Simulation in 2012

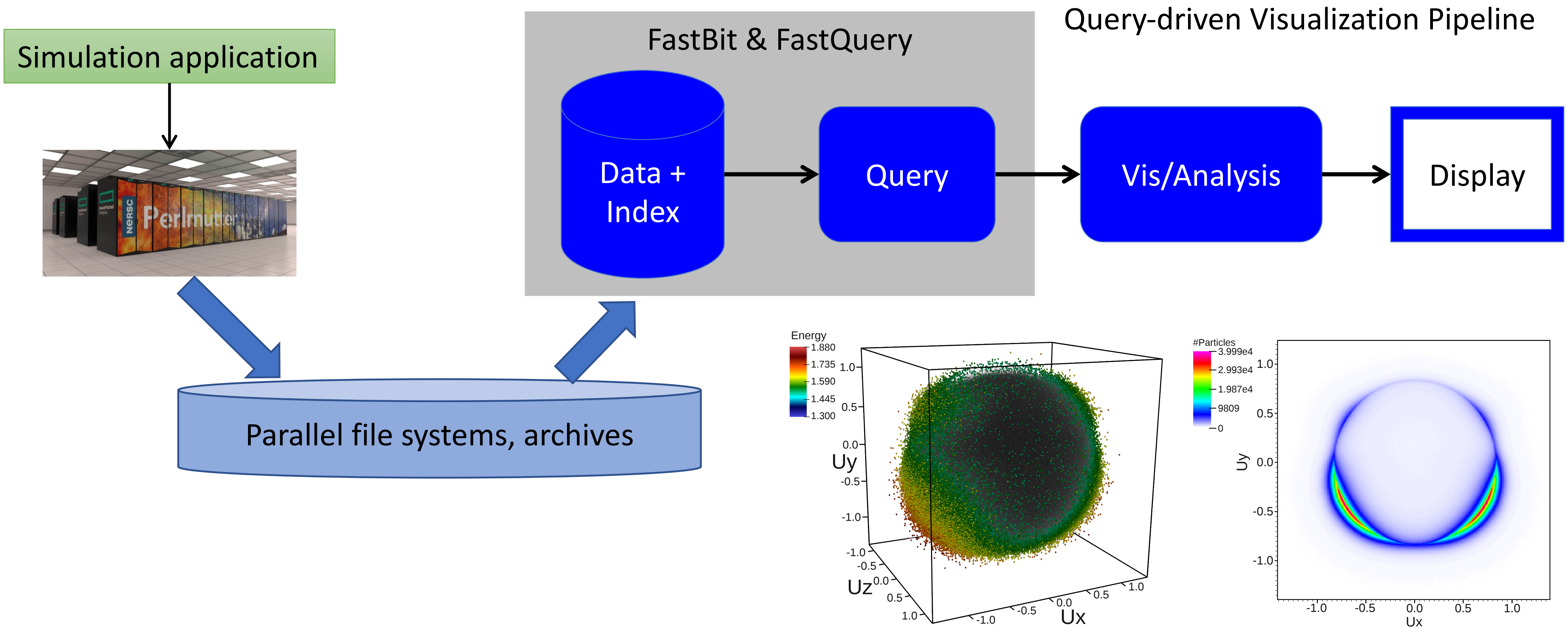
- ✧ A state-of-the-art 3D electromagnetic relativistic PIC plasma physics simulation
- ✧ It is an exascale problem and scales well on large systems
- ✧ An open boundary VPIC simulation of magnetic reconnection
- ✧ NERSC Hopper Supercomputer
  - 6,384 compute nodes; 2 twelve-core AMD 'MagnyCours' 2.1-GHz processors per node; 32 GB DDR3 1333-MHz memory per node; Interconnect with a 3D torus topology



Data life cycle of VPIC plasma physics simulation and analysis

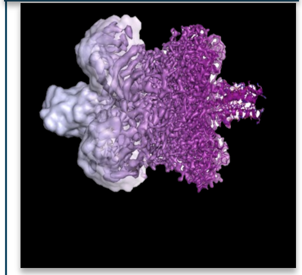


Data life cycle of VPIC plasma physics simulation and analysis



Data life cycle of experimental & observation use cases

Acquire

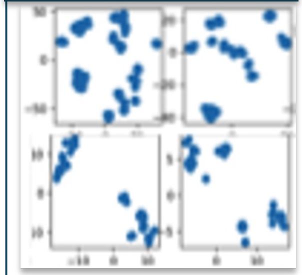


Collect from sensors, experiments, simulations

Transfer

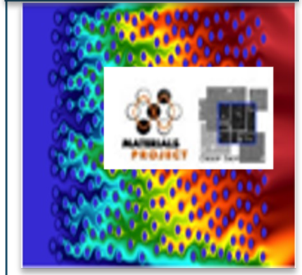
Move from instrument to computing center (supercomputing / cloud)

Clean



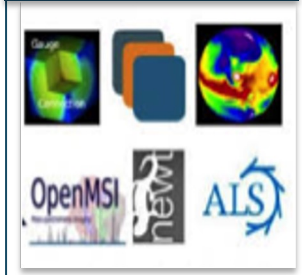
Organize, annotate, filter, encrypt, compress

Use/Reuse



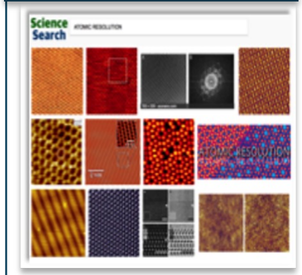
Analyze, mine, model, learn, infer, derive, predict

Publish



Disseminate & aggregate, using portals, databases

Preserve



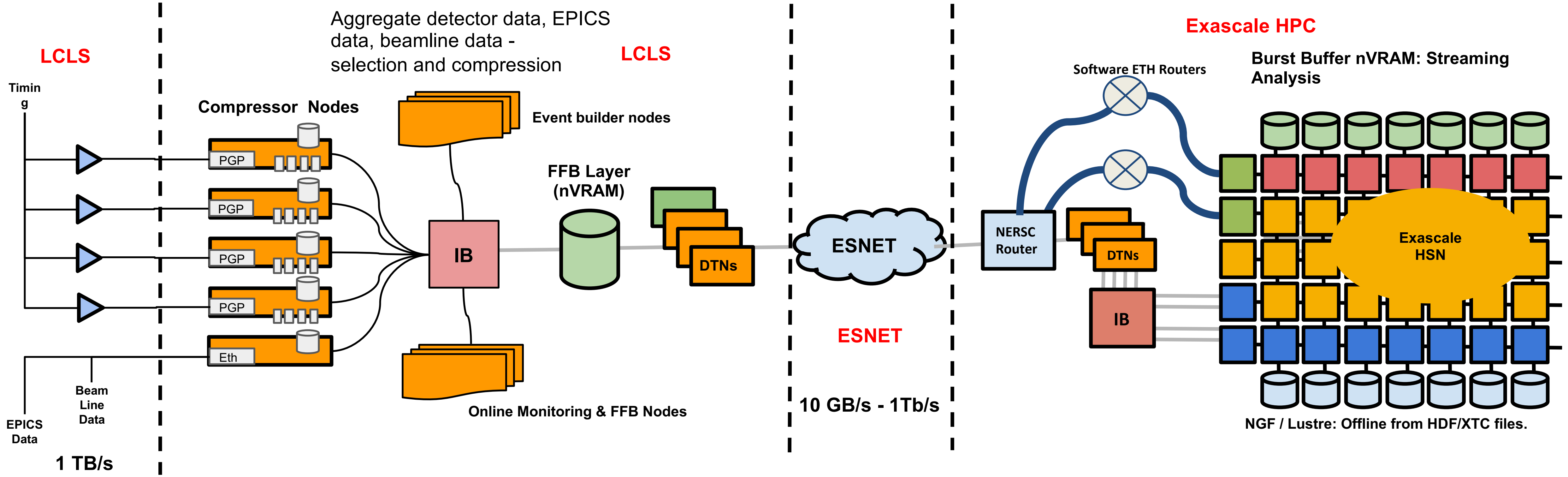
Index, curate, age, track provenance, purge

From a slide from Debbie Bard's SuperFacility presentation

An example from LCLS-II use case

https://lcls.slac.stanford.edu/lcls-ii/ https://www.youtube.com/watch?v=t7jUZwhZdd0&t=12s

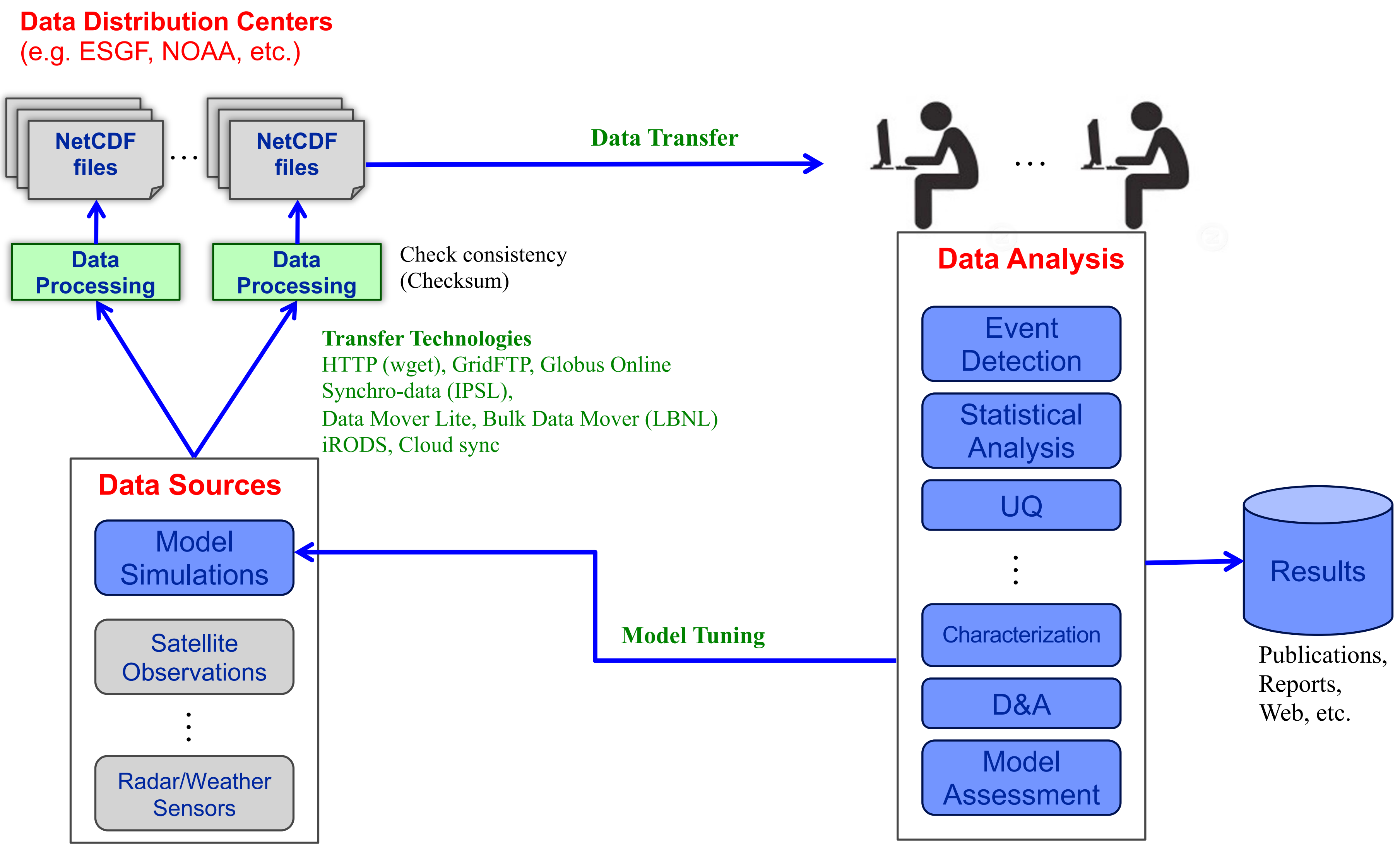
An example from LCLS-II use case



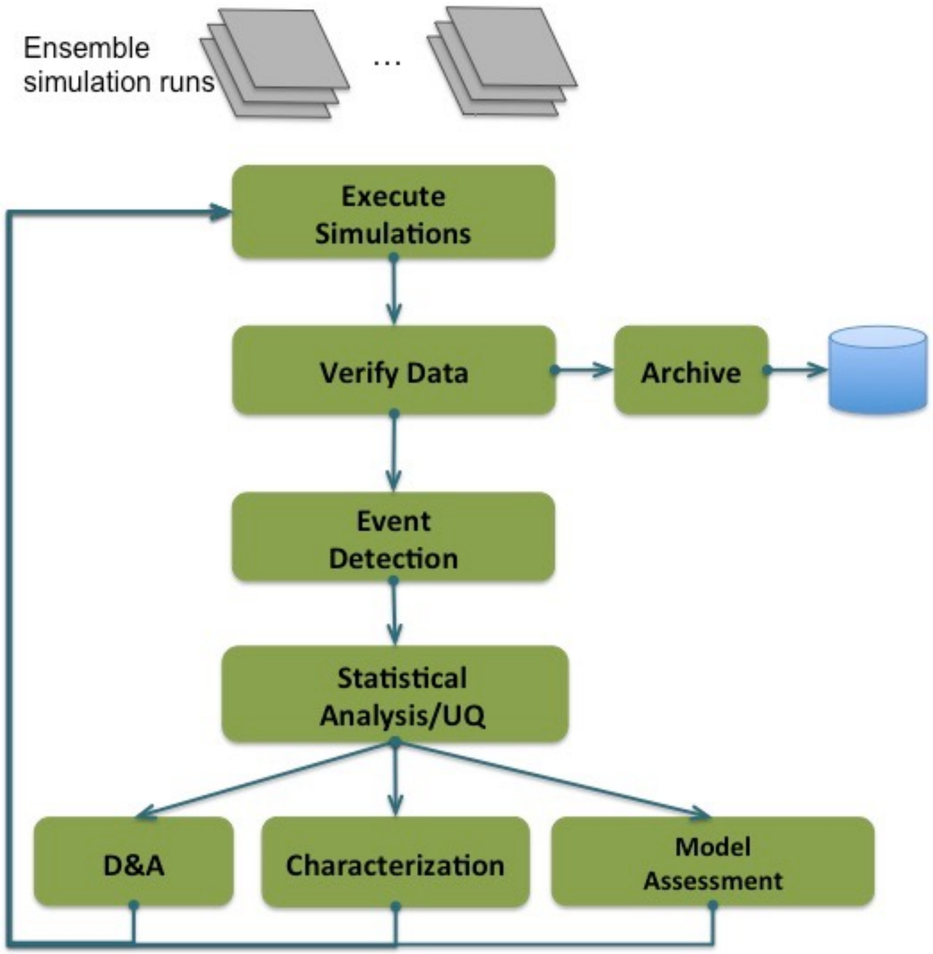
ExaFEL Data Flow

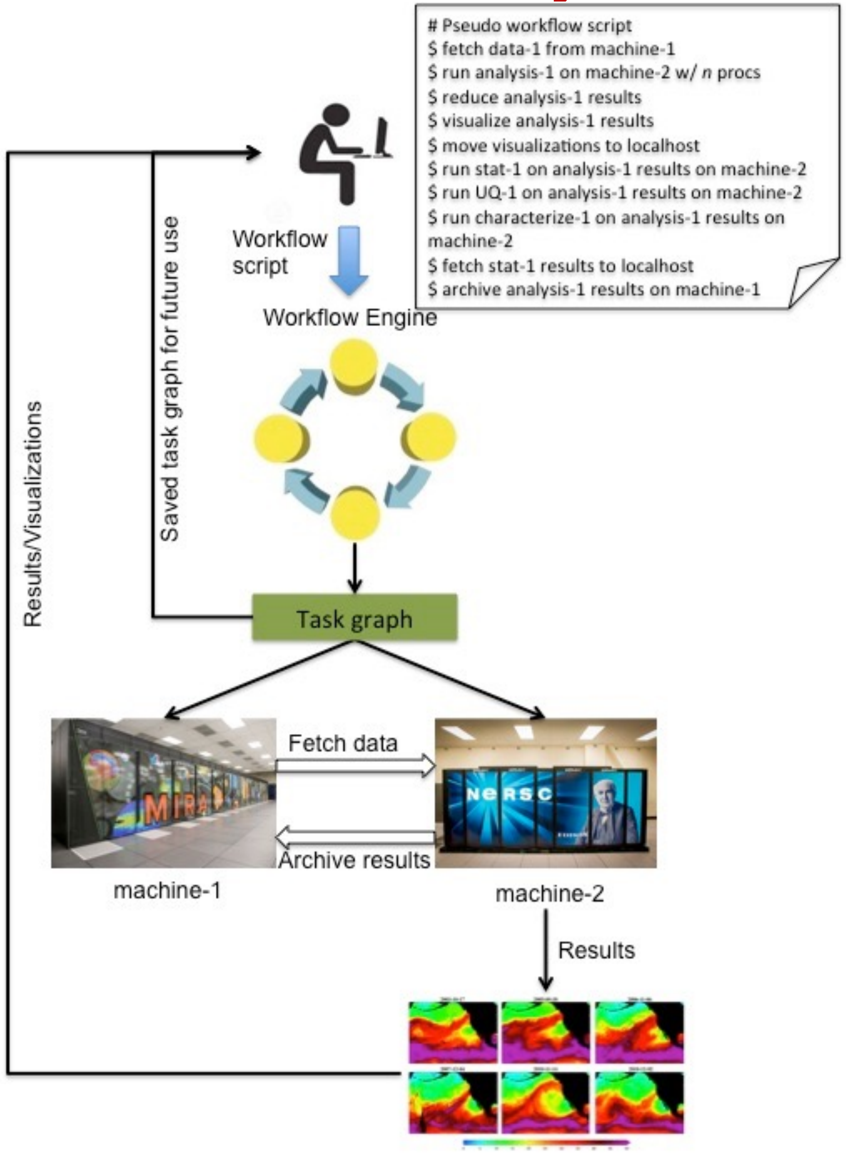
From Amedeo Perazzo, SLAC

Data life cycle of a climate use case

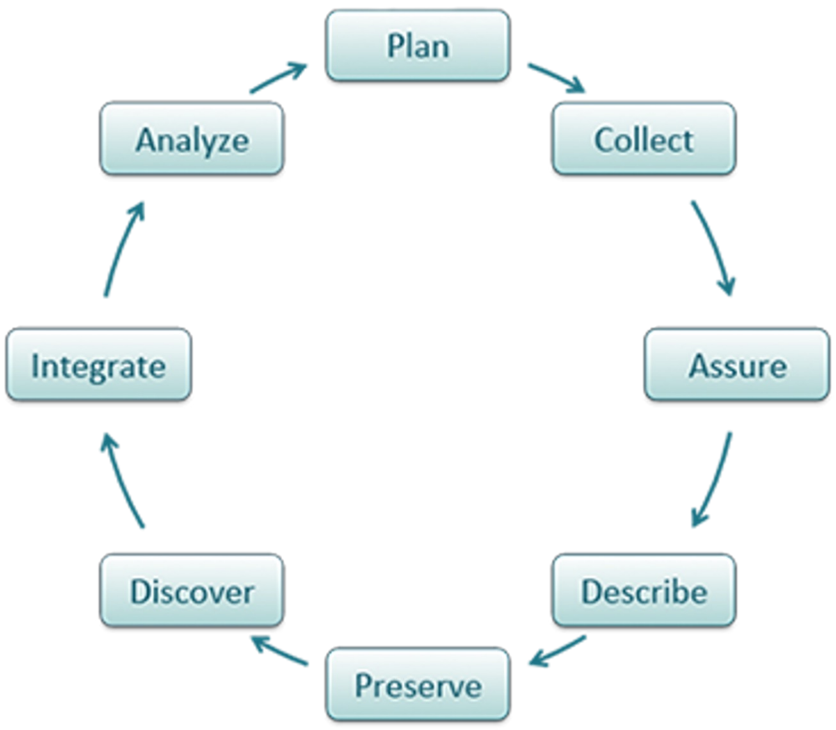


A couple more climate use cases from extreme event analysis

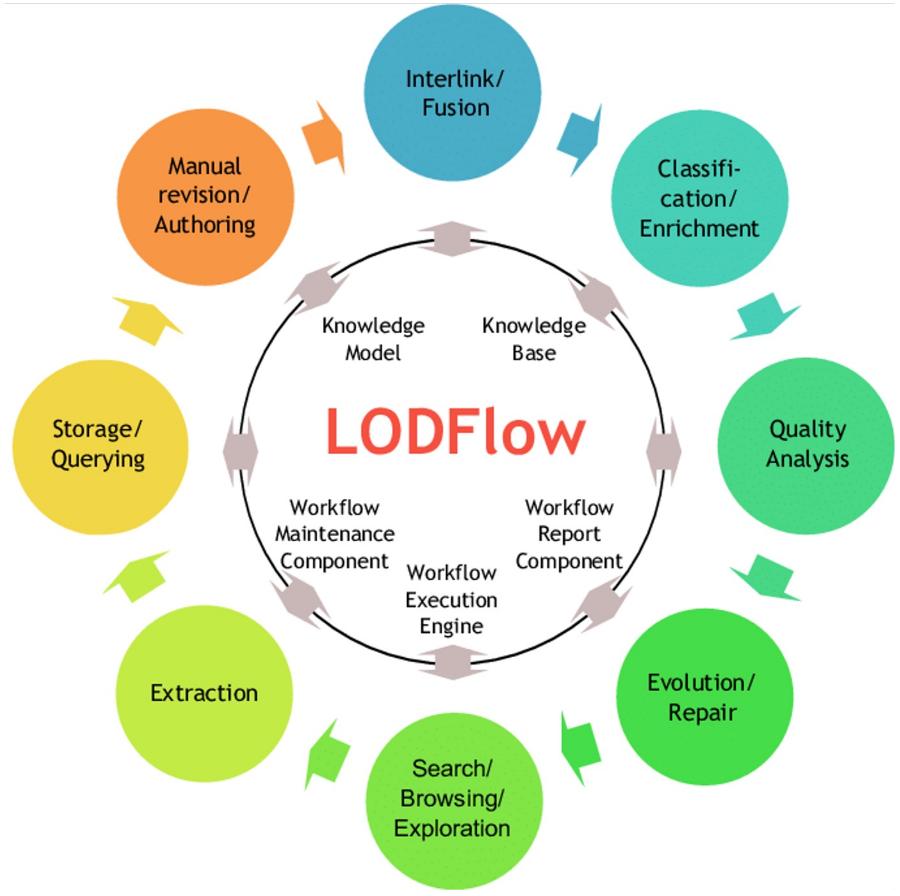




Overviews of data life cycles

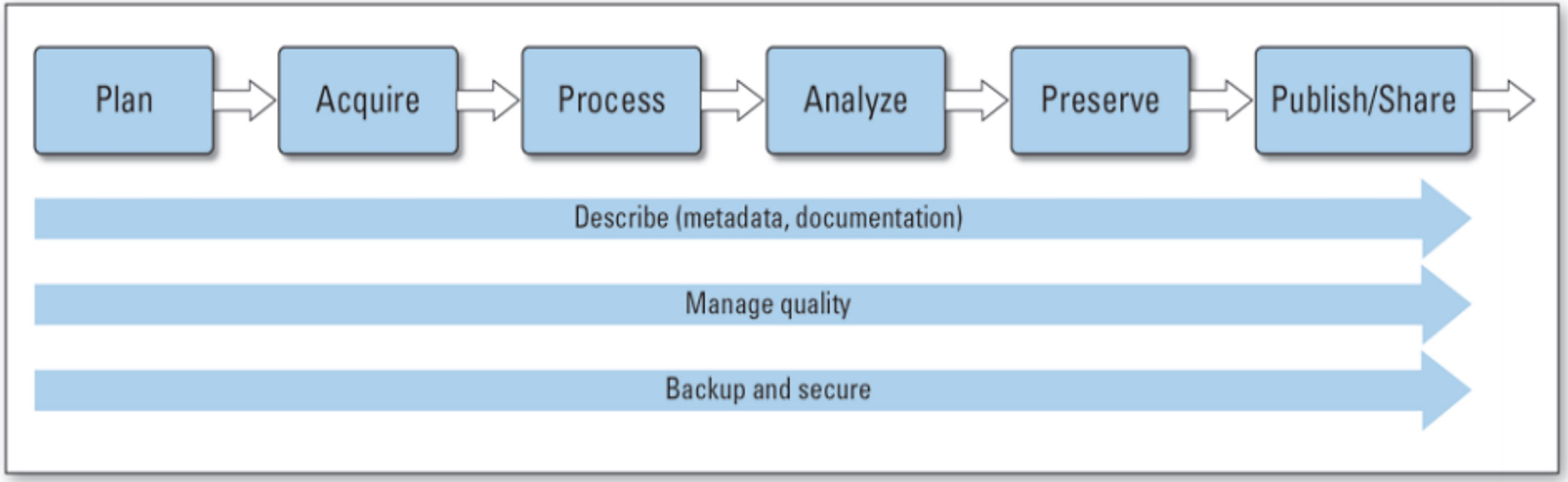


Source: https://www.dataone.org/data-life-cycle



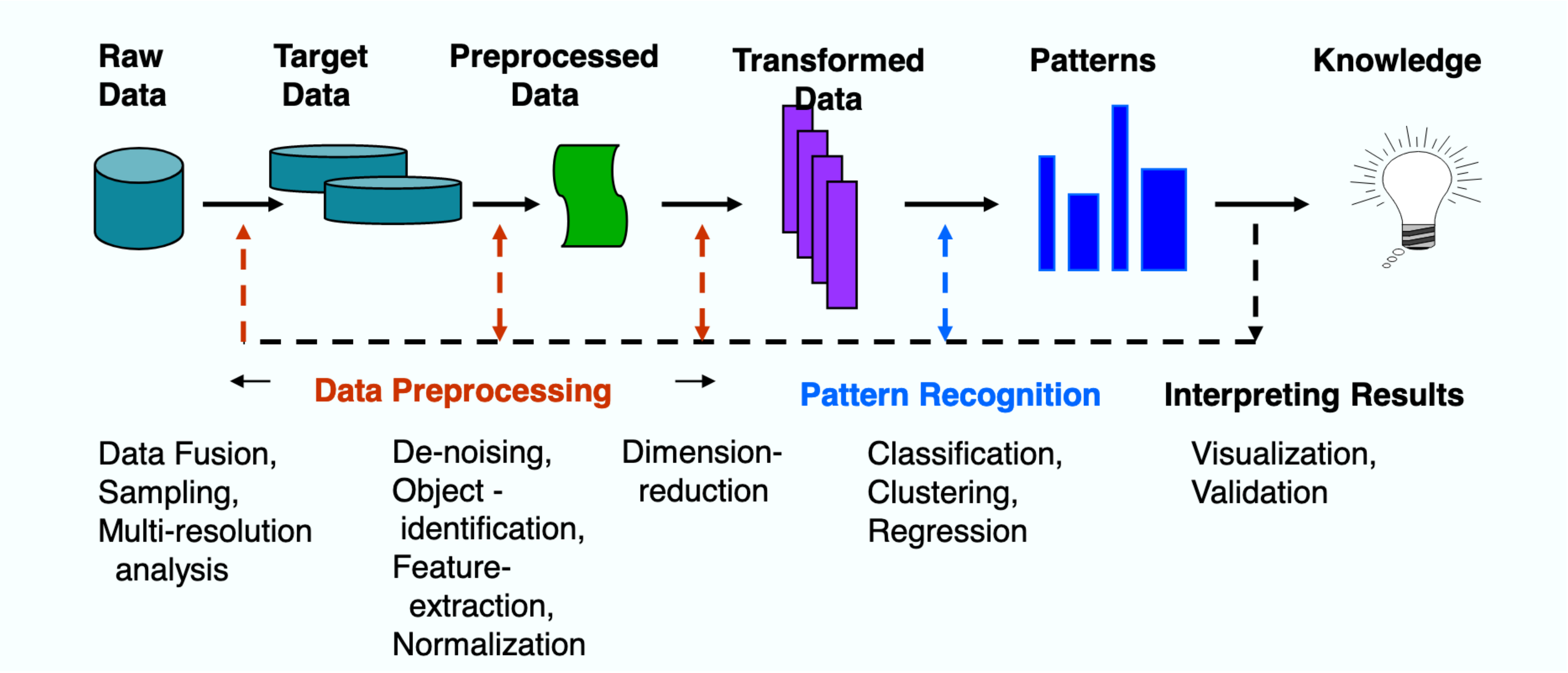
Source: Rautenberg et al. 2015 (https://doi.org/10.1145/2814864.2814882)

More data life cycle overviews



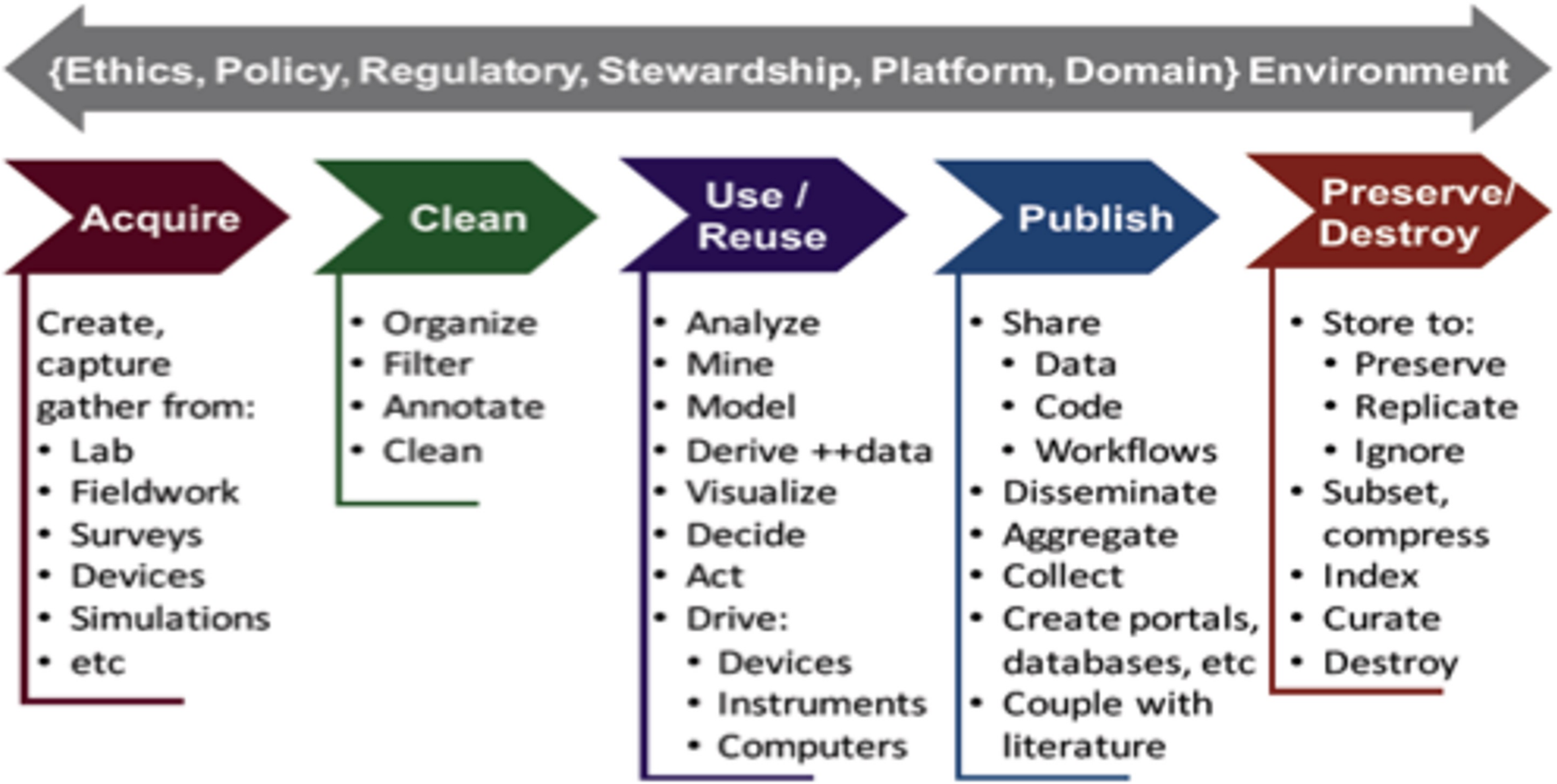
Source: The United States Geological Survey Science Data Lifecycle Model
https://pubs.usgs.gov/of/2013/1265/pdf/of2013-1265.pdf

Raw data to knowledge – Tera- / Peta-bytes to Mega-bytes



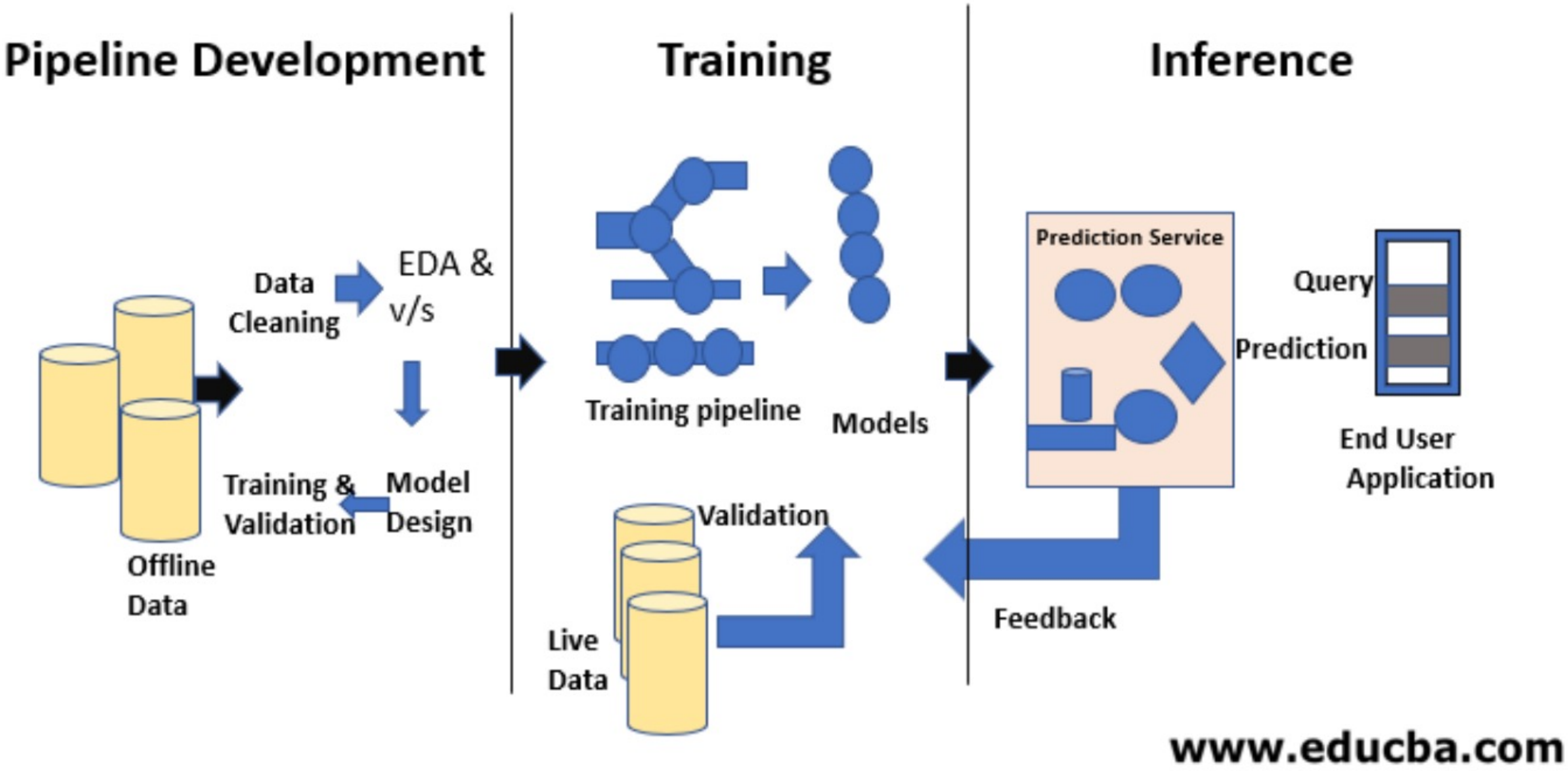
From Arie Shoshani, SDM Center presentation

From an NSF report



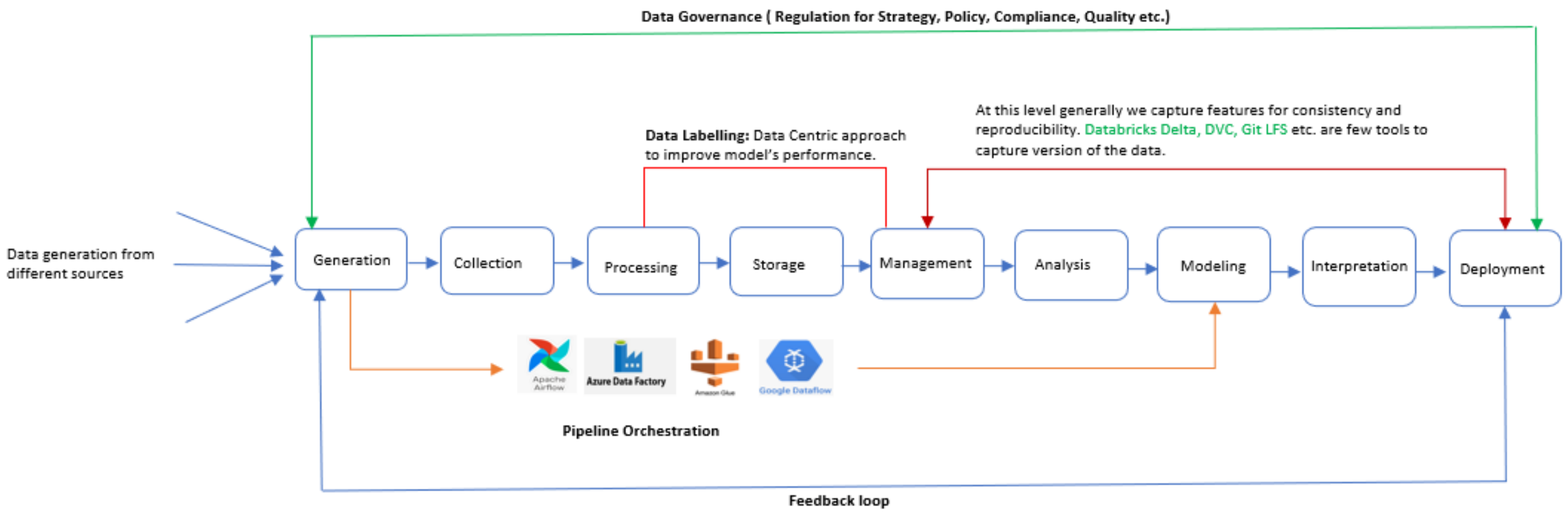
Source: NSF 2016 report: Realizing the Potential of Data Science

Machine learning life cycle



Source: https://www.educba.com/machine-learning-life-cycle/

Data life cycle in ML

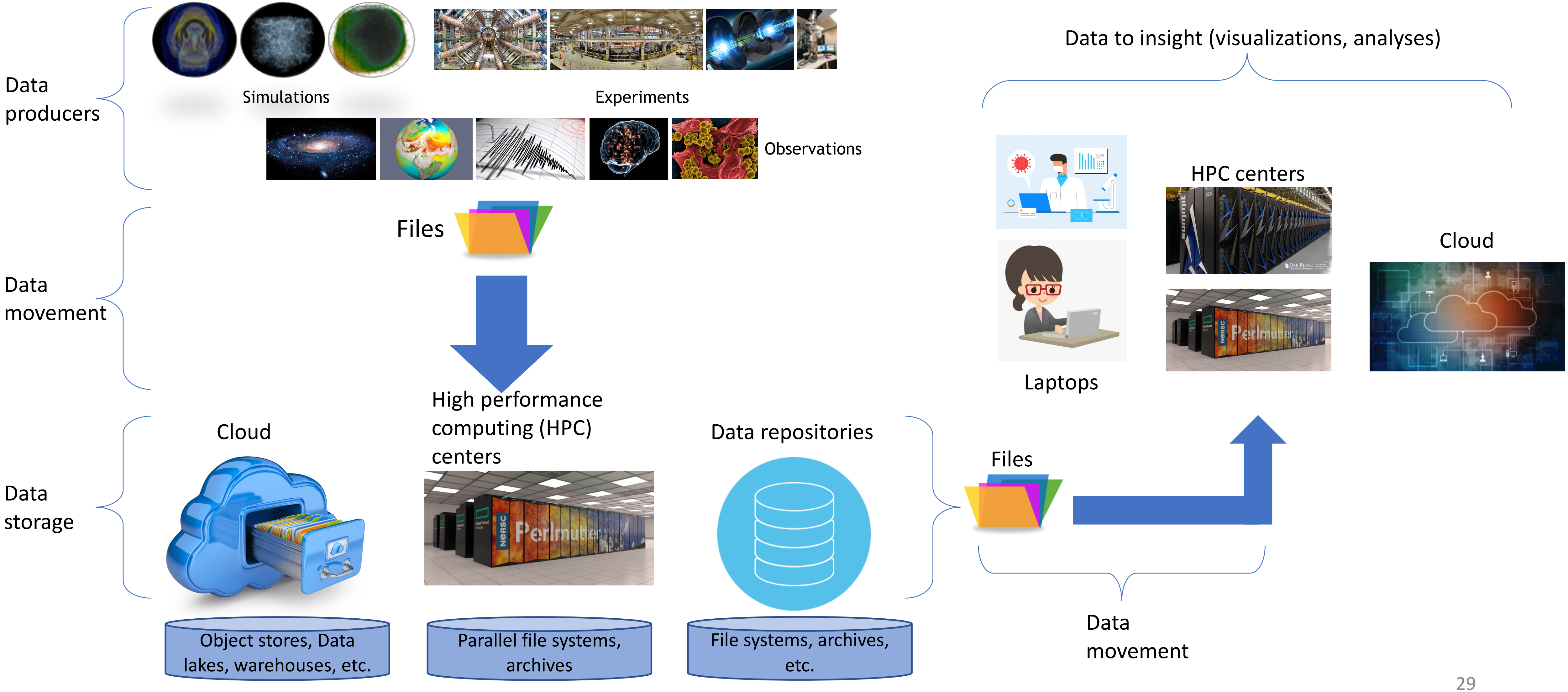


Source: https://medium.com/analytics-vidhya/data-lifecycle-of-machine-learning-project-83b2e656911d

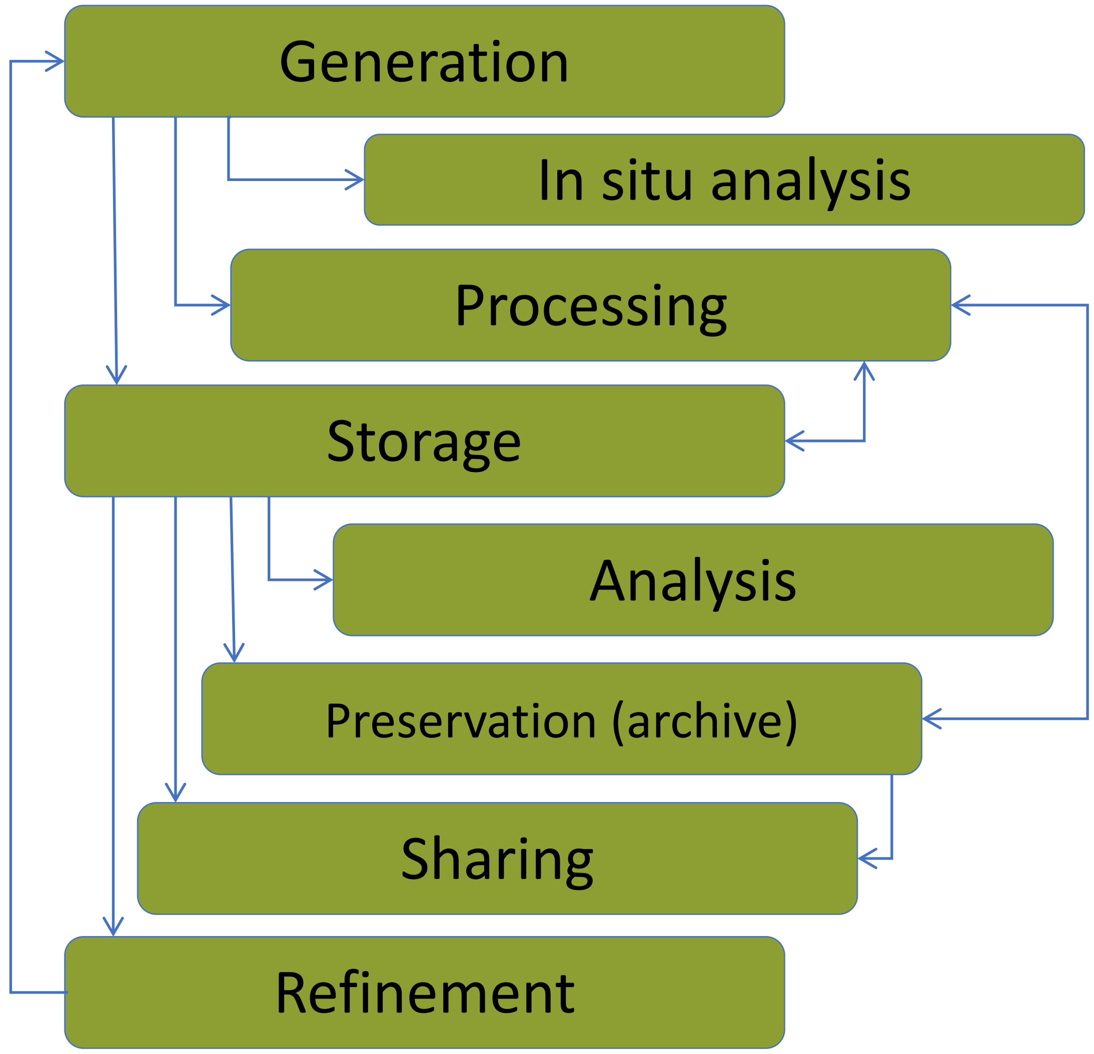
Another interesting image of data life cycle

Source: http://gluedata.com/data-lifecycle-2/

Data life cycle - An overview



Life cycle of scientific data



Additional reading

- Management, Analysis, and Visualization of Experimental and Observational Data - The Convergence of Data and Computing
  - https://www.osti.gov/servlets/purl/1366310

- The data life cycle, by Jeannette M. Wing
  - https://datascience.columbia.edu/news/2018/the-data-life-cycle/

- USGS scientific data lifecycle model
  - https://pubs.usgs.gov/of/2013/1265/pdf/of2013-1265.pdf
  - https://www.usgs.gov/data-management/data-lifecycle

Next class

- Common data and file formats used in scientific data
- Storage architectures in supercomputing systems
- Data management software stack