



Public Health
DATA LEARNING CENTER

Introduction to Data Science for Public Health

Michelle Campbell, MSHI

Director, Center for Data Modernization and
Informatics, Washington State Department of Health

Stephen Elston, PhD

Principal Consultant,
Quantia Analytics LLC



Using Zoom Q&A



If you have a question during the presentation, please click the **Q&A icon** in the Zoom toolbar to open your Q&A Pod.

About the Public Health Data Learning Center



Public Health
DATA LEARNING CENTER



SCHOOL OF PUBLIC HEALTH
UNIVERSITY *of* WASHINGTON

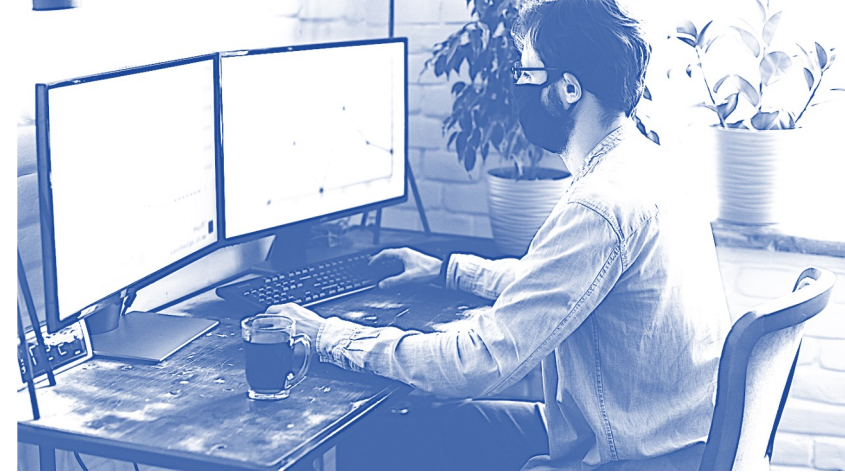


Introduction to Data Science

Part 1: Introduction to the Data Modernization Initiative

Data Modernization Initiative (DMI) Background

- COVID-19 response focused attention on challenges to public health surveillance systems
- Underfunding, outdated systems, lack of IT support hampers the ability to use and share data
 - Cumbersome processes for submitting and accessing data
 - Data spread across multiple systems that don't speak to each other
- Burnout, lack of workforce capacity intensifies challenges



DMI Goals

In line with the CDC's mission for DMI, to modernize data systems and improve the overall infrastructure for data management, governance, and analytics

- Creation of Center for Data Modernization and Informatics informed by DMI and gap analysis

DMI Goals

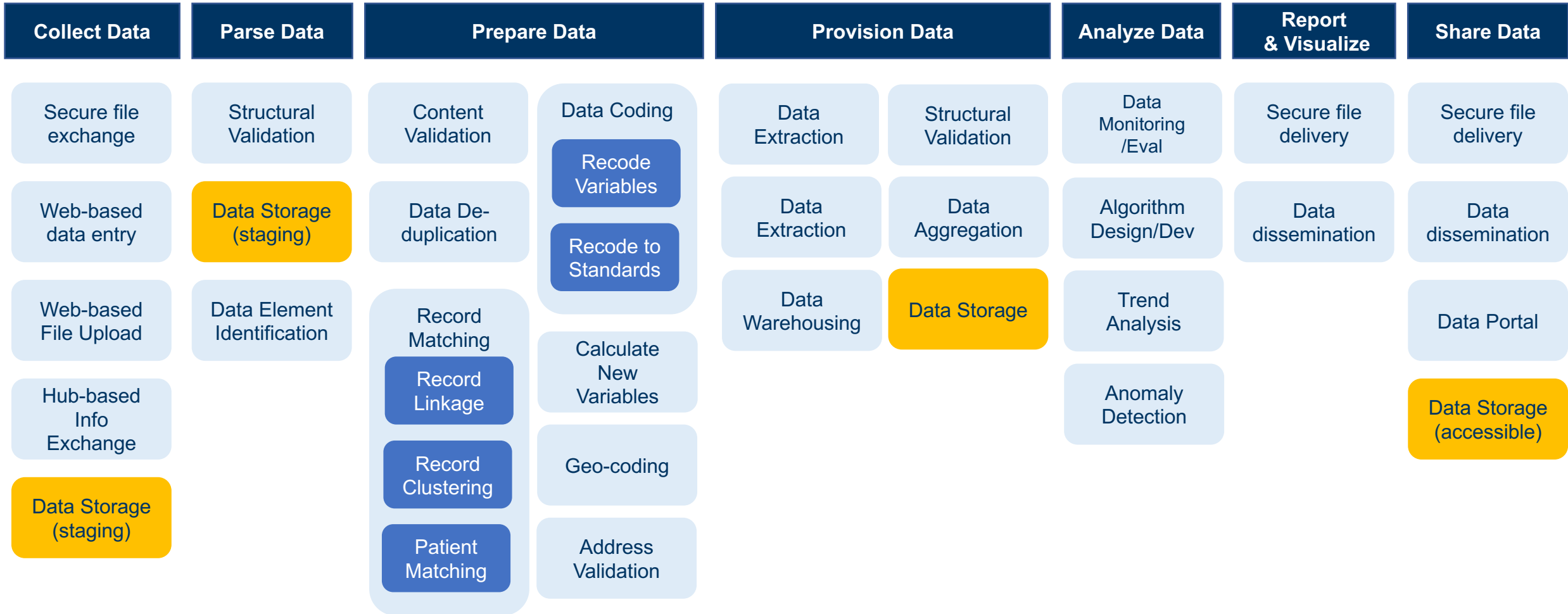
- Requires transitioning legacy, single-use, and siloed systems to reusable systems that support needs across the agency.
- Data access methods standardized; data available in a usable format to allow for longitudinal analysis
- Initial DMI workforce development plan centered on increasing data science knowledge and effectively communicating data.

Effects for the Public Health Workforce

- Public health practitioners can use and respond to public health data more quickly and accurately
- Public health surveillance systems are standards-based, coordinated, secure, and scalable to accommodate Washington's public health surveillance needs
- Staff can focus on higher-level analytical work without dealing with logistical hurdles to accessing data or the burden of data cleaning



Data Cycle – Public Health Surveillance Capabilities



Data Quality Management | Master Data Management

Data Governance | Information Governance

Grant Management | Funding Management

Data Cycle – Public Health Surveillance Capabilities

Capability	Epidemiology	Informatics	Data Governance	Information Technology	Enterprise Architecture
Secure File Exchange	A	C		R	
Web-based Data Entry	A,R	C		S	
Web-based File Upload	A,R	C		S	
Hub-Based Data Exchange	C	A,R		S	C
Data Storage (Staging)	I	C	I	A,R	C

Collect Data

Key	
R	Responsible
A	Accountable
S	Service
C	Consulted
I	Informed

Center for Data Science

Through the Data Modernization Initiative, we are working to:

- ✓ Advance public health data democratization, equity, and data-informed decisions for all Washingtonians
- ✓ Accelerate the findability, accessibility, interoperability and reusability of analytic solutions, tools, and products
- ✓ Expand Washington's public health capacity to visualize and share actionable insights to inform community-level decision making
- ✓ Democratize our data through advancing data governance and open data practices
- ✓ Strengthen core public health data for national notifiable conditions data and OHS operated data systems and applications

Reflection Questions

- How does/can your role play a part in the greater DMI effort?
- How does/can your department play a part in the greater DMI effort?
- What skills are needed within your organization to adapt to these changes?



Data Science for Public Health: From Insight to Action

Presented by: Stephen Elston

Prepared for UW School of Public Health | April 2023

Our Goal

Improve health outcomes for all people through greater insight and timely actionable results at all levels through the Data Modernization Initiative (DMI)

Three Big Questions for Today

1. What is data science?
2. How is the data science process applied?
3. How does using complex data provide greater insight?

The image features a person's hands typing on a laptop keyboard. The scene is overlaid with various digital graphics in a blue-tinted, semi-transparent style. On the left, there is a flowchart with several nodes and connecting lines. In the center and right, there are several data visualization elements: a line graph with a fluctuating blue line, a bar chart with multiple bars of varying heights, and a pie chart. The background is filled with faint, glowing binary code (0s and 1s) and hexadecimal characters (A-F, 0-9). The overall aesthetic is futuristic and data-oriented.

Part 1: What is Data Science?

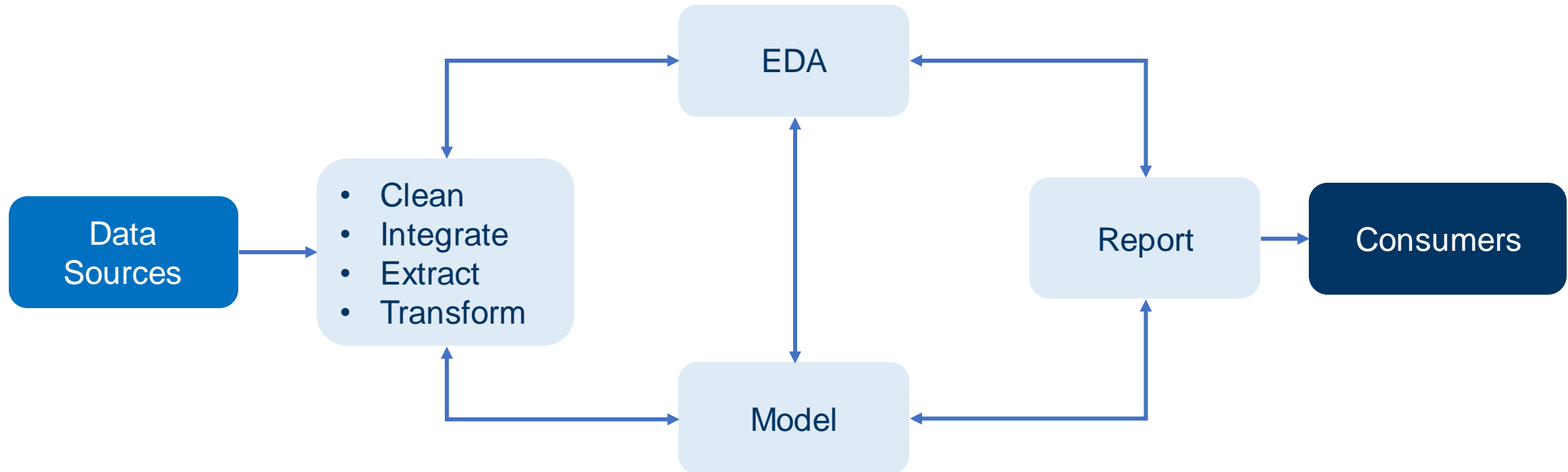
What is Data Science? Many Views



Source: [Microsoft Power BI](#)

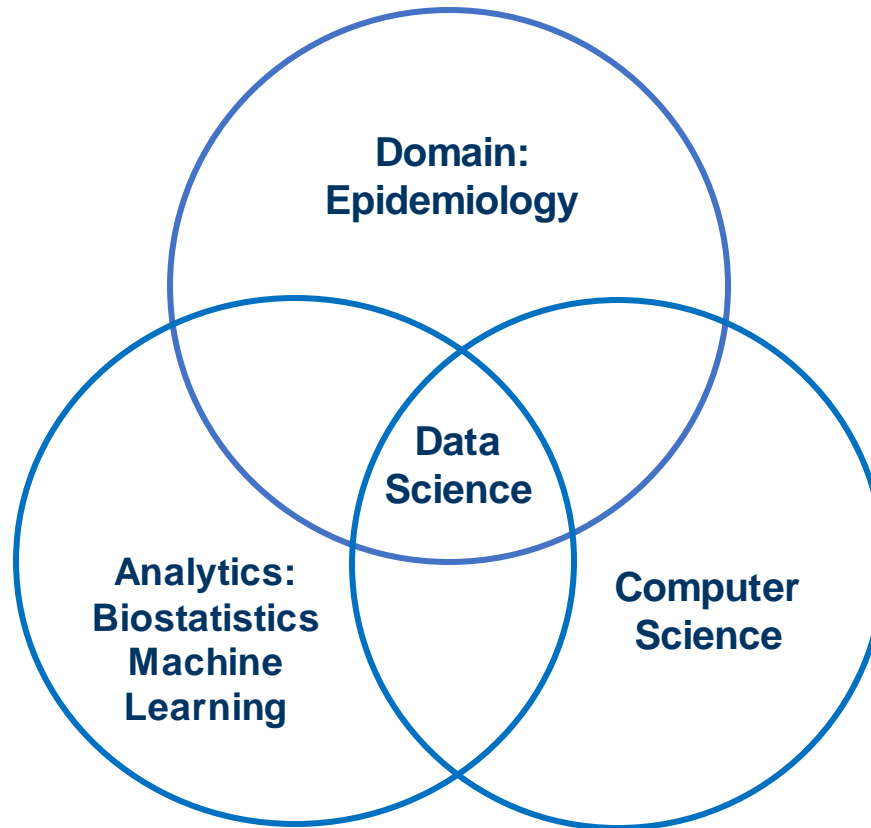
View 1: [Data science](#) is the integrated, multi-disciplinary practice of extracting **meaning** and **actionable insight** from data

What is Data Science? Many Views



View 2: Data science **unifies** statistics, data analysis, data visualization, machine learning, and related methods to understand phenomena with data

What is Data Science? Many Views



View 3: Data science **unifies** statistics, data analysis, data visualization, machine learning, programming and domain expertise

What is Data Science? Many Views

The **three Vs** of data science

- **Variety**: Data scientists create models **integrating many types of data** to provide in-depth insight
- **Velocity**: Production data science pipelines take in and process data with **low latency**, delivering timely results
- **Volume**: Data science pipelines **distill massive quantities** of data to actionable results

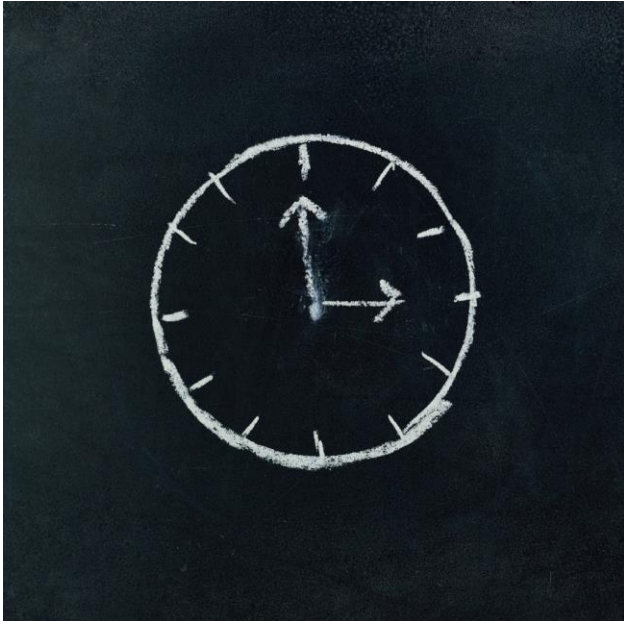
How is Data Science Different? 3 Differences



Difference 1: Data science is forward-looking

- Predictive models lead to **action**
- Action is fundamental to **data driven organizations**

How is Data Science Different? 3 Differences



Difference 2: Data science enables timely, data-driven decisions

- Data-driven organizations are grounded in **understanding at all levels**
- Models provide **understandable, trustworthy results**
- Decision-makers can receive **timely information**

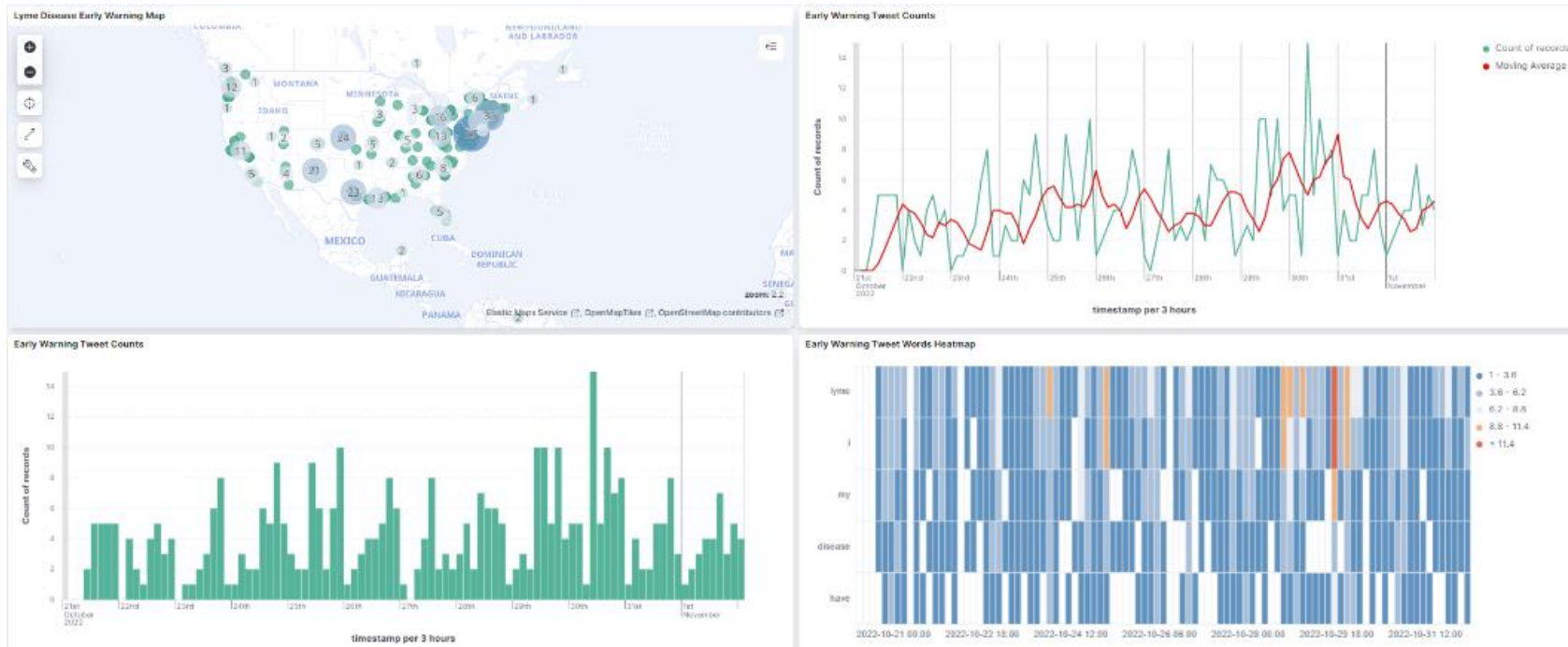
How is Data Science Different? 3 Differences



Difference 3: Data science integrates complex data

- Multidimensional views of data enable **deeper insight**
- Results are **trustworthy** and **traceable** to sources

Example: Lyme Disease Tracking




Project by [Muthuramalingam, Yi, and Yin, 2022](#), for Centre de Recherches Mathematiques and Harvard University, 2022

Goal: Create a forward-looking model to predict outbreaks

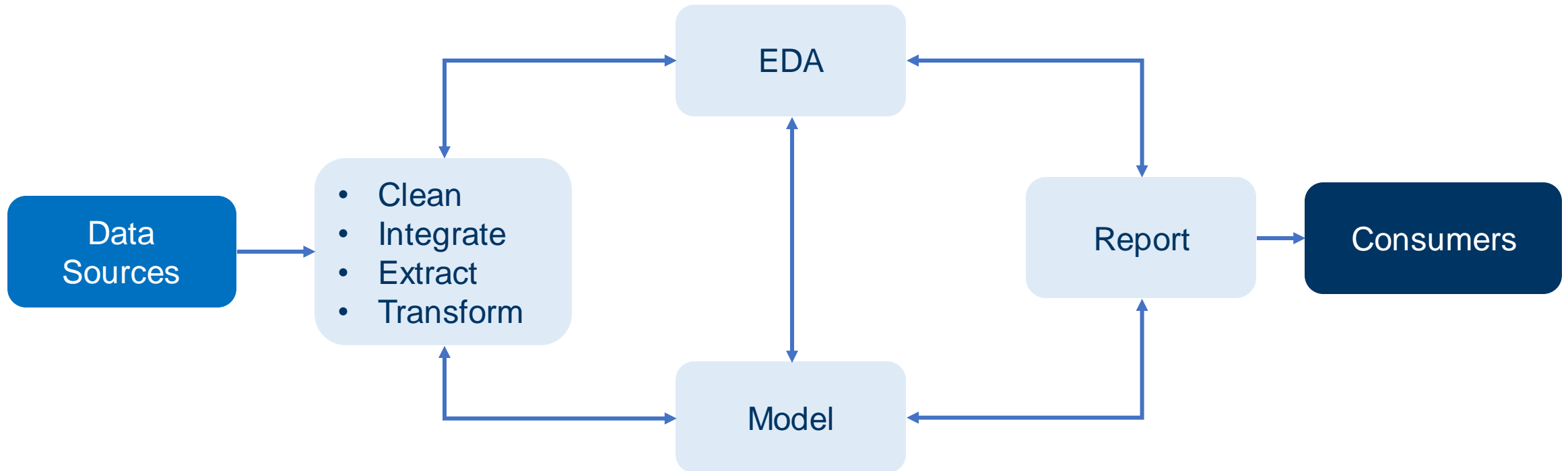
Function: Integrate clinical reports with social media posts; offer multiple views of disease spread

Result: Provide timely information source for decision-makers



Part 2: How can we apply the data science process?

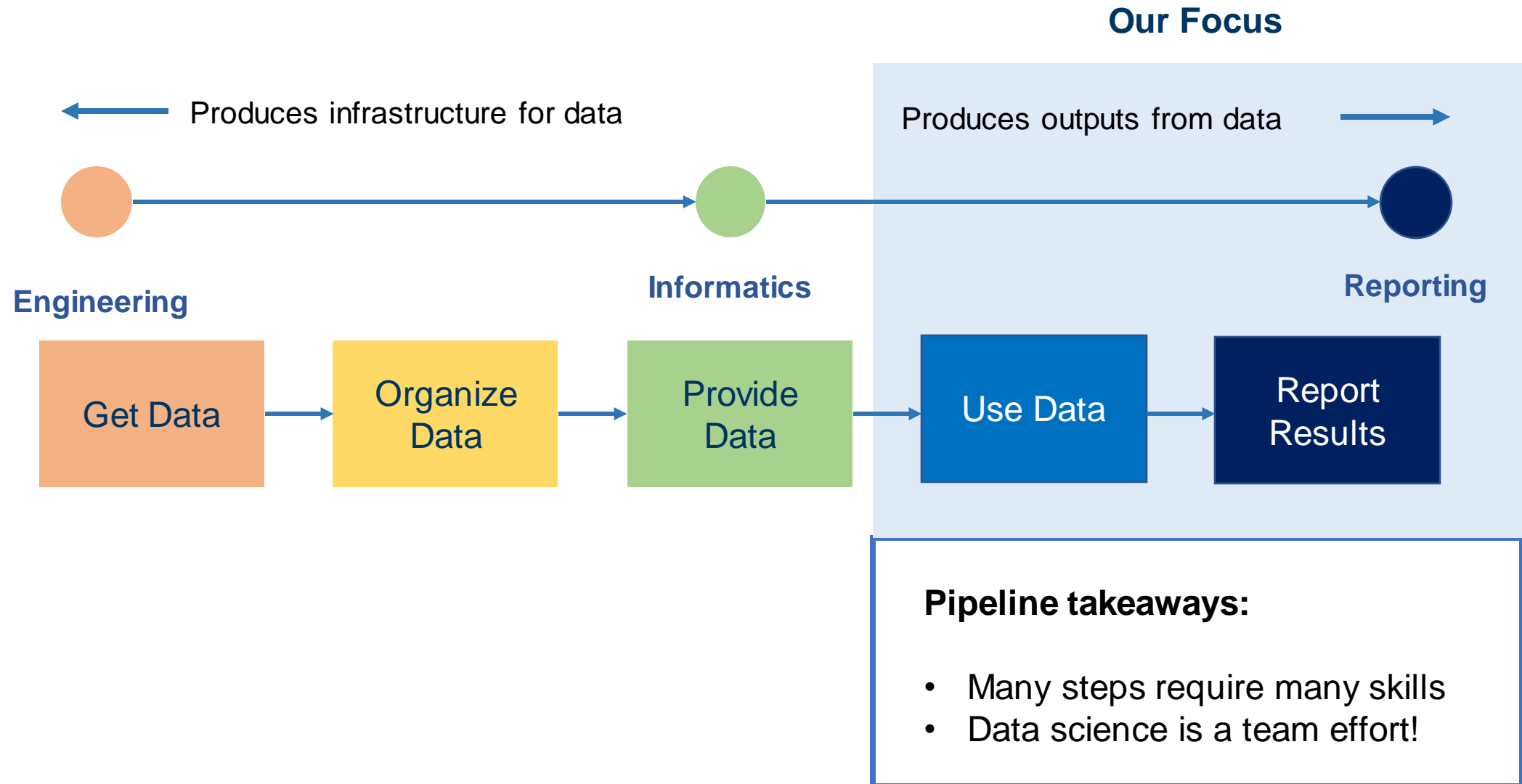
The Data Science Process



Data science is an **iterative process**. Does this process look familiar?

- **Elements are common to most analytics processes**
- **You are probably doing data science!**

The Data Science Pipeline

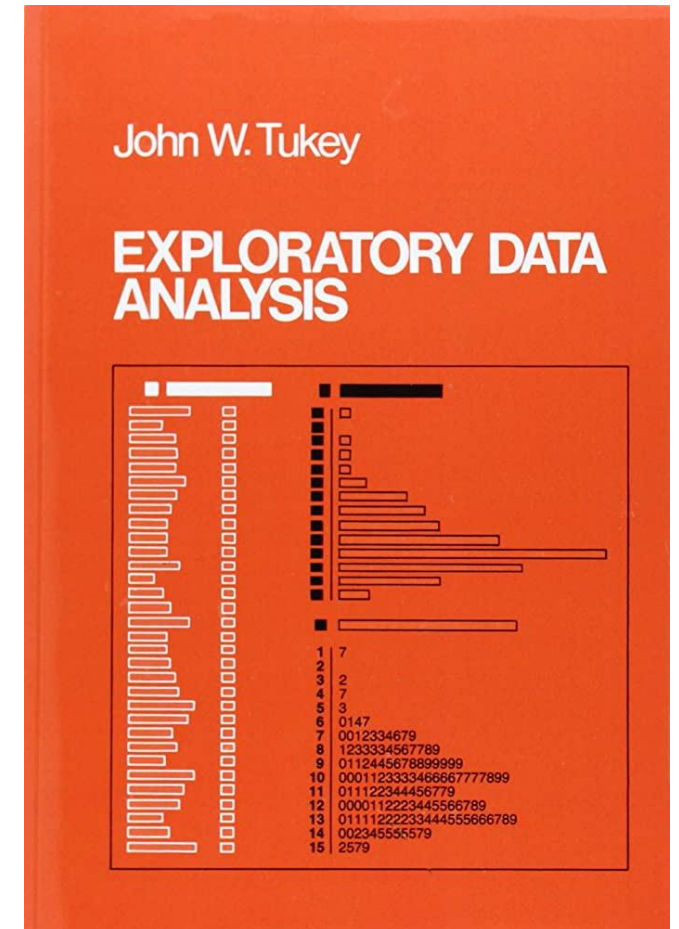


Exploratory Data Analysis and Presentation

Exploratory data analysis (EDA) builds understanding of relationships in complex data sets, using summary statistics and visualization

About EDA:

- Use begins in 18th and 19th centuries
- Well-established branch of data analysis
- Applied at **every stage of the data science process**
- Inherently **iterative process**



Modern view of Exploratory Data Analysis (EDA)
introduced by John Tukey, 1977

Exploratory Data Analysis and Presentation

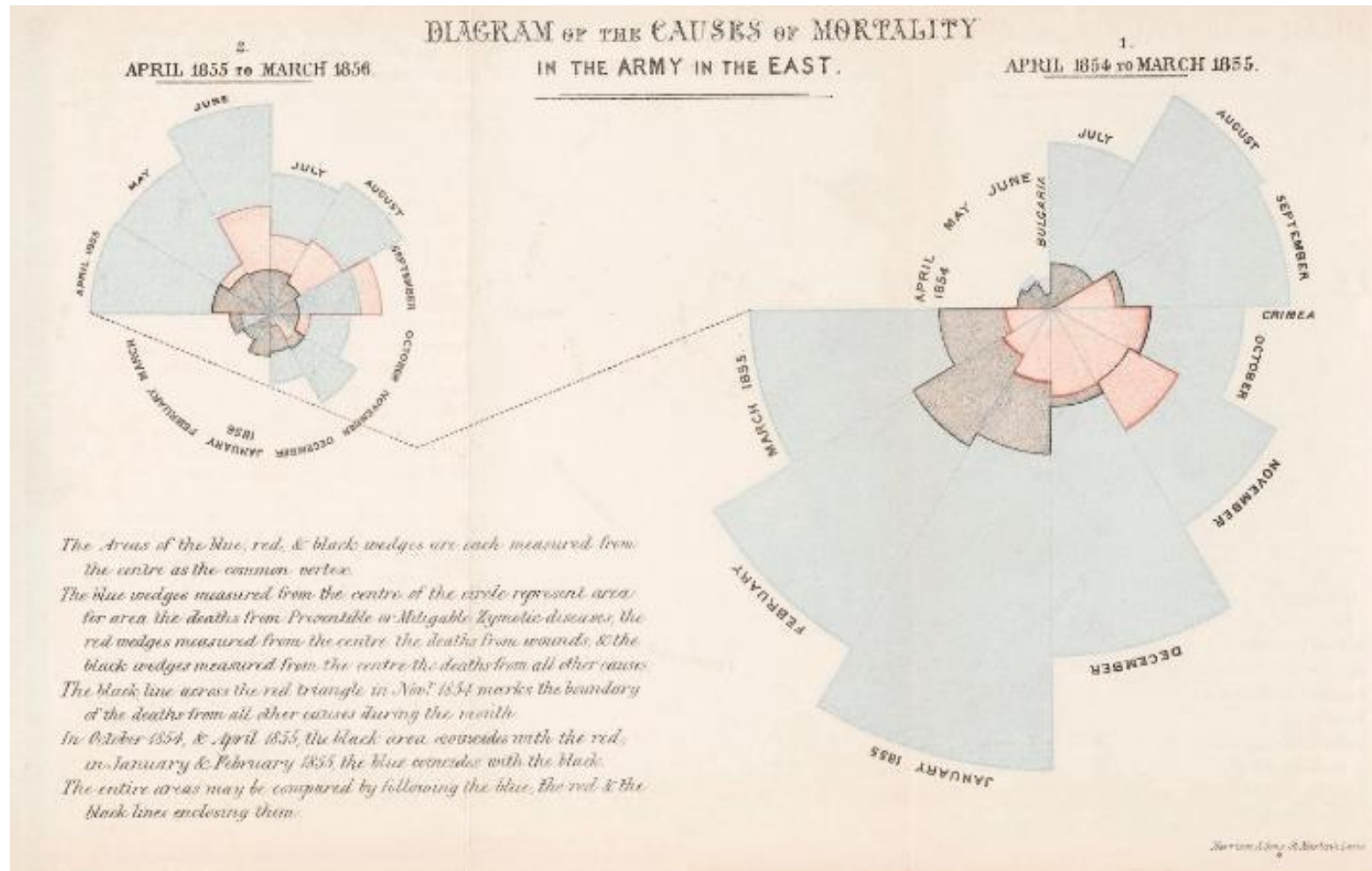
Exploratory data analysis (EDA) builds understanding of relationships in complex data sets, using summary statistics and visualization

EDA builds understanding of complex relationships:

- Summary statistics and simple models
 - Summarize key metrics of data
 - Use as measures of importance
 - Evaluate impact of actions
- Visualization of key relationships



Example: Exploratory Data Analysis



Source:

[Florence Nightingale \(1820 – 1910\): An Unexpected Master of Data, Bradshaw, 2020](#)

An early example of public health EDA - Florence Nightingale, 1859

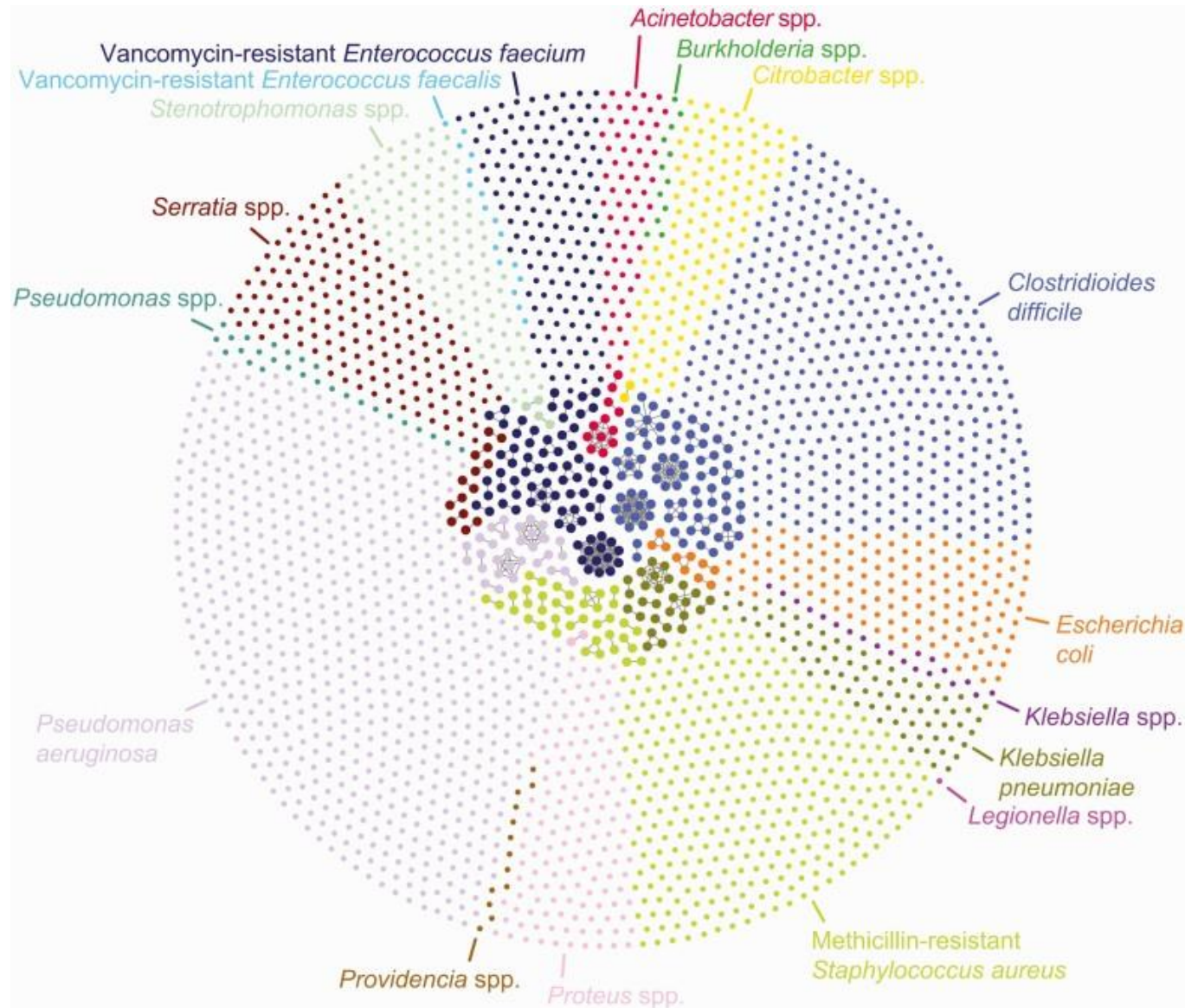
Exploratory Data Analysis and Presentation



Benefits of visualization:

- Most people have **excellent visual perception**
- Data scientists employ **visualization methods for complex data**
- Visualization **builds understanding** of relationships in data
- A few good visualizations are a powerful way **to communicate your insights**

Example: State-of-the-art example – Healthcare outbreak detection



Source:

Whole-Genome Sequencing Surveillance and Machine Learning of the Electronic Health Record for Enhanced Healthcare Outbreak Detection, [Sunderman, et.al., 2022](#).

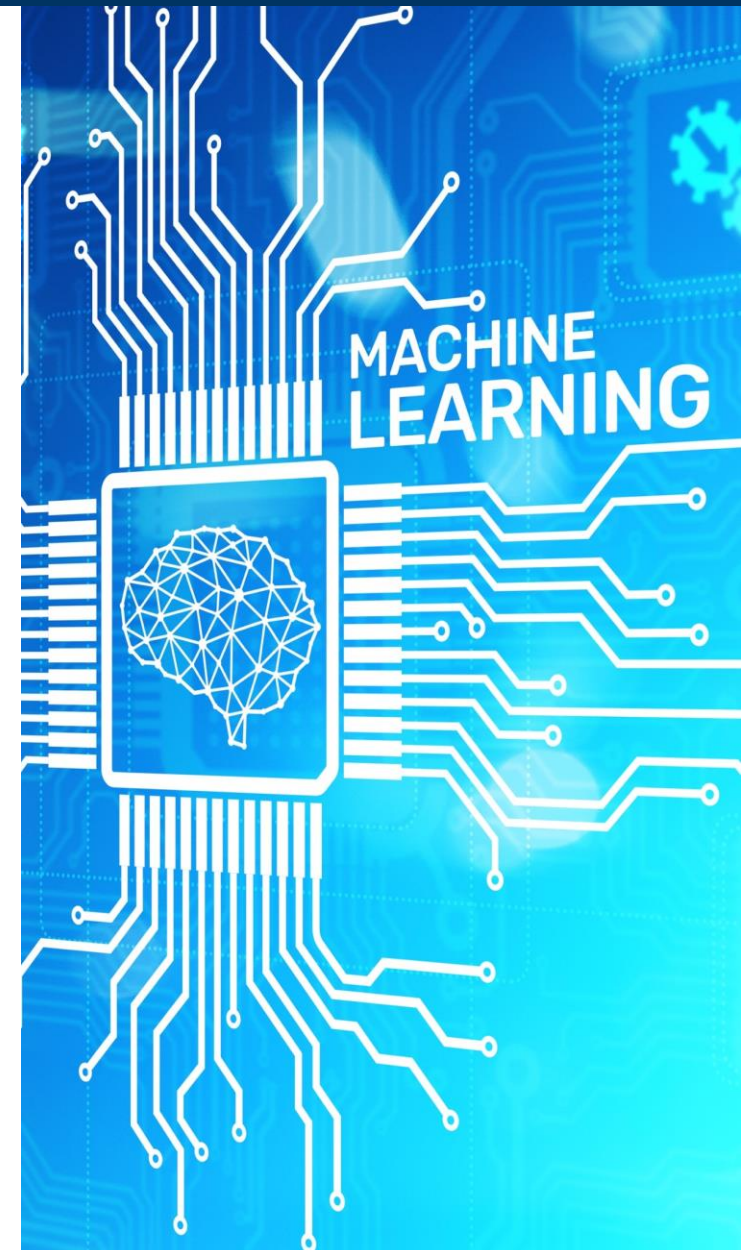
Predictive Analytics: Machine Learning View

Definition 1: Machine learning is the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.

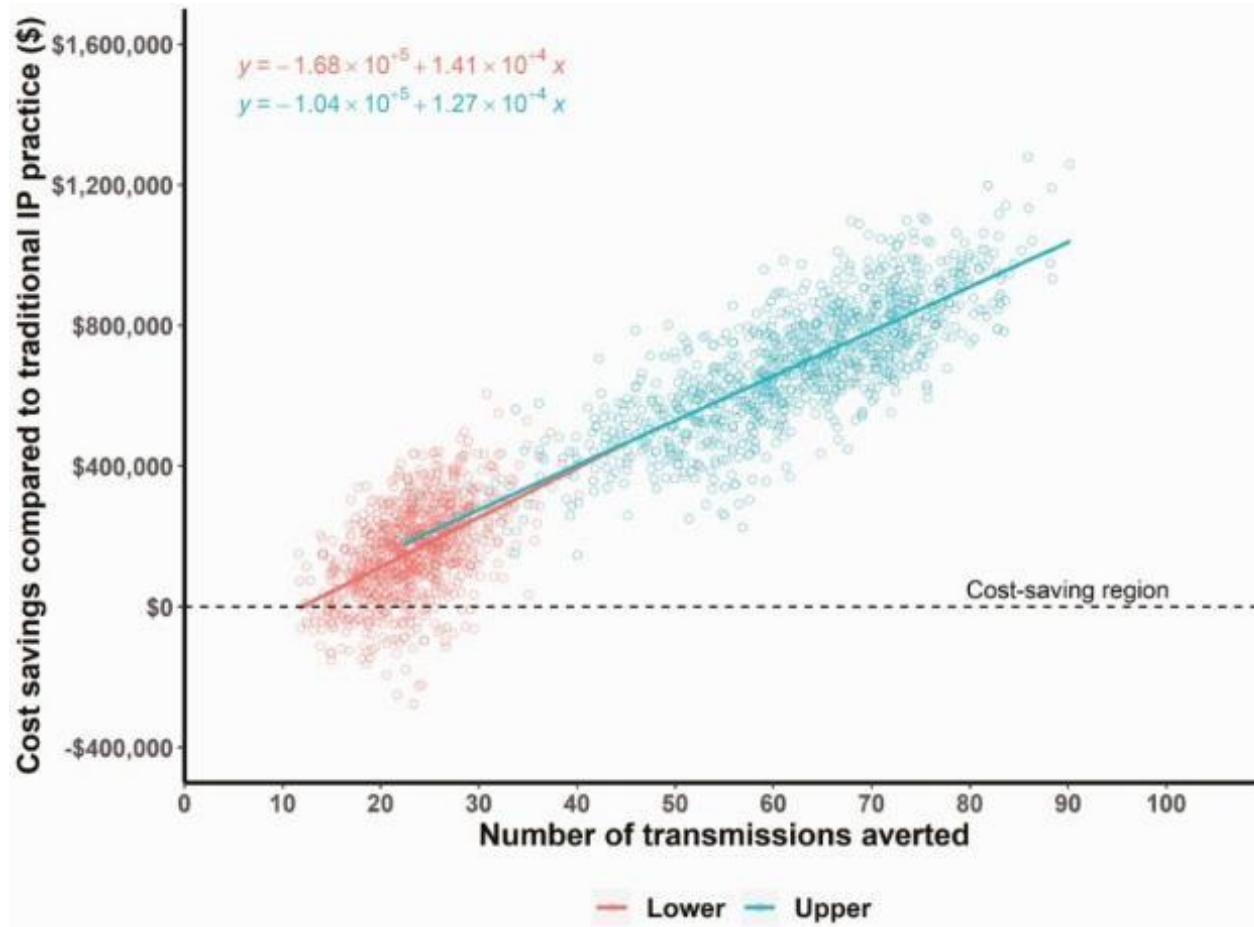
– Oxford English Dictionary

About Machine learning:

- Models are **forward looking, predictive**
- Trustworthy predictions are actionable



Example: Predictive Machine Learning — Regression to predict cost savings



Source:

Whole-Genome Sequencing Surveillance and Machine Learning of the Electronic Health Record for Enhanced Healthcare Outbreak Detection, [Sunderman, et.al., 2022](#).

Predictive Analytics: Machine Learning View

Definition 2: Machine learning models learn a function to map input features to predicted labels.

Multiple origins of ML lead to confusing terminology

- Statistics
- Computer science
- Engineering
- Economics

Label	Feature
Response	Predictor
Dependent	Independent
Outcome	Explanatory
Endogenous	Design
	Exogenous

Examples of typical input features and predicted labels

Predictive Analytics: Machine Learning View



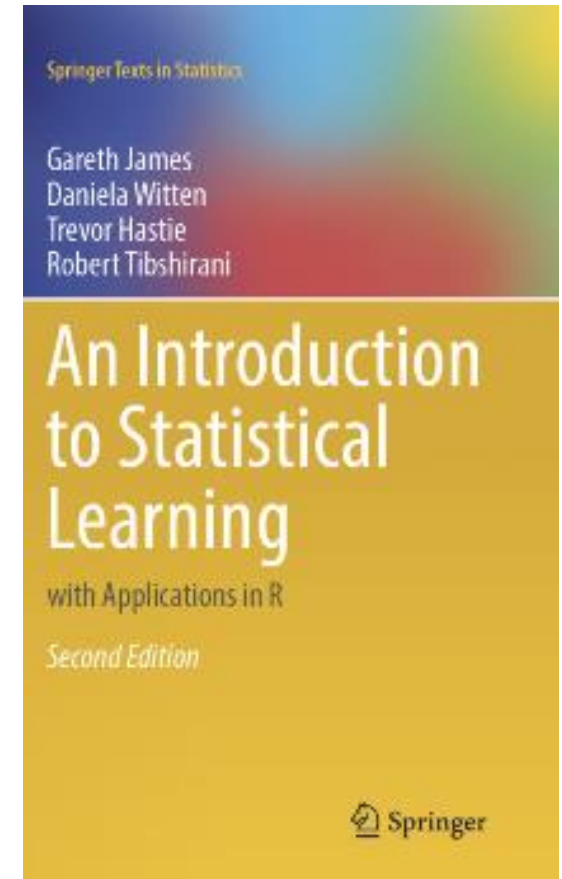
Two major categories of machine learning models

- **Supervised machine learning** train (fit) models by learning from known responses or **labeled cases**
 - Linear regression, logistic regression; familiar examples
 - Many other choices
- **Unsupervised machine learning** train (fit) models by learning from unlabeled data
 - Clustering models, familiar example

Predictive Analytics: Machine Learning View

Zoo of supervised machine learning models

- **Linear models** – regression, logistic regression
- **Tree models** – ensembles of tree models produce state of the art classifier performance
- **Support vector machines (SVM)** – Linear and nonlinear
- **Neural networks** – Complex highly nonlinear models
- Many more



An overview of the field of statistical learning focusing on R applications

Pitfalls of Machine Learning

Overfitting

- Massive number of **features** \Rightarrow **high capacity model**
- High capacity models **learn training data too well**
- **Sparse models**, regularization methods find minimal models

Pitfalls of Machine Learning

Training data problems:

- Biased samples
- Missing data
- Outliers
- Errors

Pitfalls of Machine Learning

Unbalanced cases


- Models with **rare label cases** lead to **poor predictions**
- *Example: Specific medical condition is generally **rare***
- Methods to **balance cases**: imputation, stratified sampling, etc.

Building Your Data Science Toolbox



Complex problems require a good toolbox

- Data science is a **team effort!**
 - Effective teams have **complementary skills**
- If you are doing analytical work, **you have a data science toolbox!**
- **Learn tools incrementally:** Excel, R, SAS, Python, SQL
- Use familiar methods/algorithms as ramp to learn new ones
 - ML algorithms in related families

The background features a blurred city skyline at sunset, with warm orange and yellow light. Overlaid on this are several white-outlined icons representing different data visualizations: a line graph with multiple lines, a bar chart, an atomic model, a circular gauge showing 100% and 45%, a radar chart, a bar chart with vertical error bars, and a line graph with a single line. The text is centered in a black rectangular box.

Part 3: How do we get insight from complex data?

Producing Actionable Results

To add value, deliver actionable timely results to decision makers.

- **Actionable results:**

- Address a specific problem
- Are trustworthy and traceable
- Are understandable, and explainable

- **Timely Results:**

- Available on demand



Producing Actionable Results

Presenting data science results through web/dashboards

- **Clear presentation**
 - **Uncluttered:** Limit information on page
 - **Clear charts:** no “chart junk”!
 - **Small tables:** limited human perception of tabular data
- **Explorable**
 - **Multiple views for greater insight**
 - **Drill down:** understand details and trace data sources



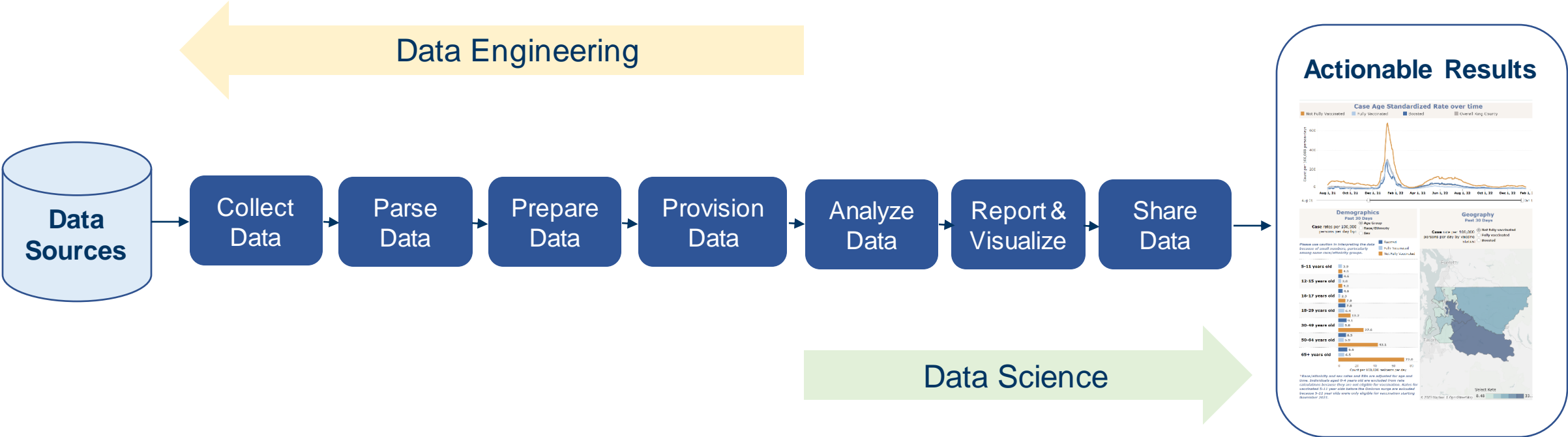
Why Integrate Complex Data?

Understanding complex problems requires rich, multidimensional, data

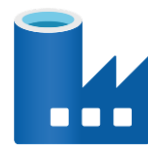
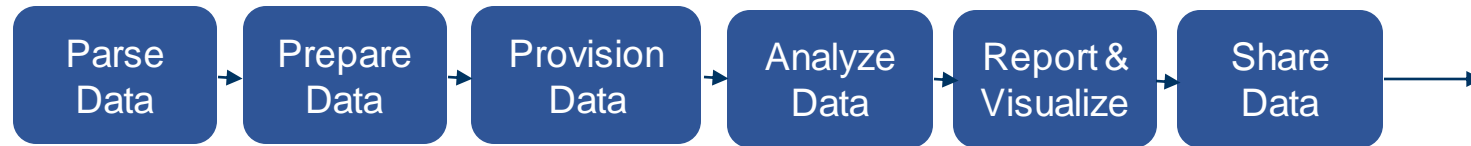
- Public health problems are complex!
- Analysis of complex problems requires integrating data
- Integrated data enables multiple views of complex problems
 - *Example; views in time, space, demographics, test results, etc.*
- Exploration of complex data leads to **deep insight**
- Machine learning models **learn from complex data**



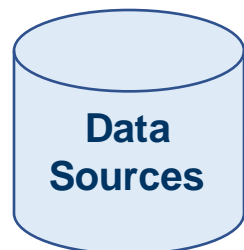
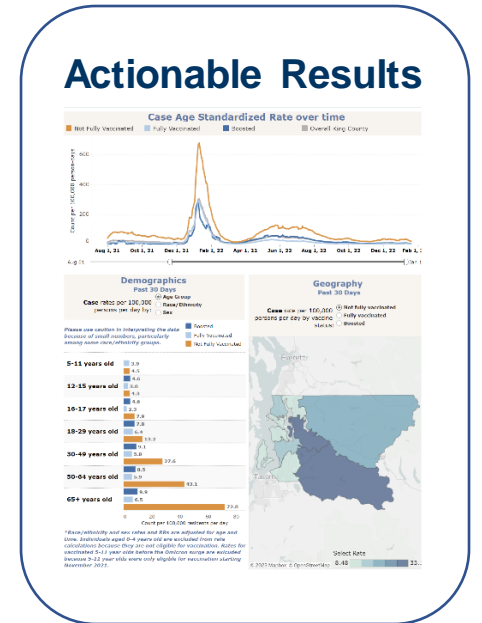
What is An “Integrated Data Process?”



Example: Vision for Integrated Data Platform



Azure Synapse Analytics

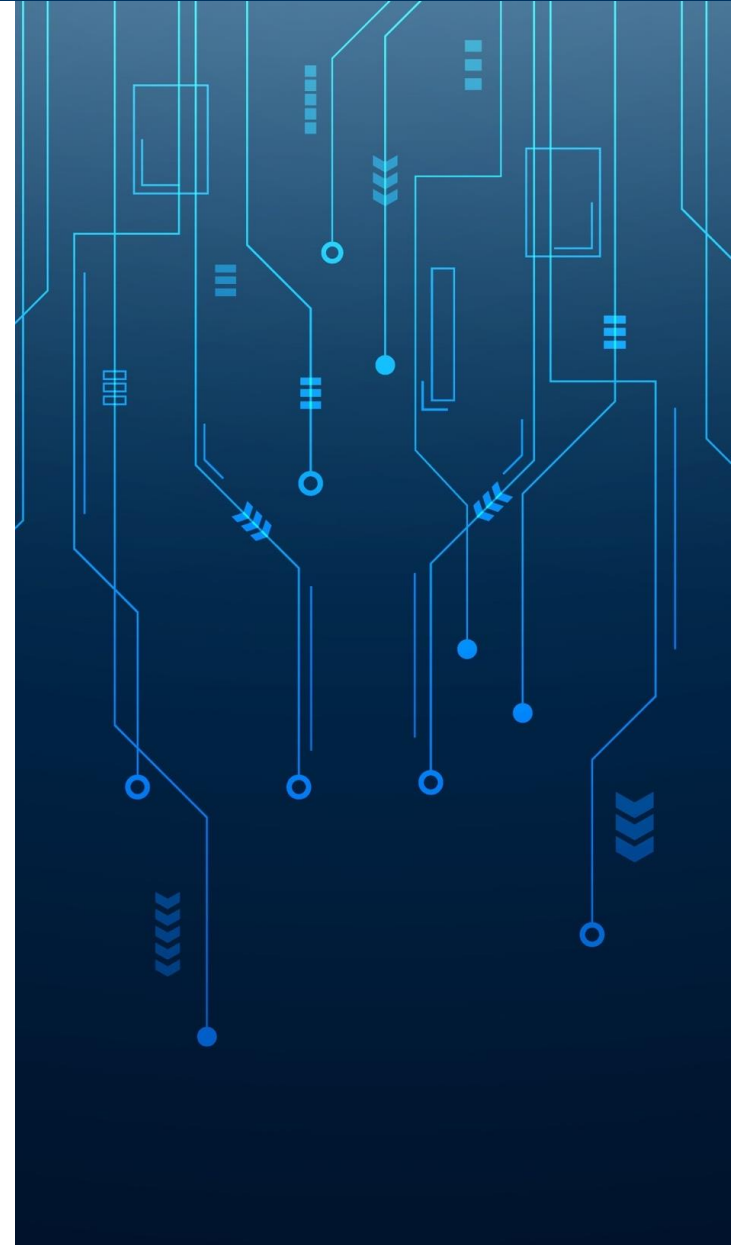


Data Lake

Data Integration is Difficult

Many pitfalls in data integration

- Management of data at massive scale
- Different coding
- Time and space mismatches
- Lack of common key
- Maintaining traceability and reliability
- Ensuring de-identification
- Access authorizations



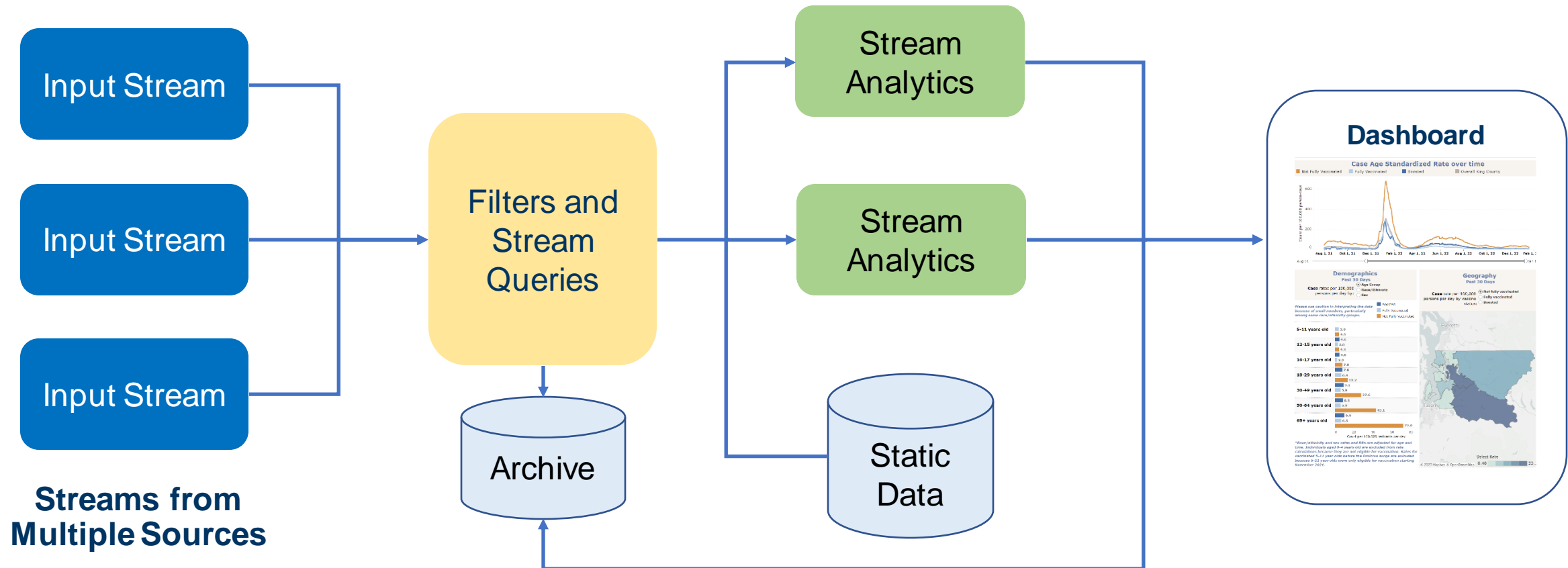
Why Streaming Analytics?

Streaming analytics deliver timely results

- Streaming data **arrive over time**
 - Multiple streaming sources
 - Sampling times often differ
- **Refresh results**
 - On new value
 - At time interval
- Streaming analytics integrate static data



Streaming Analytics



Example Streaming Analytics pipeline

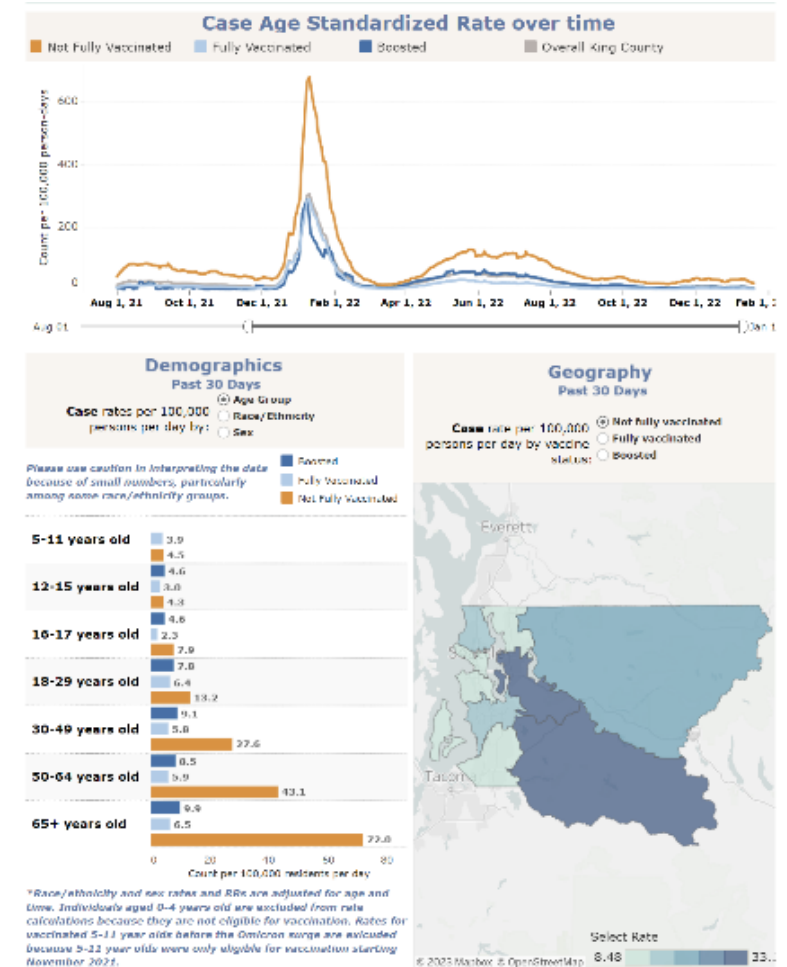
Example: Streaming Analytics

Goal: Use streaming data to enable timely updates of analytics

Function: Dynamically update dashboard as new case data arrive (5 day lag)

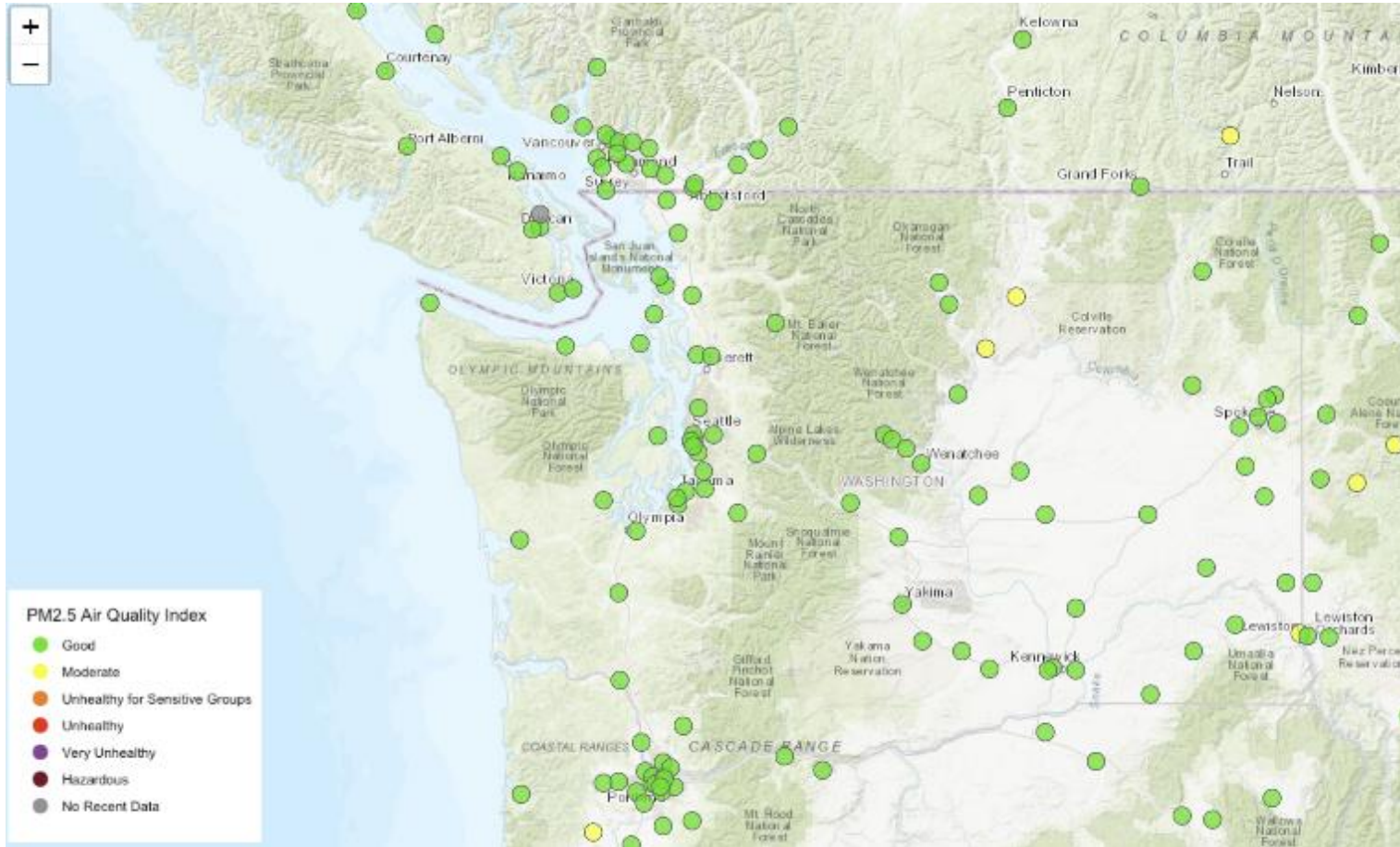
Result: Offer multiple views of vaccination status updates on dashboard:

- Temporal
- Geospatial
- Stratified



Source: [King County Outcomes by Vaccination Status Dashboard](#)

Example: Interactive AQI map using Streaming Analytics



Source:

[Washington Department of Ecology](#)

Key Points

Data science opens new exciting opportunities for **data driven** public health improvements.

Q: What is data science?

A: Data science is the integrated, multi-disciplinary practice of extracting meaning and providing **timely actionable insight** from data

Key Points

Data science opens new exciting opportunities for **data driven** public health improvements

Q: How is the data science process applied?

A: The iterative data science process explores data, builds models, and achieves **deep insight** for complex problems

Key Points

Data science opens new exciting opportunities for **data driven** public health improvements

Q: How does using complex data provide greater insight?

A: Complex public health problems are addressed by integrating complex data

Your Data Science Journey

- If you are doing analytic work, **you are doing data science!**
- Many resources to help you build your data science skills (*see Resources*)
- No one has all the required skills – it's a **team effort!**

QUESTIONS?



Q&A

To ask a question, please click the **Q&A icon** in the Zoom toolbar to open your Q&A Pod.

Appendix: Key Skill Areas

Key areas for building data science skills include:

- **R** – Widely used language with large number of statistical and machining packages
- **Python** – Primary language used for machine learning
- **SQL** – The language of data storage and access; *essential for data science after all!*
- **Visualization and dashboards** – visualization for effective presentation of results

Appendix: R Resources

R – Widely used language with large number of statistical and machining packages

- Getting started in DataCamp: [Introduction to R](#)
- Manipulating data with R in DataCamp: [Reshaping Data with tidyr](#)
- Building machine learning pipelines in R in DataCamp: [Machine Learning with Caret in R](#)
- Good source for data manipulation with R: [Tidy Modeling with R: A Framework for Modeling in the Tidyverse, Max Kuhn, Julia Silge](#)

Appendix: Python Resources

Python – Primary language used for machine learning

- Getting started in DataCamp: [Introduction to Python](#)
- Introduction to data manipulation with Python in DataCamp: [Data Manipulation with Pandas](#)
- Basic predictive analytics with Python in DataCamp: [Introduction to predictive analytics with Python](#)
- The primary source for data manipulation with Python: [Python for Data Analysis: Data Wrangling, with pandas, numpy and Jupyter, Wes McKinney](#)

Appendix: SQL Resources

SQL – The language of data storage and access; *essential for data science after all!*

- Getting started with DataCamp: [Introduction to SQL](#)
- Getting started with SQL server for Azure with DataCamp: [Introduction to SQL Server](#)
- For advanced users, the data management language for Azure Data Synapse: [Quick start on U-SQL](#)
- Advanced queries with R and Python on Azure: [Query data in Azure Synapse Analytics](#)

Appendix: Visualization and Dashboards

Visualization and dashboards – visualization for effective presentation of results

- Data visualization using R with DataCamp: [Intermediate Data Visualization with ggplot2](#)
- Data visualization using Python with DataCamp: [Intermediate Data Visualization with Seaborn](#)
- Introduction to PowerBI with DataCamp: [Introduction to PowerBI](#)
- Building dashboard with Microsoft Learning: [Getting started with building with PowerBI](#)
- Introduction to Tableau with DataCamp: [Introduction to Tableau](#)

Appendix: Local Learning Opportunities

The University of Washington Continuum College offers adult learners local online certificates, including:

- [Statistical Programming with R](#)
- [Applied Biostatistics](#)
- [Data Analytics: Techniques for Decision Making](#)
- [Data Science](#)
- [Data Visualization](#)
- [Machine Learning](#)
- [Big Data Technologies](#)