

# The Fourth Paradigm, Open Science and AI

Tony Hey

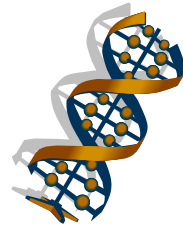
Senior Data Scientist

Rutherford Appleton Laboratory

STFC, Harwell Campus



**Earth Sciences**



**Life Sciences**



**Computer and  
Information  
Sciences**

# e-Science, the Library Community and the Support of Research



**Social Sciences**

Tony Hey  
Corporate Vice President  
Technical Computing  
Microsoft Corporation



**New Materials,  
Technologies  
and Processes**



**Multidisciplinary  
Research**

Tony Hey CNI Talk December 2005

# Jim Gray's Fourth Paradigm: Data-Intensive Scientific Discovery

# Jim Gray, Turing Award Winner

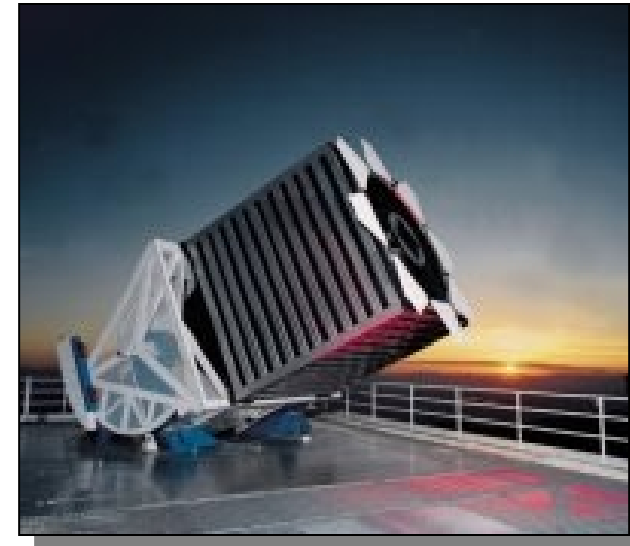


# The 'Cosmic Genome Project': The Sloan Digital Sky Survey

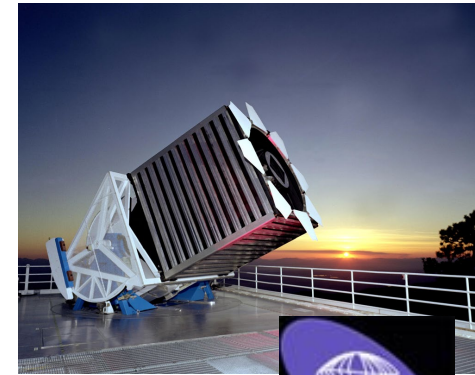
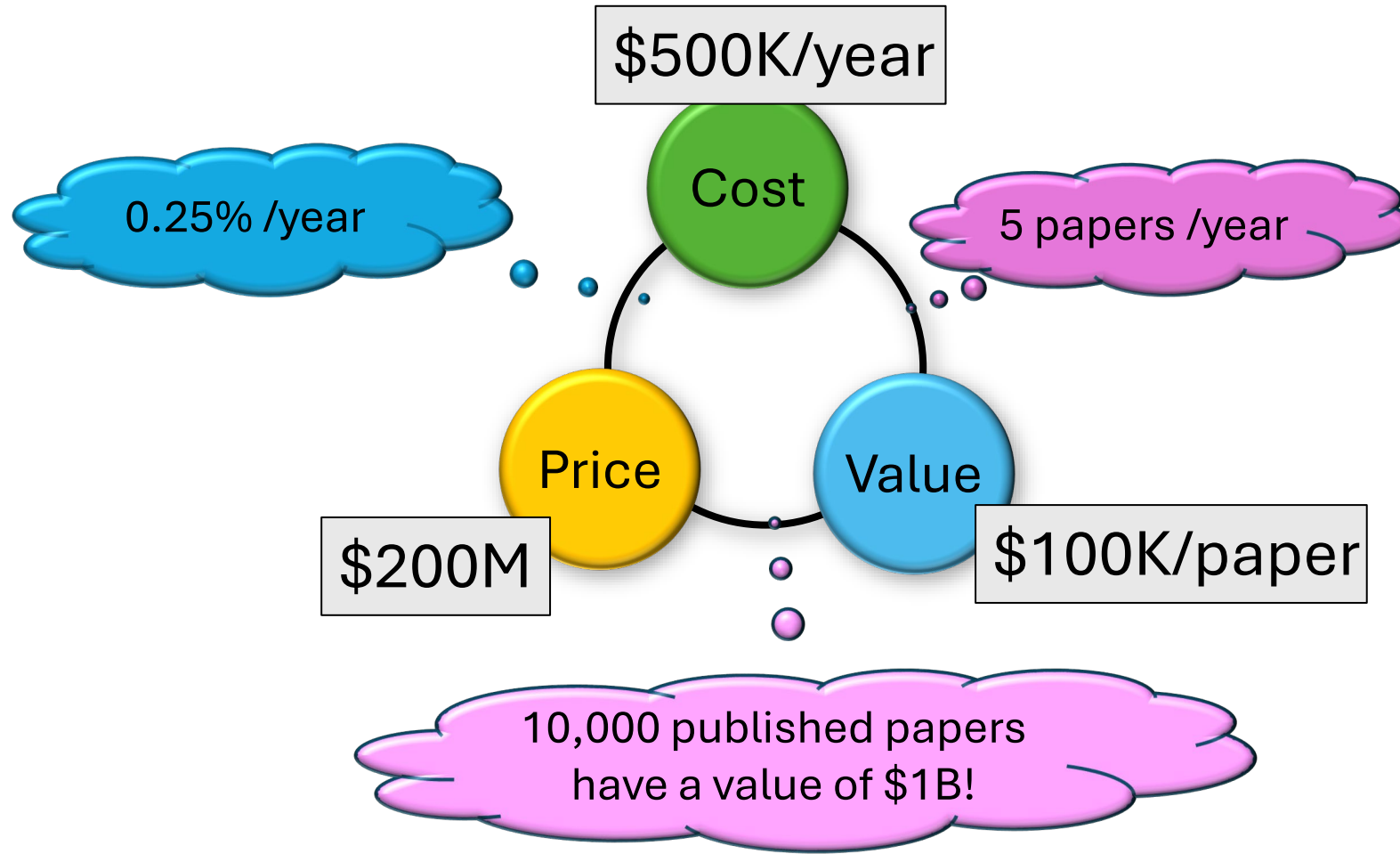


- Two surveys in one
  - Photometric survey in 5 bands
  - Spectroscopic redshift survey
- Data is public
  - 2.5 Terapixels of images
  - 40 TB of raw data => 120TB processed data
  - 5 TB catalogs => 35TB in the end
- Started in 1992, 'finished' in 2008
  - SkyServer Web Service built at JHU by team led by Alex Szalay and Jim Gray

*The University of Chicago  
Princeton University  
The Johns Hopkins University  
The University of Washington  
New Mexico State University  
Fermi National Accelerator Laboratory  
US Naval Observatory  
The Japanese Participation Group  
The Institute for Advanced Study  
Max Planck Inst, Heidelberg  
Sloan Foundation, NSF, DOE, NASA*



# Price, Value and Costs for the SDSS



Slide thanks  
to Alex Szalay

*5% of the survey price would cover the data for 20 years!*

# Jim Gray in his last talk ...

## Jim's definition of eScience: "IT meets Scientists"

"... the world of science has changed and there is no question about this. The new model is for the data to be captured by instruments or generated by simulations before being processed by software and the resulting information or knowledge stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new fourth paradigm for scientific exploration."

From the book 'The Fourth Paradigm: Data-Intensive Scientific Discovery

# The Fourth Paradigm: Data-Intensive Science

## Thousands of years ago – **Experimental Science**

- Description of natural phenomena

## Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations...

## Last few decades – **Computational Science**

- Simulation of complex phenomena

## Today – **Data-Intensive Science**

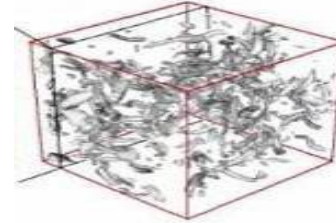
- Scientists overwhelmed with data sets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks

**eScience is the set of tools and technologies to support data federation and collaboration**

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K \frac{c^2}{a^2}$$



*(With thanks to Jim Gray)*



Big Scientific Data:  
A new Generation of Experiments

# The Vera C. Rubin Observatory: The Legacy Survey of Space and Time (LSST)

## Goal:

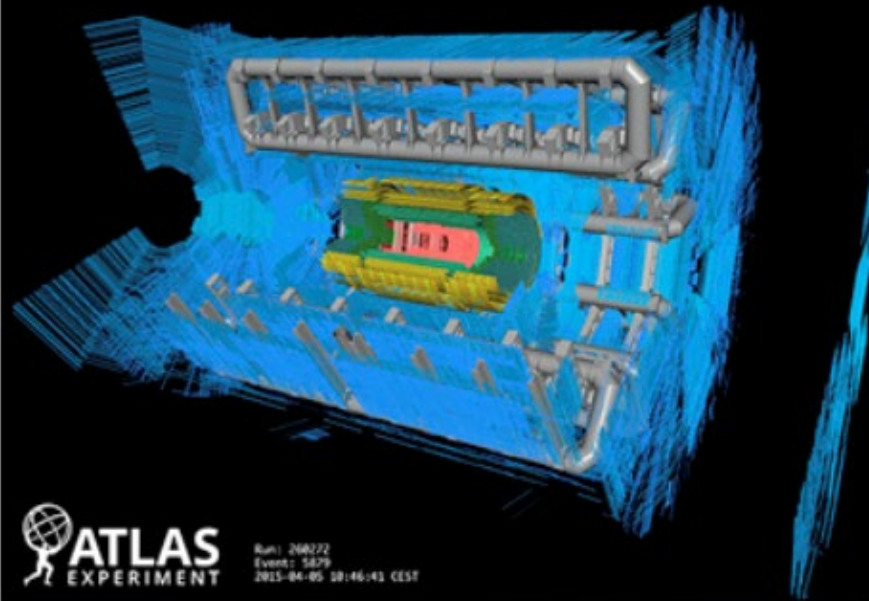
- Conduct a deep survey over an enormous area of sky with a frequency that enables images of every part of the visible sky to be obtained every few nights
- Continue in this mode for ten years to achieve astronomical catalogs thousands of times larger than have ever previously been compiled
- Deliver 500PB set of images and data products



## Four Science areas:

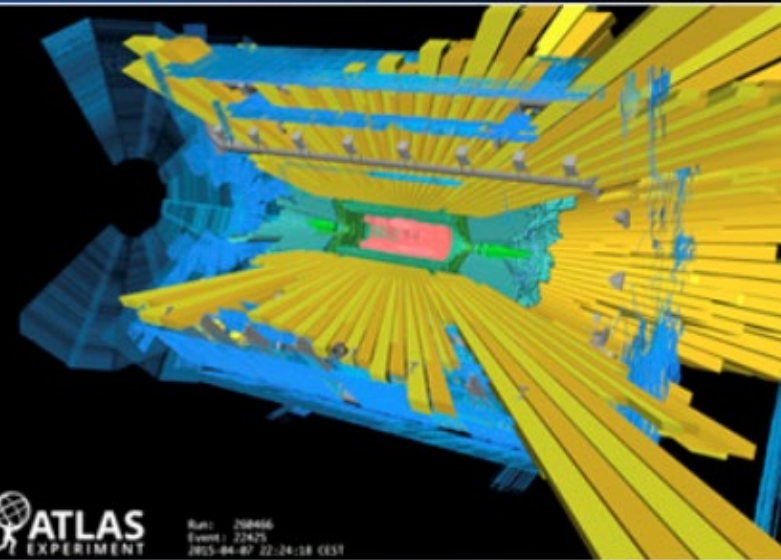
- Probing dark energy and dark matter
- Taking an inventory of the solar system
- Exploring the transient optical sky
- Mapping the Milky Way

## LHC and ATLAS Restart



ATLAS Is Ready and Waiting for Collisions

## ATLAS News



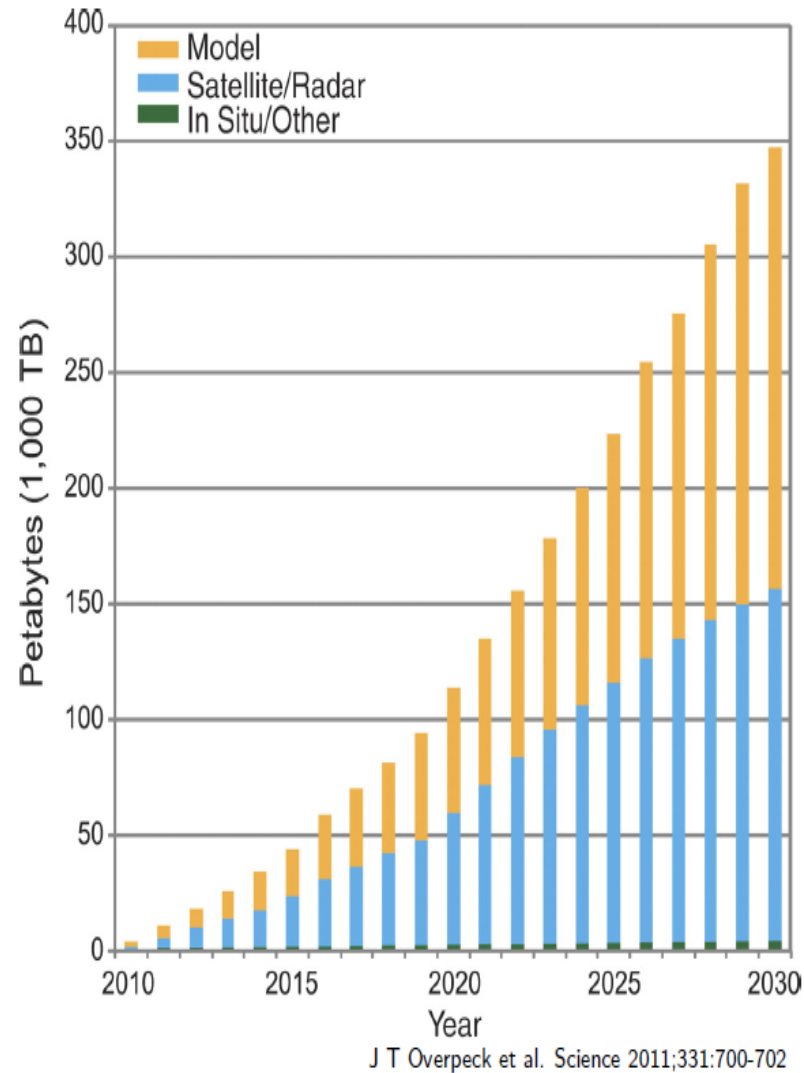
### Splashes for Synchronization

ATLAS uses "beam splash" events to provide simultaneous signals to large parts of the detector, and verify that the readout of different detectors elements are fully synchronized. [More...](#)

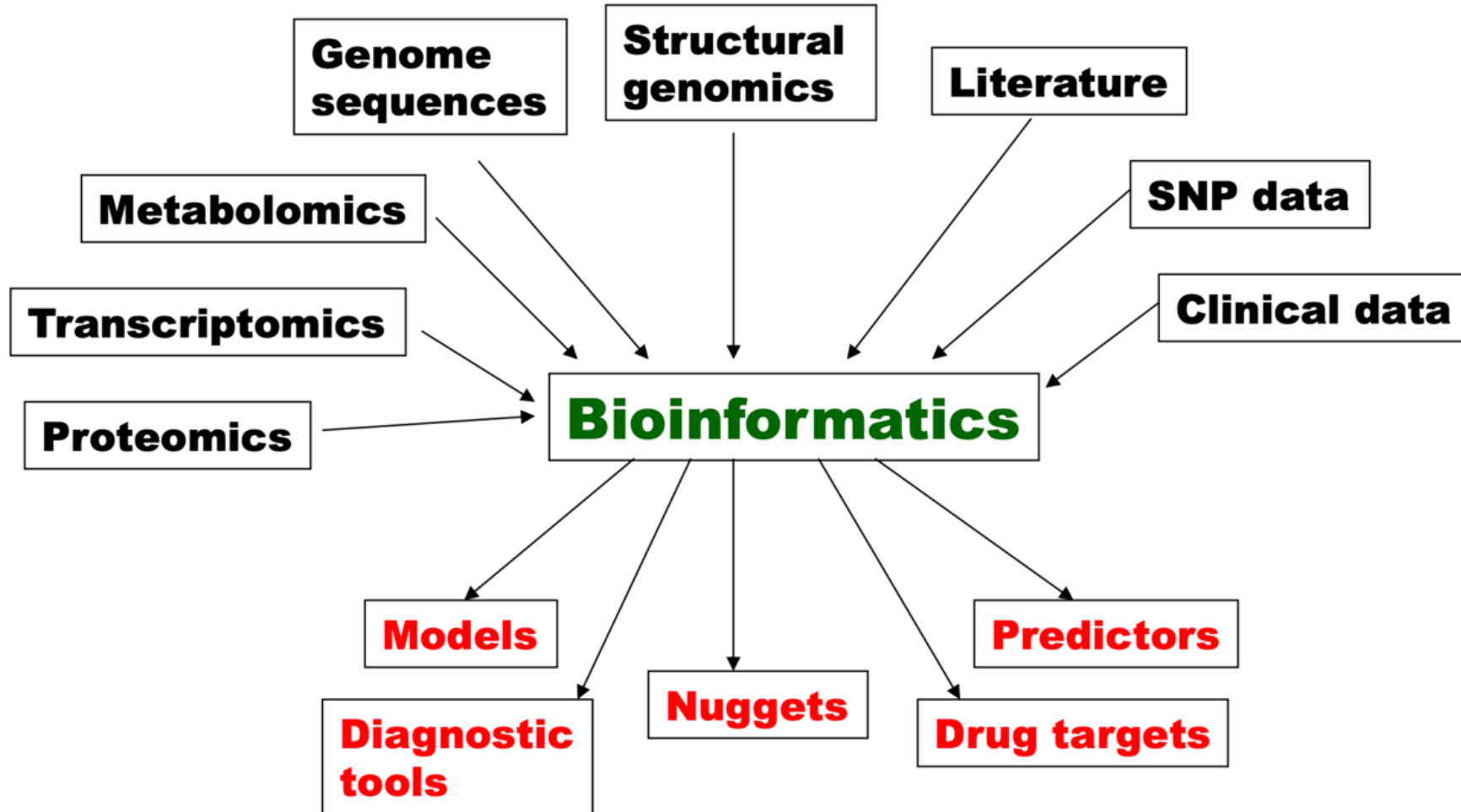
# Growth in Worldwide Climate Data

Fig. 2 The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you're not a climate scientist.

(BNL: Even if you are?)



# Explosion of data in Bioinformatics



# Background: The Deep Learning Revolution in AI

# Many Machine Learning Methods

K-means clustering

Markov random fields

Bayesian networks

Linear regression

Kalman filters

Random forests

Principal Component Analysis

Neural  
networks

Support Vector Machines

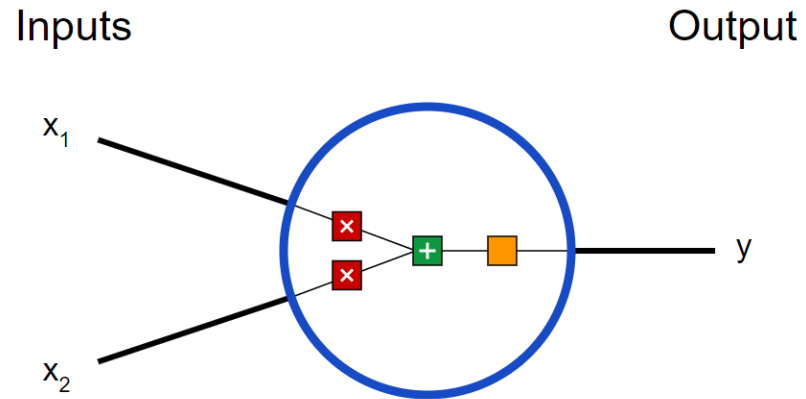
Boltzmann machines

Decision trees

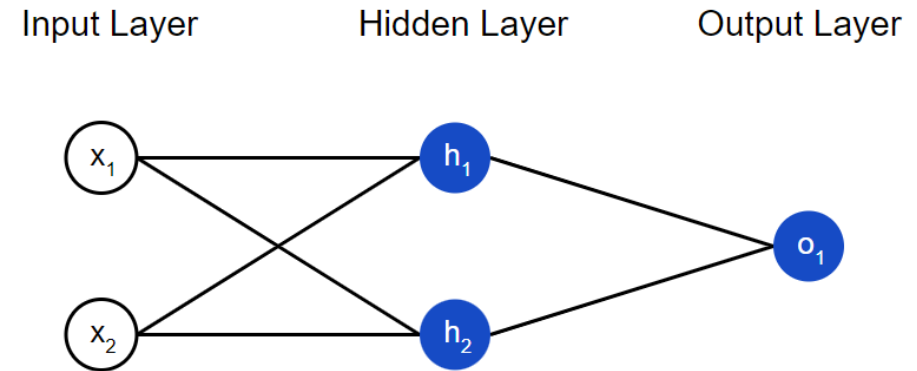
Radial basis functions

Hidden Markov Models

# Artificial Neurons and Neural Networks



An artificial neuron with two inputs, two weights and one output. The neuron only 'fires' if the combined inputs each multiplied by their weight is above a specified threshold value.



A three-layer neural network with one hidden layer. This is a 'feed forward' network in which the signals from the neurons only travel in one direction.



# IMAGENET

- ImageNet is an image dataset organized according to WordNet hierarchy. There are more than 100,000 WordNet concepts.
- ImageNet provides 1000 images of each concept that are quality-controlled and human-annotated.
- In competitions, ImageNet offers tens of millions of sorted images for concepts in the WordNet hierarchy.



What do these images have in common? *Find out!*

Check out the [ImageNet Challenge 2017](#)

- The ImageNet dataset has proved very useful for advancing research in computer vision

# Image Recognition Challenge to Computer Vision Research Groups



flamingo



cock



ruffed grouse



quail



partridge



Egyptian cat



Persian cat



Siamese cat



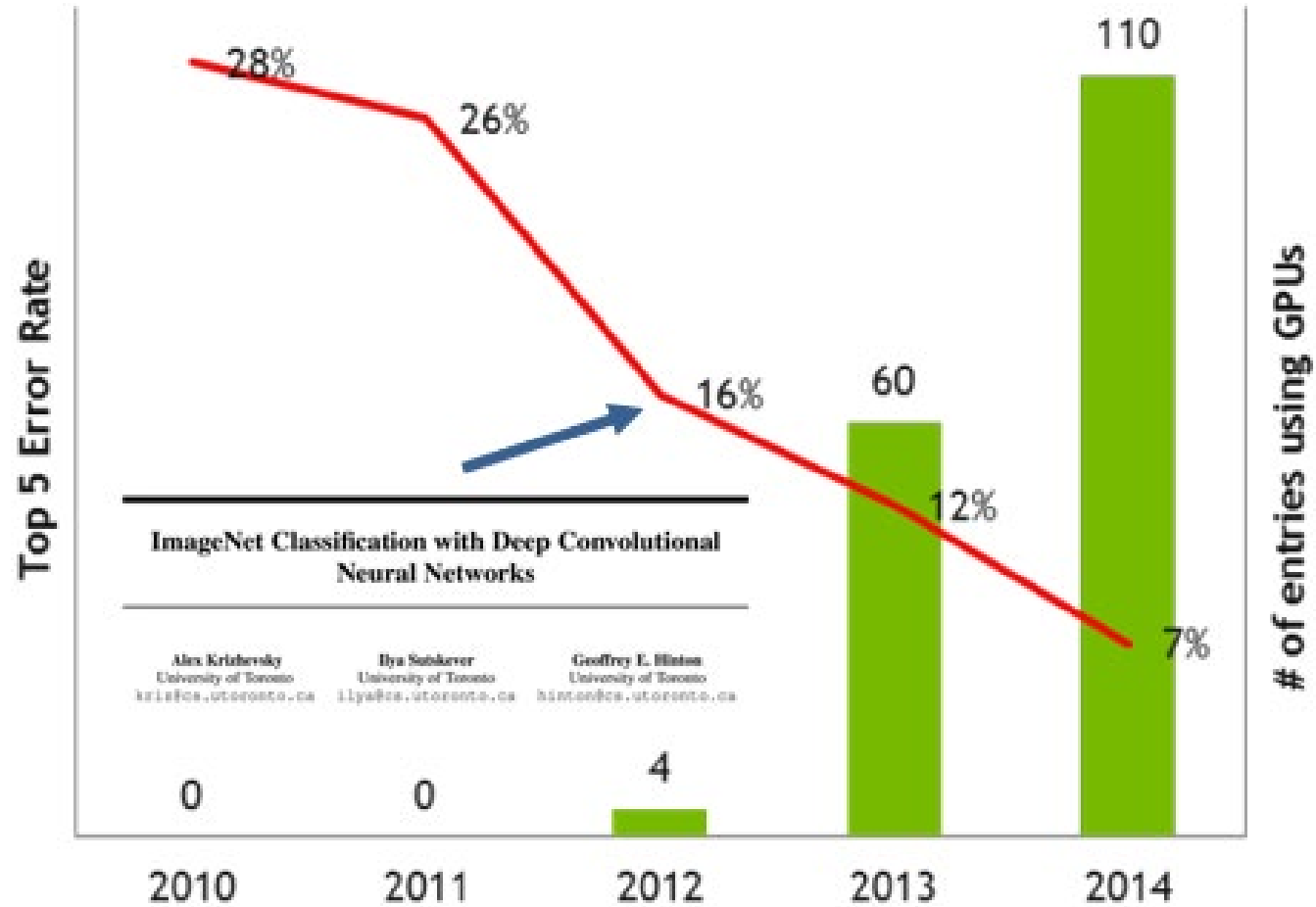
tabby



lynx

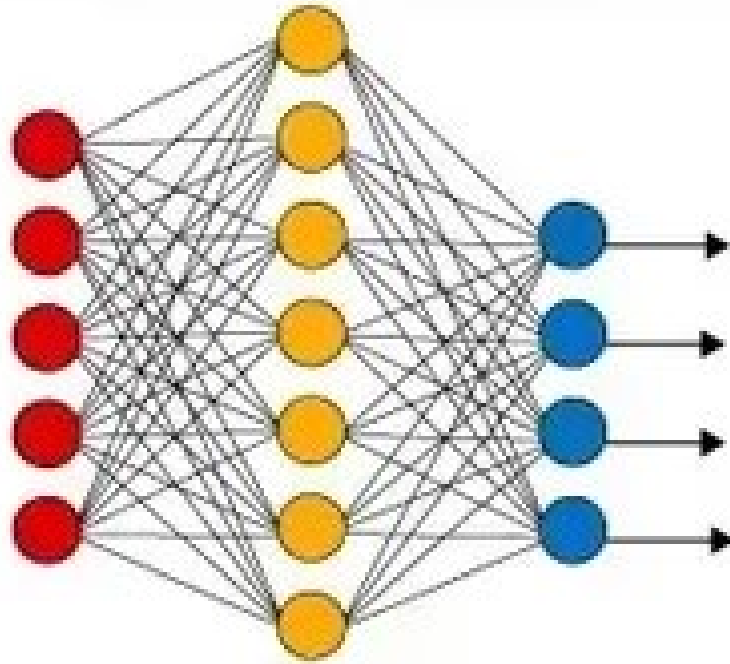
ImageNet: 1000 categories, 1.2 million images

# IMAGENET

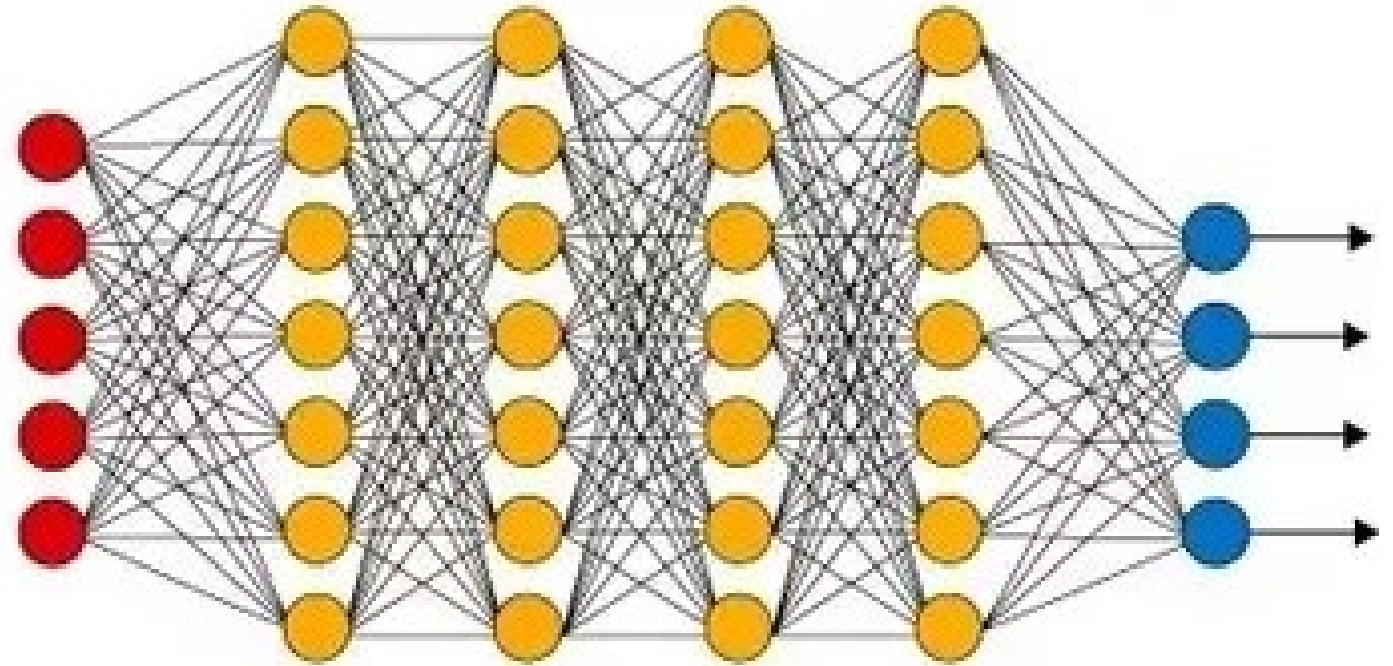


# Deep Learning Neural Networks

Simple Neural Network



Deep Learning Neural Network

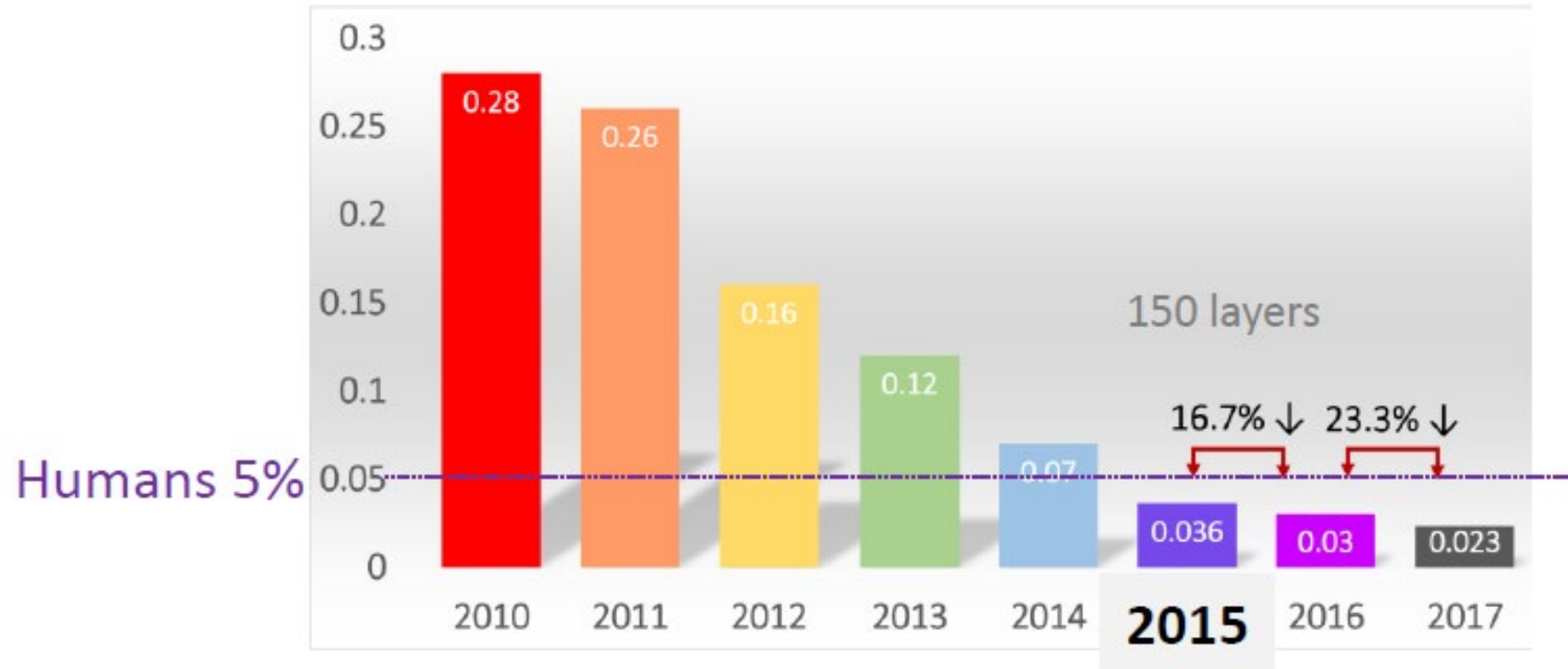


● Input Layer

● Hidden Layer

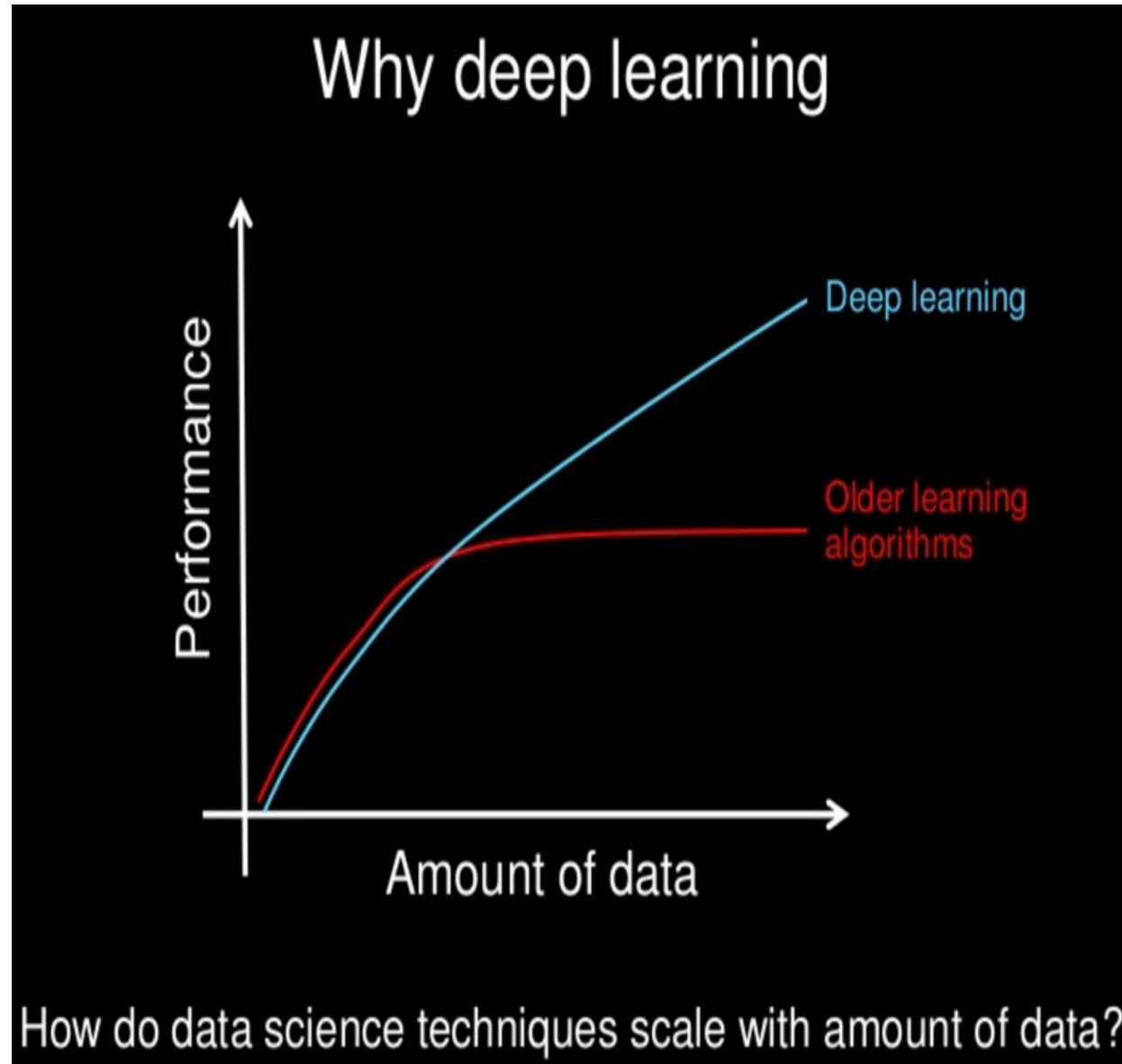
● Output Layer

# Classification Error Rate



Deep Learning error rates now lower than humans

# Why Deep Learning?





# Deep Learning now used routinely by major IT hyperscaler companies

- Near-human-level image classification
- Near-human-level speech recognition
- Near-human-level handwriting transcription
- Improved machine translation
- Improved text-to-speech conversion
- Ability to answer natural-language questions
- Near-human-level autonomous driving
- Superhuman Go playing
- Human level image analysis for Cancer diagnosis

1. Google



Illustration of a Google data center in Gresham, Ore. (The Information)

2. Microsoft



Illustration of Microsoft's largest data center in Redmond, Wash. (The Information)

3. Amazon



Illustration of Amazon's largest data center in Ashburn, Va. (The Information)

4. Meta

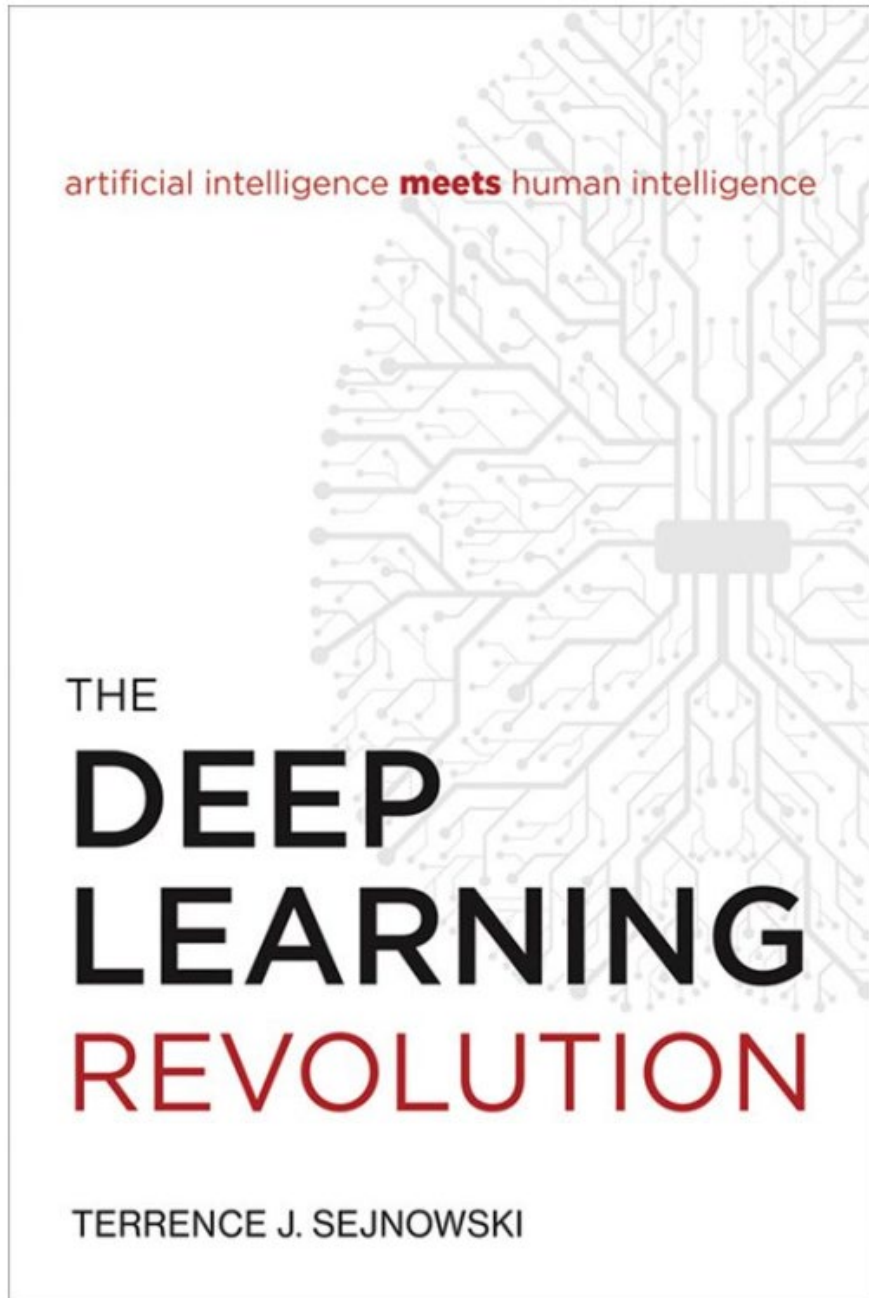


Illustration of Meta's largest data center in Prineville, Ore. (The Information)

5. Apple



Illustration of Apple's largest data center in Phoenix, Ariz. (The Information)



“What made deep learning take off was big data. ... The explosion of data is having an influence not just on science and engineering but also on every area of society.”

Terry Sejnowski



Terry Sejnowski and Geoffrey Hinton in 1980



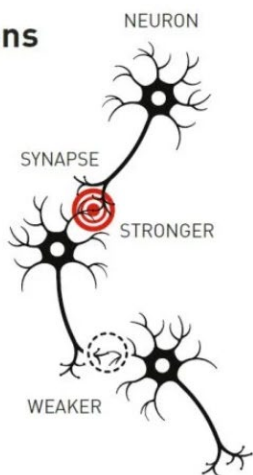
# The 2024 Nobel Prize for Physics

To John Hopfield and Geoffrey Hinton

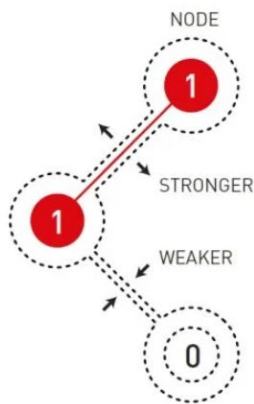
*“for foundational discoveries and inventions that enable machine learning with artificial neural networks”*

## Natural and artificial neurons

The brain's neural network is built from living cells, neurons, with advanced internal machinery. They can send signals to each other through the synapses. When we learn things, the connections between some neurons get stronger, while others get weaker.



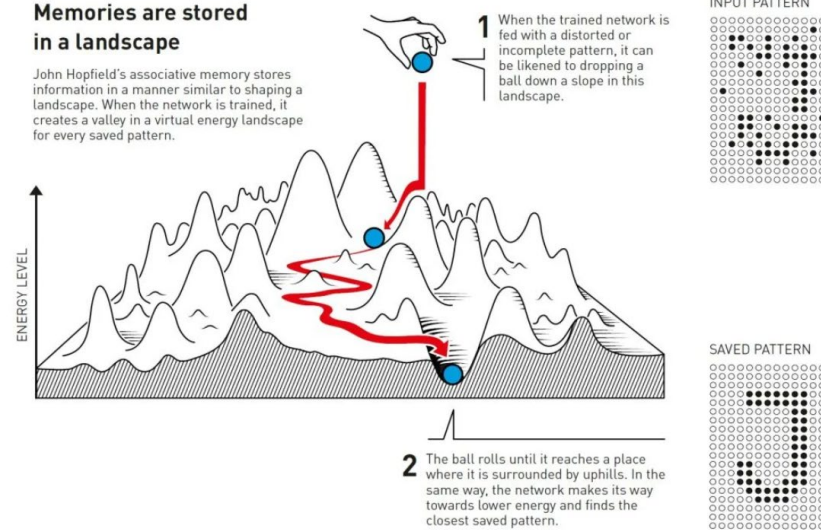
Artificial neural networks are built from nodes that are coded with a value. The nodes are connected to each other and, when the network is trained, the connections between nodes that are active at the same time get stronger, otherwise they get weaker.



The Hopfield network can be used to recreate data that contains noise or which has been partially erased.

## Memories are stored in a landscape

John Hopfield's associative memory stores information in a manner similar to shaping a landscape. When the network is trained, it creates a valley in a virtual energy landscape for every saved pattern.



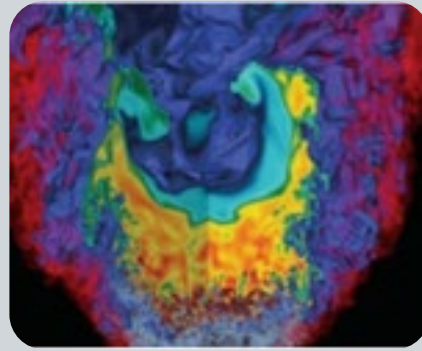
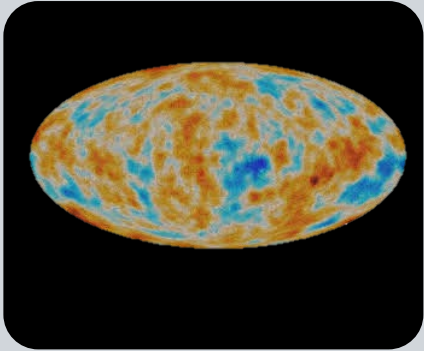
AI for Science:  
In the US and in the UK

# Lawrence's Successful Legacy of Team Science



Radiation Lab staff on the magnet yoke for the 60-in cyclotron, 1938, including:  
E. O. Lawrence  
Edwin McMillan  
Luis Alvarez  
J. Robert Oppenheimer  
Robert R. Wilson

# AI for Science at the US National Labs



## Data Sets

*Curated  
Data for  
Science*

## HPC

*Most  
powerful  
computers  
for science*

## Math and CS Research

*Science  
inspired  
foundations*

## User Facilities

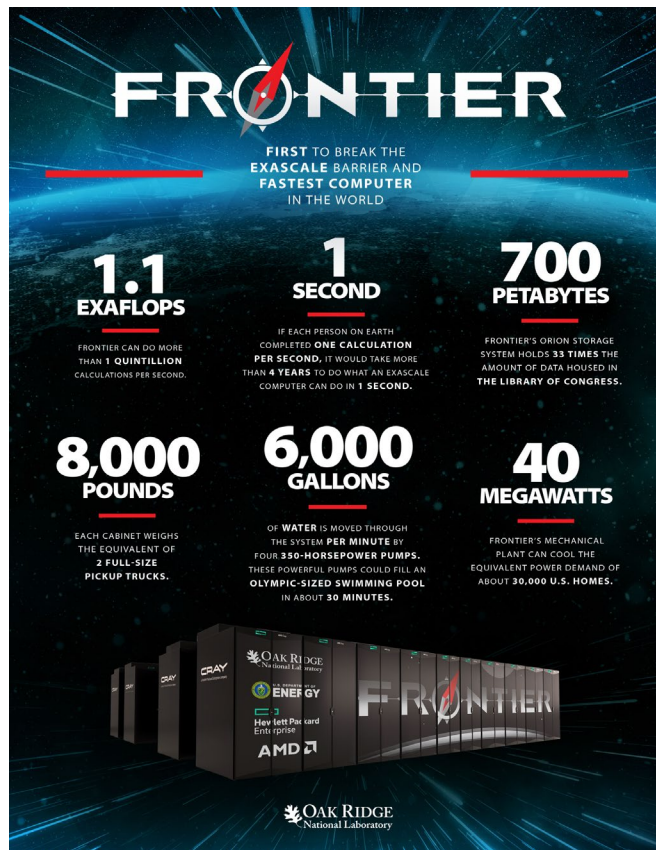
*Observation  
and  
Experiment*

## Team Science

*End-to-end  
science  
solutions*



# Frontier - First Exascale Supercomputer: Oak Ridge National Laboratory



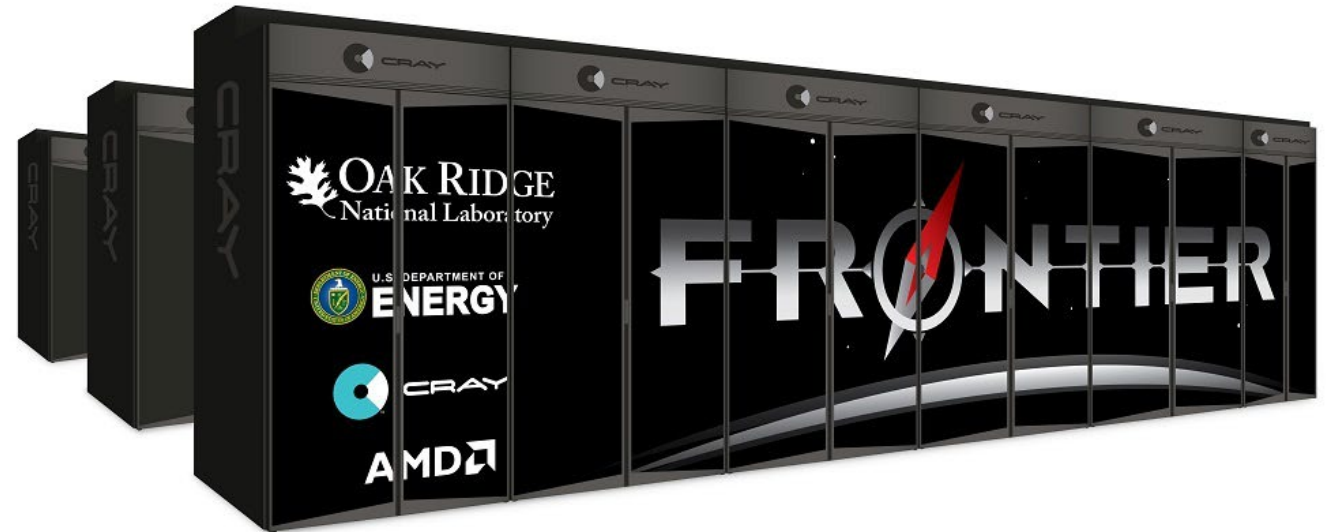
**FRONTIER**

FIRST TO BREAK THE EXASCALE BARRIER AND FASTEST COMPUTER IN THE WORLD

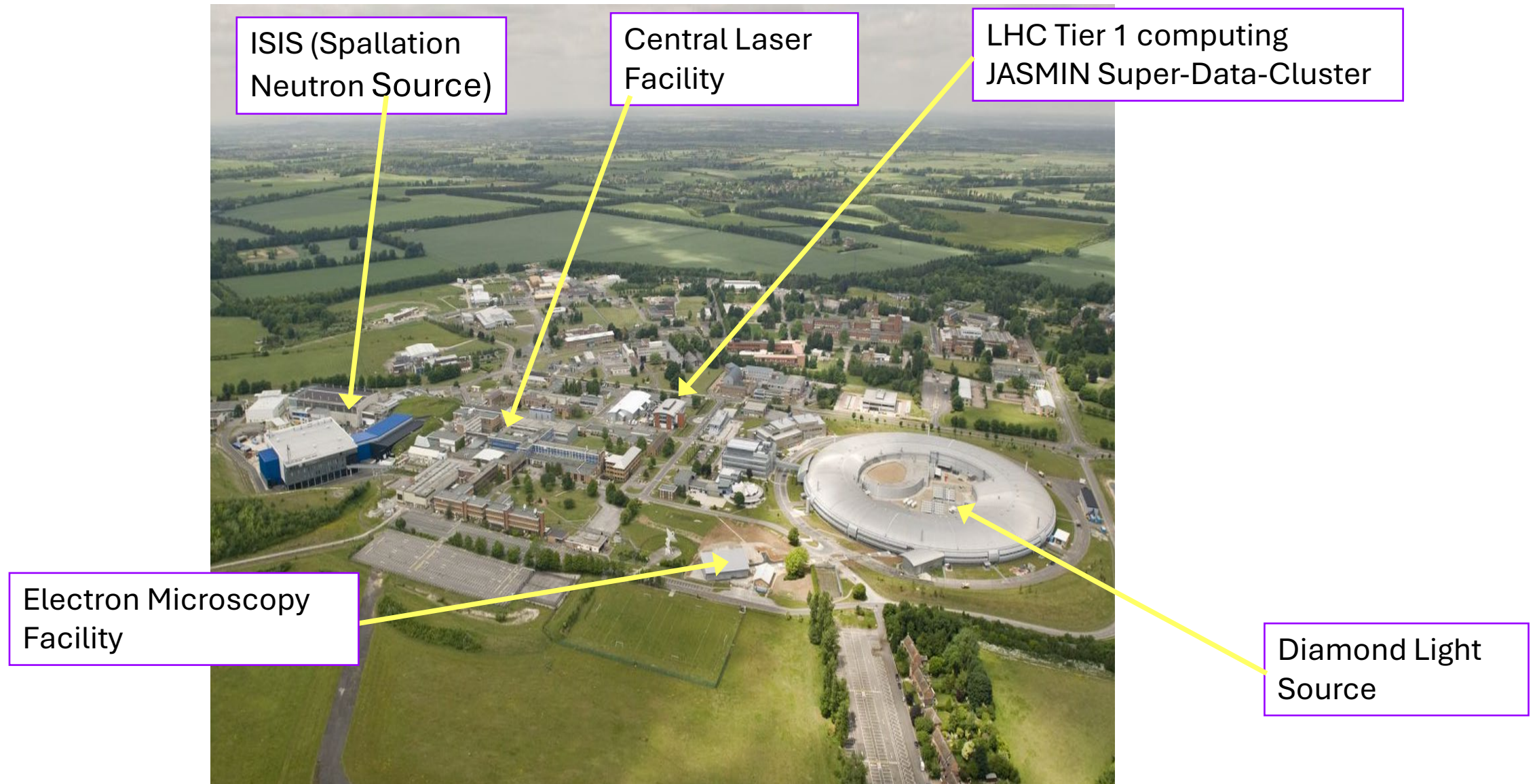
<b>1.1 EXAFLOPS</b> FRONTIER CAN DO MORE THAN 1 QUINTILLION CALCULATIONS PER SECOND.	<b>1 SECOND</b> IF EACH PERSON ON EARTH COMPLETED ONE CALCULATION PER SECOND, IT WOULD TAKE MORE THAN 4 YEARS TO DO WHAT AN EXASCALE COMPUTER CAN DO IN 1 SECOND.	<b>700 PETABYTES</b> FRONTIER'S ORION STORAGE SYSTEM HOLDS 33 TIMES THE AMOUNT OF DATA HOUSED IN THE LIBRARY OF CONGRESS.
<b>8,000 POUNDS</b> EACH CABINET WEIGHS THE EQUIVALENT OF 2 FULL-SIZE PICKUP TRUCKS.	<b>6,000 GALLONS</b> OF WATER IS MOVED THROUGH THE SYSTEM PER MINUTE BY FOUR 350-HORSEPOWER PUMPS. THESE POWERFUL PUMPS COULD FILL AN OLYMPIC-SIZED SWIMMING POOL IN ABOUT 30 MINUTES.	<b>40 MEGAWATTS</b> FRONTIER'S MECHANICAL PLANT CAN COOL THE EQUIVALENT POWER DEMAND OF ABOUT 30,000 U.S. HOMES.

OAK RIDGE National Laboratory  
ENERGY  
Heavy Ion Proton Accelerator  
AMD  
CRAY

OAK RIDGE National Laboratory



# Large-scale Experimental Facilities at the Rutherford Appleton Lab in the UK

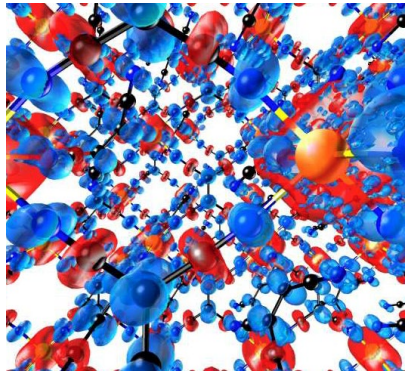




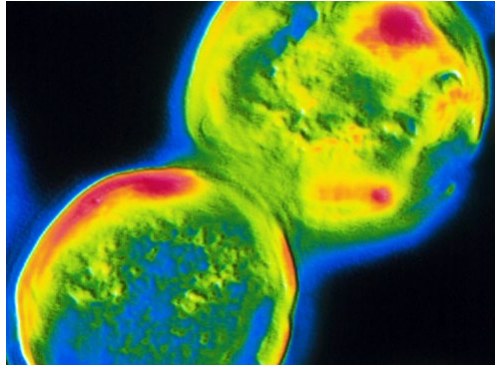
# The Scientific Machine Learning Group

Group Leader: Jeyan Thiyagalingam

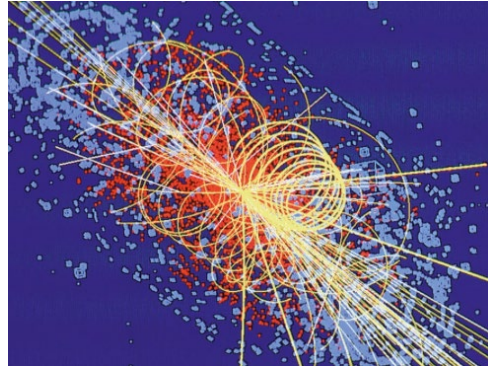
## ► Vision of *AI for Science*



*Material Sciences*



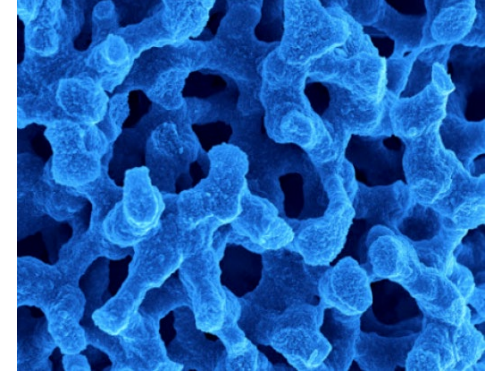
*Environmental Sciences*



*Particle Physics*



*Astronomy*

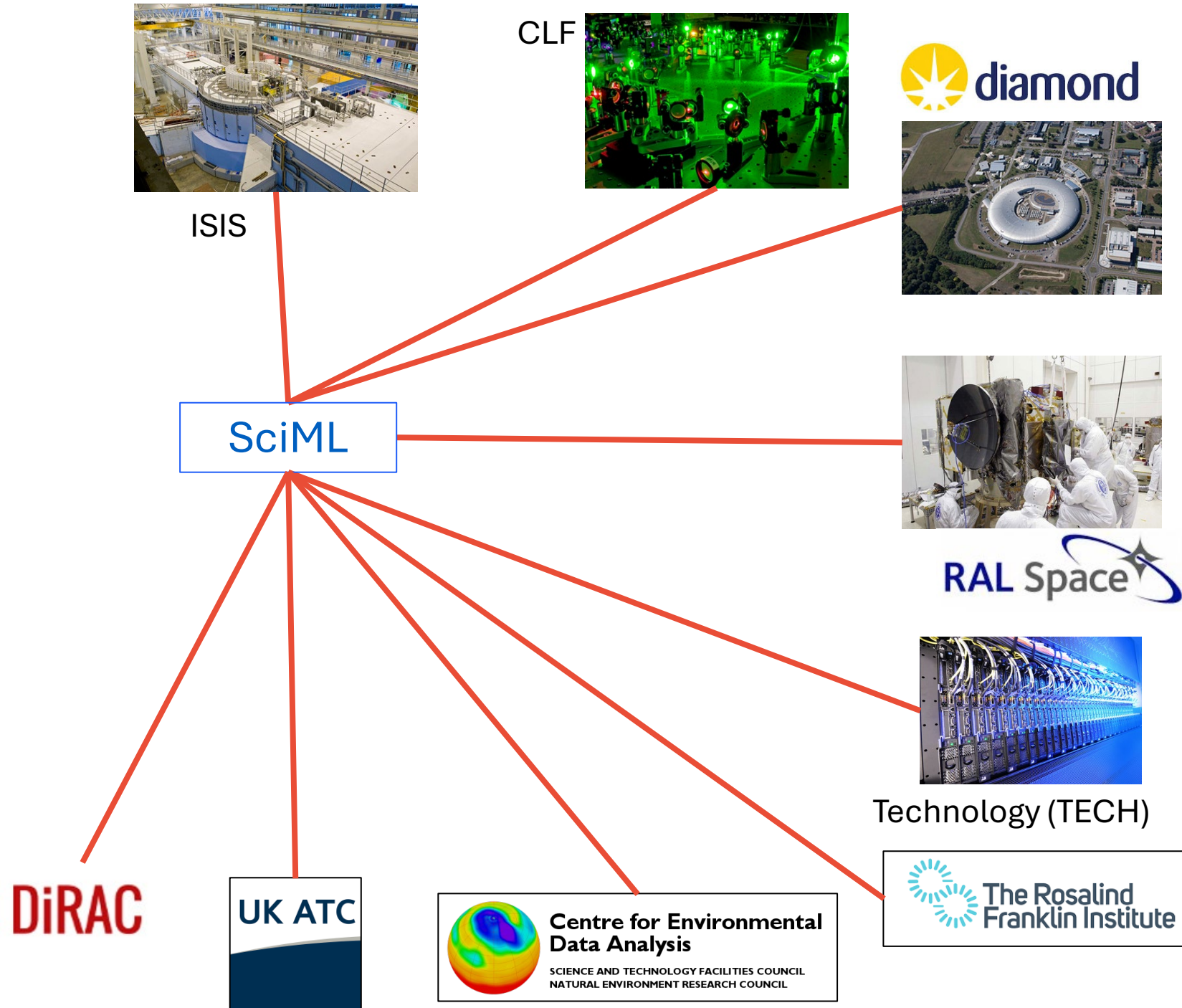


*Life Sciences*

## ► Part of the Scientific Computing Department at the STFC's National Laboratories

# SciML's Role

- Work with other national labs, facilities and STFC funded programmes
- Develop and apply AI technologies to make a difference in Science





# Open Access and Open Science

# Open Access: The Southampton Connection

1994 to 2024  
Thirty years in the trenches for OA

Some acknowledgements

- Stevan Harnad, Tony Hey, Jessie Hey
- Wendy Hall, Les Carr, Rob Tansley
- Mark Brown, Pauline Simpson, Wendy White
- Chris Gutteridge, Steve Hitchcock and many others ...

# Stevan Harnad as ‘influencer’ ...

In an ideal world of scholarly communication – all research is freely available

**Preserv** Preservation Eprint Services

- June 27<sup>th</sup> 2005 11<sup>th</sup> anniversary of Stevan Harnad’s ‘Subversive Proposal’ leading to the open access vision for scholarly material
- See also Harnad, S. and Hey, J. M. N. (1995) Esoteric Knowledge: the Scholar and Scholarly Publishing on the Net. In *Proceedings of Networking and the Future of Libraries 2: Managing the Intellectual Record, Proceedings of an International Conference, Bath, 19-21 April 1995*, 110-16. Dempsey, L., Law, D. and Mowlat, I., Eds.
- And journals still become more and more expensive



Even the work of researchers in our own institution is still often unavailable to us  
..... but we’re making progress

Harnad on the ‘Faustian Bargain’ in 1997:

‘If you wish to immortalize your words at all, you will have to surrender your copyright in exchange, so that your publisher can recover the substantial cost of getting your intellectual goods aboard the paper flotilla at all. The author must collaborate in denying access to his adverts to anyone who (or whose library) has not paid for them.’

With thanks to Jessie Hey

# eScience, Scholarly Communication and the Transformation of Research Libraries

**Tony Hey**

**Corporate Vice President for External Research  
Microsoft Research**

From a presentation to NAS in 2008

# Current Scholarly Publishing Model is in Crisis

- Journal subscriptions rising faster than library budgets
  - Cancelling subscriptions, no freedom for new journals in new and emerging fields
- Web technology and digital media now make dissemination of knowledge 'easy' and 'free' without the traditional paper journals
  - Similar dilemma to that of the music industry with MP3 and 'free' digital copies
- Curious 'crisis' in that the average academic is often unaware of these issues

# Open Access Research Repositories

- As Dean of Engineering at Southampton I was responsible for monitoring the research output of over 200 Faculty and 500 Post-Docs and Grad Students
  - University library could not afford to subscribe to all the journals that my staff published in, not to mention conference proceedings and workshop contributions ...
- Repositories will contain not only full text versions of research papers but also 'grey' literature such as workshop papers, presentations, technical reports and theses
  - In the future it is likely that repositories will also contain data, images and software

# Rob Tansley: Architect of EPrints and DSpace



## Senior Research Scientist

Hewlett Packard Laboratories  
2000 - May 2006 · 6 yrs 5 mos  
Greater Boston Area

Architect, DSpace open source digital asset management system  
Architect, China Digital Museum project  
Research lead, Digital Preservation



## Research Fellow

Southampton University  
1999 - 2000 · 1 yr  
Southampton, United Kingdom

Single-handedly launched EPrints digital document repository  
360-degree role: design, development, local service rollout, global installation package rollout, end user and technical support

## Education



### University of Southampton

PhD, Computer Science  
1996 - 2000



### University of Southampton

BSc, Computer Science  
1993 - 1996

## EPrints for Open Access

EPrints has long history in providing institutions with what it needs to publish and promote its research outputs on the Web.

[Read more](#)

## EPrints for Education

Building on the success of the educational content sharing platform EdShare, EPrints can supply a flexible platform to support your staff and students in engaging with open education practices.

[Read more](#)

## EPrints for Research Data

Building on the foundations of EPrints, coupled with community driven extensions, EPrints for your RDM solution

[Read more](#)

## EPrints for Dataset Showcases

We have leveraged the EPrints platform to supply a flexible framework to present and preserve the research output from your project.

## EPrints for REF2029

Building on the success of our REF 2014 & 2021 packages, EPrints can help institutions meet their next REF requirements.

## Building Repositories

We can build repositories that are configured to meet the particular requirements of your organisation. We work with clients to create repositories for research publications, open education resources, multimedia outputs and research data sets.

## Repository Hosting

We offer fully managed and supported EPrints based repository services at our commercially run hosting providers.

## Consultancy

We offer EPrints consultancy. We can help you set up your repository, upgrade it, or help you through a complex project.



## The DSpace Vision and Mission Statement



### Vision

The DSpace Project will produce the world's choice for repository software providing the means for making information openly available and easy to manage.



### Mission

We will create superior open source software by harnessing the skills of an active developer community, the energy and insights of engaged and active users, and the financial support of project members and registered service providers.



### DSpace Software Will:

- Focus on the Institutional Repository use case.
- Be lean, agile, and flexible.
- Be easy and simple to install and operate.
- Include a core set of functionality that can be extended to or integrated with complementary services and tools in the larger scholarly ecosystem.

# Progress towards Open Science: In the UK and in the US

# Governments are committed to open data and research

## UK National Data Library: Technical White Paper Challenge

Wellcome and the Economic and Social Research Council (ESRC) are seeking technical visions and architectures for a UK National Data Library to make public sector datasets more accessible to researchers and enable future science to thrive.

[www.wellcome.org/what-we-do/our-work/uk-data-library](http://www.wellcome.org/what-we-do/our-work/uk-data-library)

**“The G7 will collaborate in expanding open science with equitable dissemination of scientific knowledge and publicly funded research outputs including research data and scholarly publications in line with the Findable, Accessible, Interoperable, and Reusable (FAIR) principles.”**

G7 Science and Technology Ministers’  
Communique (Sendai, May 2023)



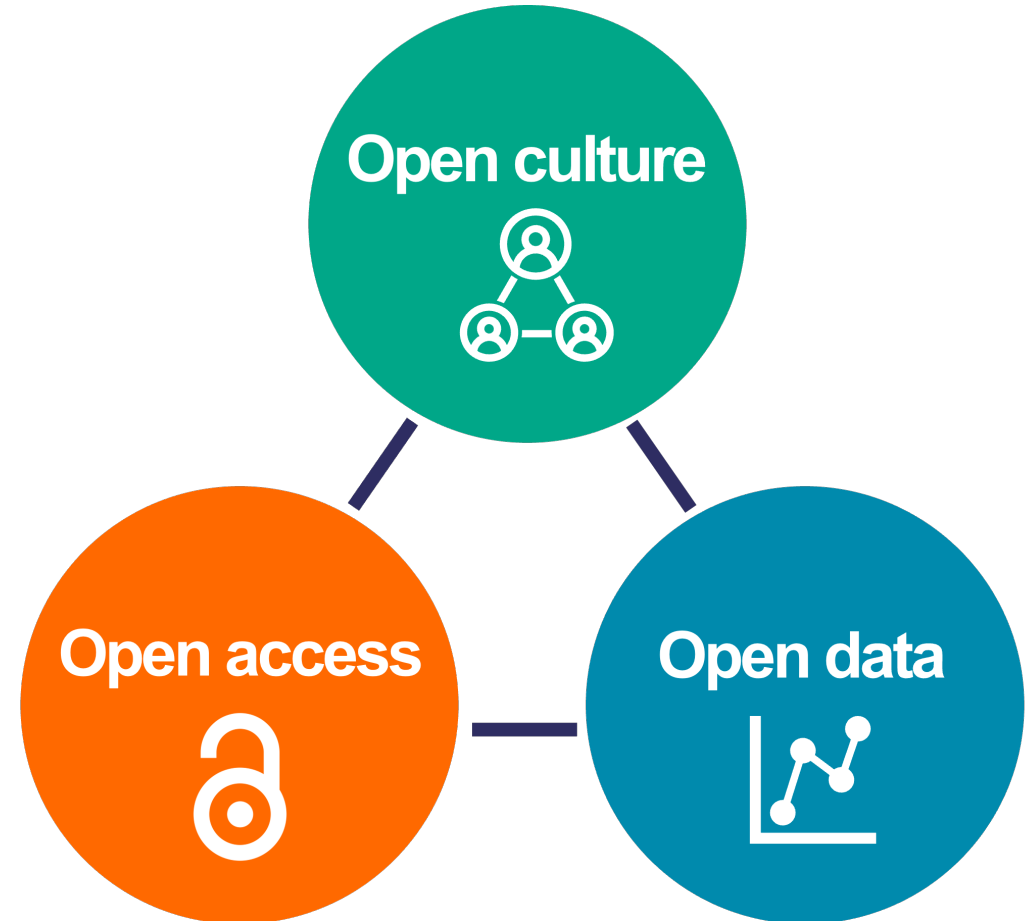
**Recommendation of the Council concerning Access to Research Data from Public Funding**

<https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347>

Slide courtesy of Rachel Bruce  
Head of Open Research  
UK Research and Innovation (UKRI)

# UKRI championing open research

- Introduced our **new open access policy** and funding for research articles and long-form outputs.
- **Developing our open data policy** to incentivise research data sharing, and where appropriate protocols, software, and code to support re-use, research integrity and collaboration.
- We are in a discovery phase to inform the best approach to progress our open data priority.
- Ambition is for a more harmonised pan-UKRI policy framework but one that recognises and supports disciplinary diversity and interdisciplinary.



Slide courtesy of Rachel Bruce  
Head of Open Research  
UK Research and Innovation (UKRI)

# Free, fast and fair: the global primary research record where researchers publish their work in full detail

Octopus is a new publishing platform for scholarly research. Funded by UKRI – the UK government research funder.

Here researchers can publish all their work for free, in full detail, enabling peer review and quality assessment, gaining credit for what they have done, and allowing the research community to build upon it.

[Learn more](#)[Author Guide](#)[Find Publications](#)

# Jisc Review of Transitional Agreements (2024)

- Transitional agreements, adopted by Jisc and UK institutions alongside the global research community were devised specially for hybrid journals operating both subscription and Open Access publishing models.
- Envisioned as a temporary mechanism to support publishers with transitioning titles to fully OA, they have the dual aim to “bring institutional investments in scholarly journal publishing under oversight and control, with an eye to cost reduction, and to drive a transition of scholarly journal publishing to Open Access”.
- In 2022, the proportion of UK Open articles was 4% higher when compared with the proportion of global open articles.
- UK OA articles accounted for 65% of UK output (including Gold, Hybrid and Green) with a continual increase in absolute numbers and proportions of Open articles over the last eight years.

# A View from the Library at Southampton (1)

## Open Access

- Use Elsevier's PURE as the research information system. Academics deposit their accepted manuscripts in PURE and these are then passed to the ePrints repository
- The University has around 20 transitional 'Read-and-Publish' deals. These are negotiated by Jisc and allow gold OA at no extra cost and are meant to aid the publishers' transition to full OA and the elimination of hybrid journals
- The transformative agreements have constrained off-the-scale increases in costs but still almost always rising above inflation. These are becoming truly unaffordable and we are starting to see actual walk-aways from these big deals

# A View from the Library at Southampton (2)

## Research Data

- Authors are increasingly linking their papers to relevant datasets
- Only a minority of authors send the Library their Data Management Plans for checking
- The Library provides training and support with metadata and storage of their data
- We use the Digital Curation Centre (DCC) in Edinburgh to provide training for our librarians
- Also engage with the RDA and Digital Preservation Coalition (DPC) as well as European initiatives such as EOSC, FAIR and Plan S

With thanks to Wendy White and Suzanne Tatham





RESEARCH DATA ALLIANCE

**The Research Data Alliance: a window the world of  
global data**

# The RDA Outputs

The global RDA community has produced over 200 standards, guidelines, best practices for different domains, infrastructures, research data management challenges. These outputs are implemented across academic, research performing, industry, private sector, funding and policy organisations as well as digital research, data and research infrastructures across the globe.

# Comments on the RDA from Hilary Hanahoe

## RDA Secretary General

- The RDA outputs cover many domains and disciplines, many areas of data management in specific institutional or disciplinary contexts.
- The success is that:
  1. The global community sees the value of an independent global platform to develop and make available RDM solutions. They are walking the walk of open science.
  2. RDA has managed to maintain its position as the only global, multidisciplinary, open initiative dealing with RDM and Open Science where all are welcome and all can get involved as well as take up the open solutions
  3. The data challenges are far from being resolved and, indeed, a forum like RDA where these challenges can be addressed is only more relevant now than it was 11 years ago when it was launched.
  4. The delivery of over 200 outputs, by groups that have an average of 70 global experts means that the distribution and ripple effect of RDA is immeasurable.

# Nelson Memorandum on Making Federally-funded Research Freely Available Without Delay

## OSTP August 2022 (1)

OSTP recommends that federal agencies:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible **without an embargo on their free and public release**
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

# Nelson Memorandum on Making Federally-funded Research Freely Available Without Delay

## OSTP August 2022 (2)

Federal agencies should, consistent with applicable law:

- Collect and make publicly available appropriate metadata associated with scholarly publications and data resulting from federally funded research, to the extent possible at the time of deposit in a public access repository.
- Such metadata should include at minimum:
  1. All author and co-author names, affiliations, and sources of funding, referencing digital persistent identifiers, as appropriate
  2. The date of publication
  3. A unique digital persistent identifier for the research output

# Comments from Peter Suber

## Director of the Harvard Open Access Project (1)

- US and UK approaches on OA have diverged since 2006 when the UK recommended gold OA policies
- Transformative agreements or 'Read-and-Publish' agreements with publishers have grown faster in UK and Europe than in the US where they are still widely unpopular
- Although the OSTP Nelson memo in 2022 requires all Federal Funding Agencies to adopt an OA policy with no embargo period by the end of 2025, there are many US university policies based on Green OA repositories since the Funding Agencies were too slow to act since the first OSTP memo on OA in 2013

# Comments from Peter Suber

## Director of the Harvard Open Access Project (2)

- Read-and-Publish agreements do not, in my view, solve the problem of rising journal prices. They do now make more new journal articles OA, which is good, but prices are rising and the charges are opaque.
  - Harvard has pioneered what are now called ‘rights-retention’ OA policies. These policies not only make new articles by affiliated authors OA through the institutional repository but they also assure that authors and the institution retain the rights to them OA.
- My opinion: A rights retention OA policy seems to be the best/only solution to getting to full open access.
- Here is an up-to-date list of universities with rights-retention OA policies, combining those that take the Harvard approach and those that take other approaches:

[https://oad.simmons.edu/oadwiki/University\\_rights-retention\\_OA\\_policies](https://oad.simmons.edu/oadwiki/University_rights-retention_OA_policies)

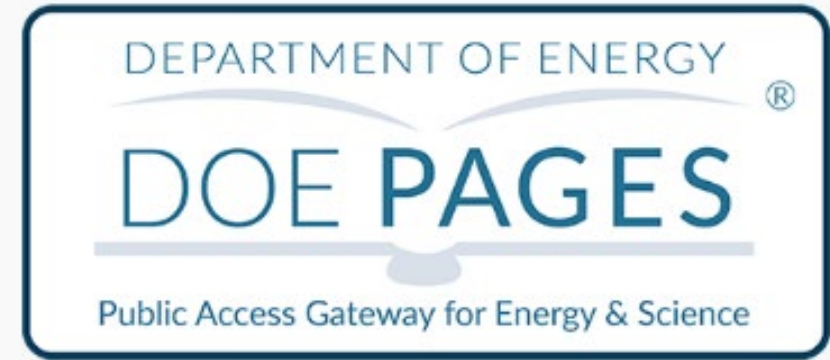
# Public Access Plan

‘Ensuring Free, Immediate and Equitable Access’ to the Results of Department of Energy Scientific Research



June 2023

<https://doi.org/10.11578/2023DOEPublicAccessPlan>



## 2023 DOE Public Access Plan Release

Key elements of the new DOE public access plan, as laid out by OSTP, will include elimination of any "embargo" period before the public gains free access to journal articles or final accepted manuscripts resulting from federal funding; immediate access to scientific data displayed in or underlying publications and expanded access to scientific data not displayed in publications; and broad adoption of persistent identifiers (PIDs) for research outputs, organizations, awards and contracts, and people.



# DOE PAGES and OSTI's Public Access Role

- The DOE PAGES service is competitive with NIH's PMC service as a solution for OA repositories based on Green OA
- OSTI is actively collaborating with the NSF and DOD funding agencies who are implementing their public access plans based on PAGES
- NSF has expanded the types of R&D outputs captured in the NSF Public Access Repository (NSF-PAR) beyond journal publications to also include metadata and links for data, software, and other outputs.
- This expanded role for NSF-PAR is similar to what OSTI has been doing for some time and is a major step towards 'open science'
- OSTI is also collaborating with the DOD to host a collection for accepted manuscripts from their grantees.

With thanks to Director Brian Hitson and OSTI staff

# Foundation Models and Large Language Models

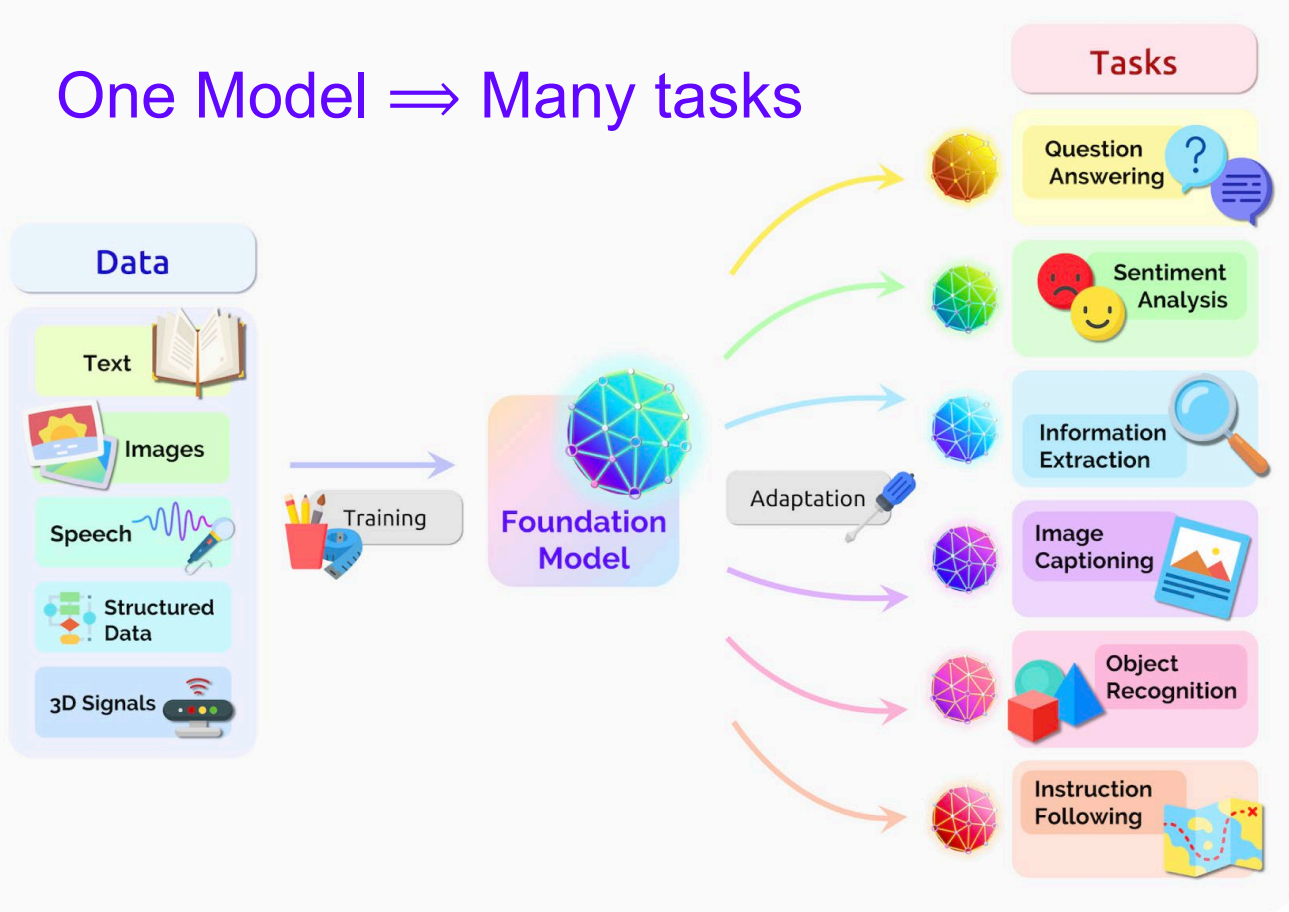
# Natural Language Processing in the Pre-Transformer Era

- Text is sequential – just series of words
- Recurrent Neural Networks (RNNs) were instrumental in the early days in which the output of a Deep Neural Network is fed back in as input to the next stage
- RNNs now replaced by ‘Transformer’ neural networks



# Foundation Models

One Model  $\Rightarrow$  Many tasks



Trained on trillions of input "tokens" for many weeks on a large-scale computers

GPT-4 has about 1 trillion parameters (1% brainscale)

Slide courtesy of Rick Stevens

# Foundation Models for Science — Opportunities

- FMs can summarize and distill knowledge – extract information from million of papers into compact computing representation
  - FMs can synthesize – combine information from multiple sources – generate small programs for specific tasks
  - FMs can generate plans, solve logic problems and write experimental protocols for robots
  - FMs with additional research, may be able to generate hypotheses to be tested and new theories for exploration
- A full-time scientific assistant that learns from across all of science?

# Difference between Foundation Models and Large Language Models (LLMs)

- Foundation models are general purpose and adaptable to a wide range of applications such as conversational AIs
    - Trained on books, articles, images, data, websites, etc.
  - LLMs are specifically trained for language-related tasks such as text generation, translation, etc.
    - Optimized for high accuracy and fluency in language tasks
- Answer from Microsoft Copilot ...

# ChatGPT on Backup and Digital Preservation

From an email from Adam Harwood,  
Research Data and Digital Preservation Technologist, University of Sussex, UK

Out of curiosity, I asked chatGPT what the difference between backup and digital preservation is. It came up with this analogy:

"Think of digital data like a rare and valuable book in a library.

A backup is like making photocopies of that book regularly. It ensures that if the original book is damaged, lost, or stolen, you have a copy that's relatively recent. Backups are like having spare copies on hand for quick access. IT professionals are great at managing these copies efficiently, making sure they are up-to-date and easily retrievable.

On the other hand, digital preservation is like being an archivist responsible for ensuring the book's survival and accessibility for generations. You're not just making copies; you're taking special measures to protect the original book and its copies over a very long time. This involves carefully storing the book in a controlled environment, documenting its history and significance, and occasionally restoring it or creating new copies in updated formats to keep it relevant and accessible. Digital preservation requires a more in-depth, long-term perspective, much like preserving historical records or artifacts.

In IT terms, backups are about data recovery in the short term, while digital preservation is about data stewardship and ensuring its availability and usability over the long term, even in a changing technological landscape."

I asked it if this was a quote or if it was something it made up. It said:

"I created the analogy provided in my previous response to help explain the difference between backup and digital preservation. It is not a direct quote from a specific source but rather a conceptual explanation based on common understanding."

# How can Academia compete with Industry on Machine Learning and AI?

Companies like Facebook, Google, Amazon, Microsoft, and Apple (and probably Baidu, Huawei, Alibaba and Tencent) and have three key advantages over academia:

1. These companies all have many, very large, private datasets that they will never make publicly available
2. Each of these companies employs many hundreds of computer scientists with PhDs in Machine Learning and AI
3. Their researchers and developers have essentially \$B's of computing power at their disposal

➤ FMs, LLMs, Machine Translation, Image Recognition, ...

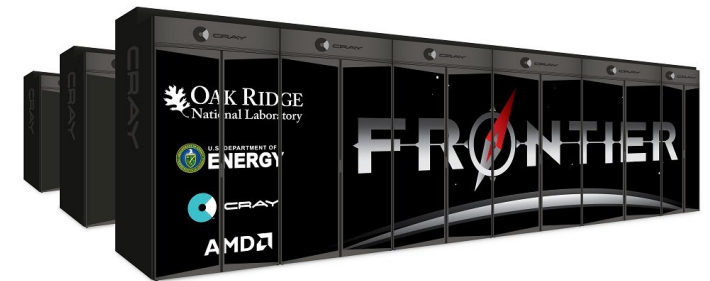


# FORGE: Pre-Training Foundation Models at ORNL

Team: Junqi Yin, Sajal Dash, Feiyi Wang, Arjun Shankar

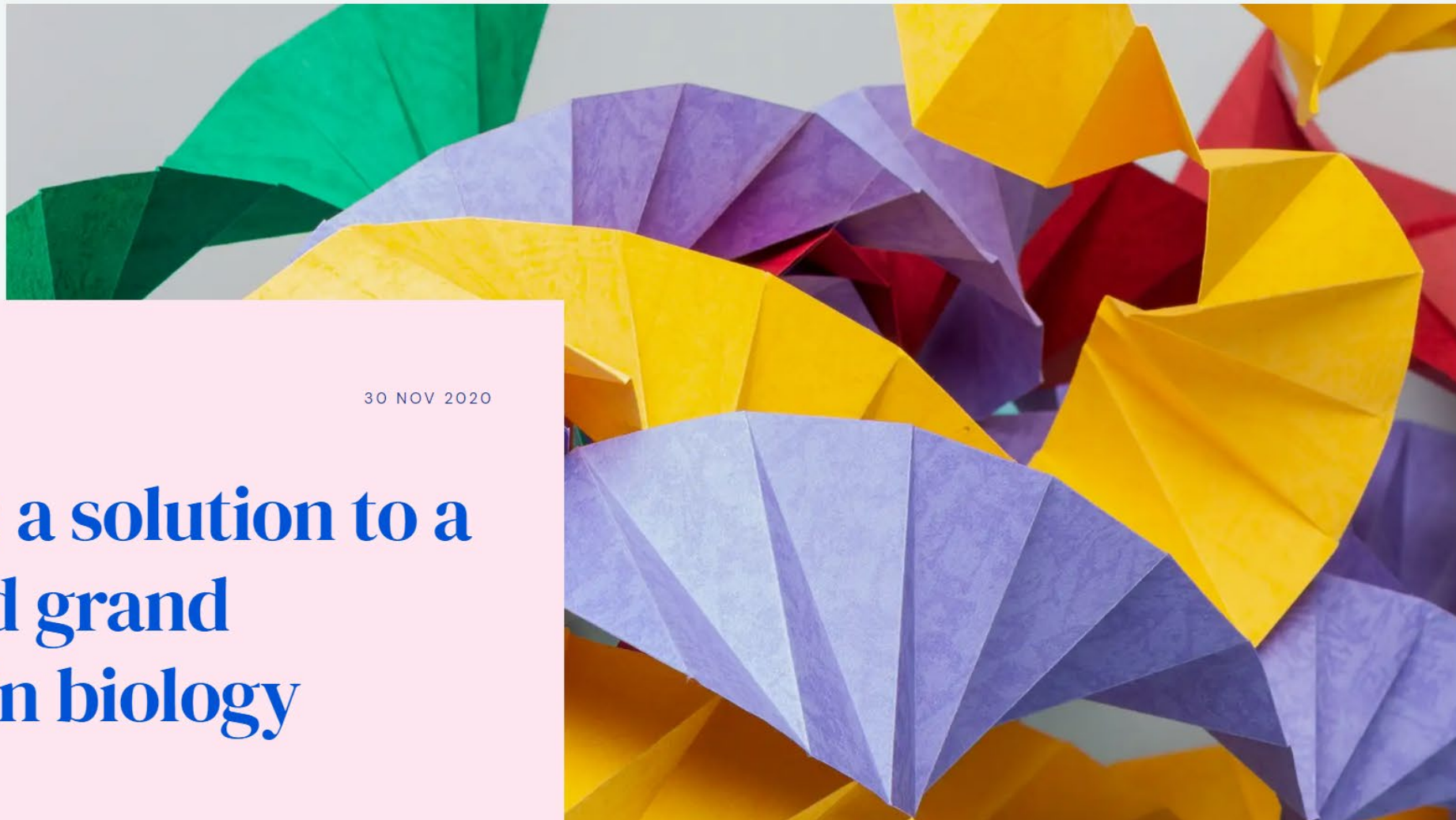
- Data Source: Scientific Texts (200M abstracts and full-text)
- Established end-to-end pipeline for building foundation models
- Ported and optimized LLM training frameworks to Frontier
- Released a suite of pre-trained foundation models (13 models in total) up to 26B parameters on over 200M scientific documents, w/ Hugging Face API

<https://github.com/at-aaims/forge>



- Demonstrated the usage of foundation models for scientific tasks.

# A Transformation of Scientific Research



BLOG POST  
RESEARCH

30 NOV 2020

# AlphaFold: a solution to a 50-year-old grand challenge in biology

**We have been stuck on this one problem – how do proteins fold up – for nearly 50 years. To see DeepMind produce a solution for this, having worked personally on this problem for so long and after so many stops and starts, wondering if we'd ever get there, is a very special moment.**

**PROFESSOR JOHN MOULT**  
CO-FOUNDER AND CHAIR OF CASP, UNIVERSITY OF MARYLAND

**AlphaFold is a once in a generation advance, predicting protein structures with incredible speed and precision. This leap forward demonstrates how computational methods are poised to transform research in biology and hold much promise for accelerating the drug discovery process.**

**ARTHUR D. LEVINSON**  
PHD, FOUNDER & CEO CALICO, FORMER CHAIRMAN & CEO, GENENTECH

**This computational work represents a stunning advance on the protein-folding problem, a 50-year-old grand challenge in biology. It has occurred decades before many people in the field would have predicted. It will be exciting to see the many ways in which it will fundamentally change biological research.**

**PROFESSOR VENKI RAMAKRISHNAN**  
NOBEL LAUREATE AND PRESIDENT OF THE ROYAL SOCIETY



DeepMind researchers report in *Nature* the creation of 350,000 predicted structures—more than twice as many as previously solved by experimental methods. The researchers say AlphaFold produced structures for nearly 44% of all human proteins, covering nearly 60% of all the amino acids encoded by the human genome.



‘This will be one of the most important data sets since the mapping of the human genome’

Ewan Birney, director of EMBL’s European Bioinformatics Institute.

‘It’s going to reset the field. It’s a very exciting time.’

David Baker, director of the University of Washington’s Institute for Protein Design.

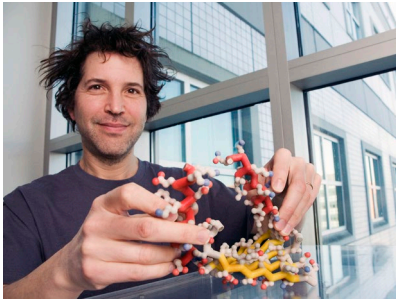


# The 2024 Nobel Prize for Chemistry

## David Baker

University of Washington, Seattle, WA, USA  
Howard Hughes Medical Institute, USA

*“for computational protein design”*



## Demis Hassabis

Google DeepMind, London, UK

*“for protein structure prediction”*



## John M. Jumper

Google DeepMind, London, UK

- The Nobel Prize in Chemistry 2024 is about proteins, life’s ingenious chemical tools.
  - David Baker has succeeded with the almost impossible feat of building entirely new kinds of proteins.
  - Demis Hassabis and John Jumper have developed an AI model to solve a 50-year-old problem: predicting proteins’ complex structures.
- These discoveries hold enormous potential.



Conclusions?

**“AI won’t replace the teacher, but teachers who use AI will replace those who don’t.”**

**From a Microsoft report “The Future Computed”**

**“AI won’t replace the scientist, but scientists who use AI will replace those who don’t.”**

**Adapted from a Microsoft report, “The Future Computed”**

**“AI won’t replace the librarian, but librarians who use AI will replace those who don’t.”**

**Adapted from a Microsoft report, “The Future Computed”**