HPC + D + A = HPDA?

**Barry Bolding, PhD
SVP & Chief Strategy Officer
Cray Inc.
bbolding@cray.com**

Era of Mobile

Age of Context
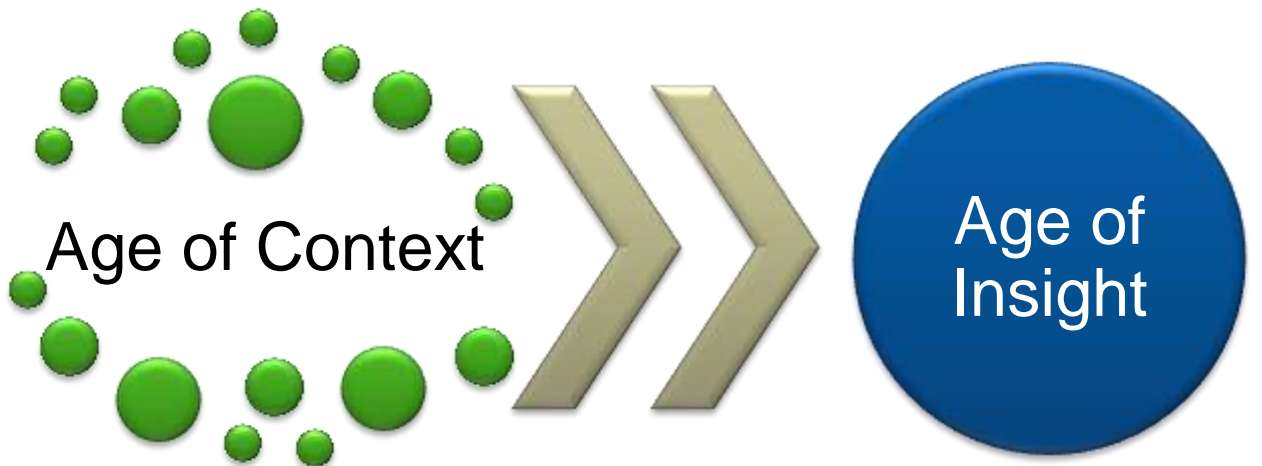
https://medium.com/crossing-the-pond/into-the-age-of-context-f0aed15171d7
Age of Context - Robert Scoble and Shel Israel

Context can be powerful, but…

Age of Context

Age of Insight

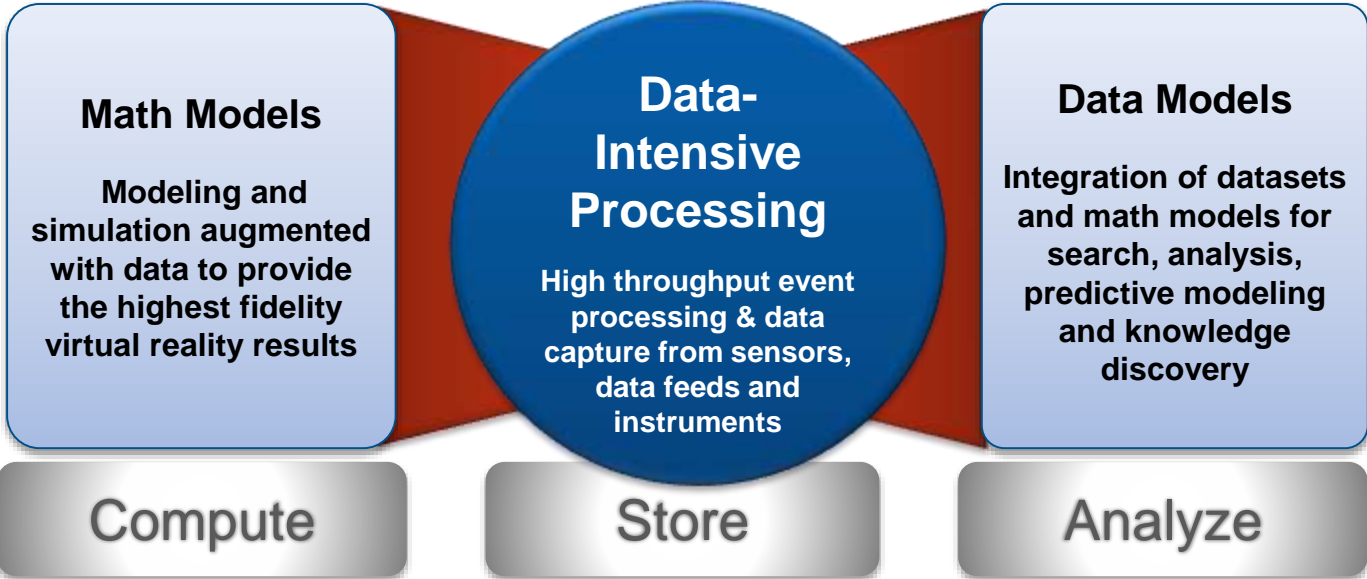Contextual interpolation

Insightful extrapolation

Cray's Vision:
The Fusion of Supercomputing and Big & Fast Data

CRAY®

Modeling The World
Cray Supercomputers solving "grand challenges" in science, engineering and analytics

Math Models

Modeling and simulation augmented with data to provide the highest fidelity virtual reality results

Data-Intensive Processing

High throughput event processing & data capture from sensors, data feeds and instruments

Data Models

Integration of datasets and math models for search, analysis, predictive modeling and knowledge discovery
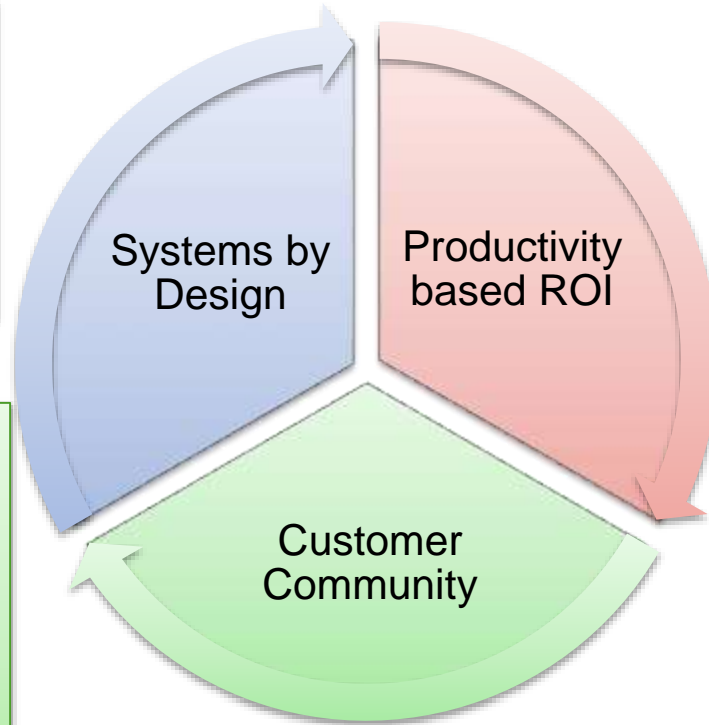
Compute

Store

Analyze

Cray's Unique Strengths in Supercomputing + Analytics

CRAY

We design our systems for **production computing and analytics at scale**.

The Cray community of customers and partners are the most demanding and enabling. Allows Cray to push the boundaries of scalable systems.

Systems by Design

Productivity based ROI

Customer Community

Cray takes the productive applications view of ROI:
- Match system designs to the applications
- Enable highly productive programming environments.
- Build solutions that are flexible and upgradeable.
- Lower TCO than commodity clusters and cloud for data-intensive applications

The Alan Turing Institute

- **UK national institute for data sciences**
  - *to break new boundaries in how we use big data in a fast moving, competitive world*
- **Joint venture: founder Universities and EPSRC**
  - Cambridge, Edinburgh, Oxford, UCL, Warwick
- **Launch partners**
  - Lloyds Register Foundation, GCHQ and Cray
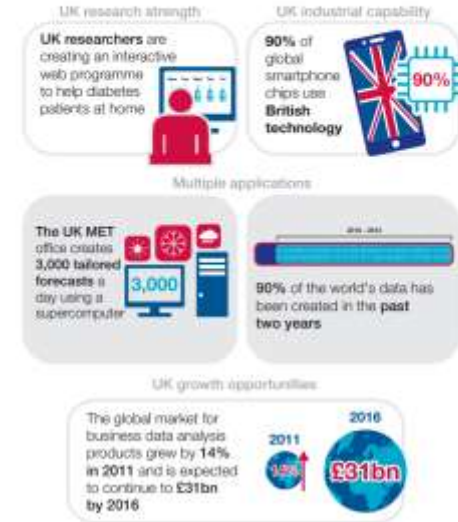- **UK Government investment in "Eight Great Technologies"**
- **https://turing.ac.uk/**

HM Government — Eight Great Technologies

Big Data
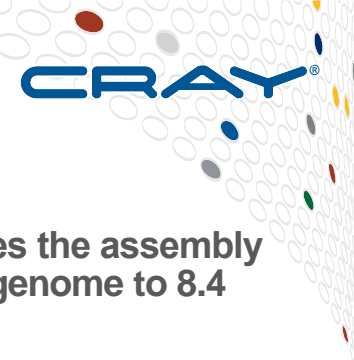Transforming the data revolution into new products and services

UK research strength — UK researchers are creating an interactive web programme to help diabetes patients at home

UK industrial capability — 90% of global smartphone chips use British technology

Multiple applications

The UK MET office creates 3,000 tailored forecasts a day using a supercomputer

90% of the world's data has been created in the past two years

UK growth opportunities

The global market for business data analysis products grew by 14% in 2011 and is expected to continue to £31bn by 2016

"We don't just want insights, we want actionable insights."

Peter Grindrod, University of Oxford

Human Genome Assembly – Under 9 Minutes on "Edison"

The end-to-end scaling of the team's HipMer genome assembler approach showing the human genome scalability on Edison on the left and the more complex wheat genome on the right (both axes are in log scale).

- Cray XC30 "Edison" reduces the assembly time of a complete human genome to 8.4 minutes

- A UPC modified version of the Meraculous code (Hip-Mer) was 170x faster.

- Eliminating the analytics back-log and fueling the drive towards precision medicine

- Cray's CX30 with Aries interconnect coupled with high memory bandwidth and memory on each node provided the I/O bandwidth required for genome assembly work

References: http://www.theplatform.net/2015/08/20/supercomputer-force-knocks-human-genome-assembly-under-9-minutes/
http://gauss.cs.ucsb.edu/~aydin/sc15_genome.pdf

"Using our HipMer technology enables for the first time assembly throughput to exceed the capability of all the world's sequencers"

HipMer: an Etreme-Scale DeNovo Genome Assembler
Georganas, Buluc, Chapman, Hofmeyr, Aluru, Egan, Oliker, Rokshar and Yelick
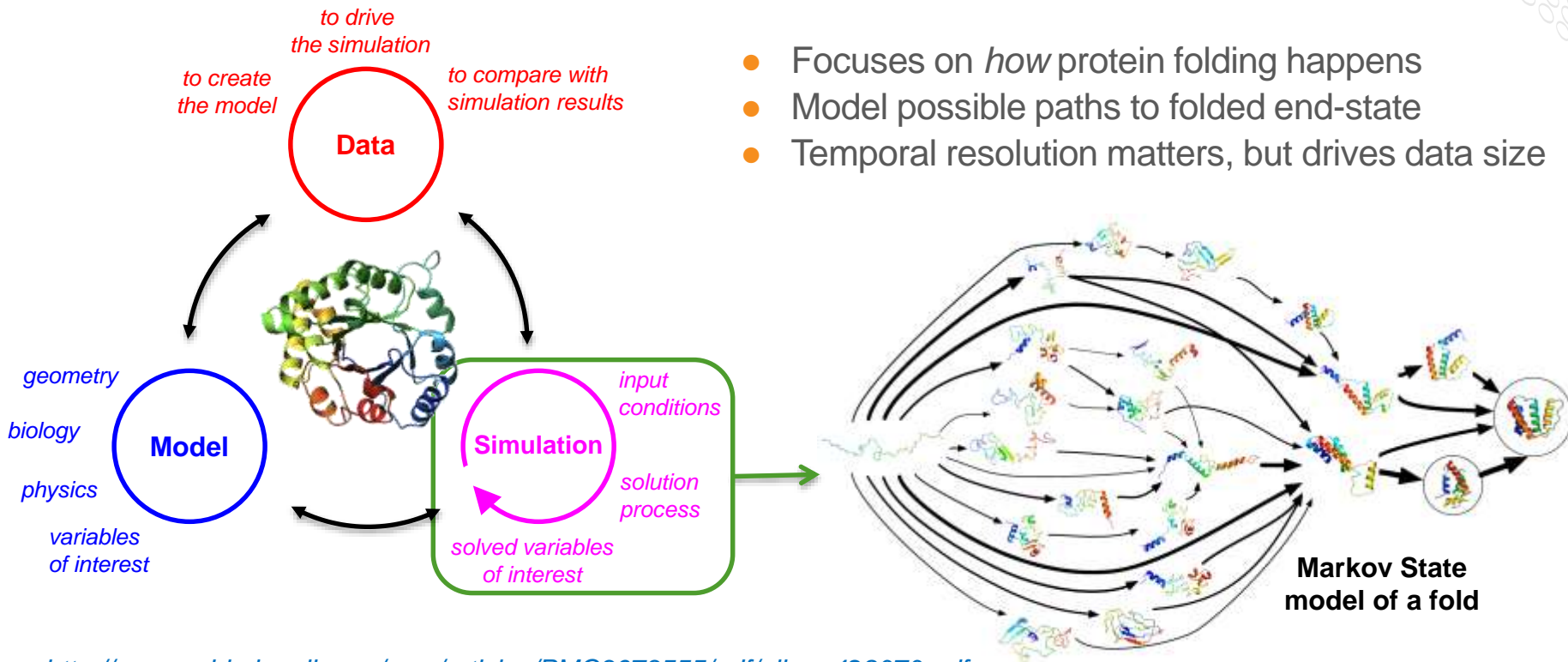
Magnus (CX30) Whitefly study at IVEC

They are battling a species complex of at least 34 morphologically indistinguishable species.

- Dr. Laura Boykin
  - funded by the Gates Foundation and a TED Fellowship.

- Awarded time on "Magnus," Her team is marrying genomics, supercomputing and evolutionary history to help African farmers develop management strategies and breed pest-resistant crops

"It's a massive problem. I'm one of 15 principal investigators working on a new project whose mission is to give farmers a cassava plant that's resistant to the viruses and the whiteflies."

Dr. Laura Boykin

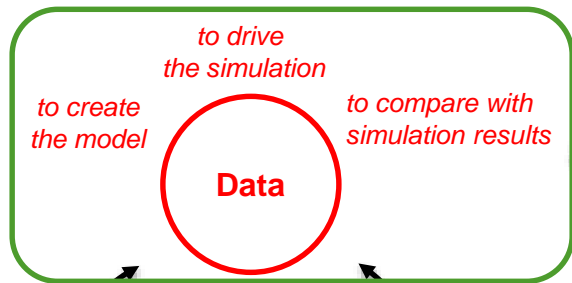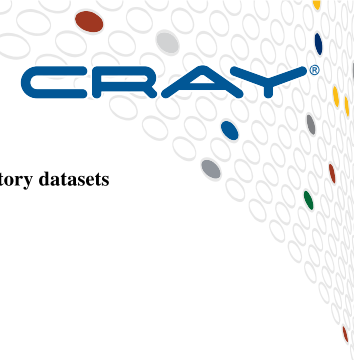# Protein Folding – Mixed Simulation and Analytics

to drive
the simulation

to create
the model

to compare with
simulation results

Data

geometry

biology

Model

physics

variables
of interest

Simulation

input
conditions

solution
process

solved variables
of interest

- Focuses on *how* protein folding happens
- Model possible paths to folded end-state
- Temporal resolution matters, but drives data size

**Markov State
model of a fold**

Protein Folding – Mixed Simulation and Analytics

Enabling in-situ data analysis for large protein-folding trajectory datasets
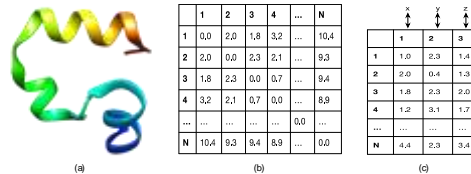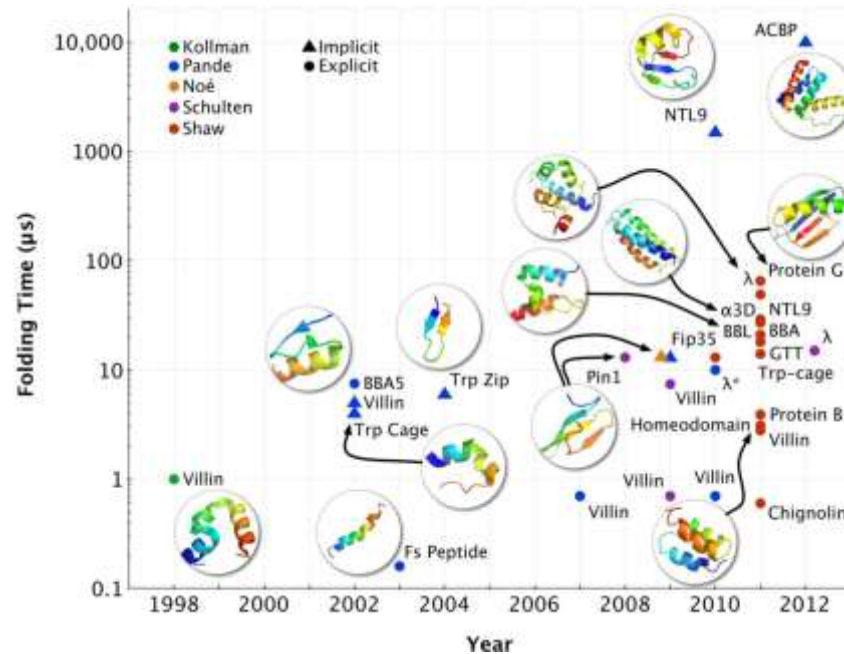
Figure 1: One conformation of the villin HP-35 protein (a); part of its distance matrix using only its backbone atoms in the conformation (b); and three eigenvectors and the associated eigenvalues capturing and synthesizing the conformation geometry (c).

Abstract—This paper presents a one-pass, distributed method that enables in-situ data analysis for large protein-folding trajectory datasets by executing sufficiently fast, avoiding moving trajectory data, and limiting the memory usage.

In-situ analysis of folds –
Dimensionality reduction using PCA & MDS

Automated classification in protein databases

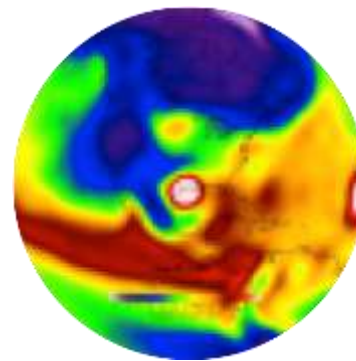*Thomas J. Lane, Diwakar Shukla, Kyle A. Beauchamp, and Vijay S. Pande*
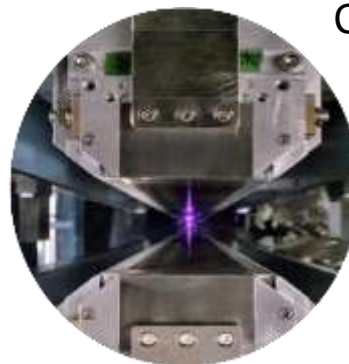
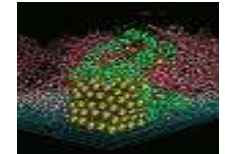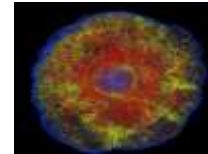# DOE Facilities are Facing a Data Deluge
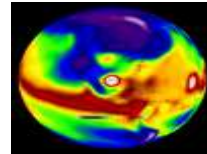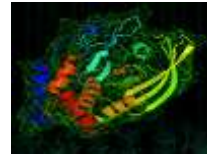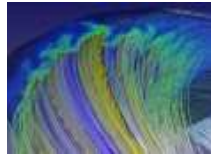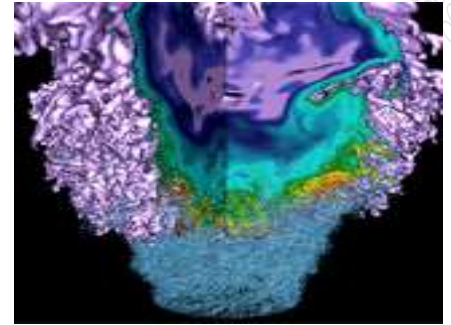

Astronomy

Genomics

Climate

Physics

Light Sources

Cray/AMPLab/LBL collaboration

Top Data Analytics Problems for NERSC users

1. **Generative Model for the Visible Universe**
   - Joint Inference across telescope images
2. **Pattern Detection for Climate**
   - Scalable Deep Learning on Cori
3. **Stephen Hawking Device**
   - Machine Learning for speech prosthetic
4. **Google Maps for BioImaging**
   - Semantic databases
5. **Genome Assembly for Wheat**
   - Graph Analytics

Data Characteristics

Science Domain	Data Source	Data Characteristics	Data Volume	Analysis Challenge
Cosmology	Multiple Telescopes	Noisy, multi-band, artifacts	O(100) TB	Data Fusion, Inference
HEP	Anti-Neutrino detectors	Noisy, artifacts, spatio-temporal	O(10) TB	Pattern/Anomaly Detection
BioImaging	*Mass-spec instruments*	*Noisy, artifacts, multi-modal*	*O(10) TB*	*Dim. Reduction Clustering, Pattern Detection*
Genomics	Sequencers	Missing data, errors	O(1-10) TB	Clustering, Pattern Detection
Neuroscience	Neural Recorders	Spatio-temporal, high dimensional, noisy	O(1) TB	Dim. Reduction, Pattern Detection
Climate	Satellites, Simulation o/p	Multi-variate, spatio-temporal	O(10) TB	Pattern/Anomaly Detection

"Current HPC machines are optimized for scientific simulations, arithmetic-intensive workloads, and regular computation, whereas data-analytics applications requires hardware optimized for bandwidth, throughput, concurrency and all-to-all communication."

Tumeo & Feo (IEEE Computer, Aug)

"The age of insight will require hardware and software infrastructures and ecosystems optimized for data analytic modeling & data-intensive simulations, including regular and irregular data access, high bandwidth, throughput, concurrency, all-to-all communication and scalability."