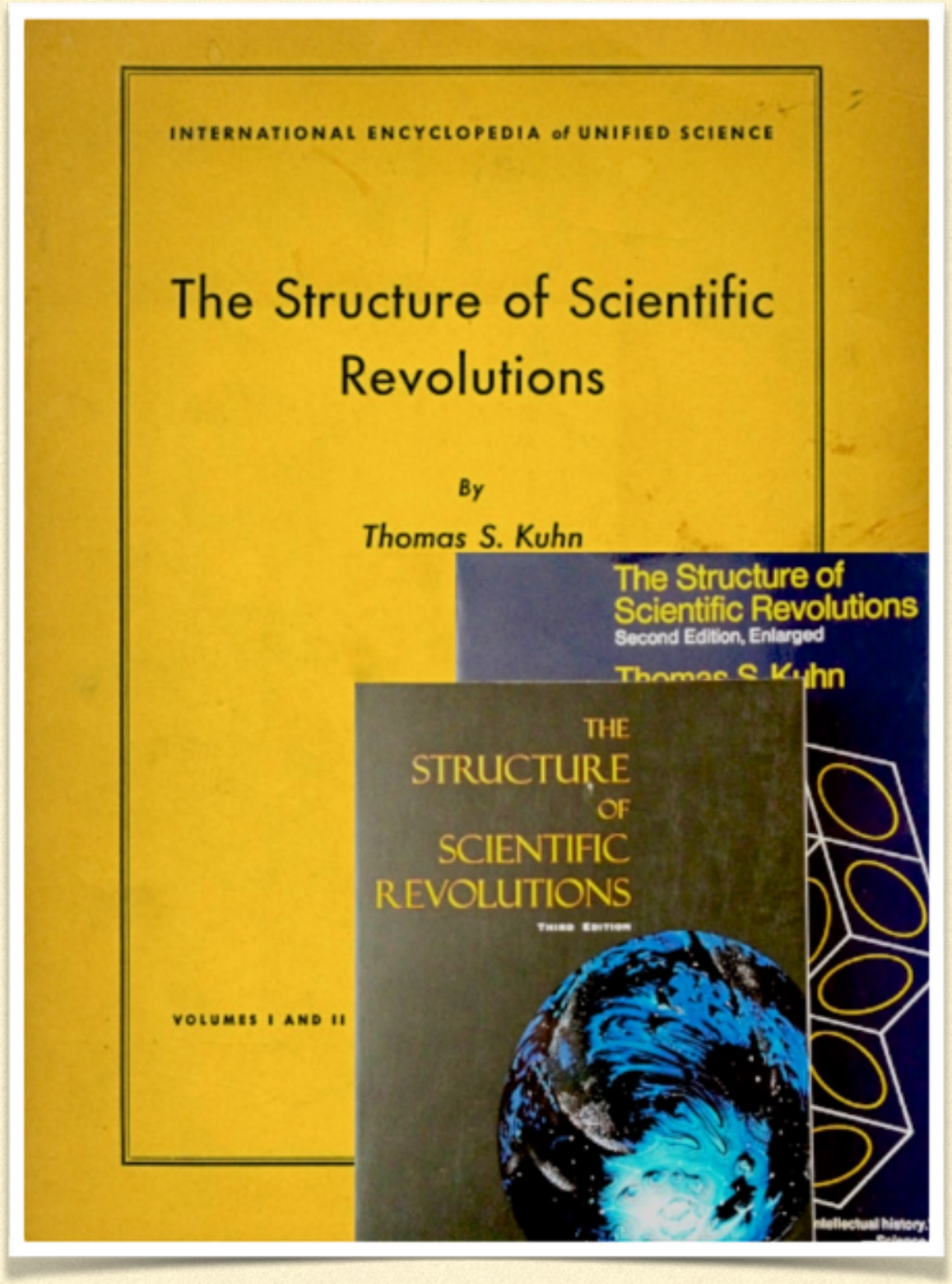
O 4º Paradigma da Pesquisa Científica

Como vamos enfrentar a tsunami de dados?

PARADIGMAS EM CIÊNCIA

Thomas Kuhn (1922–1996) apresenta em seu livro "A estrutura das revoluções científicas" (1962) a noção de *mudança de paradigma*:

a ciência não evolui gradualmente para a verdade, e sim passa periodicamente por períodos de crise e revoluções



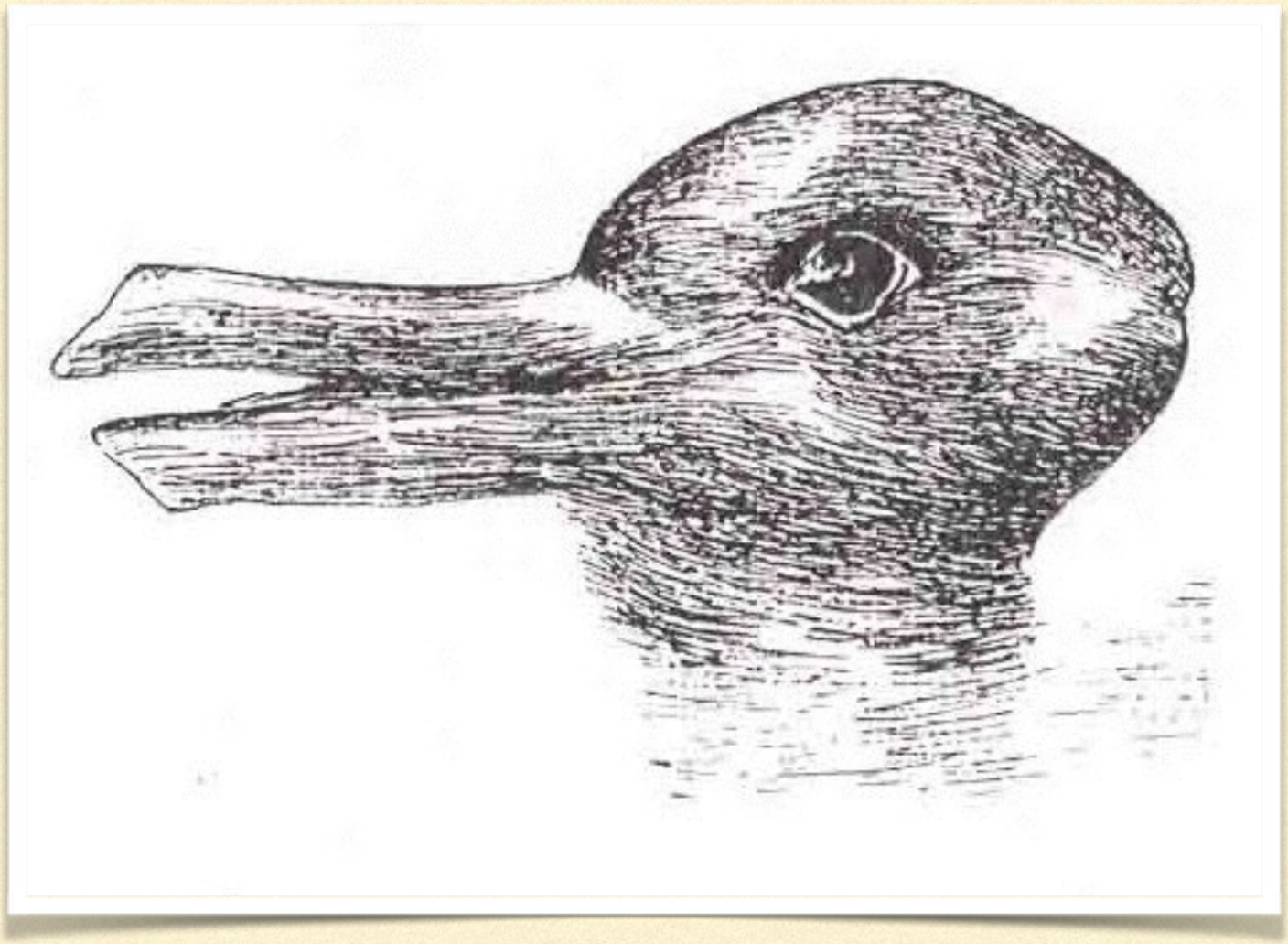
REVOLUÇÕES CIENTÍFICAS

- Há anomalias em todos os paradigmas que são interpretados como níveis aceitáveis de erro, ou simplesmente ignorados e não tratados.

- Quando um número suficiente de anomalias significativas se acumula contra um paradigma atual, a disciplina científica é lançada em um estado de crise. Durante a crise, novas idéias, talvez anteriormente descartadas, são experimentadas.

- Eventualmente um novo paradigma é formado, que ganha seus próprios novos seguidores, e uma "batalha" intelectual ocorre entre os seguidores do novo e do velho paradigma.

Mesma informação, visões diferentes



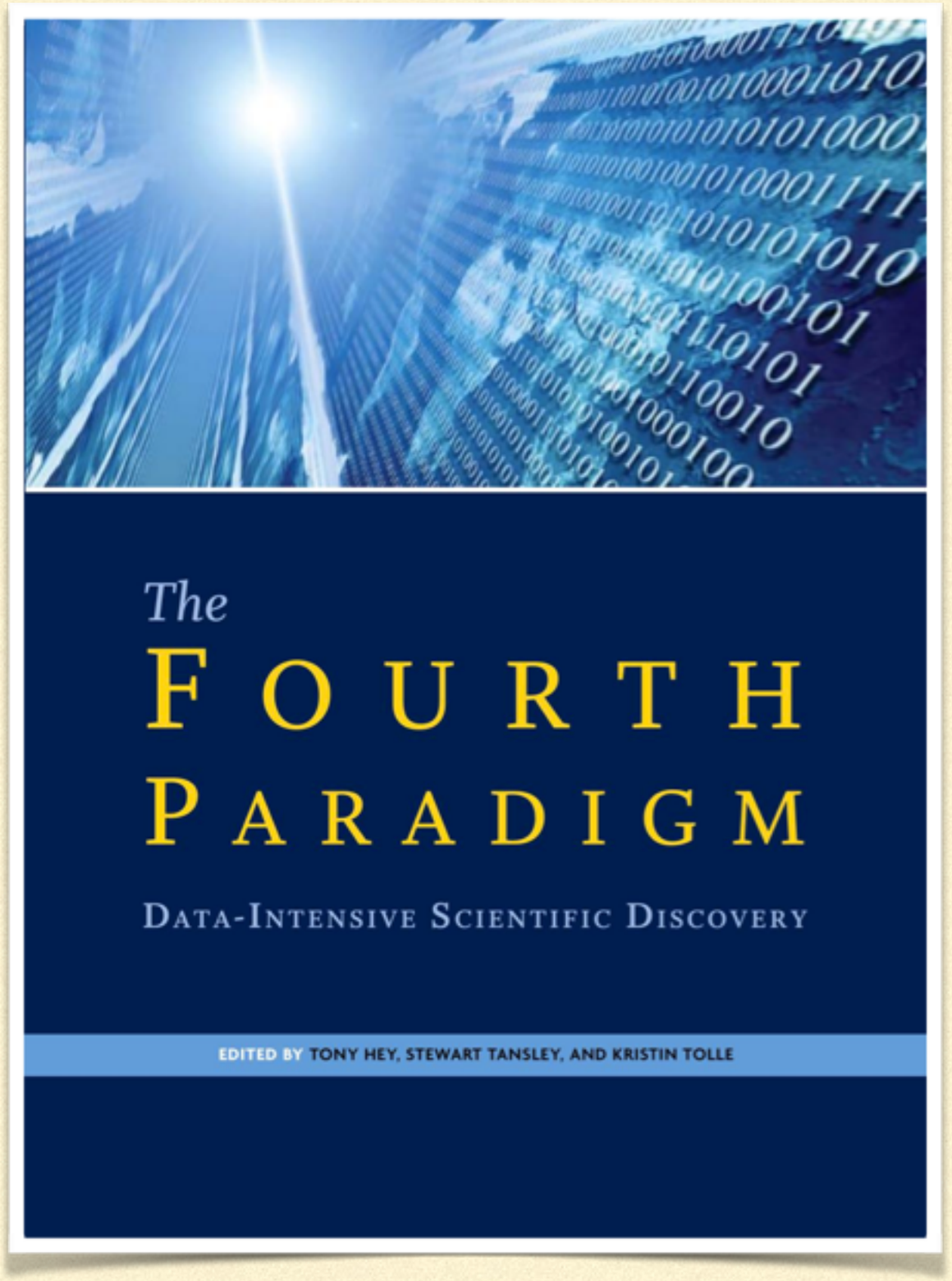
- Kuhn usou a ilusão óptica tornada famosa por Wittgenstein, para demonstrar a maneira pela qual uma mudança de paradigma poderia fazer com que se visse a mesma informação de uma maneira diferente.

Public Domain, https://commons.wikimedia.org/w/index.php?curid=667017

DATA-INTENSIVE SCIENTIFIC DISCOVERY

A maneira como fazemos a Ciência, nunca parou de evoluir ao longo de nossa história.

Na última década tem sido proposto que estamos vivenciando um quarto paradigma de como a pesquisa científica é feita em geral, independente da disciplina



O 4o paradigma é concebido como um novo método de avançar as fronteiras do conhecimento, através de novas tecnologias para coletar, manipular, analisar e exibir dados.

Jim Gray, cientista da computação no grupo Microsoft eScience, que foi perdido no mar em 2007.

"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 for the resulting information or knowledge to be 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."

Jim Gray 2007

Evolução dos paradigmas da pesquisa científica, tal como proposto por Jim Gray, 2007

Ciência Empírica

Ciência Teórica

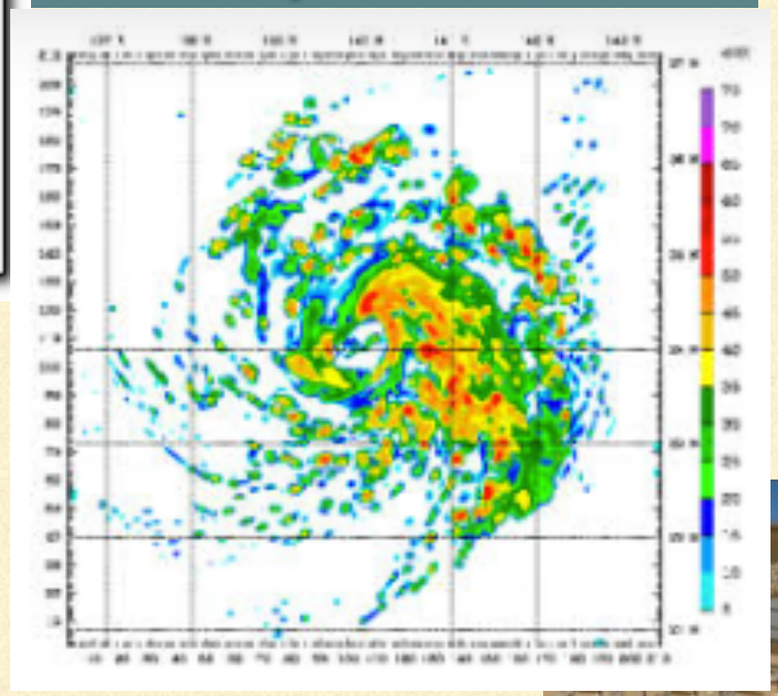
$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$$

$$\nabla \cdot \mathbf{B} = 0$$

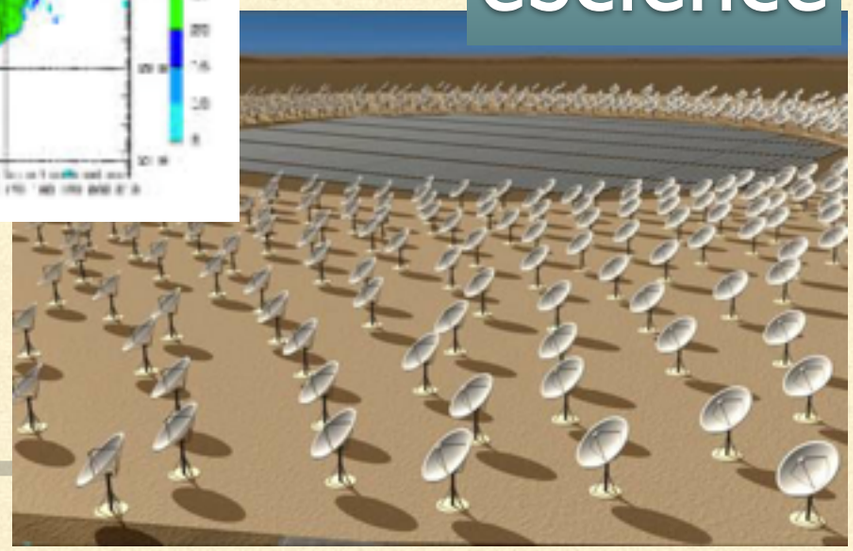
$$\nabla \mathrm{x} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial \mathrm{t}}$$

$$\nabla \mathrm{x} \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial \mathrm{t}}$$

Ciência Computacional



eScience



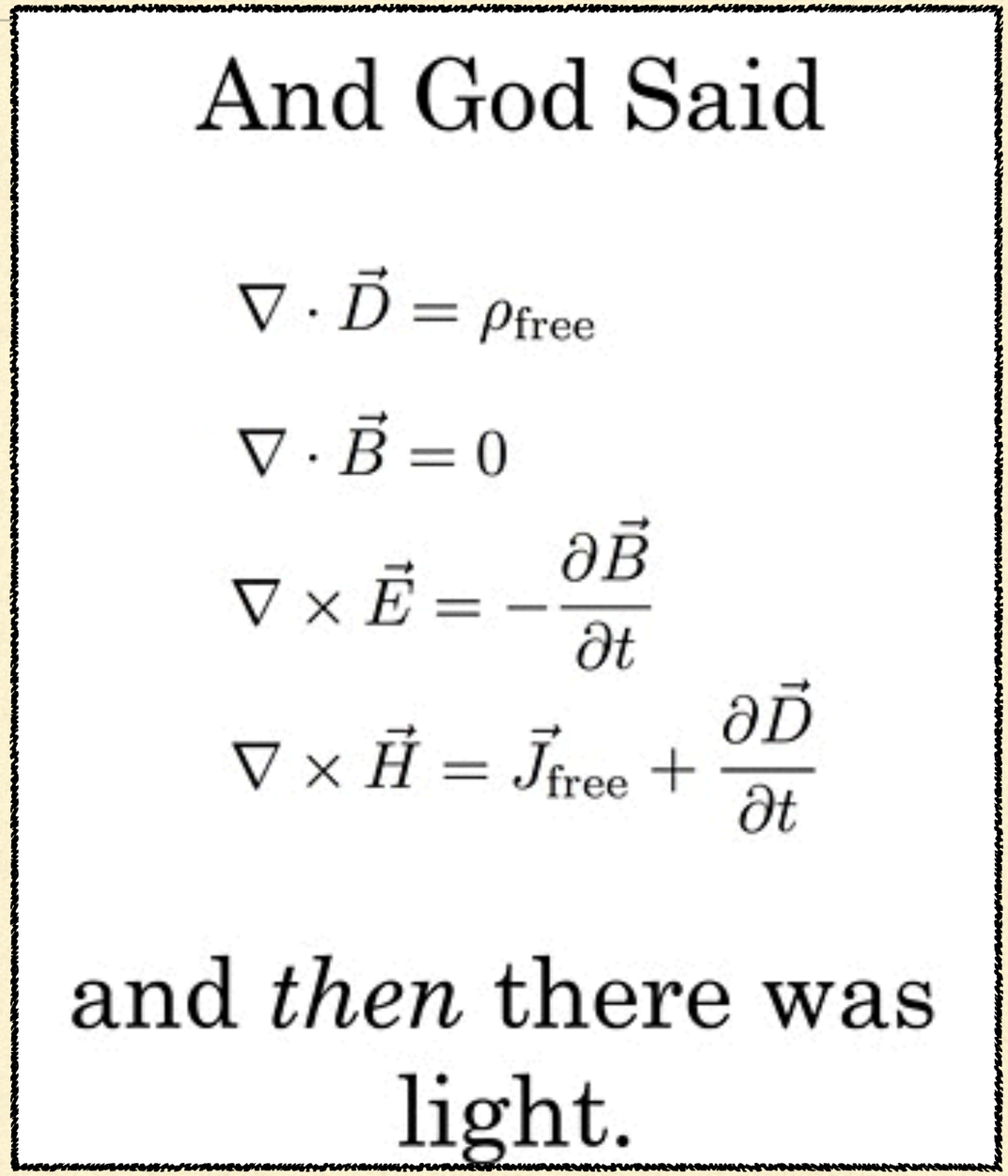
Ciência Empírica

Nicolas Camille Flammarion, Paris, 1888

Milhares de anos atrás:

a pesquisa científica era puramente empírica, baseada em observar e descrever os fenômenos naturais

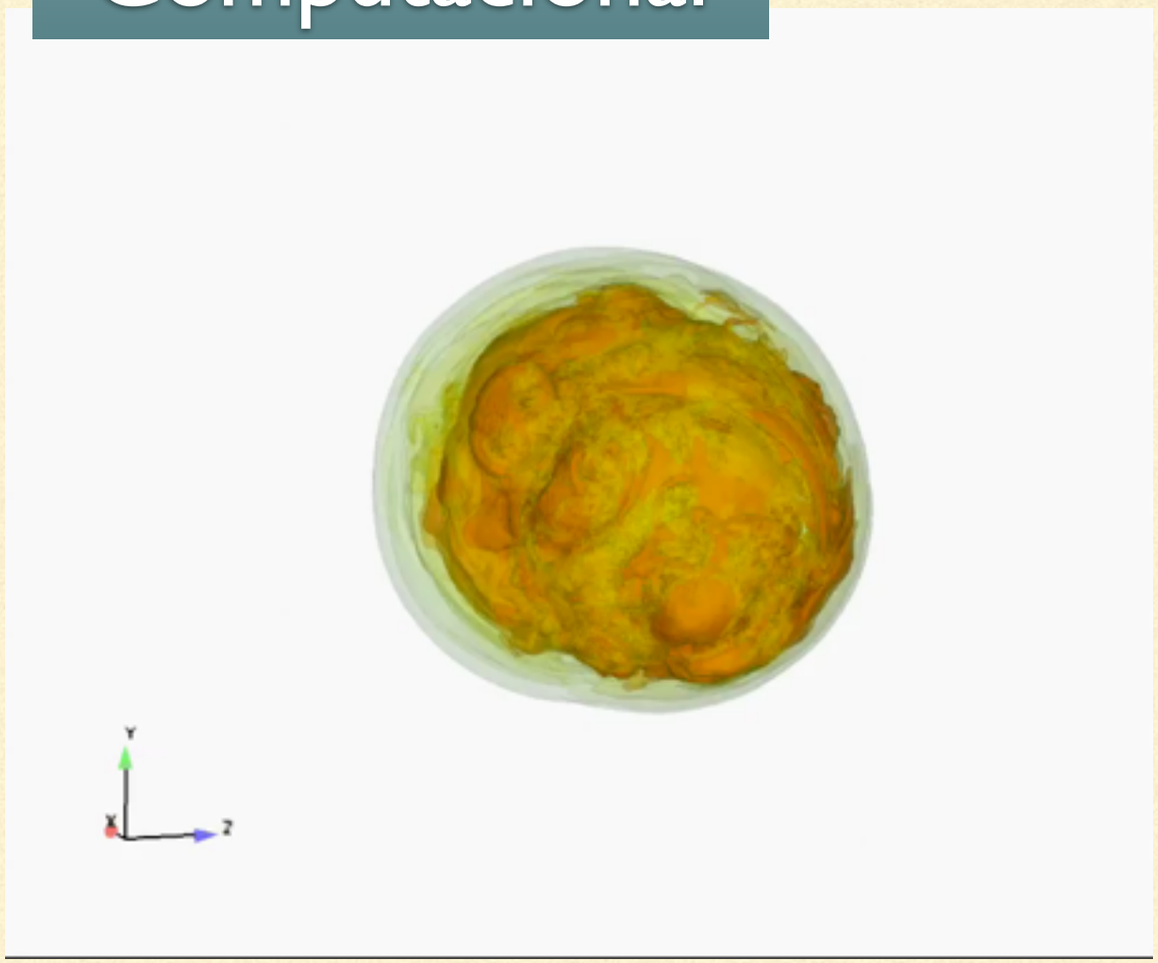
Ciência Teórica



Últimas centenas de anos:

Ciência torna-se teórica, com Leis de Kepler, Leis de Movimento de Newton, Equações de Maxwell… Usam-se modelos e generalizações.

Ciência Computacional



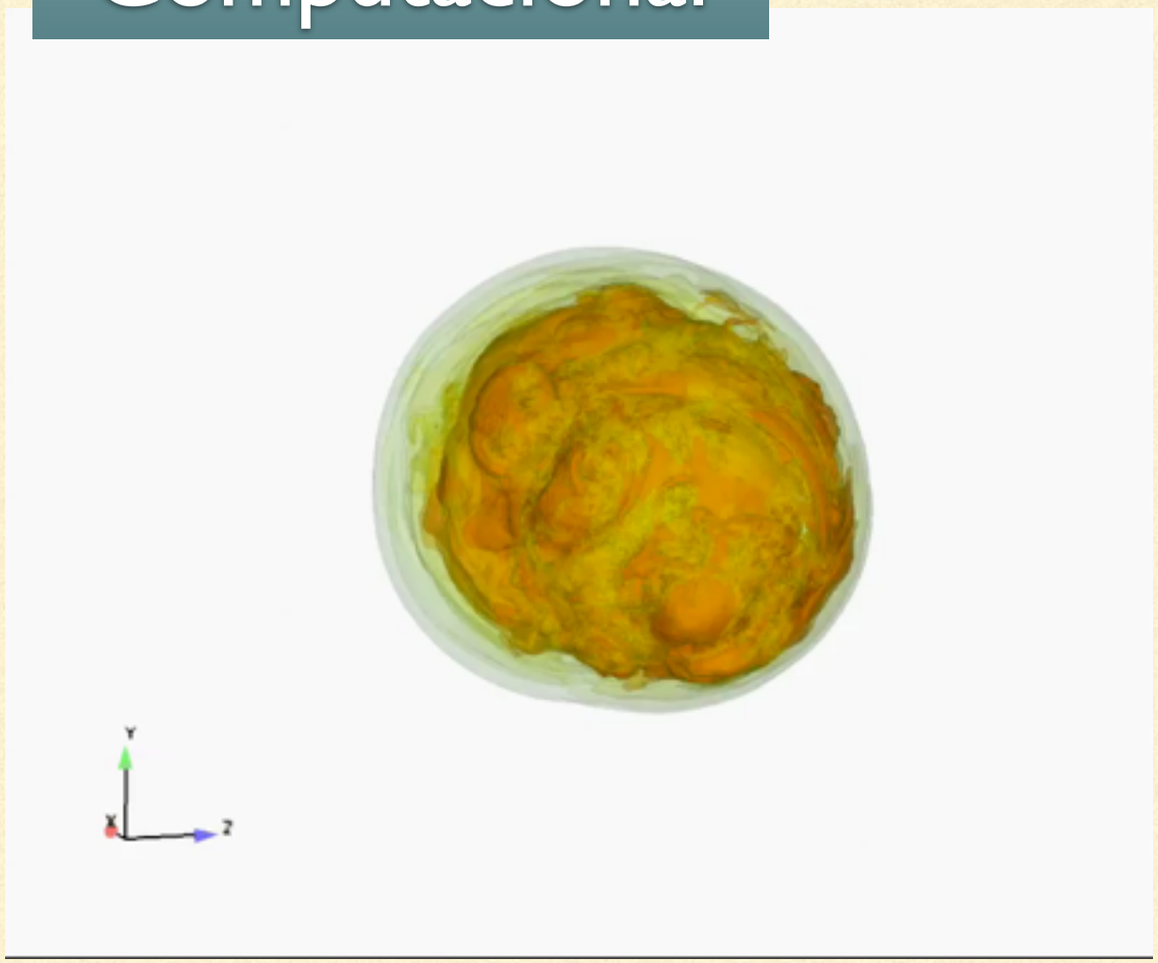
Simulação de "Spherical Accretion Shock Instability" em pulsares. Crédito: Blondin & Mezzacappa (2007).

Últimas décadas:

modelos teóricos se tornam muito complicados para serem resolvidos analiticamente, e os cientistas começaram a simular. Torna-se possível a simulação de fenômenos cada vez mais complexos.

Os resultados das simulações tornaram-se dados sintéticos, mal distinguíveis do que chamamos de "dados observacionais".

Ciência Computacional



Simulação de "Spherical Accretion Shock Instability" em pulsares. Crédito: Blondin & Mezzacappa (2007).

Últimas décadas:

modelos teóricos se tornam muito complicados para serem resolvidos analiticamente, e os cientistas começaram a simular. Torna-se possível a simulação de fenômenos cada vez mais complexos.

Os resultados das simulações tornaram-se dados sintéticos, mal distinguíveis do que chamamos de "dados observacionais".

eScience

https://skatelescope.org/

eScience

COMPUTERWORLD FROM IDG NEWS TECHNOLOGY TOOLS WHITEPAPERS

SKA telescope to generate more data than entire Internet in 2020

An exaflop-capable supercomputer, storage of at least 1.5 petabytes and data centres around the world will be required say scientists

Hamish Barwick (Computerworld)
07 July, 2011 12:05

https://skatelescope.org/

"This is a software and IT telescope in many senses because of the data challenges due to the amount it will generate, the amount of information that it passes and is going to process," Quinn said.

"This telescope will generate the same amount of data in a day as the entire planet does in a year. We estimate that there will be more data flowing inside the telescope network than the entire internet in 2020."

eScience

Hoje:

Ciência é centrada nos dados, sejam observados ou simulados, unificando teoria, experimentos e simulações = eScience.

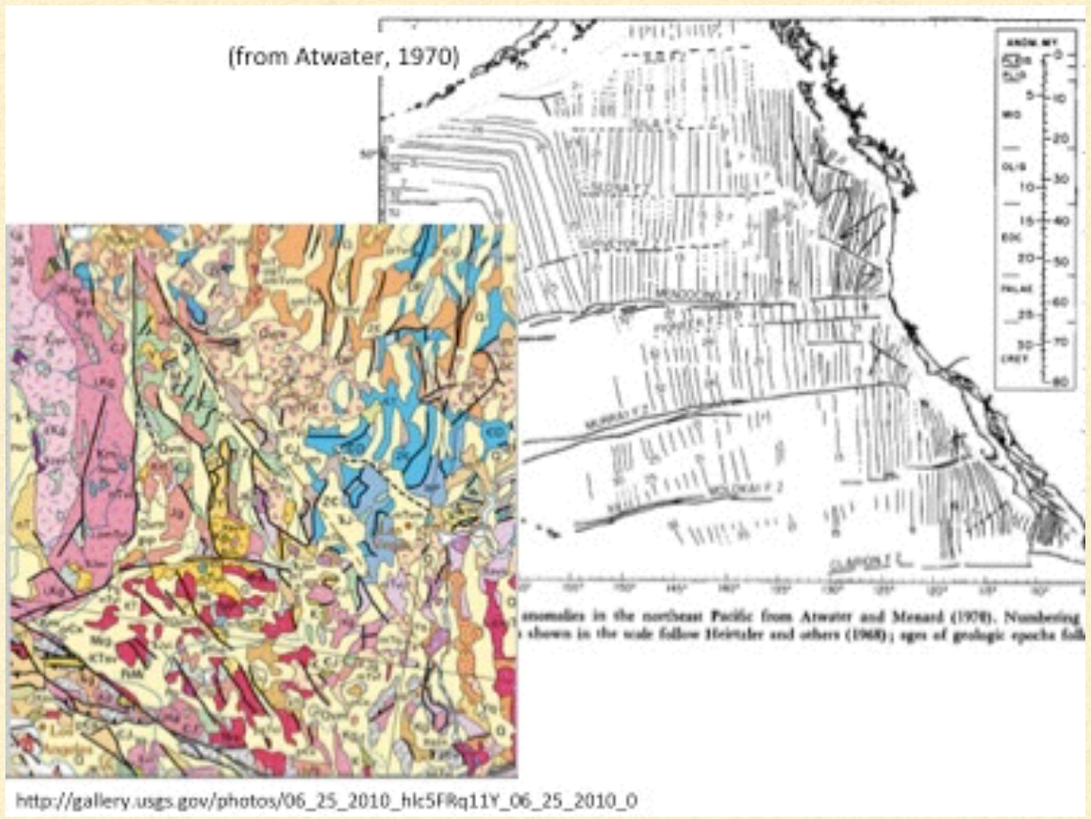
Exploração de dados, dados são capturados por instrumentos ou gerados por um simulador, processado por software, informação é armazenada em computadores, cientistas analisam bancos de dados usando gerenciamento de dados e estatística.

ASTRONOMIA E O QUARTO PARADIGMA

- A astronomia é rica em dados e o volume de dados está crescendo até um ponto em que nem sequer conseguimos armazenar ou transferir dados brutos, como é o caso do atual satélite ESA Gaia ou do futuro radiotelescópio SKA.

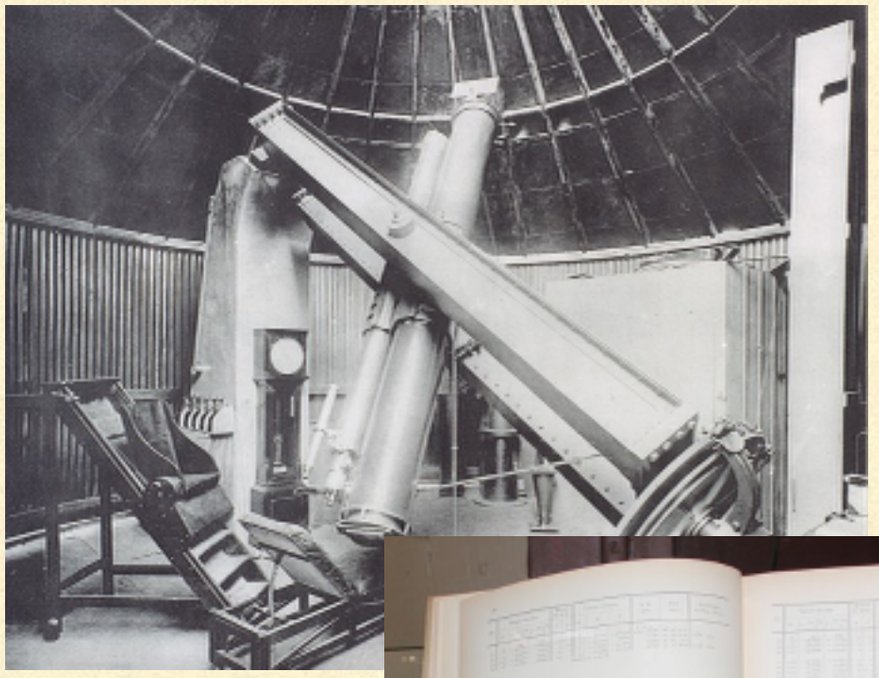
- Cada vez com mais freqüência, fazemos ciência com base na análise de imensos conjuntos de dados, com experimentos científicos, simulações e observações astronômicas atingindo PBs.

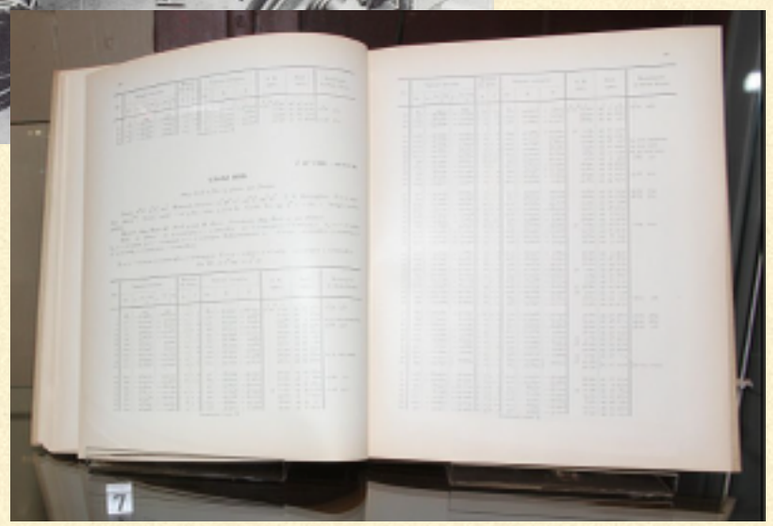
SEMENTES…



Estudo de placas tectônicas através da análise de anomalias magnéticas (Pitman & Heirtzler 1966, Atwater 1960)

Carte du Ciel e *catálogo Astrográfico* séc XIX (22 observatórios, precursor da IAU)

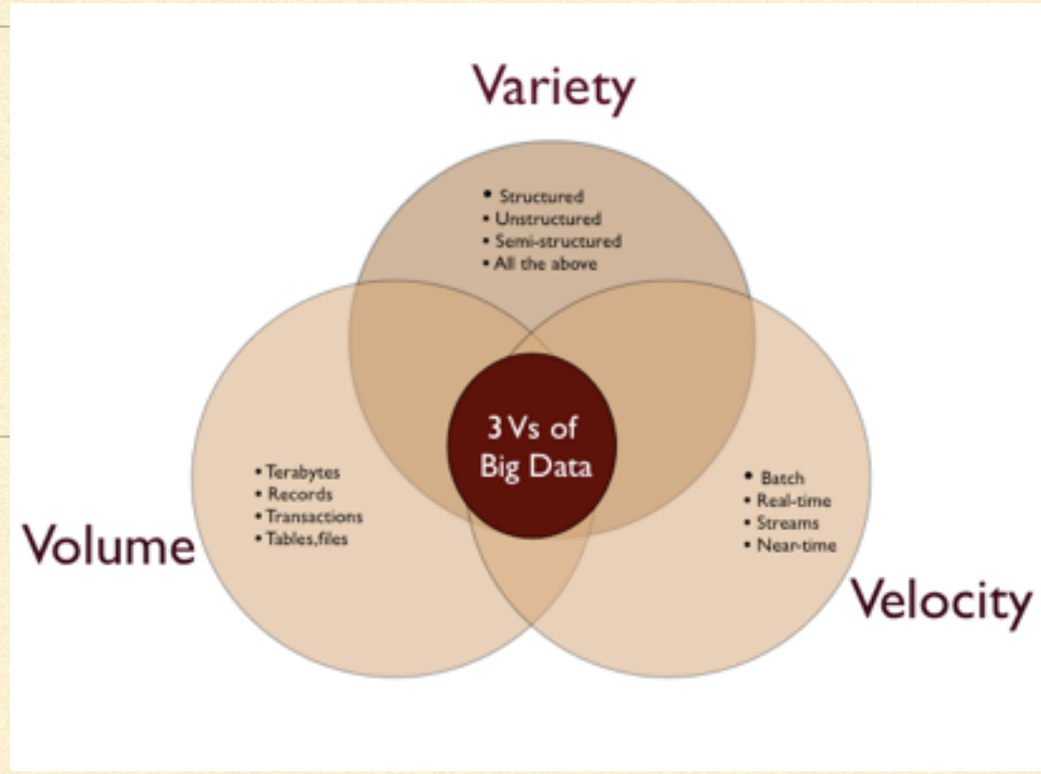


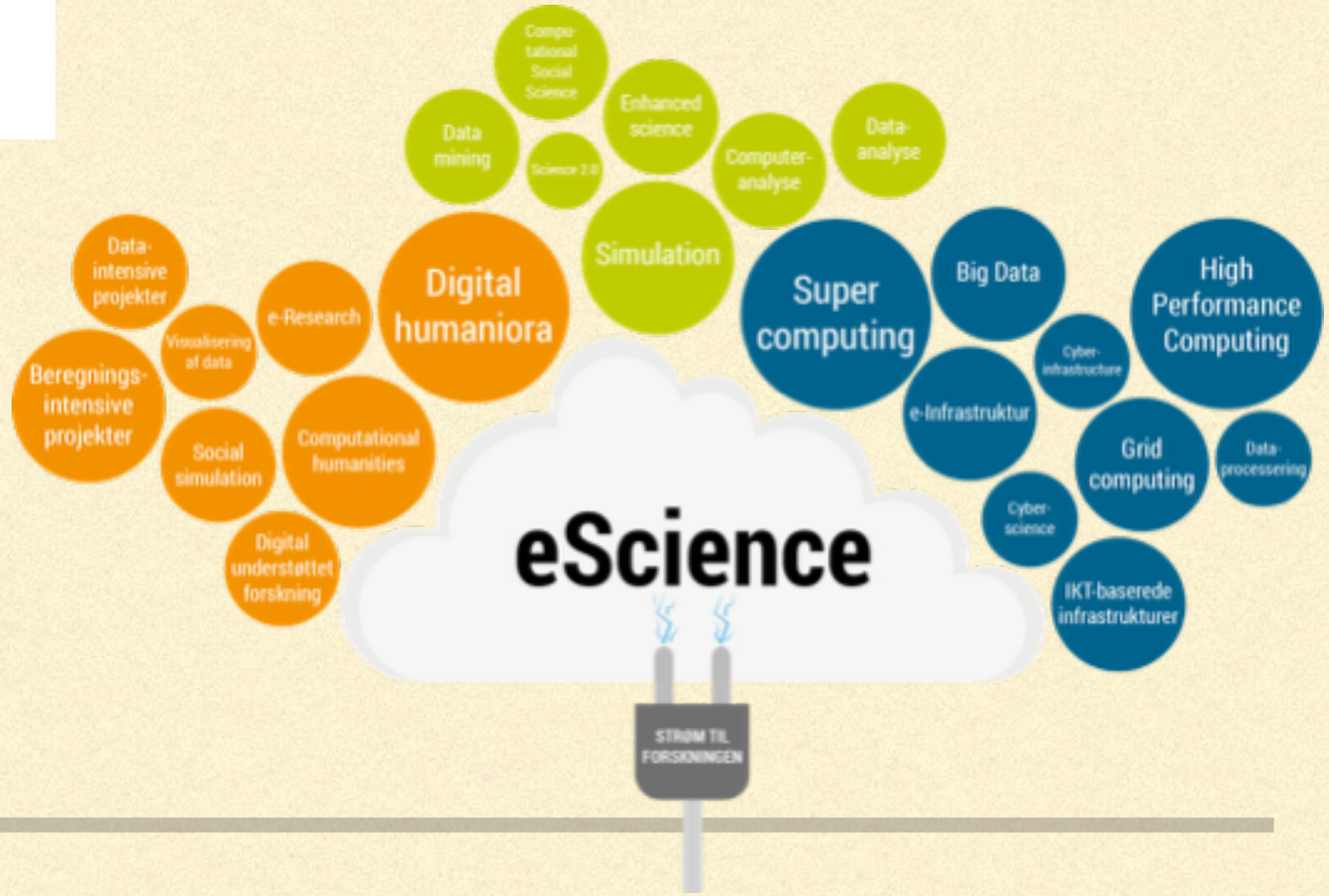


ALEXANDER SZALAY
SCIENCE IN AN EXPONENTIAL WORLD

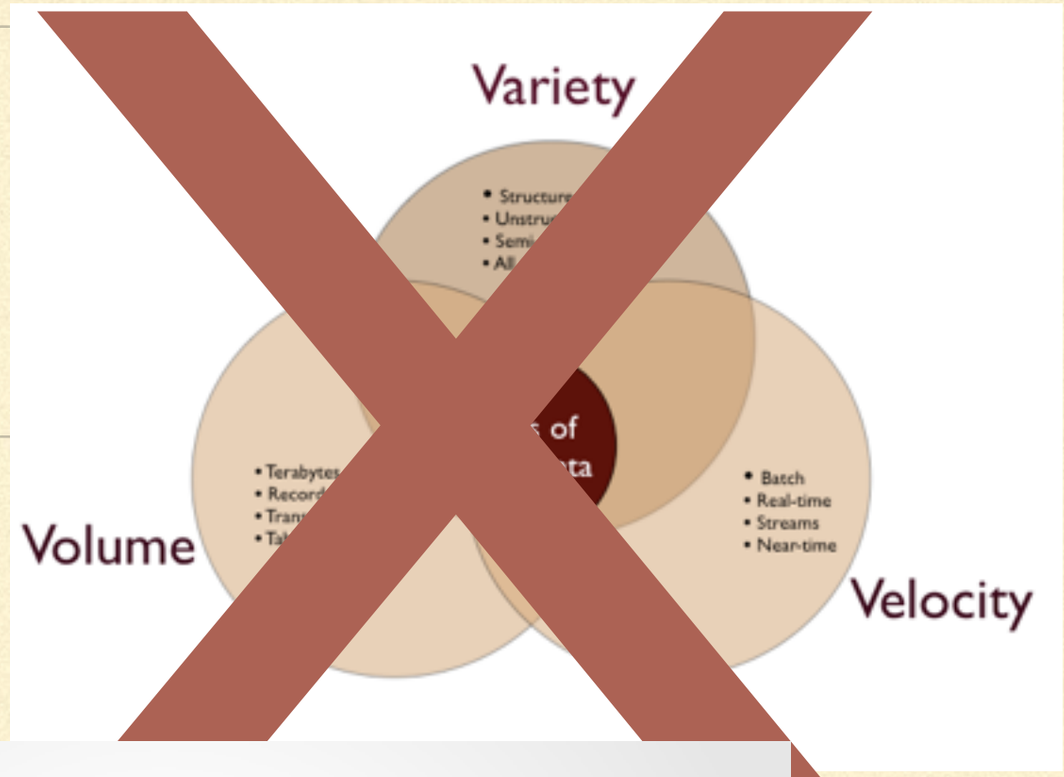
https://www.youtube.com/watch?v=hB92o4H46hc

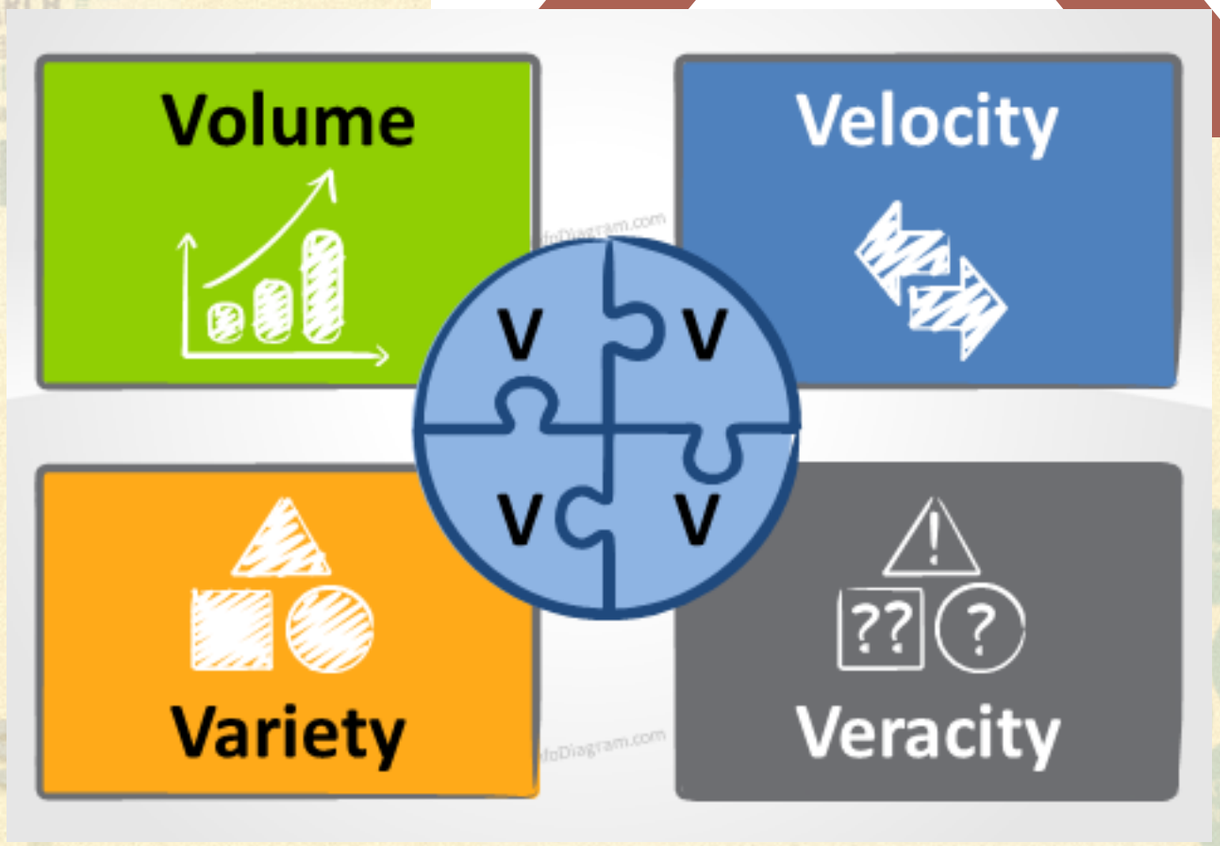
NOVOS TERMOS E CONCEITOS

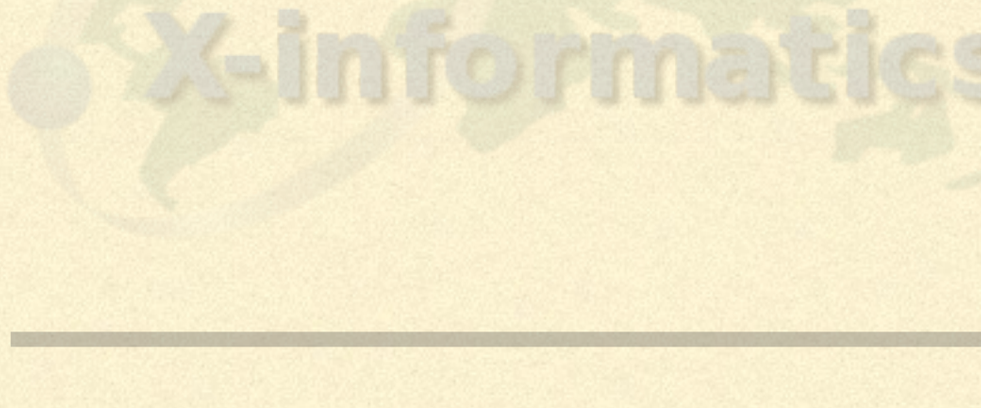




NOVOS TERMOS E CONCEITOS







NOVOS TERMOS E CONCEITOS

Conjuntos de dados que são tão grandes ou complexos que os softwares de aplicativos de processamento de dados tradicionais são inadequados para lidar com eles.

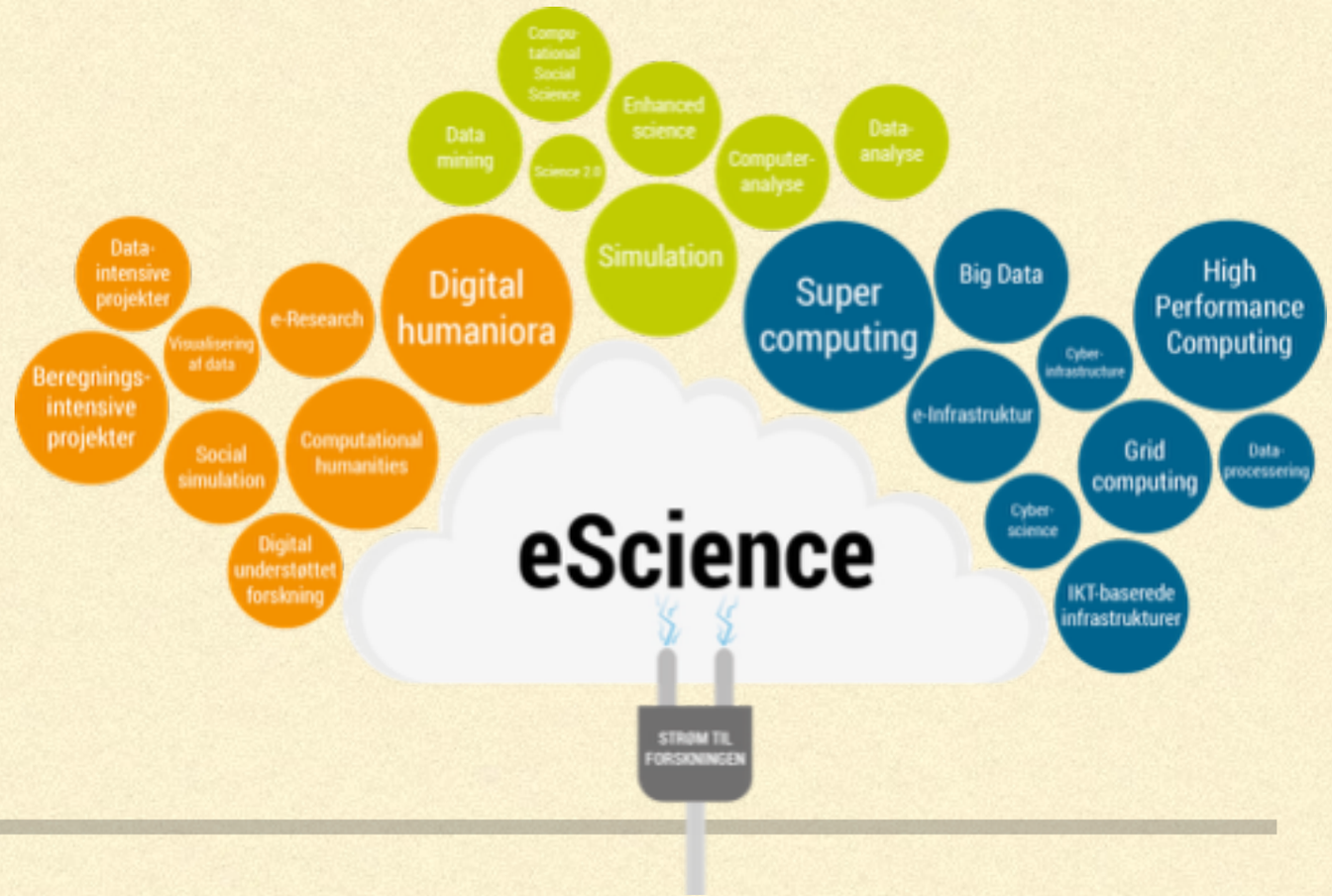
Os desafios incluem captura, armazenamento, análise, coleta de dados, pesquisa, compartilhamento, transferência, visualização, consulta, atualização e privacidade de informações.

NOVOS TERMOS E CONCEITOS

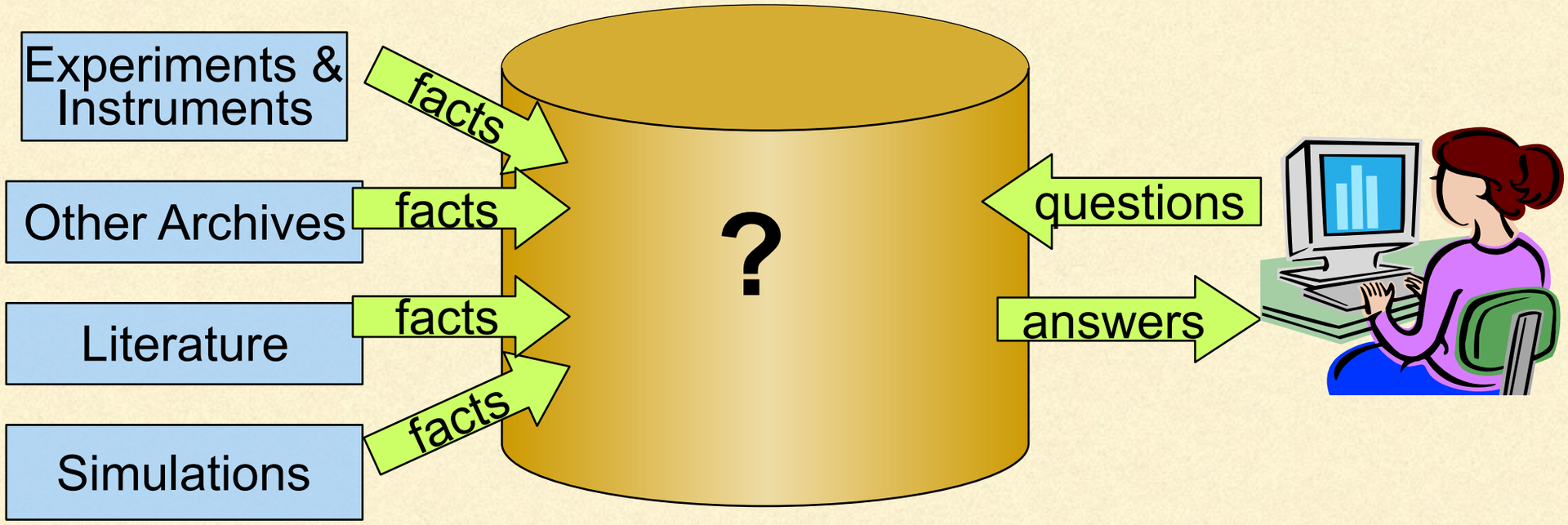
Síntese de tecnologia da informação e pesquisa científica.

eScience is where "IT meets scientists."

Cientistas da computação e estatísticos passam a ser indispensáveis para que se obtenha conhecimento dos dados que as diversas disciplinas tem acesso hoje em dia.



NOVOS TERMOS E CONCEITOS

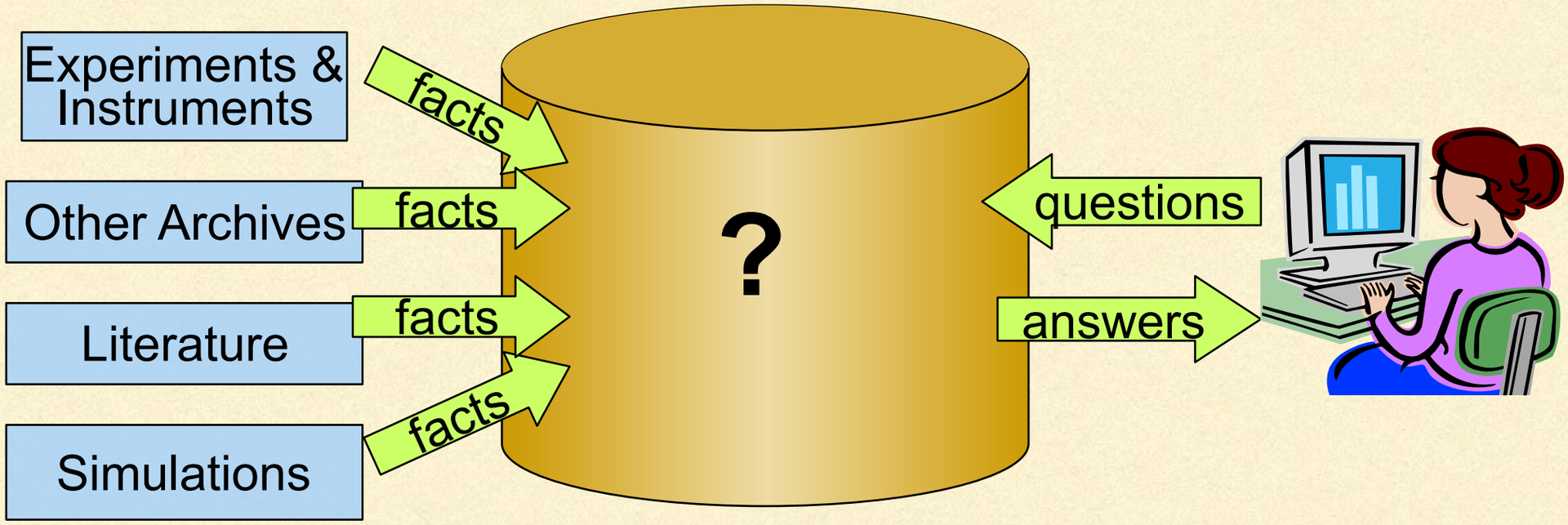


X-informatics

Evolução de X-Info e Comp-X para cada disciplina X

Como codificar e interpretar conhecimento.

E NOVOS PROBLEMAS PARA AS TECNOLOGIAS DE INFORMAÇÃO



- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to reorganize it
- How to share with others
- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

OU SEJA...

- Nesse paradigma emergente, desenvolvem-se **caminhos** em que **computadores, bancos de dados e redes** não são vistos e utilizados apenas como ferramentas, mas se tornam uma **parte fundamental do processo de descoberta de conhecimento**, tornam-se **fundamentais para a nossa interpretação dos dados**.

- Assim, este novo paradigma também **avança métodos e algoritmos** para analisar os dados armazenados nessas bases de dados de grande escala (ou entre várias bases de dados em paralelo), e nesse processo, também é necessário estabelecer **protocolos de comunicação padronizados** entre todas essas fontes de dados.

4 PILARES DA *DATA INTENSIVE SCIENTIFIC DISCOVERY*

Bancos de dados e Gerenciamento do Ciclo de Vida dos Dados

Desde a criação e armazenamento inicial até o momento em que é arquivado para a posteridade ou torna-se obsoleto e é excluído. O objetivo é garantir que os dados sejam recuperados de forma confiável para fins de pesquisa futura ou reutilização.

Workflow científicos

(Taverna, MyExperiment, Microsoft Azure) usado por vários cientistas para garantir a reprodutibilidade dos dados de modo consistente e competente.

Difusão e Troca de informações

(arXiv, Pubmed, VO) Onde a maior parte se não o todo dos resultados científicos são publicamente acessíveis.

Ferramentas avançadas de análise

(R, SciPy) com contínuos melhoramentos em ferramentas de análise e visualização de dados.

S. GEORGE DJORGOVSKI
EVOLVING SCIENCE AND TECHNOLOGY IN CYBERSPACE

https://youtu.be/FB33pV2L0Vo

I wanted to point out that **almost everything about science is changing because of the impact of information technology**.

Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, **"data-intensive" science paradigm is emerging**.

The **goal** is to have a world in which **all of the science literature is online, all of the science data is online, and they interoperate with each other**.

Lots of **new tools** are needed to make this happen.

Jim Gray 2007