Introduction to Big Data and Data Science

Anand Paul

Department of Biostatistics and Data Science

Louisiana State University Health Sciences Center

What is Big Data?

How much data are produced ?

Big Data: extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions

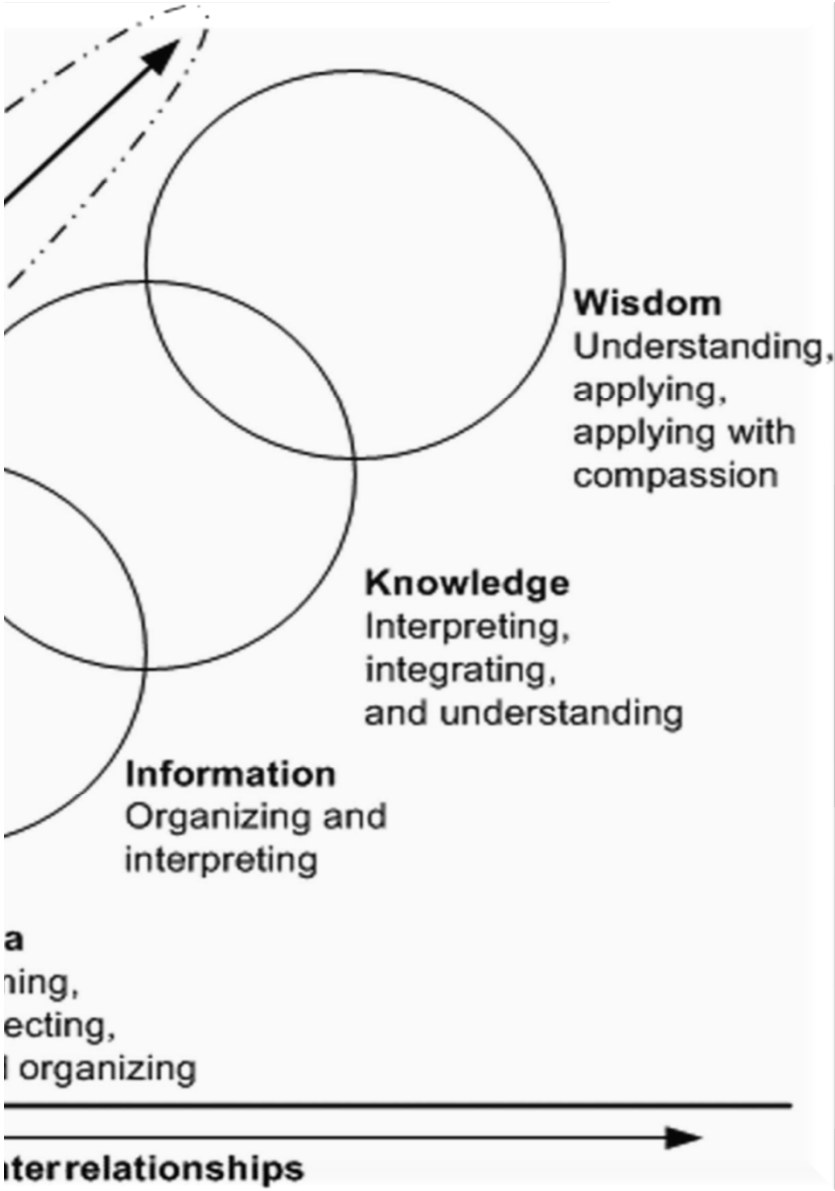
Big Data …

…more that a very large data set

Creates models to capture underlying patterns of complex systems and codifies into working applications

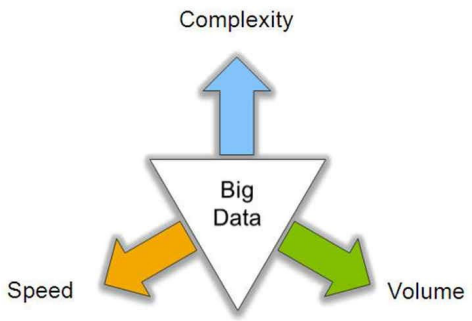
Exponential growth of structured and unstructured data

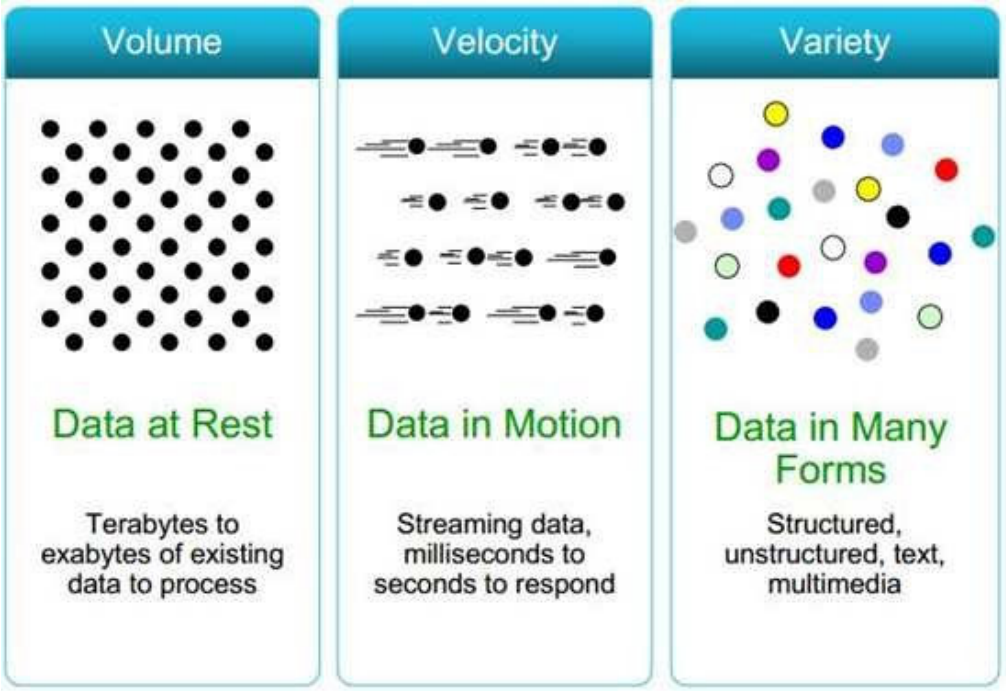
DIKW Framework



- McGonigle, D. & Mastrain, K. (2021). *Nursing Informatics and the Foundation of Knowledge* (5th ed.). Jones and Bartlett Publishers.

3v's of Big Data





https://www.bigdatanews.datasciencecentral.com/profiles/blogs/data-veracity

https://authenticredcreative.com/characteristics-of-big-data-the-3-differentiating-v/

10v's of Big Data

#1: Volume
#2: Velocity
#3: Variety
#4: Veracity
#5: Valance
#6: Validity
#7: Vulnerability
#8: Volatility
#9: Visualization
#10: Value

- now we are heading towards 100 V's

- Top 10 Big Data Challenges – A Serious Look at 10 Big Data V's", https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/

#1 Character of Big Data Volume (Scale)

1000 MB -> 1 Giga Byte
1000 GB - > 1 Tera Byte
10000 TB -> 1 Peta Byte
1000 PB -> 1 Exa Byte
1000 EB -> 1 Zetta Byte(1 ZB)

- **Data Volume**

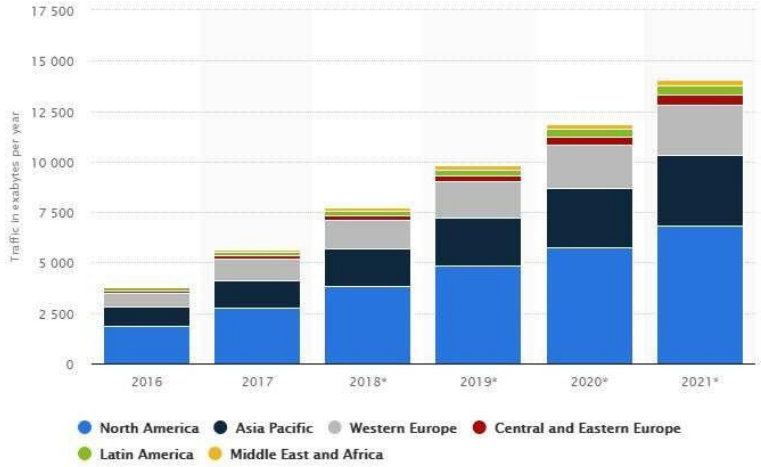
Global IP traffic from cloud data centers in 2018 10.6 zettabytes

Annual cloud data center traffic in North America in 2018 3,838 exabytes

Traffic from data centers to users worldwide in 2018 1.61 zettabytes

Data center storage capacity used for databases and analysis in 2018 230 exabytes

Data volume of global consumer web usage, e-mails and data traffic 9,059PB/mo



https://www.statista.com/topics/1464/big-data/

#2 Character of Big Data -Variety (Complexity)

- Various formats, types and structure
- Text, numerical, images, audio, (Email) video, se quences, time series, multi dimensional array, s ocial media ,Static data Vs. streaming data etc
- A single application can be generating/collecti ng many types of data
- To extract knowledge all these types of data had to linked together

3. Character of Big Data - Velocity (Speed)

- Data is begin generated fast and need to be processed fast(Real-Time Processing)
- Online Data Analytics Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:**
    Based on your current location,
    your purchase history, what you like
    ➔ send promotions right now for store next to you
  - **Healthcare monitoring:**
    sensors monitoring your activities and body
    ➔ any abnormal measurements require
    immediate reaction

4 Character of Big Data - Veracity

- Data in Doubt
  - Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception etc
  - Unpredictable content with no structure

- It may also refer to noise, and abnormality in data

#5 Valence (Electrons) : refers to the connectedness of big data in the form of graphs, just like atoms

- Data items are often directly connected to one another
- A city is connected to the country it belongs to
- Two Facebook users are connected because they are friends
- An employee is connected to his work place
- Data could also be indirectly connected
- Two scientists are connected, because they are both physicists

#6 Validity

- **Validity** in **data** collection means that your findings
- truly represent the phenomenon you are claiming to measure

- validity refers to how accurate and correct the data is for its intended use. Consistent data quality

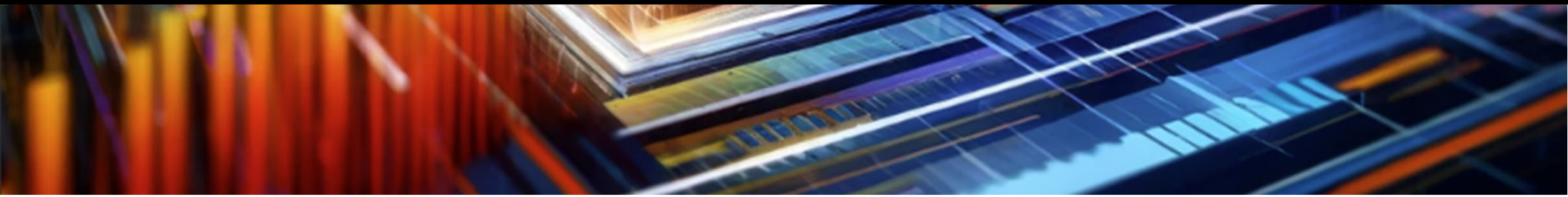
#7 Vulnerability

- Security issues in Big Data.
  - privacy
  - company behind data
  - hacking issues

#8 Volatility

- **In Big data volatility** refers to how long is
- **data** valid and how long should it be stored

#9 Visualization of Big Data



#10VALUE

Standards in Healthcare Data

Clinical Standards

LOINC

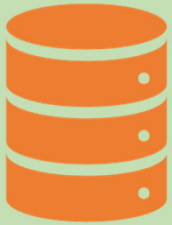
SNOMED-CT

ICD-10 / ICD-11

HL7

LOINC (Logical Observation Identifiers Names and Codes): Used for lab tests and clinical observations.S
NOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms): A comprehensive, multilingual clinical healthcare terminology.ICD-10 /
ICD-11 (International Classification of Diseases): Used for diagnosis coding in healthcare settings.
HL7 (Health Level Seven International): Standards for the exchange, integration, sharing, and retrieval of electronic health information.

Data Management Technologies



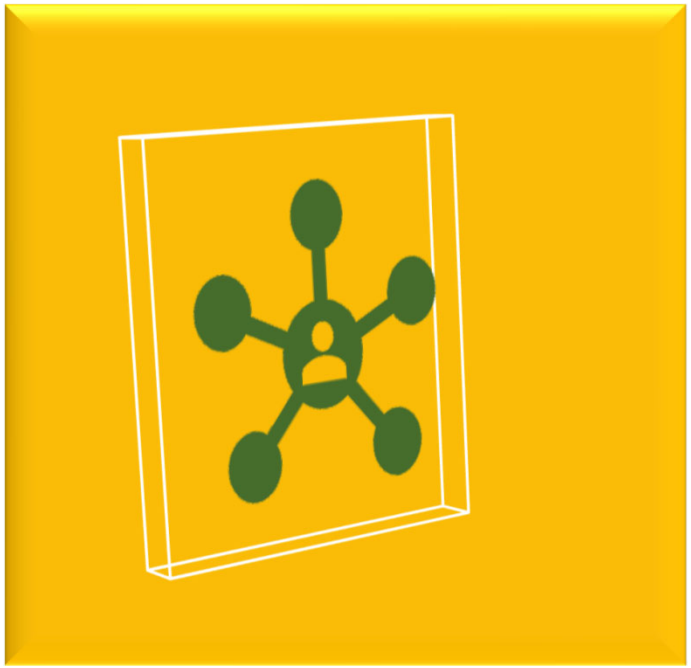
SQL

NoSQL

IHE

SQL (Structured Query Language): Used for managing relational databases.
NoSQL: Pertains to non-relational database systems designed for large-scale data storage and for handling diverse data types more flexibly.
IHE (Integrating the Healthcare Enterprise): An initiative by healthcare professionals and industry to improve the way computer systems in healthcare share information.

Data Types & Analysis



- Tabular
  - Machine learning
- Time Series
  - Time-frequency analysis
- Natural Language
  - Text mining, Sentiment analysis
- Images & Videos
  - Deep neural networks

Data Fragmentation

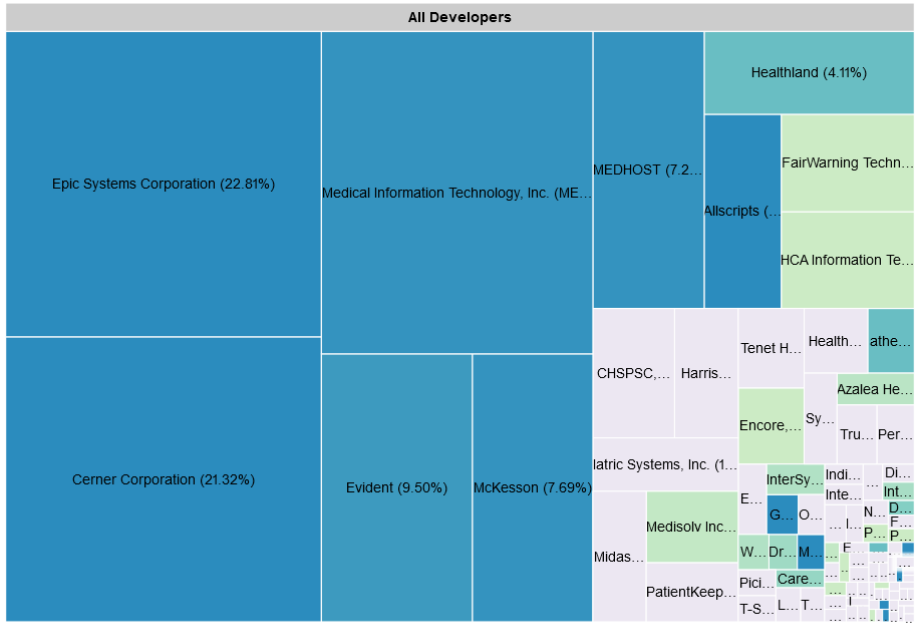
- Missing data
  - Est. 80% missing values
- Unstructured data
  - Est. to double every 3 months
- Various separated databases
  - Lack interoperability
- Exchange issues
  - Administrative, ethical, political & technical

Think about your buying habits…

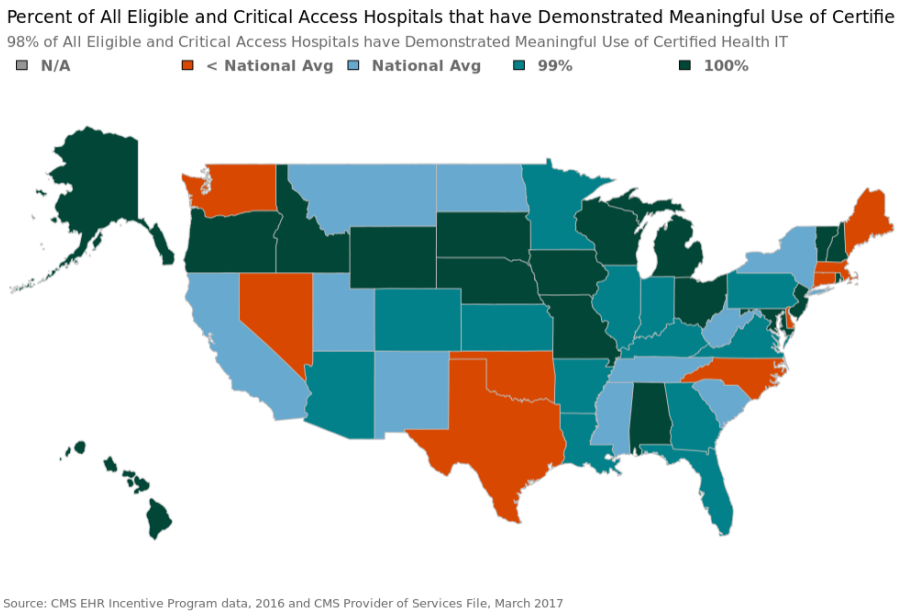
How do retailers know your shopping preference?

Financial institutions and other industries…

Healthcare needs to “catch-up” in data analytics



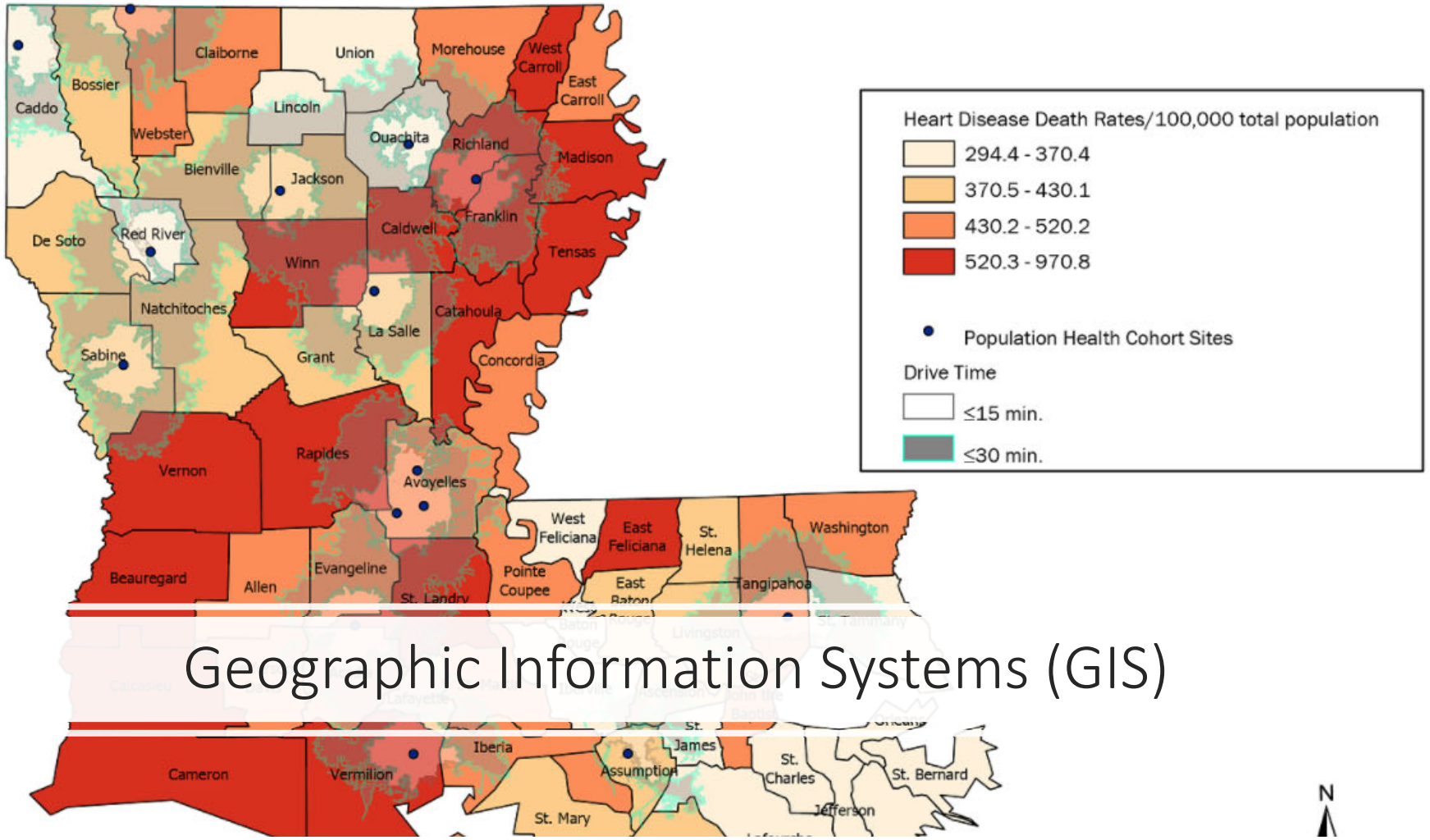
Source: Certified Health IT Product List (CHPL), March 2019; Medicare EHR Incentive Program data, 2016



Source: CMS EHR Incentive Program data, 2016 and CMS Provider of Services File, March 2017

WELL-AHEAD
WELLAHEADLA.COM

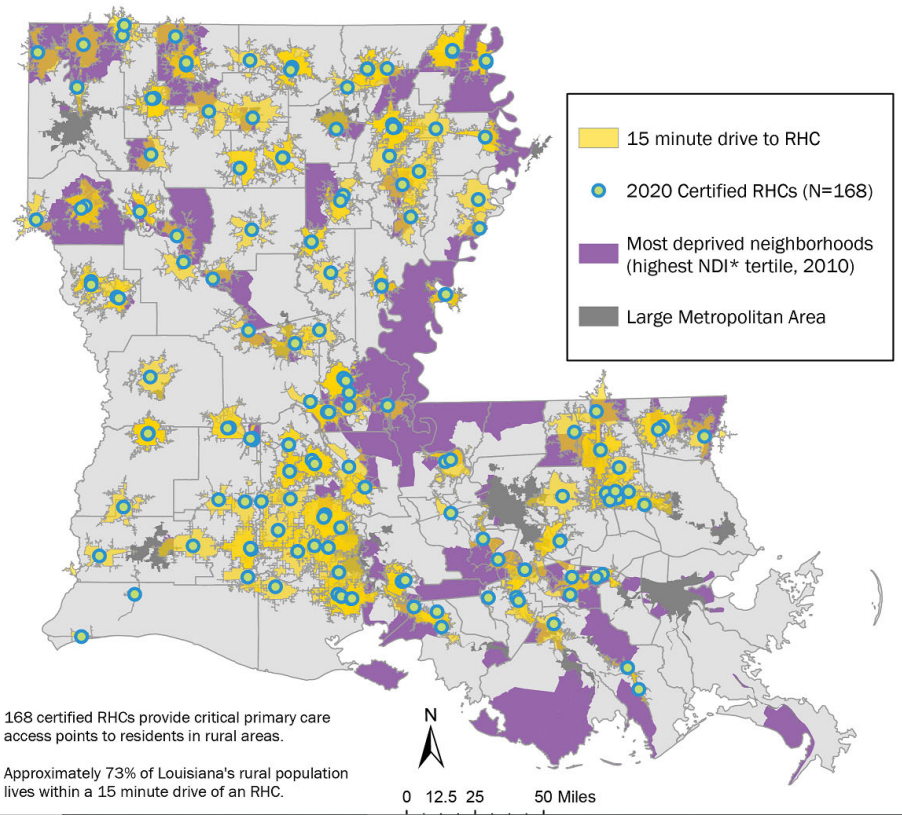
Heart Disease Death Rates by Parish and Geographic Access to Population Health Cohort Sites, Louisiana



Geographic Information Systems (GIS)

WELL-AHEAD
WELLAHEADLA.COM

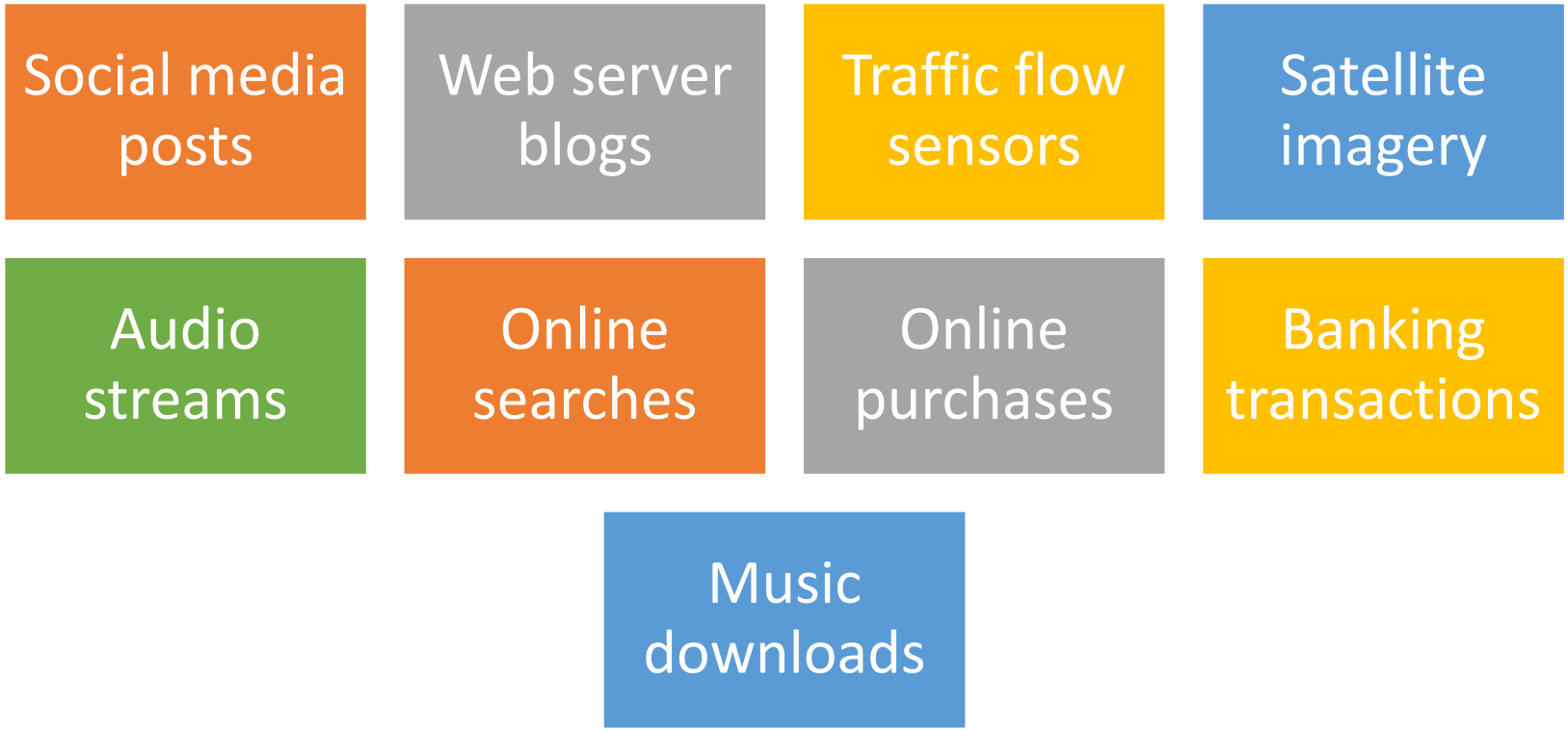
Vulnerable Areas Along Louisiana's Mississippi Delta Lack Geographic Access to Rural Health Clinics (RHC)



168 certified RHCs provide critical primary care access points to residents in rural areas.

Approximately 73% of Louisiana's rural population lives within a 15 minute drive of an RHC.

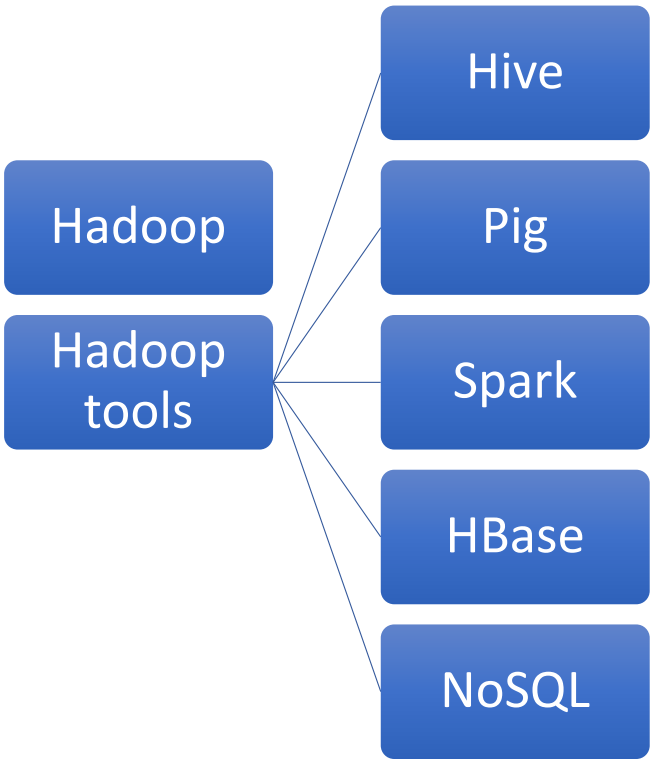
Examples of Big Data

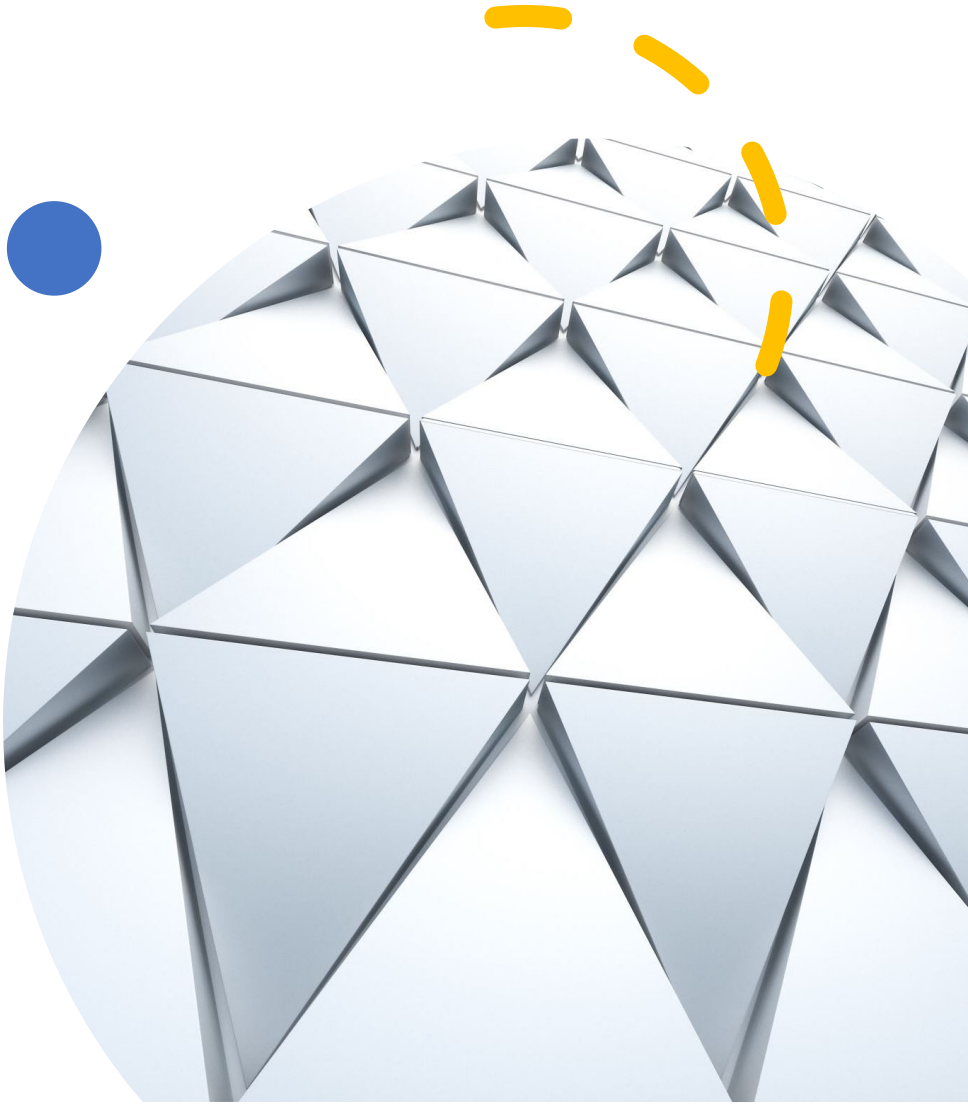



Do you have:
• Type 1 diabetes and 60 years or older
• A care partner or loved one willing to participate with you
• A smart phone
IF YOU ARE 60 OR OLDER & HAVE TYPE 1 DIABETES
What you will do:
• Wear a continuous glucose monitor for 3 months
• Share your blood sugar values via phone
• Complete short questionnaires
• Complete a short interview
What are the benefits?
• Diabetes education
• Real-time continuous glucose monitoring education, including how to share the information with your care partner
• Information on how to respond to high and low blood sugars
• $90 (per person) reimbursement
If interested, please contact Research Ernest Grigorian at ernest.grigorian@utah.edu or 801-707-2775


12:02
mobile.twitter.com
Search Twitter
Log in
Sign up
Michelle Litchman, PhD, FNP, FAA...
@MichLitch
We are recruiting 60+ year olds with #T1D and their live-in care partner (family, roommate, etc) for a #CGM study. Participants will share CGM data with the care partners and complete a survey + interview. If interested, contact Ernest (dot) Grigorian (at) utah (dot) edu. #dsma
2:20 PM · Feb 22, 2021 · Twitter for iPhone
24 Retweets 2 Quote Tweets 19 Likes
Michelle Litchman, PhD, F... · Feb 22

How do you store/process big data?





Data Science Process

Real world exists and creates data

Data are collected

Data are processed

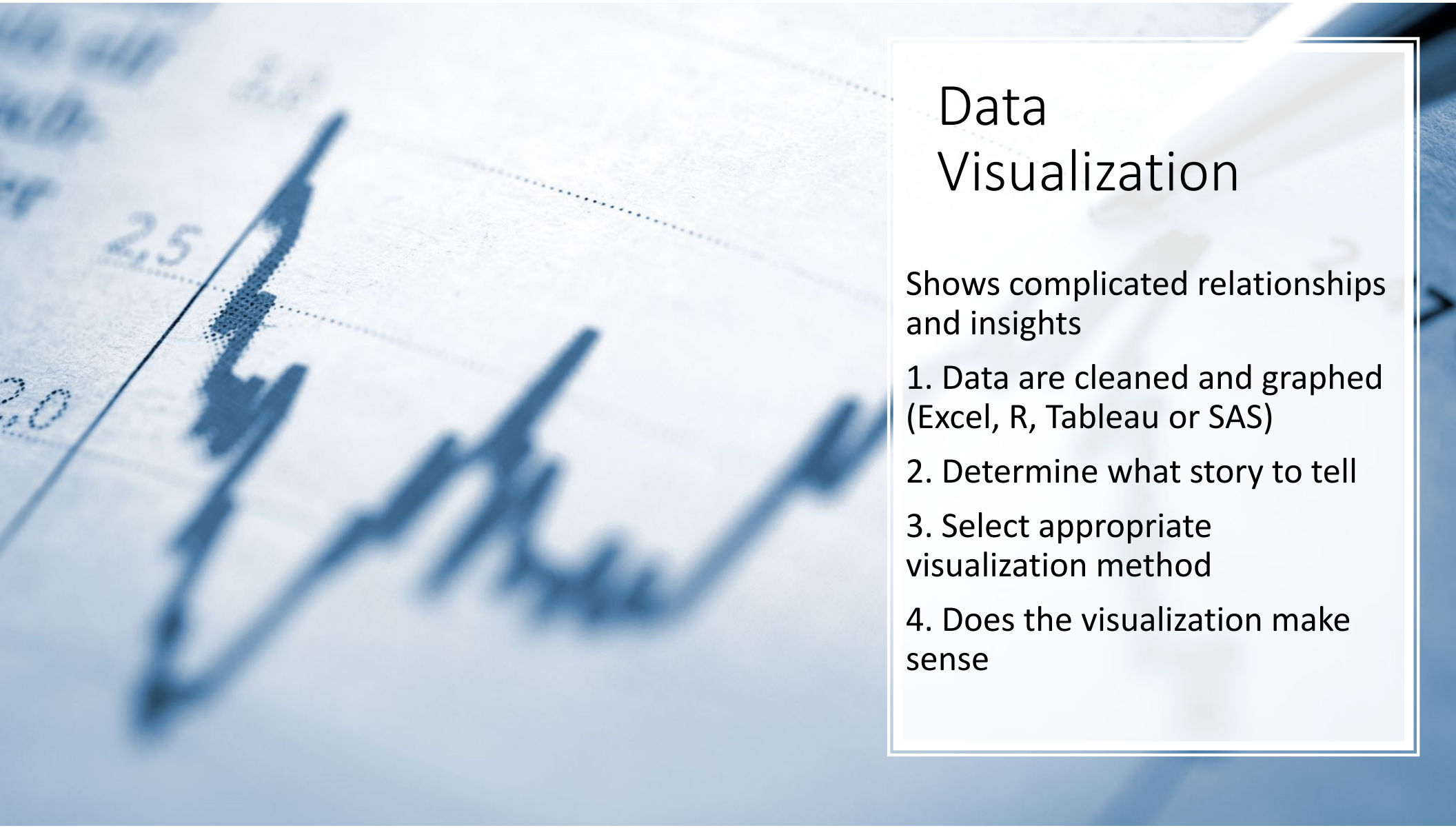
Data cleaning occurs and feeds into machine learning/algorithms, statistical models and communication/visualization/reports

Exploratory analysis feeds back into data collection

Creating models with machine learning, algorithms and statistics feeds into building a data product

Communicate results with data visualization, reports and feed back into decision making

Build a data product that is released into the real world

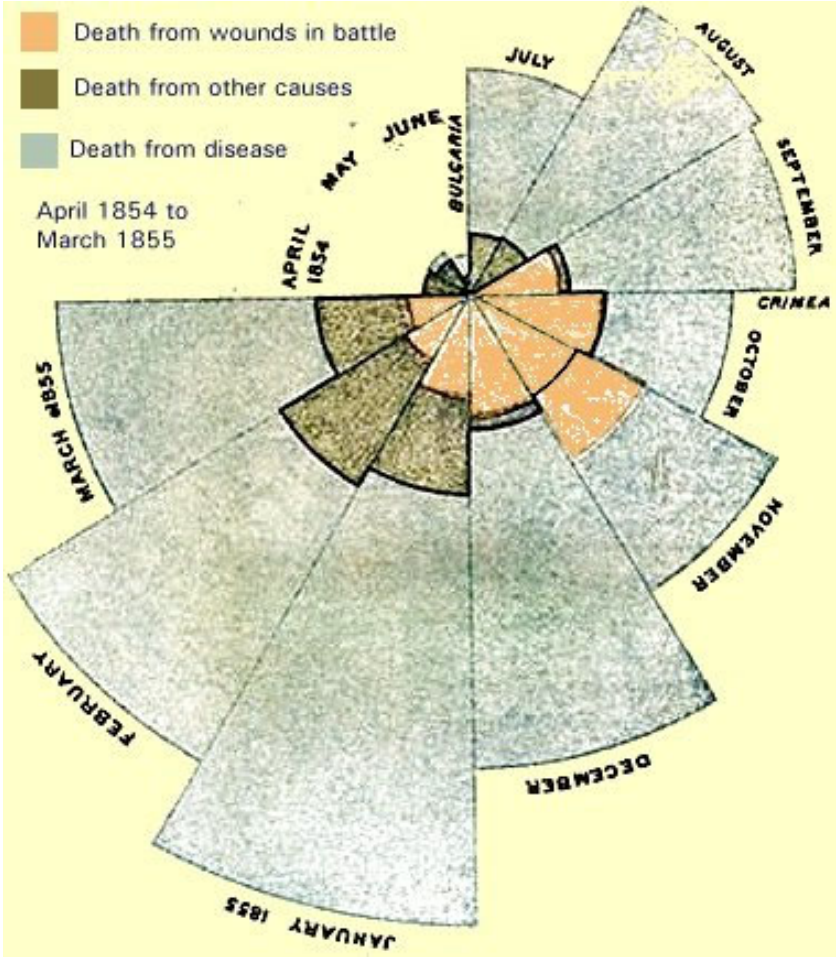


Data Visualization

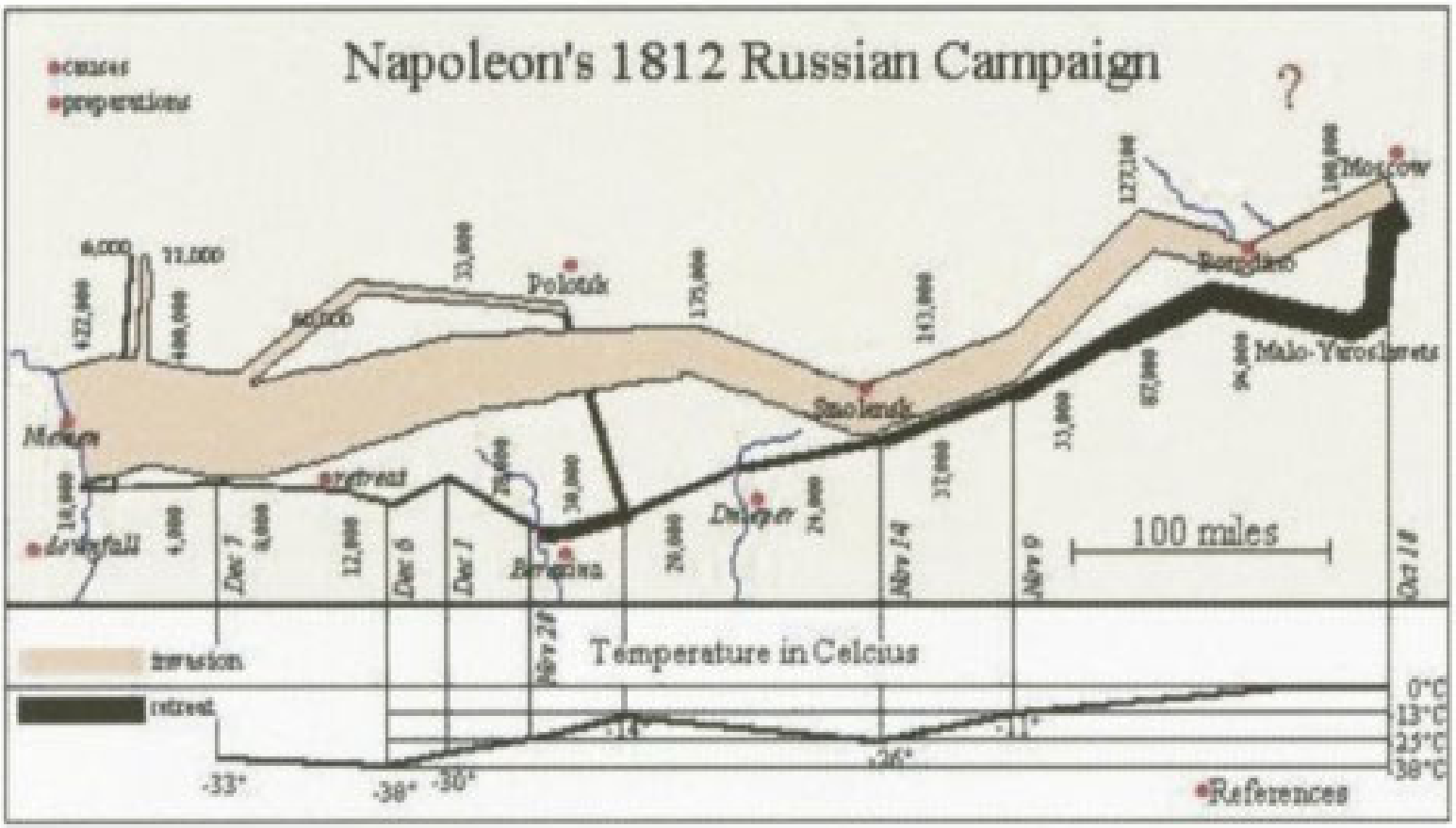
Shows complicated relationships and insights

1. Data are cleaned and graphed (Excel, R, Tableau or SAS)
2. Determine what story to tell
3. Select appropriate visualization method
4. Does the visualization make sense

Nightingale Coxcomb



Minard's Napoleon March on Moscow



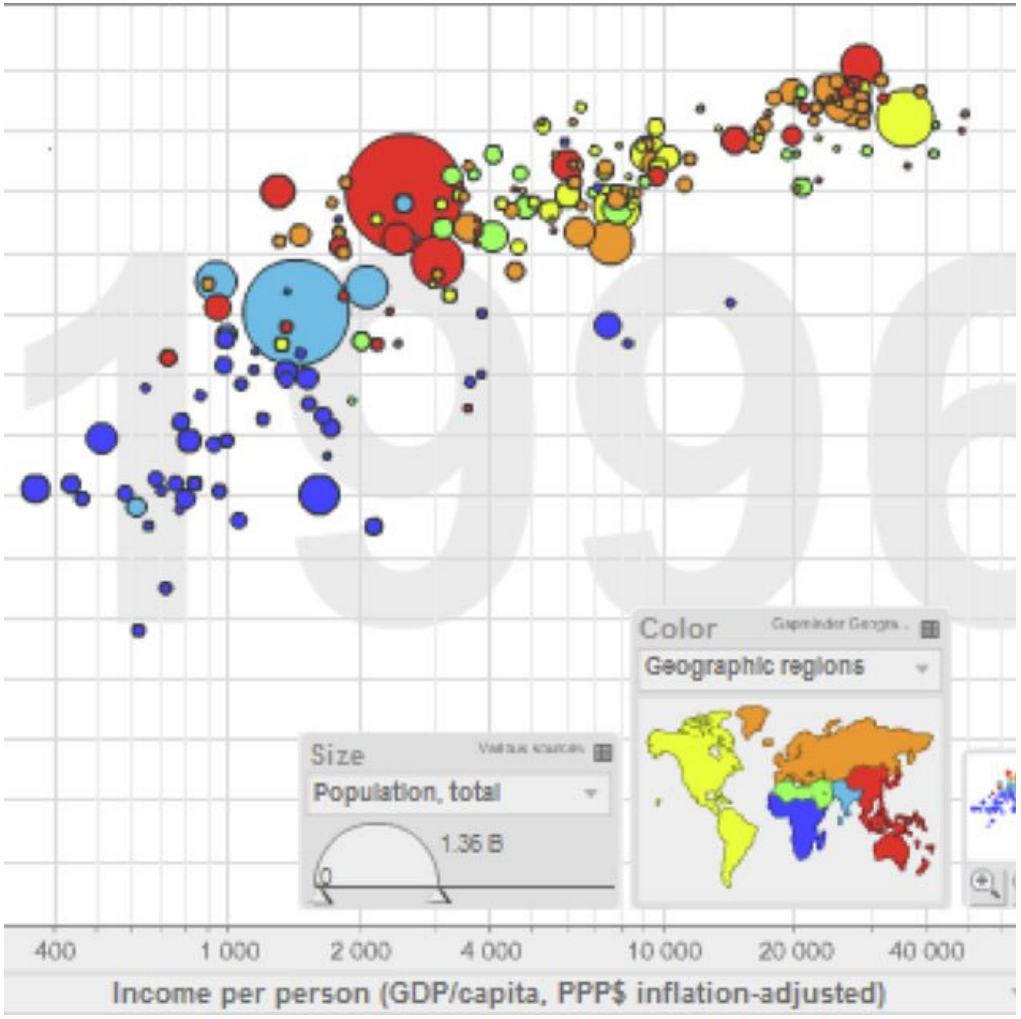
Army Size

Geographic Features

Temperature Chart

Timeline

Rosling's Gapminder



Big Data Requires a Team

- Discipline expert
- Programmers
- Database managers
- Information technology professionals
- Software engineers
- Data wranglers
- Visual designers
- statisticians

Health Data Initiatives in the US

- U.S. Department of Health and Human Services
- Health Data Initiative Forum
- Centers for Medicare and Medicaid Services
- Agency for Healthcare Research and Quality
- National Institutes of Health, National Institute on Nursing Research
- National Cancer Institute
- National Heart, Lung and Blood
- Centers for Disease Control and Prevention


Questions
?
?
Answers
?