Building a BH Data Repository: Tips and Tricks from the Trenches

Jessie Tenenbaum, PhD
Associate Professor
Duke University School of Medicine

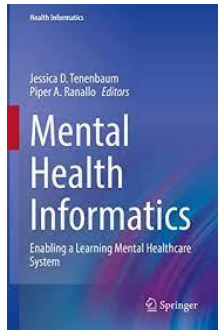
Chief Data Officer,
NC Dept of Health and Human Services

September 13, 2023

About Me

- **Researcher (2007-present)**
  - Duke SoM, Department of Biostats & Bioinformatics
  - Informatics to enable precision medicine in mental health

- **Chief Data Officer, NC DHHS (2019-present)**
  - Departmental data strategy
  - Data Office 4 pillars



- **Textbook Co-editor**

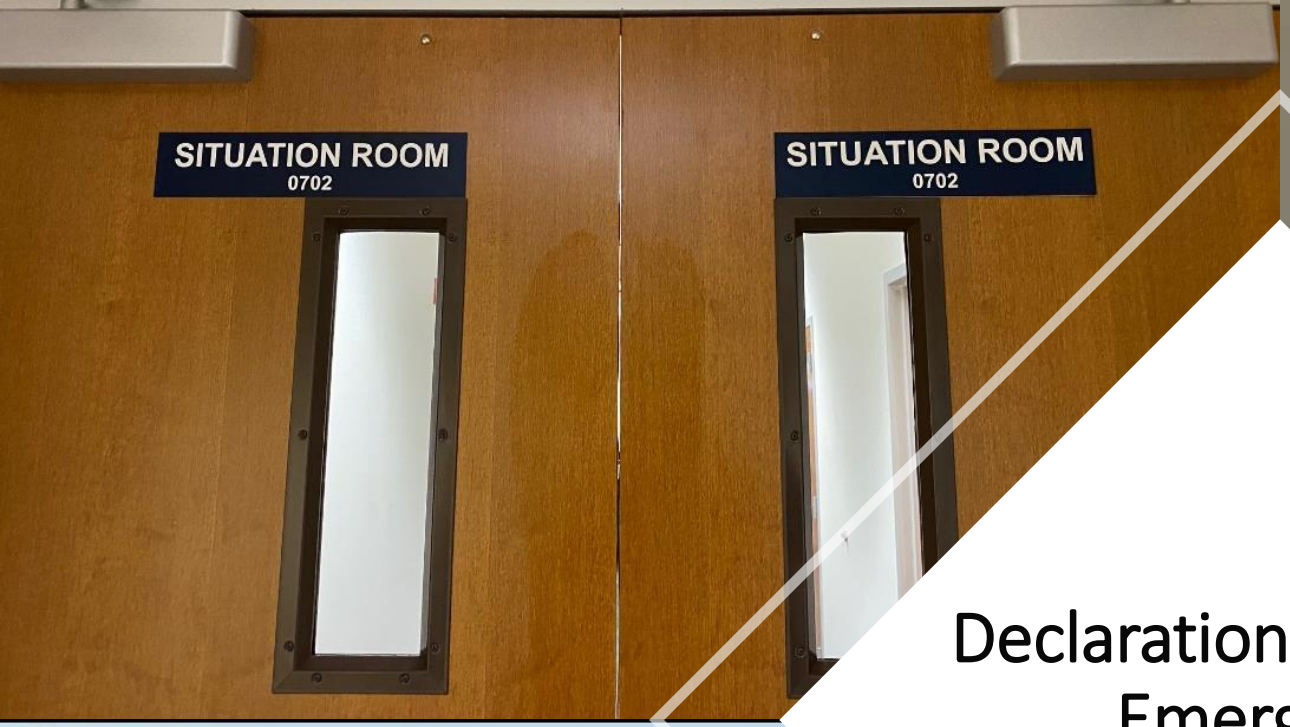
A tale of 3 repositories

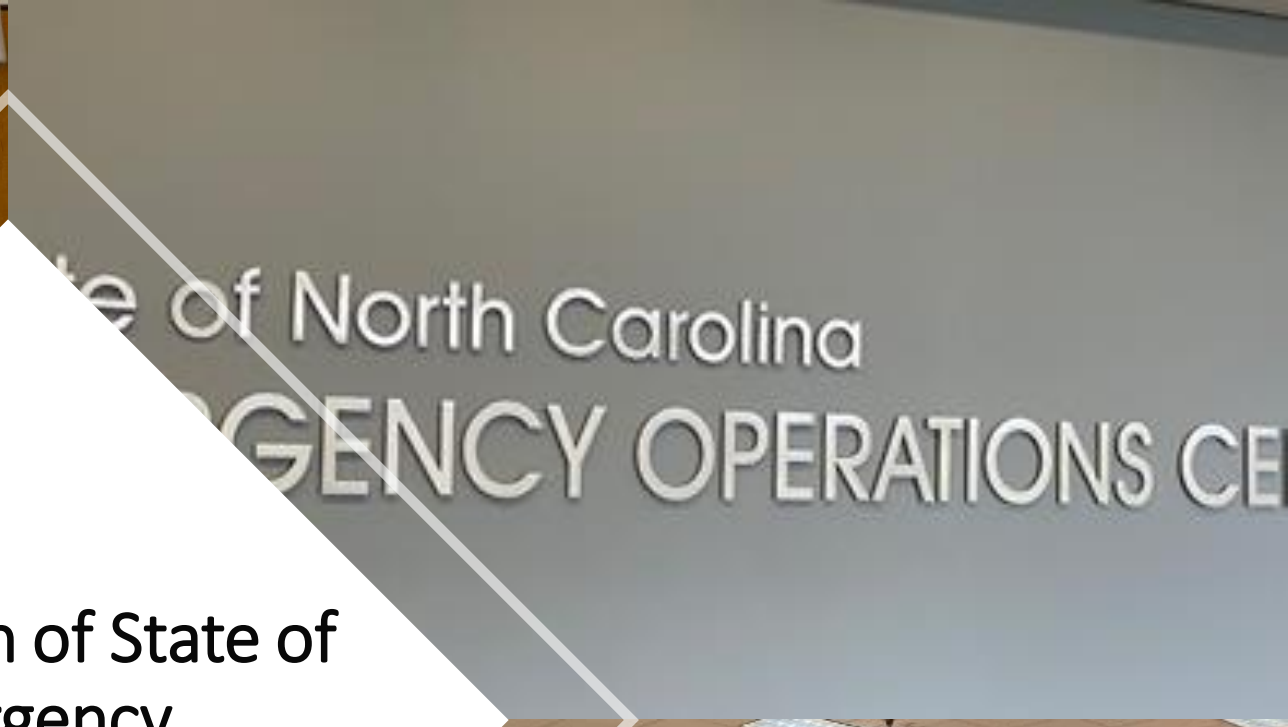
- **COVID Data**
- **NC DHHS Whole Person Health**
- **NC Longitudinal Data System**

Data Repository #1: COVID

COVID Data Repository

- **Purpose and structure:**
  - **Streamline and automate COVID reporting**
- **Process used to implement data repository system**
  - **Cloud-based**
  - **Speed- stood up in 48 hours**
- **Management structure**
  - **Inter-divisional data ownership with central data custodian**
- **Outcome of efforts**
  - **BIDP: Business Intelligence Data Platform**





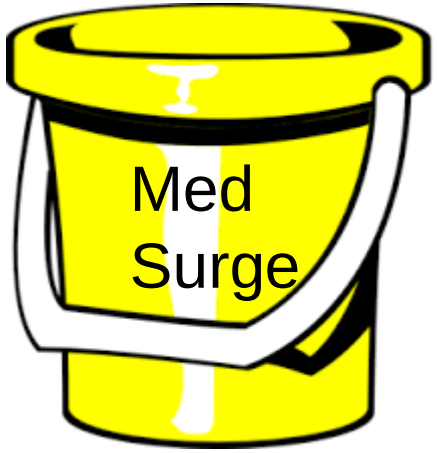
Declaration of State of Emergency

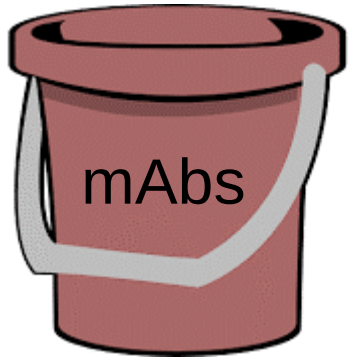
March 10, 2020

Early Actionable Questions to Enable Data Driven Policy

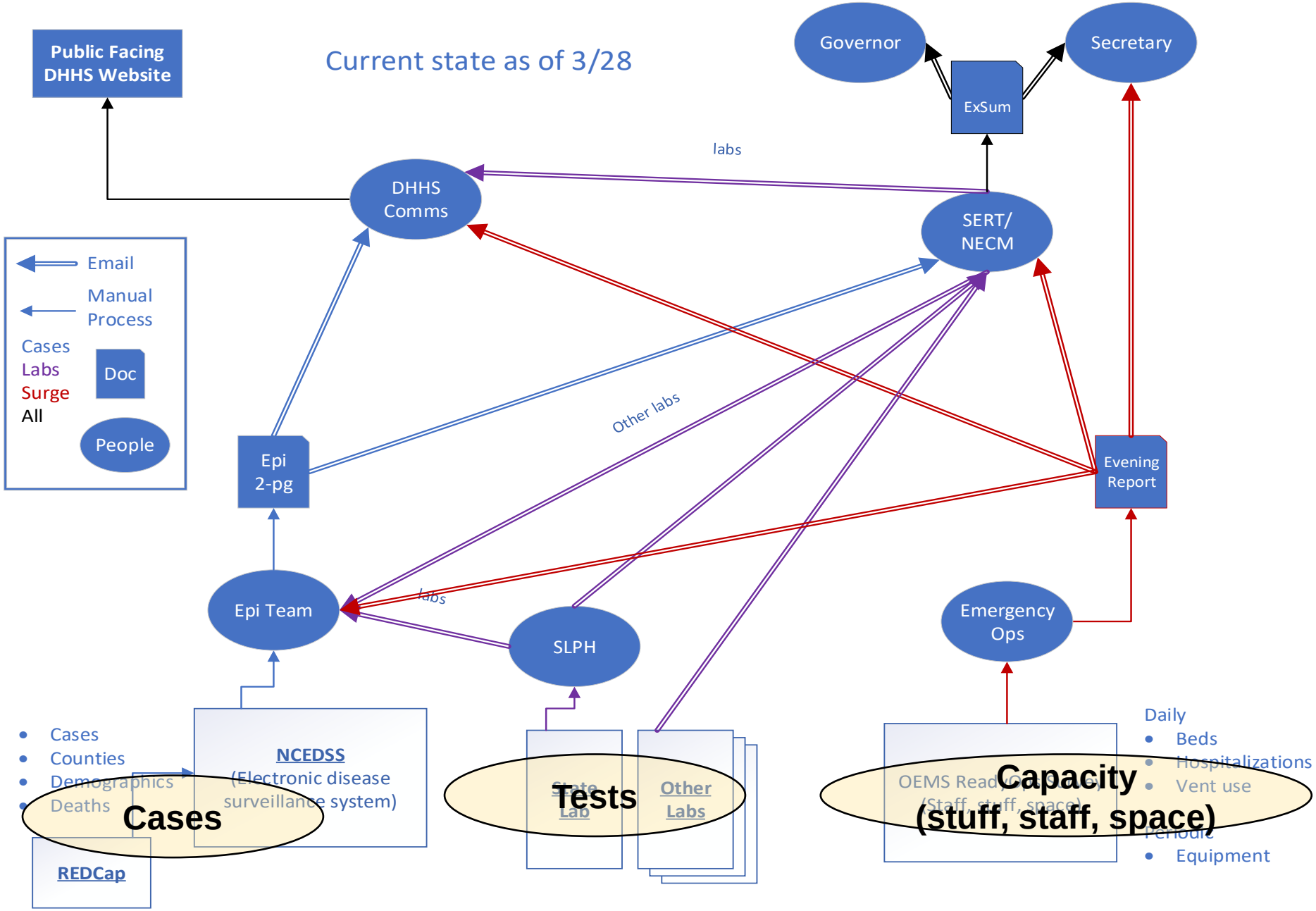
- How many cases will we see? When will our “epi curve” peak?
- How much (extra) PPE is needed, and where can we get it from?
- Will we run out of hospital beds? ICU beds? Ventilators?
- Should we shut down bars and restaurants? Schools? Businesses?
- How can we support families who are unable to work (either because workplace is shut down, or childcare is unavailable)?

Buckets of NC DHHS COVID-19 Data

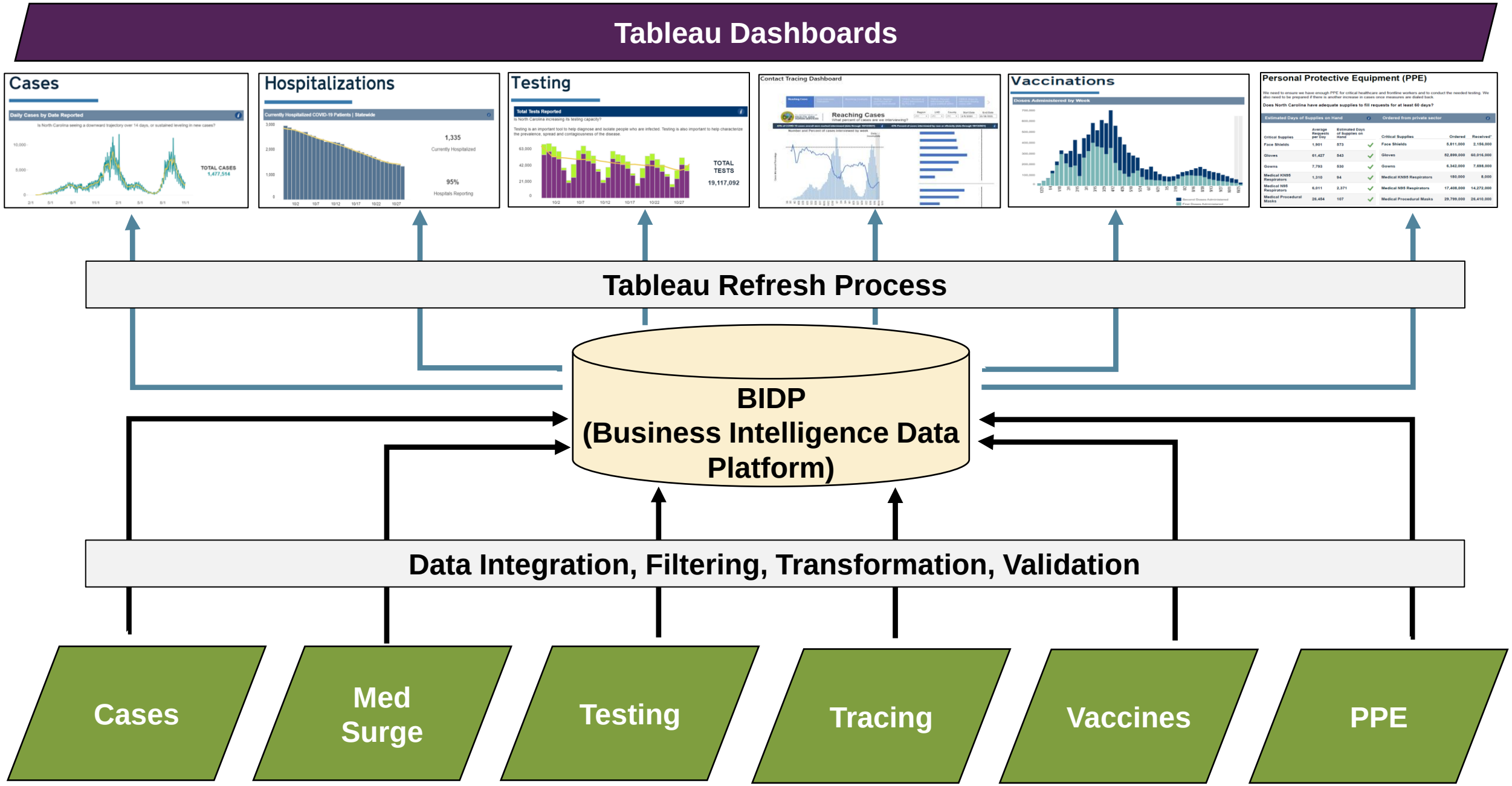


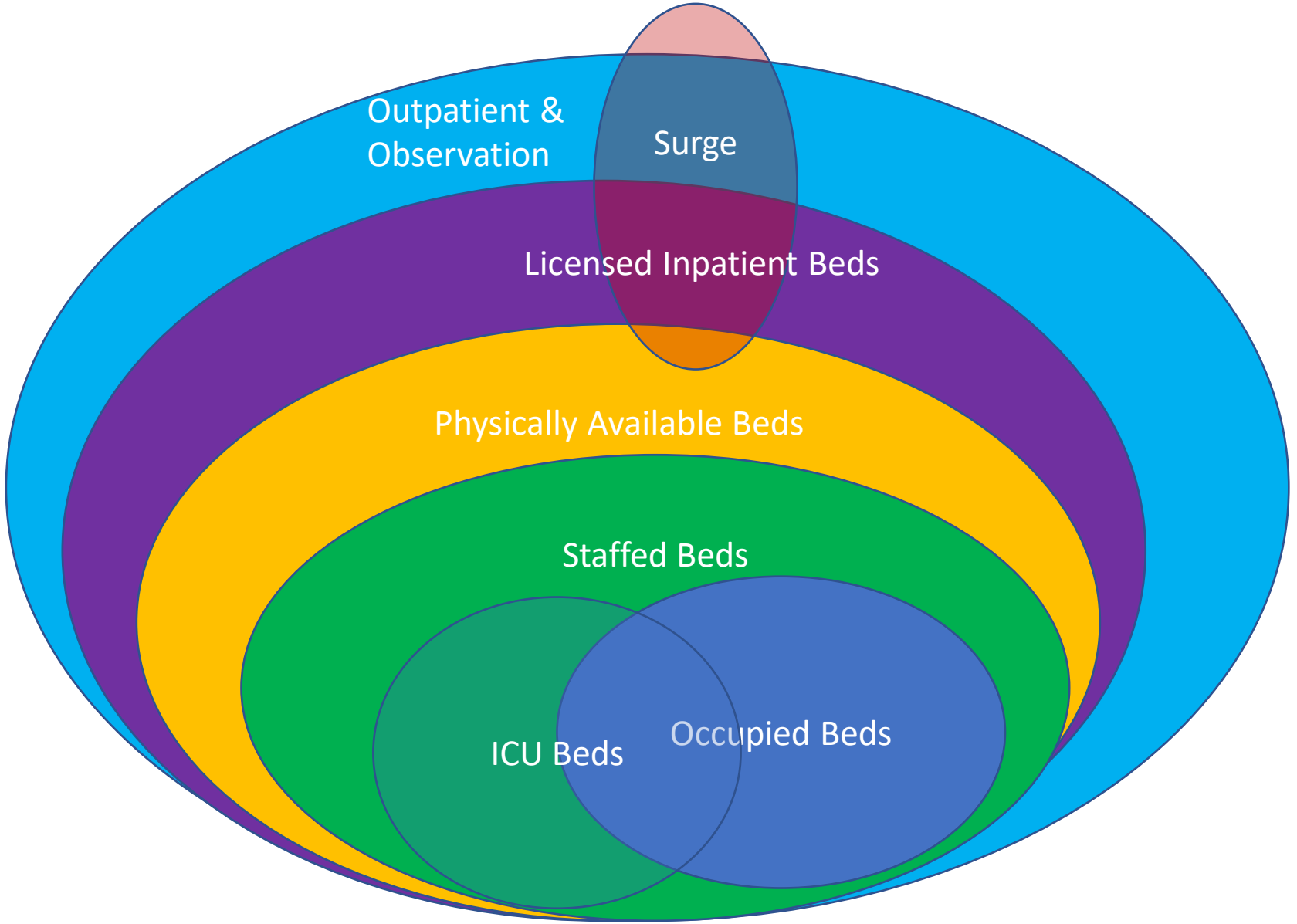


Baseline Data Flow (as of March 2020)



Today's Data Flow




Outpatient &
Observation
Surge
Licensed Inpatient Beds
Physically Available Beds
Staffed Beds
ICU Beds
Occupied Beds

Data Repository #2: Whole Person Health

NCDHHS Whole Person Health

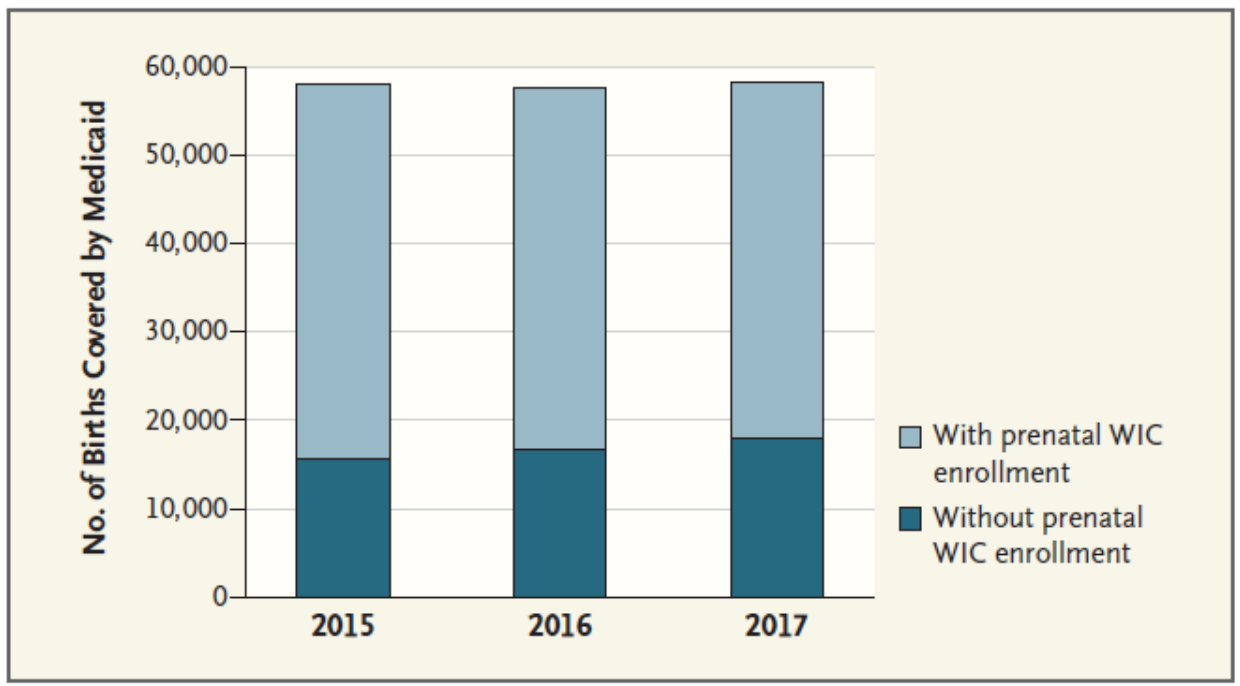
- **Purpose and structure:**
  - **360 Degree View of NC residents served through NCDHHS**
- **Process used to implement data repository system**
  - **Build upon BIDP from COVID**
- **Management structure**
  - **Inter-divisional data ownership with central data custodian**
- **Outcome of efforts**
  - **Enriched dataset for cross-enrollment analysis**
  - **More ongoing…**

The NEW ENGLAND JOURNAL of MEDICINE

Perspective FREE PREVIEW

Focusing on Population Health at Scale — Joining Policy and Technology to Improve Health

Aaron McKethan, Ph.D., Seth A. Berkowitz, M.D., M.P.H., and Mandy Cohen, M.D., M.P.H.



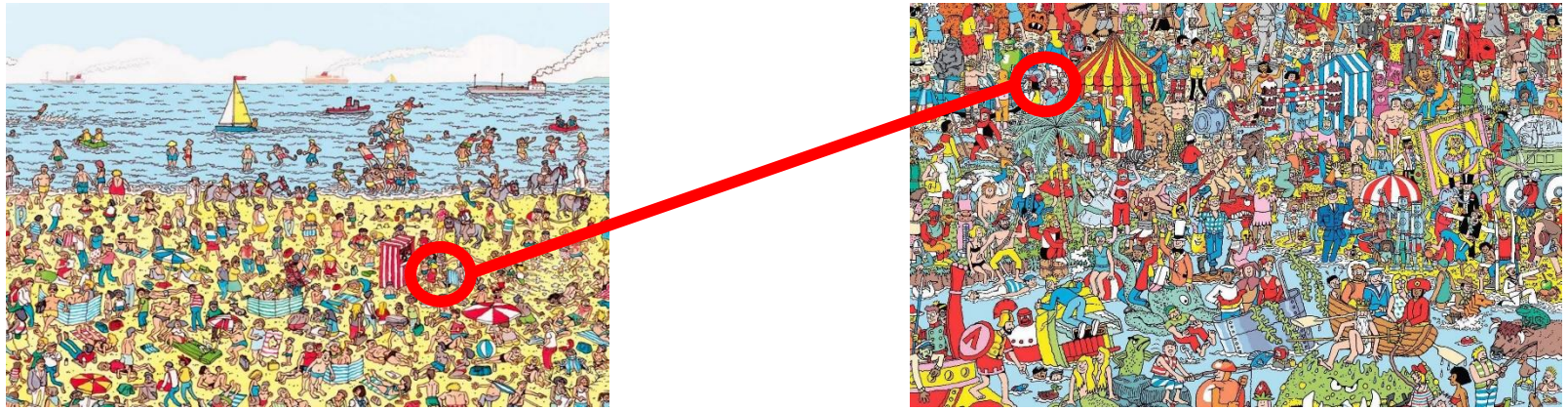
Medicaid-Covered Births with and without Concurrent Prenatal WIC Enrollment, North Carolina.

In order to gain actionable knowledge, need 2 things:

1. **The ability to integrate data between divisional silos (“Data Integration”)**

2. **The ability to identify and link the same individual from different datasets (“Entity Resolution”)**

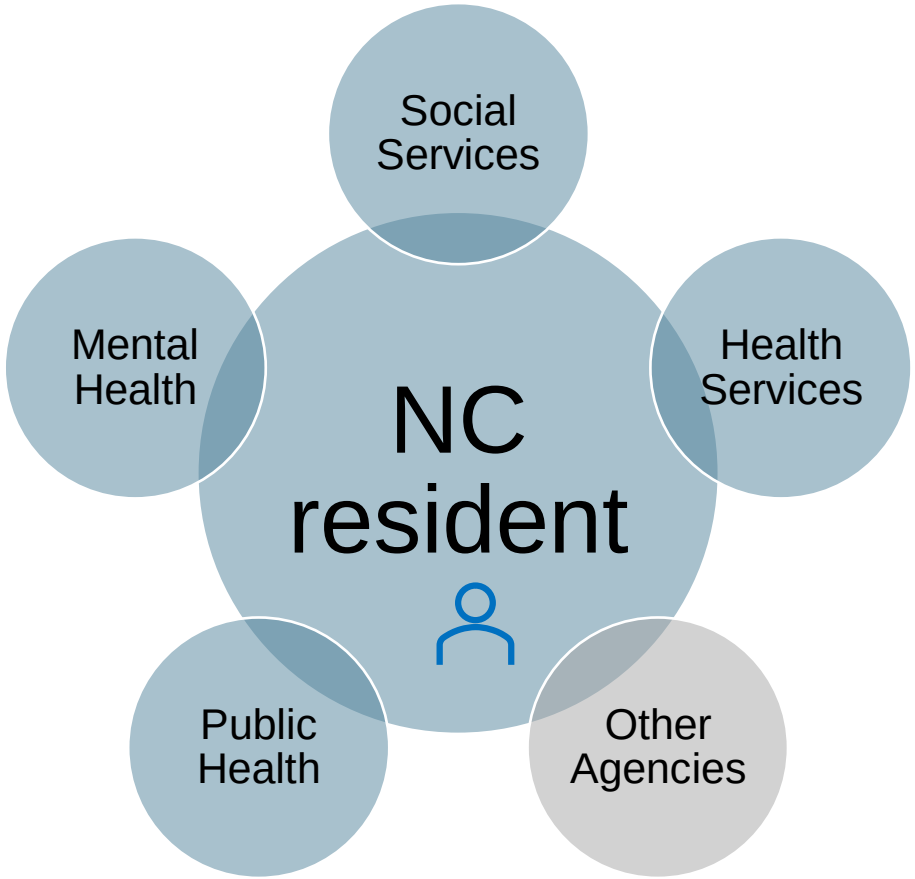
ID: 123456
Name: Waldo



ID: 123456
Name: Waldo

Whole Person Health

- Goal: link data to facilitate a “Whole Person Health” view of the people we serve.
  - Real-time individual level
  - Aggregate analysis to inform policy
- Requires the ability to integrate data across divisional silos
- Which requires ability to link records between systems



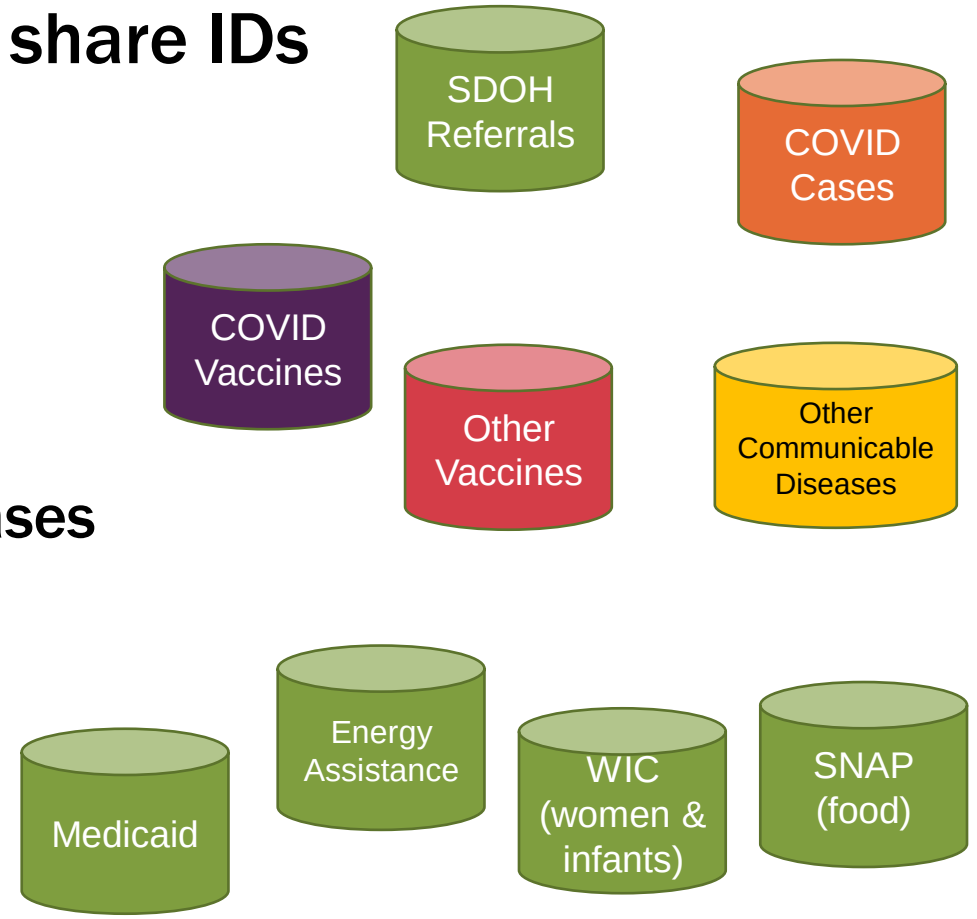
Motivating Questions (Examples)

- What % of Medicaid beneficiaries have been vaccinated?
- Which children in foster care have prescriptions for >4 psychotropic medications?
- Who receives regular food assistance referrals and therefore may benefit from SNAP but is not enrolled?
- What % of people experiencing homelessness have been vaccinated?
- What is the relationship between early grade outcomes (e.g., third grade reading) and different early childhood conditions (e.g., early learning, health, housing, child welfare)?

Answering those questions is currently difficult at best

- Data live in silos that (mostly) do not share IDs
- Probabilistic match is possible, but
  - Labor-intensive
  - Prone to error

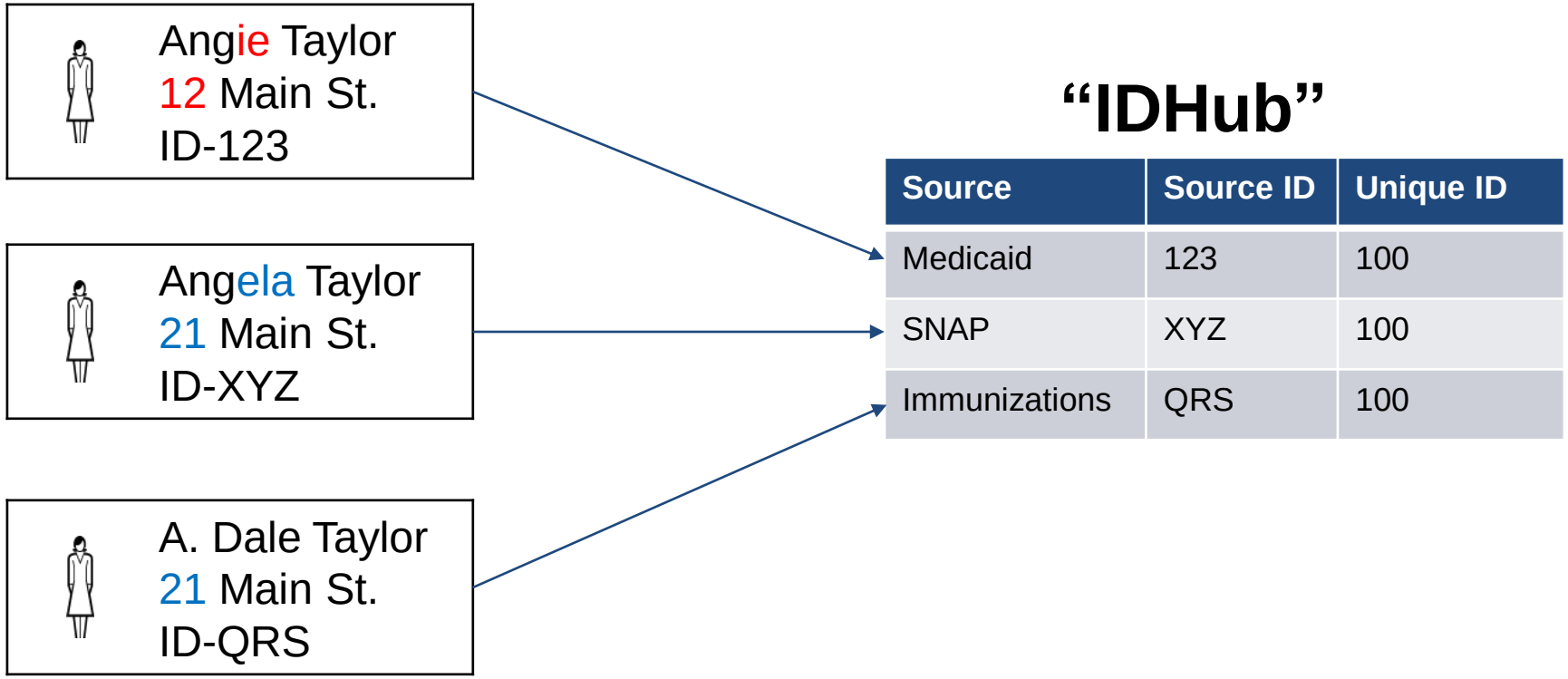
  - Example: initial approach for post-vax cases



Options for “solving” entity resolution

1. Each system uses its own ID, probabilistic “fuzzy match” between systems as needed
2. One universal ID, e.g. state-wide or National Health Identifier
3. Somewhere in between
   a. Multiple “standard” identifiers
   b. Map them to each other!
   c. Refer to that mapping for efficient data integration

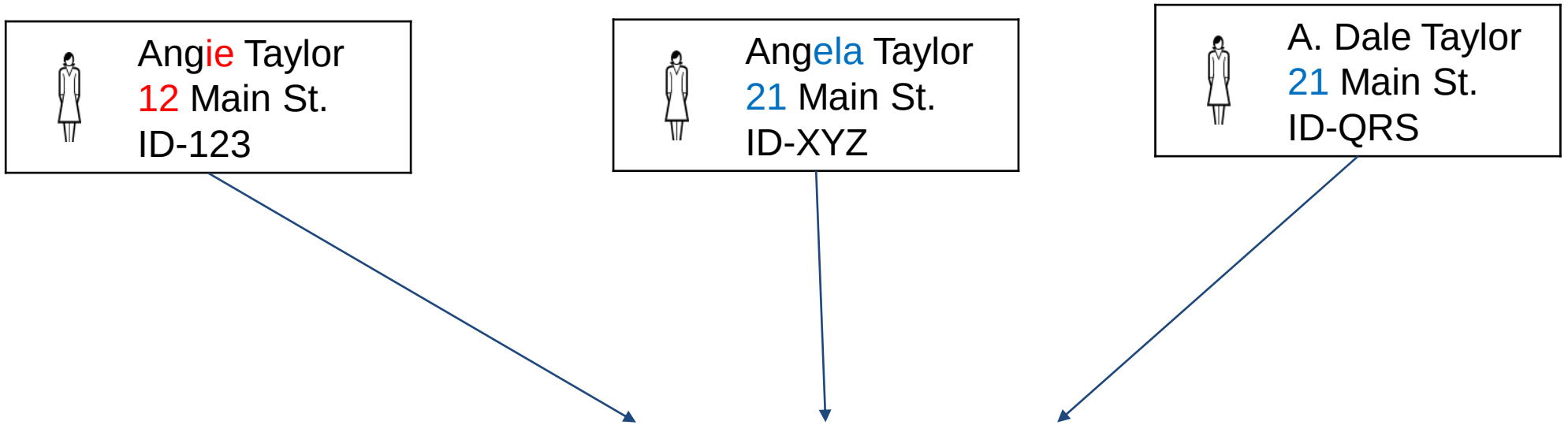
Map each ID to a universal unique identifier…



“IDHub”

Source	Source ID	Unique ID
Medicaid	123	100
SNAP	XYZ	100
Immunizations	QRS	100

…enabling mapping each separate ID to the others



EER ID	Medicaid	NCCARE360	DPI
100	123	XYZ	QRS

IDs can (and do!) change over time

- **MPI's (master patient index) are frequently merged as more data are incorporated**

Jessie
Address A
MPI 111

Jessica
Address B
MPI ~~222~~ 111

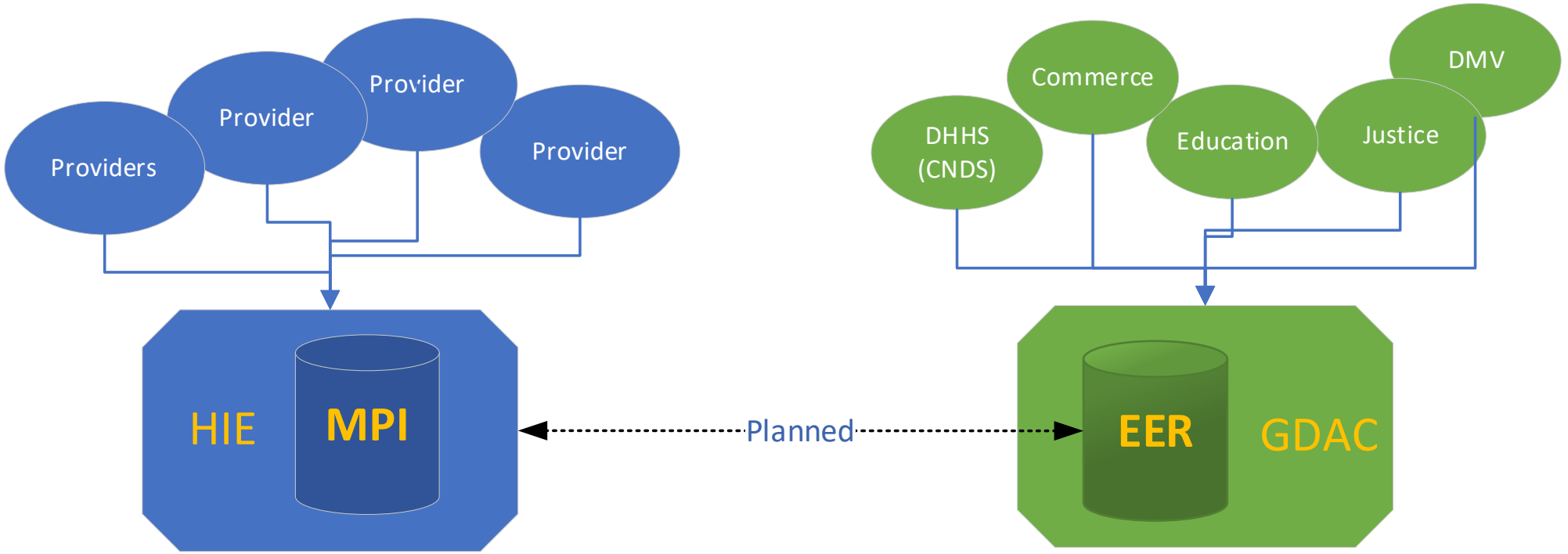
Jessie
Address B
MPI 111

- **Splits (after a "false positive" match) far rarer and more challenging to handle**

A few things about the NC Landscape

- **State-run HIE- “NC HealthConnex”**
- **Government Data Analytics Center (GDAC)- a Division of NC’s Department of Information Technology**
  - **Created by legislation for data management and analytics**
- **NC is home to SAS**

NC has 2 “Universal” Identifiers: clinical vs. non-clinical



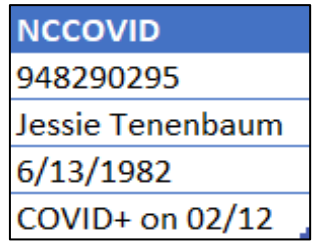
MPI ID (Master Patient Index)

- Used by Health Information Exchange (HIE) for **clinical data**
- Involves merging clinical records
- High stringency matching

EER ID (Enterprise Entity Resolution)

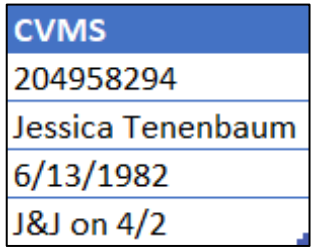
- Used by NC’s Government Data Analytics Center (GDAC) for **non-clinical data**
- Slightly lower bar for matching
- Meant for analytics

NC DHHS Source Systems send identified attributes to ID authoritative source



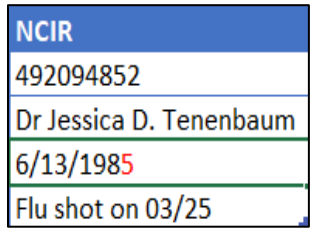
NCCOVID Test Results

NCCOVID ID, Name, DoB, etc.



CVMS Vaccinations

CVMS ID, Name, DoB, etc.



NCIR Immunizations

NCIR ID, Name, DoB, etc.

NCCARE360

GDAC

eLink ID

Source ID with assigned eLink ID

If no match found, new eLink ID is created

IDHub

UUID	CNDS ID	NCCOVID ID	CVMS ID	NCIR ID	eLink ID
0001	887	**948290295**	**204958294**	**492094852**	ABCD
0002	345	875638922	385857395	983475020	EFGH
0003	192	124834749	994749503	044759395	IJKL
0004	434	857248490	747503850	935475983	MNOP

Other options: 3rd party vendors*

- As a service
  - Data sent externally- needs DUA
  - Leverages consumer data
- Off the shelf software
  - Subset of Master Data Management
  - Internal use
  - Increased control

verato

Informatica

Mphasis
The Next Applied

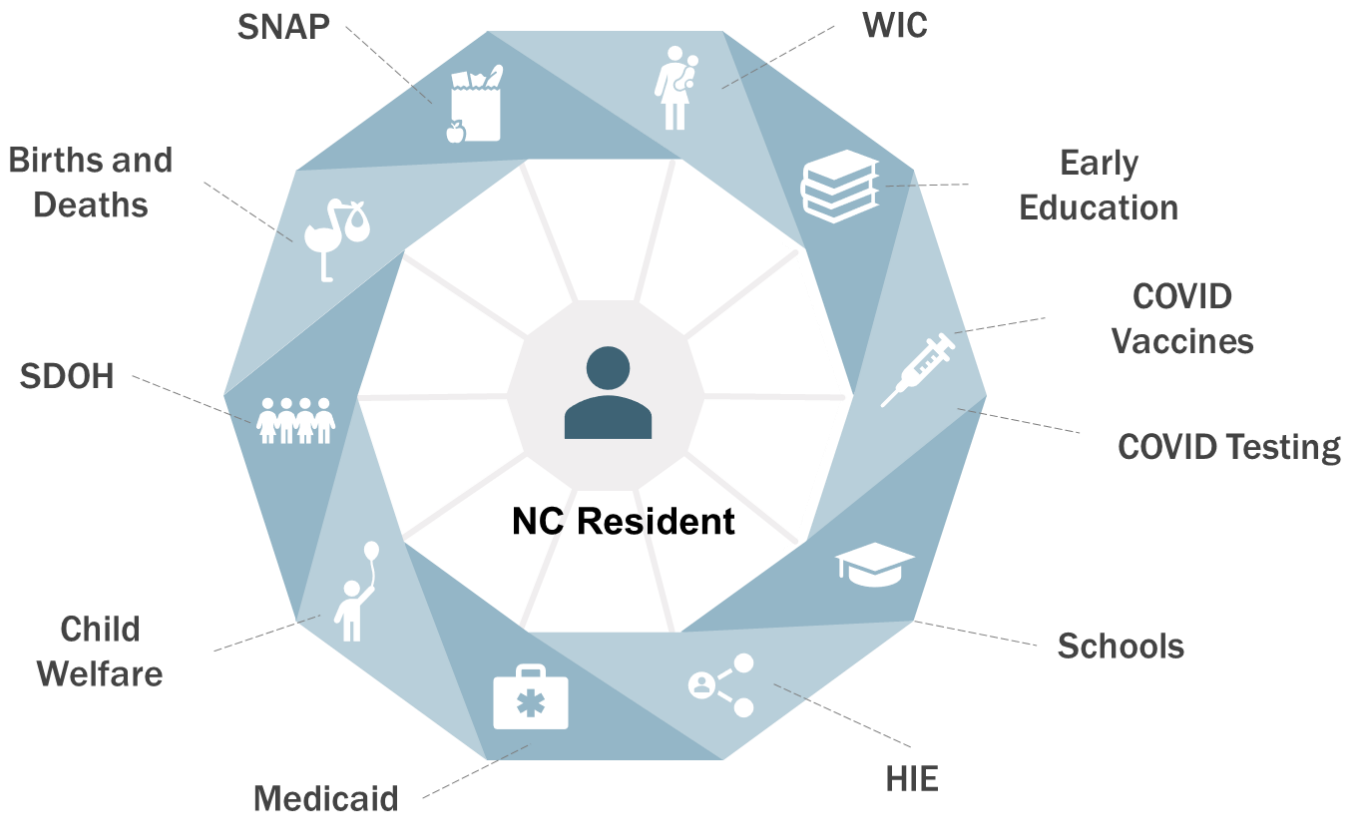
InterSystems
TrakCare

LexisNexis

4medica
One Patient...One Record

*No endorsement intended or implied!

…enabling Whole Person Health



Data Repository #3:
NC Longitudinal
Data ~~System~~ Service

NC Longitudinal Data Service (LDS)

- **Purpose and structure:**
  - Longitudinal view from early childhood to workforce
  - System of systems
- **Process used to implement data repository system**
  - Roadmap built over many years
- **Management structure**
  - Exec Director with Governance Board (on which I serve), multiple
  - committees
- **Outcome of efforts**
  - Development and governance still evolving

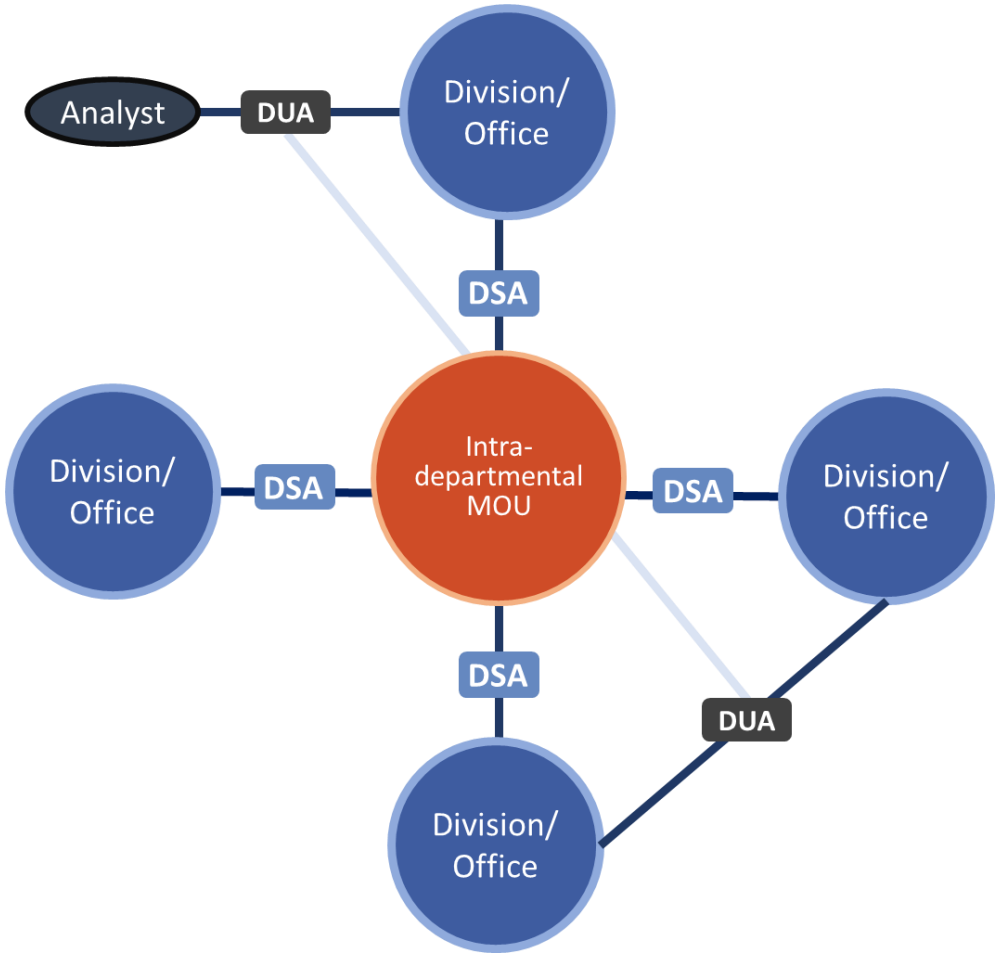
Insights and Challenges

Be Use Case Driven

Governance is hard

- Trust is key
- Involve legal counsel from the start
- Opt out considerations
- Data quality
  - Check!
  - Fit for use
- Legal framework

Foundational legal agreements: visual representation



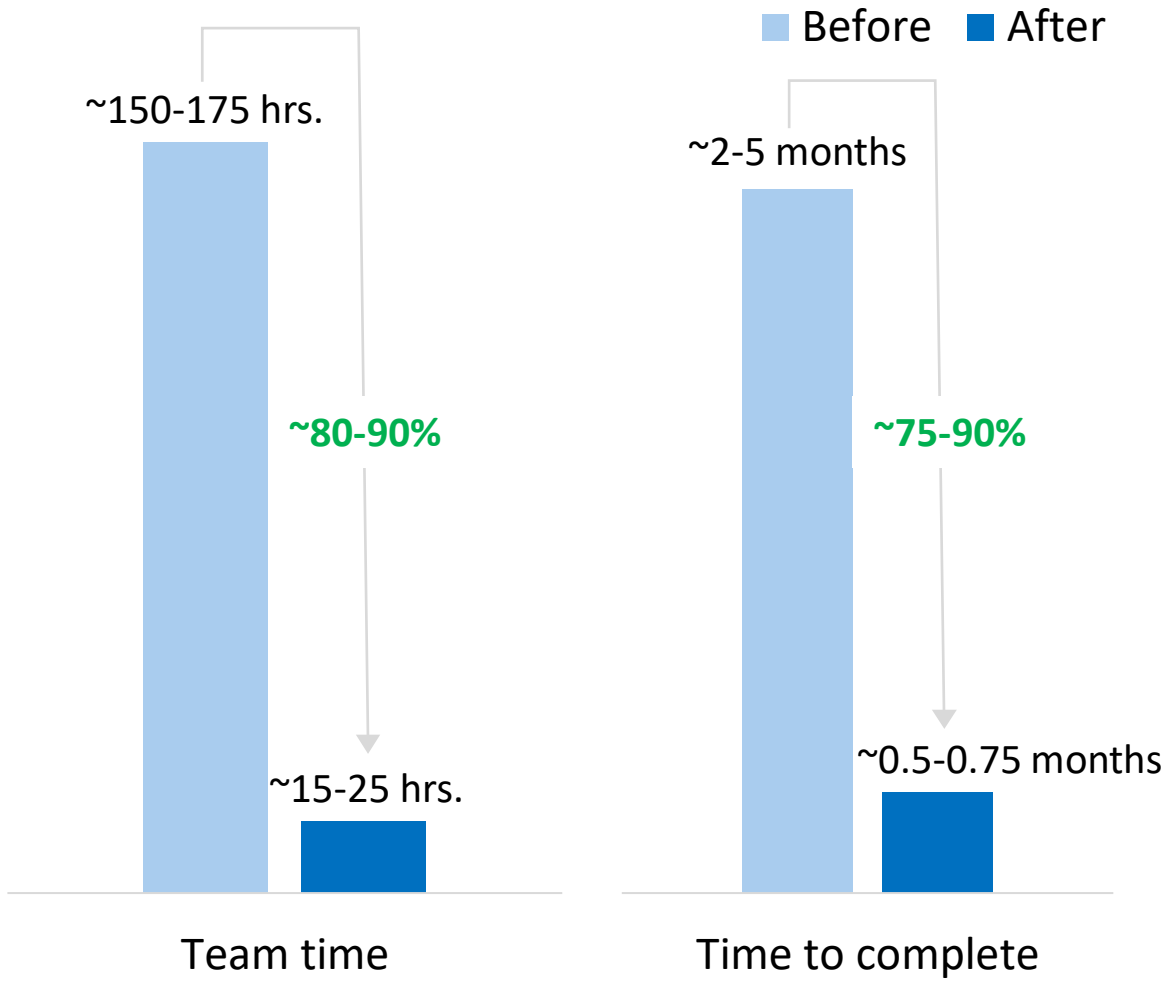
INTRA-DHHS DATA SHARING: BENEFITS OF NEW LEGAL / DATA GOVERNANCE FRAMEWORK

Illustrative

Key benefits

1. **Clarifies requirements and guidelines** (e.g., who the permitted signatories are)
2. Provides **approved language & templates** for agreements, preventing rework while also **mitigating risk**
3. **Saves team time & effort for business and legal**, often by not requiring an additional Data Use Agreement
4. **Gets to data insights & program action faster**, given quicker time to completion

Time savings per use – estimated median experience*



*Use cases that do not fit into DSA or are external to DHHS may be outside of this construct and require more time

Lessons Learned

- This stuff is hard- the devil is in the details, and the edge cases
- Data standards are important
- Takes longer and costs more than one might expect
- Do agreements early!
- Get stakeholder buyin
- Concrete wins can help keep momentum
- Motivate business support through targeted use cases

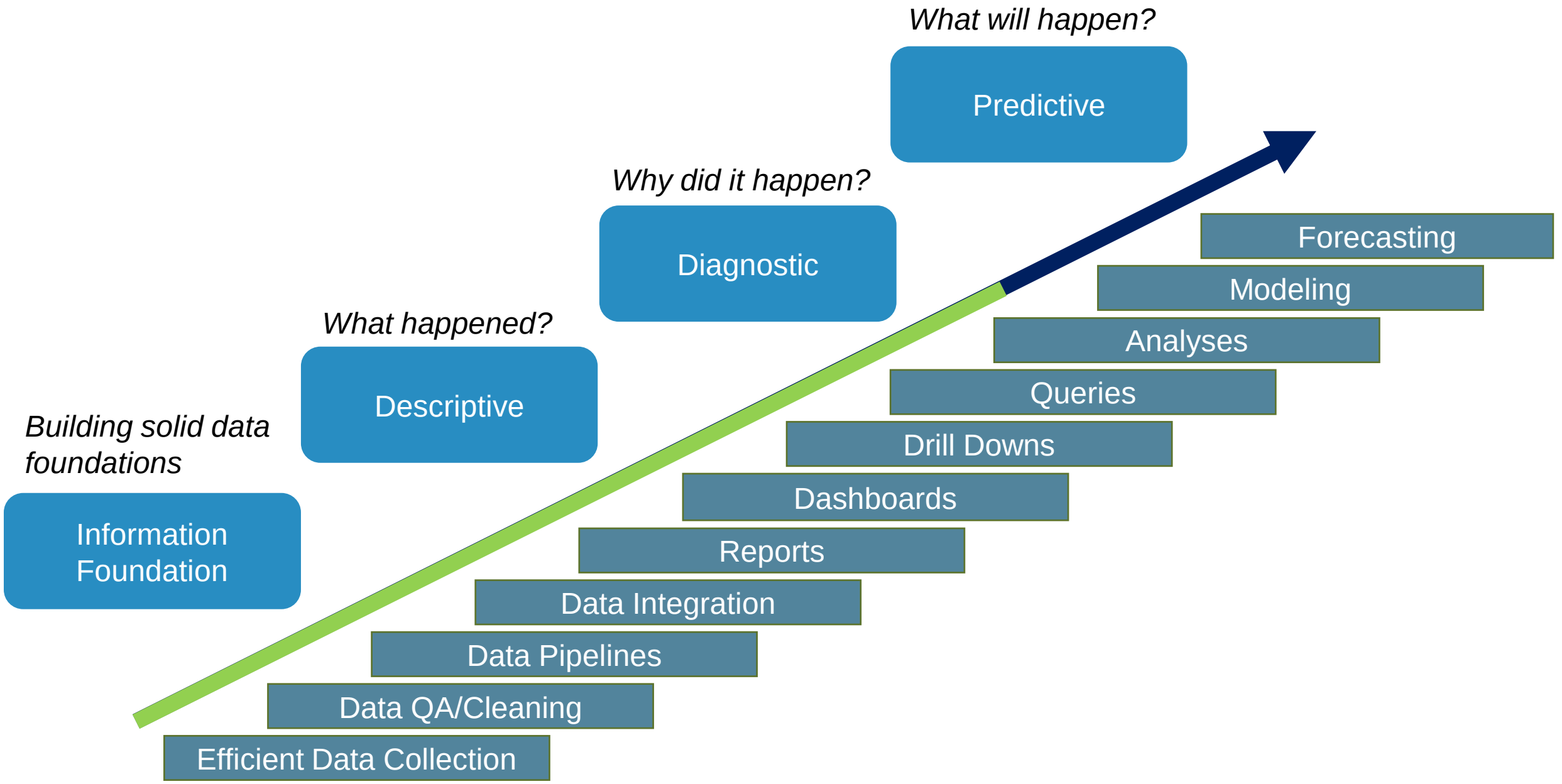
Thanks!

@jessiet1023

Data Pillars

- **Data Infrastructure:** Technology used to store, exchange, and access data
- **Data Governance:** People, processes, and technology for data quality, security, management, and access
- **Data Use:** Reports, visualization, and analysis
- **Data Literacy:** Workforce training across all levels of baseline knowledge

Moving from descriptive to predictive



KEY SOLUTIONS: DATA INTEGRATION

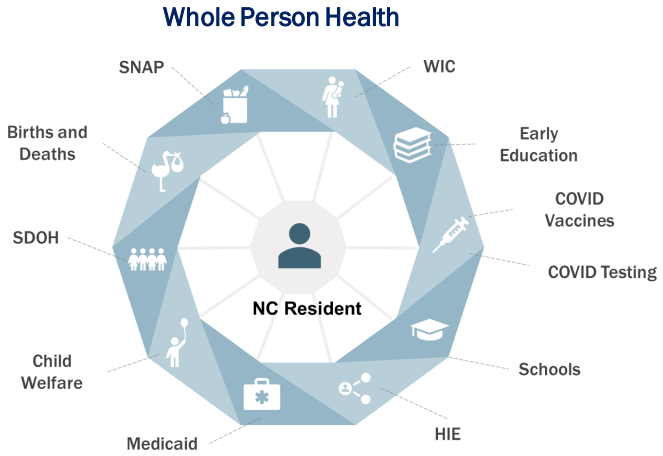
Limited availability of data integrated across programs for whole-person health

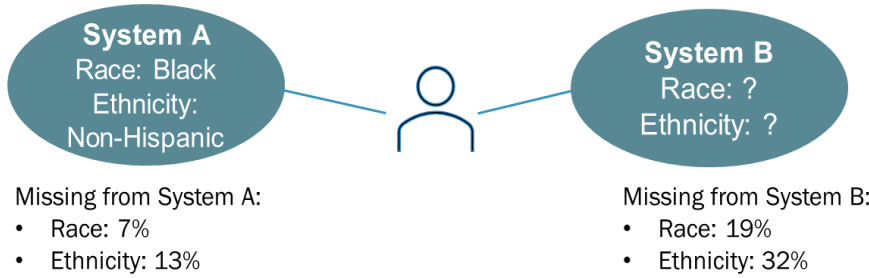
Reusable integrated data: Data structures have been built to integrate data across programs in a way that enables repeatable and extensible analysis of whole-person health data

Objectives

Leverage data integrated across programs for 1) enabling repeatable and extensible whole-person health use cases and 2) equity analysis and action

Some datasets have relatively complete data on race and ethnicity, others do not.





Whole person view enables “filling in the gaps.”