The Convergence of HPC, BDA, and AI in Future Workflows

Henry A Gabb, PhD

Senior Principal Engineer

Intel Core and Visual Computing Group

Legal Information

This presentation contains the general insights and opinions of Intel Corporation (“Intel”). The information in this presentation is provided for information only and is not to be relied upon for any other purpose than educational. Statements in this document that refer to Intel’s plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel’s results and plans is included in Intel’s SEC filings, including the annual report on Form 10-K.

Any forecasts of goods and services needed for Intel’s operations are provided for discussion purposes only. Intel will have no liability to make any purchase in connection with forecasts published in this document. Intel accepts no duty to update this presentation based on more current information. Intel is not liable for any damages, direct or indirect, consequential or otherwise, that may arise, directly or indirectly, from the use or misuse of the information in this presentation. Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

WHAT IS HPC?

"High-Performance Computing," or HPC, is the application of "supercomputers" to computational problems that are either too large for standard computers or would take too long.

- NICS

High-performance computing (HPC) is the use of super computers and parallel processing techniques for solving complex computational problems.

- Techopedia

The term high performance computing (HPC) refers to any computational activity requiring more than a single computer to execute a task.

- HPC Wales

HPC is an ***activity*** characterized by the workload's *nature, intent, and response to scale.*

HPC IS EVOLVING, EXPANDING…

SCOPE
ANALYTICS AND AI

SCALE
EXASCALE

DELIVERY
HPC IN THE CLOUD

DELUGE
THE FLOOD OF DATA

By 2020...

The average internet user will generate
~1.5 GB OF TRAFFIC PER DAY

Smart hospitals will generate over
3,000 GB PER DAY

Self driving cars will generate over
4,000 GB PER DAY... EACH

A connected plane will generate over
40,000 GB PER DAY

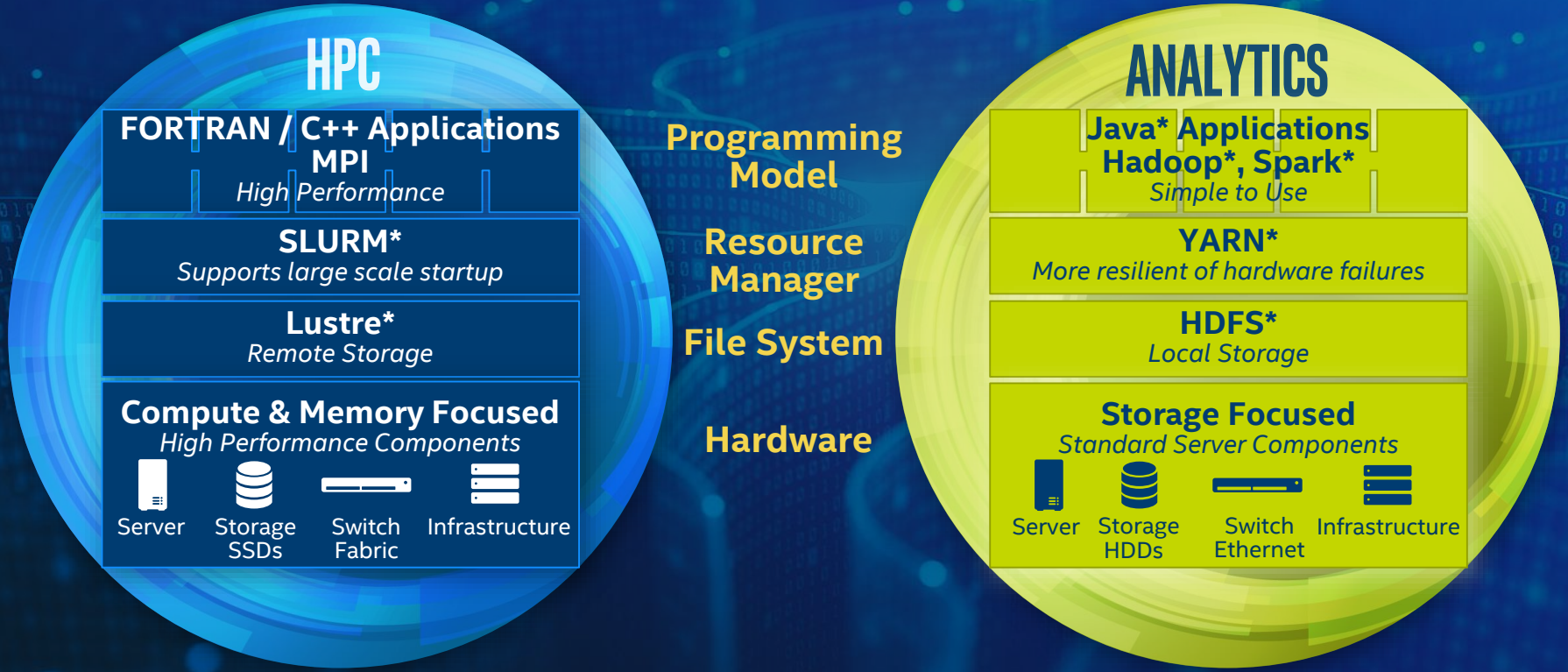
A connected factory will generate over
1,000,000 GB PER DAY

RADAR	~10-100 KB	PER SECOND
SONAR	~10-100 KB	PER SECOND
GPS	~50 KB	PER SECOND
LIDAR	~10-70 MB	PER SECOND
CAMERAS	~20-40 MB	PER SECOND

All numbers are approximated
http://www.cisco.com/c/en/us/solutions/service-provider/vni-network-traffic-forecast/infographic.html http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html
https://datafloq.com/read/self-driving-cars-create-2-petabytes-data-annually/172
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html,

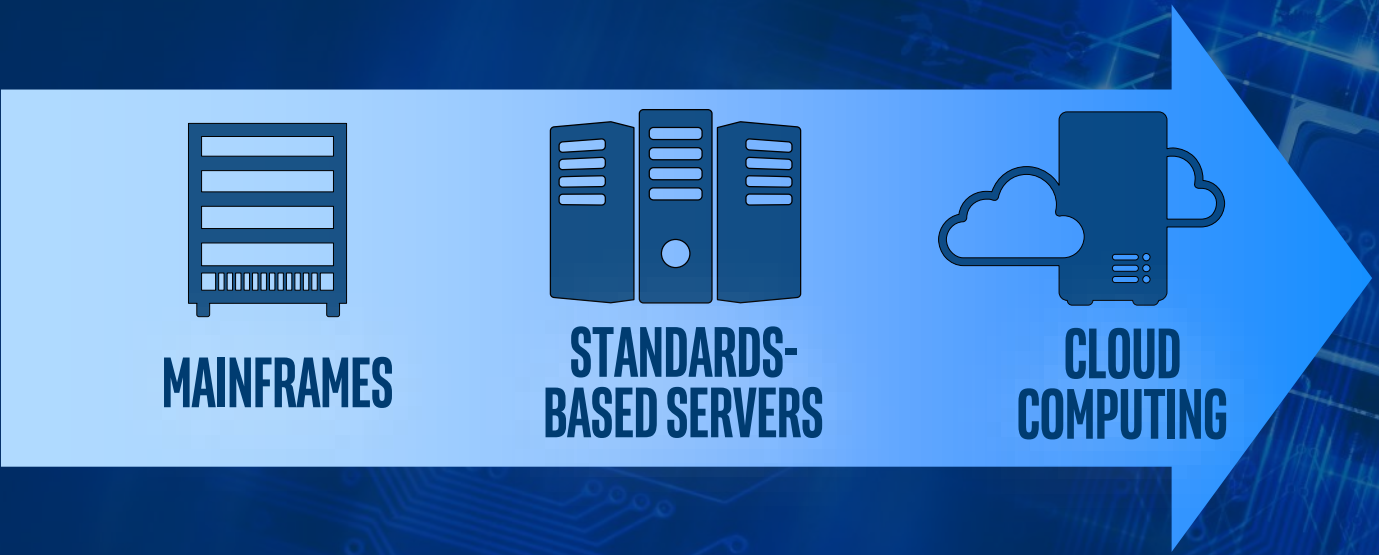
intel

SYSTEM PERSPECTIVE: TWO SEPARATE WORLDS



Daniel Reed and Jack Dongarra, *Exascale Computing and Big Data* in Communications of the ACM journal, July 2015 (Vol 58, No.7), and Intel analysis

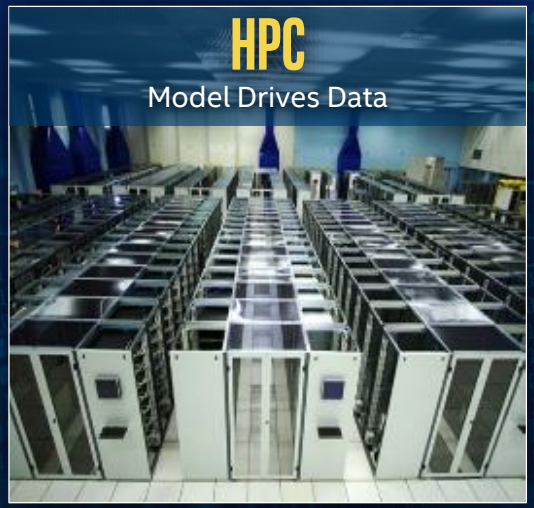
THE NEXT BIG WAVE OF COMPUTING

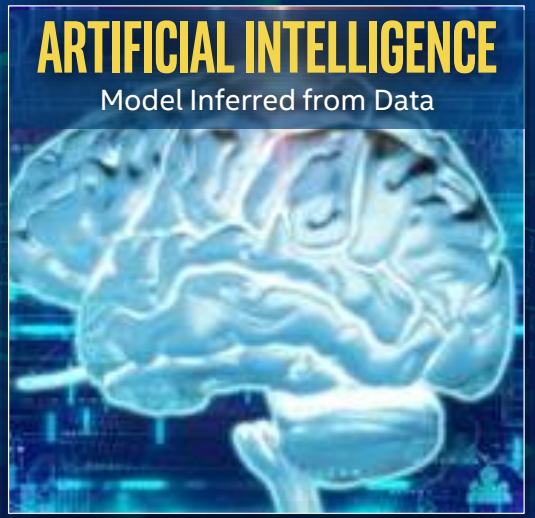


AI COMPUTE CYCLES WILL GROW 12X BY 2020

Source: Intel

THREE PILLARS OF THE EXASCALE ERA





Computational image: CERN

intel

Defining the “Three Pillars”

Traditional HPC

Big Data Analytics

Artificial Intelligence

Traditional High-Performance Computing

Equation-driven simulation

Performance is critical (and largely the user's responsibility)

- Low-level languages: Fortran, C, and C++
- Established parallel methods: MPI, OpenMP, TBB, CUDA
- Established numerical methods and math libraries

Batch processing

- Assumption of dedicated compute nodes with closed-ended, single-user processes
- Checkpoint/restart already part of most HPC applications
- CPU-hour is the allocation unit

Big Data

Overloaded, hyped term

Threshold used in the present study:

- *The data is large enough that I/O, transfer, storage, and archiving are first-order considerations in the workflow, e.g.:*
  - I/O consumes a significant portion of the runtime and stresses the filesystem
  - Repeated transfer between disk and memory is problematic
  - Computation must be brought to the data rather than vice versa
  - Archiving is an important part of the workflow

Artificial Intelligence

Another overloaded, hyped term

Threshold used in the present study:

- *Somewhere in the workflow, an automated process augments human judgment in the data processing, decision-making, or interpretation of results.*
- *It can be anything from rule-based decision support to deep neural networks.*

“Three Pillars” Use-Cases

Traditional high-performance computing, artificial intelligence, and big data analytics are converging.

"Three Pillars" Bonus Features

Application/Project Use-Cases	Domain	AI	HPC	BDA	SciVis	Interactive Steering
Sequence Variant Calling (CCBGM)	Genetics	✓	✓	✓		
Cryo-Electron Microscopy (LBNL)	Structural Biology	✓	✓	✓	✓	
Turbulent Flow Data (TACC)	Fluid Dynamics		✓	✓	✓	✓
HathiTrust Research Center (IU, UIUC)	Digital Humanities	✓	✓	✓		
Google Ngram	Linguistics		✓	✓		
Stellar Magnetism (TACC)	Astrophysics		✓	✓	✓	
ModelCenter Explore (Phoenix Integration)	Manufacturing	✓	✓		✓	✓
Tox21 Consortium (EPA, NIH, FDA), Robot Scientist (University of Manchester)	Toxicology, Pharmacology	✓		✓		✓
Big Data to Knowledge (NIH)	Medicine	✓	✓	✓	✓	
ALCF Data Science Program (ANL)	Physics, Energy, Brain	✓	✓	✓	✓	✓
Pittsburgh Science of Learning Center DataShop (CMU)	Education	✓		✓		

Generic Converged Use-Case #1

Many scientific applications fit this workflow.

Data-Collection Instrument

The instrument can be anything from a DNA sequencer to a medical device to a radio telescope to an IoT network, etc.

Streaming Data

Compute Resource

Store Data

Process Data

Distributed Database, Parallel Filesystem

Adjustments

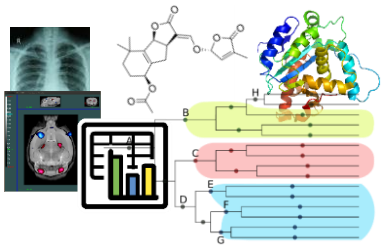
Interactive Steering

Discard Bad Data

Researcher/Analyst

Augment human analysis and interactive steering with artificial intelligence for better performance.

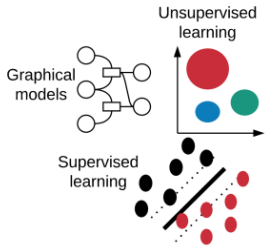
Results

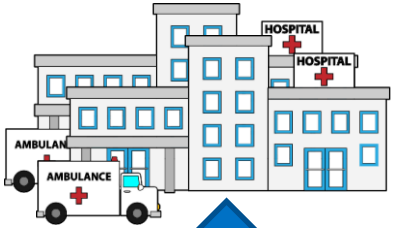


Better performance blurs the line between pre- and post-processing, allowing results while the sample is still in the instrument.

Generic Converged Use-Case #2

Collaborative environments are common in medicine.





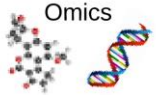
Clinical trials/EHR

Image

Analytics

Clinical Validation

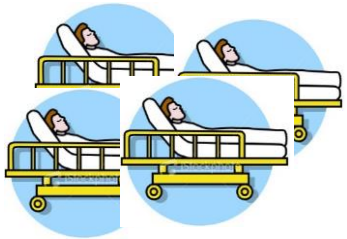
Data Types



Omics

Clinician assessment

Symptom assessment



EHR Data

Actionable Information

Physician, Medical Informaticist

PATIENT-SPECIFIC THERAPEUTIC CHOICES

BIOMARKER DISCOVERY

Patients

Integrated Compute/Storage Platform

Biotech StartupX

Use-Case #1

AWS API Gateway

Submit Order

AWS S3

Genes

Genomes

gRNA

Retrieve Sequences

AWS Lambda

$gRNA_1$ $Organism_1$
$gRNA_2$ $Organism_2$
... ...
$gRNA_N$ $Organism_M$

N x M Sequence Alignments

Candidate gRNA

AWS ML (Inference)

Candidate gRNA

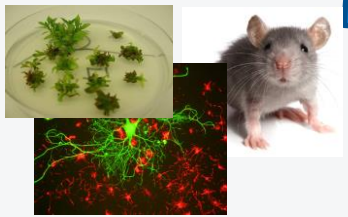
Deploy Model

gRNA ($)
KO cell line ($)

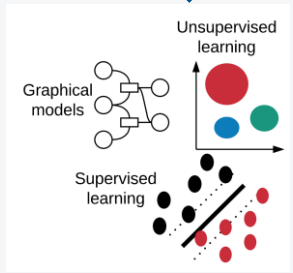
Use-Case #2

Validation

Analytics



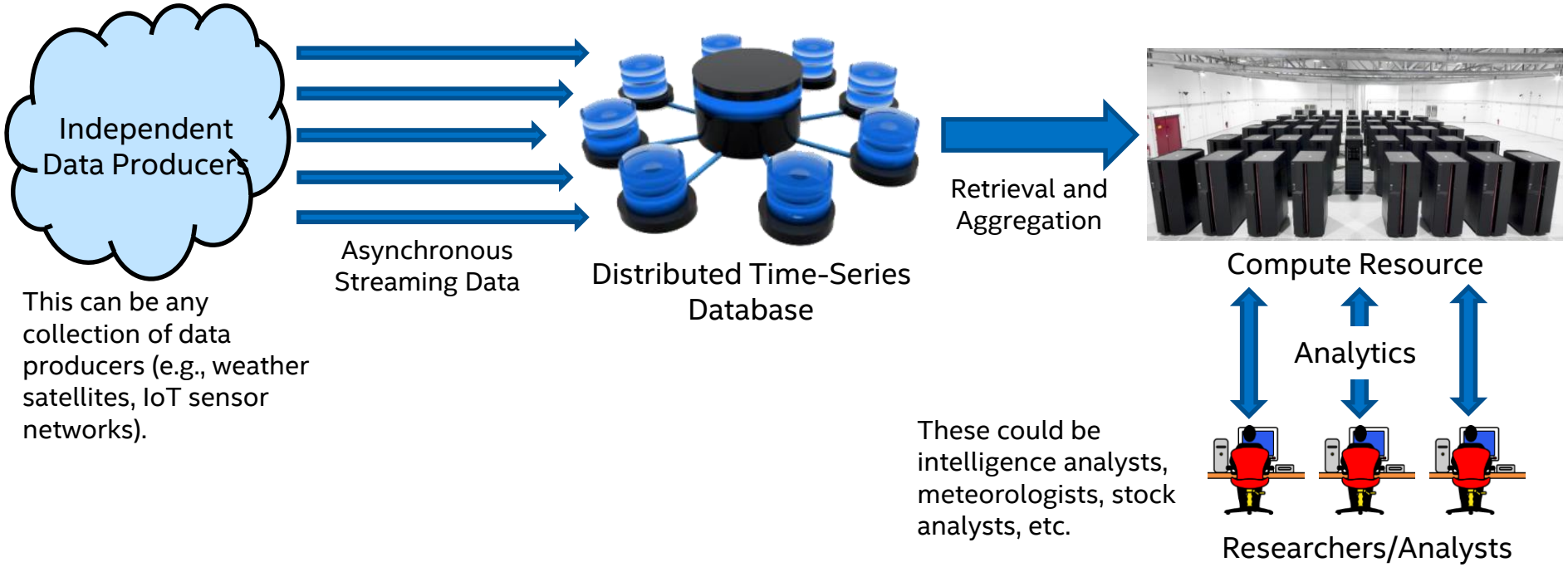
AWS Rekognition



AWS ML (Training)

Generic Converged Use-Case #3

Real-time monitoring and surveillance.



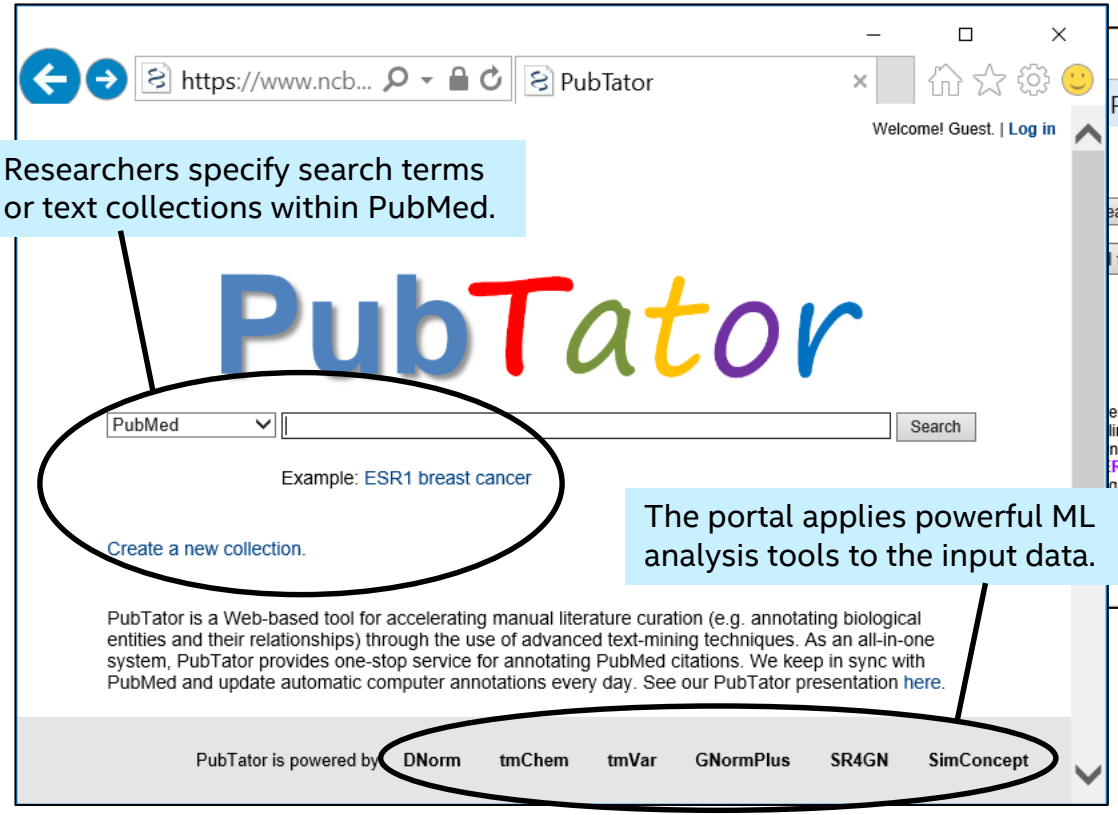
Generic Use-Case #4: Portals

Defining characteristics:

- Raising the level of abstraction
  - Computations initiated through intuitive, high-level interfaces
  - Complex HPC, BDA, and AI accessible to computing novices
  - Compute and data environments hidden from users
- Separation of concerns
  - Users focus on answering research questions
  - Maintainers focus on computational efficiency and data integrity
- Data are too large and/or dynamic to easily copy and maintain locally
  - Bring users to data instead of distributing data to users

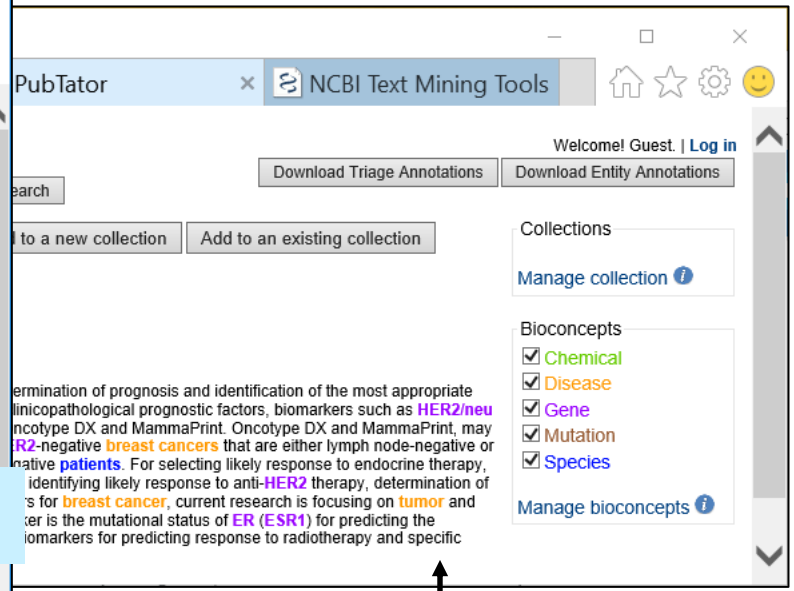
Generic Converged Use-Case #4 (example 1)

The National Library of Medicine has embraced abstraction for its vast text corpora.



Researchers specify search terms or text collections within PubMed.

The portal applies powerful ML analysis tools to the input data.



The underlying computations are hidden from the user. Analysis results are simply returned.

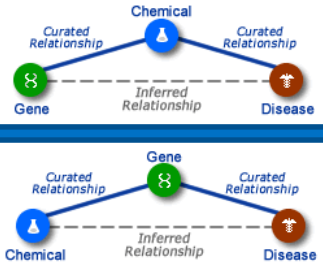
Generic Converged Use-Case #4 (example 2)

Comparative Toxicogenomics Database project mines massive scientific text corpora for chemical-gene-disease (C-G-D) relations.

Scientific Literature

Topic Selection

"Biocurators"



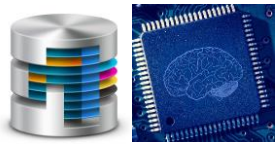
Compute Resource

Text Corpus

G-C, C-D, G-D Pairs

Assembling the semantic network is a *continuous* compute-, data-, and human-intensive process with AI augmenting human curation.

Network Assembly



AI and Controlled Vocabularies

Query Portal

Users

Queries

Hypotheses

(e.g., chemical toxicity, drug therapies, genetic tests)

Semantic Network of C-G-D Relations

Future Platforms

Traditional HPC platforms stressed processor and interconnect performance.

Newer platforms must balance processor, memory, interconnect, and I/O performance to accommodate both equation- and data-driven applications.

Resource Allocation in Future Environments

Is the processor-hour still the primary allocation unit?

Do tiered memory- and disk-hours have to be counted?

Is dedicated compute still a valid assumption, or is multi-tenancy the new norm?

Do single-user jobs give way to multi-user jobs (where some of the users are event-triggered AI agents)?

The Future of High-Performance Computing

Heterogeneity (CPU, GPU, FPGA, ASIC, etc.)

- Blessing: target algorithms to architectures
- Curse: *must* target algorithms to architectures for efficiency
- Demands separation of concerns between domain experts and tuning experts

Productivity and interactivity

- Rise of high-level languages: Python, R, Julia, Go, etc.
- Increasing reliance on frameworks: Apache Spark, TensorFlow, etc.

Workflows are both data- and equation-driven

- Data streaming from appliances and edge devices
- Persistent, memory-resident, and distributed databases
- Not necessarily closed-ended, single-user batch processes on dedicated resources
- Artificial intelligence augments human judgment to improve final results

Curation and Archiving

The Afterthought of Computational Research

Data plans not discussed in use-cases

Funding agencies now require data plans

- Necessary funding rarely provided
- Archiving is a continuous effort
- Curation is labor-intensive and requires specialized skills and knowledge
- Little agreement on metadata standards

Sometimes easier to repeat an experiment

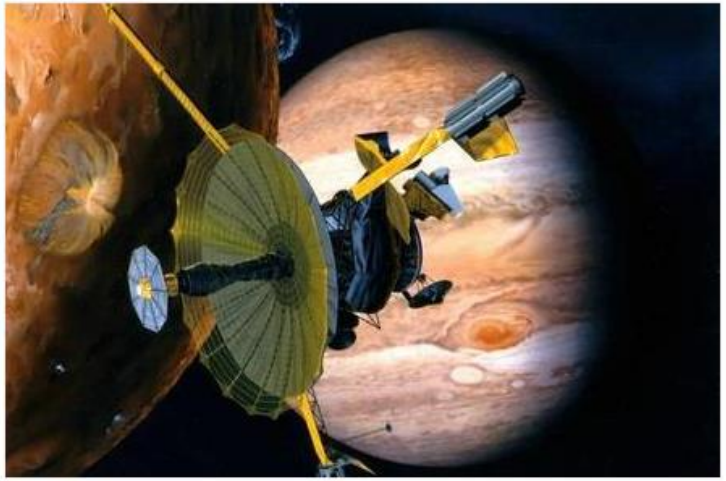
Sometimes easier to reprocess original data

Science https://www.theregister.co.uk

NASA dusts off FORTRAN manual, revives 20-year-old data on Ganymede

Analysing Galileo's Jovian moon results

By Richard Speed 1 May 2018



NASA Galileo Probe (Courtesy NASA/JPL-Caltech)

NASA scientists have made some new discoveries about Jupiter's giant moon Ganymede, thanks to a dedicated team, an elderly VAX machine and 20-year-old data from the long-defunct Galileo probe.