MACHINE LEARNING AND AI FOR HEALTHCARE

By Arjun Panesar

DATA

#1

“The plural of anecdote is data.” — Raymond Wolfinger

What is Data?

- Data itself can take many forms
  - *Character*
  - *Text*
  - *Words*
  - *Numbers*
  - *Pictures*
  - *Sound*
  - *…*

What is Data?

- Types
  - *Structured*
  - *Unstructured*
  - *Semi Structured*

- Values
  - *Qualitative*
  - *Quantitative*

What is Data?

- For “Data” to become “information”, it requires interpretation
- Information is organized or classified data, which has some meaningful value (or values) for the receiver.
- Information is the processed data on which decisions and actions should be based
- Data -> information -> insights -> making better decisions

DATA IN HEALTHCARE

#2

Revolution in Healthcare

- Growing cost pressures
- Digital Health
- Personalized care
- Evidence-based medicine

Data in Healthcare

- Patient and population wellness
- Patient education and engagement
- Prediction of disease and care risks
- Medication adherence
- Disease management
- Disease reversal/remission
- Individualization and personalization of treatment and care
- Financial, transactional, and environmental forecasting, planning, and accuracy

TYPES OF DATA

#3

Types of Data

- Types based on following a model or schema
  - *Structured*
    - Blood glucose monitor
  - *Unstructured*
    - Text messages, social media posts
  - *Semi Structured*
    - X-rays

Types of Data

Structured Data

Semi-Structured Data

Unstructured Data

Types of Data

- The Gartner report has indicated that data volume is set to grow 800% over the next 5 years, and 80% of this data will be in the form of unstructured data

Key Differences

- Volume
- Analyze and Interpretation

Definitions in ML domain

- Instance
  - *A single row of data or observation.*
- Feature
  - *A single column of data. It is a component of the observation.*
- Data type
  - *This refers to the kind of data represented by the feature (e.g., Boolean, string, number)*
- Dataset
  - *A collection of instances used to train and test machine learning models.*
- Training dataset
  - *Dataset used to train the machine learning model*
- Testing dataset
  - *Dataset used to determine accuracy/performance of the machine learning model.*

Dataset

	Feature 1 (data type)	Feature 2	Feature 3
Instance 1	Boolean	String	Number
Instance 2			
Instance 3			
Instance 4			
Instance 5			

BIG DATA

#4

Big Data

- Laney 2001
- Volume
- Variety
- Velocity

Big Data

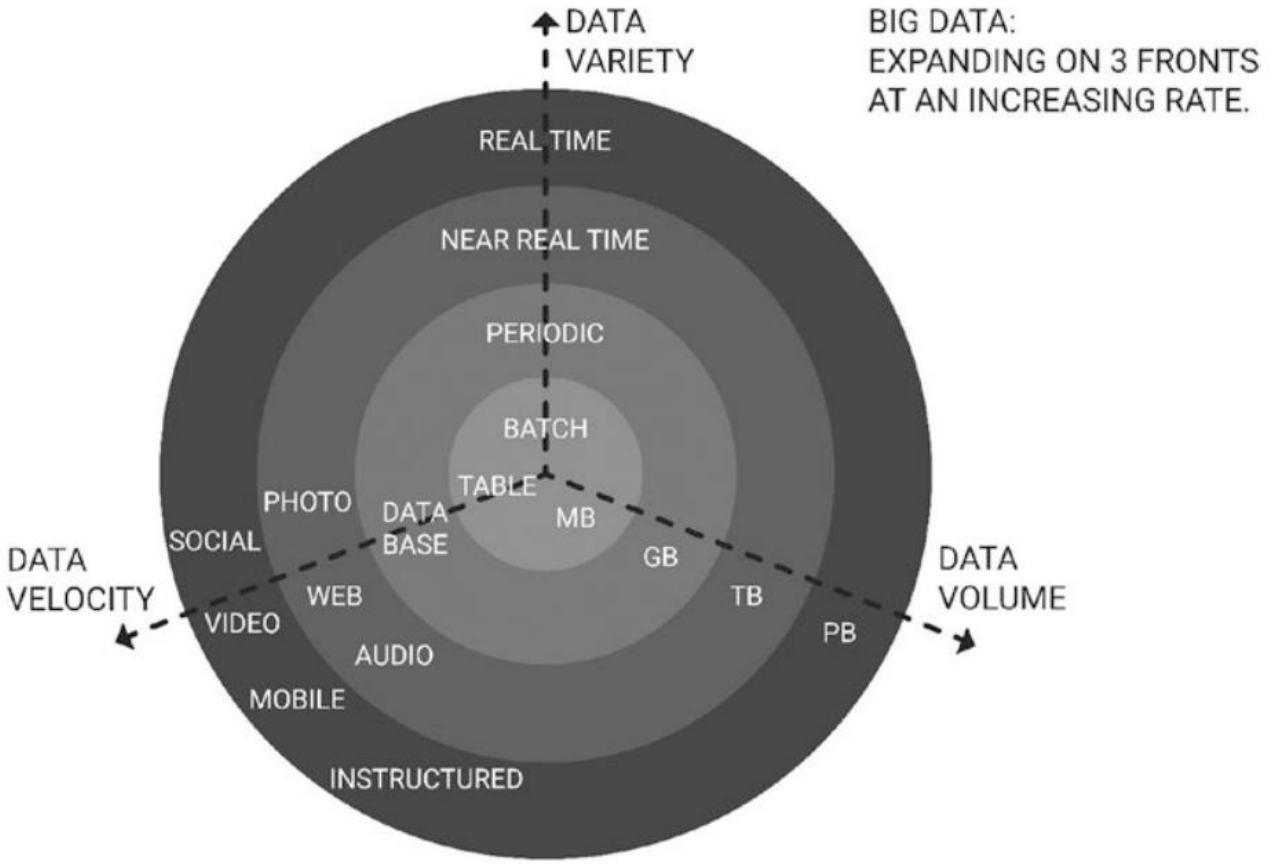
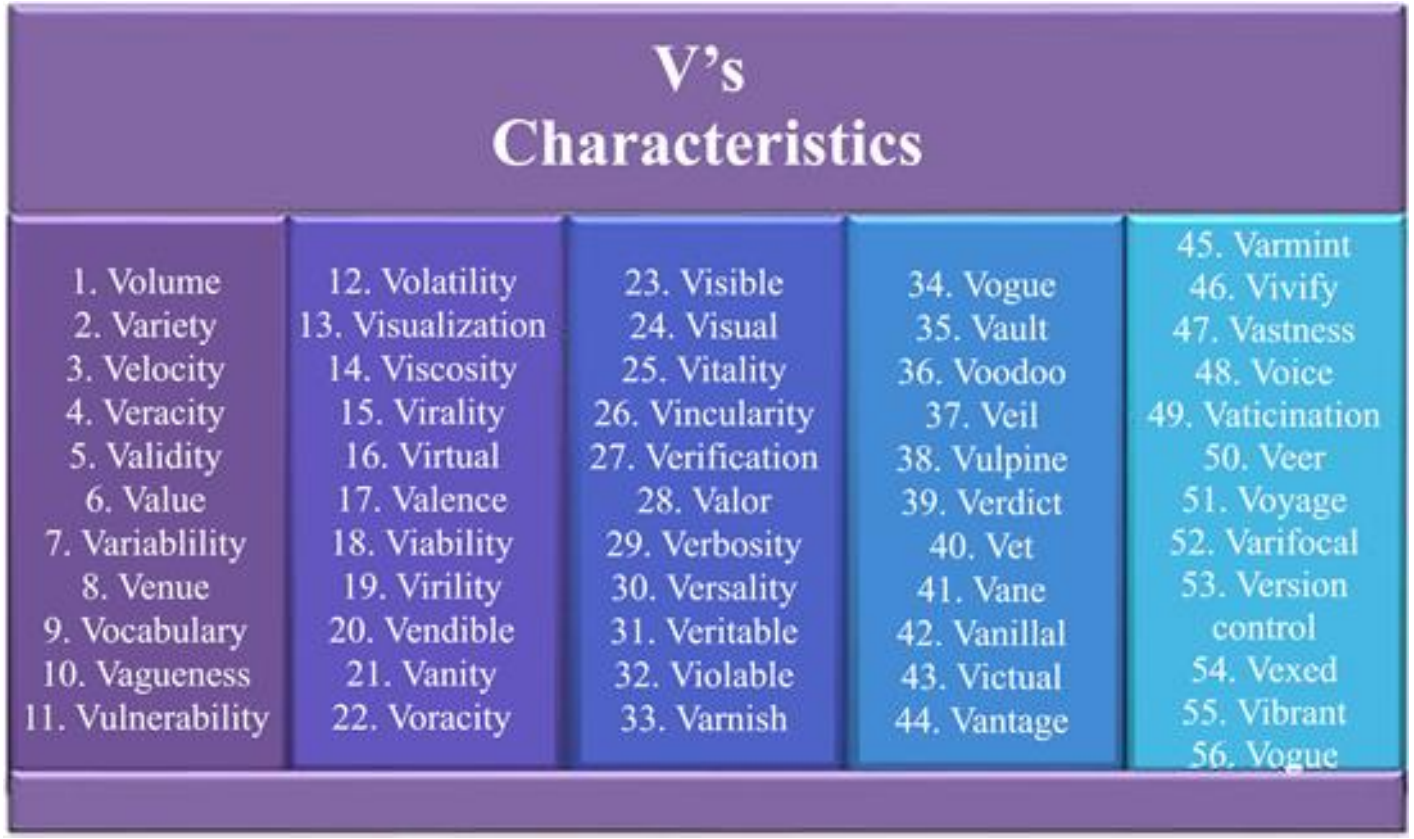


Figure 2-1. *The 3 Vs of data*

The 10 Vs of Data

56 Vs of Data


V's
Characteristics
1. Volume
2. Variety
3. Velocity
4. Veracity
5. Validity
6. Value
7. Variablility
8. Venue
9. Vocabulary
10. Vagueness
11. Vulnerability
12. Volatility
13. Visualization
14. Viscosity
15. Virality
16. Virtual
17. Valence
18. Viability
19. Virility
20. Vendible
21. Vanity
22. Voracity
23. Visible
24. Visual
25. Vitality
26. Vincularity
27. Verification
28. Valor
29. Verbosity
30. Versality
31. Veritable
32. Violable
33. Varnish
34. Vogue
35. Vault
36. Voodoo
37. Veil
38. Vulpine
39. Verdict
40. Vet
41. Vane
42. Vanillal
43. Victual
44. Vantage
45. Varmint
46. Vivify
47. Vastness
48. Voice
49. Vaticination
50. Veer
51. Voyage
52. Varifocal
53. Version control
54. Vexed
55. Vibrant
56. Vogue

Volume

- Demand for storage and respective cost
- The shift from locally housed data to internet and cloud
- Moore’s Law by Gordon Moore

Volume

- Moore states that the number of transistors that are able to fit onto an integrated circuit doubles approximately every 18 months.

- In 1981, the price of a gigabyte of storage was USD 300,000. In 2004, this was $1.00; and it was $0.10 in 2010. Today, 1GB can be rented in the cloud for $0.023 per month, with the first year of storage free

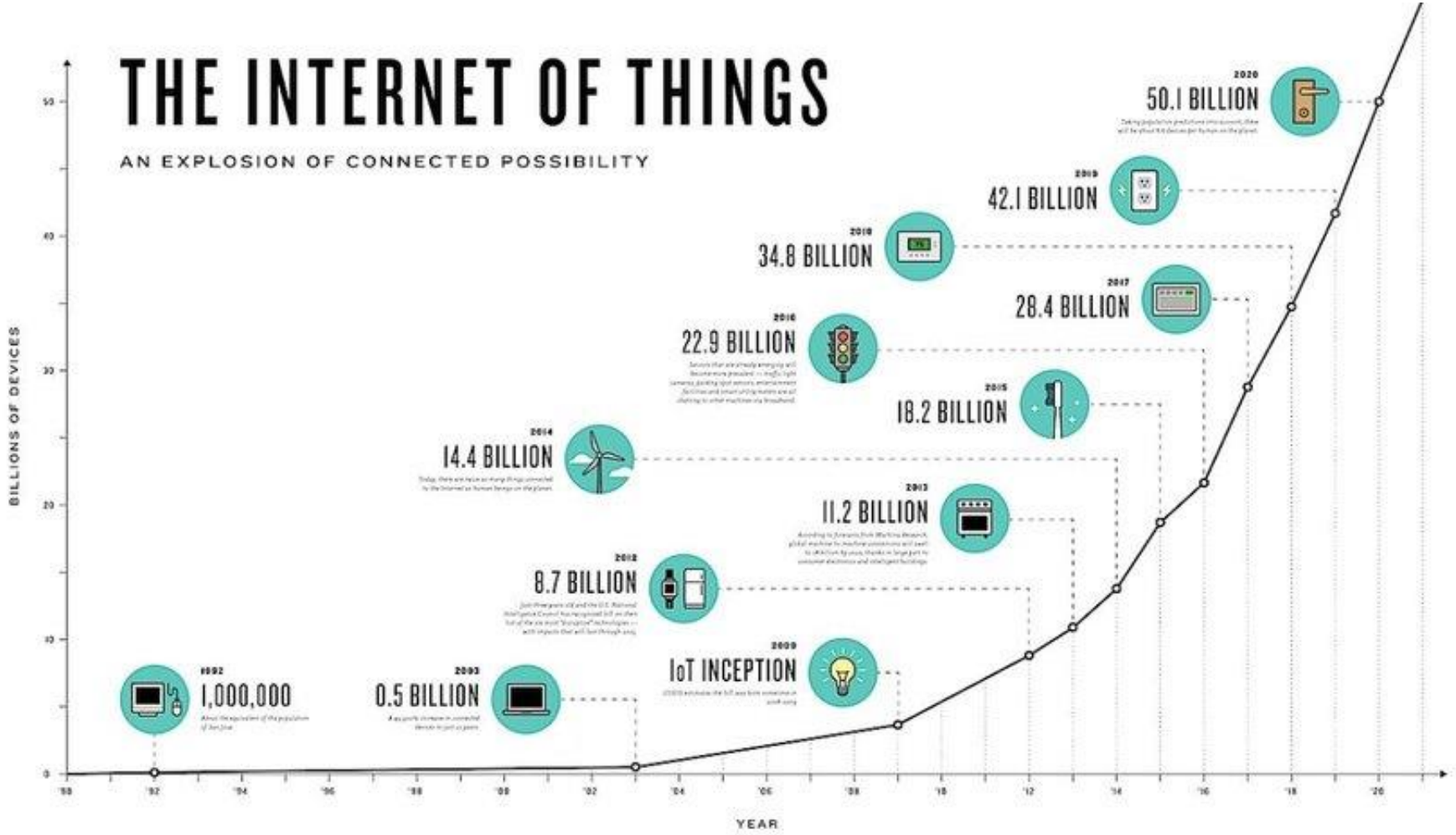
Coping with Data Volume

- Approach
- Types of Data
- Deployment
- Access
- Operations
- Future Use

Variety

- This refers not only to variations in the types of data but also sources and use cases
- Twenty years ago, we used to store data in the form of spreadsheets and databases. Today, data may be in the form of photographs, sensor data, tweets, encrypted files, and so on
- This variety of unstructured data creates problems for storage, mining, and analyzing data

The Internet of Things (IoT)



Velocity

- The third V of big data is velocity, referring to the speed at which data is created, stored, and prepared for analysis and visualization

- The key to maximum clinical value is in the integration of various and heterogeneous data sources.

Wearable biomedical sensors

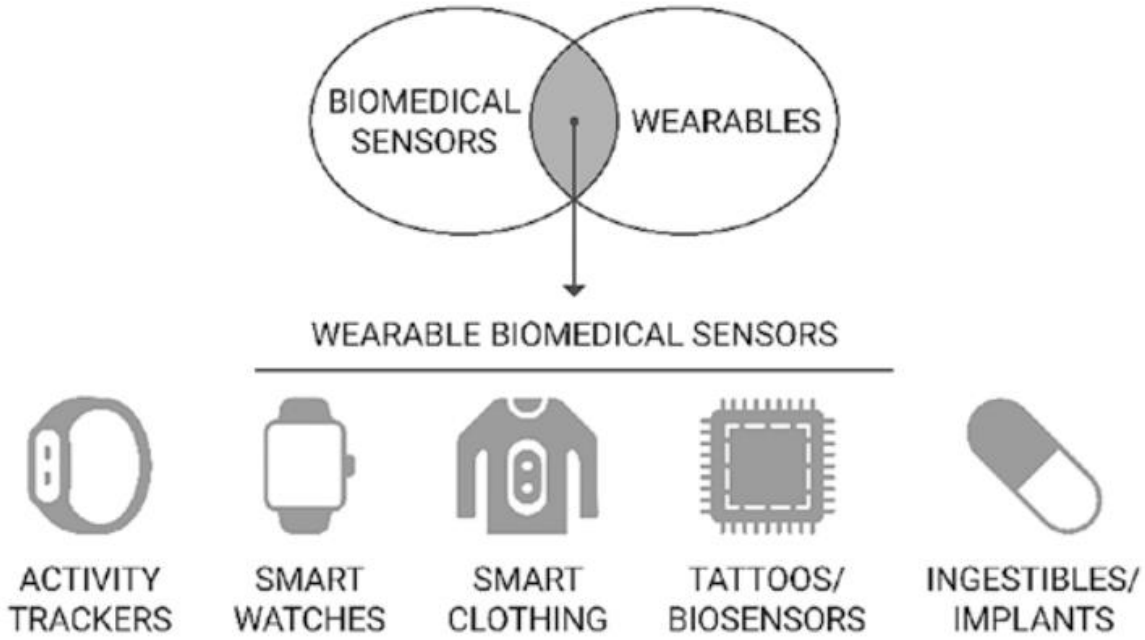


Figure 2-4. *Wearable biomedical sensors*

Value

- Value refers to the usefulness of data
- Accuracy and usability
- Data’s value comes from the analyses conducted on it. It is this analysis of data that enables it to be turned into information, with the aim of eventually turning it into knowledge
- **Value is not found, it is made. The key is to make the data meaningful to your users and your organization by managing it well**

Big Data and Value

- The power of big data was best demonstrated by monolith Google in 2009 when it was able to track the spread of influenza with only a one-day delay across the United States, faster than the Centers for Disease Control and Prevention (CDC), through the analysis of associated search terms.[19]
- However, in 2013, Google's Flu Trends got it wrong—which spurred questions on the concepts of value and validity of data.

- Google did a perfect job in tracking the searches for flu symptoms in 2013. However, it had no way of knowing that its results weren't valid regarding predicting flu prevalence. Google Flu Trends, at least in this instance, wasn't providing valuable results for this use case.
- Google was never to know that real-world cases of flu failed to match the search distribution and frequency—so it could never do anything about it.

VERACITY

#5

صداقت، صحت - truthfulness and reliability

Worthless data is worthless, no matter how plentiful

Veracity refers to data **truthfulness** and whether data is of optimal quality and suitability in the context of its use relating to the **biases, noise, and abnormality in data**.

Factors:

- Data entry
- Data management
- Integration quality
- Staleness
- Usage

6 Cs of trusted data

- **Clean data:** deduplication, standardization, verification, matching, and processing
  - *Biggest challenge*
- **Complete data**
- **Current data**
- **Consistent data**
- **Compliant data**
- **Collaborative data**

Complete veracity; Ideal but not achievable?

- While most advocate complete veracity in data for clinical care, the gold standard of clinical care is harder to achieve in real-world scenarios with such a varied input of data.
- Enforcing data quality may be considered too difficult to achieve cost-effectively.
- In this case, training an intelligent system to learn to **estimate for parameters not seen** may be of use.
- Veracity can refer to ensuring sizeable training samples for rich model-building and validation, which empowers whole-population analytics.

Veracity applies to models too

- Not only does data need to be trustworthy, but so too do the algorithms and systems interpreting it.

Humans make mistakes or to err is human!

- Data entry can be a risky point that goes unnoticed without good data science. This is typically a human-based problem. Even in small data settings, humans make mistakes.

- To be clinically relevant, data requires maximum veracity. As the veracity of data improves, machine learning on this data enables more veracious conclusions.

A hypothetical scenario

- An AI model is being developed to predict the likelihood of sepsis in ICU patients based on electronic health records (EHRs).

How to ensure veracity?

To ensure veracity, the data must be accurate, complete, and trustworthy. For instance:

- **Data Sources:** The EHR data should come from reputable hospitals with consistent data collection practices. Data from poorly maintained records, or institutions with low standards for data entry, could introduce errors.
- **Data Quality Checks:** Procedures like removing duplicate records, addressing missing values, and cross-checking lab results against clinician notes are necessary to improve the data's reliability.
- **Clinical Accuracy:** The timestamps of events (e.g., time of antibiotic administration) must be recorded correctly. If the time is inaccurate, the AI model could make incorrect predictions about sepsis onset.

VALIDITY

#6

روایی، اعتبار - accuracy and appropriateness

What is validity?

- Referring to whether is data is correct and accurate for the intended use.
- To ensure that only useful and relevant data is used.
- Truthfulness or veracity of data is absolute, whereas validity is contextual.

The scenario: how to ensure validity?

For data to be valid, it needs to be relevant and appropriately used in the model's context. For example:

- **Predictor Variables:** Using lab results like white blood cell count and vital signs (heart rate, temperature) is valid for sepsis prediction, while using irrelevant factors (like a patient's marital status) would lack validity.
- **Population Match:** If the model is intended for adult ICU patients, but the training data includes pediatric cases, the predictions may not be valid for the intended adult population.
- **Outcome Consistency:** The definition of sepsis should be consistent across the dataset (e.g., following established clinical guidelines). Inconsistent labeling could lead to the model learning from incorrect or ambiguous examples.

VARIABILITY

#7

تغییرپذیری

Big data is variable

- Variability defines data where the meaning is regularly changing. Variability is very relevant in performing sentiment analyses.
- For example, in a series of tweets, a single word can have a completely different meaning.

Variability vs. variety

- Variability is often confused with variety. To illustrate, a florist may sell five types of roses. That is variety. Now, if you go to the florist for two weeks in a row and buy the same white rose ever day, each day it will have a subtly different form and fragrance. That is variability.

Model should understand the context

- To perform proper sentiment analyses, algorithms need to be able to understand the context of texts and be able to decipher the exact meaning of a word in that particular context. This is still very difficult, even with progress in natural language processing abilities.

VISUALIZATION

#8

Make data understandable

- Referring to the appropriate analyses and visualizations required on big data to make it readable, understandable, and actionable.
- Visualization may not sound complex; however, overcoming the challenge of visualizing complex datasets is crucial for stakeholder understanding and development.
- Quite often, it is data visualization that becomes the principal component in **transferring knowledge learned from datasets to stakeholders.**

SMALL DATA

#9

Small vs. Big data

- Big data is distributed, varied, and comes in real time;
- Small data is data that is accessible, informative, and actionable as a result of the format and volume.
- Examples of small data include patient medical records, prescription data, biometric measurements, a scan, or even Internet search histories.
- In comparison to organizations and services such as Google and Amazon, the amount of data is far smaller in comparison.

Going back with a new mindset

- Going back to traditional datasets and applying more modern techniques such as machine learning should not be overlooked and is a great place for a data scientist to start.
- **It’s not the size of the data; it’s what you do with it that counts.**

METADATA

#10

Metadata = Data about Data

- Which is descriptive data about each asset, or individual piece, of data.
- Metadata provides granular information about a single file supporting the facility to discover patterns and trends from metadata as well as the data it is supporting.
- Metadata gives information about a file’s origin, date, time, and format; and it may include notes or comments.

Importance of Metadata

- It is key to investing resources, predominantly time, in strategic information management to make sure assets are correctly named, tagged, stored, and archived in a taxonomy consistent with other assets in the collection.
- This facilitates quicker dataset linkage and maintains consistent methodology for asset management to ensure files are easy to find, retrieve, and distribute.

With big data comes big metadata

- For example, Google and Facebook use taxonomy languages such as Open Graph that help create a more structured web, enabling more robust and descriptive information to be provided to users. This, in turn, provides human-friendly results to users, optimizing click through and conversion.

Source: https://ahrefs.com/blog/open-graph-meta-tags/

ahrefs

Example

- A machine learning algorithm could use the metadata belonging to a piece of music, rather than (or alongside) the actual music itself, to be able to suggest further relevant music.
- Features of the music such as genre, artist, song title, and year of release could be obtained from metadata for relevant results.

Uses of metadata

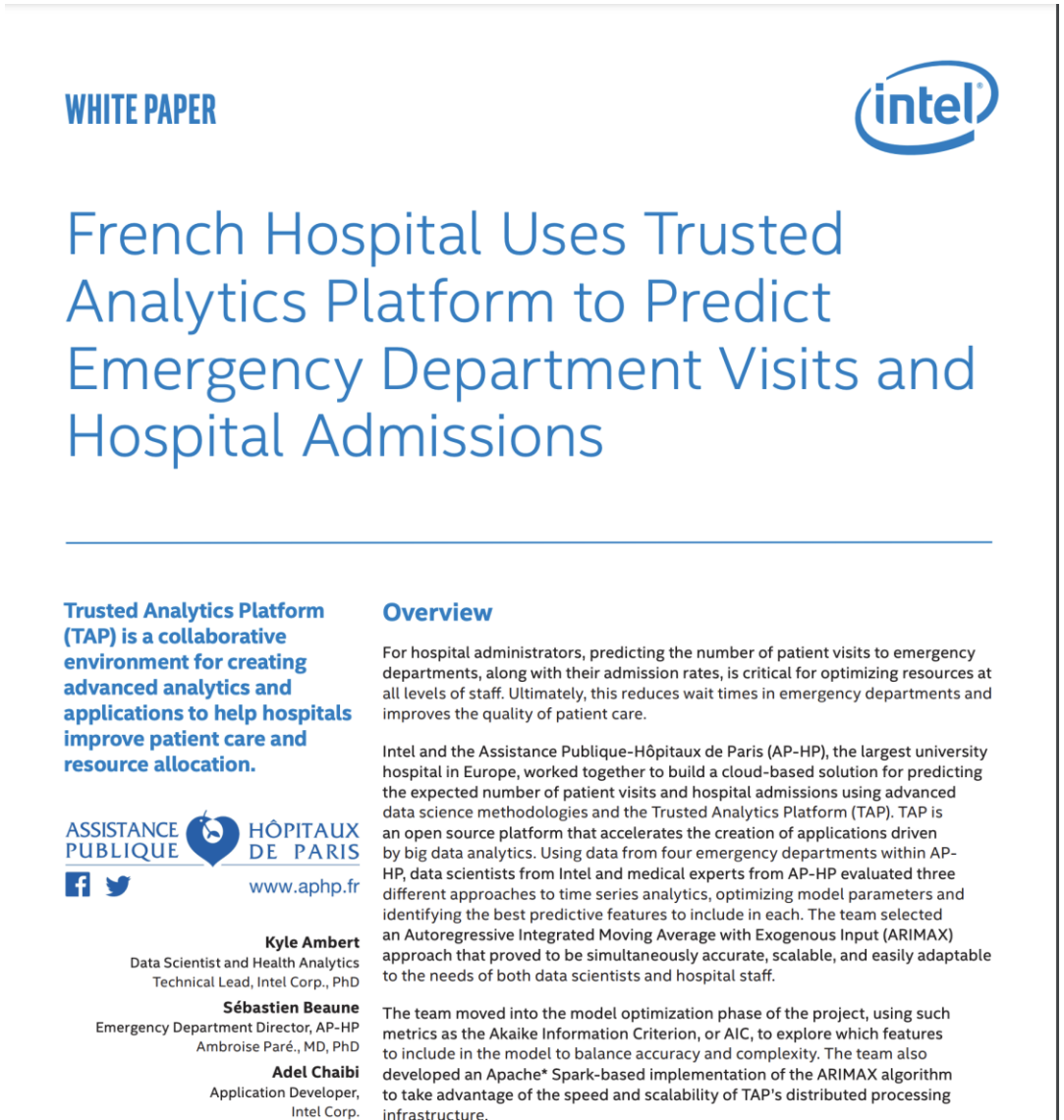
- Handling Missing or Inconsistent Data: If a creatinine level is missing because the patient was discharged, the model should treat this differently than if it were missing due to lab issues.
- Ensuring Data Quality and Reliability: heart rate was measured using a manual technique versus a continuous monitoring device
- Enabling Transfer Learning and Model Generalization: an AI model initially trained on adult ICU data can be adapted to work with pediatric ICU data by utilizing metadata that distinguishes the patient populations.

HEALTHCARE DATA USE CASES

#11

#1 Number of visitors and admission

- Assistance Publique-Hôpitaux de Paris (AP-HP) + Intel



WHITE PAPER

intel

French Hospital Uses Trusted Analytics Platform to Predict Emergency Department Visits and Hospital Admissions

Trusted Analytics Platform (TAP) is a collaborative environment for creating advanced analytics and applications to help hospitals improve patient care and resource allocation.

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

www.aphp.fr

Kyle Ambert
Data Scientist and Health Analytics Technical Lead, Intel Corp., PhD

Sébastien Beaune
Emergency Department Director, AP-HP Ambroise Paré., MD, PhD

Adel Chaibi
Application Developer, Intel Corp.

Overview

For hospital administrators, predicting the number of patient visits to emergency departments, along with their admission rates, is critical for optimizing resources at all levels of staff. Ultimately, this reduces wait times in emergency departments and improves the quality of patient care.

Intel and the Assistance Publique-Hôpitaux de Paris (AP-HP), the largest university hospital in Europe, worked together to build a cloud-based solution for predicting the expected number of patient visits and hospital admissions using advanced data science methodologies and the Trusted Analytics Platform (TAP). TAP is an open source platform that accelerates the creation of applications driven by big data analytics. Using data from four emergency departments within AP-HP, data scientists from Intel and medical experts from AP-HP evaluated three different approaches to time series analytics, optimizing model parameters and identifying the best predictive features to include in each. The team selected an Autoregressive Integrated Moving Average with Exogenous Input (ARIMAX) approach that proved to be simultaneously accurate, scalable, and easily adaptable to the needs of both data scientists and hospital staff.

The team moved into the model optimization phase of the project, using such metrics as the Akaike Information Criterion, or AIC, to explore which features to include in the model to balance accuracy and complexity. The team also developed an Apache* Spark-based implementation of the ARIMAX algorithm to take advantage of the speed and scalability of TAP's distributed processing infrastructure.

#2 Reducing readmissions

- UT Southwestern hospital in the United States, EHRs analytics led to a drop in the readmission rate of cardiac patients from 26.2% to 21.2% through successful identification of at-risk patients.

Comparative Study > BMJ Qual Saf. 2013 Dec;22(12):998-1005.
doi: 10.1136/bmjqs-2013-001901. Epub 2013 Jul 31.

Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study

Ruben Amarasingham [1], Parag C Patel, Kathleen Toto, Lauren L Nelson, Timothy S Swanson, Billy J Moore, Bin Xie, Song Zhang, Kristin S Alvarez, Ying Ma, Mark H Drazner, Usha Kollipara, Ethan A Halm

Affiliations + expand

PMID: 23904506 PMCID: PMC3888600 DOI: 10.1136/bmjqs-2013-001901
Free PMC article

Abstract

Objective: To test a multidisciplinary approach to reduce heart failure (HF) readmissions that tailors the intensity of care transition intervention to the risk of the patient using a suite of electronic medical record (EMR)-enabled programmes.

Methods: A prospective controlled before and after study of adult inpatients admitted with HF and two concurrent control conditions (acute myocardial infarction (AMI) and pneumonia (PNA)) was performed between 1 December 2008 and 1 December 2010 at a large urban public teaching hospital. An EMR-based software platform stratified all patients admitted with HF on a daily basis by their 30-day readmission risk using a published electronic predictive model. Patients at highest risk received an intensive set of evidence-based interventions designed to reduce readmission using existing resources. The main outcome measure was readmission for any cause and to any hospital within 30 days of discharge.

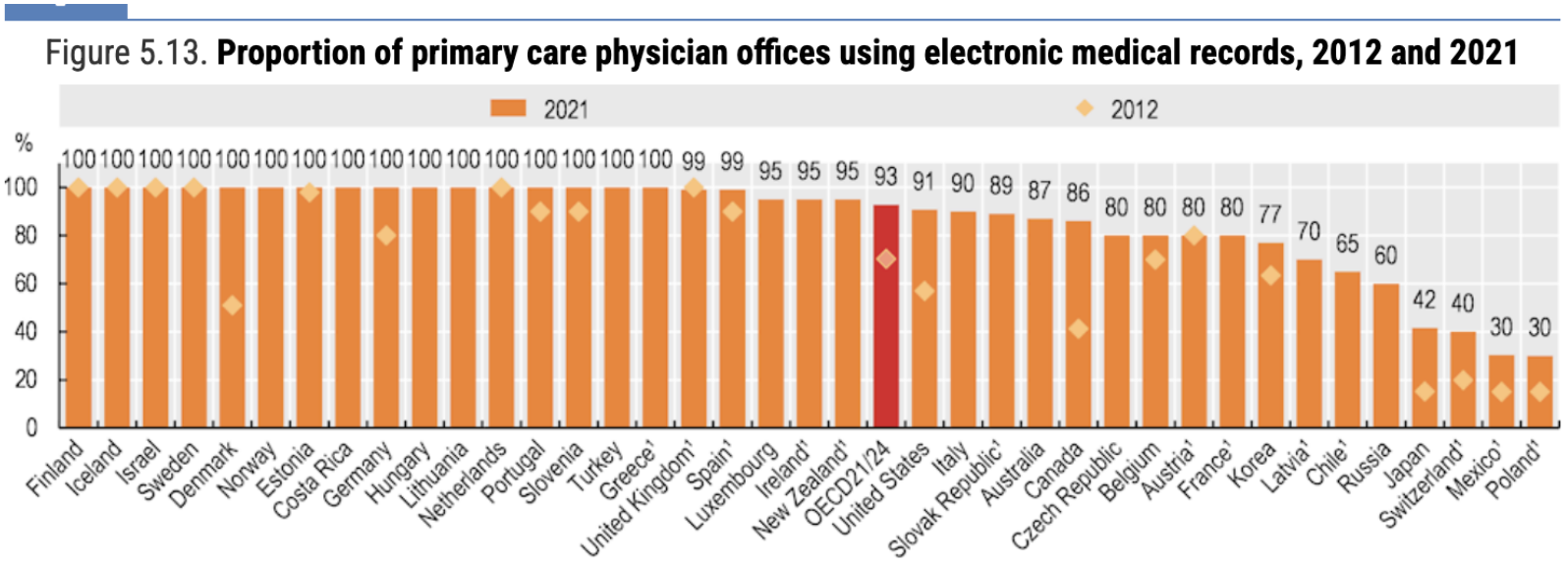
Results: There were 834 HF admissions in the pre-intervention period and 913 in the post-intervention period. The unadjusted readmission rate declined from 26.2% in the pre-intervention period to 21.2% in the post-intervention period (p=0.01), a decline that persisted in adjusted analyses (adjusted OR (AOR)=0.73; 95% CI 0.58 to 0.93, p=0.01). In contrast, there was no significant change in the unadjusted and adjusted readmission rates for PNA and AMI over the same period. There were 45 fewer readmissions with 913 patients enrolled and 228 patients receiving intervention, resulting in a number needed to treat (NNT) ratio of 20.

#3 Predictive Analytics

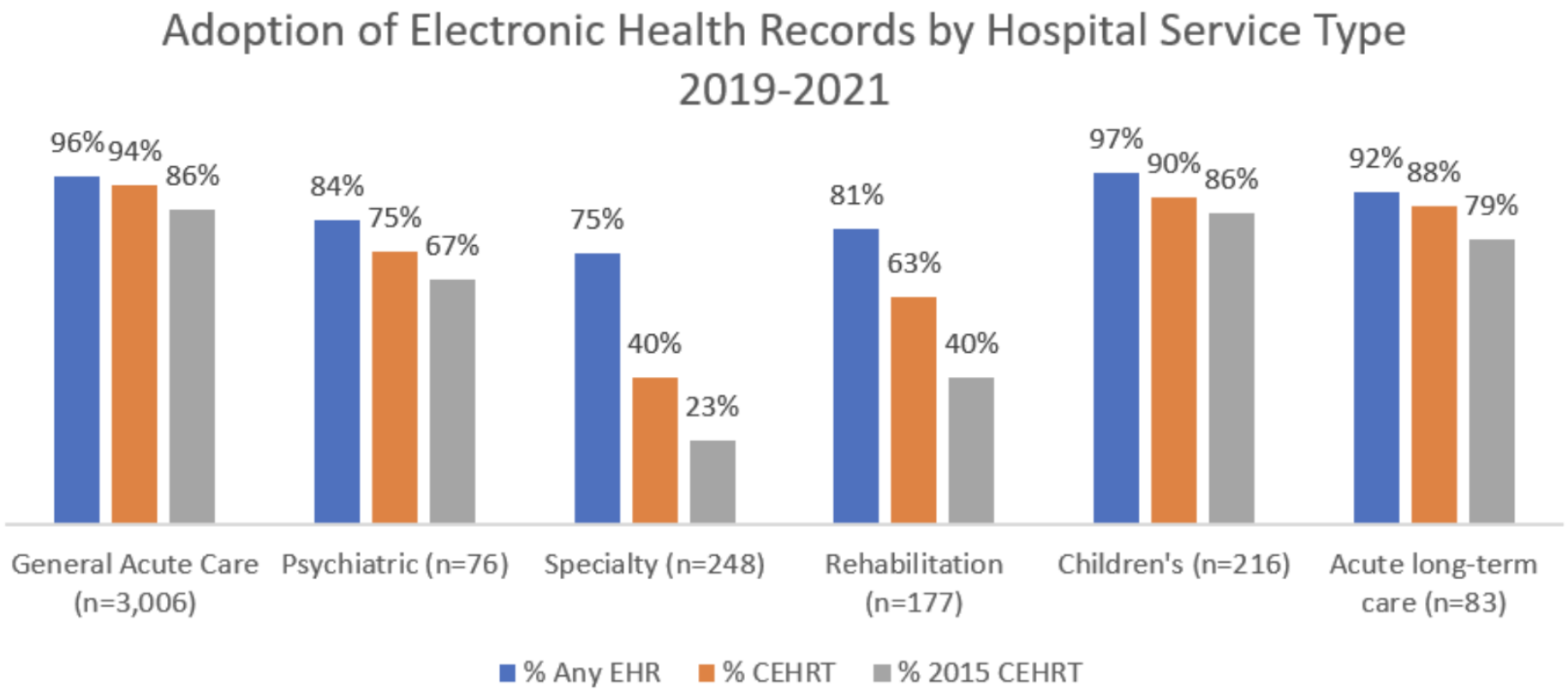
- OptumLabs has a history of building predictive AI models with regression and machine-learning methods. For example, we've constructed models to predict the onset of Alzheimer's disease, of diabetes, as well as models of patient clusters with heart failure and COPD.
- Optum labs have over 160 million de-identified medical records

#4 EHR

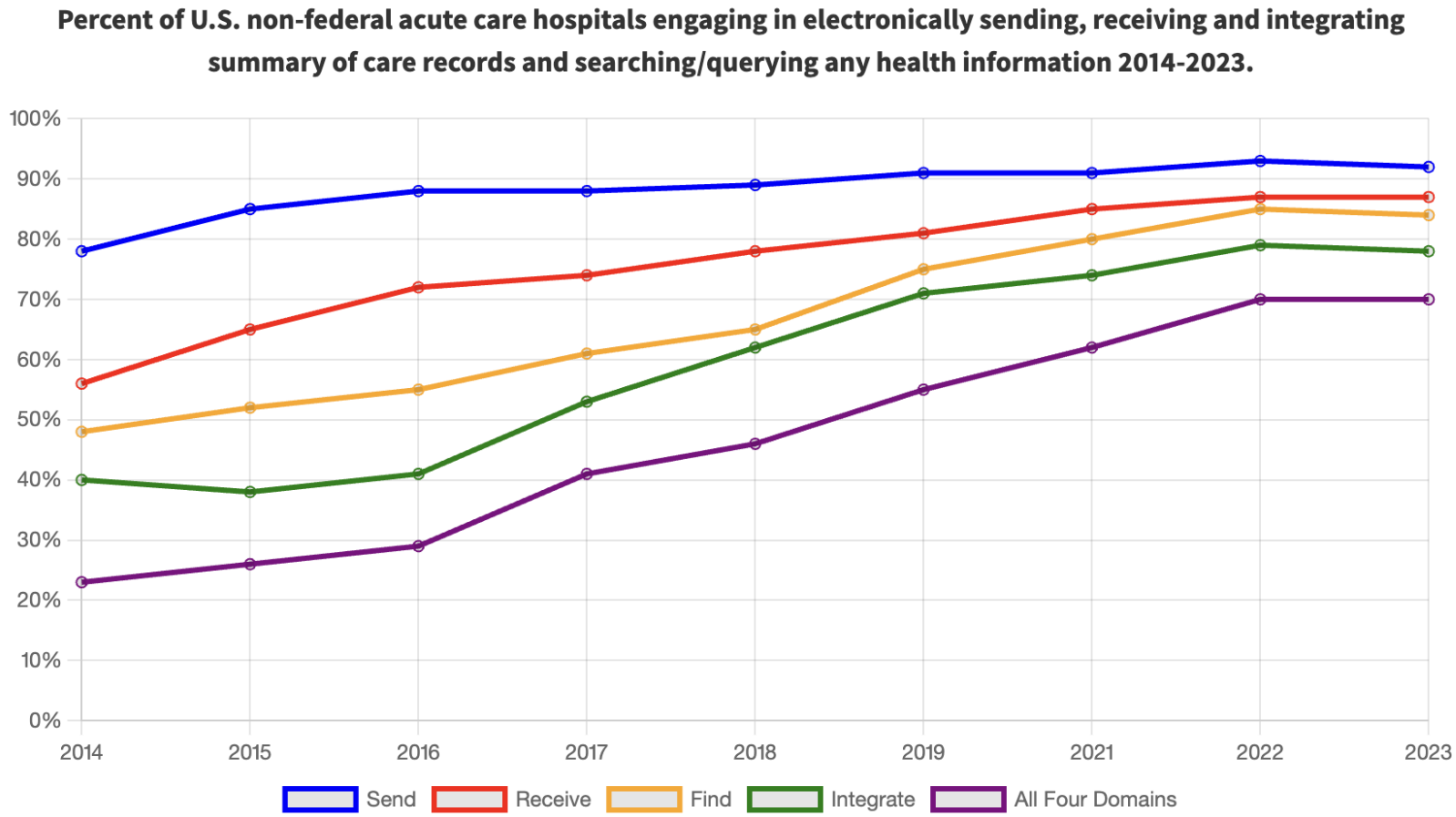
- Starting in 2015, hospitals and doctors will be subject to financial penalties under Medicare if they are not using electronic health records.



#4 EHR


Adoption of Electronic Health Records by Hospital Service Type
2019-2021
96%
94%
86%
84%
75%
67%
75%
40%
23%
81%
63%
40%
97%
90%
86%
92%
88%
79%
General Acute Care (n=3,006)
Psychiatric (n=76)
Specialty (n=248)
Rehabilitation (n=177)
Children's (n=216)
Acute long-term care (n=83)
% Any EHR
% CEHRT
% 2015 CEHRT

#4 EHR


Percent of U.S. non-federal acute care hospitals engaging in electronically sending, receiving and integrating summary of care records and searching/querying any health information 2014-2023.
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2014
2015
2016
2017
2018
2019
2021
2022
2023
Send
Receive
Find
Integrate
All Four Domains

#5 Value-Based Care/Engagement

- No longer are patients considered passive recipients of care.
- Better patient engagement enhances trust between patients, treatment providers, and bill payers. Moreover, it leads to better health outcomes and cost savings (or some other benefit) to the provider.
- Blue Shield of California is improving patient outcomes by developing an integrated system that connects doctors, hospitals, and health coverage to the patient's broader health data to deliver evidence-based, personalized care.

#6 Healthcare IoT

- Millions of people use devices that data-fy their lives toward the quantified self.
- The data recorded could be used to detect the risk of disease, alert doctors, or request emergency services depending on the biometrics received.
- **With sophisticated devices come sophisticated solutions to novel problems.**
- an innovative program from the University of California, Irvine, gave patients with heart disease the opportunity to return home with a wireless weighing scale and weigh themselves at regular intervals. Predictive analytics algorithms determined unsafe weight gain thresholds and alerted physicians to see the patient proactively before an emergency readmittance was necessary.

#6 Healthcare IoT: problems

- Demotivating: According to several randomized trials, Fitbit wearers do exercise more, but not enough to guarantee weight loss and improved fitness. In fact, some studies ahave determined they can be demotivating.
- Accuracy: A Cleveland Clinic study in 2016 found that heart rate monitors from four brands on the market were reporting inaccurate readings 10 to 20% of the time.
- Abandonment: an average abandonment rate of more than 30% after a period of engaged usage.

#6 Healthcare IoT: solutions

- Offering users of devices such as these tangible incentives like discounts on health or life insurance
- Warnings and alerts

#7 Evidence based medicine

- Clinical trials work on a small scale, testing new treatments in small groups with internal validity (i.e., no other conditions or concerns other than those specified), and looking at how well treatments work and establishing if there are any side effects. Warnings and alerts
- With growing datafication, there is also increasing “real-world evidence” or data, which can be analyzed at an individual level to create a patient data model and aggregated across populations to derive larger insights around disease prevalence, treatment, engagement, and outcomes.

#8 public health

- Analysis of disease patterns and outbreaks allows public health to be substantially improved through an analytics-driven approach
- In West Africa, mobile phone location data proved invaluable in tracking the spread of the population—and as a result, helped to predict the Ebola virus's expanse.
- After the Haiti earthquake in 2010, a team from Karolinska Institute in Sweden and Columbia University in the United States analyzed calling data from two million mobile phones on the Digicel Haiti network.[35] Phone records were used to understand population movements and for the United Nations to allocate resources more efficiently. The data was also used to identify areas at risk of the subsequent cholera outbreak.

AIHRC

THANK YOU