

Welcome to the Sentinel Innovation and Methods Seminar Series

The webinar will begin momentarily

- Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.
- Note: closed-captioning for today's webinar will be available on the recording posted at the link above.

Machine Learning in Distributed Data Networks like the FDA Sentinel System: Opportunities, Challenges, and Considerations

Jenna Wong, PhD
jenna_wong@harvardpilgrim.org

Sentinel Innovation and Methods Seminar Series
December 16, 2022

DEPARTMENT OF POPULATION MEDICINE



HARVARD
MEDICAL SCHOOL



Harvard Pilgrim
Health Care Institute





Volume 45, issue 5, May 2022

Role of Artificial Intelligence and Machine Learning in Pharmacovigilance

Issue editors

Andrew Bate & Yuan Luo

16 articles in this issue


Drug Safety (2022) 45:493–510

<https://doi.org/10.1007/s40264-022-01158-3>

REVIEW ARTICLE



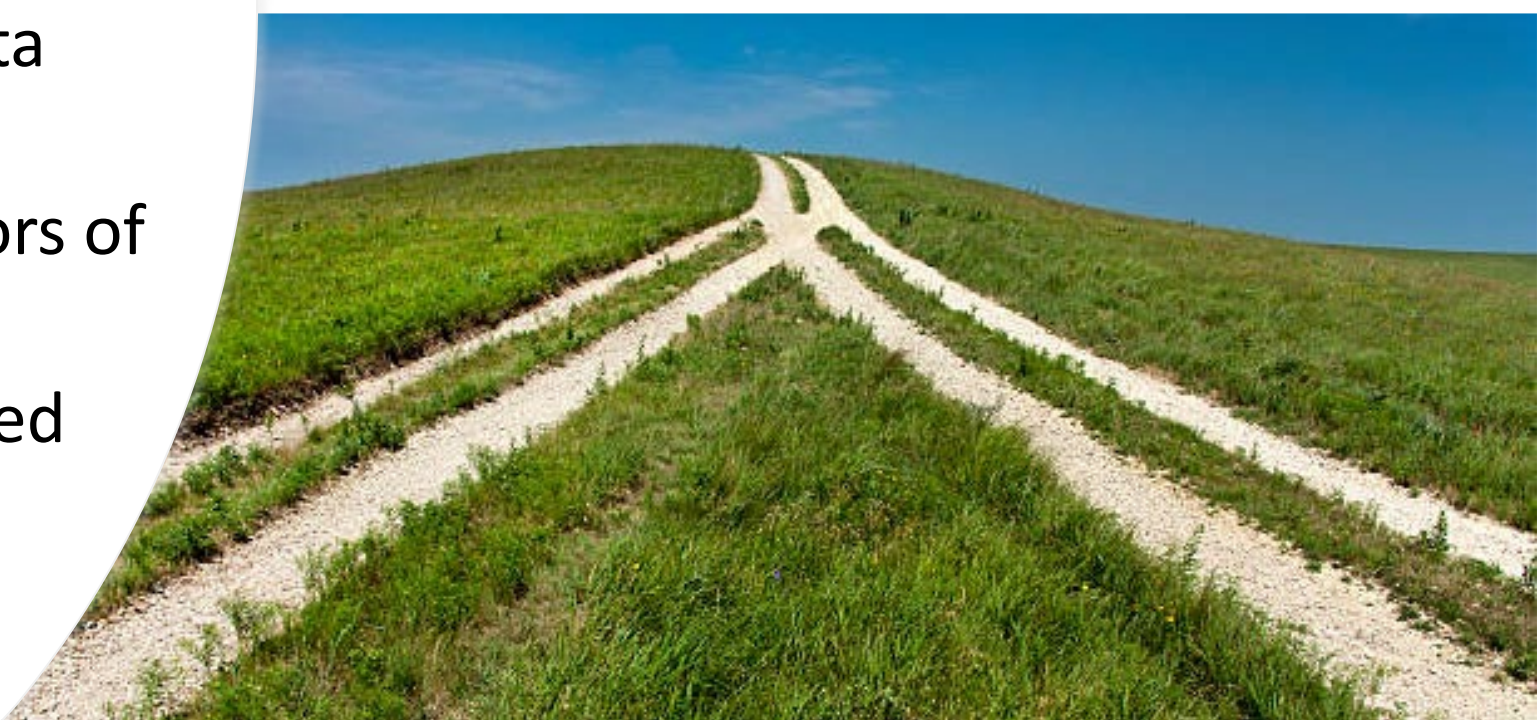
Applying Machine Learning in Distributed Data Networks for Pharmacoepidemiologic and Pharmacovigilance Studies: Opportunities, Challenges, and Considerations

Jenna Wong¹ · Daniel Prieto-Alhambra^{2,3} · Peter R. Rijnbeek³ · Rishi J. Desai⁴ · Jenna M. Reps⁵ · Sengwee Toh¹ 



Overview

1. Introduction
2. Machine learning and key activities of distributed data networks
3. Practical data-related factors of distributed data networks
4. Four scenarios of distributed data networks
5. Additional considerations



Original Investigation

July 6, 2022

Machine Learning–Based Models Incorporating Social Determinants of Health vs Traditional Models for Predicting In-Hospital Mortality in Patients With Heart Failure

Matthew W. Segar, MD, MS¹; Jennifer L. Hall, PhD²; Pardeep S. Jhund, MBChB, MSc, PhD³; [et al](#)

[Author Affiliations](#) | [Article Information](#)

JAMA Cardiol. 2022;7(8):844-854. doi:10.1001/jamacardio.2022.1900

Original Investigation | Public Health

July 11, 2022

Machine Learning Analysis of Handgun Transactions to Predict Firearm Suicide Risk

Hannah S. Laqueur, PhD, MA, MPA^{1,2}; Colette Smimiotis, PhD, MS^{1,2}; Christopher McCort, MS^{1,2}; [et al](#)

[Author Affiliations](#) | [Article Information](#)

JAMA Netw Open. 2022;5(7):e2221041. doi:10.1001/jamanetworkopen.2022.21041

Original Investigation | Pediatrics

October 27, 2022

Newborn Cry Acoustics in the Assessment of Neonatal Opioid Withdrawal Syndrome Using Machine Learning

Andrew W. Manigault, PhD¹; Stephen J. Sheinkopf, PhD²; Harvey F. Silverman, PhD³; [et al](#)

[Author Affiliations](#) | [Article Information](#)

JAMA Netw Open. 2022;5(10):e2238783. doi:10.1001/jamanetworkopen.2022.38783

Article | [Open Access](#) | [Published: 21 November 2022](#)

Machine learning models for prediction of HF and CKD development in early-stage type 2 diabetes patients

Eiichiro Kanda, [Atsushi Suzuki](#), [Masaki Makino](#), [Hiroo Tsubota](#), [Satomi Kanemata](#), [Koichi Shirakawa](#) & [Toshitaka Yajima](#) [✉](#)

Scientific Reports 12, Article number: 20012 (2022) | [Cite this article](#)

813 Accesses | 5 Altmetric | [Metrics](#)

Article | [Open Access](#) | [Published: 04 May 2022](#)

Machine learning-based prediction of relapse in rheumatoid arthritis patients using data on ultrasound examination and blood test

[Hidemasa Matsuo](#) [✉](#), [Mayumi Kamada](#), [Akari Imamura](#), [Madoka Shimizu](#), [Maiko Inagaki](#), [Yuko Tsuji](#), [Motomu Hashimoto](#), [Masao Tanaka](#), [Hiromu Ito](#) & [Yasutomo Fujii](#)

Scientific Reports 12, Article number: 7224 (2022) | [Cite this article](#)

2978 Accesses | 1 Citations | 56 Altmetric | [Metrics](#)



Original Investigation

November 17, 2022

Association of Machine Learning–Based Assessment of Tumor-Infiltrating Lymphocytes on Standard Histologic Images With Outcomes of Immunotherapy in Patients With NSCLC

Mehrdad Rakaee, PhD^{1,2,3}; Elio Adib, MD^{1,4}; Biagio Ricciuti, MD⁵; [et al](#)

[Author Affiliations](#) | [Article Information](#)

JAMA Oncol. Published online November 17, 2022. doi:10.1001/jamaoncol.2022.4933

Viewpoint

April 3, 2018

Big Data and Machine Learning in Health Care

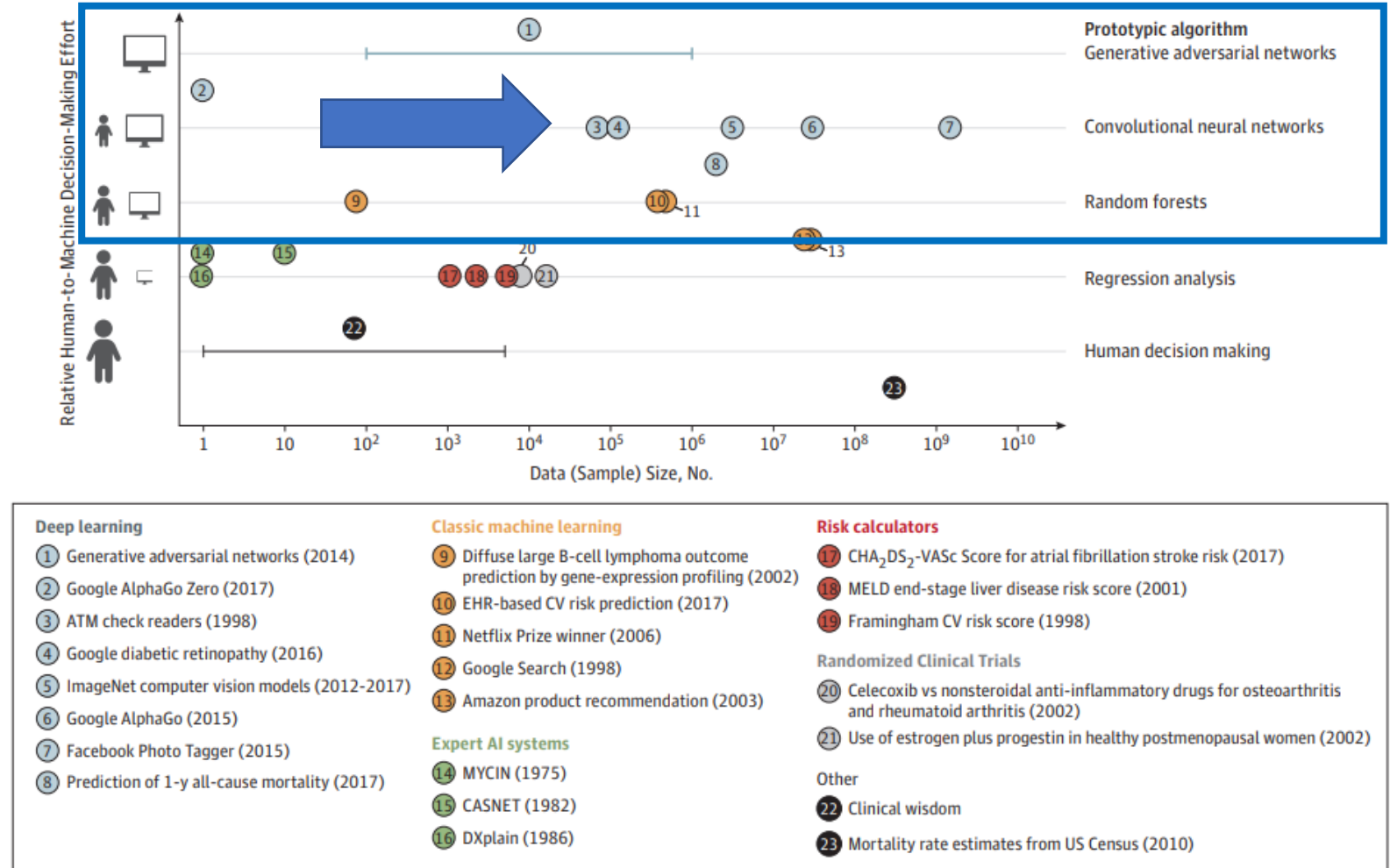
Andrew L. Beam, PhD¹; Isaac S. Kohane, MD, PhD¹

□ Author Affiliations | Article Information

JAMA. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391

- Algorithms exist along a **continuum** between fully human-guided versus fully machine-guided data analysis.
- The degree to which an algorithm can be considered an instance of machine learning depends on how much of the algorithm's structure or parameters are predetermined by humans.

Figure. The Axes of Machine Learning and Big Data

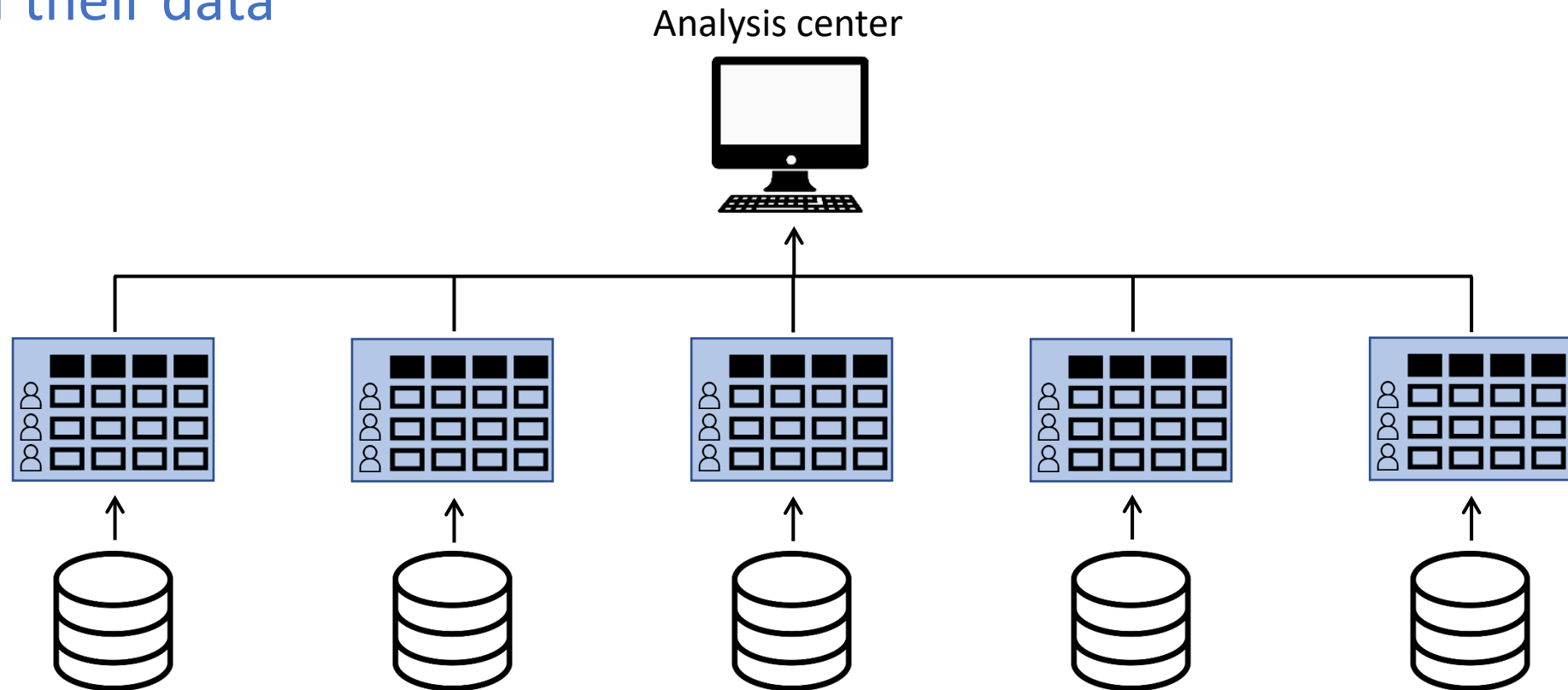


Multi-database studies

- Access to larger and more diverse study populations
 - More precise and generalizable findings
 - Greater capture of rare exposures and outcomes
 - Better suited to investigate heterogenous treatment effects
 - More data for machine learning algorithms

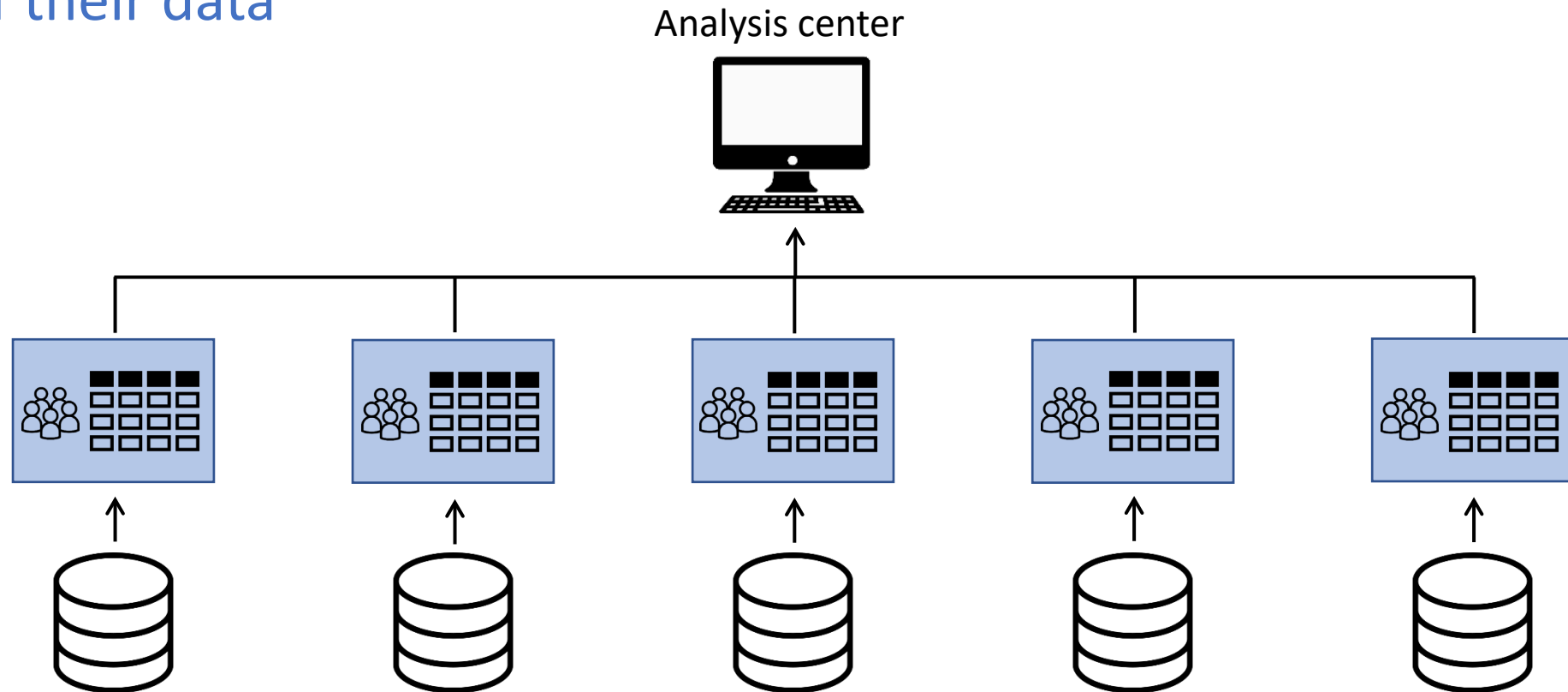
Distributed data networks (DDNs)

- Network of data partners whose databases are **not pooled centrally**, and data partners maintain **full control over the physical storage and use of their data**



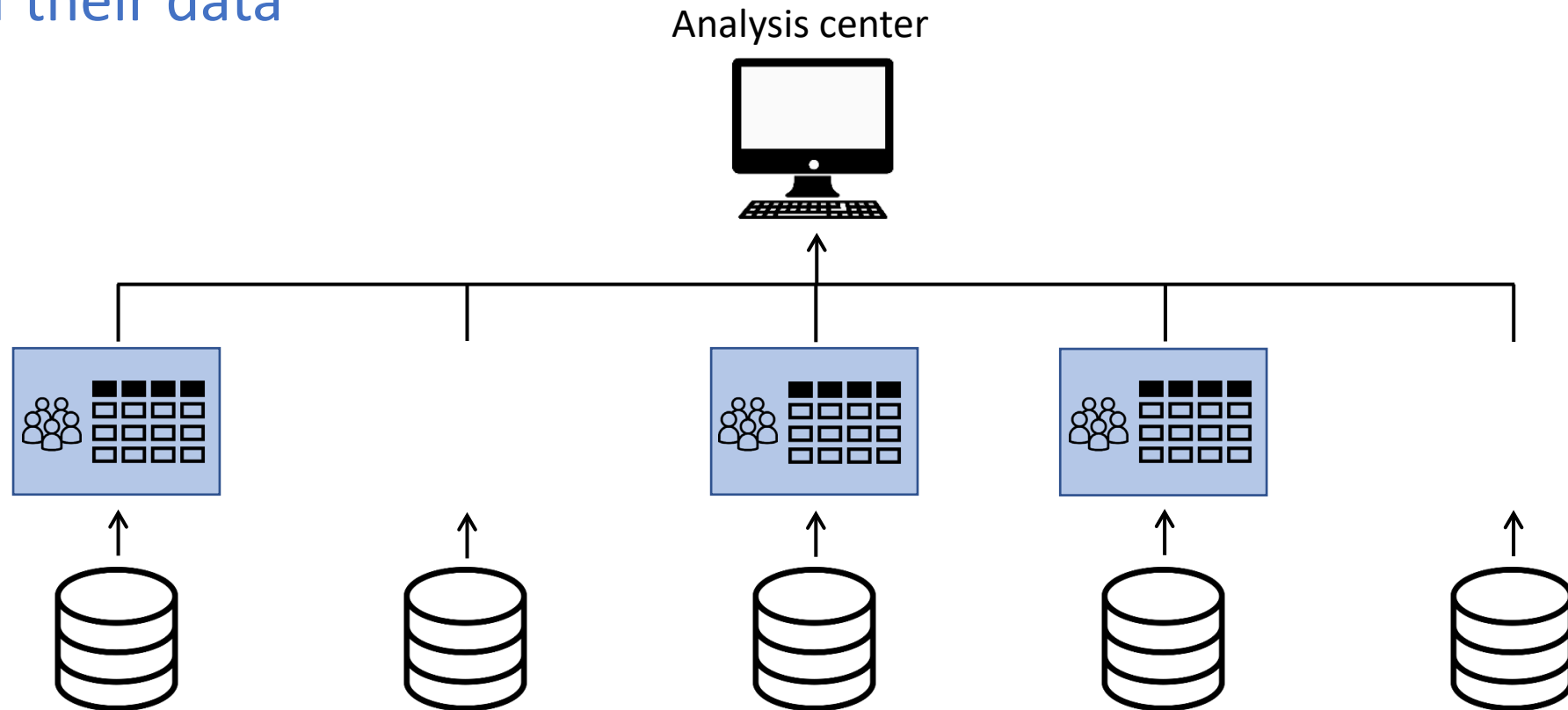
Distributed data networks (DDNs)

- Network of data partners whose databases are **not pooled centrally**, and data partners maintain **full control over the physical storage and use of their data**



Distributed data networks (DDNs)

- Network of data partners whose databases are **not pooled centrally**, and data partners maintain **full control over the physical storage and use of their data**



DDNs in pharmacoepidemiology



Network name

Asian Pharmacoepidemiology Network (AsPEN)

Canadian Network for Observational Drug Effect Studies (CNODES)

European Health Data & Evidence Network (EHDEN)

Health Care Systems Research Network (HCSRN)

National Patient-Centered Clinical Research Network (PCORnet)

Observational Health Data Sciences and Informatics (OHDSI) collaborative

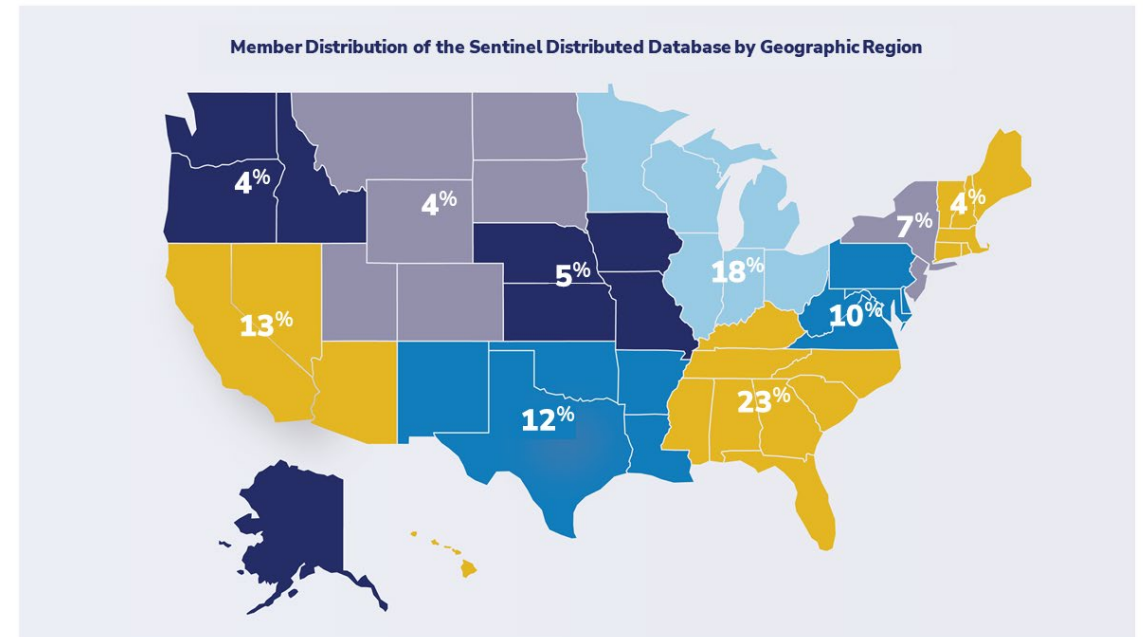
Sentinel System

Vaccine Safety Data Link



The FDA Sentinel System, 2000-2022

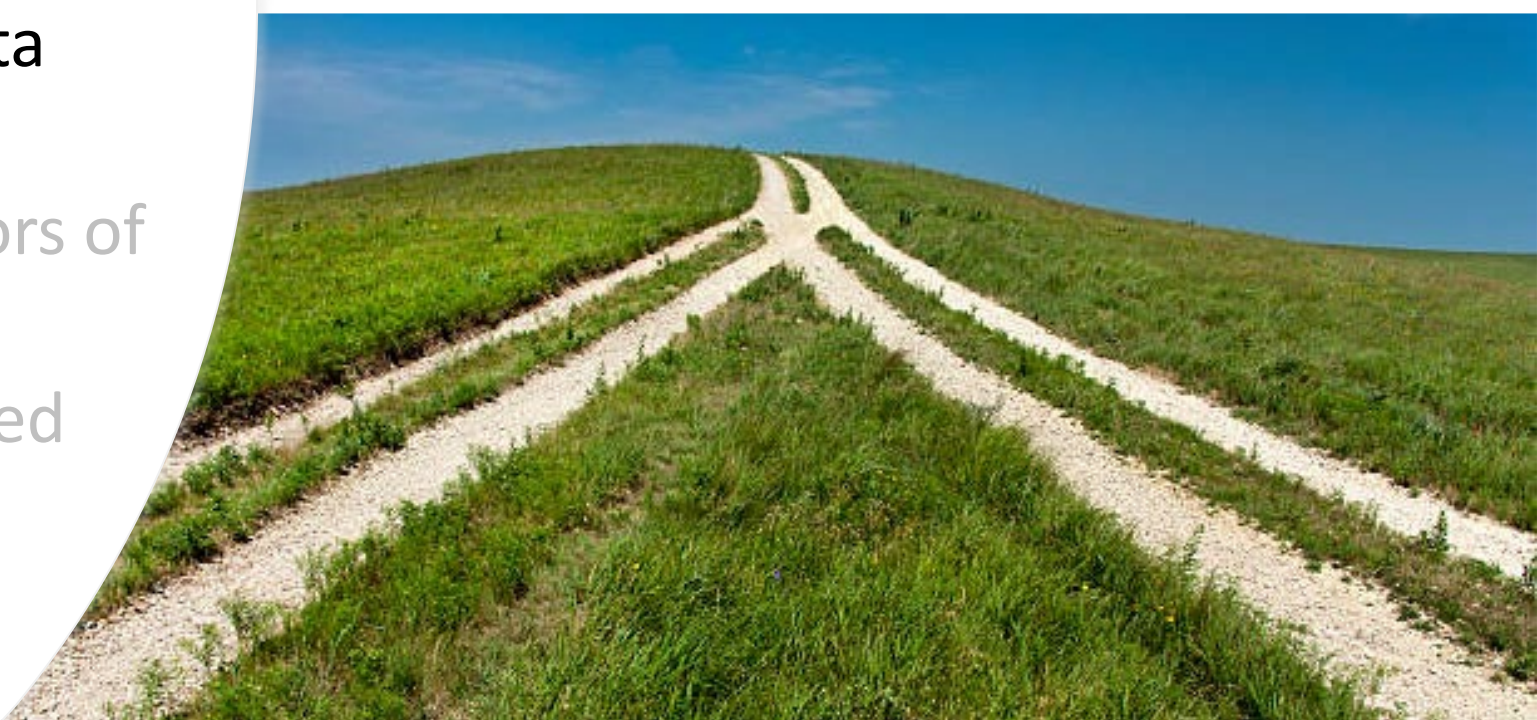
- Number of data-contributing sites: **13**
- Number of individuals currently accruing new data: **64 million**
- Total person-years of data: **874 million**
- Unique medical encounters: **16 billion**
- Pharmacy dispensings: **17 billion**
- Types of electronic health data
 - Administrative data
 - Registry data
 - Inpatient data
 - Clinical data
 - Patient-reported measures



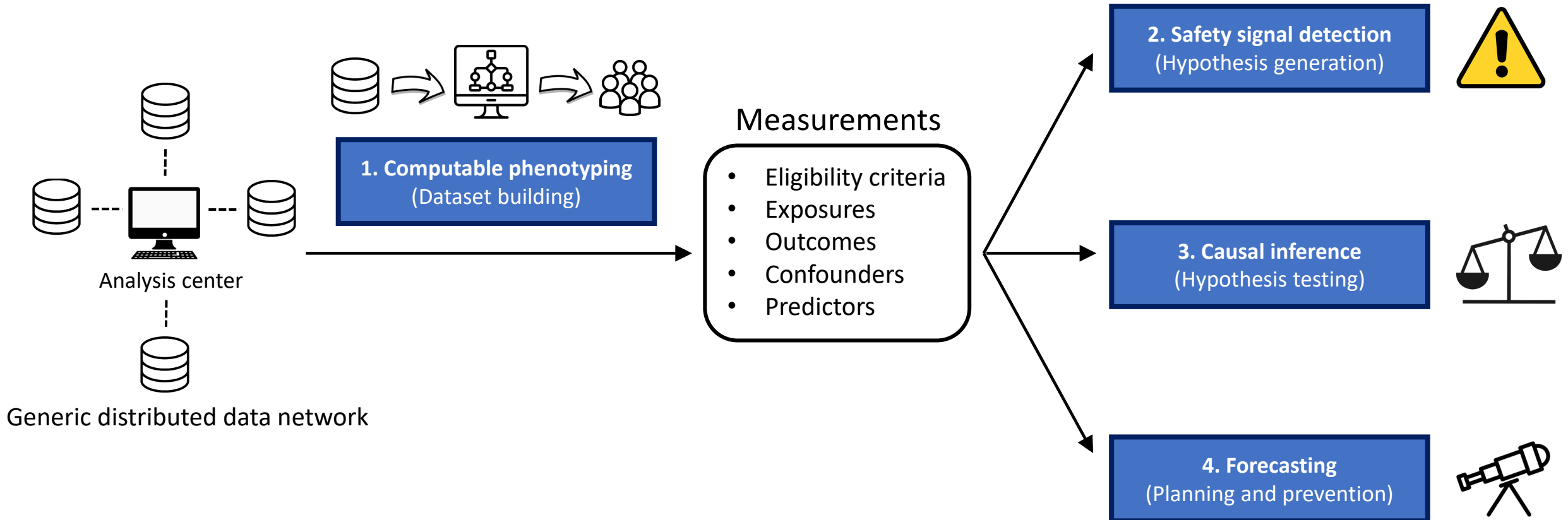
Source: <https://www.sentinelinitiative.org/about/key-database-statistics#member-distribution-of-the-sentinel-distributed-database-by-geographical-region>

Overview

1. Introduction
2. Machine learning and key activities of distributed data networks
3. Practical data-related factors of distributed data networks
4. Four scenarios of distributed data networks
5. Additional considerations



Key activities of distributed data networks



How can the use of machine learning
enhance these activities?

1. Computable phenotyping

- Phenotyping definition
 - Determine mapping from inputs (e.g., biological, behavioral, or clinical features) to phenotype status using [machine-guided process](#)
- Information extraction
 - Extract potentially relevant phenotypic information from unstructured data (e.g., text, images, etc.) via an [automated process](#)

2. Safety signal detection

- Disproportionality measures
 - Information Component estimated using Bayesian Confidence Propagation Neural Networks
- Reduce potential confounding
 - Estimate general propensity scores (e.g., propensity score-matched tree-based scan statistic)
- Other innovative approaches with longitudinal observational data
 - E.g., random forest classifier trained to signal adverse drug reactions from features derived from various cohort designs addressing Bradford Hill's causality considerations
- Information extraction
 - Extract adverse events and drug-adverse event pairs from unstructured text via an automated process

3. Causal inference

- Automate the high-dimensional confounding adjustment process
 - Properly specified “nuisance functions”
 - Reduce dimensionality of candidate covariates
 - Prioritize candidate covariates
 - Obtain correct functional form of covariates
 - Simultaneously consider multiple covariate sets
 - Information extraction
 - Extract candidate covariates for nuisance functions from unstructured data

Weberpals J et al. Deep Learning-based Propensity Scores for Confounding Control in Comparative Effectiveness Research: A Large-scale, Real-world Data Study. *Epidemiology*. 2021 May 1;32(3):378-388.

Wyss R et al. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*. 2018;29:96–106.

Ju C et al. Propensity score prediction for electronic healthcare databases using Super Learner and High-dimensional Propensity Score Methods. *J Appl Stat*. 2019;46(12):2216-2236.

Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology*. 2021;32:393–401.

Wyss R et al. Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database studies: An overview of the current literature. *Pharmacoepidemiol Drug Saf*. 2022 Sep;31(9):932-943.

4. Forecasting

Diagnostic modeling

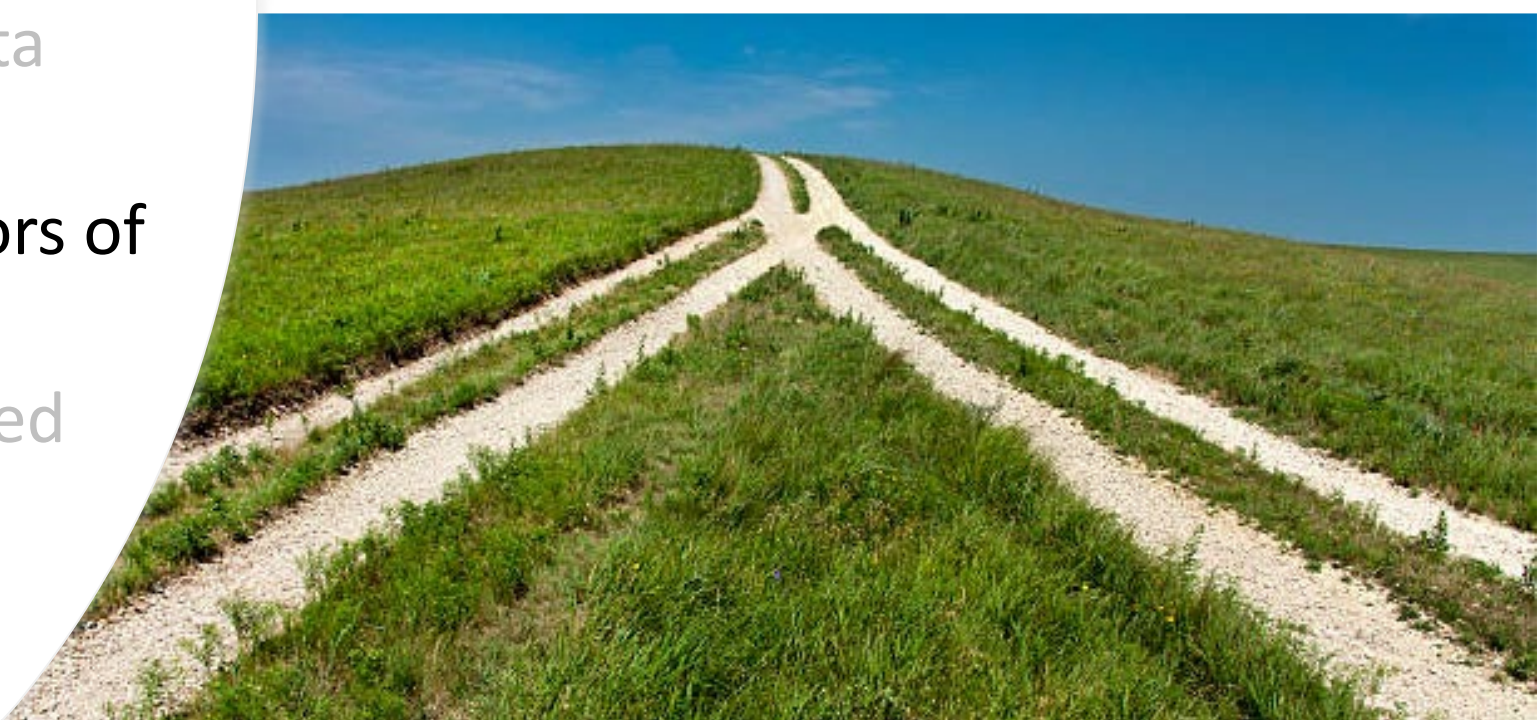
- **Phenotyping** algorithm
 - Use inputs (e.g., biological, behavioral, or clinical features) to **determine phenotype status** using machine-guided process
- Information extraction
 - Extract potentially relevant **phenotypic information** from unstructured data (e.g., text, images, etc.) via an automated process

Prognostic modeling

- **Prognostic** algorithm
 - Use inputs (e.g., biological, behavioral, or clinical features) to **predict future health events** using machine-guided process
- Information extraction
 - Extract potentially relevant **prognostic information** from unstructured data (e.g., text, images, etc.) via an automated process

Overview

1. Introduction
2. Machine learning and key activities of distributed data networks
- 3. Practical data-related factors of distributed data networks**
4. Four scenarios of distributed data networks
5. Additional considerations



Data-related factors

Spectrum of possibilities

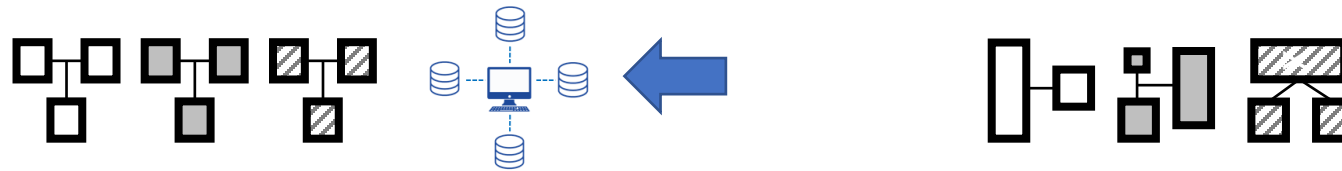
Modality of source data



Structured data

Unstructured data

Degree of data standardization



Common data model

No common data model

Granularity of shared data

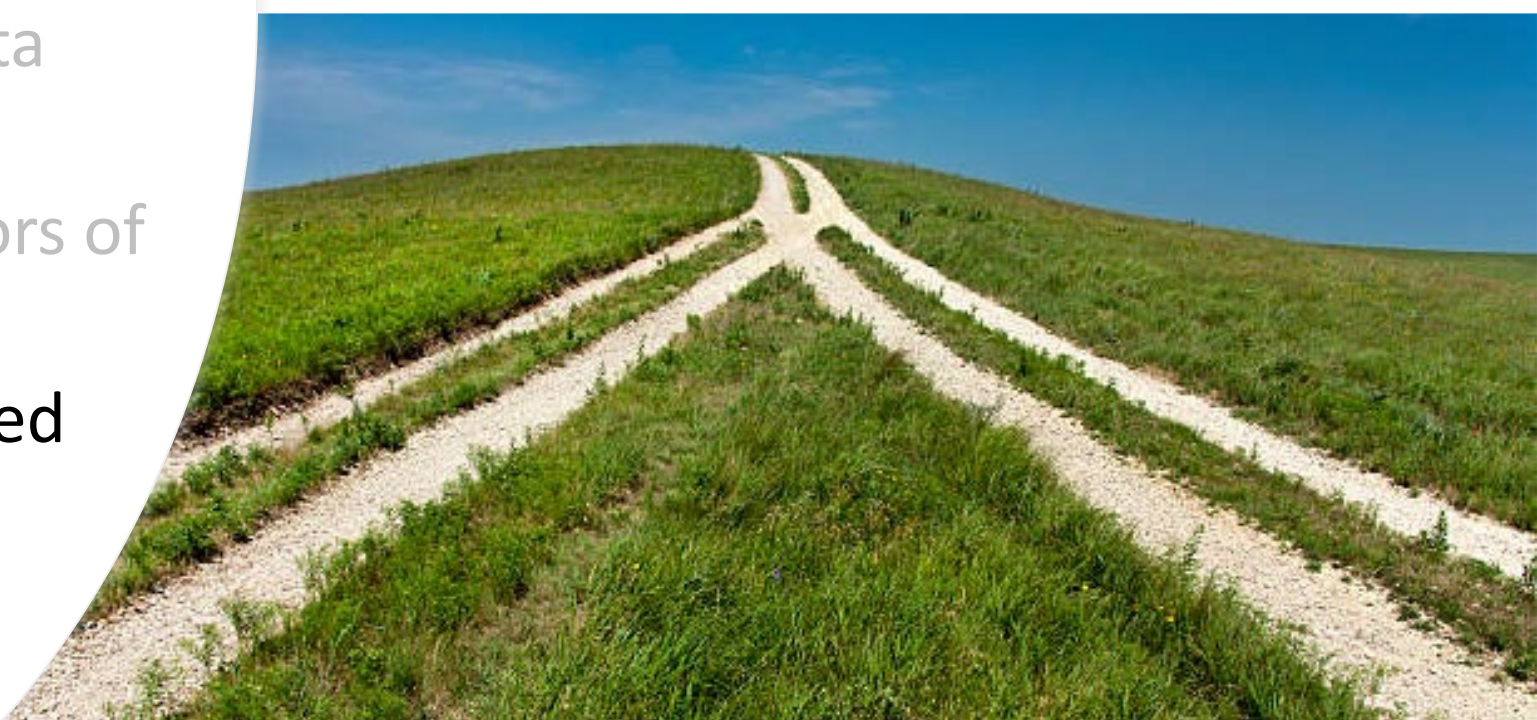


Individual-level

Summary-level

Overview

1. Introduction
2. Machine learning and key activities of distributed data networks
3. Practical data-related factors of distributed data networks
4. Four scenarios of distributed data networks
5. Additional considerations



Modality of source data

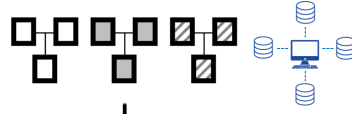


Structured data

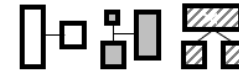


Unstructured data

Degree of data standardization



Common data model



No common data model

Granularity of shared data



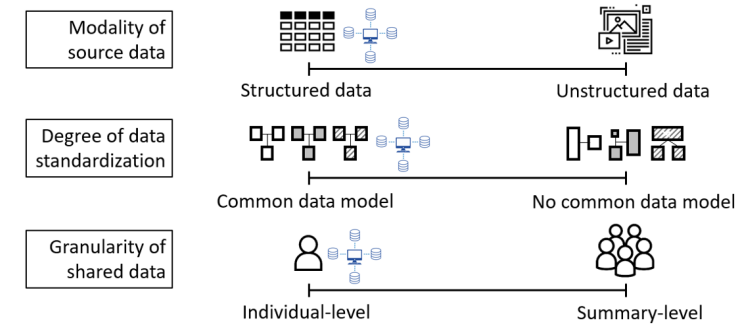
Individual-level



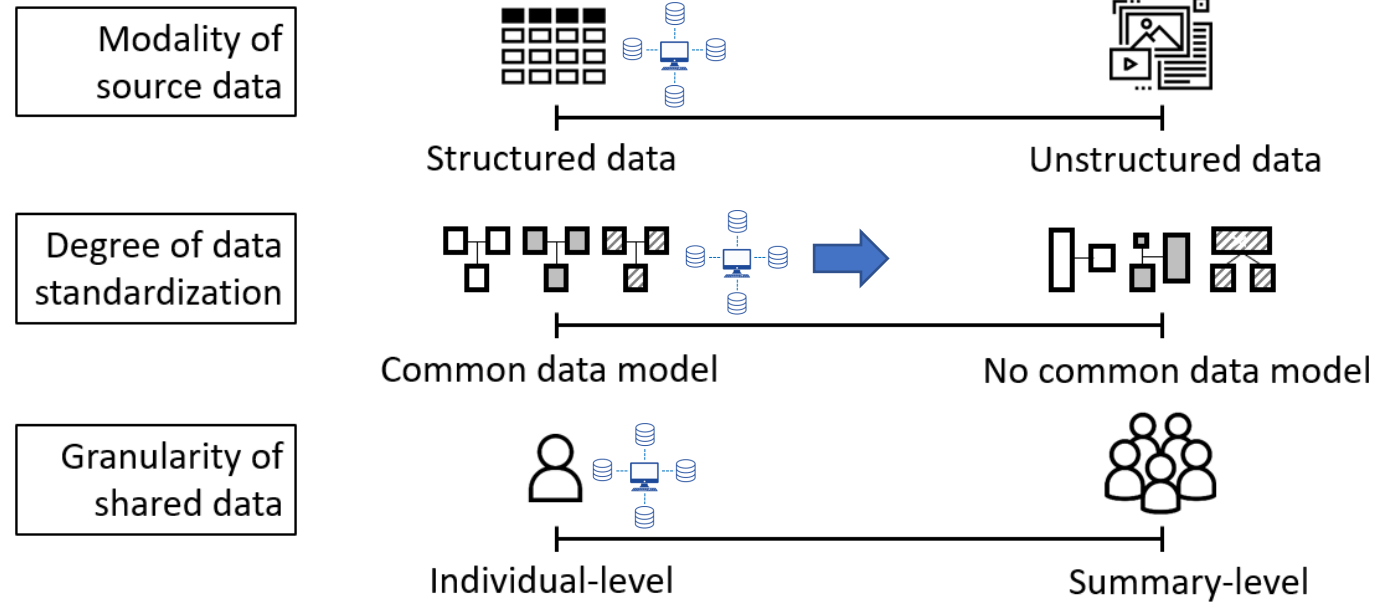
Summary-level

Scenario	Modality of source data	Degree of data standardization	Granularity of shared data
1 – Base case	Structured data only	Common data model for all inputs	Individual-level data for all sites
2 – Less standardized data available	Structured data only	No common data model for some inputs	Individual-level data for all sites
3 – More complex data modalities used	Structured and unstructured data	Common data model for all inputs	Individual-level data for all sites
4 – Less granular data shared	Structured data only	Common data model for all inputs	Summary-level data for all sites

Scenario 1. Base case



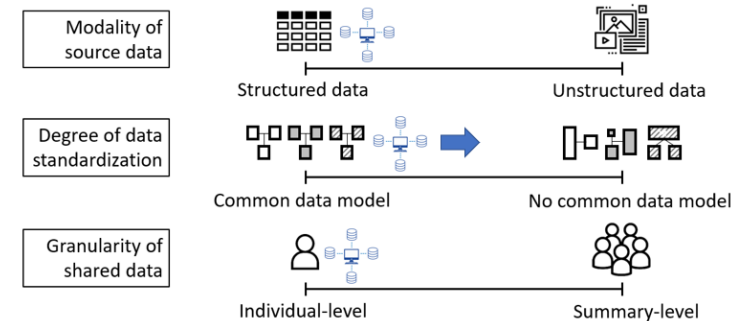
- Simplest and most straightforward setting for machine learning
 - Structured data only → facilitates creation of a common data model (CDM)
 - CDM-derived inputs only → facilitates curation of analytic dataset
 - Sharing of individual-level data → enables modeling to proceed with same flexibility as in single database setting
- Although it is technically possible to apply machine learning to a centralized dataset, *should* it be done?
 - What is the purpose of the machine learning model?
 - What is the extent of heterogeneity between data partners?



Scenario	Modality of source data	Degree of data standardization	Granularity of shared data
1 – Base case	Structured data only	Common data model for all inputs	Individual-level data for all sites
2 – Less standardized data available	Structured data only	No common data model for some inputs	Individual-level data for all sites
3 – More complex data modalities used	Structured and unstructured data	Common data model for all inputs	Individual-level data for all sites
4 – Less granular data shared	Structured data only	Common data model for all inputs	Summary-level data for all sites

Scenario 2. Less standardized data available

- Creates challenges for the feature engineering process
- Information of interest exists outside the CDM in the native (and thus unstandardized) structured data within data partners' source systems
- Relevant for data-adaptive machine learning models because often of interest to consider *more* features



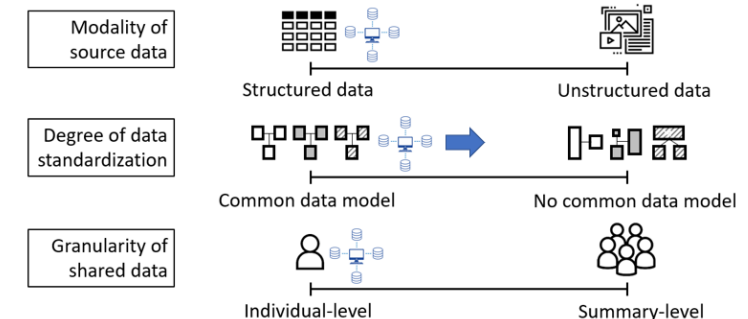
Scenario 2. Less standardized data available

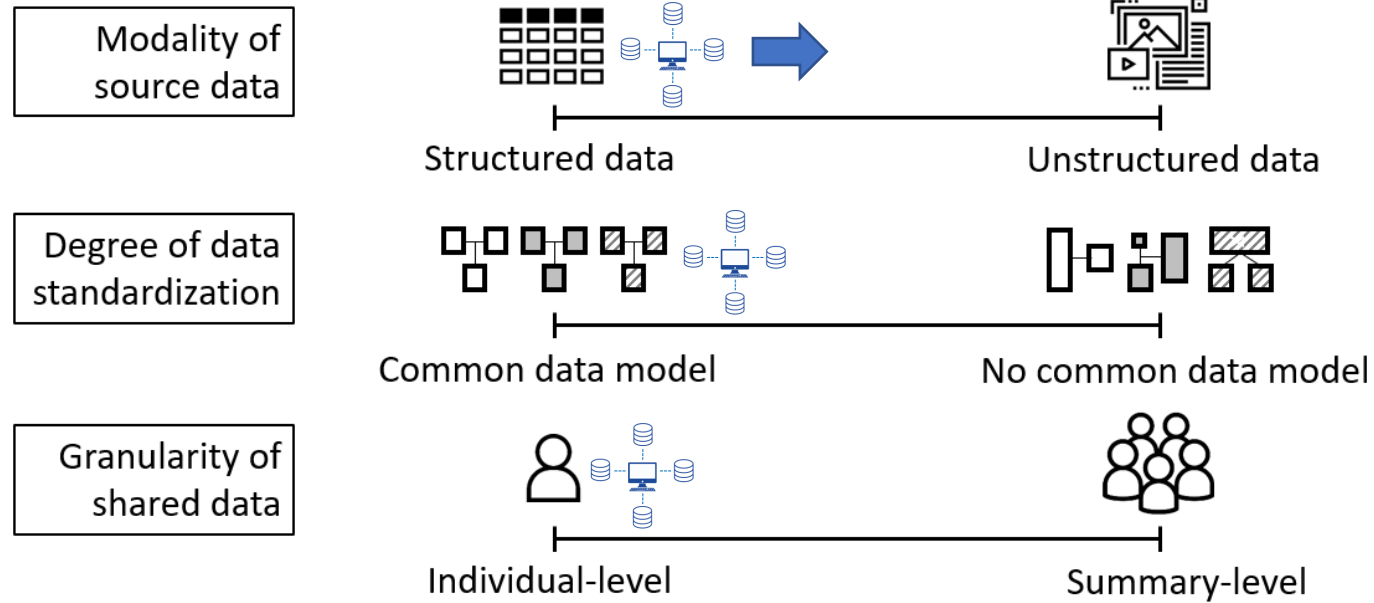
- **Approach 1: Standardize the unstandardized information**

- Resource-intensive, but may be warranted if information is **easily obtained**, will be **frequently used**, or is **urgently required**
 - E.g., Additions to latest Sentinel CDM (8.1.0)
 - Patient-Reported Measures Table
 - SARS-CoV-2 lab test results

- **Approach 2: Do a site-specific analysis**

- May be preferred when
 - Additional information is available in select sites only
 - Added value of unstandardized information is uncertain

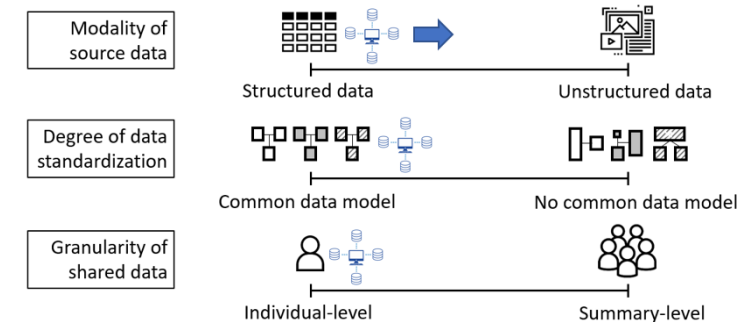




Scenario	Modality of source data	Degree of data standardization	Granularity of shared data
1 – Base case	Structured data only	Common data model for all inputs	Individual-level data for all sites
2 – Less standardized data available	Structured data only	No common data model for some inputs	Individual-level data for all sites
3 – More complex data modalities used	Structured and unstructured data	Common data model for all inputs	Individual-level data for all sites
4 – Less granular data shared	Structured data only	Common data model for all inputs	Summary-level data for all sites

Scenario 3: More complex data modalities

- Creates challenges for the feature engineering process
- Information of interest is unstructured data that exists outside the CDM in the data partners' source systems
 - Essentially an extension of challenges in Scenario 2
- Relevant because many opportunities for machine learning involve extraction and use of information from unstructured data



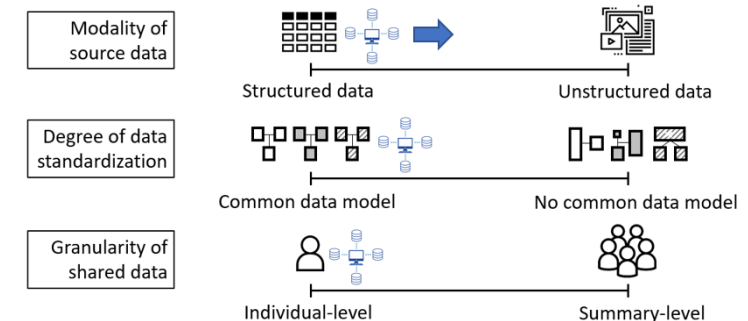
Scenario 3: More complex data modalities

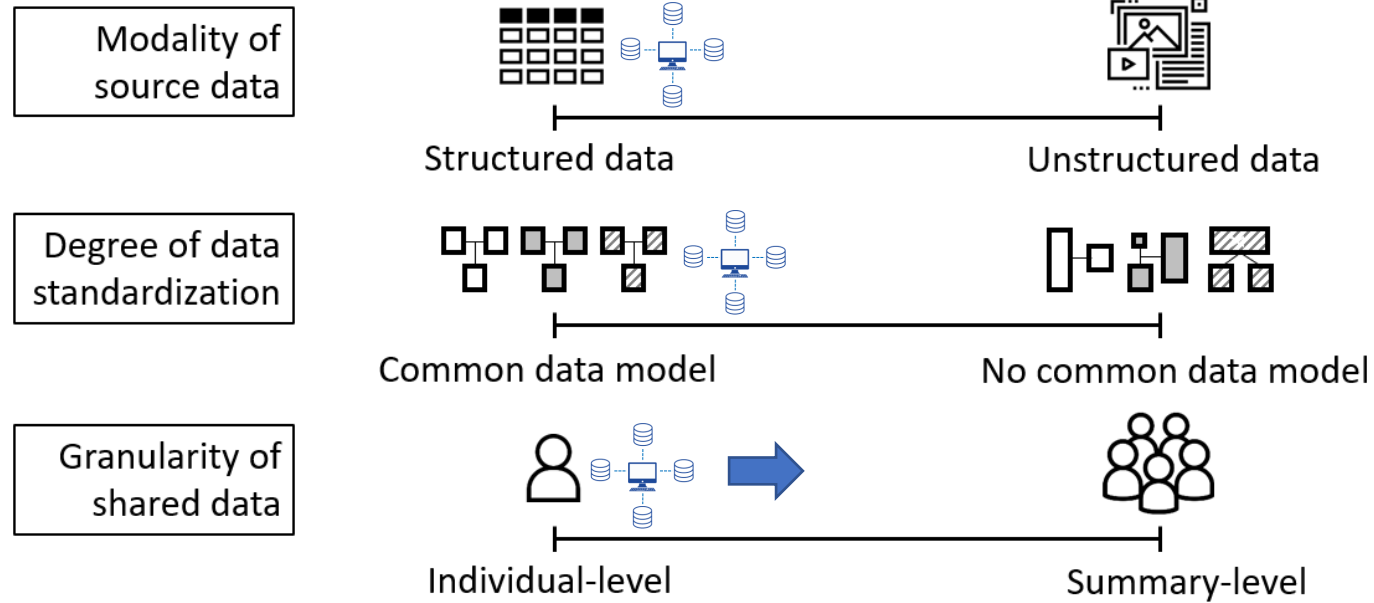
- **Approach 1: Do a site-specific analysis**

- All data processing and information extraction on the unstructured data done outside CDM according to a pre-defined protocol
 - Completed Sentinel Project: “Validation of Anaphylaxis using Machine Learning”

- **Approach 2: Incorporate the unstructured data into the CDM**

- Store raw text as a single field in the CDM
- Perform information extract upfront and encode output into the CDM
 - Ongoing Sentinel Project: “Representation of Unstructured Data Across Common Data Models”

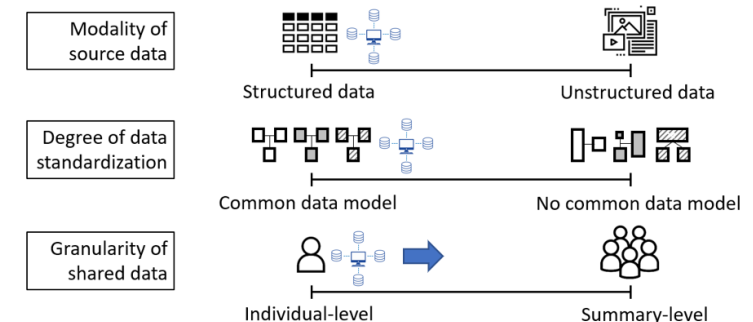




Scenario	Modality of source data	Degree of data standardization	Granularity of shared data
1 – Base case	Structured data only	Common data model for all inputs	Individual-level data for all sites
2 – Less standardized data available	Structured data only	No common data model for some inputs	Individual-level data for all sites
3 – More complex data modalities used	Structured and unstructured data	Common data model for all inputs	Individual-level data for all sites
4 – Less granular data shared	Structured data only	Common data model for all inputs	Summary-level data for all sites

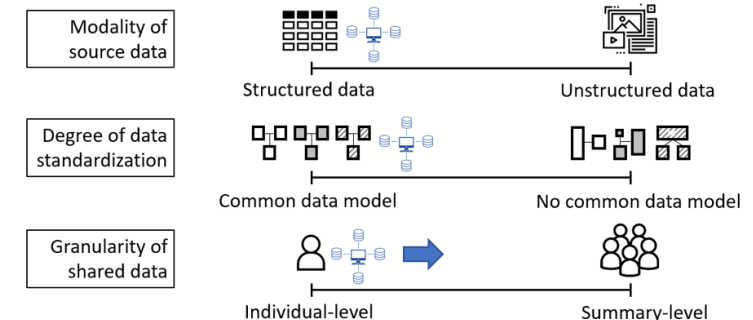
Scenario 4: Less granular data shared

- Create challenges for the machine learning model fitting process
- Possible analytic options are constrained by the inability of data partners to share individual-level data with the analysis center



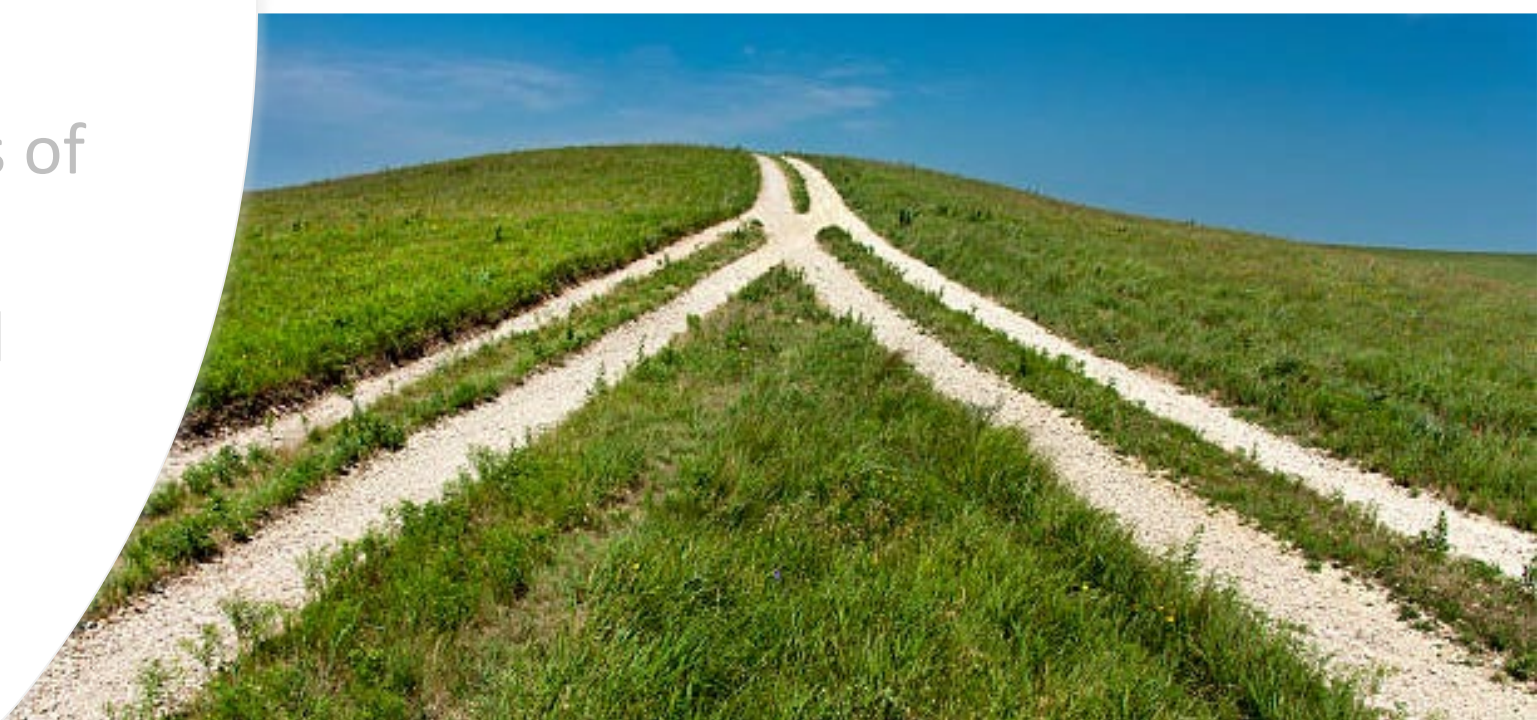
Scenario 4: Less granular data shared

- **Approach 1: Fit site-specific machine learning models**
 - Each site fits a custom model
 - Fit model in one site, apply model in other site(s)
 - Sentinel Methods Project: “Validation of Anaphylaxis Using Machine Learning”
- **Approach 2: Collaboratively learn a global model**
 - Has been successfully demonstrated for regression analyses
 - Emerging and actively developing area of research for more complex machine learning models

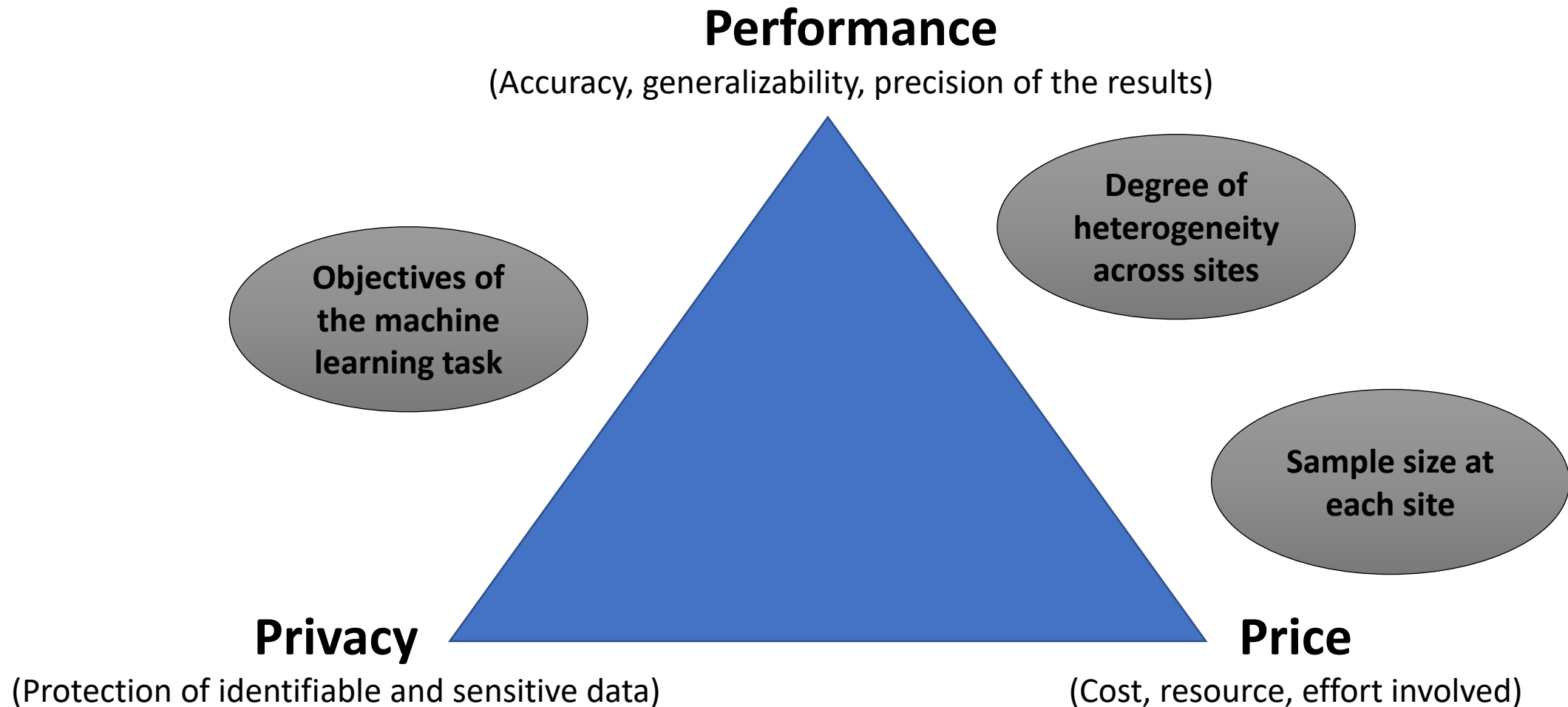


Overview

- Introduction
- Machine learning and key activities of distributed data networks
- Practical data-related factors of distributed data networks
- Four scenarios of distributed data networks
- **Additional considerations**



Choice of approach is a balancing act



Additional opportunities

Issues for machine learning	Single database settings	Distributed data network settings
Generalizability	External model validation is rare and slow	External model validation can be done quickly and easily
Transparency	Lower impetus to provide finer details	High transparency required to enable data partners to replicate process
Interpretability	Lower impetus to interpret and explain model outputs	Unusual or discrepant results across data partners may create need to interpret and explain model outputs

Conclusions


- There are many opportunities to use machine learning in distributed data networks
- Distributed data networks face unique challenges over and above those encountered in single-database settings
- Various approaches may be considered to address these challenges
- Utility of machine learning in distributed data networks will likely continue to increase in the coming years

Drug Safety (2022) 45:493–510
<https://doi.org/10.1007/s40264-022-01158-3>

REVIEW ARTICLE



Applying Machine Learning in Distributed Data Networks for Pharmacoepidemiologic and Pharmacovigilance Studies: Opportunities, Challenges, and Considerations

Jenna Wong¹ · Daniel Prieto-Alhambra^{2,3} · Peter R. Rijnbeek³ · Rishi J. Desai⁴ · Jenna M. Reps⁵ · Sengwee Toh¹ 

Contact:
jenna_wong@harvardpilgrim.org

DEPARTMENT OF POPULATION MEDICINE

