# Big data em vigilância em saúde

**Wildo Navegantes de Araújo**
**wildo@unb.br**

Brasília - DF
10/08/23

# O que é “big data”?

A quantidade de informação potencial gerada no primeiro dia de vida de um bebê é equivalente a 70 vezes a quantidade de informação (> 155M de itens em 470 idiomas) contida na biblioteca do congresso americano. (Fonte: PBS 2016)

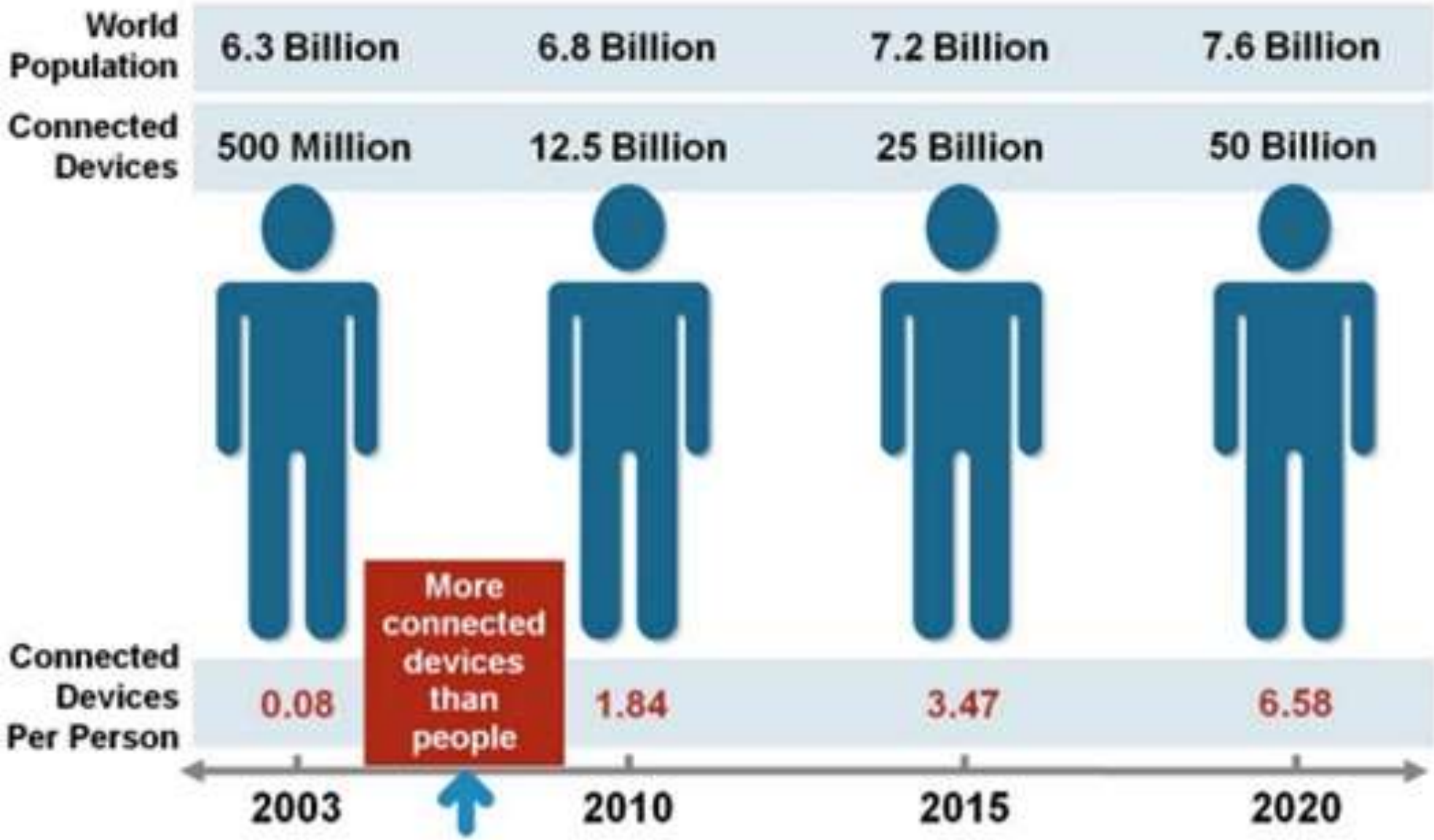
# Netflix

- 500 bilhões de eventos/dia
  - Pico: 8 milhões de eventos/s
- 1.3 petabyte dados/dia (1.000.000 GB)
  - Pico: 24 GB/s
- Uso intenso de aprendizagem de máquina (todas as camadas)

# Twitter

- ~320 milhões de usuários
  - 100 milhões ativos/dia
- 500 milhões de tweets/dia
  - 6 mil/s
  - Pico: 618.725 em 60 segundos (após final da copa 2014)
  - Brasil x Alemanha: 35.6 milhões de tweets
    - Por que?

7 X 1

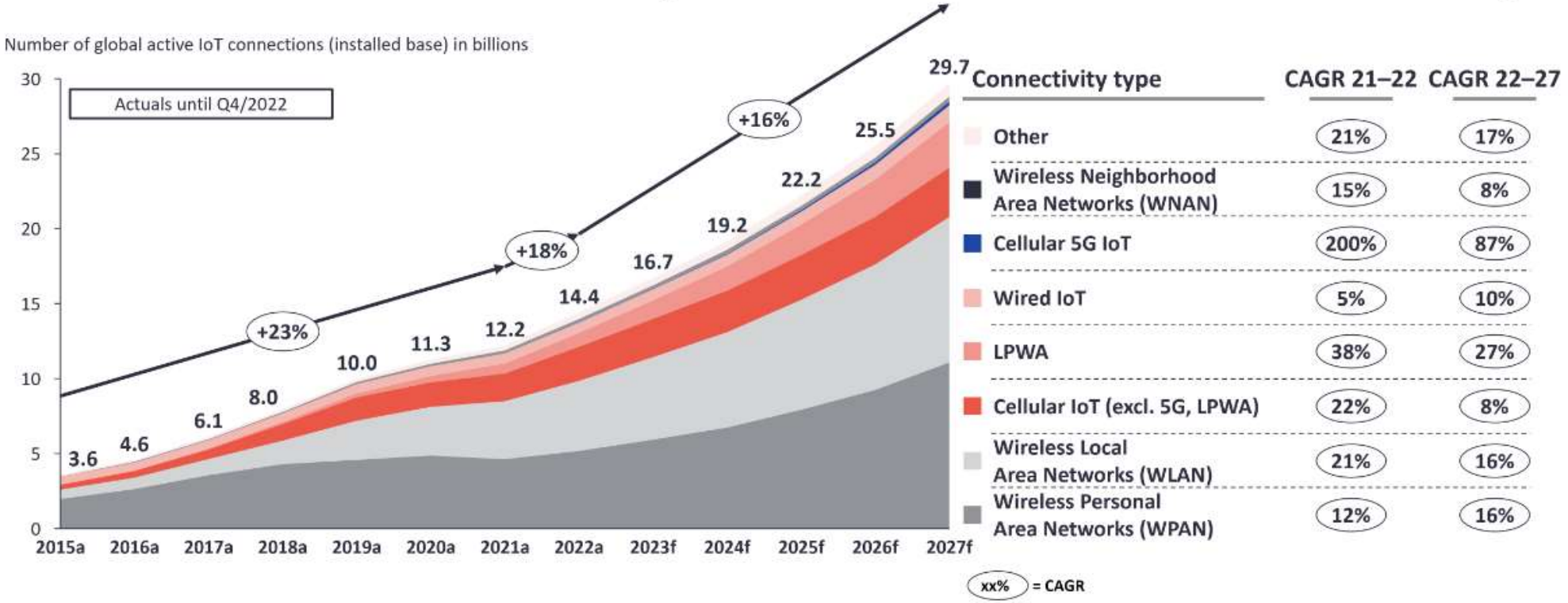

World Population
6.3 Billion
6.8 Billion
7.2 Billion
7.6 Billion
Connected Devices
500 Million
12.5 Billion
25 Billion
50 Billion
More connected devices than people
Connected Devices Per Person
0.08
1.84
3.47
6.58
2003
2010
2015
2020

IOT ANALYTICS

May 2023

Your Global IoT Market Research Partner

# Global IoT market forecast (in billions of connected IoT devices)




**Note**: IoT connections do not include any computers, laptops, fixed phones, cellphones, or consumers tablets. Counted are active nodes/devices or gateways that concentrate the end-sensors, not every sensor/actuator. Simple one-directional communications technology not considered (e.g., RFID, NFC). Wired includes ethernet and fieldbuses (e.g., connected industrial PLCs or I/O modules); Cellular includes 2G, 3G, 4G, 5G; LPWA includes unlicensed and licensed low-power networks; WPAN includes Bluetooth, Zigbee, Z-Wave or similar; WLAN includes Wi-Fi and related protocols; WNAN includes non-short-range mesh, such as Wi-SUN; Other includes satellite and unclassified proprietary networks with any range.

**Source**: IoT Analytics Research 2023. We welcome republishing of images but ask for source citation with a link to the original post and company website.

https://iot-analytics.com/number-connected-iot-devices/

https://ediscoverytoday.com/2023/04/20/2023-internet-minute-infographic-by-ediscovery-today-and-ltmg-ediscovery-trends/

## Usuários de Internet no Brasil (jan/2023)

- 181.8 milhões de pessoas
- Aumentou para 7.1 milhões novos usuários em 2023, em 2022 foram 4.1
- 15.7% da população brasileira sem usar nada da internet
- 152.4 milhões usam mídias sociais

## Uso de rede social no Brasil (jan/2023)

1. WhatsApp (169 mi)
2. YouTube (142 mi)
3. Instagram (113 mi)
4. Facebook (109 mi)
5. TikTok (82 mi)
6. LinkedIn (63 mi)
7. Messenger (62 mi)
8. Kwai (48 mi)
9. Pinterest (28 mi)
10. Twitter (24 mi)

https://datareportal.com/reports/digital-2023-brazil

## Usuários de Internet no Brasil (jan/2023)

- 181.8 milhões de pessoas
- Aumentou para 7.1 milhões novos usuários em 2023, em 2022 foram 4.1
- 15.7% da população brasileira sem usar nada da internet
- 152.4 milhões usam mídias sociais

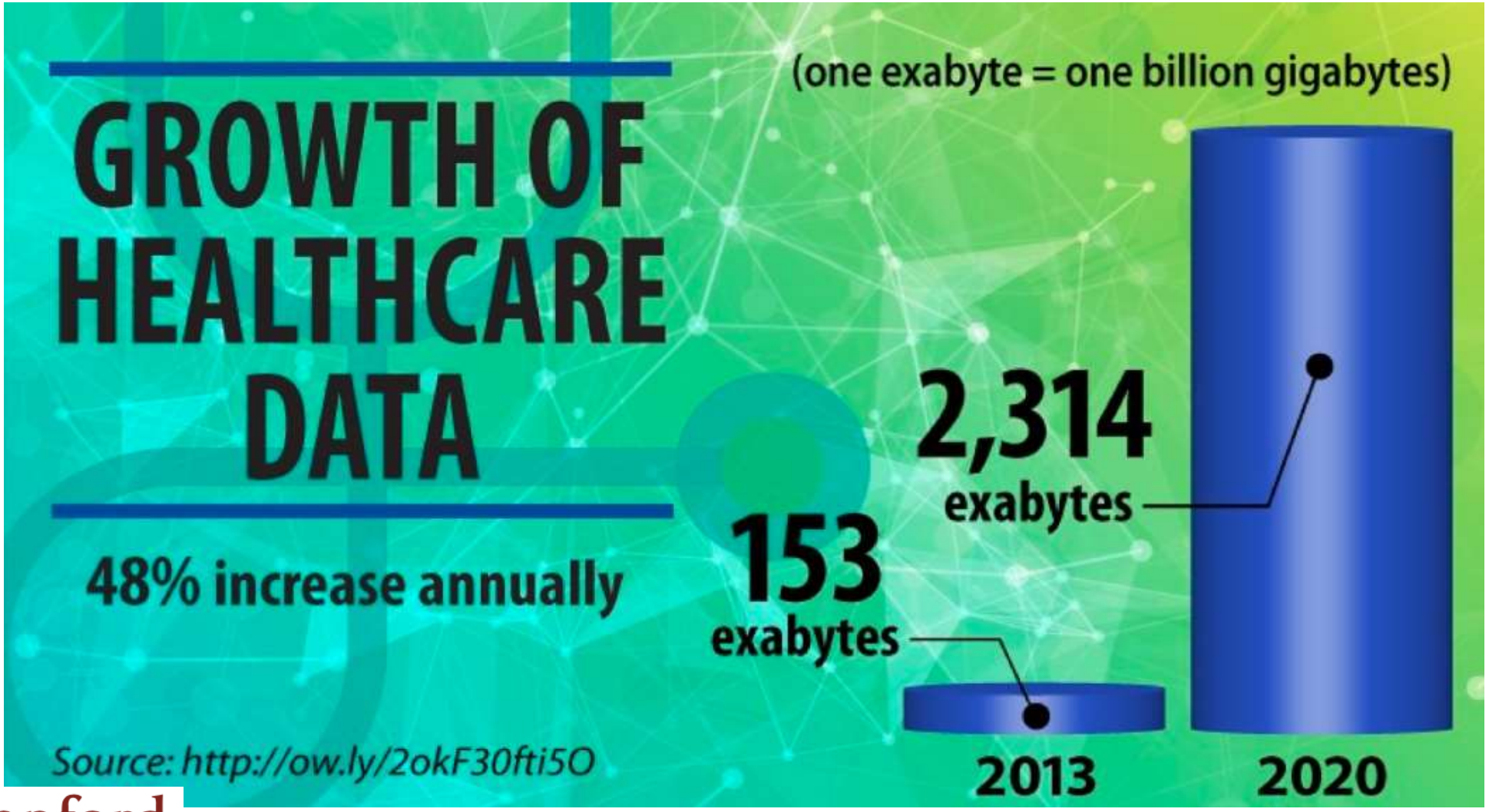
## Uso de rede social no Brasil (jan/2023)

1. WhatsApp (169 mi)
2. YouTube (142 mi)
3. Instagram (113 mi)
4. Facebook (109 mi)
5. TikTok (82 mi)
6. LinkedIn (63 mi)
7. Messenger (62 mi)
8. Kwai (48 mi)
9. Pinterest (28 mi)
10. Twitter (24 mi)

203 milhões hab. (IBGE, 2023)

https://datareportal.com/reports/digital-2023-brazil

UnB



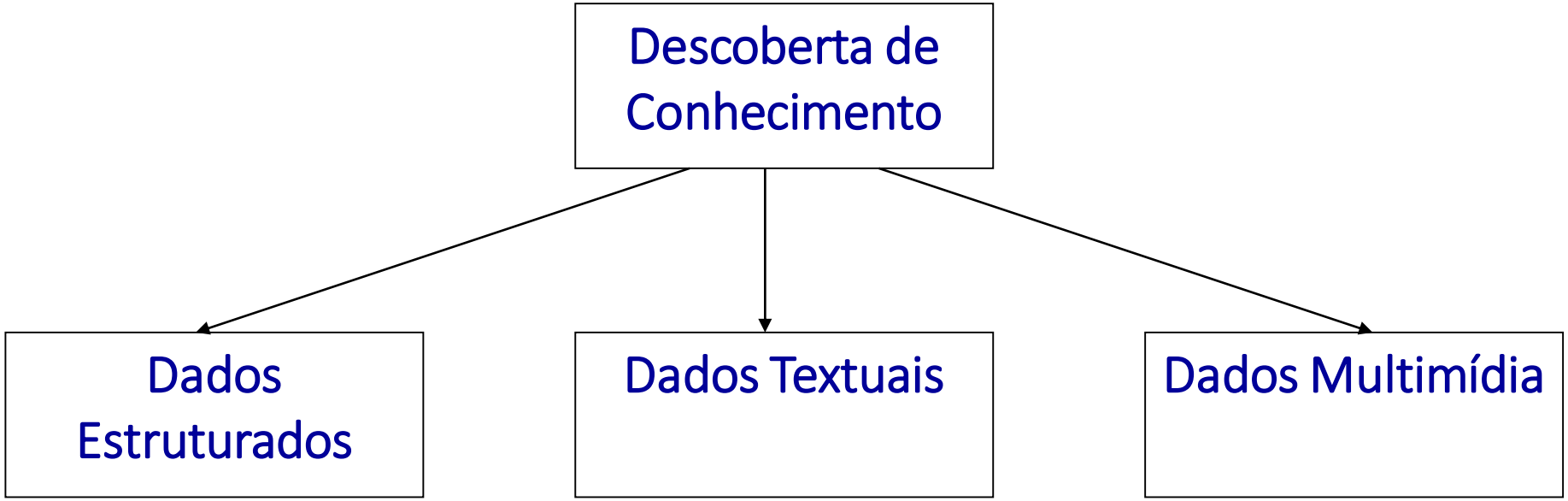

Stanford MEDICINE

# O que é “big data”?

- Big data é um enorme volume de dados estruturados* e não estruturados. O volume é tão grande que é impossível processar com técnicas de banco de dados e software tradicionais.
- Big data é o termo utilizado para uma coleção de conjuntos de dados tão grande e complexo que se torna impossível processar usando ferramentas de gerenciamento de banco de dados ou aplicações de processamento de dados tradicionais.
- Big data são dados cuja escala, diversidade e complexidade exigem novas arquiteturas, técnicas, algoritmos e análises para gerenciá-los e extrair valor e conhecimento oculto deles.
- Big data é a quantidade de dados suficientemente grande que leve a uma mudança nas formas tradicionais de análise de dados (Chiavegatto, 2015).

Chiavegattto, 2015. *Epidemiol. Serv. Saúde*, vol.24, n.2, pp.325-332

UnB




Dbf, Csv, xls

Ex.: Sinan
SIM
SIH

-Twitter
-Koo
-Documentos do Word

Fotos (Instagram)
Telegram
Sons & Vídeos (Youtube, Tiktok, Sopro cardíaco, Discord, batimento cardíaco fetal)

| P | Q | R | S | T | U | V |
|---|---|---|---|---|---|---|
| SEXUAL_ORIENTATION | CLADE | HOSPITALISED | CONCURRRENT_STI | ICU | DATE_RASH | OUTCOME |
| | | NO | UNK | | 15/05/2022 | A |
| | | NO | UNK | | 26/05/2022 | A |
| | | YUNK | UNK | | 02/06/2022 | A |
| | | NO | UNK | | 15/06/2022 | A |
| MSM | WA | YISOL | | | 29/05/2022 | |
| MSM | | NO | | | 03/06/2022 | |
| UNK | | | | | | |
| BISEXUAL | WA | NO | NO | | 03/06/2022 | |
| | | | | | | |
| MSM | WA | NO | NO | | 07/06/2022 | |
| | | NO | | | | |
| MSM | | YTREAT | UNK | | 10/06/2022 | |
| | | NO | | | | |
| | | NO | | | | |
| | | NO | | | | |
| | | NO | | | | |
| | | NO | | | | |
| | | NO | | | | |
| MSM | | | SYPH | | 22/06/2022 | |
| MSM | | | | | 19/06/2022 | |
| MSM | WA | | | | 22/06/2022 | |
| | | NO | | | 02/06/2022 | |
| | | | | | | |
| | | | | | | |
| MSM | WA | NO | SYPH | | 17/06/2022 | A |
| MSM | WA | NO | SYPH | | 07/06/2022 | A |
| MSM | WA | NO | SYPH | | 07/06/2022 | A |
| UNK | WA | NO | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| MSM | | NO | SYPH | | 18/06/2022 | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

# O que é “big data”?

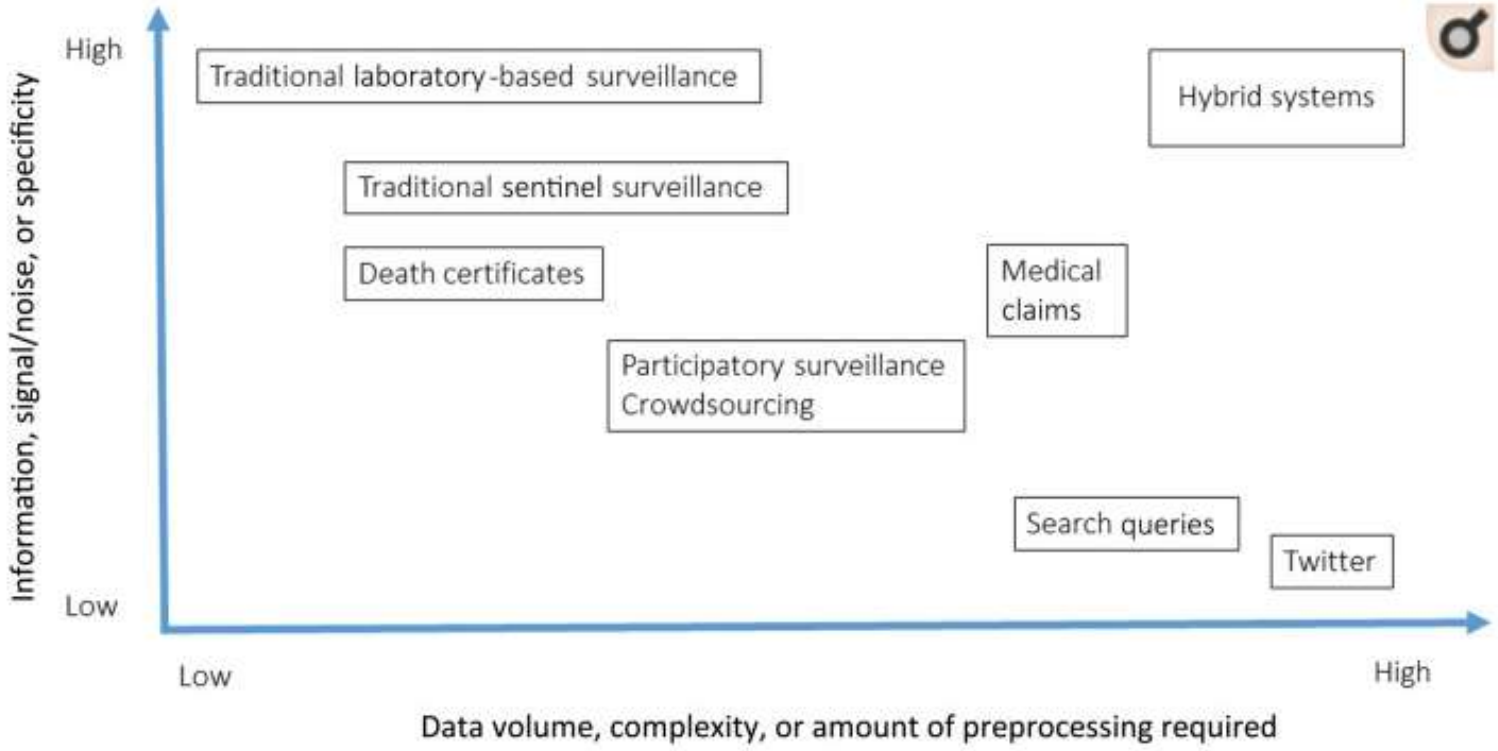
- Não existe um consenso quanto a sua definição
- Refere-se a capacidade de explorar uma grande quantidade de dados, analisá-la e extrair conclusões.
- Representa uma nova era na exploração e utilização de dados[1]
- Ajuda a conhecer a monitorar ameaças e benefícios ao longo prazo
- Tem/Terá impacto sobre a economia, ciência, saúde e sociedade em geral[2]

1.Zikopoulos, P. Eaton, C. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 2011

2. Mayer-schönberger, V; Cukier, K. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, 2013

# O que é “big data”?

- Envolve a gestão e análise de grandes volumes de dados:

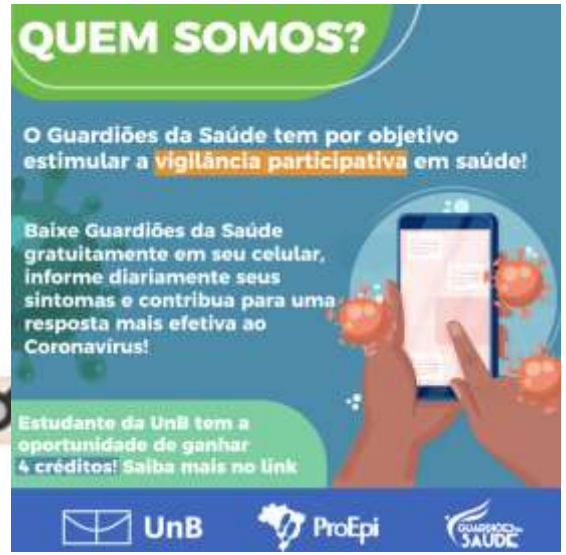


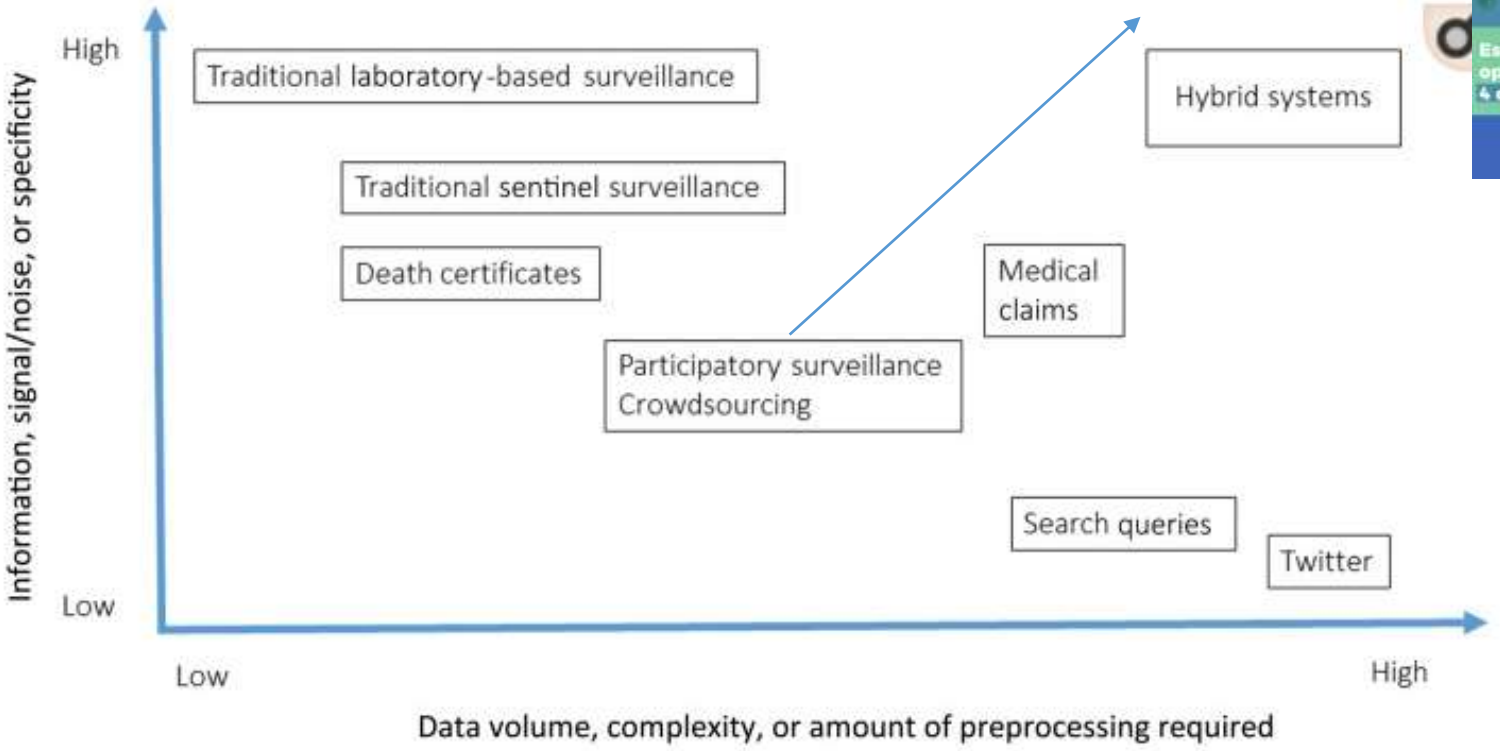

Overview of the characteristics of infectious disease surveillance systems. Hybrid systems combining traditional surveillance with big data streams fall in the desirable zone associated with high information return and high data volume.

J Infect Dis. 2016 Dec 1; 214(Suppl 4): S380–S385.

# O que é “big data”?

- Envolve a gestão e análise de grandes volumes de dados:



Overview of the characteristics of infectious disease surveillance systems. Hybrid systems combining traditional surveillance with big data streams fall in the desirable zone associated with high information return and high data volume.

J Infect Dis. 2016 Dec 1; 214(Suppl 4): S380–S385.

# O que está por trás de Big Data (7 V):

# Por que eu devo usar big data?

- Entender padrões
- Modelagem e Predição
  - Aumentar a precisão de estimativas
- Estabelecer limites e fronteiras
- Usar dados secundários:
  - Os usuários “pagam” para “ofertar” dados
  - O Estado “cobra” para se responsabilizar pelos dados
- Estimar parâmetros desconhecidos

# Por que eu devo usar big data em vigilância em saúde?

- Entender padrões
- Modelagem e Predição
  - Aumentar a precisão de estimativas
- Usar dados secundários:
  - É o país com mais de 100M habitantes com sistema universal de saúde com mais dados em saúde estruturados, padronizados, e “públicos” para análise do mundo.
- Ofertar possibilidades de (avaliar) intervenção e melhoria de qualidade de vida

UnB

Volume 214, Issue suppl_4
1 December 2016

Cover image

ISSN 0022-1899
EISSN 1537-6613

**BIG DATA FOR INFECTIOUS DISEASE SURVEILLANCE AND MODELING**

On the cover: Big data tools can help track and predict the global spread of infectious diseases (red = intense disease activity).

https://academic.oup.com/jid/issue/214/suppl_4

A Brief History of Disease Surveillance and Beginnings of Big Data

| Year | Disease(s) | Key Thinker/Surveillance System | Institution(s) |
|---|---|---|---|
| 1662 | Plague | John Graunt | Bills of Mortality, UK [1] |
| 1817 | Smallpox | J. C. Moore | [2] |
| 1847 | Influenza | William Farr | General Registrar Office, UK [3] |
| 1854 | Cholera | John Snow | General Registrar Office, UK [4] |
| ~1900 | All | International Classification of Diseases | WHO [5] |
| 1918 | Influenza | 121 Cities Mortality Reporting System | CDC [6] |
| 1976 | Influenza | Weekly viral surveillance | CDC [6] |
| 1984 | Influenza, viral hepatitis, acute urethritis, measles, mumps | First computerized disease surveillance network | French Sentinelles Network [7] |
| ~2000 | All | Medical claims forms are first used | Private companies and government [8, 9] |
| 2008 | Influenza | Google Flu Trends launched | Google [10, 11] |
| 2015 | Influenza | Birth of hybrid systems | Public/private partnership [12] |

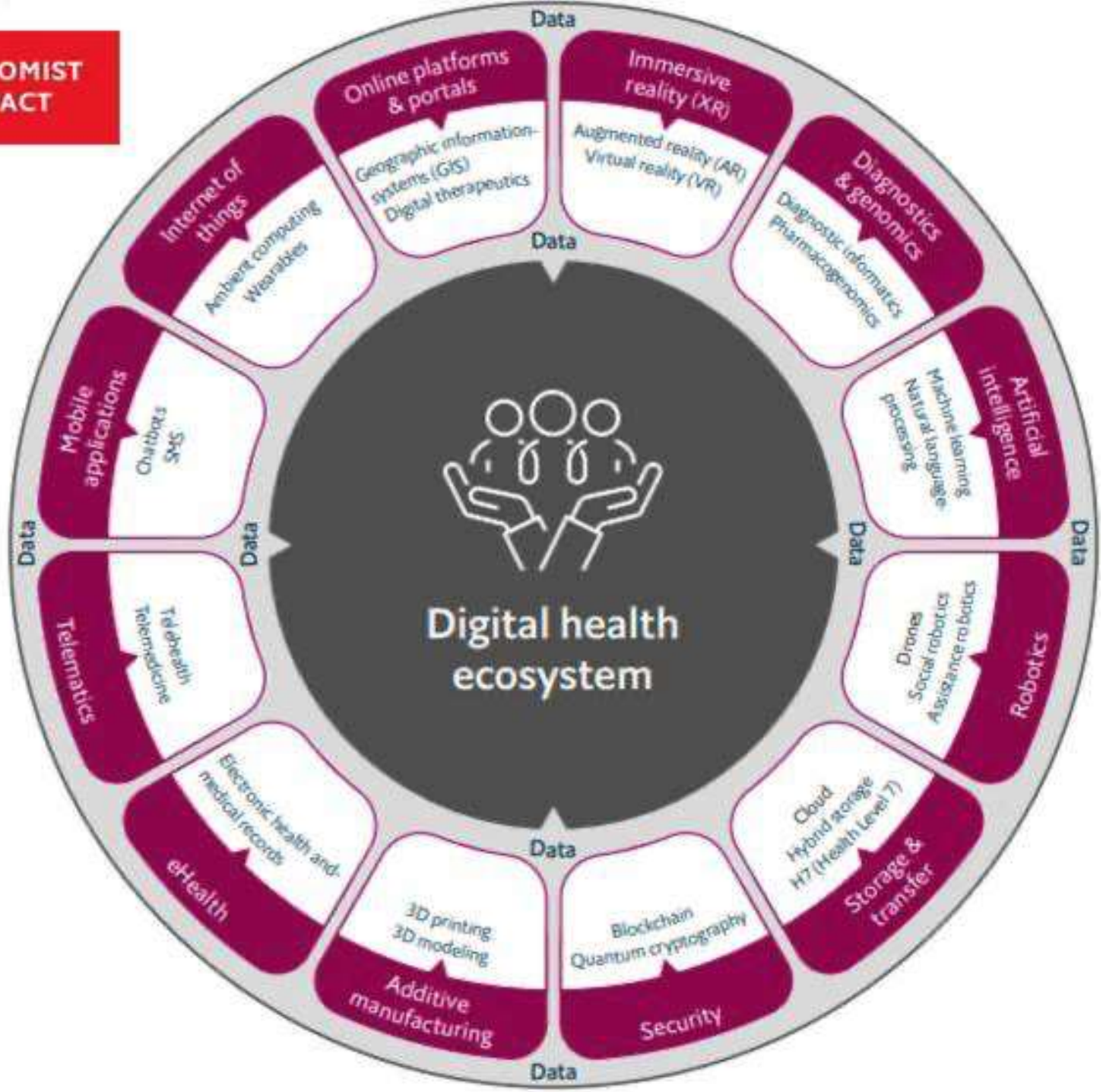
J Infect Dis. 2016 Dec 1; 214(Suppl 4): S380–S385.

UnB

# Aplicações conhecidas:

UnB

ECONOMIST IMPACT



# Quais ferramentas usar?

UnB

# Quais ferramentas usar?

Depende...e
Não existe uma única ferramenta

UnB

# Big Data Landscape 2016 (Version 3.0)

## Infrastructure

- Hadoop On-Premise
- Hadoop in the Cloud
- Spark
- Cluster Services
- NoSQL Databases
- NewSQL Databases
- Graph Databases
- MPP Databases
- Cloud EDW
- Data Transformation
- Data Integration
- Management / Monitoring
- Security
- Storage
- App Dev
- Crowd-sourcing

## Analytics

- Analyst Platforms
- Analytics Platforms
- Data Science Platforms
- Visualization
- BI Platforms
- Statistical Computing
- Log Analytics
- Social Analytics
- Real-Time
- Machine Learning
- Speech & NLP
- Horizontal AI
- Search
- Data Services
- For Business Analysts
- Web / Mobile / Commerce

## Applications

- Sales & Marketing
- Customer Service
- Human Capital
- Legal
- Ad Optimization
- Security
- Vertical AI Applications
- Publisher Tools
- Govt / Regulation
- Finance
- Education/ Learning
- Life Sciences
- Industries

## Cross-Infrastructure/Analytics

## Open Source

- Framework
- Query / Data Flow
- Data Access
- Coordination
- Real-Time
- Stat Tools
- Machine Learning
- Search
- Security
- Visualization

## Data Sources & APIs

- Health
- IOT
- Financial & Economic Data
- Air / Space / Sea
- Location / People / Entities
- Other

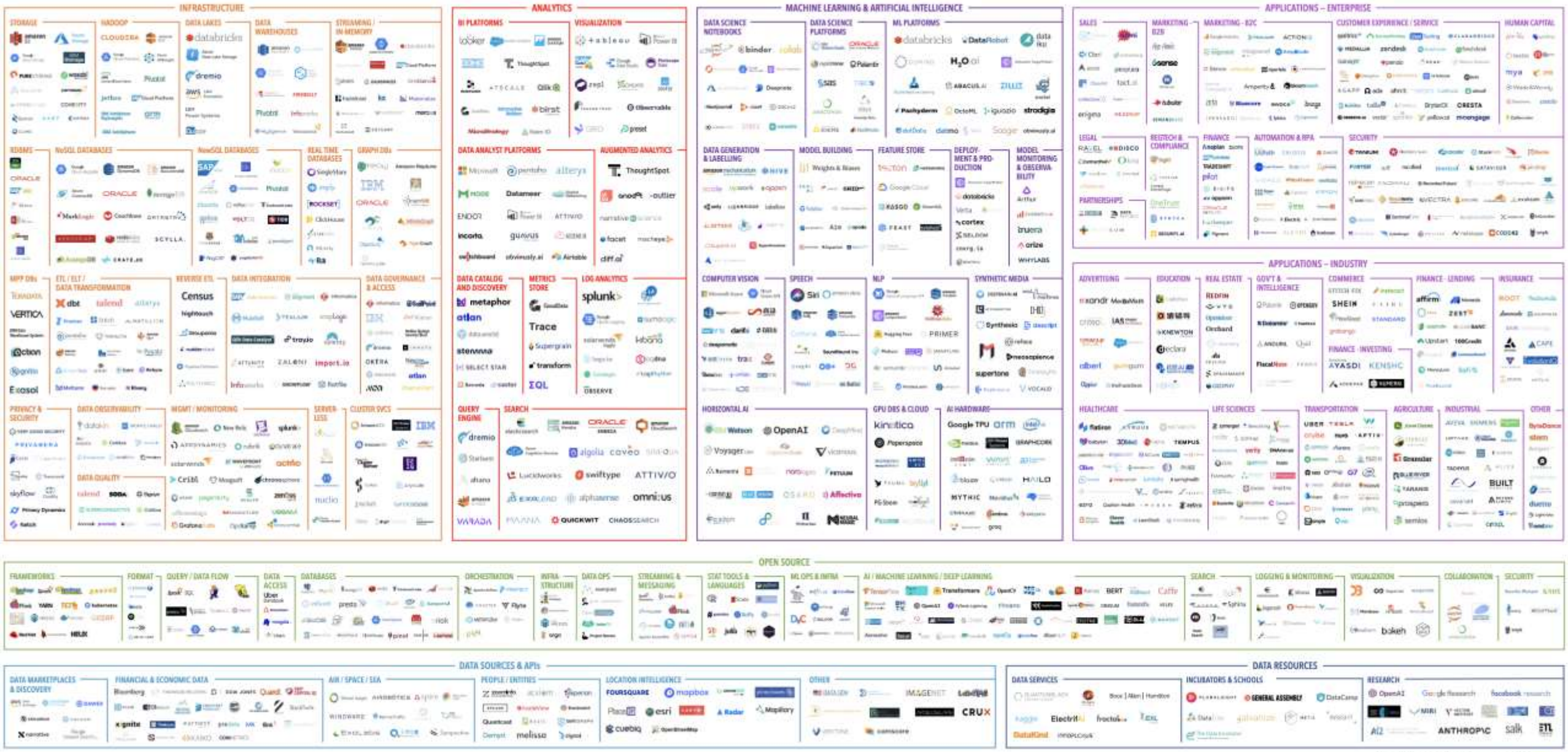
## Incubators & Schools

*Last Updated 3/23/2016*

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

# MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



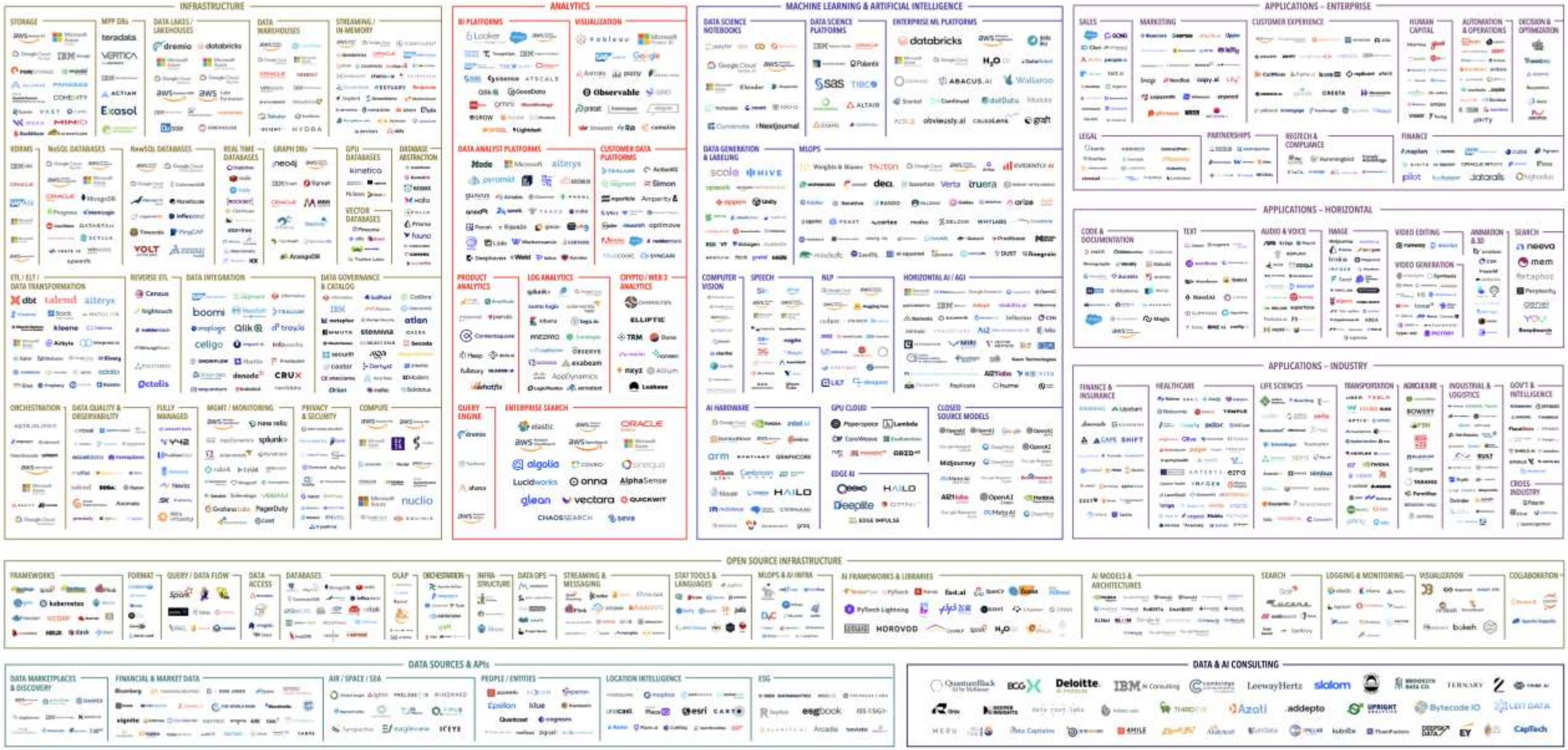

Version 3.0 - November 2021

© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

mattturck.com/data2021

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

# THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE






© Matt Turck (@mattturck), Kevin Zhang (@ykevinzhang) & FirstMark (@firstmarkcap)

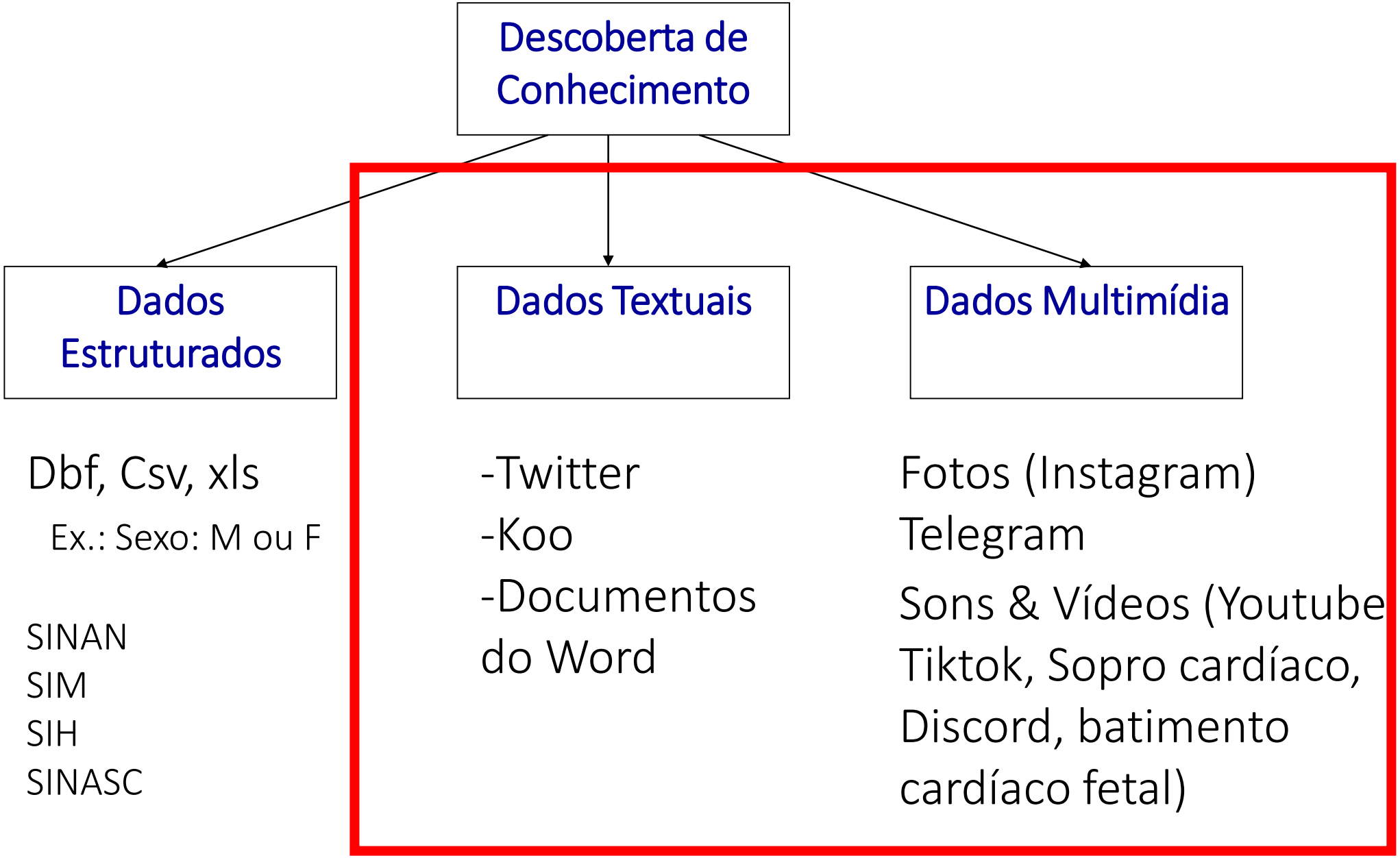
Blog post: mattturck.com/MAD2023

Interactive version: MAD.firstmarkcap.com

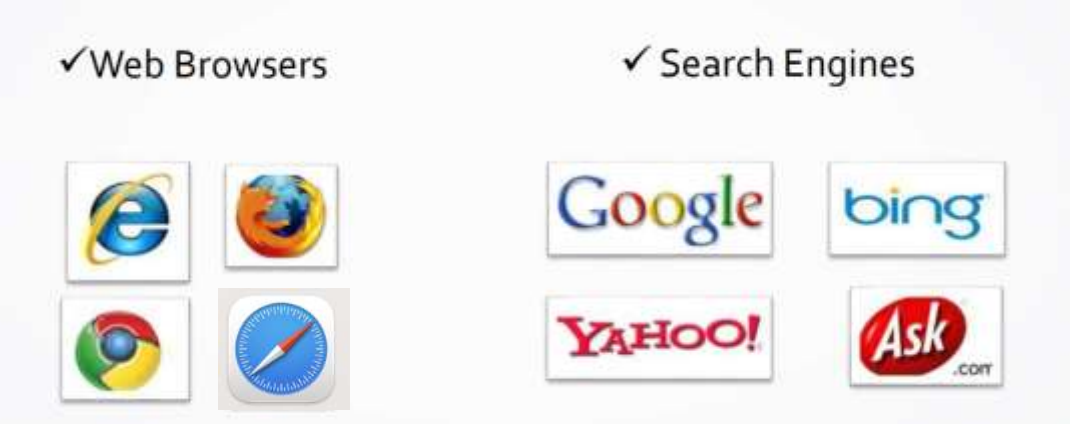
Comments? Email MAD2023@firstmarkcap.com

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

# Exemplos:


Descoberta de Conhecimento
Dados Estruturados
Dados Textuais
Dados Multimídia
Dbf, Csv, xls
Ex.: Sexo: M ou F
SINAN
SIM
SIH
SINASC
-Twitter
-Koo
-Documentos do Word
Fotos (Instagram)
Telegram
Sons & Vídeos (Youtube Tiktok, Sopro cardíaco, Discord, batimento cardíaco fetal)

# Quem coleta dados para além do setor saúde ?



# Desafios para o Big Data

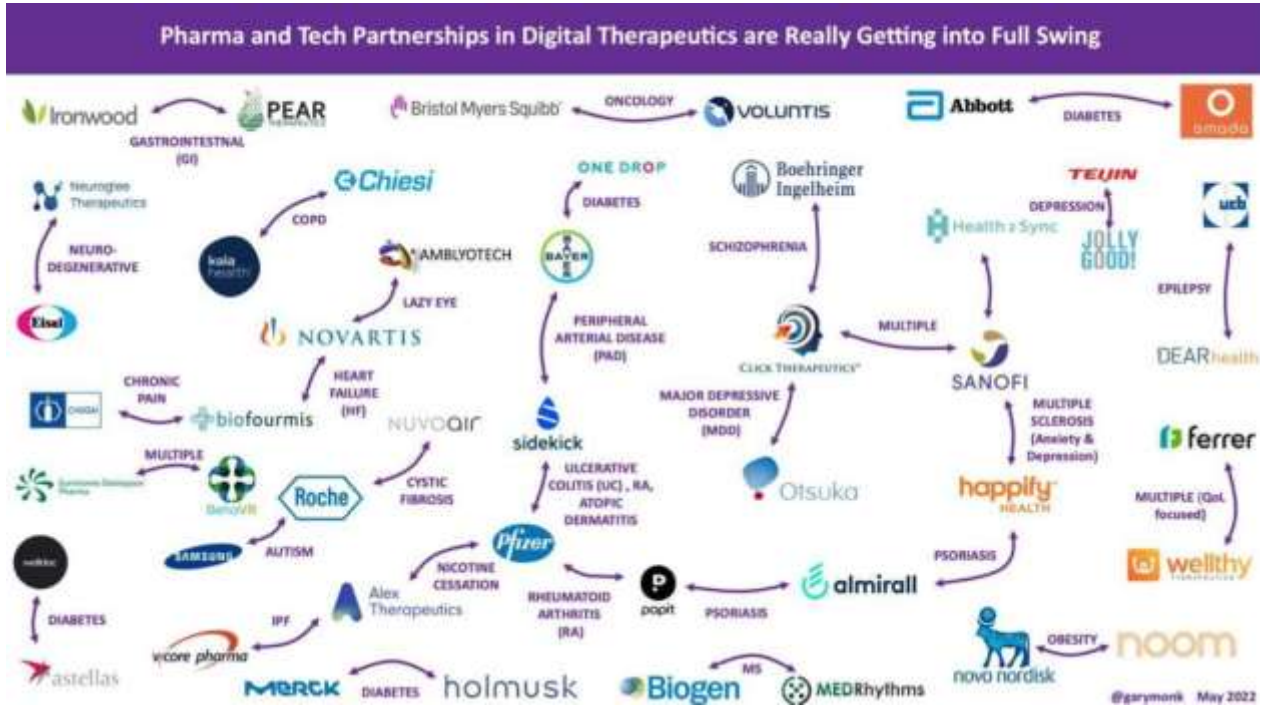
- Formação de equipes multiprofissionais, hardwares e nuvens
- A automatização de pesquisas altera a definição do conhecimento
  - A história dos “algoritmos impulsionadores”
- Objetividade e precisão das informações
- Grande volume de dados não garante informações melhores
- Nem todos os dados são equivalentes
- Só porque o dado é acessível não significa que seja ético
- O acesso limitado ao Big Data cria novas divisas digitais

Borges do Nascimento et al, https://www.jmir.org/2021/4/e27275

# Exemplos de aplicações:

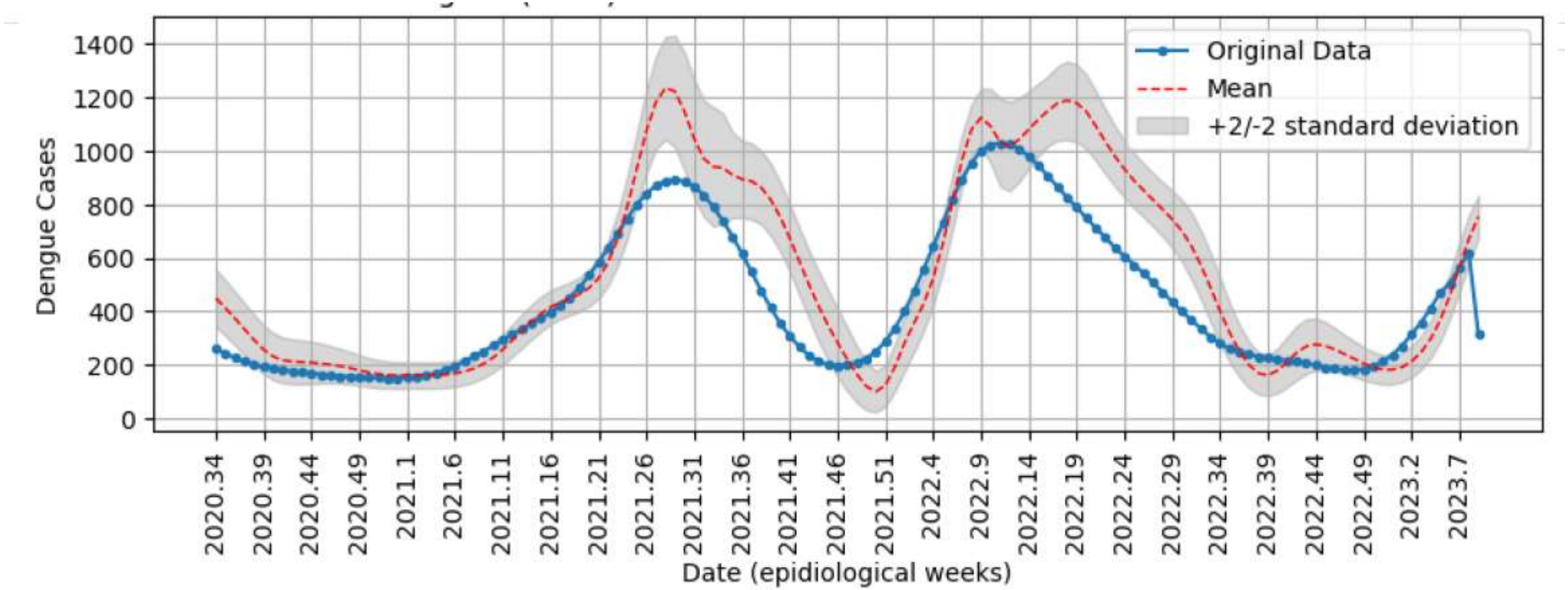
# Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health

Marcel Salathé



Monitoramento do horizonte tecnológico !

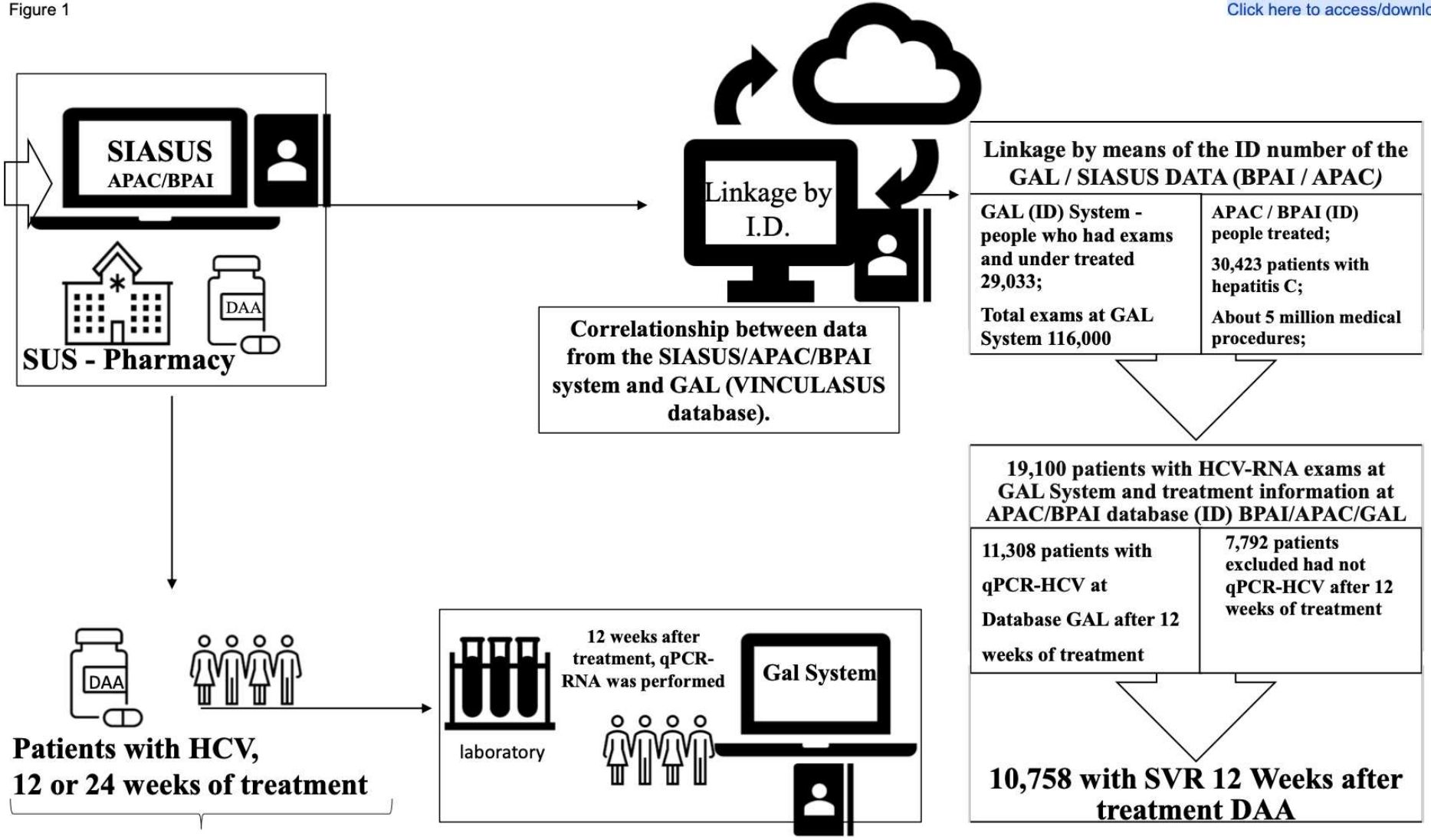
# Nowcasting e forecasting usando bigdata por inteligência artificial



# Bigdata para estudar um política pública:
# - Efetividade das DAAs em HCV

Figure 1

Click here to access/download;Figure;f




Vivaldini et al (2021)

UnB

## Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health

Marcel Salathé

## Participatory Syndromic Surveillance of Influenza in Europe

Caroline Guerrisi,[1] Clement Turbelin,[1] Thierry Blanchon,[1] Thomas Hanslik,[1,2] Isabelle Bonmarin,[3] Daniel Levy-Bruhl,[3] Daniela Perrotta,[4] Daniela Paolotti,[4] Ronald Smallenburg,[5] Carl Koppeschaar,[5] Ana O. Franco,[6] Ricardo Mexia,[7] W. John Edmunds,[8] Bersabeh Sile,[9] Richard Pebody,[9] Edward van Straten,[10] Sandro Meloni,[11] Yamir Moreno,[11] Jim Duggan,[12] Charlotte Kjelso,[13] and Vittoria Colizza[1,4]

346 IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 22, NO. 2, MARCH 2018

## On the Accuracy and Scalability of Probabilistic Data Linkage Over the Brazilian 114 Million Cohort

Robespierre Pita, Clícia Pinto, Samila Sena, Rosemeire Fiaccone, Leila Amorim, Sandra Reis, Mauricio L. Barreto, Spiros Denaxas, and Marcos Ennes Barreto

## A Platform for Monitoring Regional Antimicrobial Resistance, Using Online Data Sources: ResistanceOpen

Derek R. MacFadden,[1,2] David Fisman,[1] Jeff Andre,[2] Yuki Ara,[2] Maimuna S. Majumder,[2,4] Isaac I. Bogoch,[1] Nick Daneman,[1] Annie Wang,[1] Marianna Vavitsas,[2] Lucas Castellani,[1] and John S. Brownstein[2,3]

[1]Division of Infectious Diseases, University of Toronto, Canada; [2]Boston Children's Hospital, [3]Harvard Medical School, Boston, and [4]Massachusetts Institute of Technology, Cambridge

2017 IEEE 30th International Symposium on Computer-Based Medical Systems

## Probabilistic integration of large Brazilian socioeconomic and clinical databases

Clicia Pinto*, Robespierre Pita*, George Barbosa*, Bruno Araújo*, Juracy Bertoldo*, Samila Sena*, Sandra Reis*, Rosemeire Fiaccone*, Leila Amorim*, Maria Yuri Ichihara*, Mauricio Barreto*, Marcos Barreto*†, Spiros Denaxas†

*Centre for Data and Knowledge Integration for Health (CIDACS), FIOCRUZ, Salvador, Brazil

The Journal of Infectious Diseases

SUPPLEMENT ARTICLE

## Big Data for Infectious Disease Surveillance and Modeling

Shweta Bansal,[1,2] Gerardo Chowell,[1,3] Lone Simonsen,[1,4] Alessandro Vespignani,[5] and Cécile Viboud[1]

UnB

# Aplicações conhecidas em pandemia

nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-021-26459-6 OPEN

## Effectiveness of ChAdOx1 vaccine in older adults during SARS-CoV-2 Gamma variant circulation in São Paulo

Matt D. T. Hitchings[1,19], Otavio T. Ranzani[2,3,19], Murilo Dorion[4], Tatiana Lang D'Agostini[5], Regiane Cardoso de Paula[5], Olivia Ferreira Pereira de Paula[5], Edlaine Faria de Moura Villela[5], Mario Sergio Scaramuzzini Torres[6], Silvano Barbosa de Oliveira[7,8], Wade Schulz[9], Maria Almiron[7], Rodrigo Said[7], Roberto Dias de Oliveira[10], Patricia Vieira Silva[11], Wildo Navegantes de Araújo[7,8,12], Jean Carlo Gorinchteyn[13], Jason R. Andrews[14], Derek A. T. Cummings[15,16], Albert I. Ko[4,17] & Julio Croda[4,11,18]✉

https://www.nature.com/articles/s41467-021-26459-6.pdf

UnB

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-26459-6 OPEN

Effectiveness of ChAdOx1 vaccine in older adults during SARS-CoV-2 Gamma variant circulation in São Paulo

Matt D. T. Hitchings[1,19], Otavio T. Ranzani[2,3,19], Murilo Dorion[4], Tatiana Lang D'Agostini[5], Regiane Cardoso de Paula[5], Olivia Ferreira Pereira de Paula[5], Edlaine Faria de Moura Villela[5], Mario Sergio Scaramuzzini Torres[6], Silvano Barbosa de Oliveira[7,8], Wade Schulz[9], Maria Almiron[7], Rodrigo Said[7], Roberto Dias de Oliveira[10], Patricia Vieira Silva[11], Wildo Navegantes de Araújo[7,8,12], Jean Carlo Gorinchteyn[13], Jason R. Andrews[14], Derek A. T. Cummings[15,16], Albert I. Ko[4,17] & Julio Croda[4,11,18]

# Aplicações conhecidas:

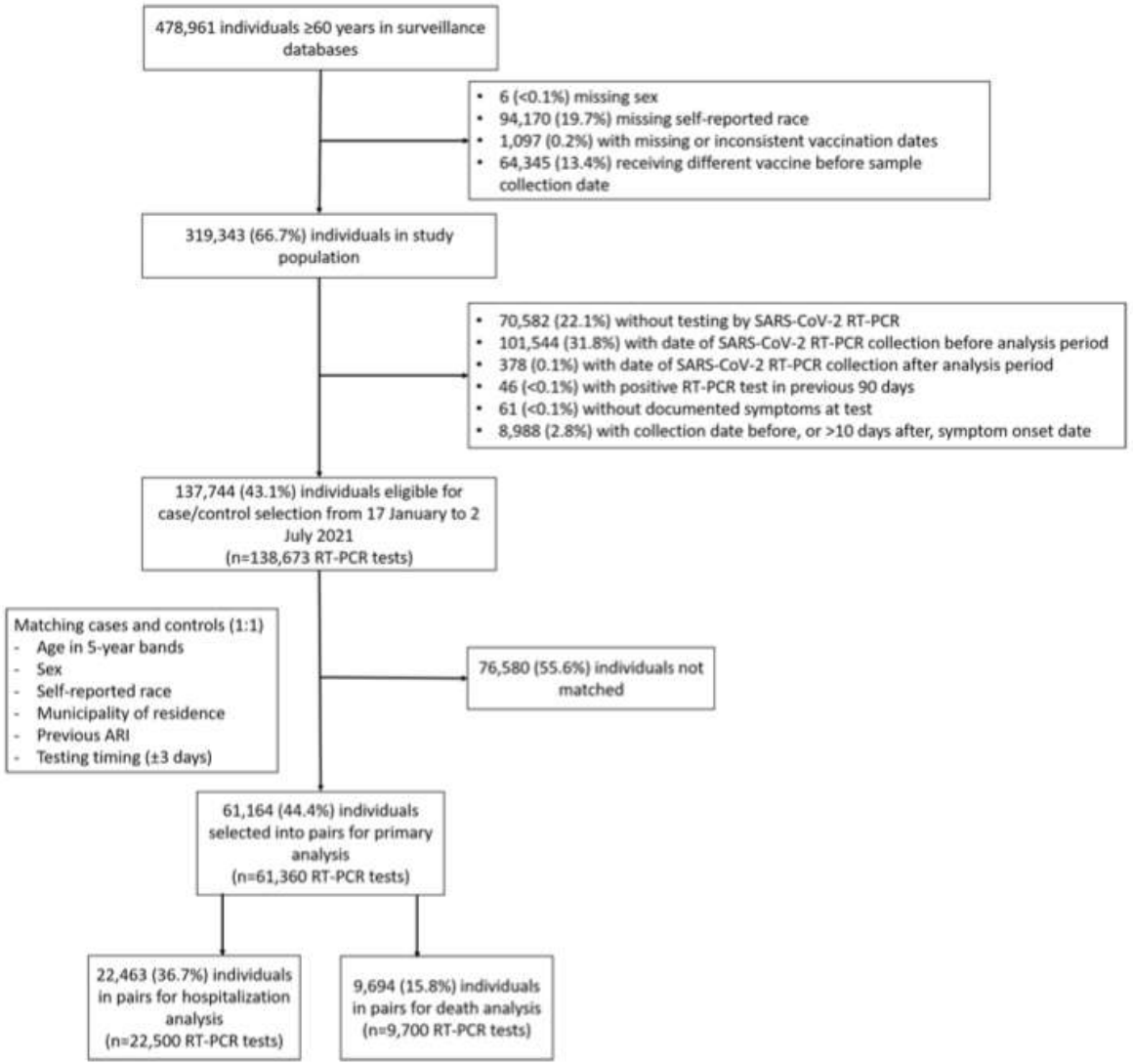



**Fig. 2 Study flowchart.** Flowchart of the identification of the study population from surveillance databases and selection of matched cases and controls.

https://www.nature.com/articles/s41467-021-26459-6.pdf

UnB

# Aplicações conhecidas:

ARTICLES
https://doi.org/10.1038/s41591-022-01701-w

nature medicine

Check for updates

OPEN

## Vaccine effectiveness of heterologous CoronaVac plus BNT162b2 in Brazil

Thiago Cerqueira-Silva[1,2,11], Srinivasa Vittal Katikireddi[3,4,11], Vinicius de Araujo Oliveira[2,5], Renzo Flores-Ortiz[5], Juracy Bertoldo Júnior[2,5], Enny S. Paixão[6], Chris Robertson[4,7], Gerson O. Penna[8], Guilherme L. Werneck[9], Maurício L. Barreto[2,5], Neil Pearce[6], Aziz Sheikh[10,12], Manoel Barral-Netto[1,2,5,12] and Viviane S. Boaventura[1,2,12] ✉

https://www.nature.com/articles/s41591-022-01701-w.pdf

# Aplicações conhecidas:

ARTICLES
https://doi.org/10.1038/s41591-022-01701-w
nature medicine
Check for updates

OPEN

## Vaccine effectiveness of heterologous CoronaVac plus BNT162b2 in Brazil

Thiago Cerqueira-Silva[1,2,11], Srinivasa Vittal Katikireddi[3,4,11], Vinicius de Araujo Oliveira[2,5], Renzo Flores-Ortiz[5], Juracy Bertoldo Júnior[2,5], Enny S. Paixão[6], Chris Robertson[4,7], Gerson O. Penna[8], Guilherme L. Werneck[9], Maurício L. Barreto[2,5], Neil Pearce[6], Aziz Sheikh[10,12], Manoel Barral-Netto[1,2,5,12] and Viviane S. Boaventura[1,2,12] ✉

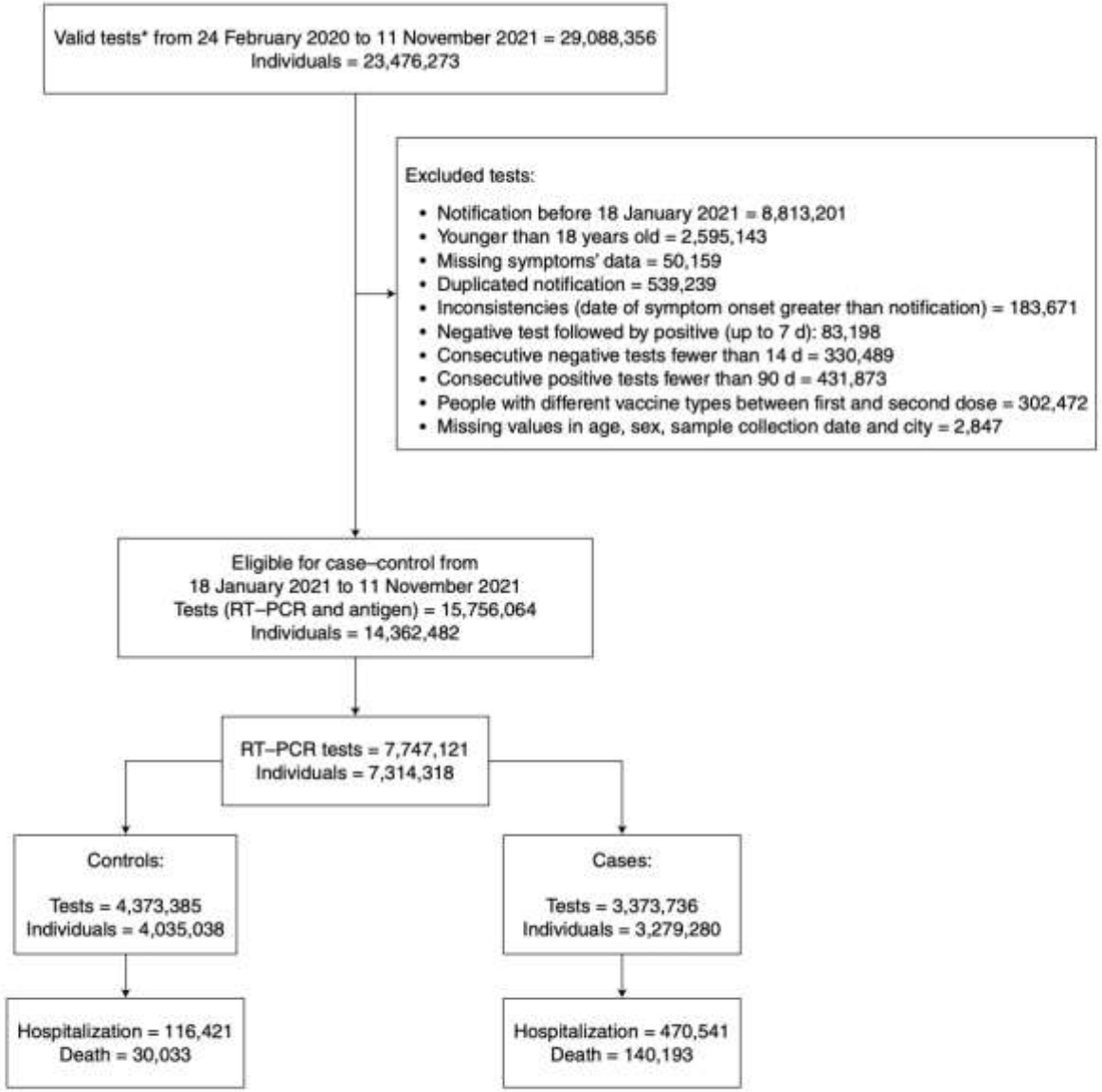



**Fig. 1 | Flowchart of the study population from surveillance databases and selection of cases and controls.** *Antigen or RT-PCR—sample collected 10 or fewer days after symptom onset.

https://www.nature.com/articles/s41591-022-01701-w.pdf

UnB

cidacs
Centro de Integração de Dados e Conhecimentos para Saúde
FIOCRUZ

### Coorte de 100 Milhões de Brasileiros

O principal objetivo da plataforma é criar uma coorte inédita e inovadora que viabilize o estudo das determinantes sociais e dos efeitos de políticas e programas sociais sobre os diferentes aspectos da saúde na sociedade brasileira.

Mais detalhes

### Equidade e Sustentabilidade

A Plataforma para Estudos e Pesquisas sobre Equidade e Sustentabilidade Urbana e seus Efeitos na Saúde desenvolve estudos e pesquisas relacionadas à equidade e a sustentabilidade de áreas urbanas e seus efeitos na saúde.

Mais detalhes

### Tecnologias e Inovações para o SUS

A "Plataforma de Incorporação de Tecnologias e Inovações em sistemas de Informação para Apoiar os Programas e Ações do SUS" está centrada na construção, avaliação e oferta de soluções em Saúde Digital para os Programas e ações do Sistema Único de Saúde.

Mais detalhes

### Epidemiologia Genômica

O Epigen-Brasil é a maior iniciativa latino-americana em genômica populacional e epidemiologia genética e o núcleo de pesquisas situado na Bahia atualmente se constitui em uma plataforma do Cidacs.

Mais detalhes

### Plataforma Zika

Visa o aprimoramento do conhecimento científico sobre a doença e apoio na adoção de medidas de saúde pública mais adequadas para o enfrentamento da tríplice epidemia ocasionada pelos vírus Zika, Dengue e Chikungunya.

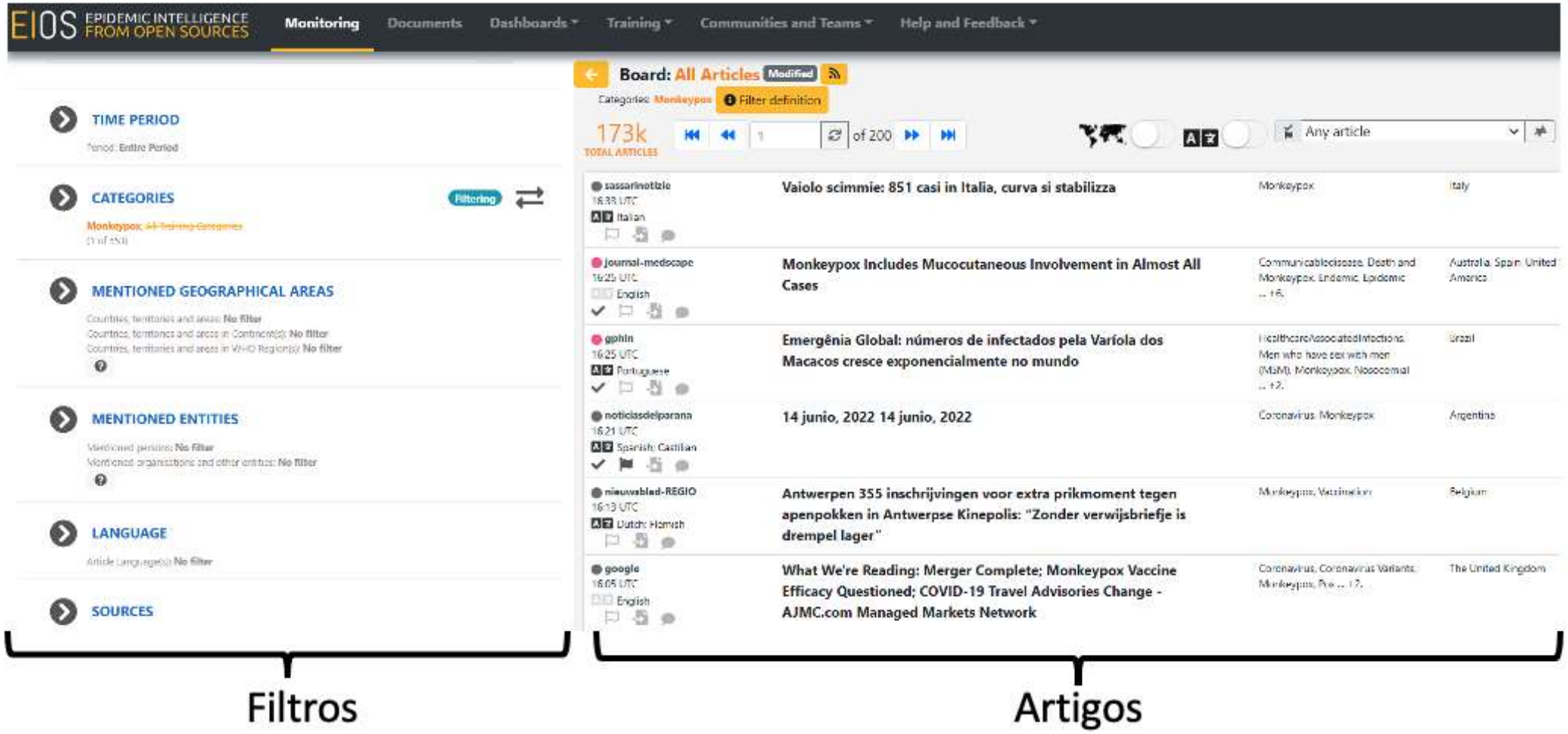
Mais detalhes

### Bioinformática

A "Plataforma Bioinformática de Alto Rendimento de Dados Biológicos" busca viabilizar a execução de projetos multidisciplinares envolvendo dados biológicos e populacionais provenientes de novas tecnologias.

Mais detalhes

# Detecção digital de doenças



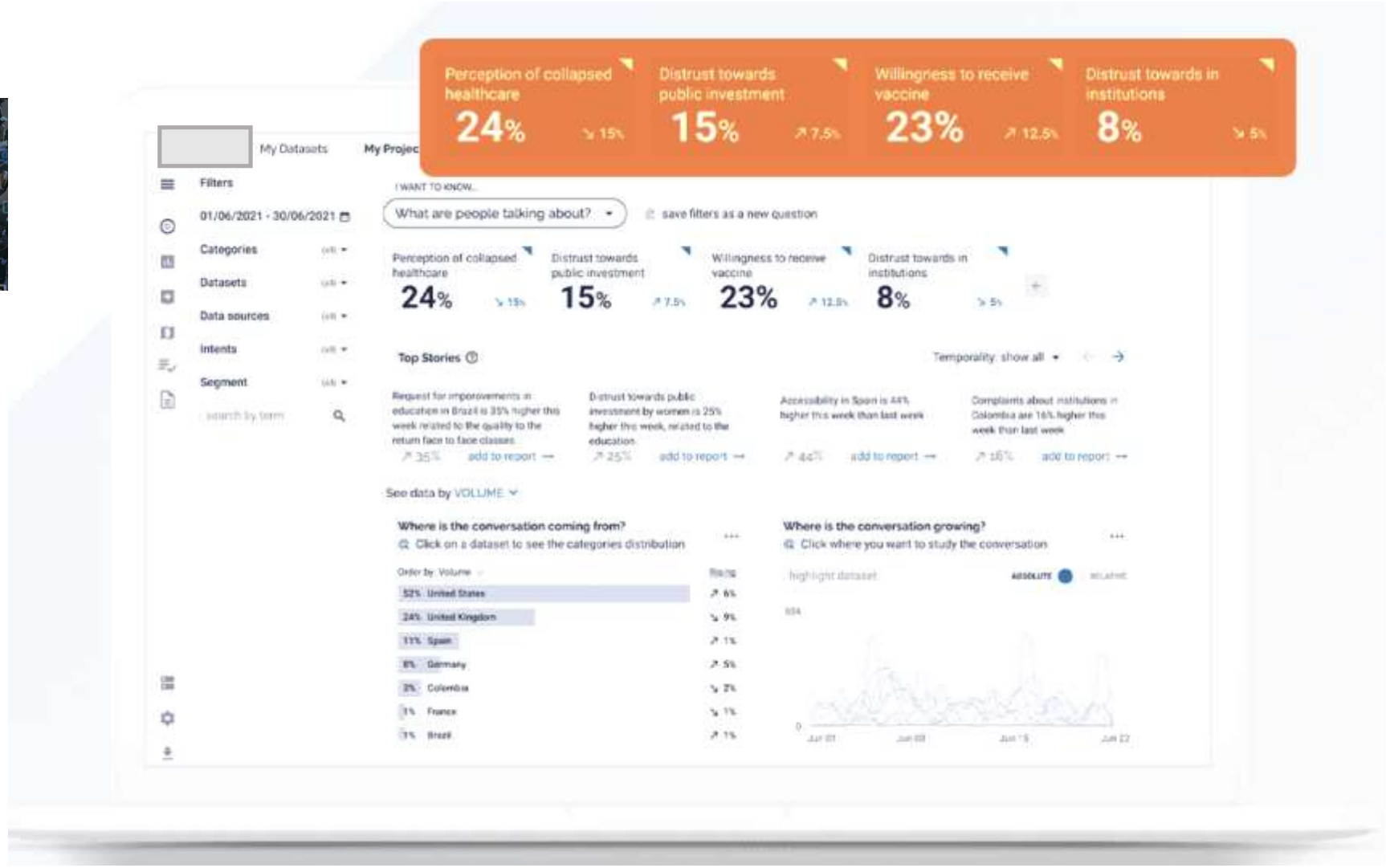

https://www.who.int/initiatives/eios

# Monitoramento de Comportamento

- Envolve análise de grandes volumes de dados para comportamentos:

# Monitoramento de Comportamento

- Envolve análise de grandes volumes de dados para comportamentos:



Even though the importance of social media and Internet-based data to epidemiological surveillance is clear, health agencies have been reluctant to incorporate these data sources into their systems because many technical issues have not yet been addressed. The technologies used in event-based systems must be adapted to the individual perceptions of and interactions with their own epidemiological data and to social media and other data from the Internet. Future work in this field will have wide-reaching implications for investments in systems for early warnings of and responses to health threats across the globe and for optimal public health surveillance in the 21st century.

Velasco et al (2014)

The Milbank Quarterly, Vol. 92, No. 1, 2014 (pp. 7-33)

Embora a importância das mídias sociais e dos dados baseados na Internet para a vigilância epidemiológica seja clara, as agências de saúde têm sido relutantes em incorporar essas fontes de dados em seus sistemas, porque muitas questões técnicas ainda não foram abordadas. As tecnologias utilizadas nos sistemas baseados em eventos devem ser adaptadas às percepções individuais e interações com seus próprios dados epidemiológicos e à mídia social e outros dados da Internet. O trabalho futuro neste campo terá implicações abrangentes para os investimentos em sistemas para alertas antecipados e respostas às ameaças à saúde em todo o mundo e para uma vigilância de saúde pública ótima no século XXI.

Velasco et al. (2014)

The Milbank Quarterly, Vol. 92, No. 1, 2014 (pp. 7-33)

Obrigado !