Scientific Applications and Heterogeneous Architectures – Data Analytics and the Intersection of HPC and Edge Computing

Michela Taufer

THE UNIVERSITY OF TENNESSEE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®

Acknowledgements


T. Estrada


H. Weinstein


M. Cuendet

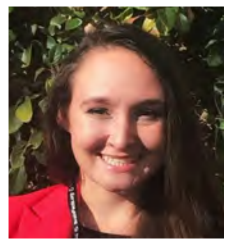

E. Deelman


R. Vargas


R. da Silva


T. Johnston


T. Do


B. Mulligan

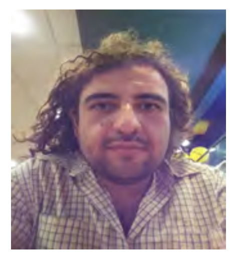

D. Rorabaugh


S. Thomas


H. Carrillo


A. Razavi

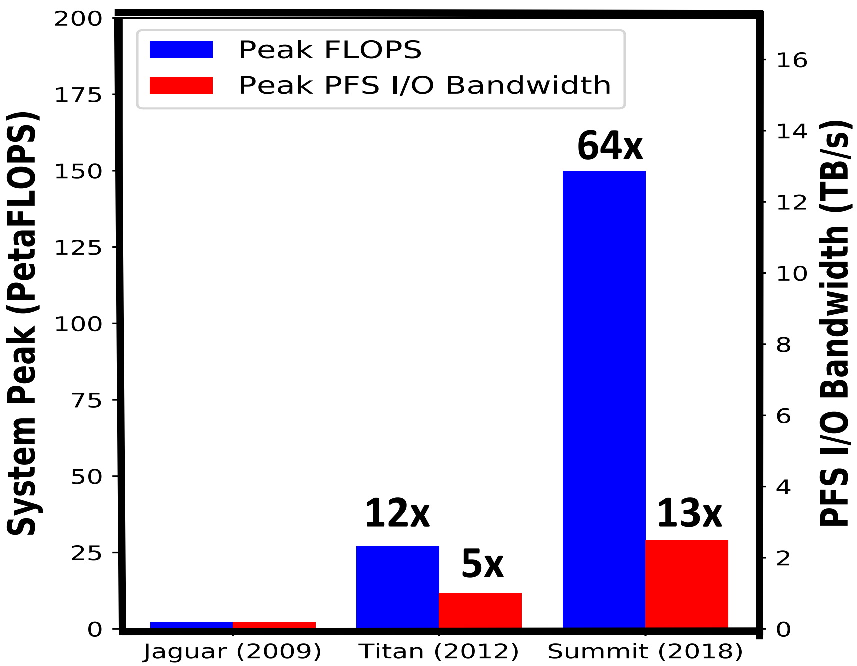

R. LLamas


M. Guevara

Sponsors:

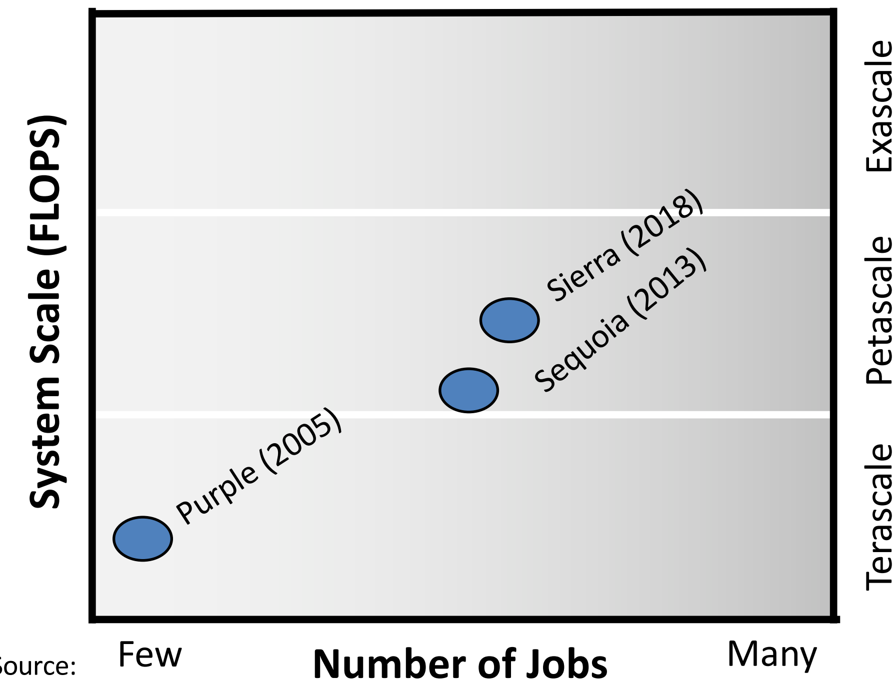
Trends in Next-Generation Systems: IO Gap and Ensembles

Widening IO Gap



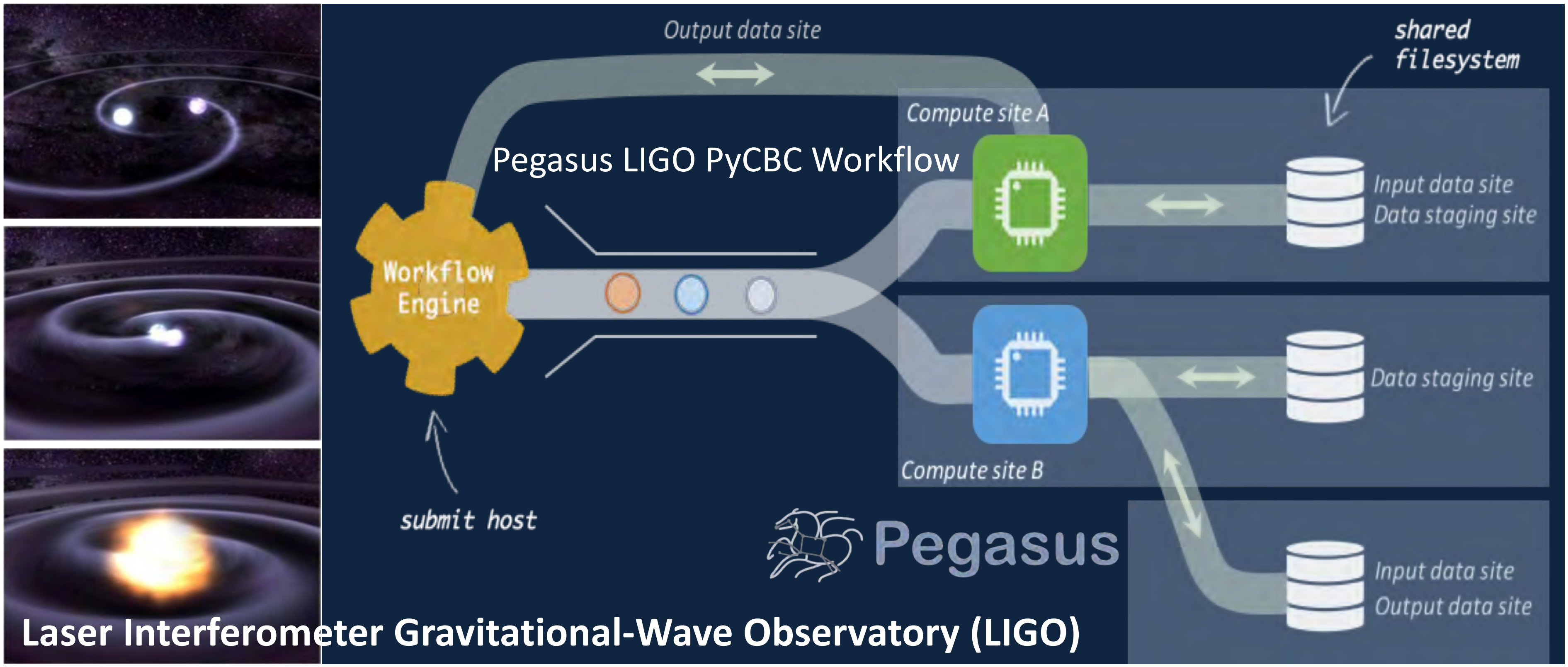
Source: Lucy Nowell (DOE)

Rising Importance of Ensembles



Source: https://wci.llnl.gov/simulation/computer-codes/uncertainty-quantification

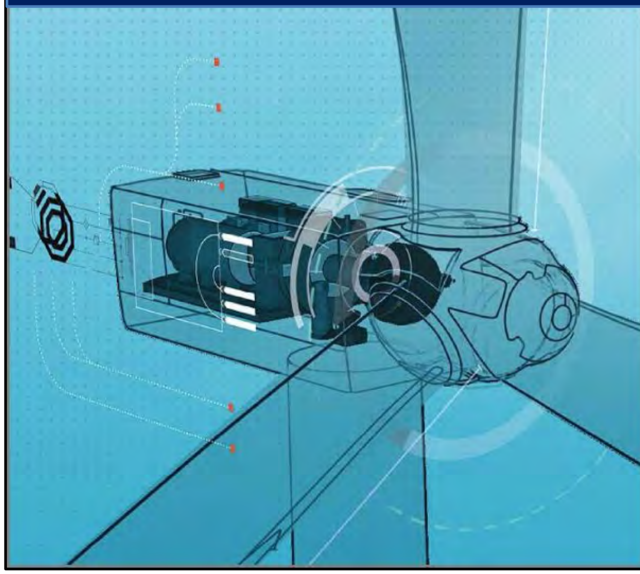
Trends in Workflows: Compute + Analytics + Data



Source: Ewa Deelman, ISI - USC

Extending HPC to Connect to the “Edge”

Simulations	Data Analytics	Real System with Sensors



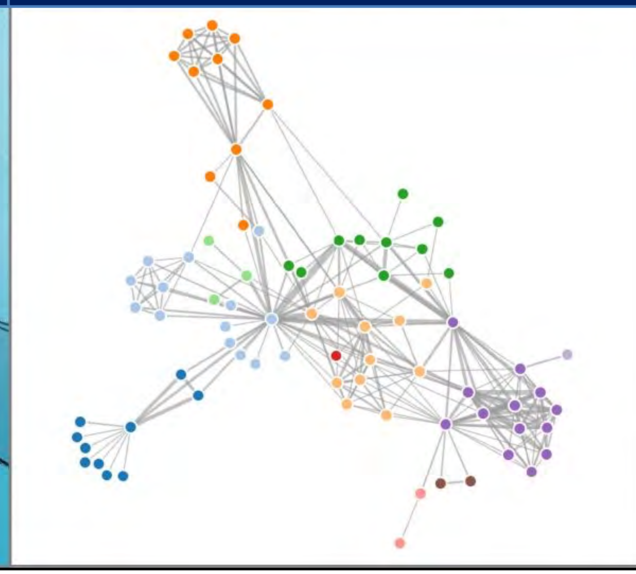


Image Source: https://iiot-world.com/digital-disruption/the-right-representation-of-digital-twins-for-data-analytics/

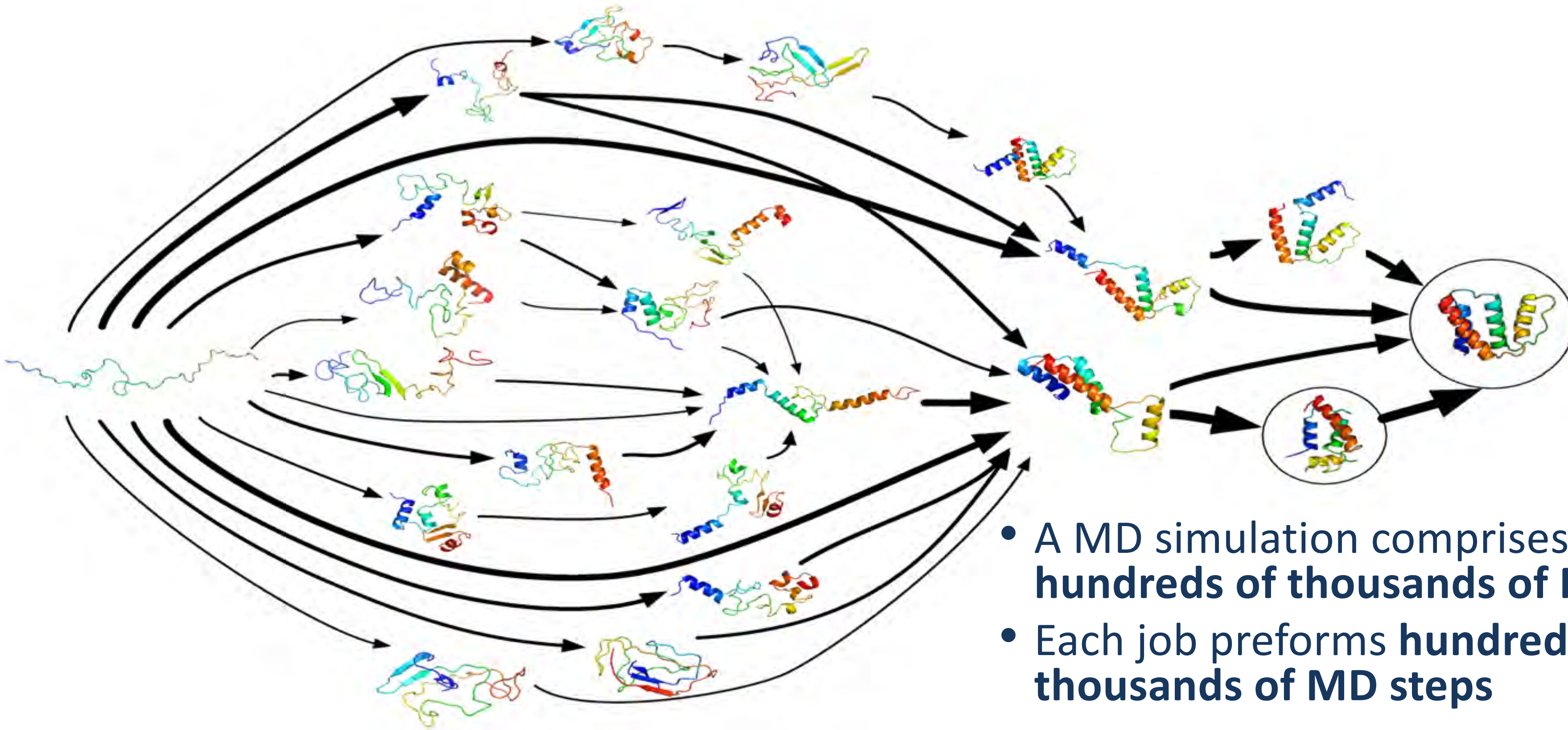
Two Use Cases

- Extending HPC to integrate data analytics
  - Next generation MD workflows
  - Molecular structures
  - ***Data transformation – i.e.,*** *capturing information*
  - ***Dataflow modeling –*** *i.e., lost information*
- Extending HPC to connect to the "Edge"
  - Next generation precision farming
  - Soil moisture data
  - ***Data prediction*** *– i.e., from coarse- to fine-grained information*

Two Use Cases

- Extending HPC to integrate data analytics
  - Next generation MD workflows
  - Molecular structures
  - ***Data transformation – i.e.,*** *capturing information*
  - ***Dataflow modeling –*** *i.e., lost information*
- Extending HPC to connect to the “Edge”
  - Next generation precision farming
  - Soil moisture data
  - ***Data prediction*** *– i.e., from coarse- to fine-grained information*

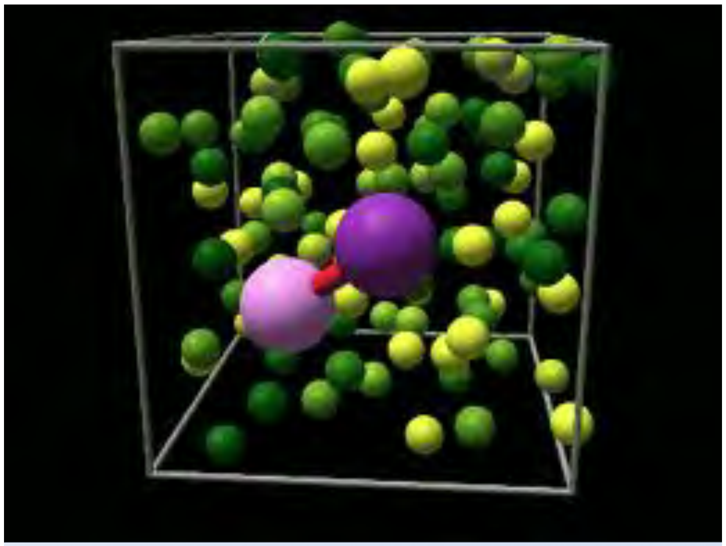
Classical Molecular Dynamics Simulations



- A MD simulation comprises of **hundreds of thousands of MD job**
- Each job preforms **hundreds of thousands of MD steps**

Source: Vincent Voelz Gorup - CC BY-SA 3.0, http://www.voelzlab.org/

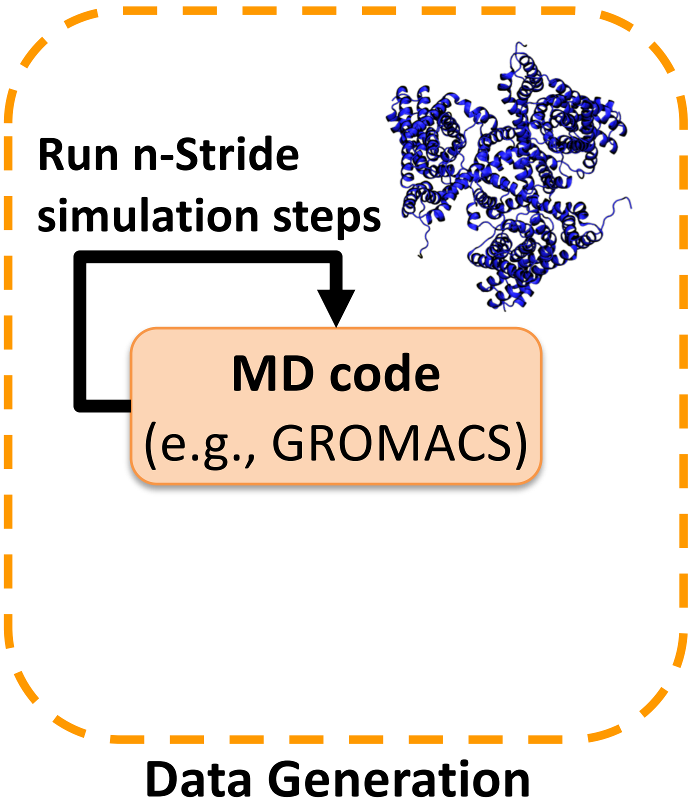
Classical Molecular Dynamics Simulations



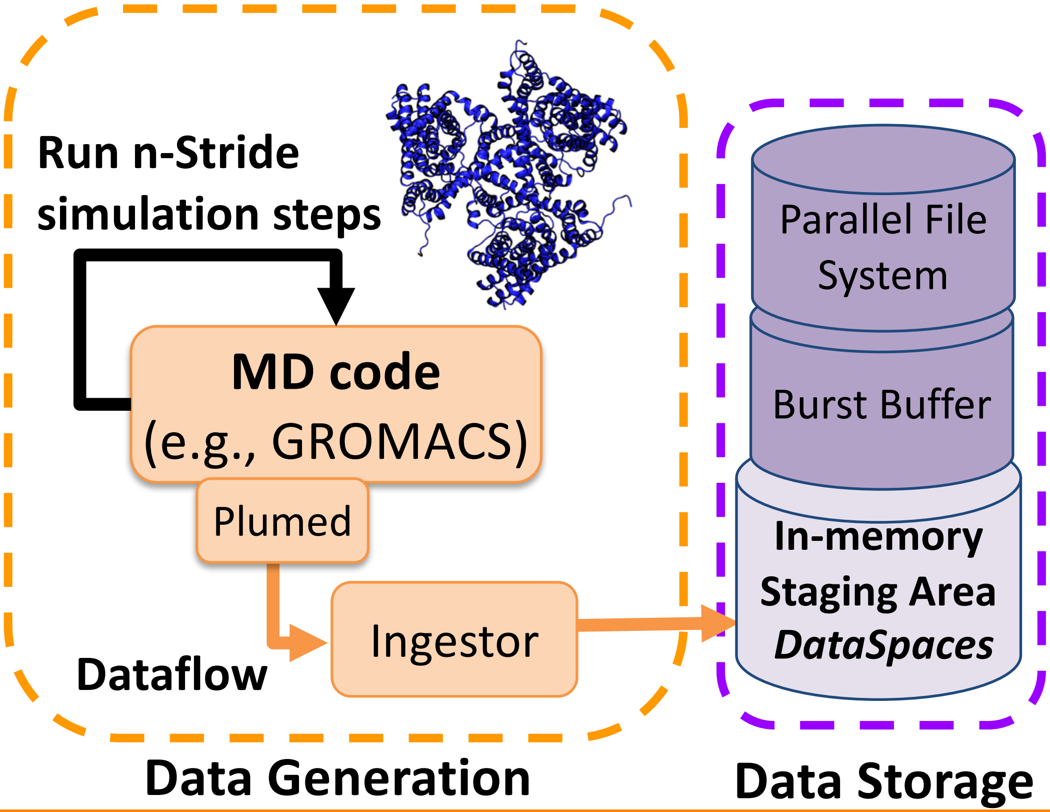
→Forces on single atoms
→ Acceleration
→ Velocity
→ Position

- MD step computes **forces** on single atoms (e.g., bond, dihedrals, nonbond)
- Forces are added to compute **acceleration**
- Acceleration is used to update **velocities**
- Velocities are used to update the **atom positions**
- Every *n* steps (Stride)
  → ***Store 3D snapshot or frame***

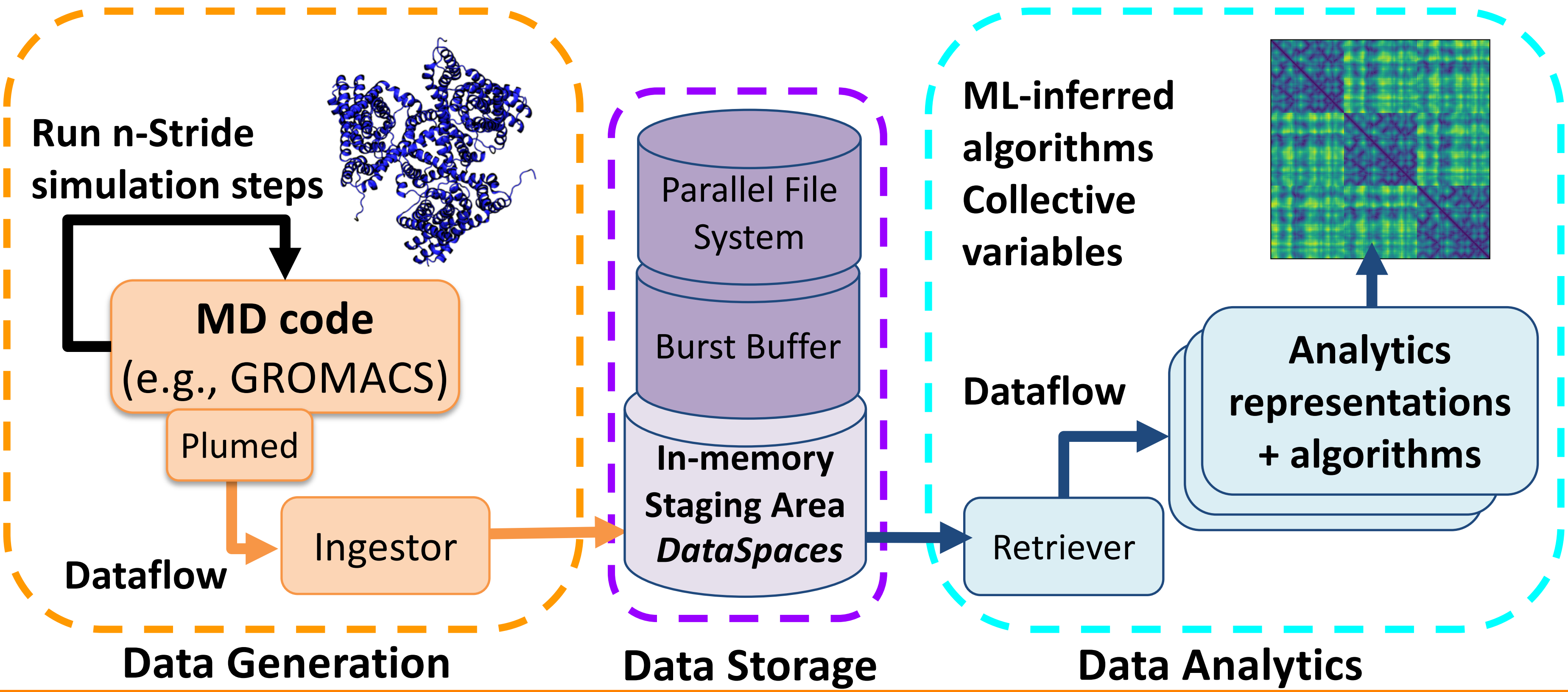
Building a Closed-loop Workflow



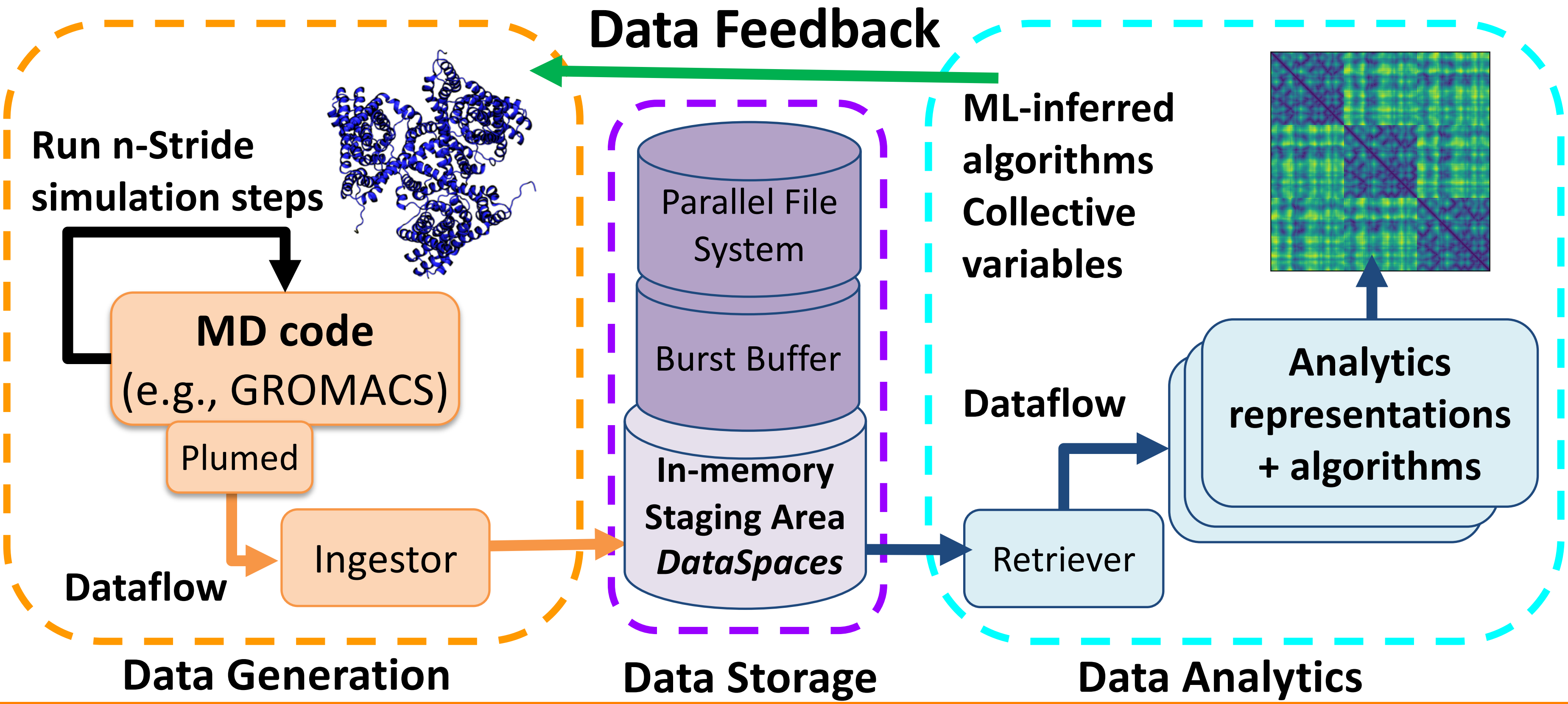
Building a Closed-loop Workflow



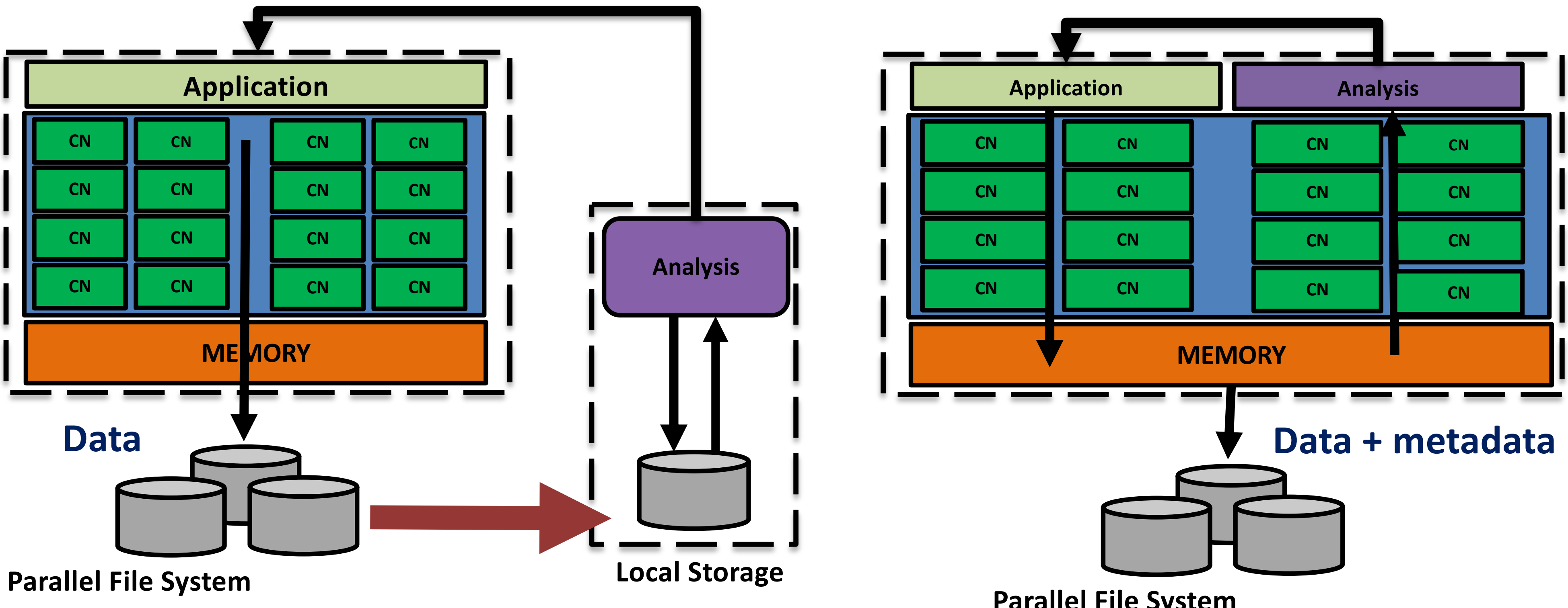
Building a Closed-loop Workflow



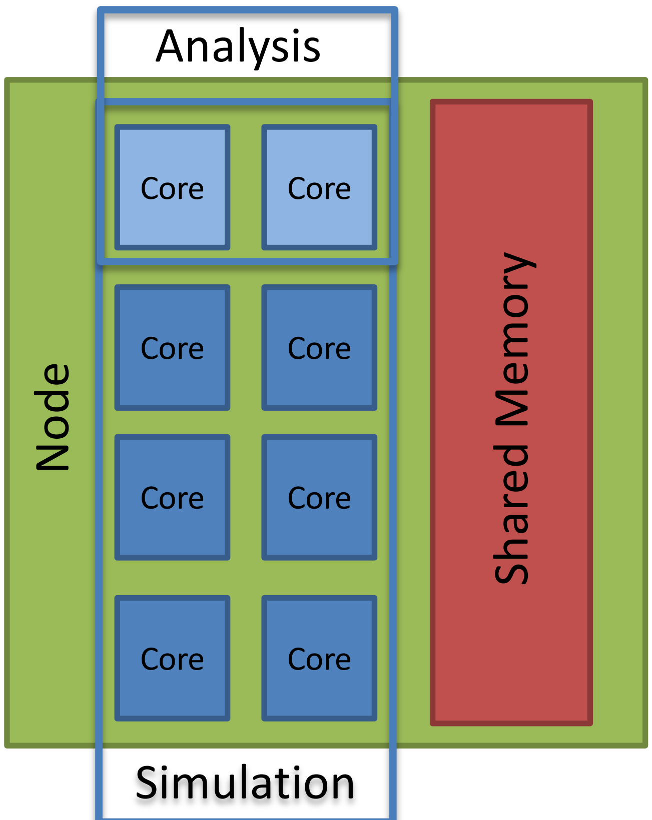
Building a Closed-loop Workflow

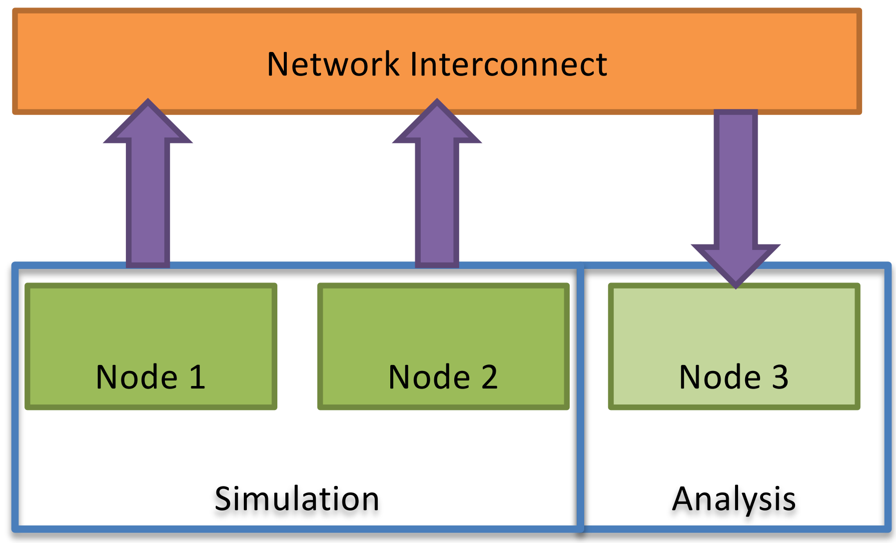


Extending HPC to Integrate Data Analytics



Augmenting HPC with In Situ and *In Transit* Analytics



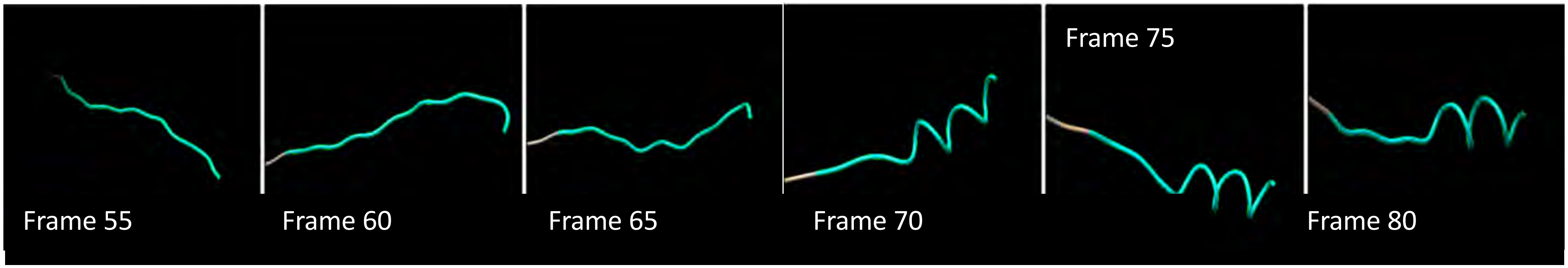


Example of tools:

- DataSpaces (Rutgers U.)
- DataStager (GeorgiaTech)

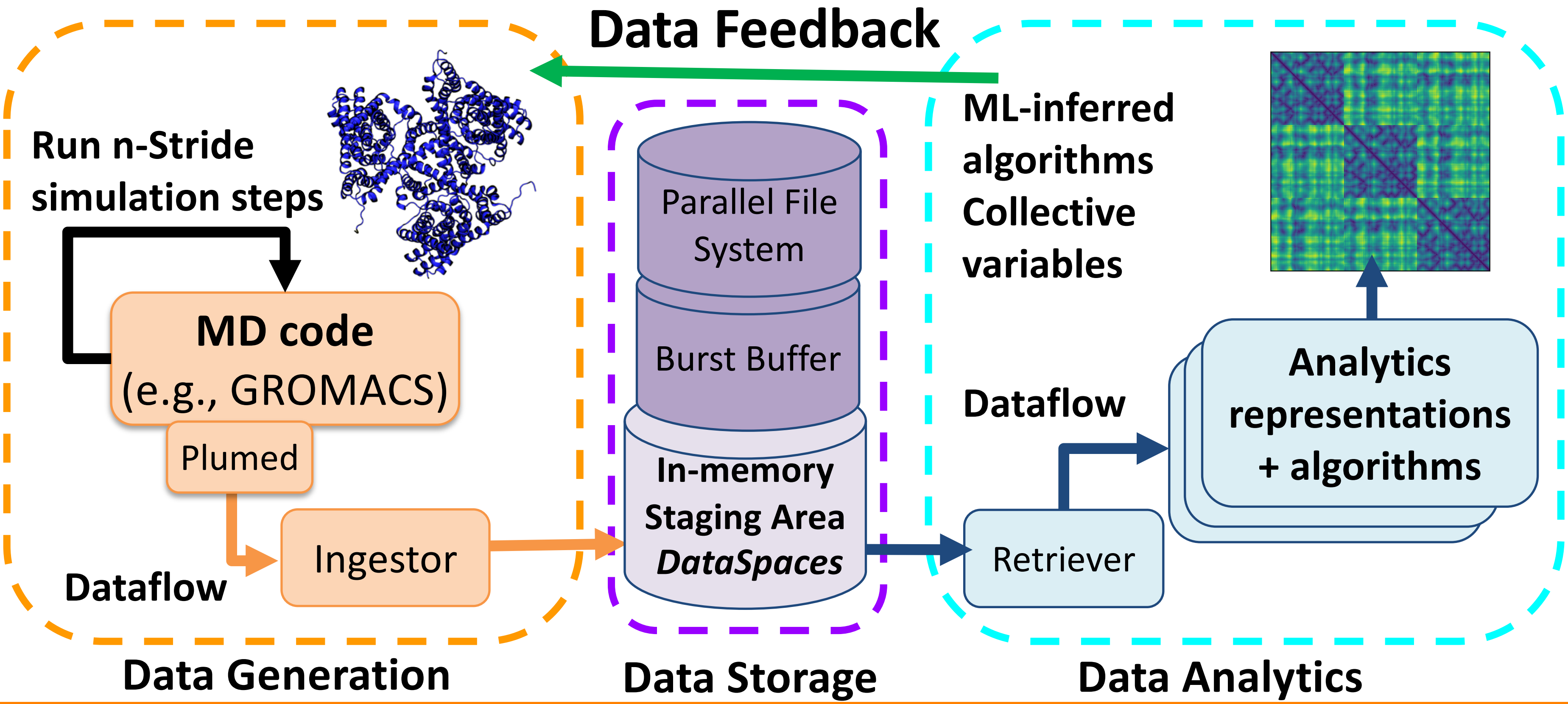
In Situ and *In Transit* Analytics for MD Simulations

Frames (or snapshots) of an MD trajectory with a stride of 5 steps:

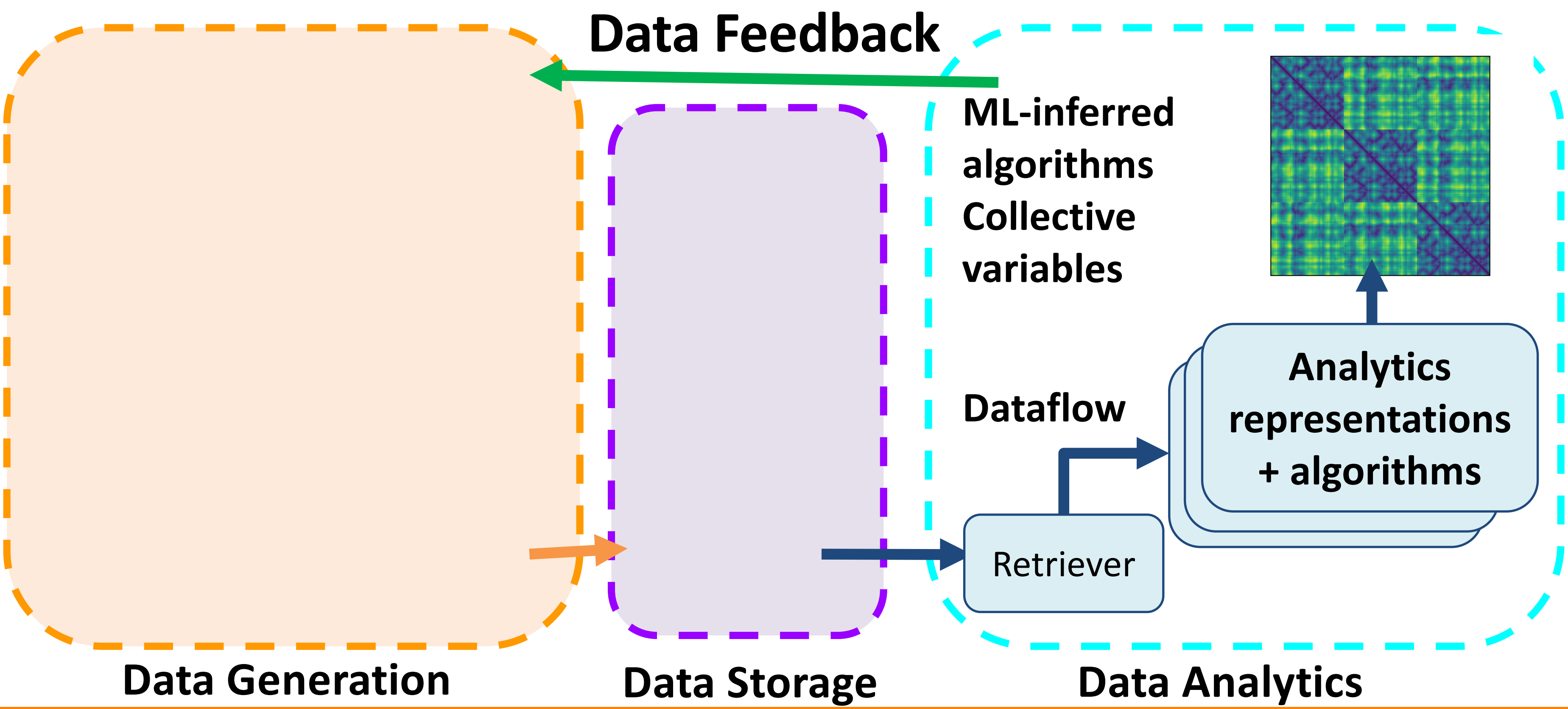


- We want to capture what is going on in each frame **without**:
  - Disrupting the simulation (e.g., stealing CPU and memory on the node)
  - Moving all the frames to a central file system and analyzing them once the simulation is over
  - Comparing each frame with past frames of the same job
  - Comparing each frame with frames of other jobs

Building a Closed-loop Workflow



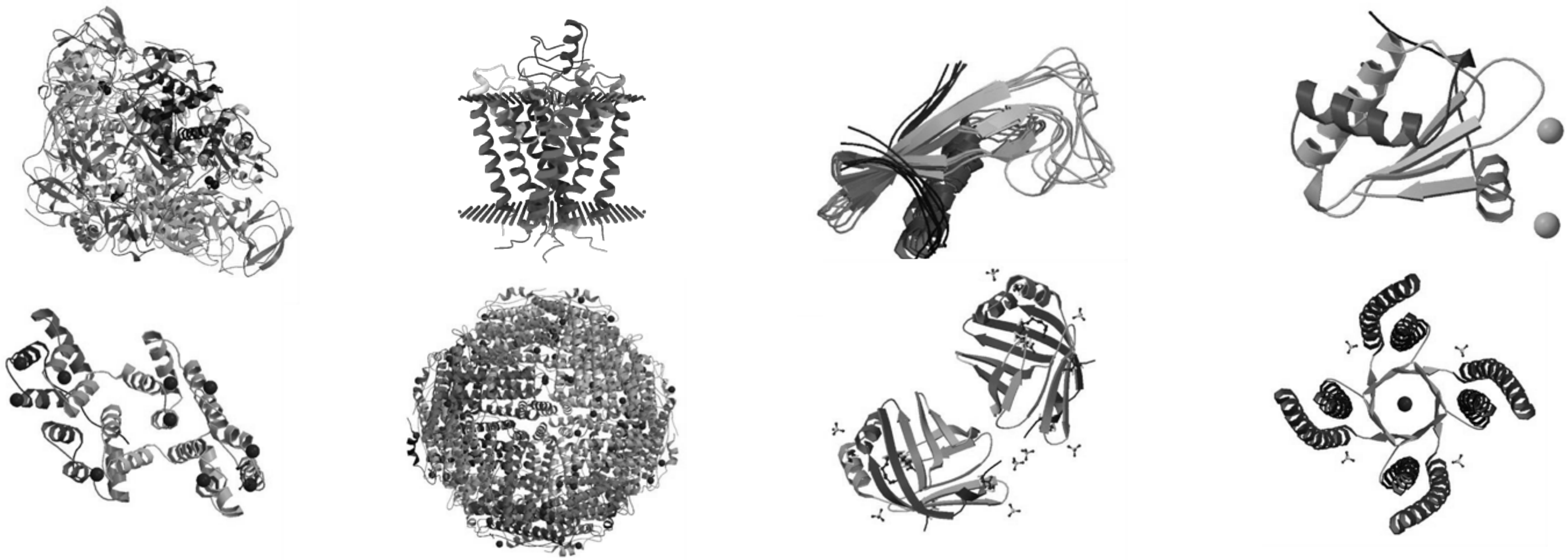
Building a Closed-loop Workflow



Proteins with Similar Functions

Key principle: proteins with similar structures have similar functions

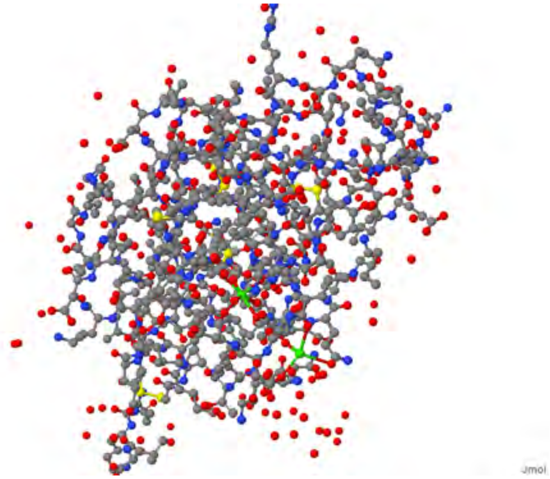
- Measure millions of protein variants expressed from yeast or bacteria
- Structure proteins to produce desired properties (protein engineering)



T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer. ***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

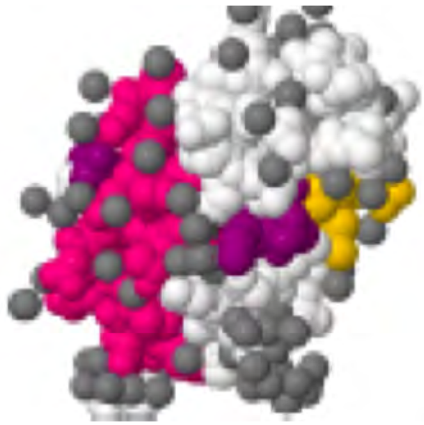
BIG ORANGE BIG IDEAS

Protein Representations



3D Cartesian representation

Multi-fold representation

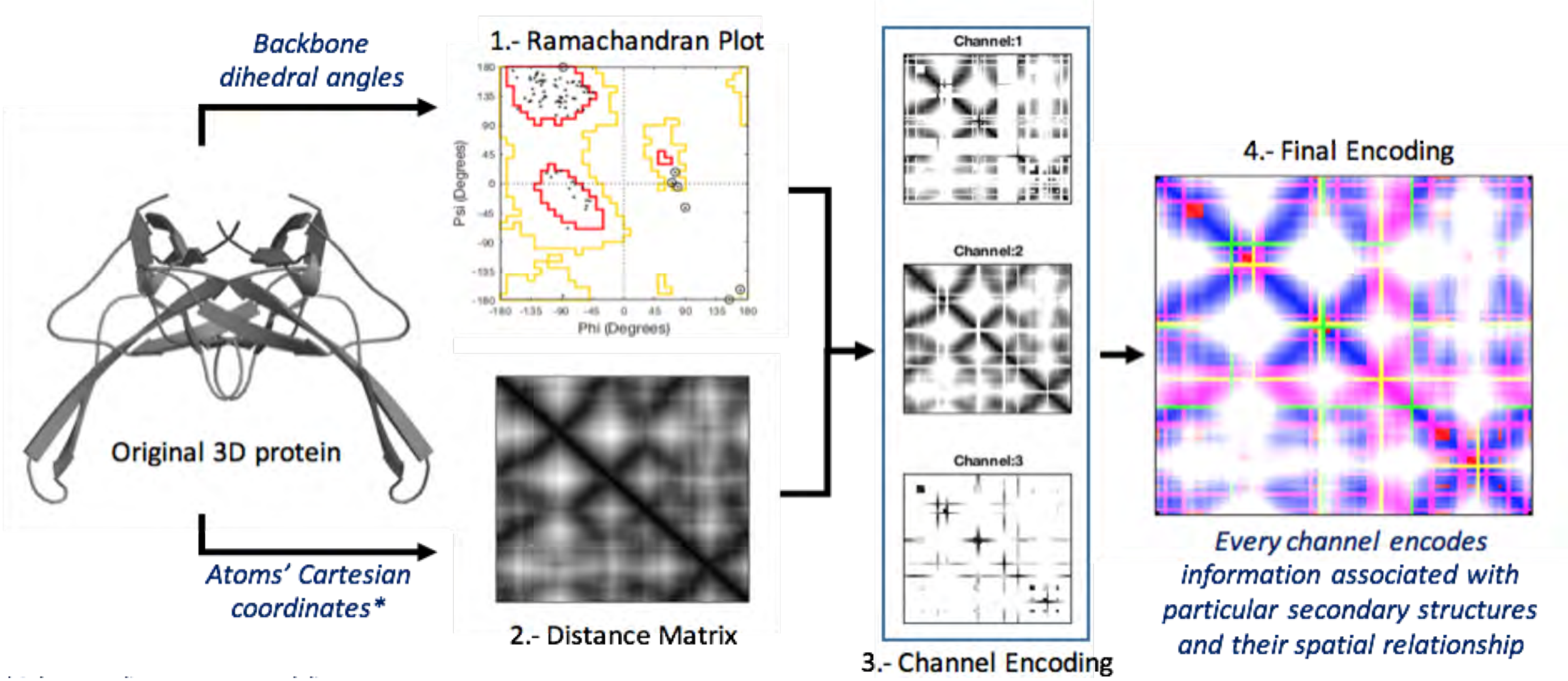


Surface representation

T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer. ***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

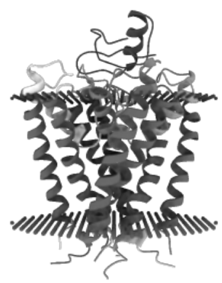
BIG ORANGE BIG IDEAS

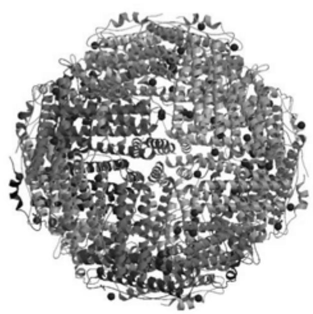
From Multi-fold Representation to Image Encoding

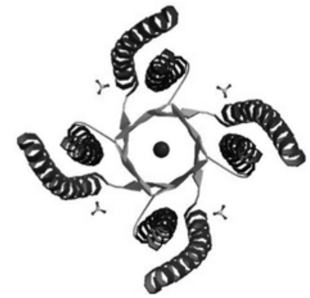


T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer. ***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

From Multi-fold Representation to Image Encoding

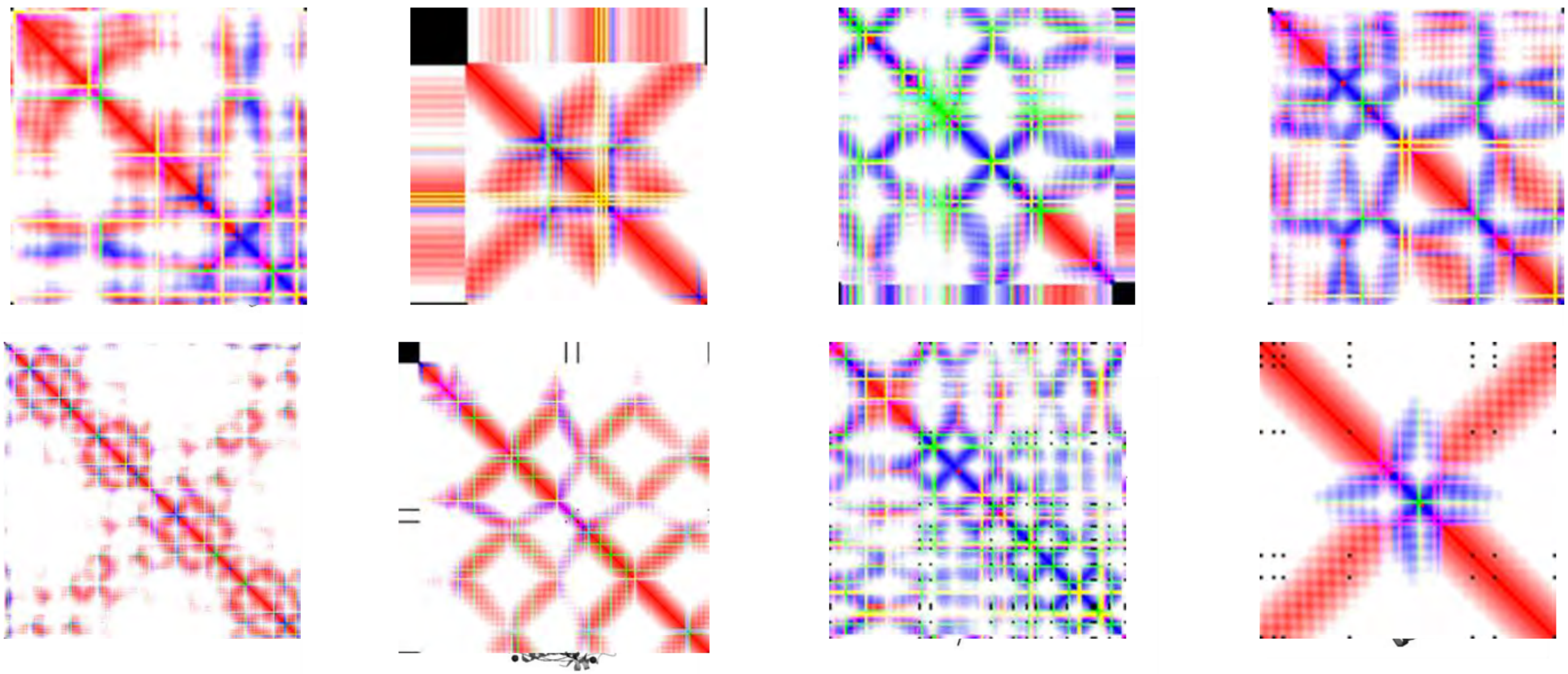






T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer.
Graphic Encoding of Proteins for Efficient High-Throughput Analysis. *ICPP 2018.*

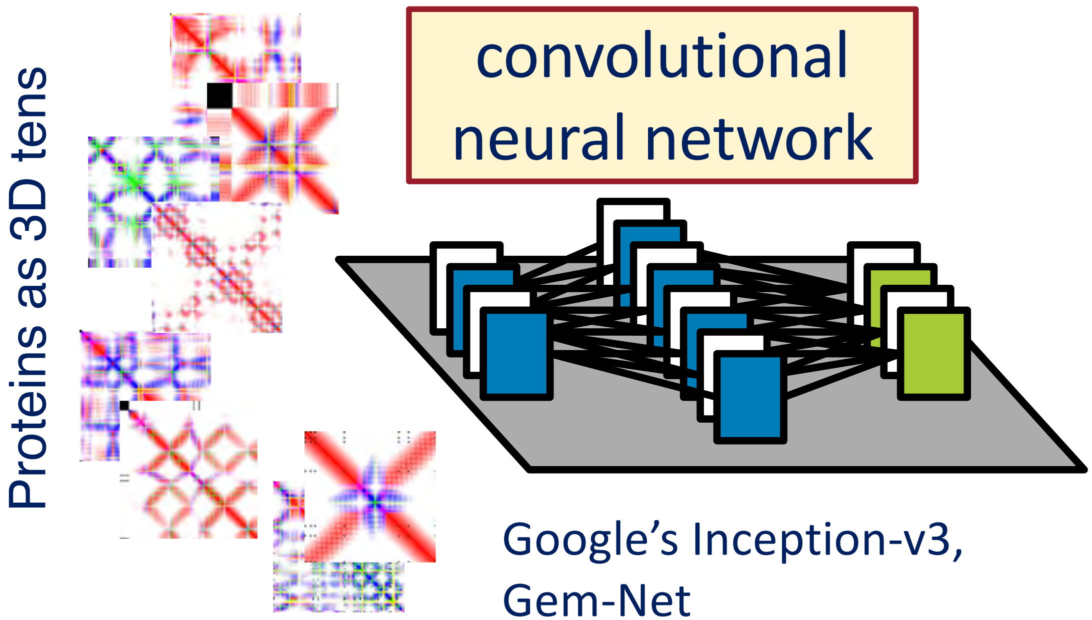
From Multi-fold Representation to Image Encoding



T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer.
Graphic Encoding of Proteins for Efficient High-Throughput Analysis. *ICPP 2018.*

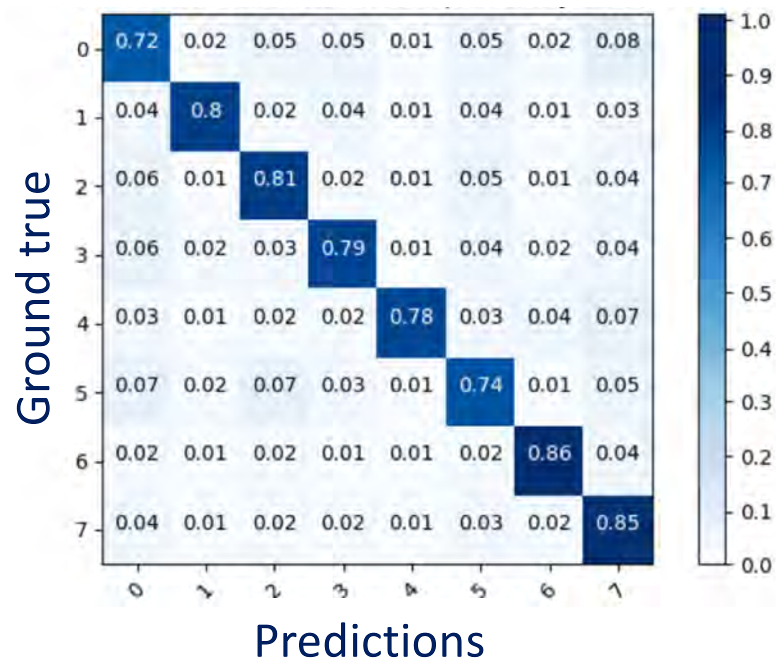
High-Throughput Protein Analysis

- Eight biological processes from biological process taxonomy in RCSB-PDB
- 62,991 proteins from the PDB



Google's Inception-v3,
Gem-Net

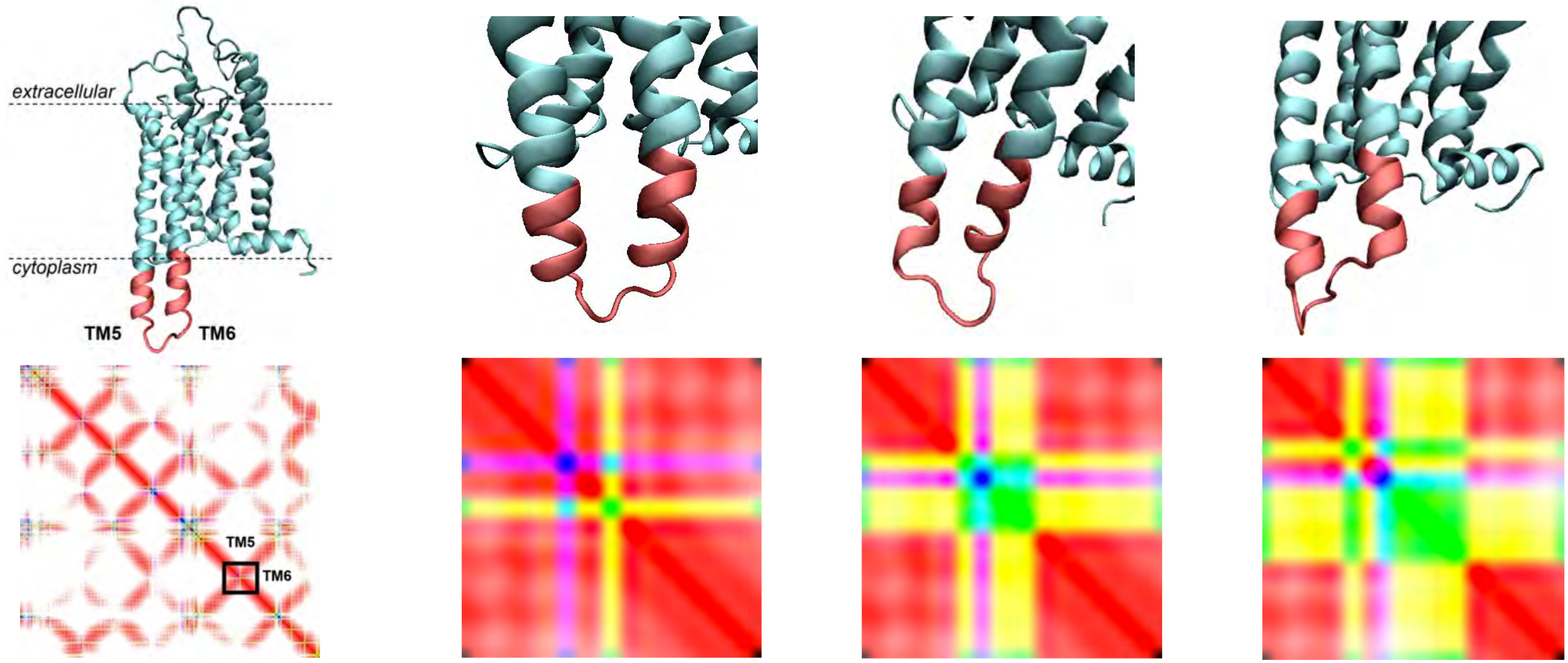
Normalized Confusion Matrix - **Accuracy 80.66%**



T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer. ***Graphic Encoding of Proteins for Efficient High-Throughput Analysis***. *ICPP 2018.*

Capturing Changes in Folding with Transfer Learning

Protein: Opsin | Frame 50 | Frame 1500 | Frame 1950

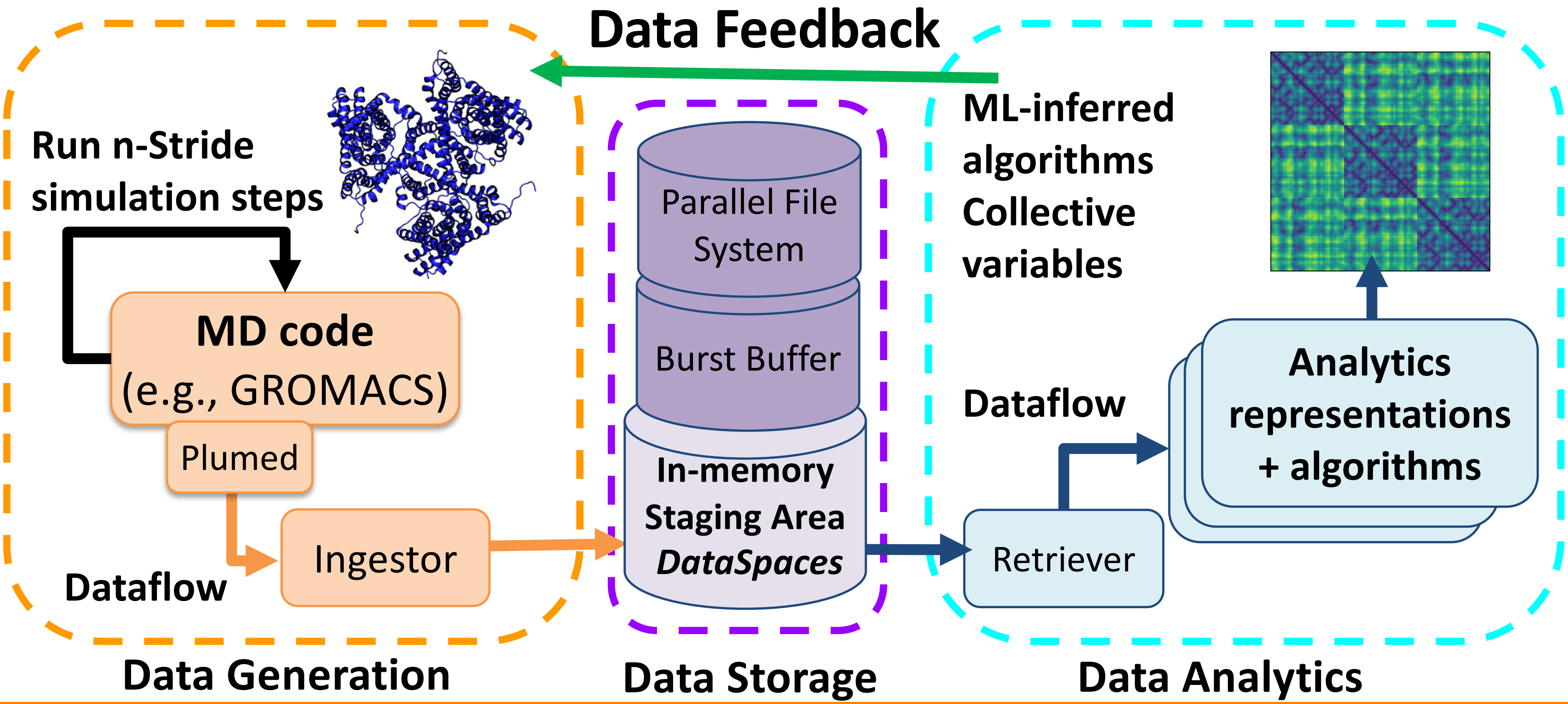


T. Estrada,et al. **A Graphic Encoding Method for Quantitative Classification of Protein Structure and Representation of Conformational Changes**. 2019

Two Use Cases

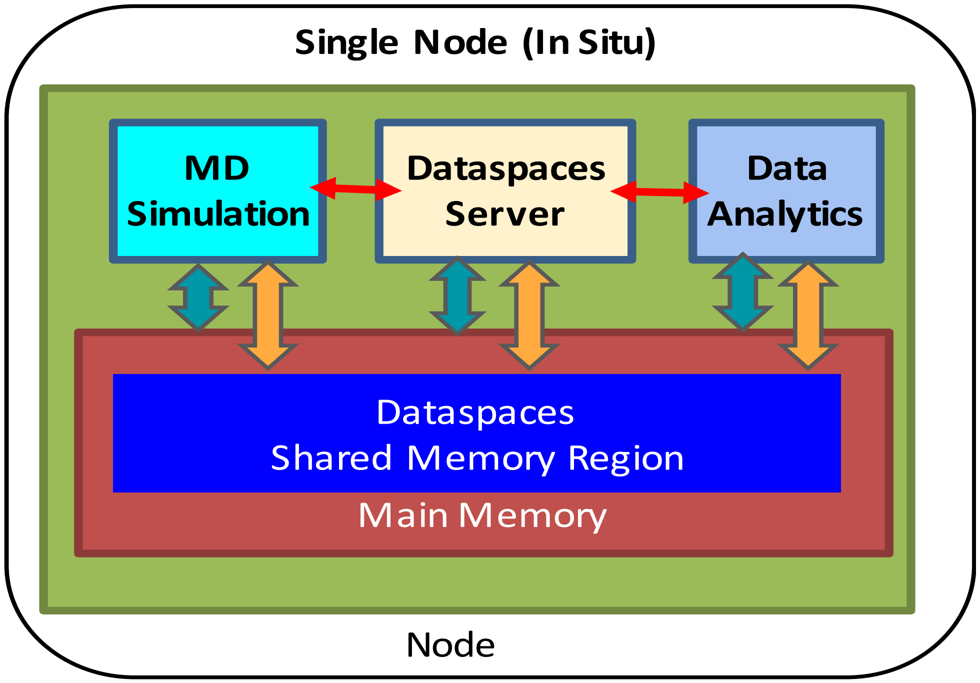
- Extending HPC to integrate data analytics
  - Next generation MD workflows
  - Molecular structures
  - ***Data transformation – i.e.,*** *capturing information*
  - ***Dataflow modeling –*** *i.e., lost information*
- Extending HPC to connect to the "Edge"
  - Next generation precision farming
  - Soil moisture data
  - ***Data prediction*** *– i.e., from coarse- to fine-grained information*

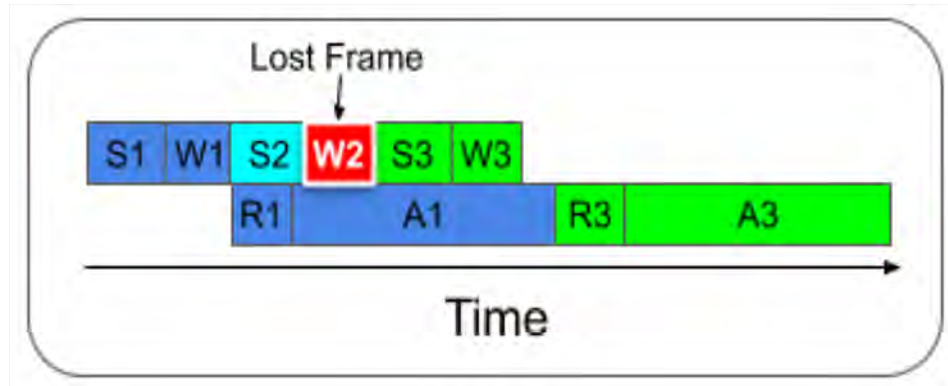
Building a Closed-loop Workflow



BIG ORANGE BIG IDEAS

Modeling Lost Frames

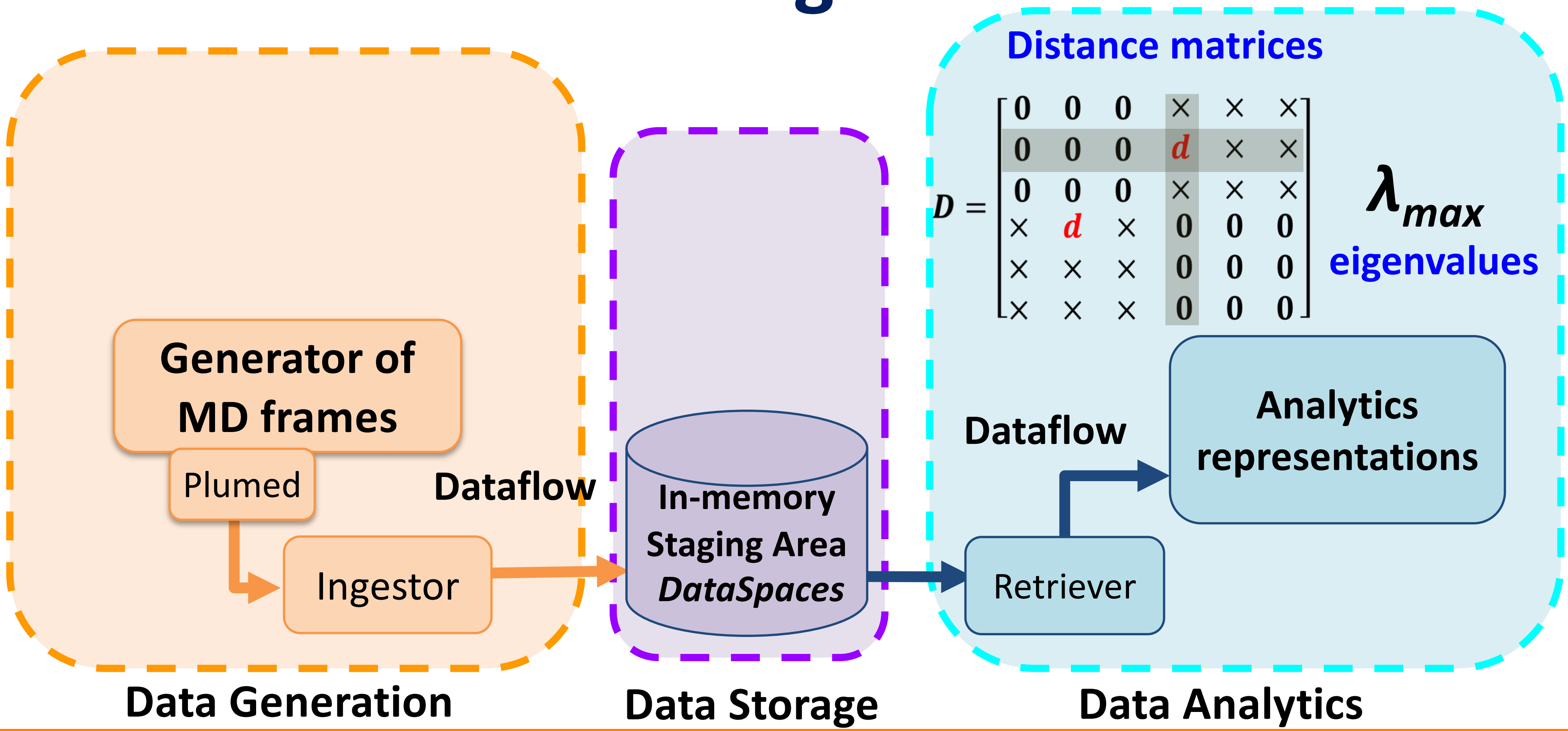




S1, S2, S3: generate MD frame
W1, **W2**, W2: write to shared memory
R1, R2, R3: read from shared memory
A1, A3: analyze frame

S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019.

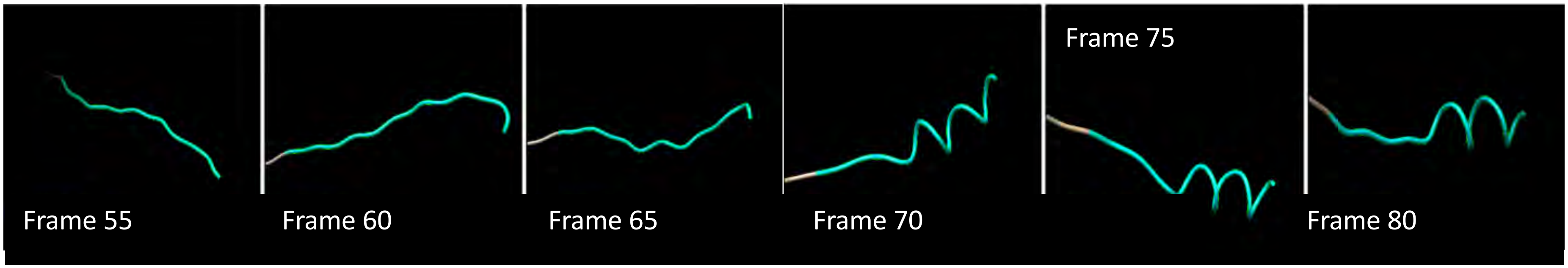
Dataflow for Modeling Lost Frames



T. Johnston et al. In-Situ Data Analytics and Indexing of Protein Trajectories. *Journal of Computational Chemistry (JCC),* 38(16):1419-1430, 2017.

Eigenvalues: Proxy for Structural Changes

Frames (or snapshots) of an MD trajectory with a stride of 5 steps:



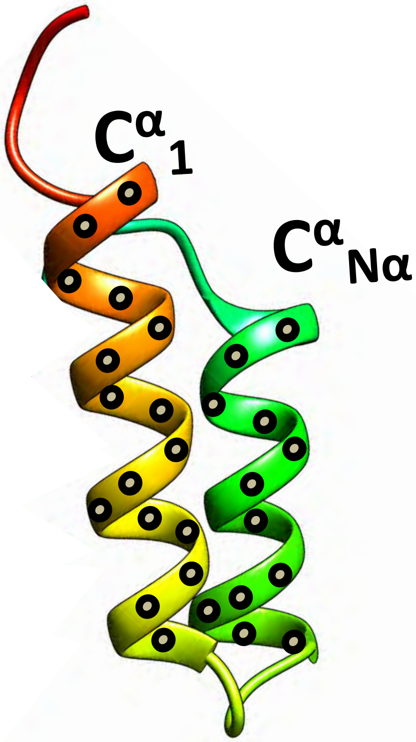
λmax_{55} λmax_{60} λmax_{65} λmax_{70} λmax_{75} λmax_{80}

The distance between two max eigenvalues can serve as a proxy for distance between the two associated conformations

T. Johnston et al. In-Situ Data Analytics and Indexing of Protein Trajectories. *Journal of Computational Chemistry (JCC),* 38(16):1419-1430, 2017.

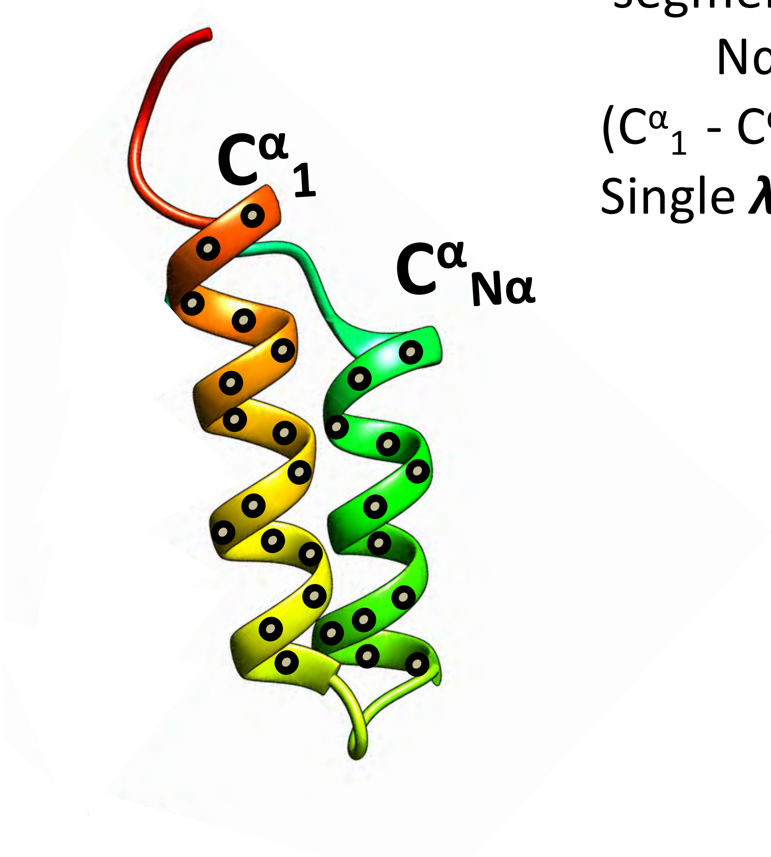
Single frame at time *t*

Nα C^{α} atoms



S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019.

Single frame at time *t*
$N\alpha$ C^{α} atoms



Distance of two segments with segment length:

$N\alpha/2 \times C^{\alpha}$ atoms

$(C^{\alpha}_{1} - C^{\alpha}_{N\alpha/2\,-1})\ (C^{\alpha}_{N\alpha/2} - C^{\alpha}_{N\alpha/2})$

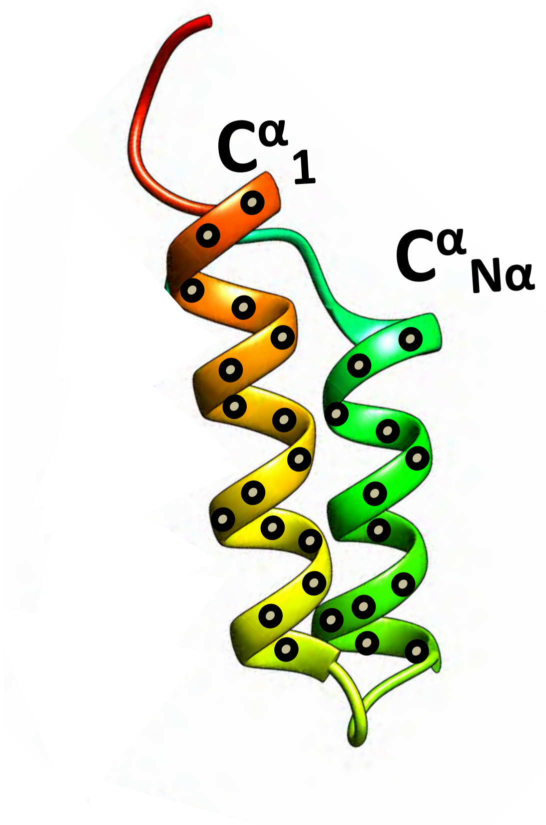
Single $\boldsymbol{\lambda_{max}}$

$$
\begin{array}{c|cc}
 & C^{\alpha}_{1} & C^{\alpha}_{N\alpha} \\
\hline
C^{\alpha}_{1} \ \ldots \ C^{\alpha}_{N\alpha/2\,-\,1} & 0 & d_{ij} \\
C^{\alpha}_{N\alpha/2} \ \ldots \ C^{\alpha}_{N\alpha} & dji & 0
\end{array}
$$

S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019.

Single frame at time *t*

$N\alpha$ C^{α} atoms

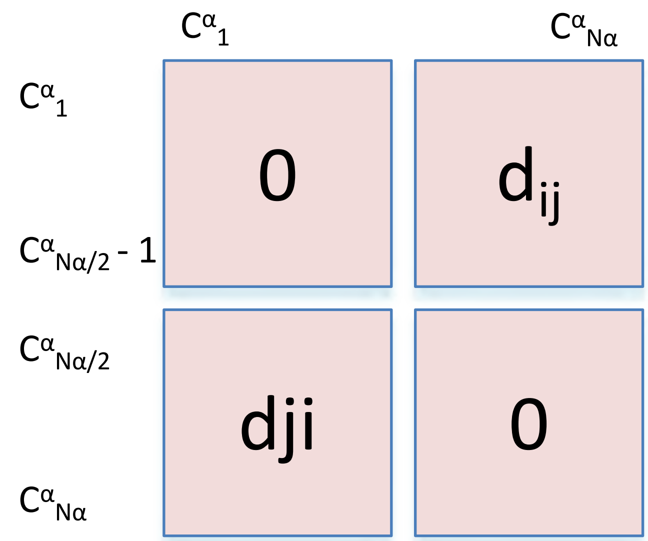


Distance of two segments with segment length:

$N\alpha/2 \times C^{\alpha}$ atoms

$(C^{\alpha}_{1} - C^{\alpha}_{N\alpha/2-1})\ (C^{\alpha}_{N\alpha/2} - C^{\alpha}_{N\alpha/2})$

Single $\boldsymbol{\lambda}_{max}$



Distances of $N\alpha/2$ segments with segment length: $2 \times C^{\alpha}$ atoms

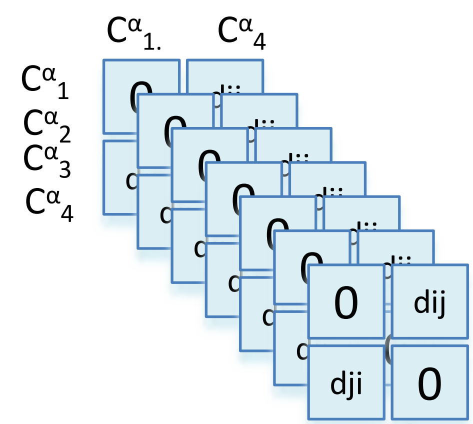
$(C^{\alpha}_{1} - C^{\alpha}_{2})\ (C^{\alpha}_{3} - C^{\alpha}_{4})$

$(C^{\alpha}_{5} - C^{\alpha}_{6})\ (C^{\alpha}_{7} - C^{\alpha}_{8})$

....

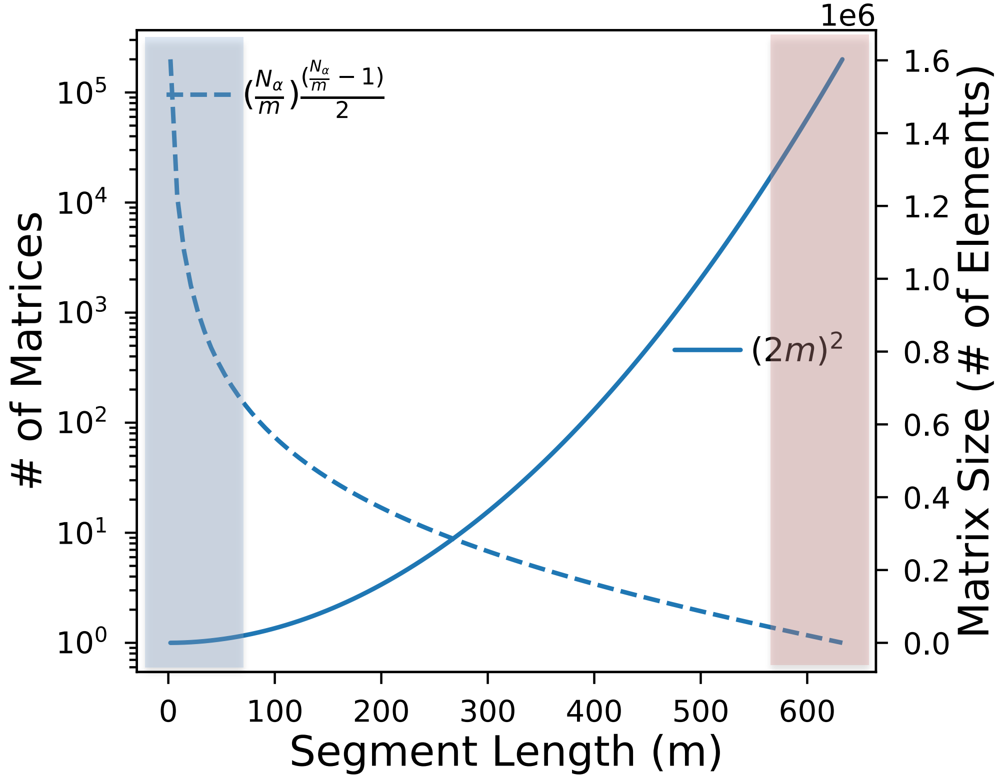
$(C^{\alpha}_{N\alpha/2-3} - C^{\alpha}_{N\alpha/2-2})\ (C^{\alpha}_{N\alpha/2-1} - C^{\alpha}_{N\alpha/2})$

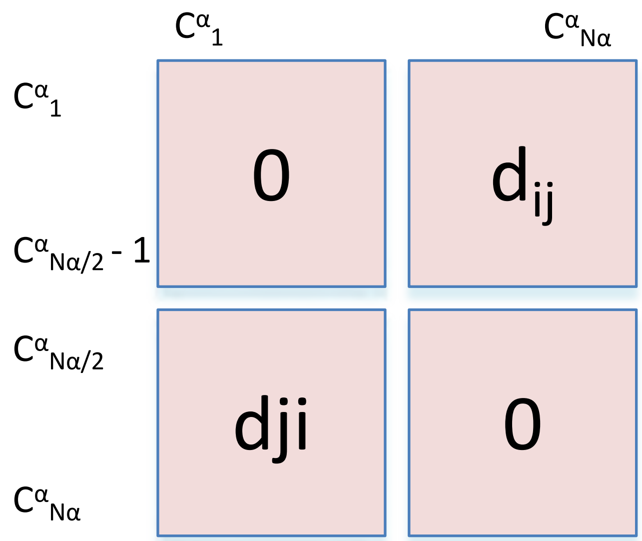
$\boldsymbol{\lambda}_{max,\,1}\ \boldsymbol{\lambda}_{max,\,2} \ldots\ \boldsymbol{\lambda}_{max,\,N\alpha/2}$

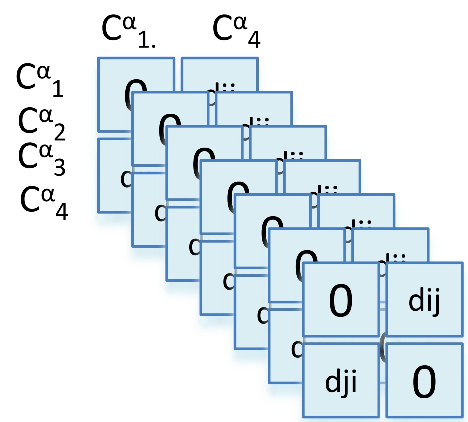


S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019.

Segment size = proxy of **number of matrices** and **matrix sizes**







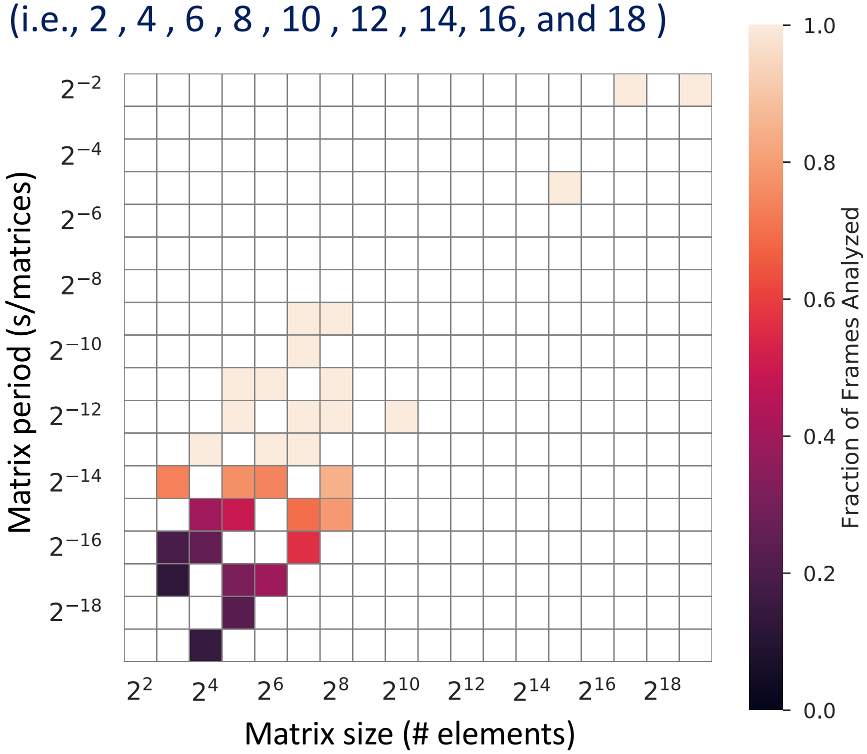
Many small matrices

Few large matrices

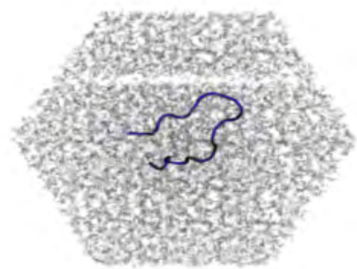
S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019.

2-step Model: Fraction of Analyzed Frames

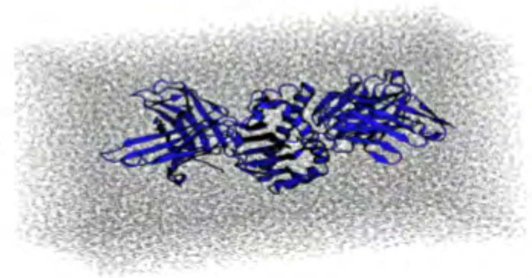
Observables: small segment lengths
(i.e., 2 , 4 , 6 , 8 , 10 , 12 , 14, 16, and 18)



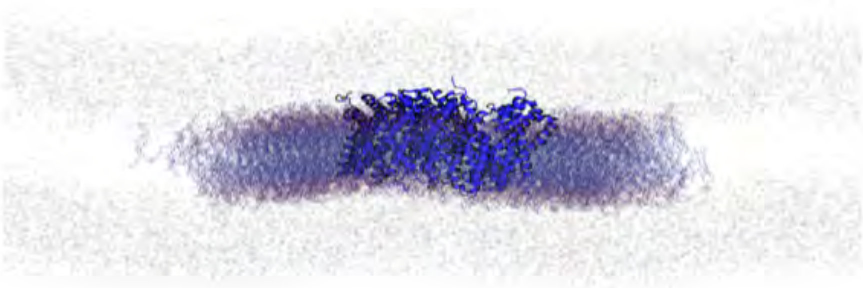
Trp cage 12,619 atoms



T cell receptor 81,092 atoms



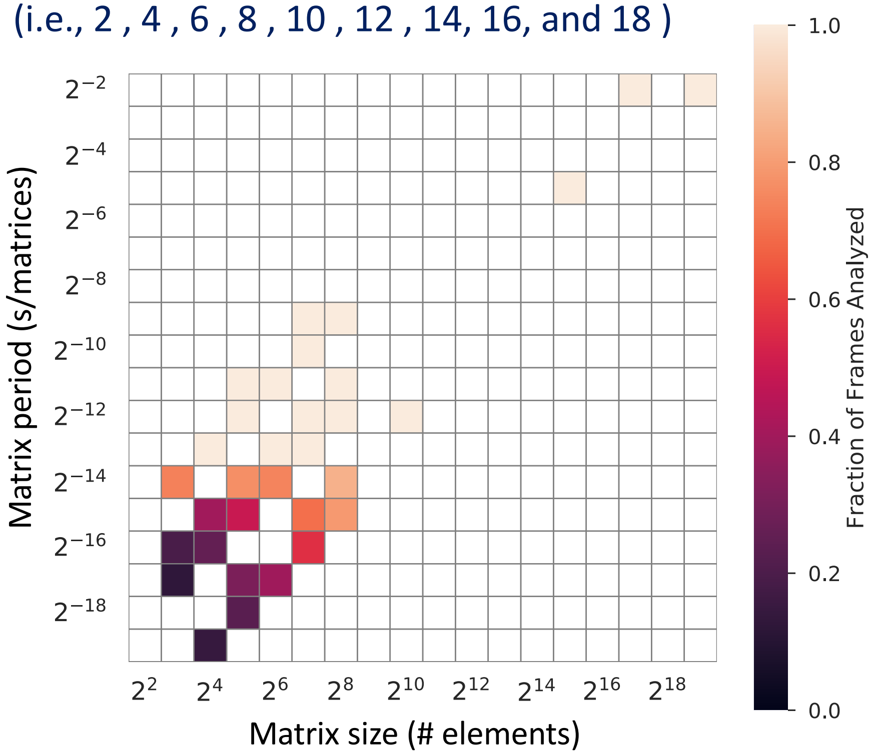
Gltph 270,088 atoms



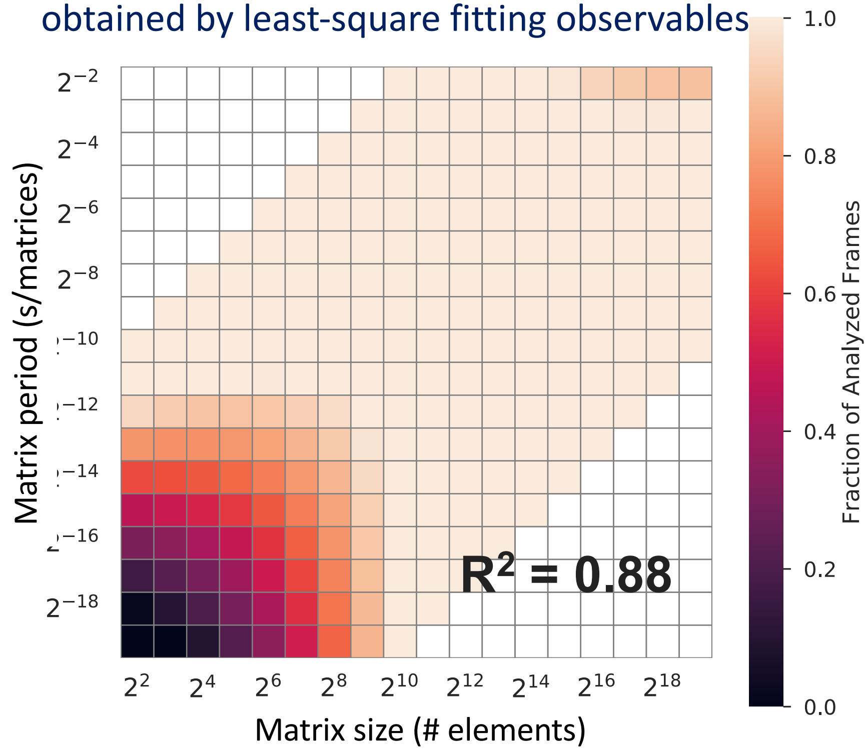
S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019.

2-step Model: Fraction of Analyzed Frames

Observables: segment lengths
(i.e., 2 , 4 , 6 , 8 , 10 , 12 , 14, 16, and 18)



Model: polynomial model of degree 2
obtained by least-square fitting observables

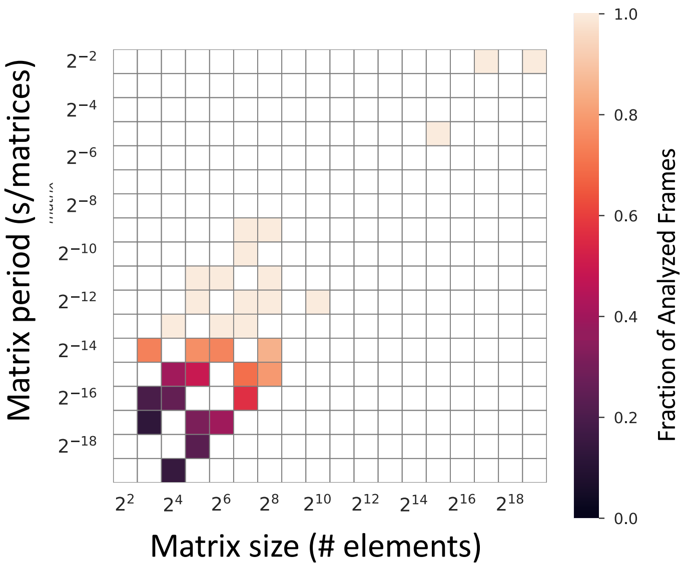


S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019. *(Submitted)*

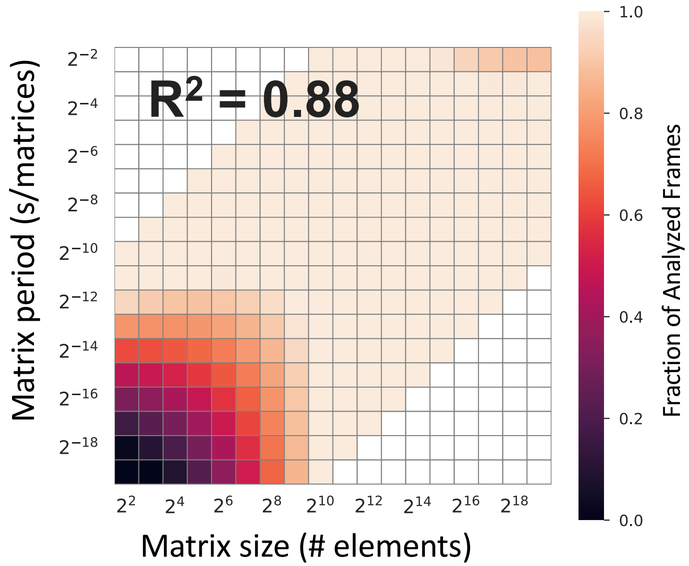
BIG ORANGE BIG IDEAS

2-step Model: Fraction of Analyzed Frames

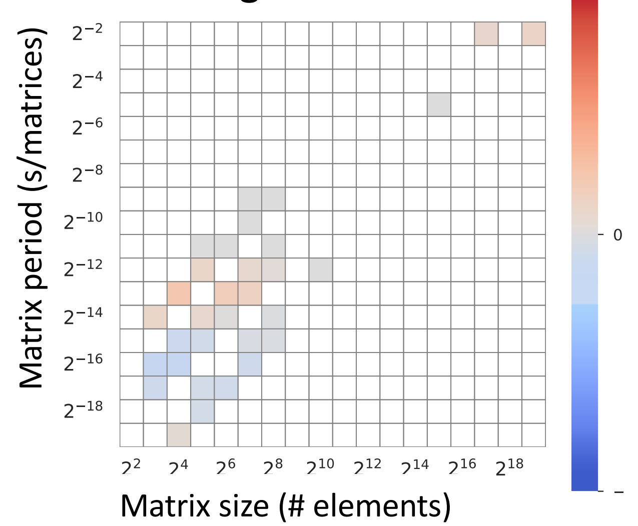
Observables



Degree 2 polynomial model

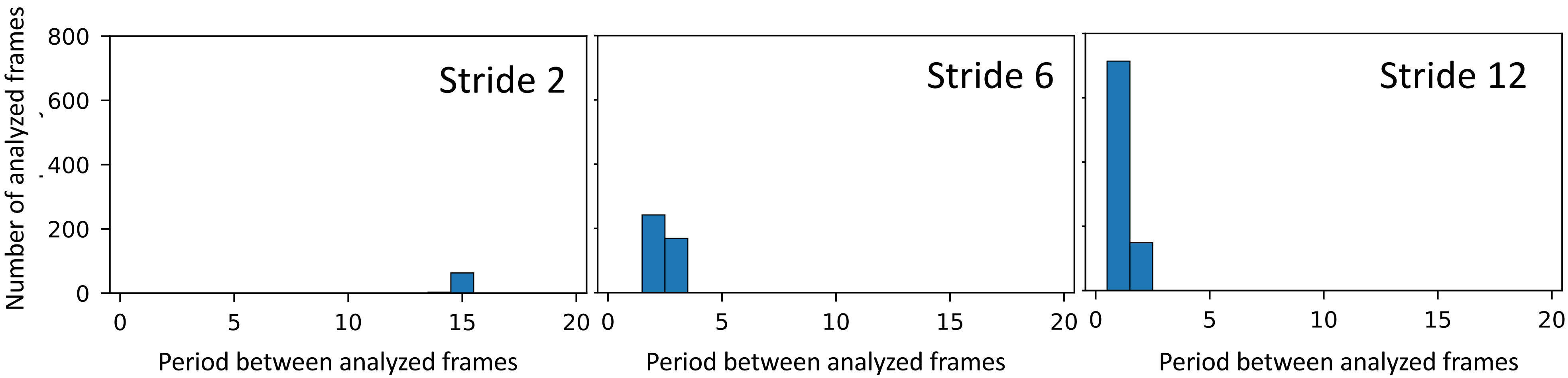


Absolute error between data and fitting model



2-step Model: Frames Distribution

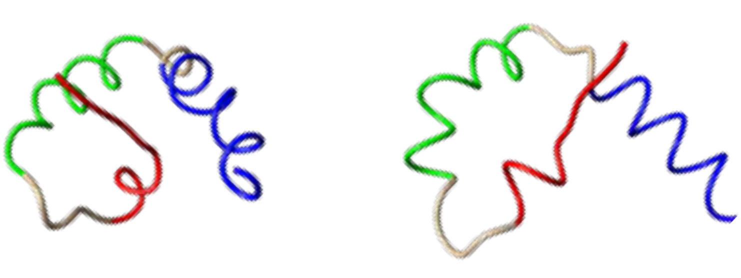
- Given a trajectory, we model the proportions *p* and *q* of analyzed frames (*f*) with periods *k* and *k+1*
  - Example: Gltph (27,000 atoms and TPS 318), trajectory of 1,000 frames.



S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019.

Case Study: 1BDD Protein Conformations

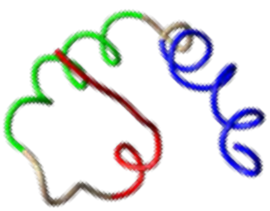
Frame 1330 Frame 1360 Frame 1390



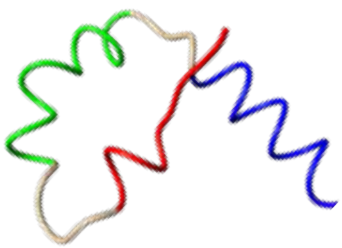
S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019. *(Submitted)*

Case Study: 1BDD Protein Conformations

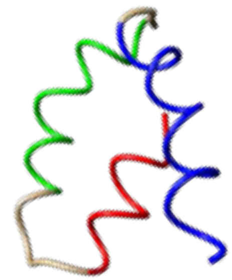
Frame 1330



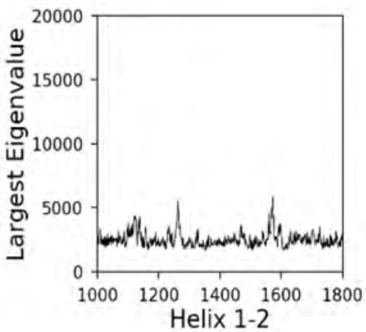
Frame 1360



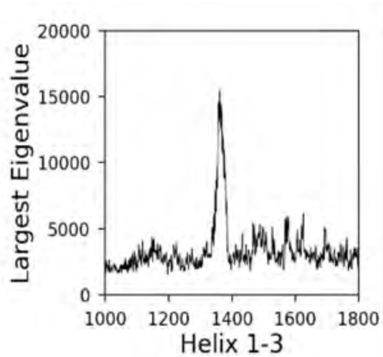
Frame 1390



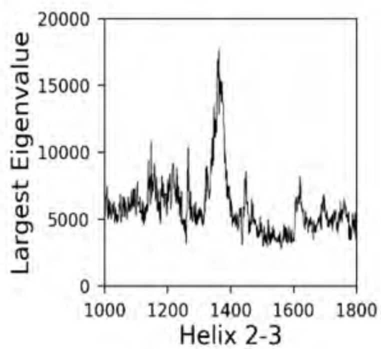
Helix 1-2: $\boldsymbol{\lambda}_{max}$



Helix 1-3: $\boldsymbol{\lambda}_{max}$



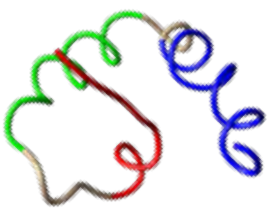
Helix 2-3: $\boldsymbol{\lambda}_{max}$



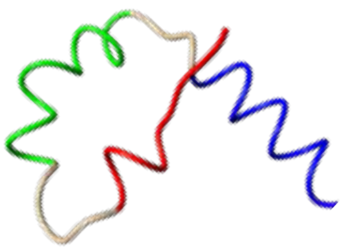
S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019. *(Submitted)*

Case Study: 1BDD Protein Conformations

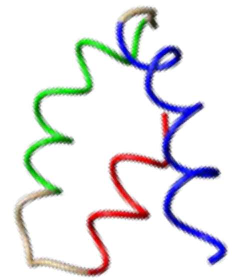
Frame 1330

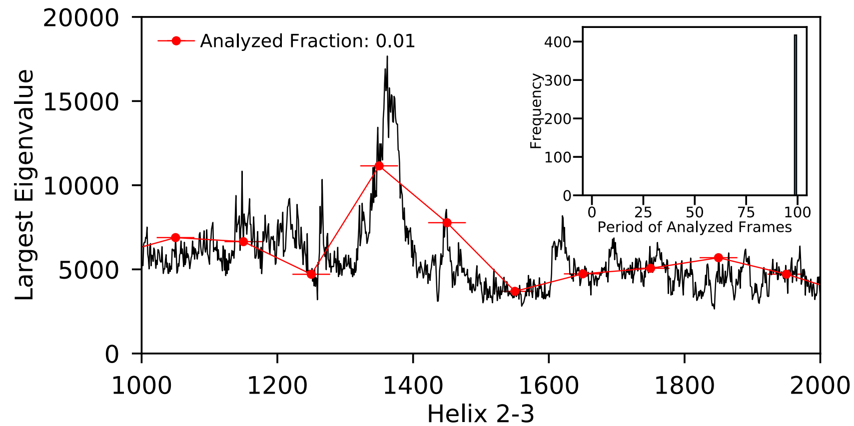


Frame 1360

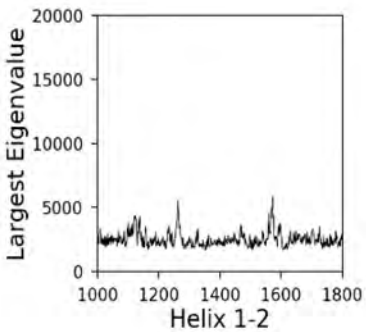


Frame 1390

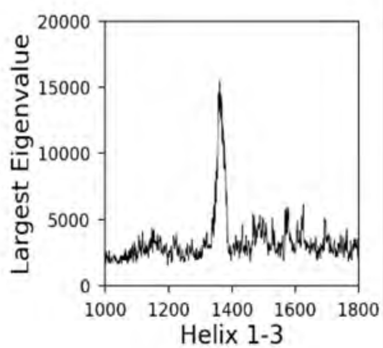




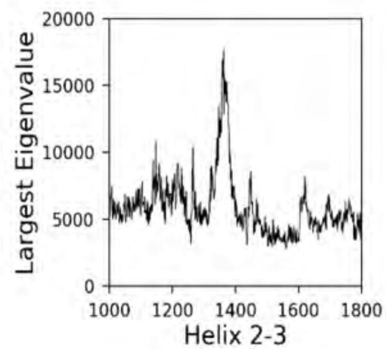
Helix 1-2: $\boldsymbol{\lambda}_{max}$

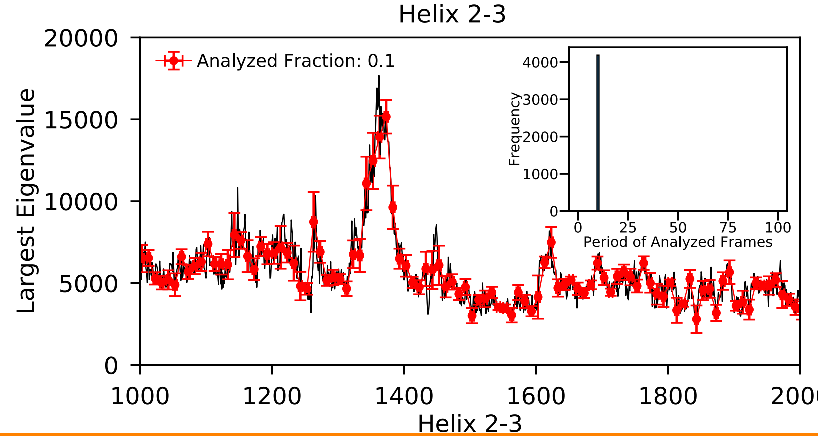


Helix 1-3: $\boldsymbol{\lambda}_{max}$



Helix 2-3: $\boldsymbol{\lambda}_{max}$





S. Thomas et al. Characterization of In Situ and In Transit Analytics of Molecular Dynamics Simulations for Next-generation Supercomputers. *eScience*, 2019. *(Submitted)*

Two Use Cases

- Extending HPC to integrate data analytics
  - Next generation MD workflows
  - Molecular structures
  - ***Data transformation – i.e.,*** *capturing information*
  - ***Dataflow modeling –*** *i.e., lost information*
- Extending HPC to connect to the "Edge"
  - Next generation precision farming
  - Soil moisture data
  - ***Data prediction*** *– i.e., from coarse- to fine-grained information*

Soil Moisture Data for Precision Farming

- Satellites collect raster data across the surface of the Earth

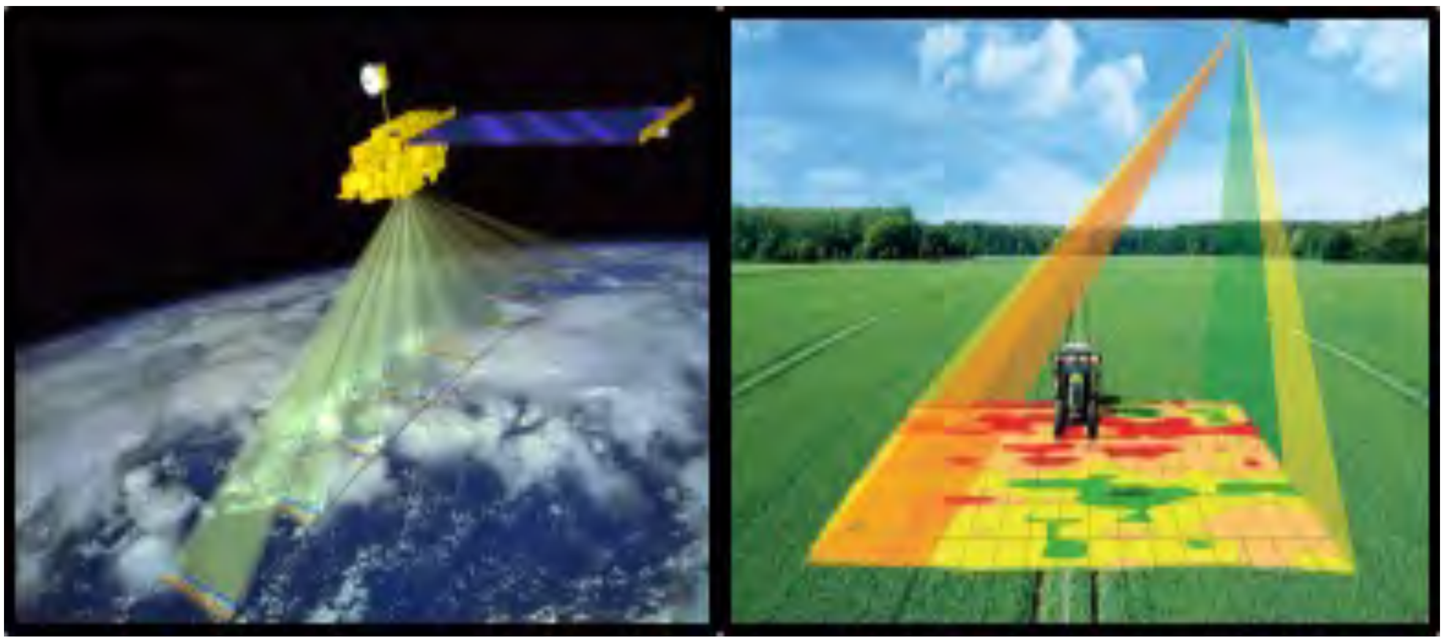


Image Source: http://www.esa-soilmoisture-cci.org/

Soil Moisture: Incomplete Data

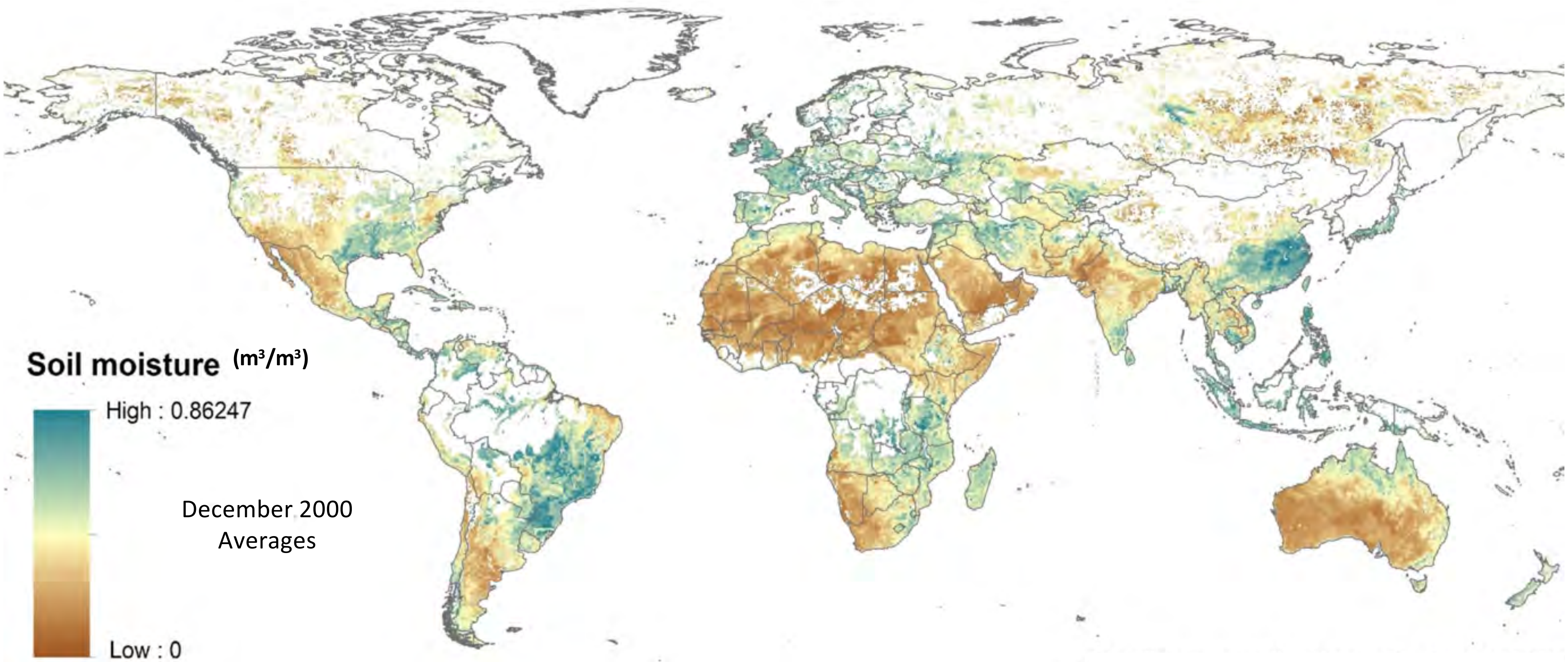
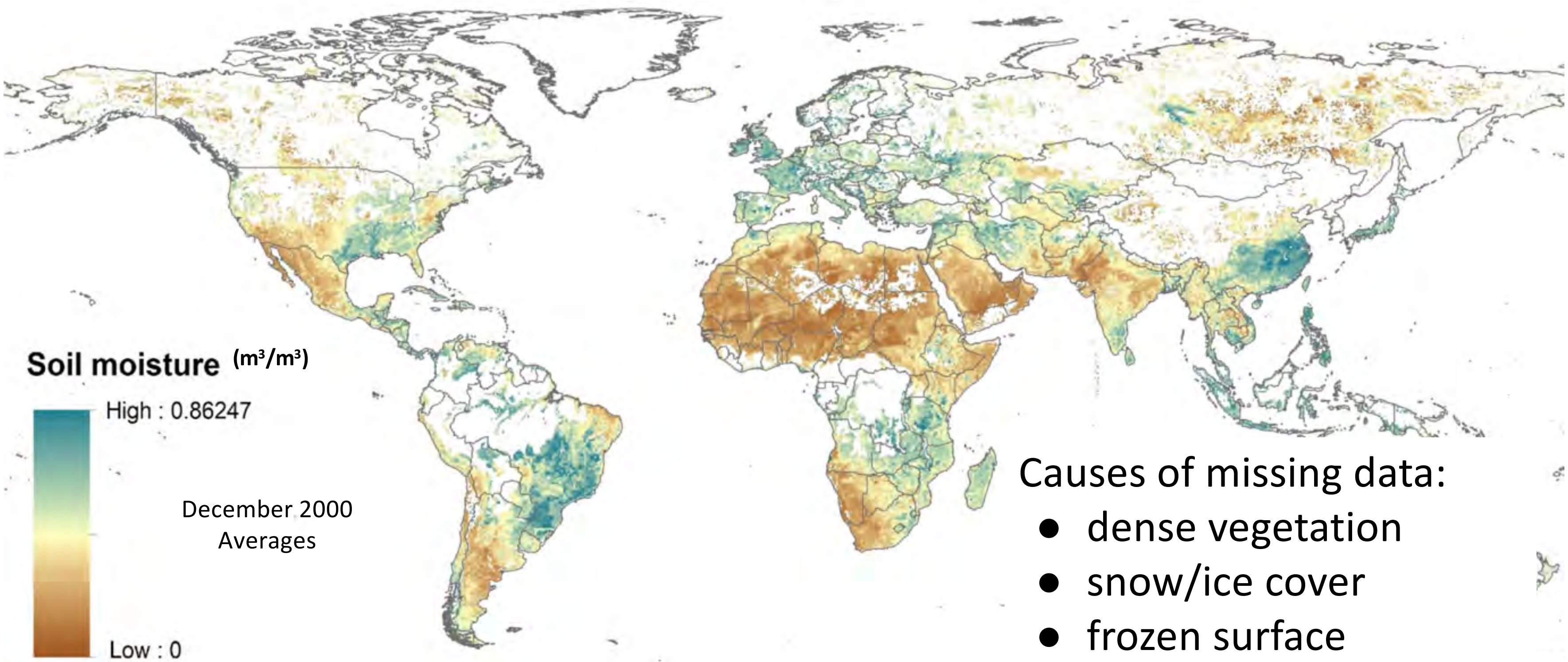


Image source: Ricardo Llamas, University of Delaware
Data source: ESA-CCI soil moisture database (http://www.esa-soilmoisture-cci.org/)

Soil Moisture: Incomplete Data

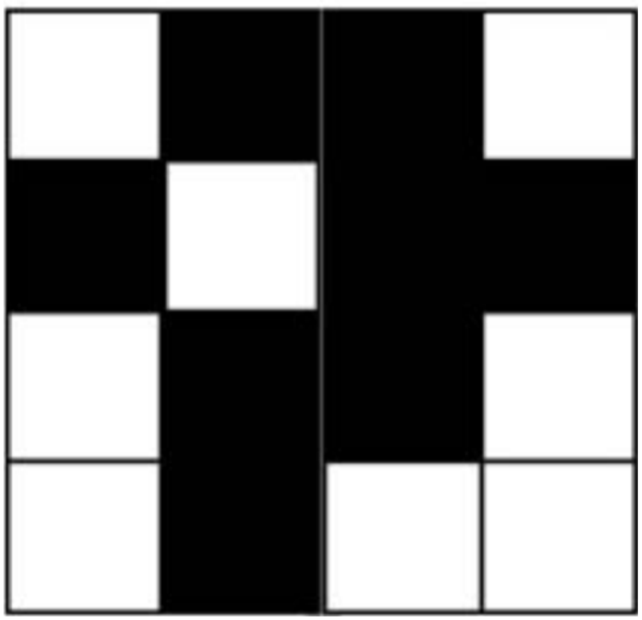


Causes of missing data:

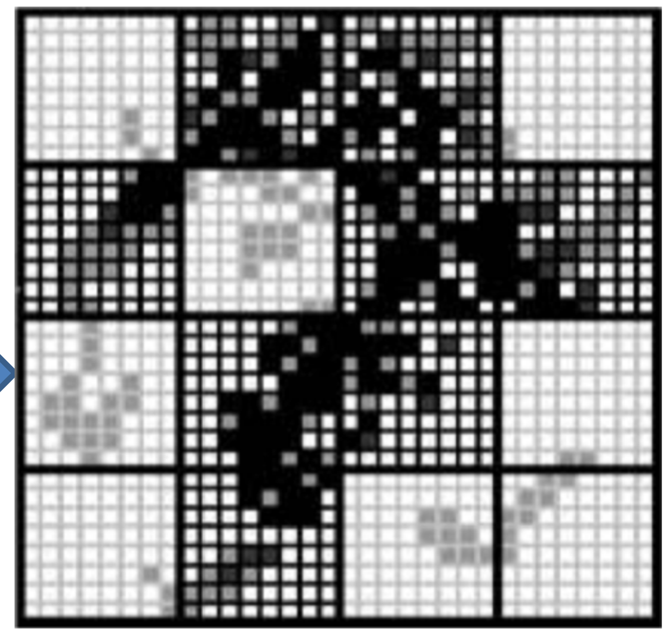
- dense vegetation
- snow/ice cover
- frozen surface
- extremely dry surface

Image source: Ricardo Llamas, University of Delaware
Data source: ESA-CCI soil moisture database (http://www.esa-soilmoisture-cci.org/)

Soil Moisture: Coarse-grained Data



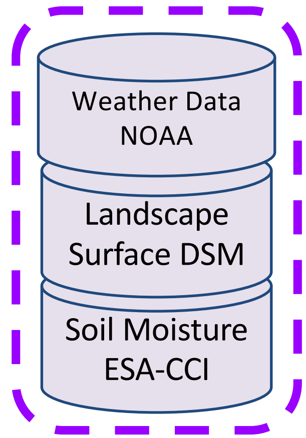
Original Resolution
27 km × 27 km



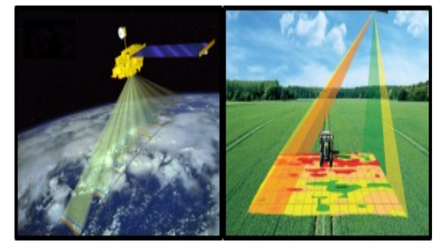
Desired Resolution
1 km × 1 km

Image source: McPherson et al., *Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions—possibilities and limitations*, Ecological Modeling **192**:499-522, 2006.

Building a Closed-loop Workflow



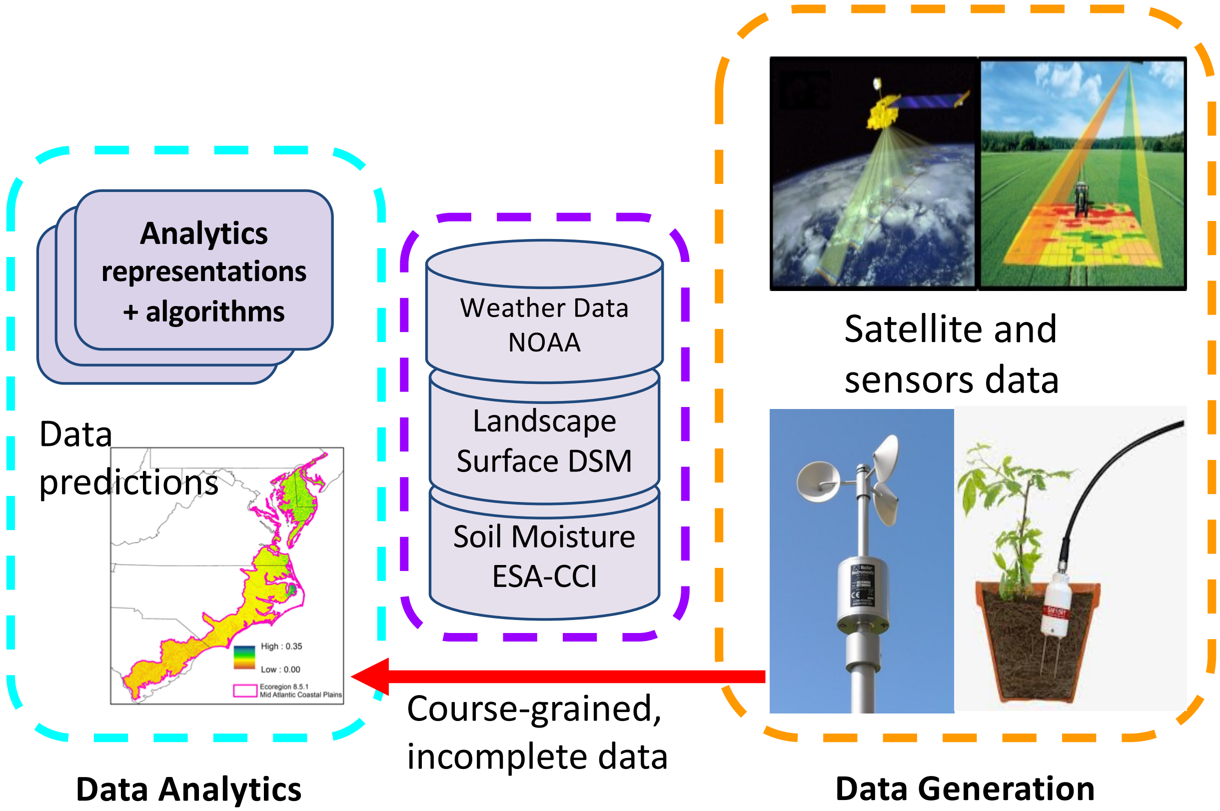
Course-grained, incomplete data



Satellite and sensors data

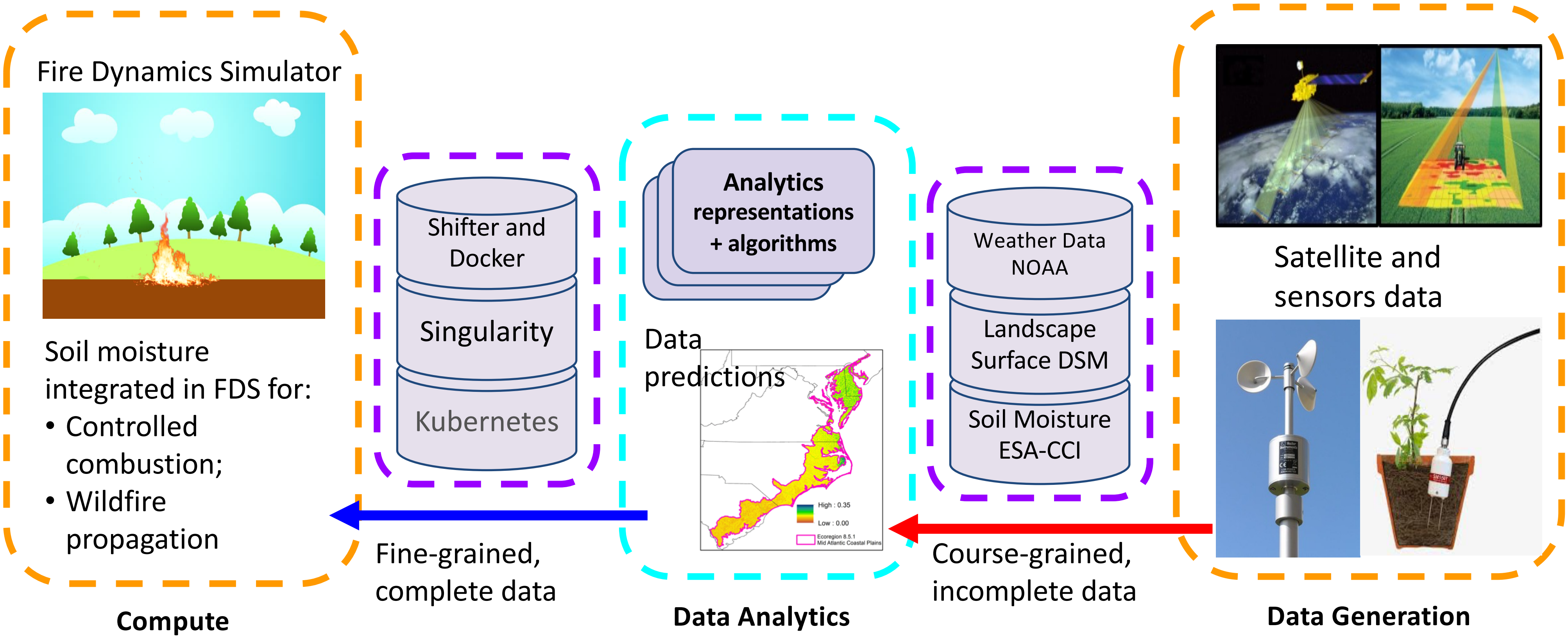
Data Generation

Building a Closed-loop Workflow

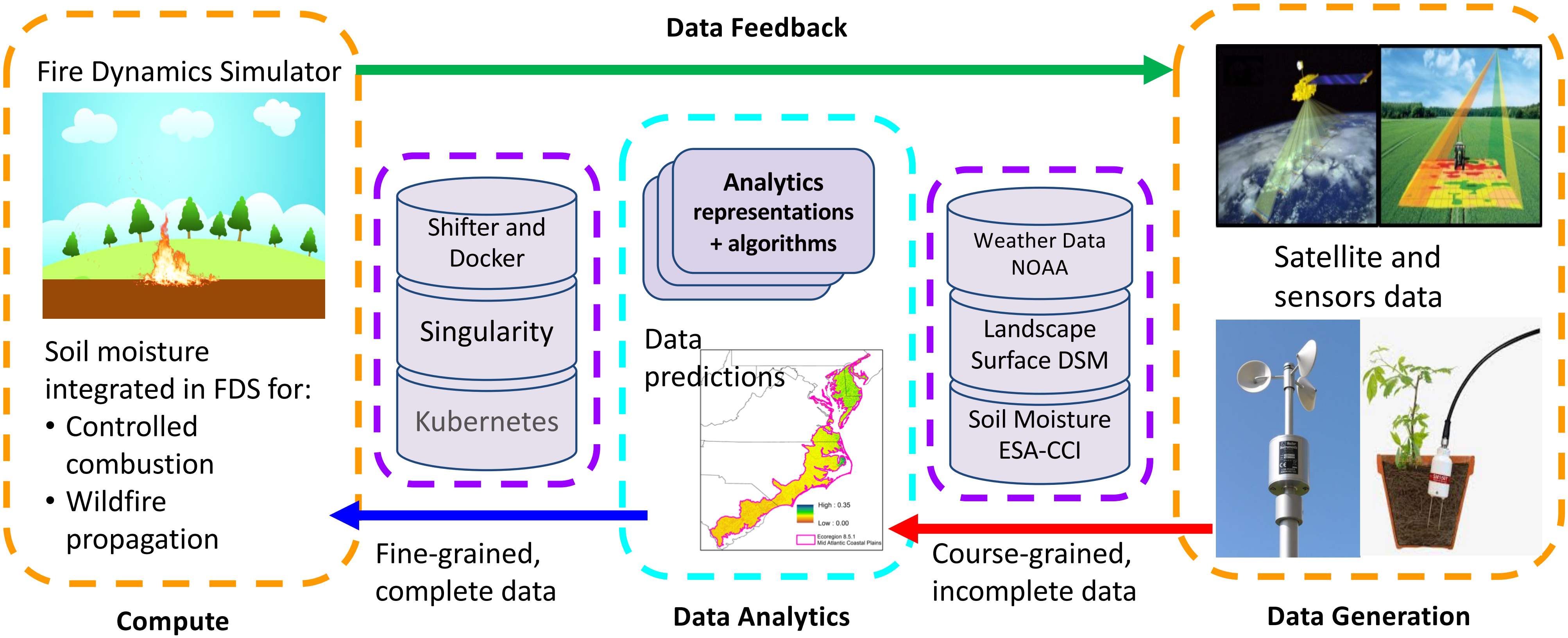


BIG ORANGE BIG IDEAS

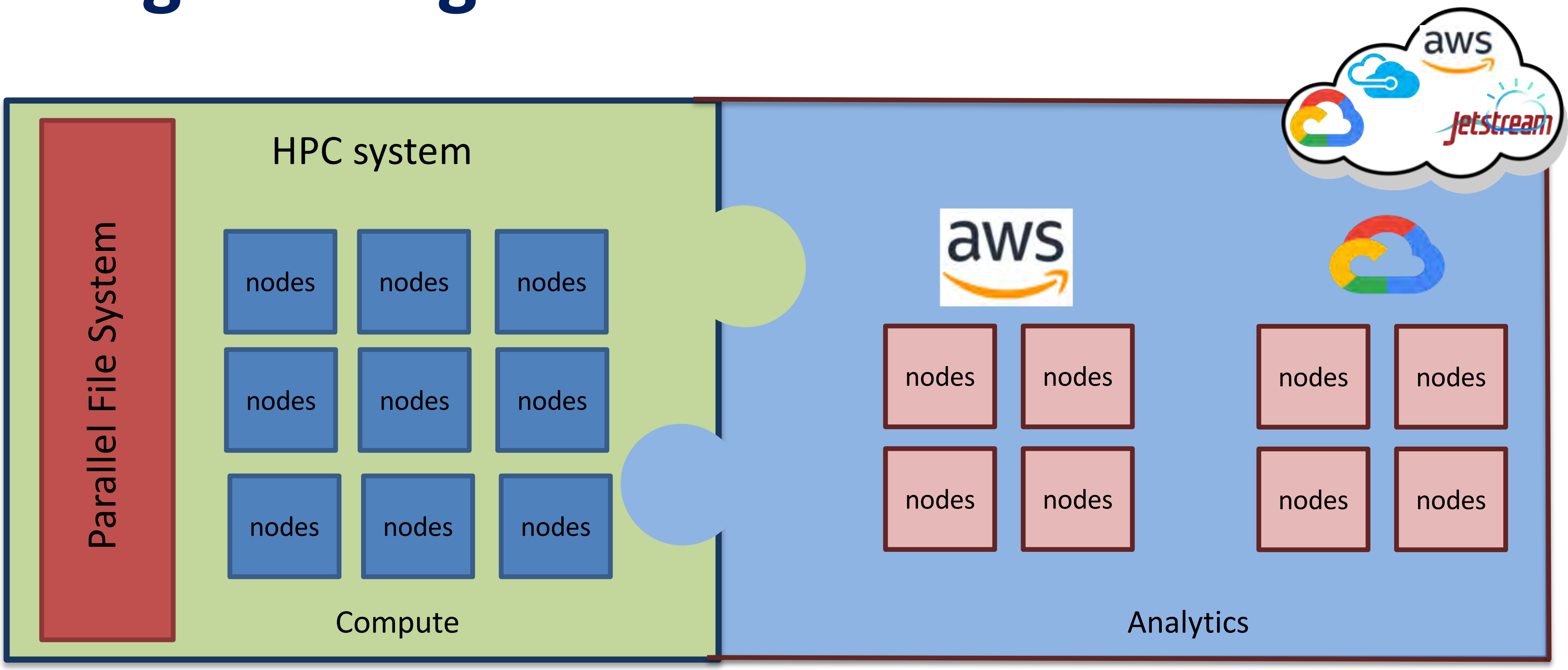
Building a Closed-loop Workflow



Building a Closed-loop Workflow

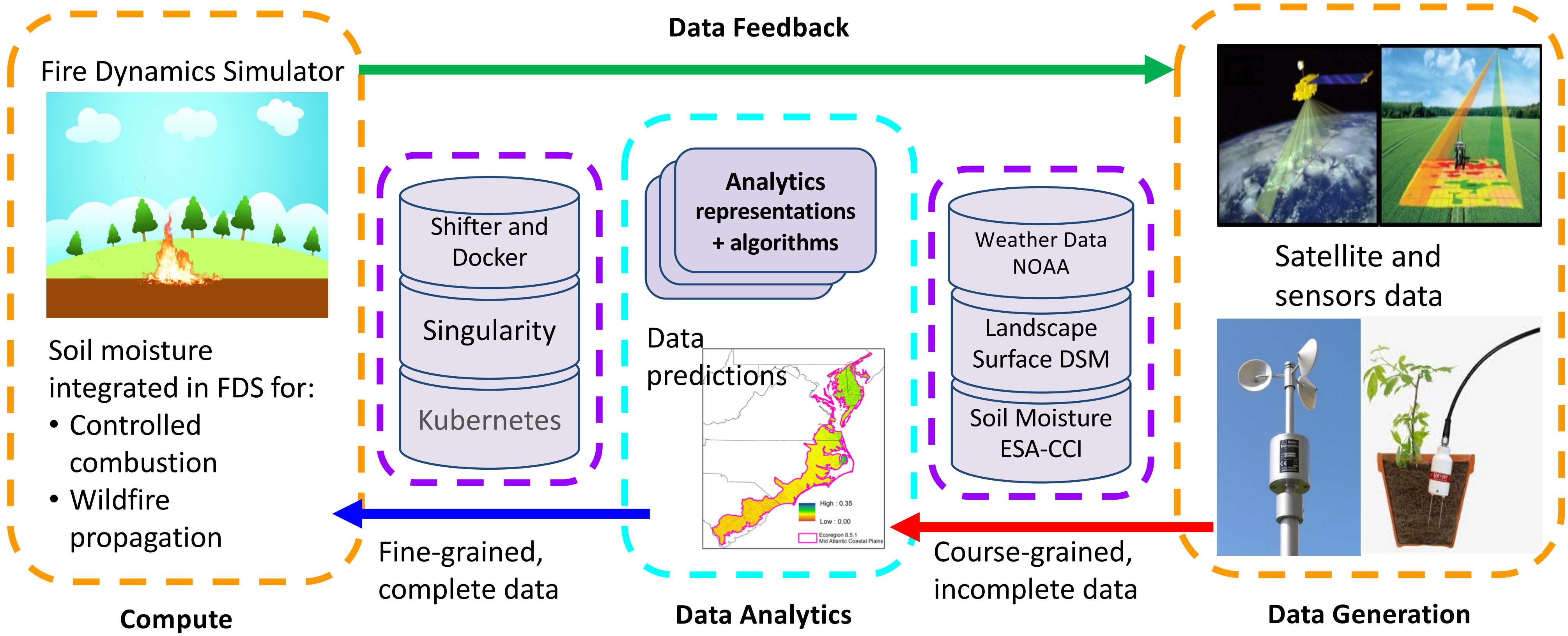


Augmenting HPC with the Cloud

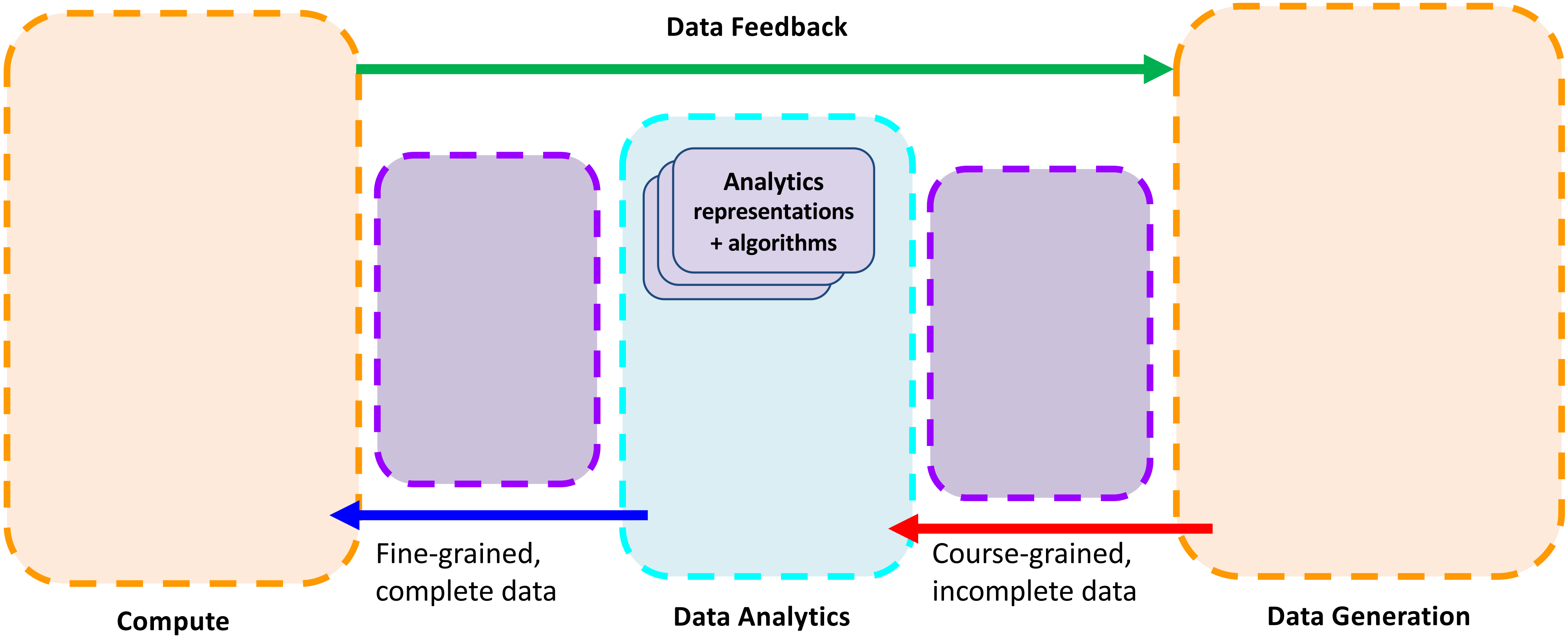


Source: https://www.nextplatform.com/2018/02/26/adaptive-approach-bursting-hpc-cloud/

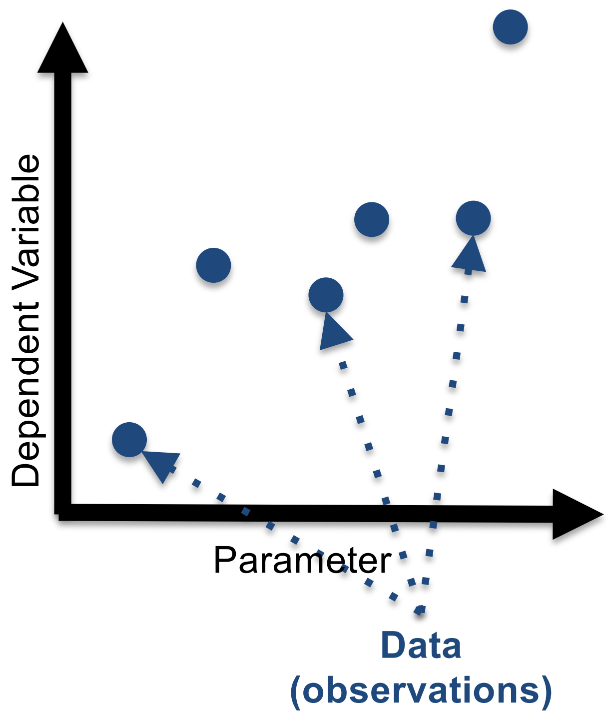
Building a Closed-loop Workflow

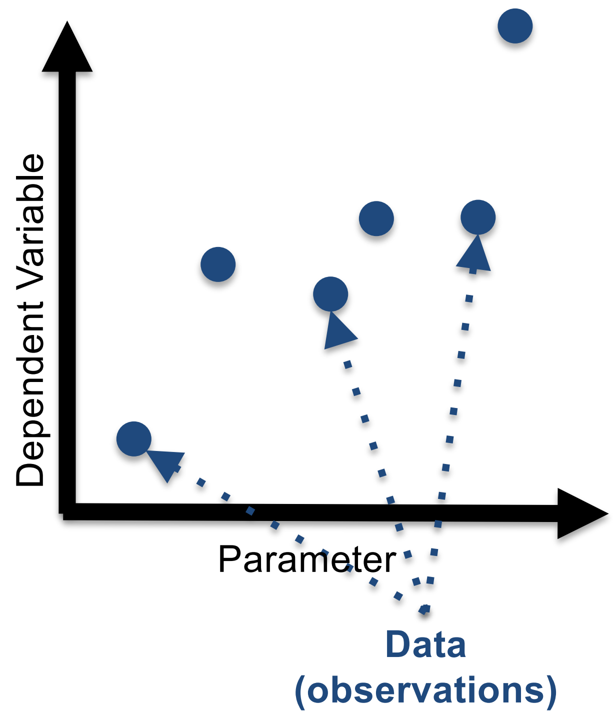


Hybrid Algorithms for Analytics

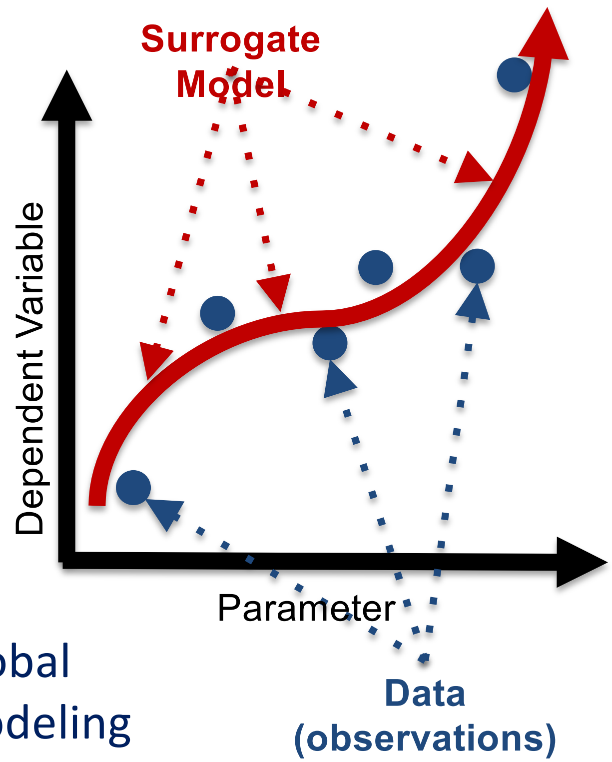


Global versus Local Data Modeling

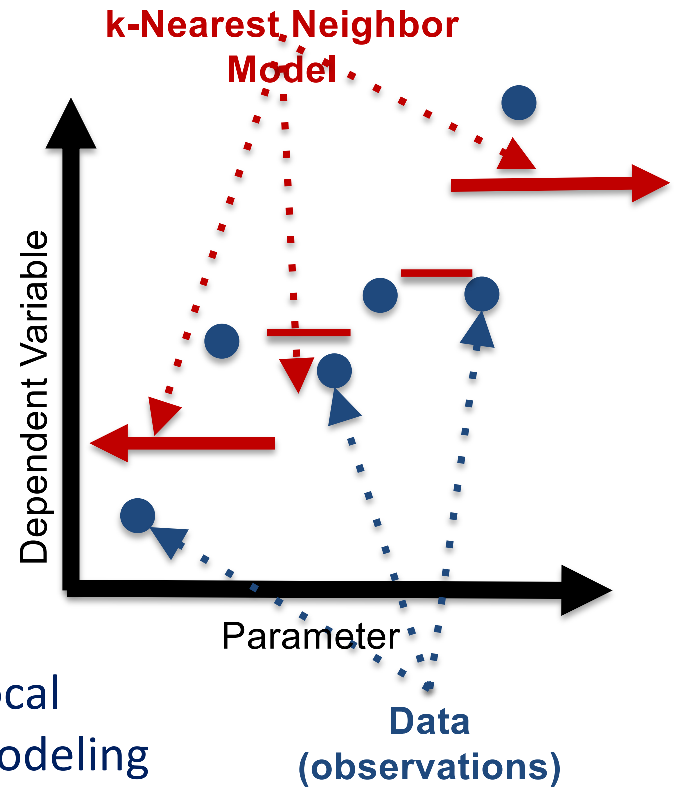




Global versus Local Data Modeling



Global modeling



Local modeling

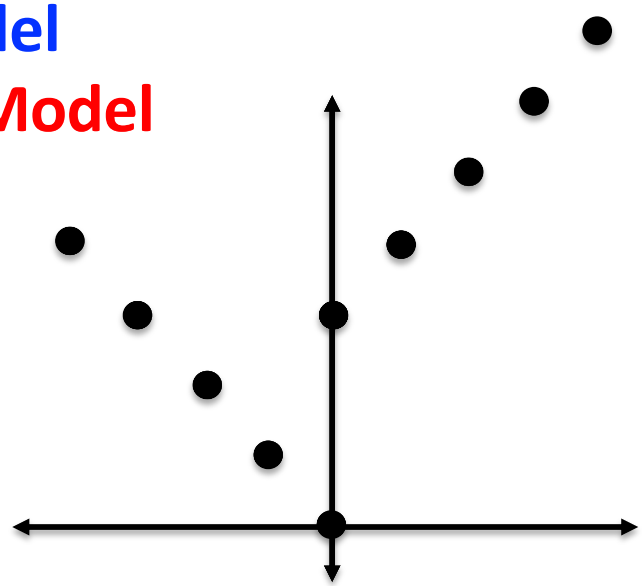
SBM vs. k-NN

Observations

Piecewise linear data

Surrogate-based model

2-Nearest Neighbor Model



Johnston et al., “HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces.” SBAC-PAD 2016: 26-33

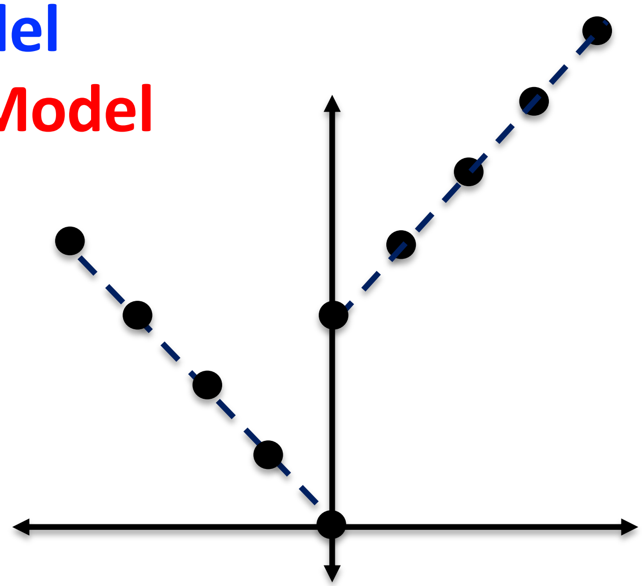
SBM vs. k-NN

Observations

Piecewise linear data

Surrogate-based model

2-Nearest Neighbor Model



Johnston et al., “HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces.” SBAC-PAD 2016: 26-33

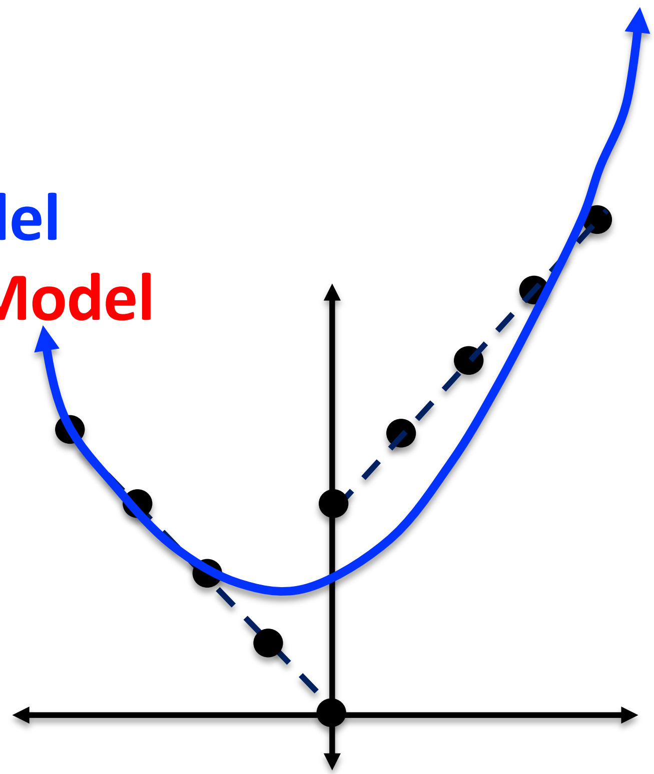
SBM vs. k-NN

Observations

Piecewise linear data

Surrogate-based model

2-Nearest Neighbor Model



Johnston et al., "HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces." SBAC-PAD 2016: 26-33

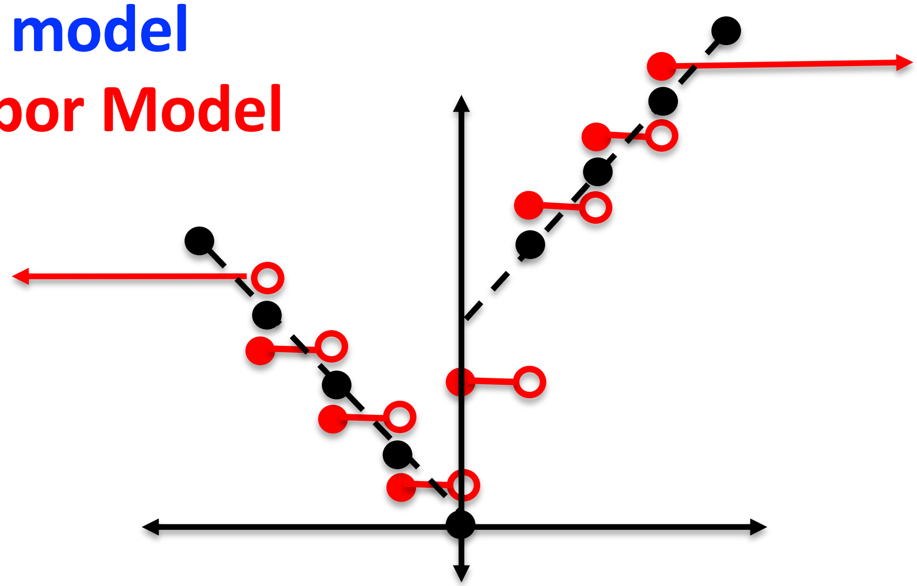
SBM vs. k-NN

Observations

Piecewise linear data

Surrogate-based model

2-Nearest Neighbor Model



Johnston et al., "HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces." SBAC-PAD 2016: 26-33

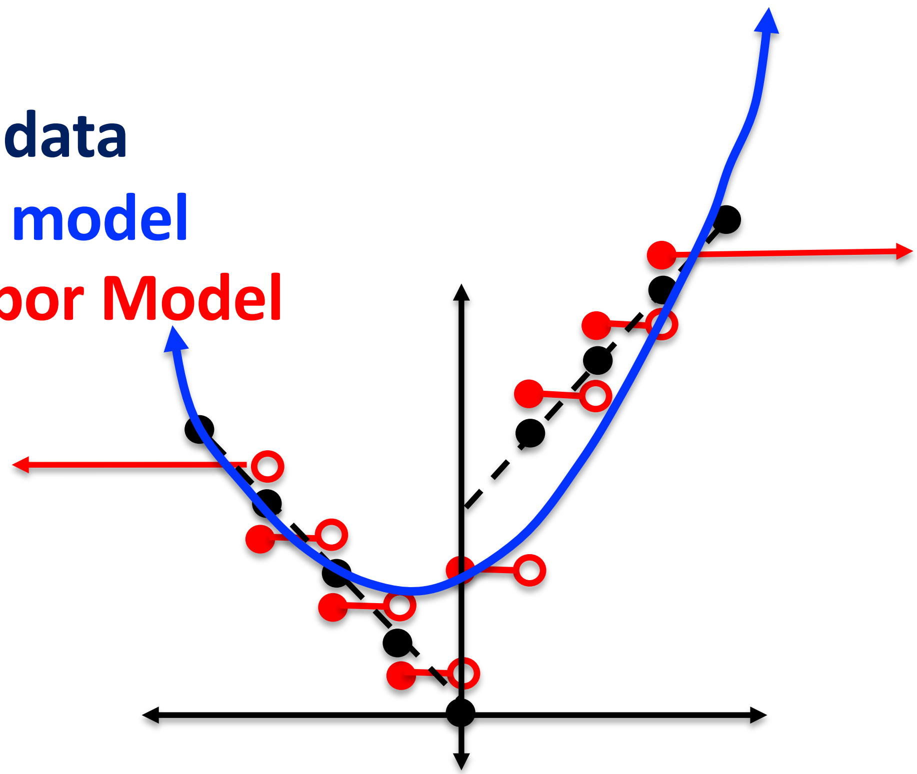
SBM vs. k-NN

Observations

Piecewise linear data

Surrogate-based model

2-Nearest Neighbor Model



Johnston et al., "HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces." SBAC-PAD 2016: 26-33

Hybrid Piecewise Polynomial Modeling (HYPPO)

k Nearest Neighbors

- Use **local** data
- Compute the average
  (*many simple local models*)

Surrogate-Based Modeling

- Use **all** sampled data
- Construct **one** **polynomial**
  (*single complex global model*)

Johnston et al., "HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces." SBAC-PAD 2016: 26-33

Hybrid Piecewise Polynomial Modeling (HYPPO)

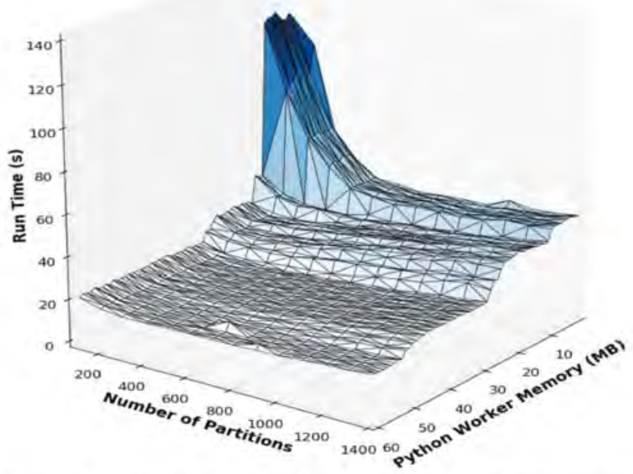
k Nearest Neighbors
- Use **local** data
- Compute the average
  (*many simple local models*)

Surrogate-Based Modeling
- Use **all** sampled data
- Construct **one** **polynomial**
  (*single complex global model*)

Hybrid modeling → HYPPO
- Use **local** data
- Construct many **polynomials**
  (*many complex local models*)

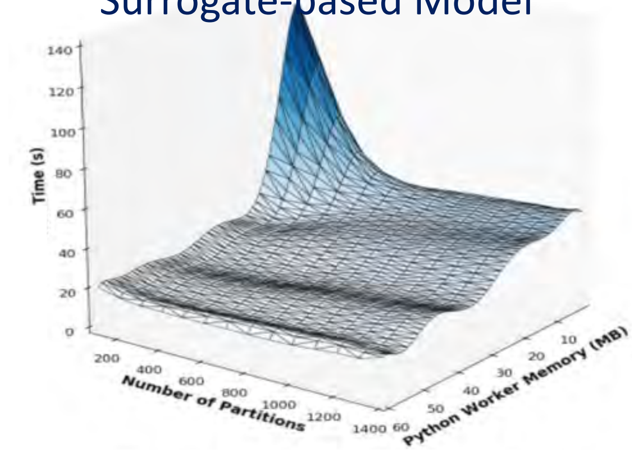
Johnston et al., "HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces." SBAC-PAD 2016: 26-33

Observations

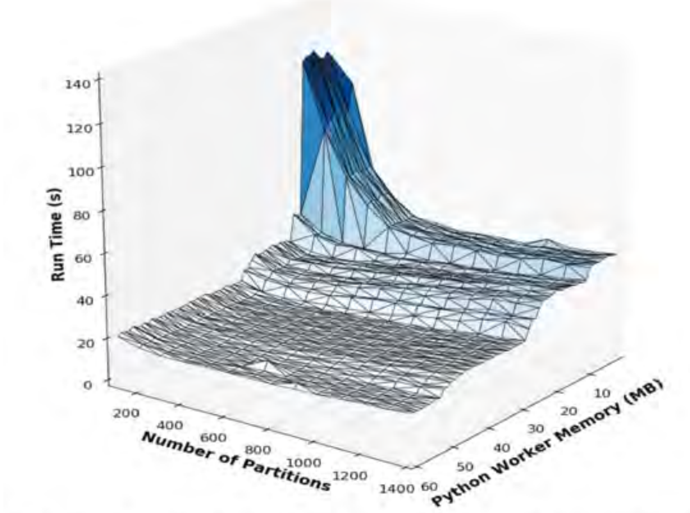


Johnston et al., “HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces.” SBAC-PAD 2016: 26-33

Surrogate-based Model

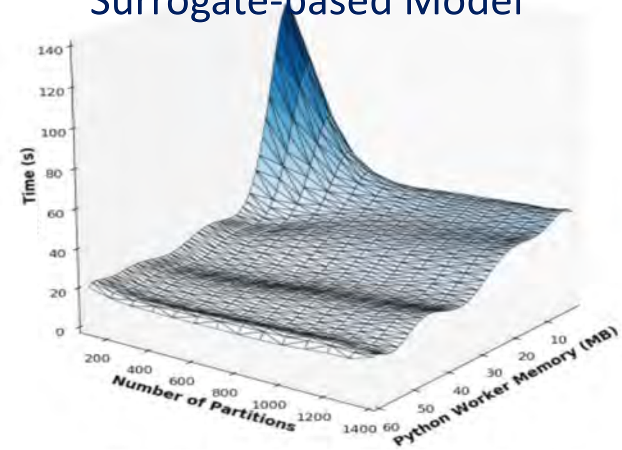


Observations

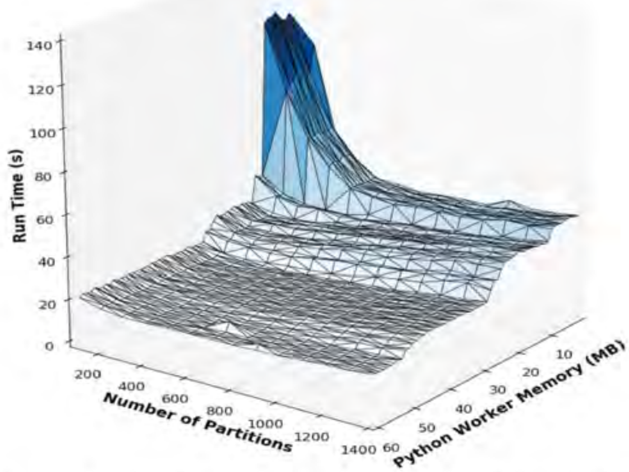


Johnston et al., "HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces." SBAC-PAD 2016: 26-33

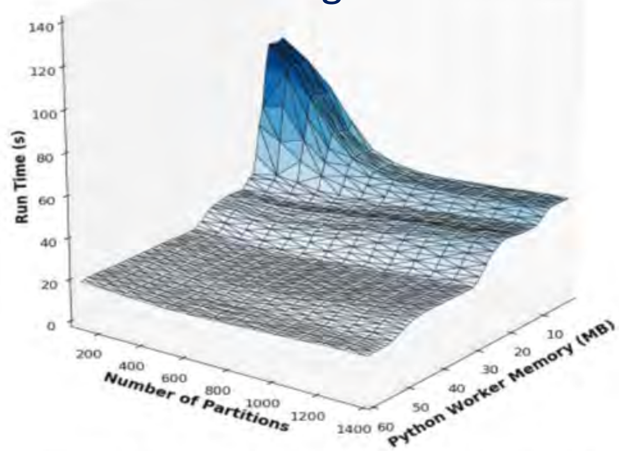
Surrogate-based Model



Observations

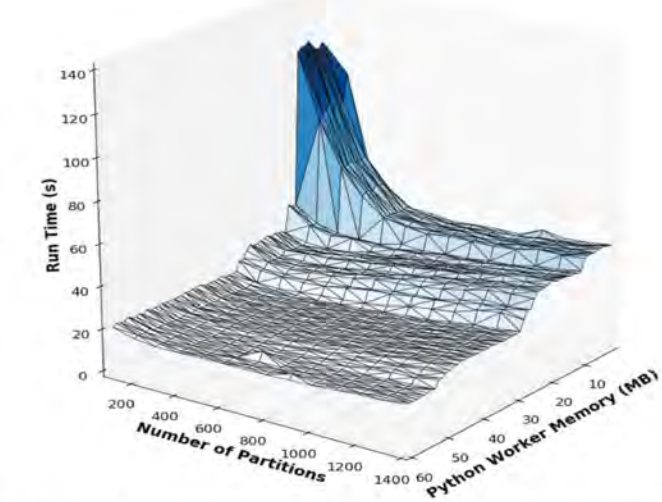


k Nearest Neighbors Model

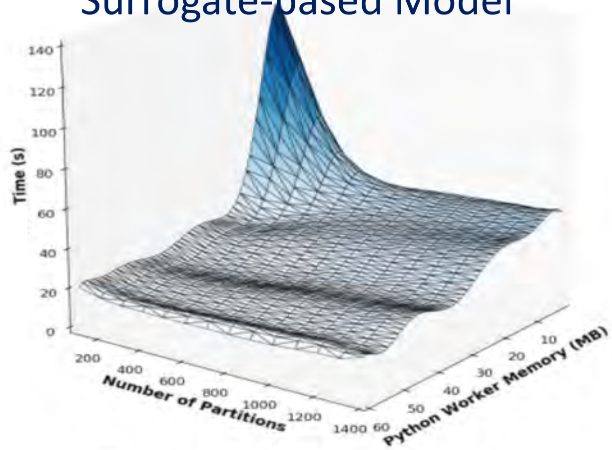


Johnston et al., “HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces.” SBAC-PAD 2016: 26-33

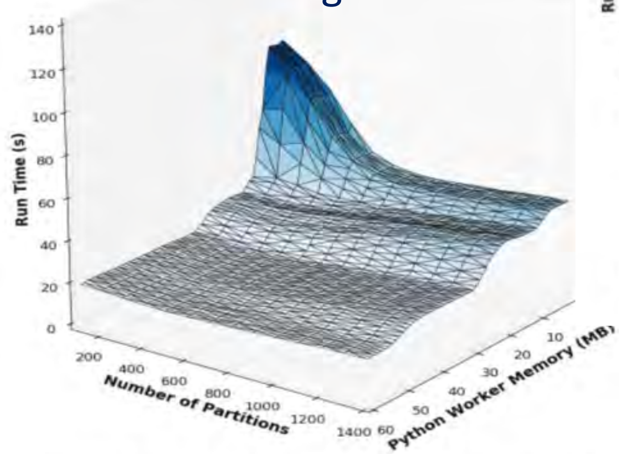
Observations



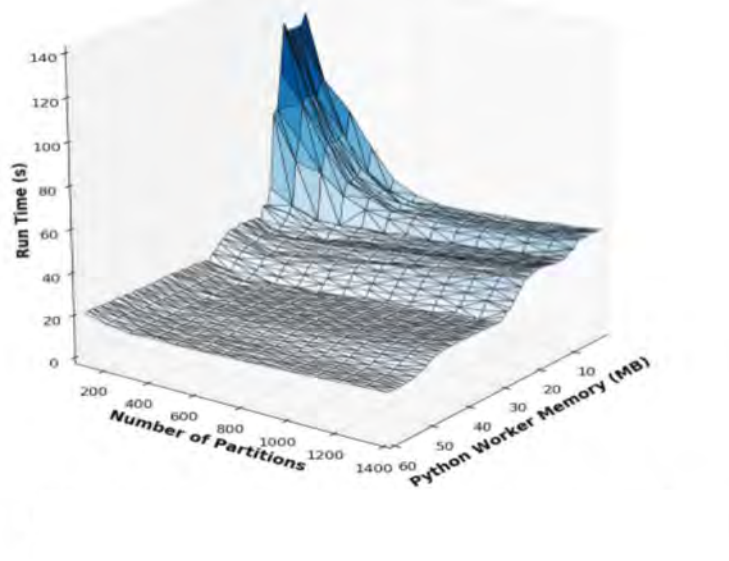
Surrogate-based Model



k Nearest Neighbors Model

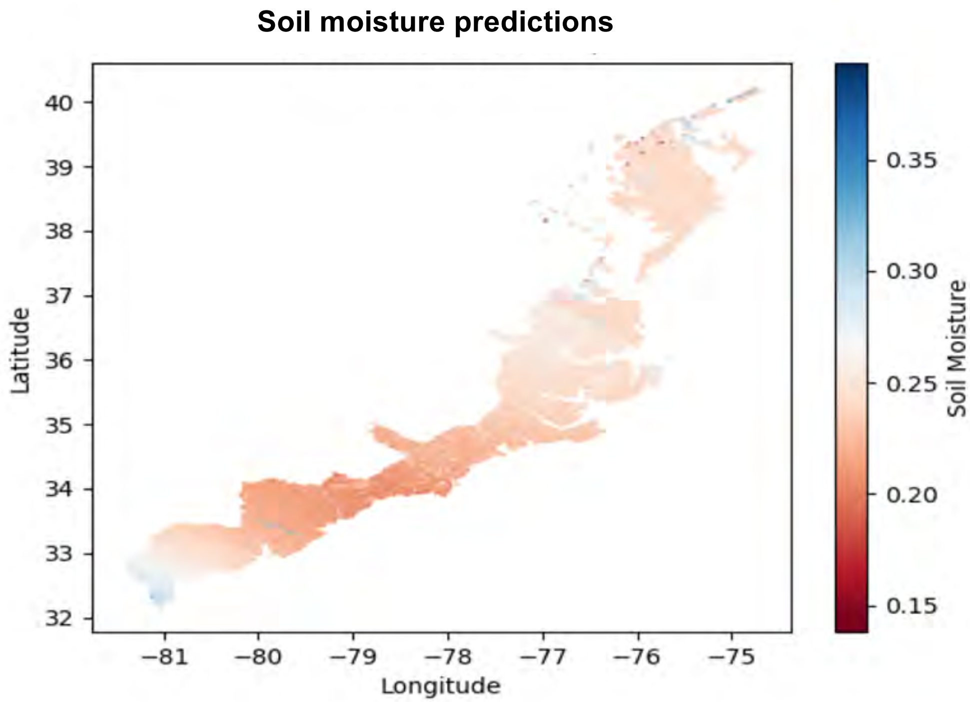


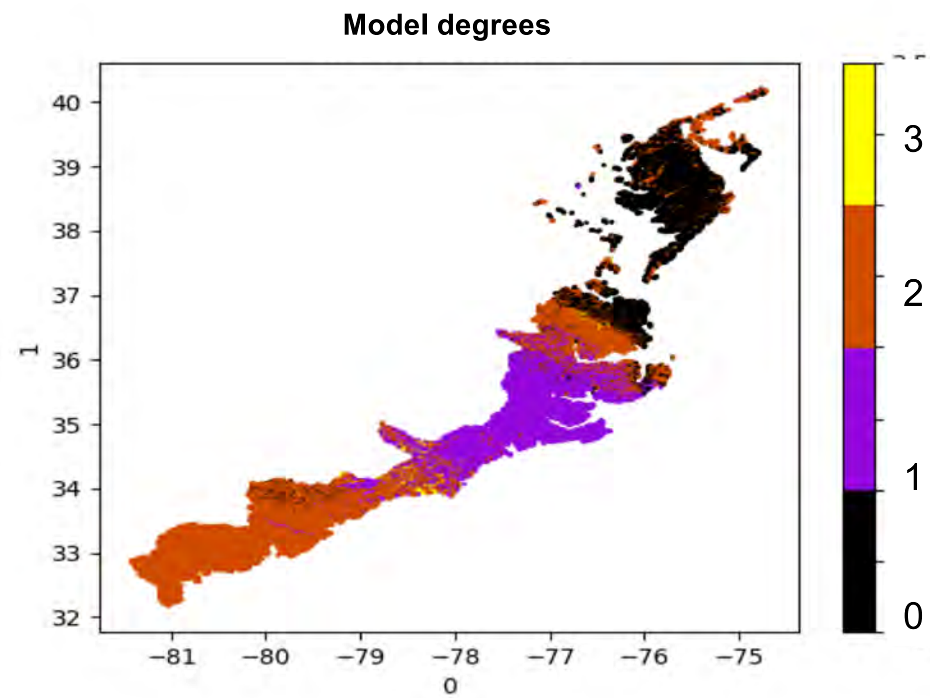
Hybrid Piecewise Polynomial Model (HYPPO)



Johnston et al., “HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces.” SBAC-PAD 2016: 26-33

Case Study: Fine-grained Modeling of Mid-Atlantic Region





D. Rorabaugh, M. Guevara, R. Llamas, J. Kitson, R. Vargas, M. Taufer. SOMOSPIE: A modular SOil MOisture SPatial Inference Engine based on data driven decision. **arXiv:1904.07754, 2019**

Challenges and Opportunities (I)

- Two trends in HPC are impacting scientific applications:
  - Convergence of simulations and data analytics
  - Emergence of edge computing
- Applications in precision medicine and precision farming are leveraging these trends
- There is the need to further integrate these trends in HPC
  - New challenges and opportunities for the HPC community

Challenges and Opportunities (II)

- *Efficiency:* Optimize performance and power usage associated to data generation, movement, and analytics
- *Non-invasive:* Capture knowledge from data without rewriting simulations' legacy codes or simulations' scripts
- *Generality:* Build workflows that support different types of analytics across different applications and different data
- *Portability:* Execute combined compute and analytics across different systems, including the edge, and with heterogenous resources
- *Scalability:* Design methods for knowledge discovery at scale (e.g., scalable ML algorithms) for "compute + analytics + data" workflows