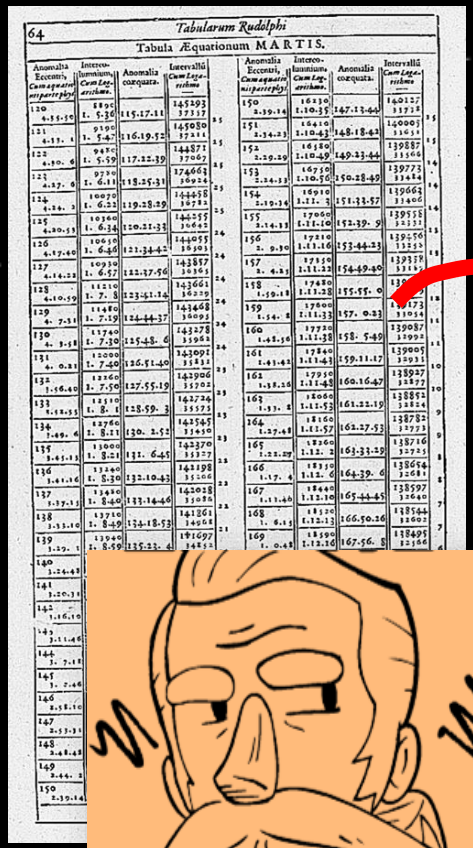
A emergência dos dados de pesquisa na ciência contemporânea

ou

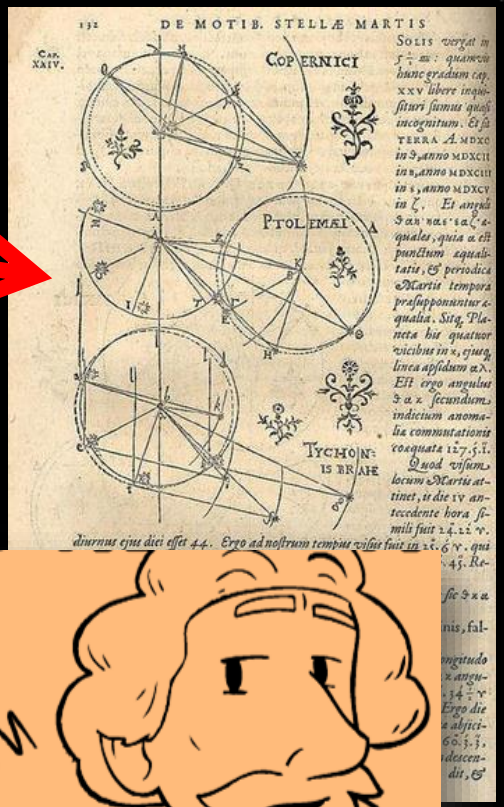
O fim da teoria

DADOS EXPERIMENTAIS

TEORIA


64 Tabularum Rudolphi

Tabula Æquationum MARTIS.

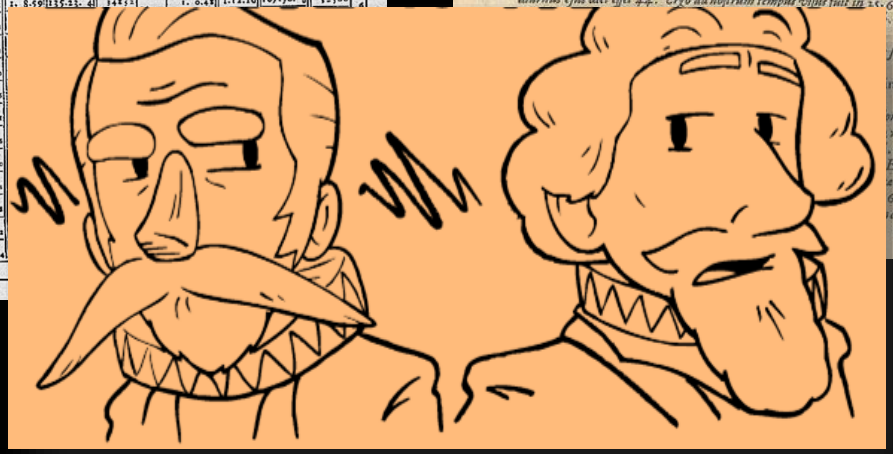

132 DE MOTIB. STELLÆ MARTIS

CAP. XXIV.

COPERNICI

PTOLEMÆI

TYCHONIS BRAHE



TICHO BRAHE

JOHANNES KEPLER

Dados de pesquisa sempre foram valorizados na ciência

KEPLER que era assistente de TICHO BRAHE pegou o catálogo de observações astronômicas sistemáticas do TICHO e descobriu as leis do movimento planetário.

Este fato estabeleceu a divisão entre a mineração e análise de dados experimentais, cuidadosamente arquivados, e a criação de teorias

GALILEU GALILEI

Os registros claros e cuidadosos de suas observações e seu estilo de publicação não somente permitiu que **ele compreendesse** o Sistema Solar **como permitiu também que seus leitores compreendessem como ele chegou as suas descobertas**. Isto por que o **caderno de notas de Galileu integravam seus dados (desenhos de Júpiter e suas luas), metadados chaves (cronometragem de cada observação, condições meteorológicas, propriedades do telescópio) e texto (descrição dos métodos, análises e conclusões).** Quando Galileu inclui as informações de suas notas no **Siderius Nuncius,** a integração entre texto, dado e metadado foi preservada.

De forma diferente de como Galileu reportou em Siderius Nuncius o resultado de suas pesquisas, a quantidade de dados reais e de descrição de dados nas publicações modernas quase nunca são suficientes para repetir ou mesmo estatisticamente verificar o estudo que está sendo apresentado (Goodman, 2014; Sayão e Sales. 2018)

O ACESSO ABERTO A DADOS DE PESQUISA TEM RAIZES ANTIGAS (Borgman, 2017)

O **World Data Center** - foi estabelecido na **década de 1950 para arquivar e distribuir dados coletados dos programas observacionais do Ano Geofísico Internacional** de 1957-1958 (Korsmo 2010; Shapley and Hart 1982).

CODATA foi fundado em 1966 pelo International Council for Science para **promover a cooperação em gestão e uso de dados** (Lide and Wood 2012).

MUSEU DE HISTÓRIA NATURAL

Antes das práticas acadêmicas se deslocarem para o reino digital ou para o paradigma do *big data*, os museus de história natural já tinham ampliado o seu conceito de curadoria antecipando a demanda por gestão e aprimoramento dos dados digitais (PALMER et al., 2013, p. 2).


CENÁRIOS
HIPERINFORMAÇÃO
DILÚVIO DE DADOS NA CIÊNCIA

PROTAGONISMO DOS DADOS NA CIÊNCIA CONTEMPORÂNEA

eScience

Ciência aberta

Cauda longa

Humanidades Digitais

BIG DATA CIENTÍFICO
Grandes projetos
Observatórios
Instalações complexas
Dados distribuídos
Simulação por computador

DADOS ABERTOS
Metodologias
Equipamentos
Software
Cadernos de laboratório
Roteiro de entrevistas
Resultados negativos

DADOS DOS DO GRANDE NÚMERO DE PEQUENOS LABORATÓRIOS
Heterogêneos
Não tratados
Invisíveis
Coletivamente é o maior volume

TECNOLOGIA COMPUTACIONAL APLICADAS A ESTUDOS EM HUMANIDADES.
Humanidades estudando Tecnologias digitais (Bobley)

OS PARADIGMAS CIENTÍFICOS

1º PARADIGMA:

Ciência experimental ou empírica estuda a relação entre fenômenos por meio de experimentos

2º PARADIGMA:

Ciência teórica ou **descritiva** formula modelos para descrição e explicação dos fenômenos naturais

3º PARADIGMA:

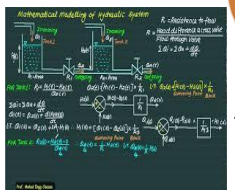
Ciência baseada em **simulação** uso de *software* e grande **geração de dados**

SIMULAÇÃO POR COMPUTADOR

É análogo a um experimento físico, mas usa equações matemáticas para representar o mundo real

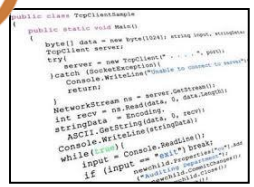
REALIDADE

ABSTRAÇÃO



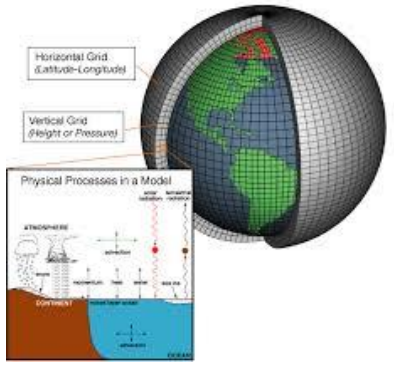
MODELO MATEMÁTICO

ALGORÍTMO NUMÉRICO



SOFTWARE

Muitos dados!!!

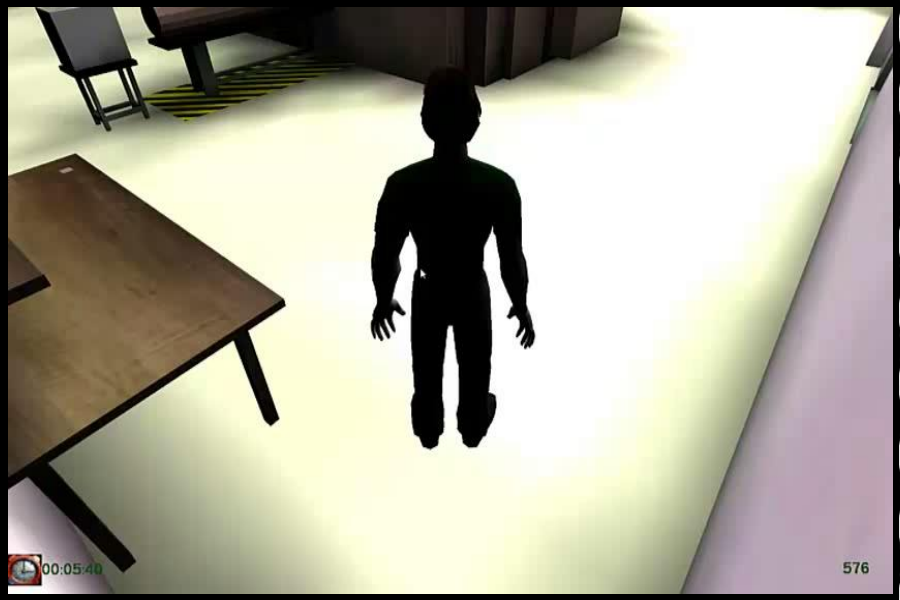


PREDIÇÕES CLIMÁTICAS

No eScience a simulação deixa de ser uma ferramenta que auxilia o pesquisador a fazer ciência para transformar o modo de fazer ciência e definir um novo PARADIMA CIENTÍFICO.

EXEMPLO DE RESULTADO DE PESQUISA NA ÁREA NUCLEAR (CNEN/IEN)

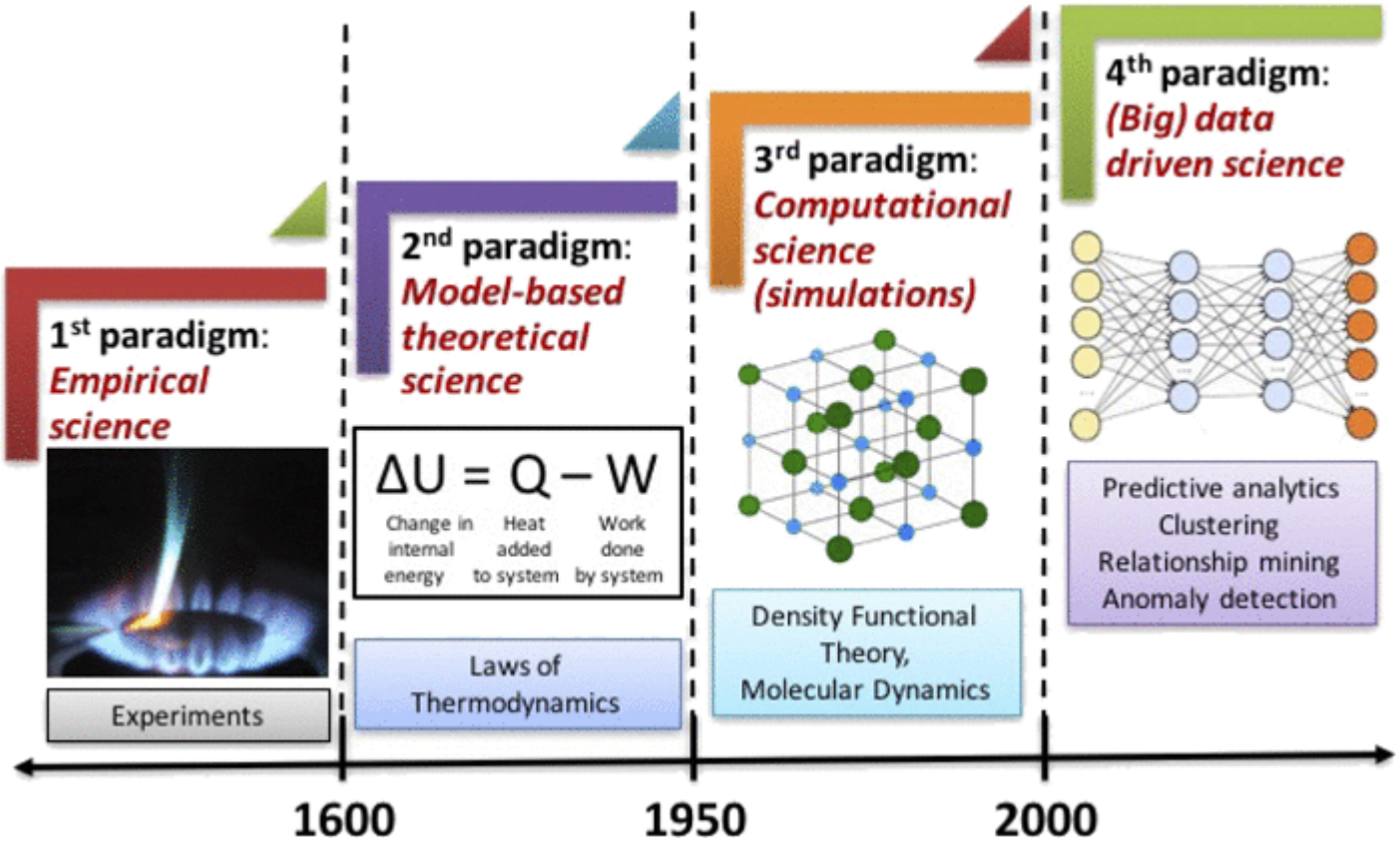
VIRTUAL SIMULATION OF A NUCLEAR POWER PLANT'S CONTROL ROOM AS A TOOL FOR ERGONOMIC EVALUATION.



Leandro Barbosa S. Gatto [a], Antônio Carlos A. Mól [a,b,c],
Isaac J.A. Luquetti dos Santos [a],
Carlos Alexandre F. Jorge [a,*], Ana Paula Legey [c]

INSTITUTO DE ENGENHARIA NUCLEAR - CNEN/IEN

MATERIALS INFORMATICS AND BIG DATA: REALIZATION OF THE "FOURTH PARADIGM" OF SCIENCE IN MATERIALS SCIENCE



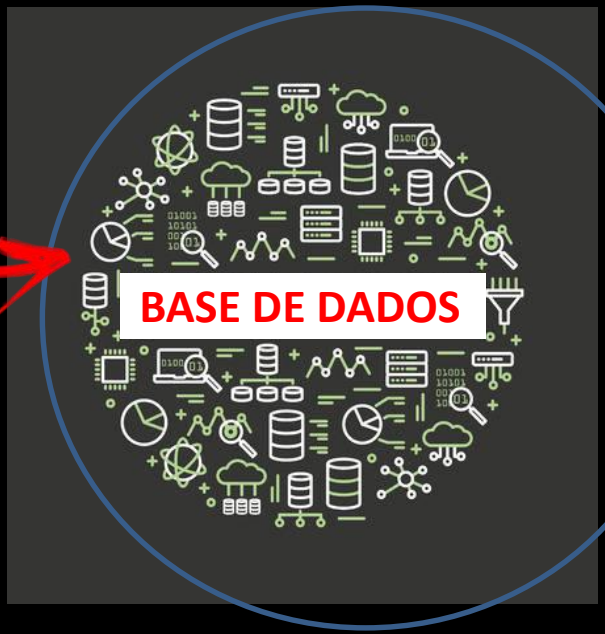
Materials informatics is a field of study that applies the principles of informatics to materials science and engineering to better understand the use, selection, development, and discovery of materials.

BIG DATA


COMPUTAÇÃO
analisar
integrar
visualizar
NOVAS
DESCOBERTAS
DADOS ESTRUTURADOS
E NÃO ESTRUTURADOS

FONTES DE DADOS

Combinação de **múltiplas fontes de dados provenientes de domínios diferentes,**



COMPUTAÇÃO

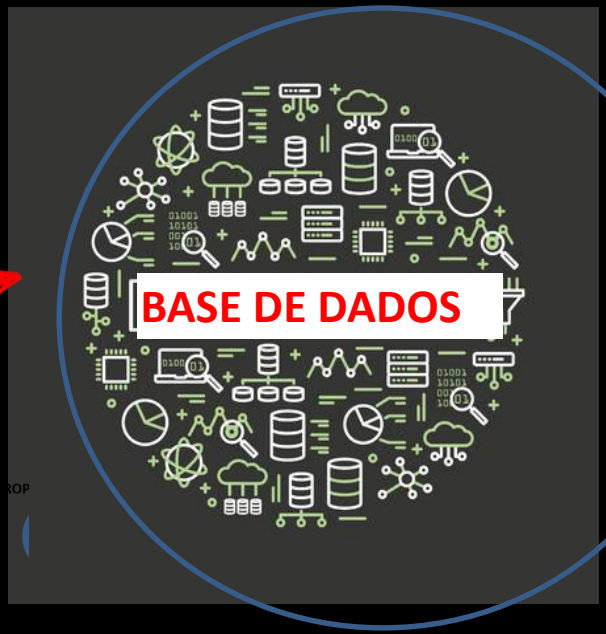
PADRÕES
RELAÇÕES
HIPÓTESES
TEORIAS

CAPTURA GESTÃO/CURADORIA ANÁLISE

NOVOSS DESCOBERTAS

Combinação de múltiplas fontes de dados provenientes de domínios diferentes,

FONTES DE DADOS



COMPUTAÇÃO

Análises exploratórias
Exploração de coleções de dados
Mineração de dados
Modelagem
Simulação interativa
Realidade virtual
Workflow científico

PADRÕES
RELAÇÕE
HIPÓTESE
TEORIAS

CAPTURA

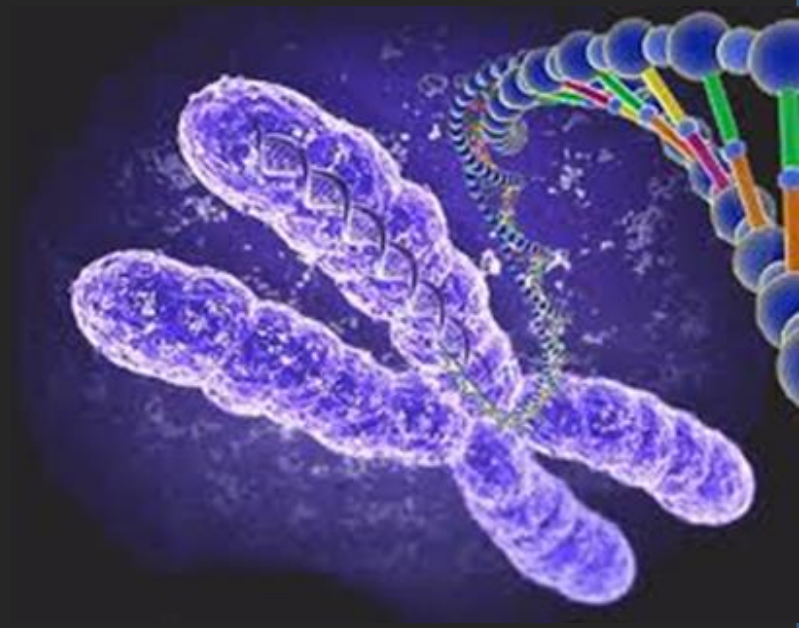
GESTÃO/CURADORIA

ANÁLISE

NOVOSS DESCOBERTAS

O DILÚVIO DE DADOS

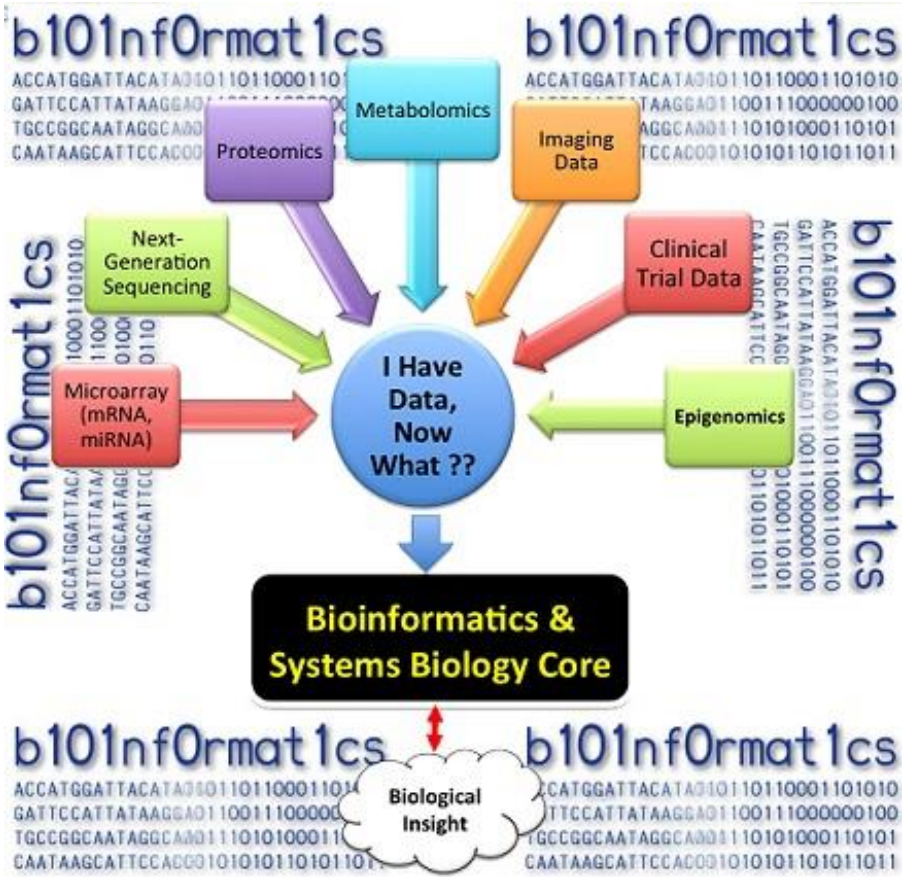
Uma nova geração de instrumentos científicos, sensores, satélites, software de simulação, colaboratórios produzem em ritmo exponencial quantidades imensas e diversificadas de dados brutos de pesquisa



Subprodutos dos processos de pesquisa

Existem hoje disciplinas científicas totalmente orientadas por dados, por exemplo:
BIOINFORMÁTICA e ASTROINFORMÁTICA

BIOINFORMÁTICA



BIOINFORMÁTICA é um campo interdisciplinar que desenvolve métodos e **ferramentas de software** para compreender dados biológicos.
Como um campo interdisciplinar da ciência, a bioinformatica combina **ciência da computação**, **estatistica**, **matemática**, e **engenharia** para processar, analisar e i**nterpretar dados biológicos** .

ASTROINFORMÁTICA

Astroinformática está focada em desenvolver ferramentas, métodos e **aplicações computacionais**, da **ciência de dados** e da **estatística** para pesquisa e educação na área de astronomia orientada por dados.

O QUARTO PARADIGMA CIENTÍFICO

eScience

ACELERAR A PESQUISA CIENTÍFICA E GERAR CONHECIMENTO COM BASE NA EXPLORAÇÃO DESSE ACÚMULO DE DADOS

Ferramentas avançadas de **software e de mineração** de dados ajudam a interpretar e transformar os dados brutos em **configurações ilimitadas de informação e conhecimento**.

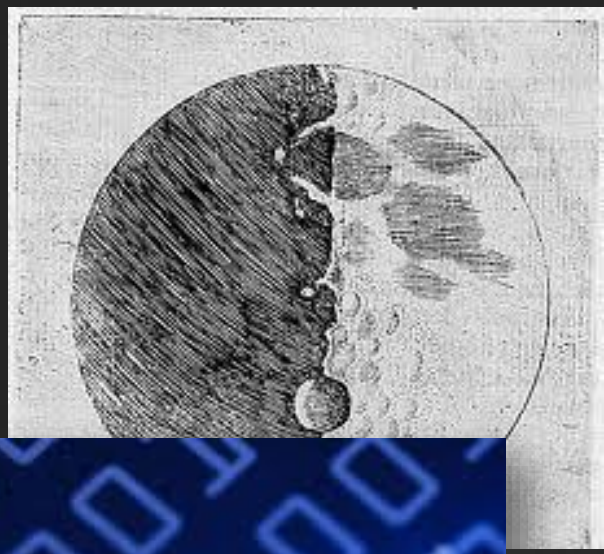
Perguntas instigantes e recursivas colocadas perante os vários segmentos científicos podem agora ser endereçadas, pela combinação de **múltiplas fontes de dados provenientes de domínios diferentes**, através da aplicação de modelos complexos e de métodos inéditos de análise.

O QUARTO PARADIGMA CIENTÍFICO

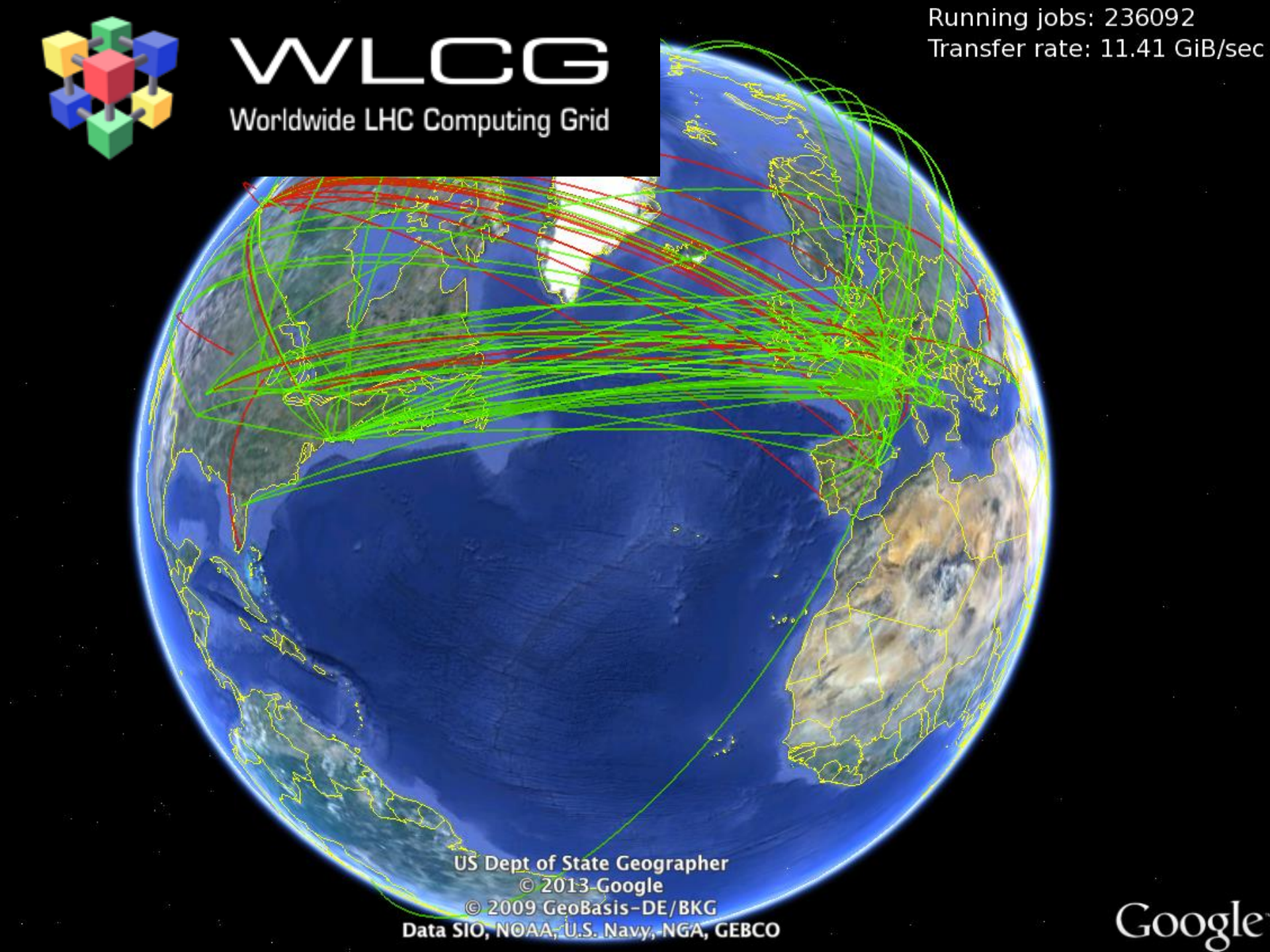
eScience

“

CIÊNCIA PRODUZIDA A PARTIR DO USO, ARMAZENAMENTO, PROCESSAMENTO, ANÁLISE E COMPARTILHAMENTO DE DADOS DE PESQUISA (GRAY, 2005).



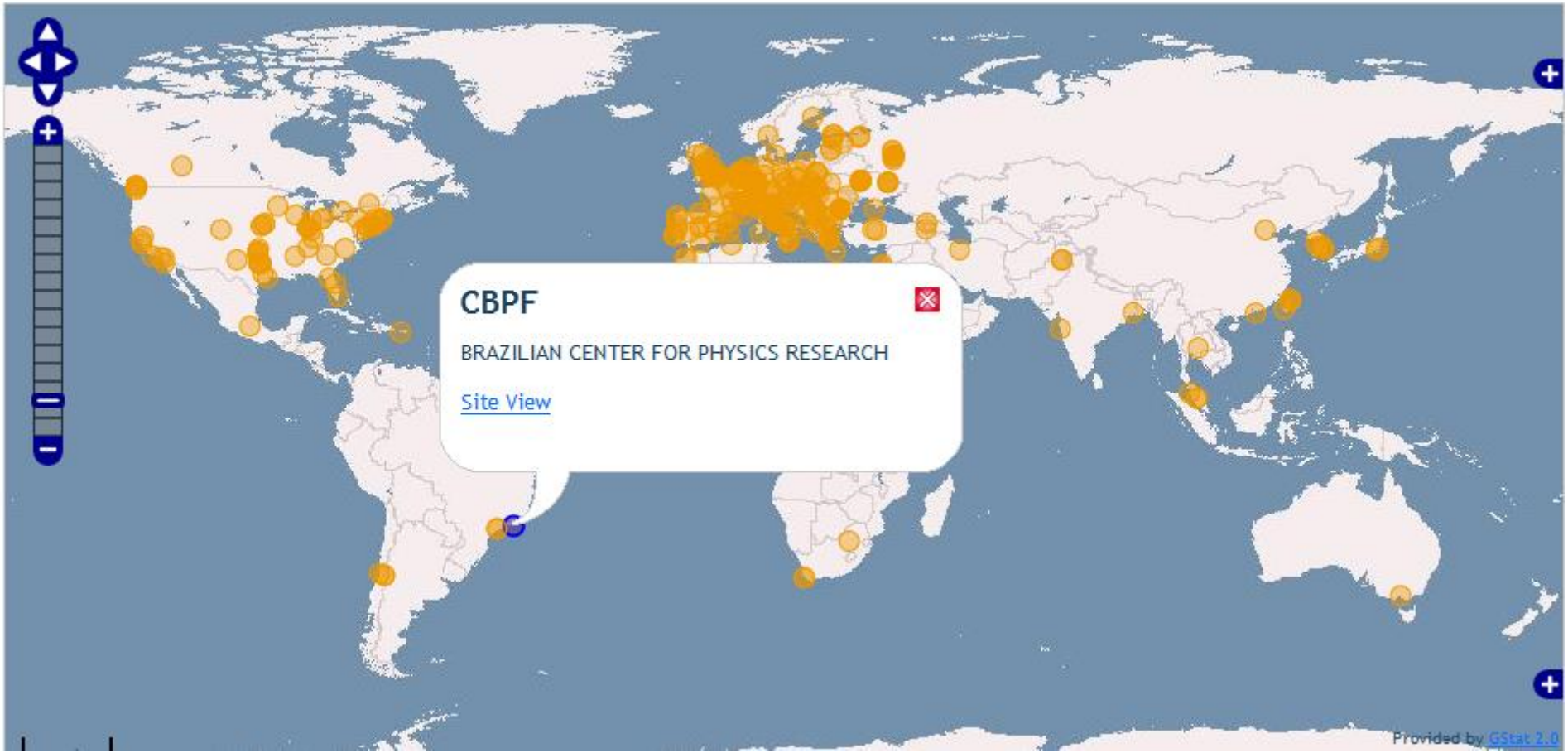
O poder dos computadores modernos permite que **relações altamente complexas** e até então despercebidas possam ser identificadas e se tornem o motor do quarto paradigma


WLCG
Worldwide LHC Computing Grid
Running jobs: 236092
Transfer rate: 11.41 GiB/sec
US Dept of State Geographer
© 2013 Google
© 2009 GeoBasis-DE/BKG
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Google

WORLDWILDE LHC COMPUTING GRID

GLOBAL COLABORATION

42 COUNTRIES – 170 COMPUTING CENTRES



O QUARTO PARADIGMA CIENTÍFICO

eScience

O MODO DE FAZER CIÊNCIA MUDA....

A computação não é mais meramente um suporte para o padrão tradicional de se conduzir a investigação científica em determinadas disciplinas, mas pode mudar fundamentalmente o desenvolvimento dessas disciplinas.

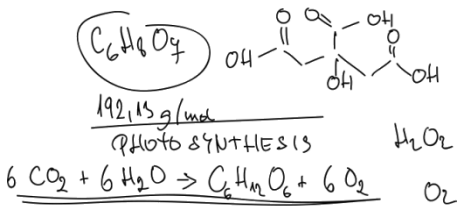
FORMULAÇÃO DE HIPÓTESES

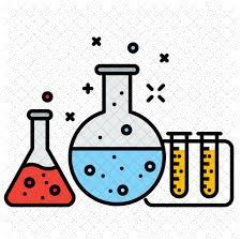
Ao invés de hipóteses serem testadas e desenvolvidas a partir de dados coletados para este propósito, **hipóteses são construídas após a identificação relações nos conjuntos de dados**. Neste abordagem os dados vem primeiro, incorporados numa sequencia de captura de dados, curadoria e análises

TEORIAS/ HIPOTESES

EXPERIMENTOS

OBSERVAÇÕES





PESQUISA ORIENTADA POR HIPÓTESES

PESQUISA ORIENTADA POR DADOS

COLETA DE DADOS

BASES DE DADOS

- ANÁLISE PREDITIVA
- EXTRAÇÃO DE CLUSTER
- DETEÇÃO DE ANOMALIAS
- ANALISE DE CORRELAÇÕES
- ETC.

ALGORÍTMOS SOFISTICADOS FERRAMENTAS ESTATÍSTICAS

PADRÕES, MODELOS HIPÓTESES

O fim da teoria

O DILÚVIO DE DADOS TORNA O MÉTODO CIENTÍFICO OBSOLETO

Chris Anderson (2008)

Illustration: Marian Bantjes "All models are wrong, but some are useful."

Como o título indica, Anderson afirmou que, **na era da informação petabyte e da supercomputação, o método científico tradicional baseado em hipóteses se tornaria obsoleto.** Não há mais teorias ou hipóteses, nem mais discussões se os resultados experimentais refutam ou apoiam as hipóteses originais. **Nesta nova era, o que conta são algoritmos sofisticados e ferramentas estatísticas** para filtrar uma enorme quantidade de dados para encontrar informações que poderiam ser transformadas em conhecimento.

Em vez de buscar resultados precisos sob condições controladas e de campo simplificado, os cientistas são levados a ver na **desordem dos dados um reflexo da complexidade da natureza**

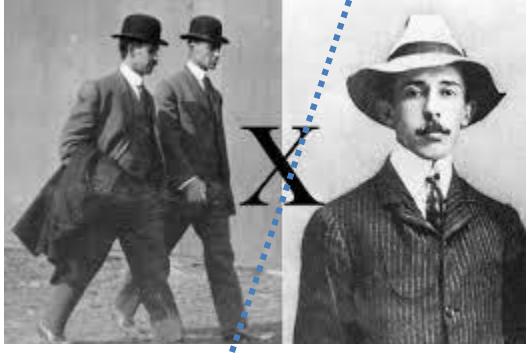
O método científico tradicional está superado?

O big data pode substituir a ciência orientada por hipótese por sofisticados algoritmos e massivas coleções de dados?

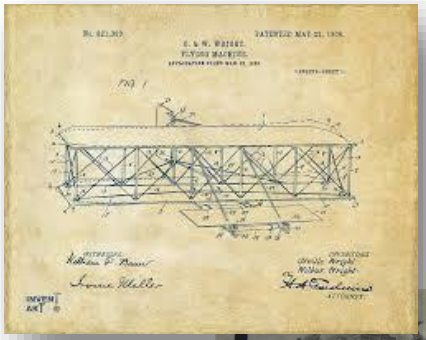
Dada a quantidade de dados científicos disponíveis é possível descartar o papel das formulações teóricas e de hipóteses?

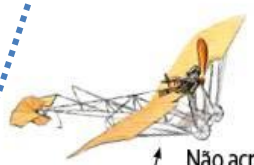
CIÊNCIA ABERTA

No século XX houve um movimento na direção do fechamento do conhecimento patrocinado por grandes corporações no sentido de privar parte das pessoas do acesso ao conhecimento como forma de gerar receita financeira (KON, 2013)

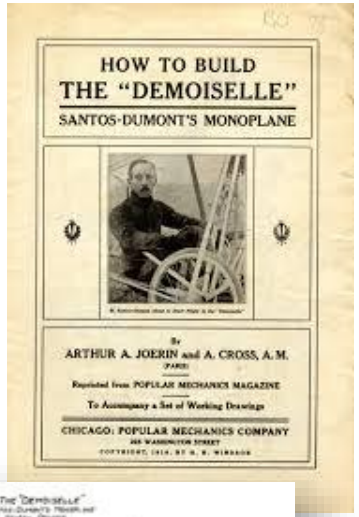


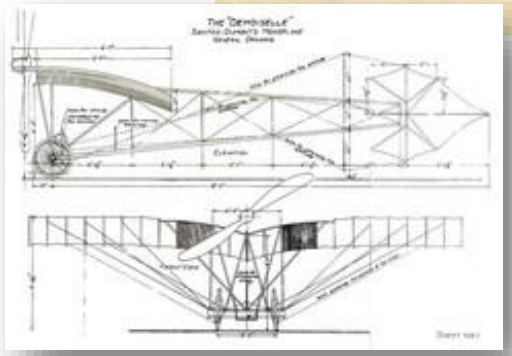
Com a patente, os irmãos receberam propostas lucrativas e passaram a comercializar as máquinas. Fundaram a **Wright Company** em 1909 e, no ano seguinte, fizeram o primeiro voo comercial da história. Ficaram ricos vendendo aviões.





Não acreditava em patentes e divulgou seus estudos para empresas copiarem de graça, vendo suas invenções como um presente à humanidade. Seu modelo **Demoiselle Nº 20** foi o primeiro produzido em série e traz conceitos seguidos até hoje pela indústria aeronáutica

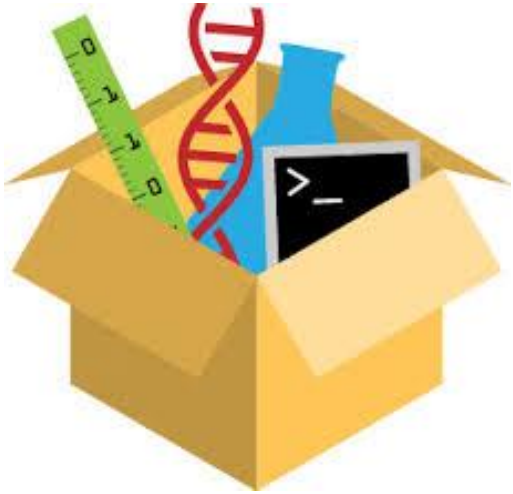




Muitos dos crescimentos marcantes da ciência nos últimos séculos é devido a práticas abertas...

Ciência aberta

O conhecimento científico é um patrimônio da humanidade que, portanto, deve estar disponível livremente para que as pessoas, cientistas ou não, possam usá-lo, reusá-lo, distribui-lo sem constrangimentos tecnológicos, econômicos, sociais ou legais



Quando há compartilhamento de ideias e abertura do conhecimento a ciência avança mais rapidamente

O compartilhamento e o intercâmbio permitem descobrir conexões no que estava antes desconectado

REPRODUTIBILIDADE

Reprodutibilidade dos experimentos científicos é um dos fundamentos da ciência.

TRANSPARÊNCIA NAS METODOLOGIAS

Códigos fontes para reproduzir dos dados; uso de **software livres e formatos abertos**; **ferramentas de pesquisa abertas**; Dados de entrada e metadados **Cadernos de pesquisa abertos**

DISPONIBILIDADE DOS DADOS

Os dados científicos devem estar disponíveis para qualquer pessoa **sem restrições de copyright, patentes ou outros mecanismos de controle**. Dados abertos incentivam o reuso em outras áreas diferentes da original, o que pode levar a descobertas surpreendentes.

ACESSO AOS RESULTADOS

Os pesquisadores devem **divulgar suas descobertas** de forma que elas estejam acessíveis para todos os usuários potenciais sem qualquer barreira.

PESQUISA 2.0

Colaboração crescente entre cientistas efetivada por meio das mídias sociais e da internet. Um número crescente de cientistas estão encontrando novas estratégias para comunicar seus trabalhos usando wikis, blogs, twitter

AVALIAÇÃO

A **avaliação** pelas instituições de pesquisa, bem como a aprovação de **financiamento pelas agências** deve levar em conta a preparação dos dados para disponibilidade na mesma escala em que considera artigos de periódicos e outras publicações, ou seja o nível de transparência.

A ABERTURA DOS DADOS E O SEU IMPACTO NA COMUNICAÇÃO CIENTÍFICA

Novos tipos de publicação

- Publicação ampliada
- Data journal
- Periódico de resultados negativos
- Cadernos eletrônicos de laboratório

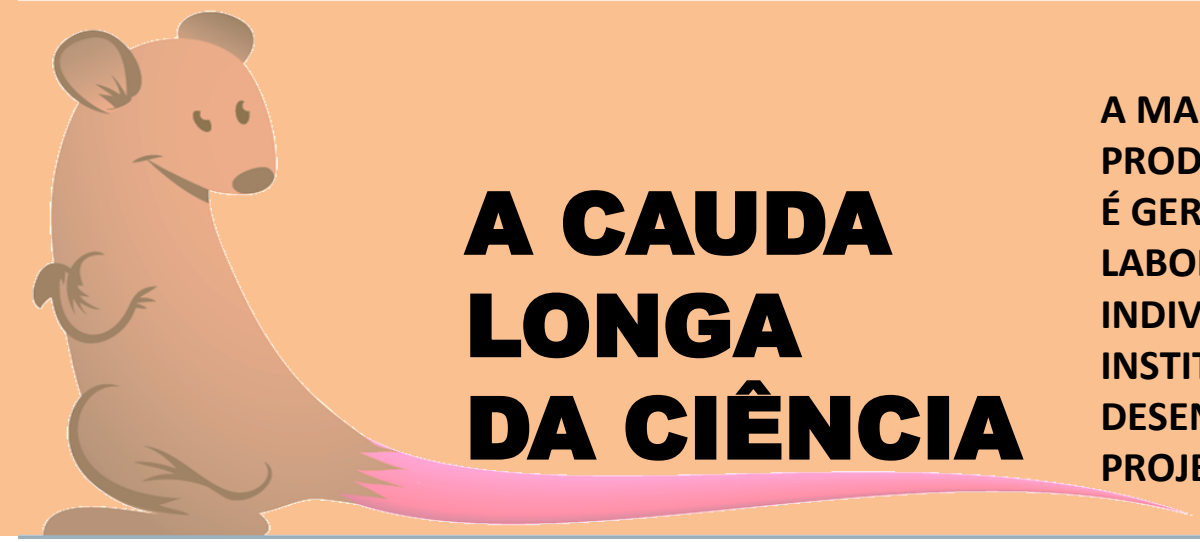
Encurtamento do ciclo de comunicação científica

- Compartilhamento
- Reuso
- Organização dos dados
- Novas formas de socialização

Novos modalidades de revisão

- Consulta aos dados
- Revisão em rede
- Revisão interativa
- Anotações
- Revisão dos dados

CAUDA LONGA DA CIÊNCIA



A CAUDA LONGA DA CIÊNCIA

A MAIORIA DAS COLEÇÕES DE DADOS PRODUZIDAS PELA PESQUISA CIENTÍFICA É GERADO/COLETADO POR PEQUENOS LABORATÓRIOS E PESQUISADORES INDIVIDUALMENTE NAS UNIVERSIDADES E INSTITUTOS DE PESQUISA, QUE DESENVOLVEM UM GRANDE NÚMERO DE PROJETOS CIENTÍFICOS

DOMÍNOS ESPECÍFICOS

ASTRONOMIA
FISICA NUCLEAR
GENOMA
PROTEINA
SENSORIAMENTO REMOTO

Dados da Grande Ciência são fáceis de manipular, compreender e arquivar;
A Pequena Ciência é terrivelmente heterogênea e muito mais vasta e gera 2-3 vezes mais dados do que a Big Science (MacColl, 2010)

Volume dos dados

VÁRIOS DOMÍNIOS E INSTITUIÇÕES

Número de datasets

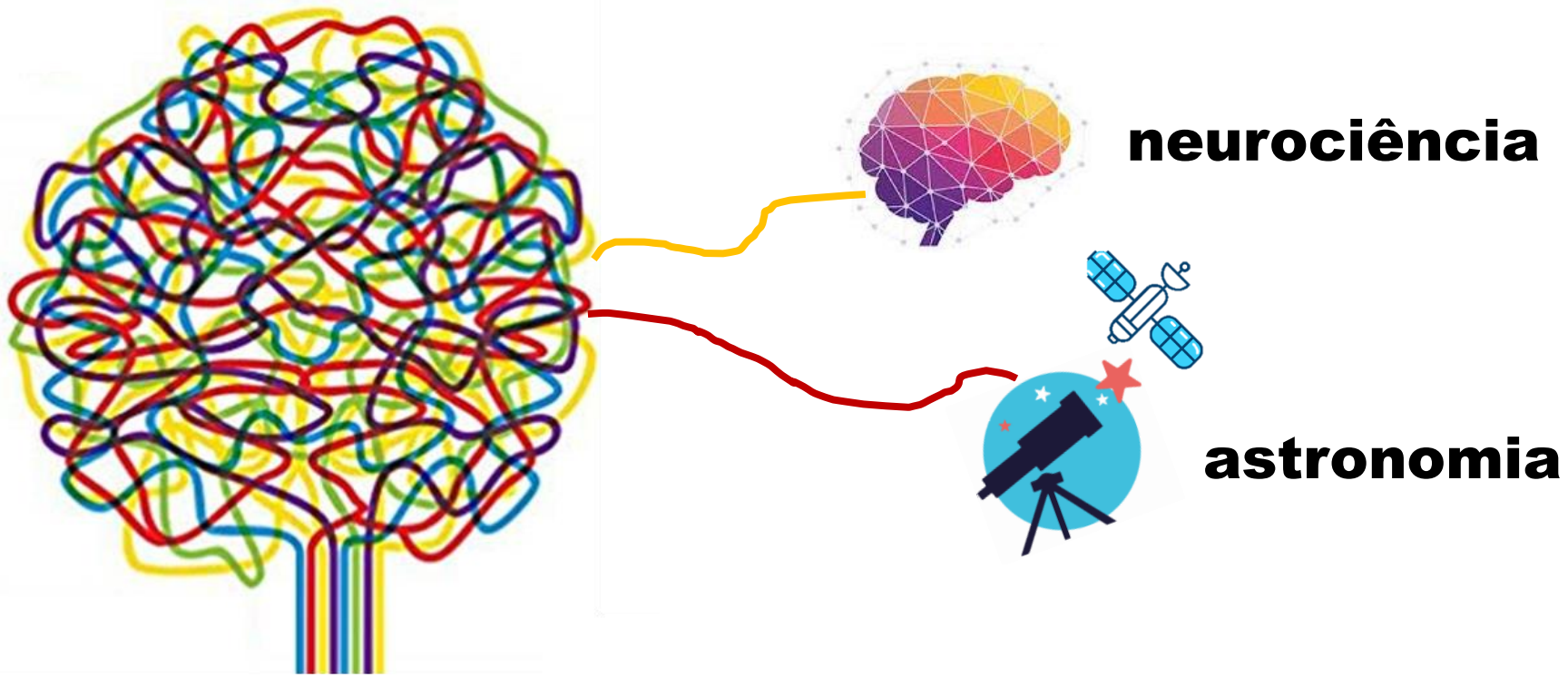
PEQUENOS LABORATÓRIOS, EQUIPES E PESQUISADORES INDIVIDUAIS

HUMANIDADES DIGITAIS

Transformando dados da nossa herança cultural, arquivos históricos, arte, literatura e mídias sociais em novos conhecimentos sobre o mundo que vivemos

DIVERSIDADE DOS DADOS

Os dados da cauda longa, com **sua natureza heterogênea e diversificada**, devem se **integrar a homogeneidade da grande ciência** formando **uma ecologia ou diversidade de dados**. Isto por que nem sempre a grande ciência, definida por predicados homogêneos e estáveis **é o modelo mais adequado para algumas das áreas mais avançadas** e inovadoras da pesquisa científica. Na maioria das vezes, integrar dados formando uma diversidade de dados transversalmente rica, estabelece modelos eficientes de geração de conhecimento



A **perspectiva sistêmica do espaço de dados** torna a integração desses ativos chave **para respostas a novas indagações da ciência**. Isso acontece especialmente ao vincular a estabilidade da grande ciência ao território de alto coeficiente de autonomia e independência da cauda longa, cujas condutas desafiadoras favorecem a inovação e a geração de conhecimentos multi e interdisciplinar.