

TDU

INTRODUCTION TO DATA MANAGEMENT FOR APPLIED PUBLIC HEALTH

November 21, 2022



Public Health
Agency of Canada

Agence de la santé
publique du Canada

WELCOME TO THE COURSE

1. Welcome and housekeeping
2. Data management basics
3. Data management horror stories
4. Take away activity
5. Caveats and recommendations
6. Wrap-up

LAND ACKNOWLEDGEMENT

Marc-André Bélair:
Vancouver, BC

Unceded traditional territories
of the Musqueam, Squamish
and Tsleil-Waututh Nations

Alanah Jansen:
Victoria, BC

Land of the Coast Salish
People of the Hul'qumi'num
Treaty Group

Ben: Winnipeg, MB
Treaty One Territory, Land of
the Anishinabe, Ininew, Oji-Cree, Dene,
Dakota, and Métis people

Anja: Toronto, ON
Wendat, the Anishnaabeg,
Haudenosaunee, Métis, and
the Mississaugas of the Credit First
Nation

Joanne: St. John's, NL

- NFLD: Mi'kmaq and Beothuk
- Labrador: Innu (Nitassinan), Inuit (Nunatsivut), Southern Labrador Inuit (NunatuKavut)

Mirna: Montreal, QC

Kanien'kehá:ka Nation, with ties to
the Haudenosaunee Confederacy and
Anishinaabeg peoples (unceded)

Michael: Ottawa, ON
Land of the Algonquin Anishinabe Nation

<https://native-land.ca/>

LEARNING OBJECTIVES

- By the end of the training participants will be able to:
 - Explain principles of data management in the context of applied public health practice;
 - Identify common ways in which data are stored for use in public health investigations;
 - Compose caveats for interpretation and formulate recommendations to improve data quality; and
 - Execute descriptive epidemiological analyses to summarize and interpret data in the context of a public health investigation.

EXPERIENTIAL AIM

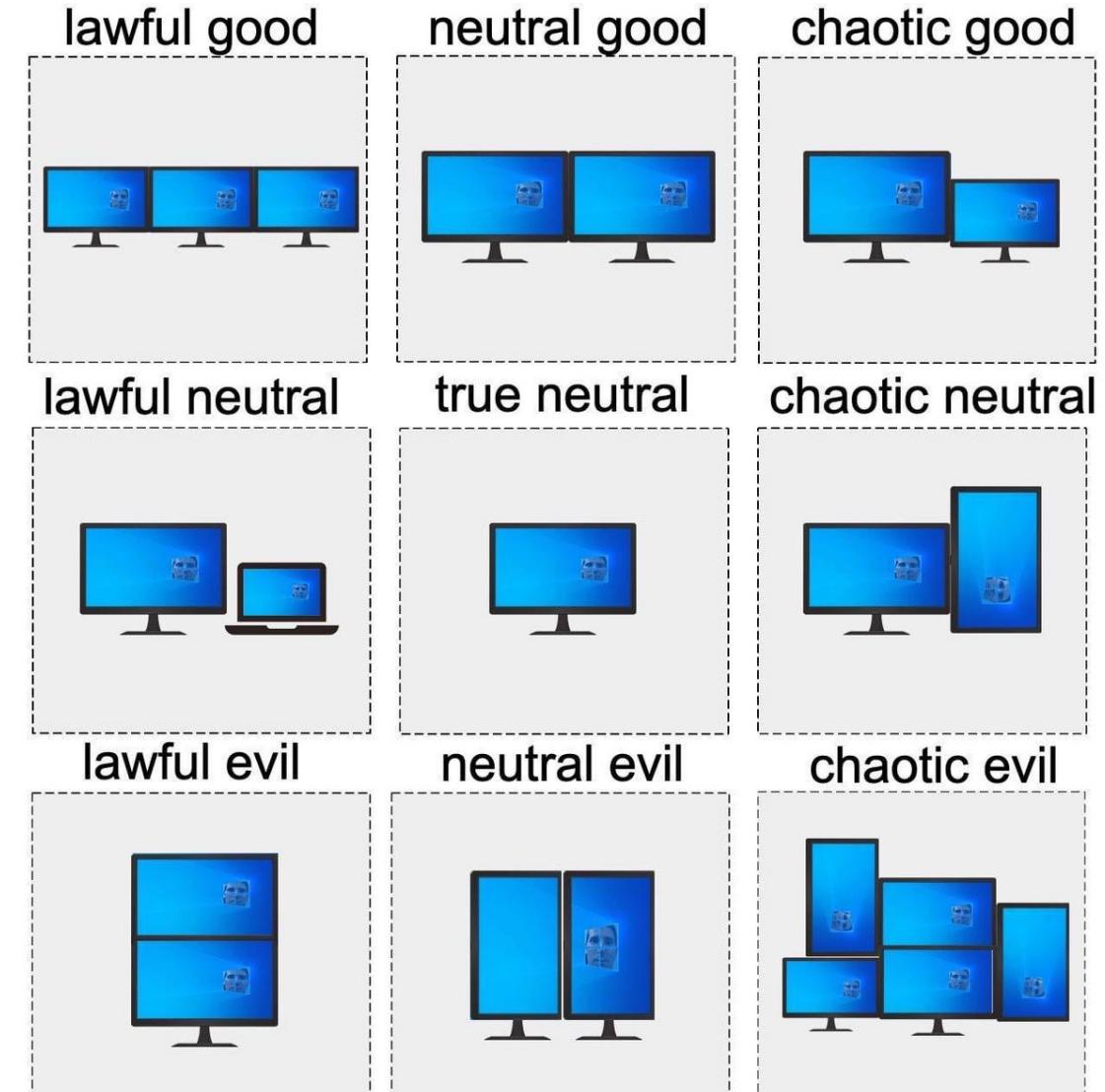
- To provide a safe place and relaxed environment for learning, practice, and discussion.
- To have fun!

HOUSEKEEPING

- Virtual classroom environment: distractions are a given.
- Please keep muted unless you have something to discuss.
- Please have cameras on for discussions, though please feel free to turn your cameras off in between activities or during lectures.
- Feel free to step away at anytime. Use the coffee cup to let us know!
- Maintain an inclusive, respectful and fun environment.

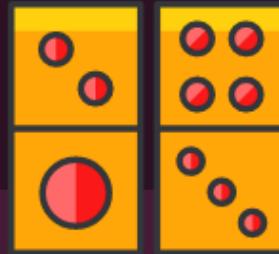
ICEBREAKER!

- What is your data-alignment?
- Add your name to the square in **the rubric on Slide 9** that best describes your data-alignment! (Note: these can be completely fictional!)
- We'll go around the room and give everyone the opportunity to introduce themselves, and describe the persona they think fits within their data-alignment of choice.
- (See next slide for some examples!)



Icebreaker challenge:

ROLL FOR INITIATIVE!



Do you see yourself as:



"LAWFUL" DATA ALIGNMENT

- Performs all data requests in advance
- Ensures corporate-approved formatting and analysis

"I use only IT-approved software and assets because it is the right thing to do."

or



"CHAOTIC" DATA ALIGNMENT

- Frequently requests an extension
- "One more thing"
- Requests all available data, filters in analysis stage

"Sometimes you just have to wait for the data speak to you."



"GOOD" DATA ALIGNMENT

- Extensive documentation for easy hand-overs
- Involves community liaisons

"Our primary goal is ensuring transparency, reproducibility, and that data do no harm."

or



"EVIL" DATA ALIGNMENT

- No documentation
- Doesn't involve others
- Analysis conforms to self-interests

"Just leave the request with me and I'll take care of it when I can fit it in."

Lawful Good

Neutral Good

Chaotic Good

Lawful Neutral

Neutral

Chaotic Neutral
[Ben]

Lawful Evil

Neutral Evil

Chaotic Evil

DATA MANAGEMENT BASICS AND SOUND PRACTICES

OVERVIEW

- Public health data
- How public health data are often organized for use
- Data management in public health
- Considerations for data management in applied public health practice

DATA AND PUBLIC HEALTH

And you and you and you

WHAT ARE* PUBLIC HEALTH DATA?

- Data are *everywhere* in public health, and the use of data to inform evidence-based recommendations are central to our roles as epidemiologists
- Effective data collection, analysis and interpretation are key tools for **evidence-based decision making**
 - Effective interventions
 - Responsibility to public
 - Measure impact of outcomes
 - Identify additional risk factors

EXAMPLES OF DATA IN PUBLIC HEALTH?

DATA AND THE ROLE OF THE EPIDEMIOLOGIST

- Questionnaires and surveys
- Outbreak line lists
- Vital statistics
- Laboratory
- Biomonitoring
- Environmental reports
- Dashboards
- Free text
- Routine surveillance
- Hospital admissions / discharges
- Medical charts
- Insurance billing
- Emergency medical services
- Pharmaceuticals – OTC sales, dispensations
- Border movements

DATA AND THE ROLE OF THE EPIDEMIOLOGIST

- **What do epidemiologists do with regards to these data?**

- Identify sources of data
- Develop indicators
- Design and implement questionnaires and surveys
- Data collation and analysis planning
- Data quality assurance
- Database design and construction
- Data analysis and visualisation
- Communication and reporting
- Interpretation and formulate recommendations
- Evaluation

All this and more!

EXAMPLES OF DATA IN PUBLIC HEALTH

- Outbreak line lists

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	natid	locid	healthunit	casestatus	interstatus	province	sexcase	agecase	symp1dt	hospad	hus	death	recovered	trav	earliestdate	earliesttype
2	AB-77	X-93118		27	Confirmed Not Interviewed	AB	F	7	43997	N					43997	Onset Date
3	MB-75	X-75922		58	Confirmed Interviewed: Routine Followup	MB	M	55	43999	Y				N	43999	Onset Date
4	MB-29	X-40899		44	Confirmed Interviewed: Routine Followup	MB	F	26	44001	Y		N			44001	Onset Date
5	AB-9	X-57925		5	Confirmed Interviewed: Routine Followup	AB	M	27	44002	Y		N			44002	Onset Date
6	AB-88	X-83124		16	Confirmed Not Interviewed	AB	F	65	44004	N		N			44004	Onset Date
7	AB-53	X-81940		46	Confirmed Interviewed: Routine Followup	AB	F	32	44004	N		N			44004	Onset Date
8	MB-44	X-99018		13	Confirmed Interviewed: Routine Followup	MB	M	78	44004	N				N	44004	Onset Date
9	AB-6	X-43439		49	Confirmed Interviewed: Routine Followup	AB	M	44	44005	DK		N		Y	44005	Onset Date
10	AB-67	X-73906		82	Confirmed Interviewed: Routine Followup	AB	F	28	44006	N		N		N	44006	Onset Date
11	AB-21	X-64891		69	Confirmed Not Interviewed	AB	F	70	44006	N		N			44006	Onset Date
12	AB-44	X-26587		60	Confirmed Interviewed: Routine Followup	AB	M	68	44006	N		N		Y	44006	Onset Date
13	AB-52	X-93773		7	Confirmed Not Interviewed	AB	F	35	44007	N		N			44007	Onset Date
14	AB-40	X-6675		84	Confirmed Interviewed: Routine Followup	AB	F	77	44008	N		N			44008	Onset Date
15	AB-37	X-88850		14	Confirmed Interviewed: Routine Followup	AB	F	16	44008	N		N			44008	Onset Date
16	AB-94	X-36242		94	Confirmed Not Interviewed	AB	F	79	44009	Y		N			44009	Onset Date
17	AB-77	X-81867		10	Confirmed Interviewed: Routine Followup	AB	M	6	44009	N		N			44009	Onset Date
18	AB-42	X-35231		93	Confirmed Interviewed: Routine Followup	AB	F	20	44010	N		N			44010	Onset Date
19	BC-89	X-95694		72	Confirmed Interviewed: Routine Followup	BC	M	29	44010	Y				N	44010	Onset Date
20	AB-28	X-52707		39	Confirmed Interviewed: Routine Followup	AB	M	84	44011	N		N			44011	Onset Date
21	AB-84	X-96436		84	Confirmed Interviewed: Routine Followup	AB	M	5	44011	N		N			44011	Onset Date
22	AB-29	X-14410		21	Confirmed Interviewed: Routine Followup	AB	F	89	44011	Y		N			44011	Onset Date
23	AB-25	X-29445		96	Confirmed Interviewed: Routine Followup	AB	M	30	44012	N		N			44012	Onset Date
24	AB-51	X-51344		75	Confirmed Not Interviewed	AB	F	86	44012	N		N			44012	Onset Date
25	AB-11	X-70393		11	Confirmed Interviewed: Routine Followup	AB	F	64	44012	N		N			44012	Onset Date
26	AB-20	X-89372		53	Confirmed Interviewed: Routine Followup	AB	M	44	44012	N		N			44012	Onset Date

EXAMPLES OF DATA IN PUBLIC HEALTH

- Surveillance databases

Government of Canada Gouvernement du Canada Search Canada.ca

Jobs Immigration Travel Business Benefits Health Taxes More services

Home → Health → Science, research & data → Public Health Infobase → Data Tools

Data Tools

The Data Tools help users visualize public health data through simple drop-down menus. Users can access geographic comparisons, trends, age distributions, and disaggregated data layered by socioeconomic and sociodemographic variables.

Search the Data Tools: Enter keyword(s) here Sort search results by: A/Z Category Relevance Date

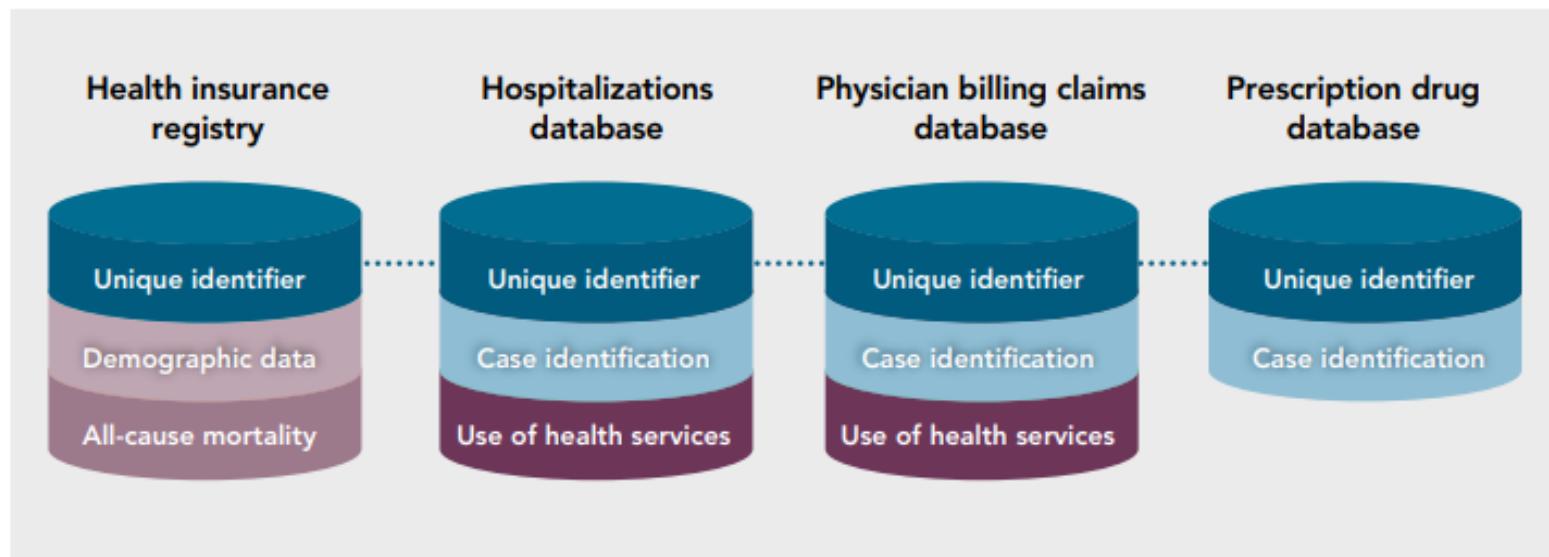
Nutrition Surveillance Data Tool
The Nutrition Surveillance Data Tool was developed by the Bureau of Food Surveillance and Science Integration (BFSSI) at Health Canada. This tool presents usual intakes of energy, nutrients and other dietary components using data collected in the 2015 Canadian Community Health Survey – Nutrition. It highlights where Canadians' intakes fall relative to Dietary Reference Intakes established by the Institute of Medicine of The National Academies of Science, Engineering and Medicine.
[View](#)

Canadian Risk Factor Atlas (CRFA)
The Canadian Risk Factor Atlas (CRFA) is an interactive mapping interface presenting data on key risk factors for chronic disease and indicators of poor mental health integrated into dynamic Canadian maps at several geographic levels. Maps can be filtered by age groups and by sex. Data is also presented by socioeconomic characteristics for all three geographic layers and by population centres or rural areas of residence for the provinces and territories.
[View](#)

Cancer in Canada Data Tool (CCDT)
The Canadian Cancer Data Tool (CCDT) provides comprehensive data on the incidence and mortality of cancer in Canada over time by age and sex in an easy to use, flexible format. Information on 24 different categories of cancer types.
[View](#)

Cancer in Young People in Canada Data Tool
The Cancer in Young People in Canada (CYP-C) Data Tool provides pan-Canadian surveillance data on children and youth with cancer to inform research and planning for cancer control efforts. The CYP-C surveillance system operates

The Canadian Chronic Disease Surveillance System (CCDSS) Data Tool presents incidence, prevalence and all-cause mortality



Select item(s) from each list below:

Geography

Canada

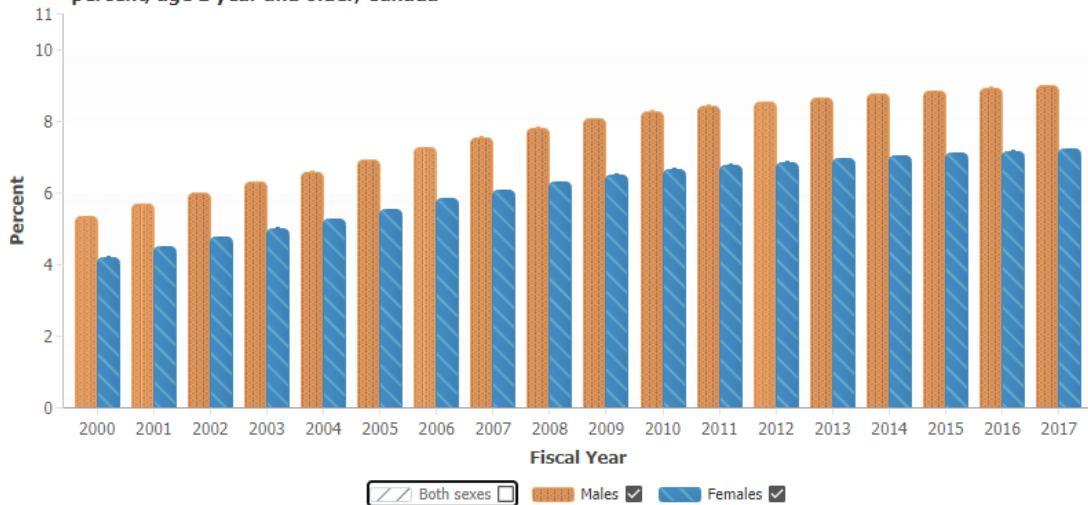
Conditions

Diabetes mellitus (types comb)

Measure

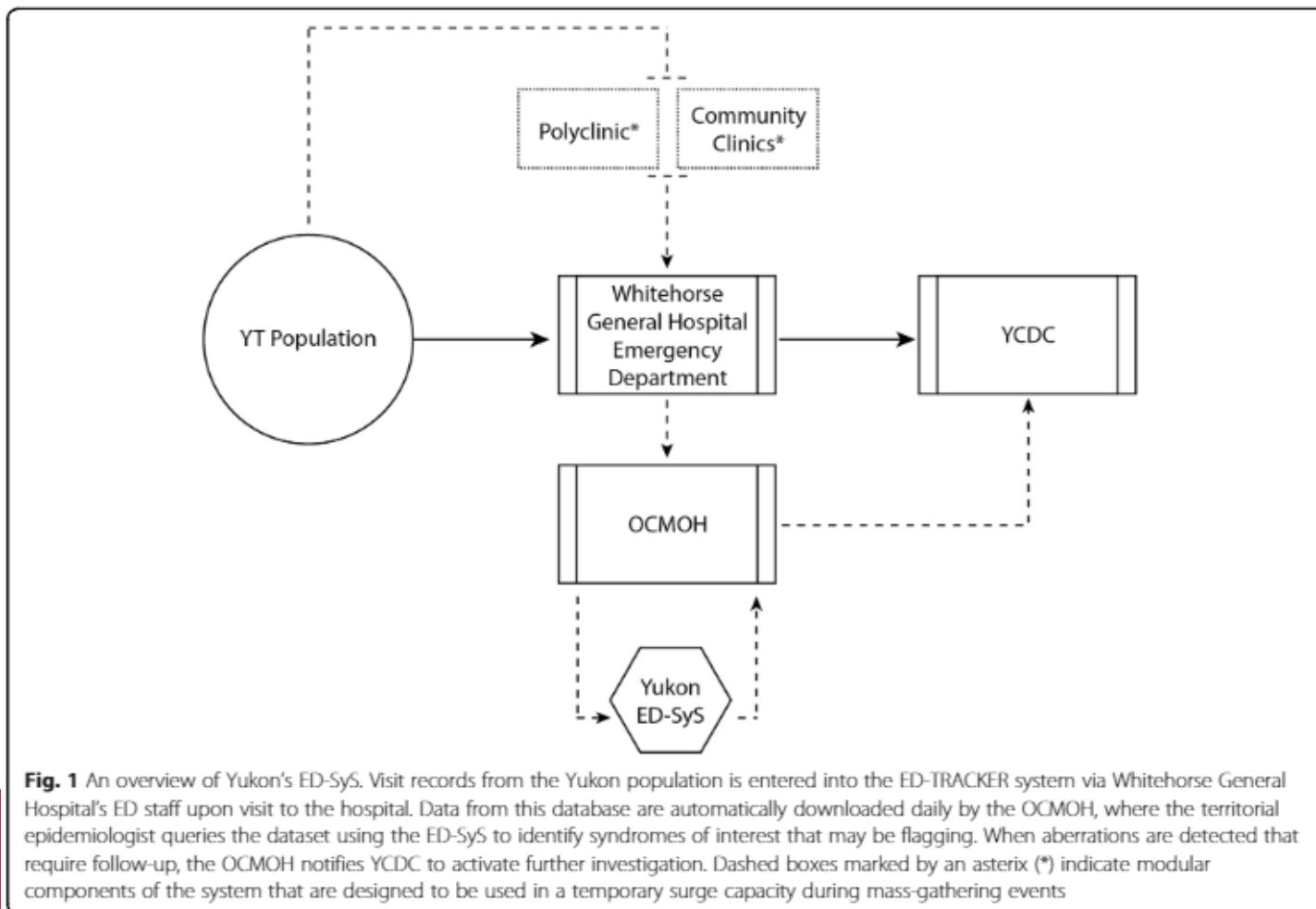
Age-standardized prevalence

Diabetes mellitus (types combined), excluding gestational diabetes, age-standardized prevalence, percent, age 1 year and older, Canada*



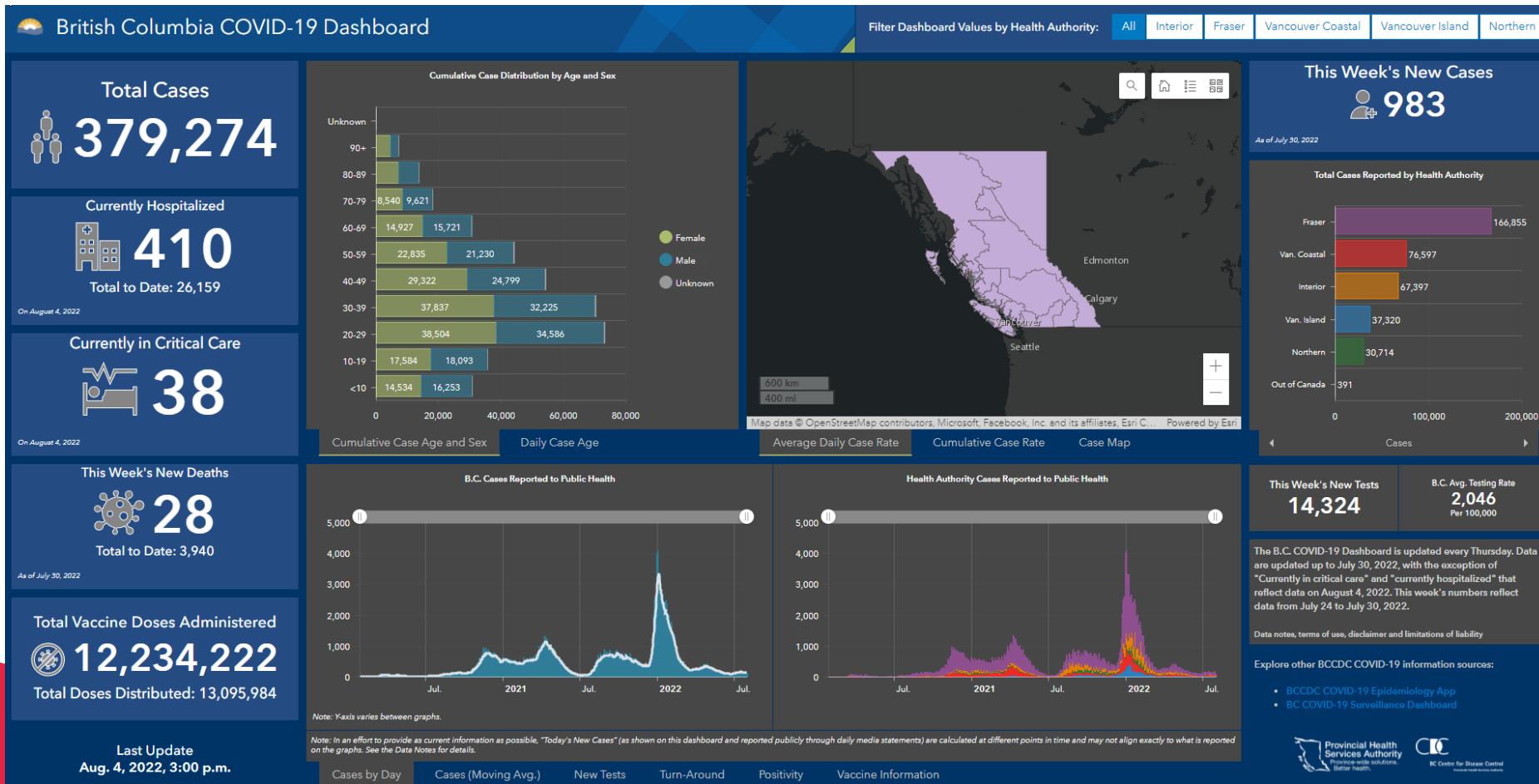
EXAMPLES OF DATA IN PUBLIC HEALTH

- Syndromic Surveillance Systems

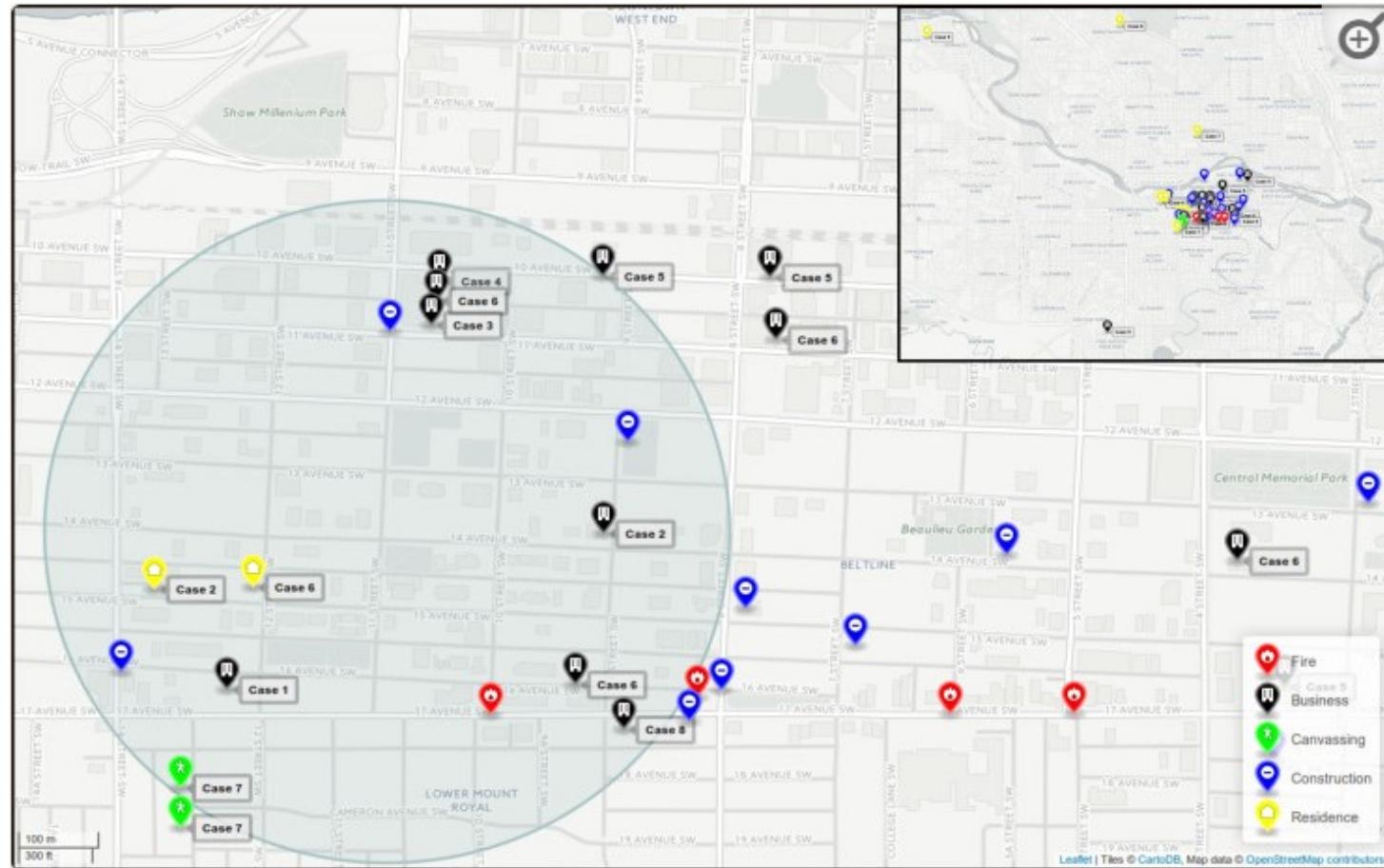


EXAMPLES OF DATA IN PUBLIC HEALTH

- Dashboards



EXAMPLES OF DATA IN PUBLIC HEALTH



Unusual Legionnaires' outbreak in cool, dry Western Canada: an investigation using genomic epidemiology

HOW PUBLIC HEALTH DATA ARE ORGANIZED FOR USE

From here to there

FLAT FILES

- Simple format, easy to create and manage as spreadsheets (I.e., (Excel))
- Can contain wide data (case linelists); or long data (multiple observations per unit)

Case	Sex	Age	Symptom
01	M	32	Headache
01	M	32	Fever
02	F	45	Headache
02	F	45	Coughing
02	F	45	Fever

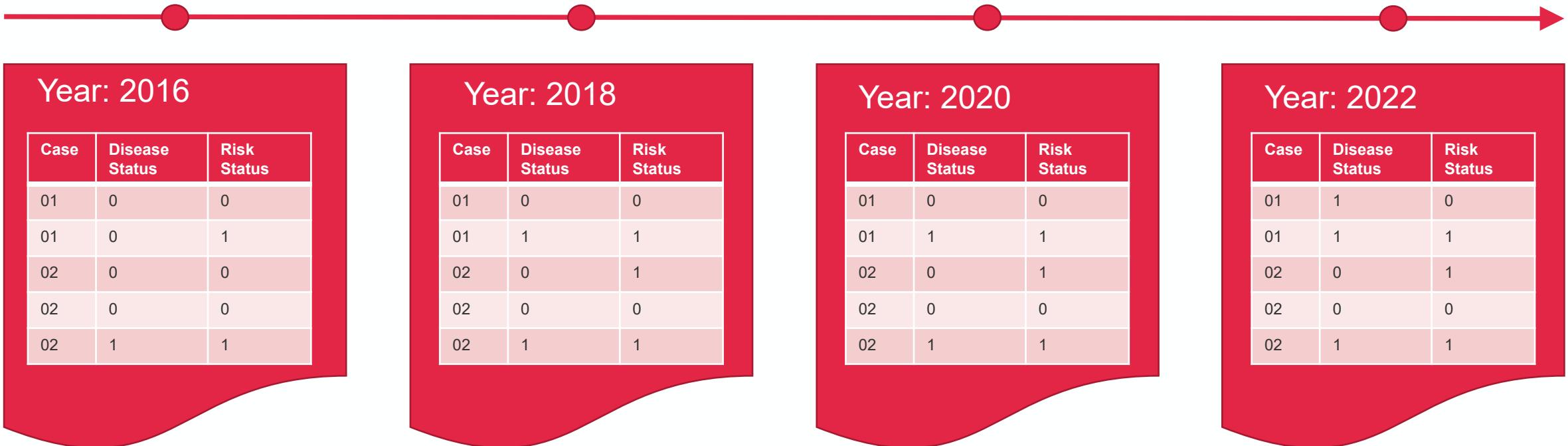
Long
format

Case	Sex	Age	Headache	Fever	Coughing
01	M	32	Y	Y	N
02	F	45	Y	Y	Y

Wide
format

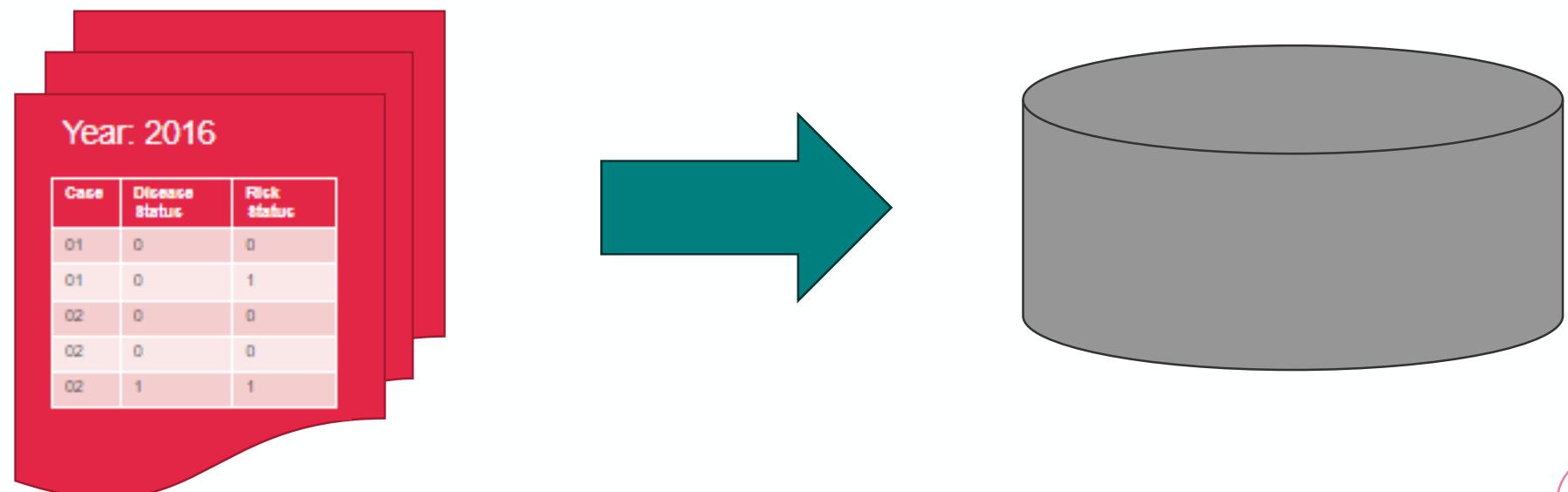
LONGITUDINAL DATA

- The same data collected from the same individuals at multiple points over time



DATABASES

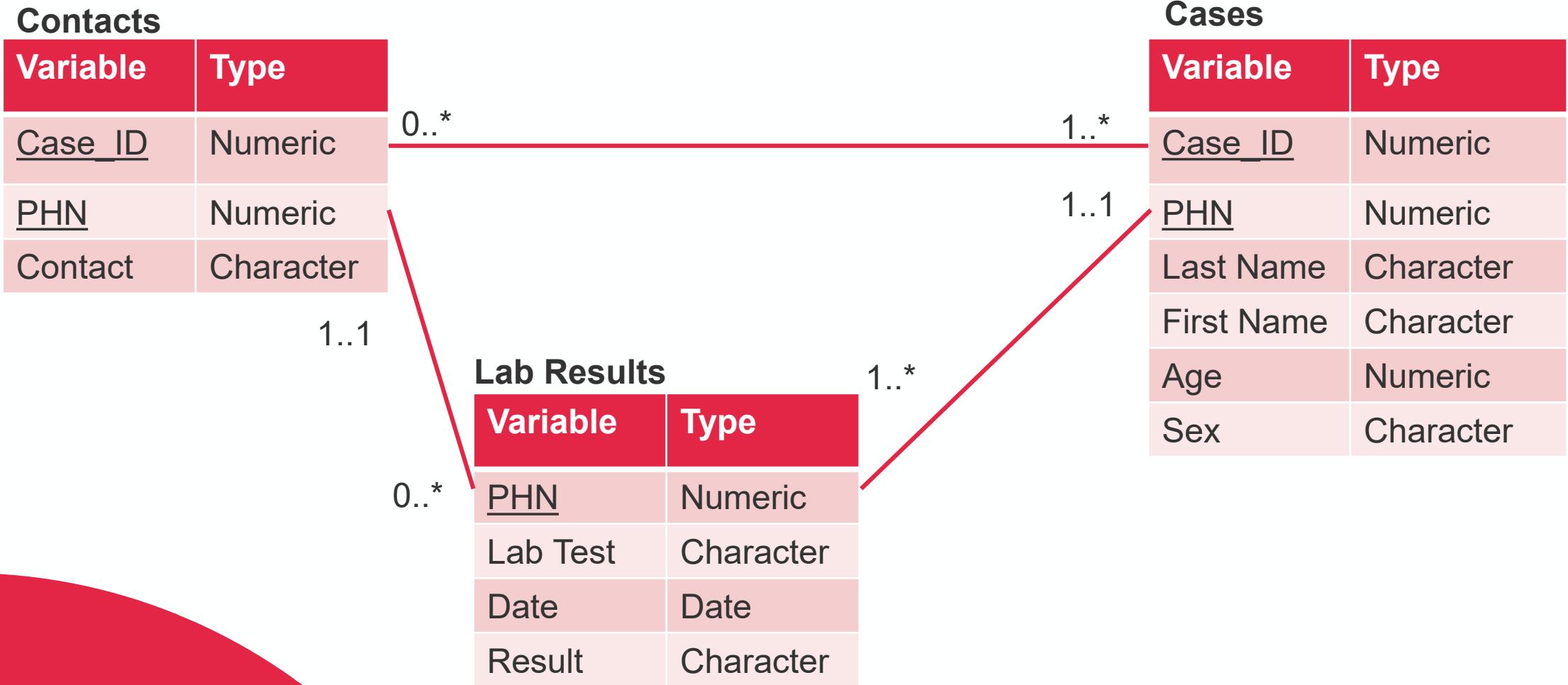
- Databases are often the result of working with more data than is reasonable to keep in one flat file
- We therefore split the data into multiple smaller tables that can be recombined when needed through the use of a *unique identifier*, or *key variable*



RELATIONAL DATABASE

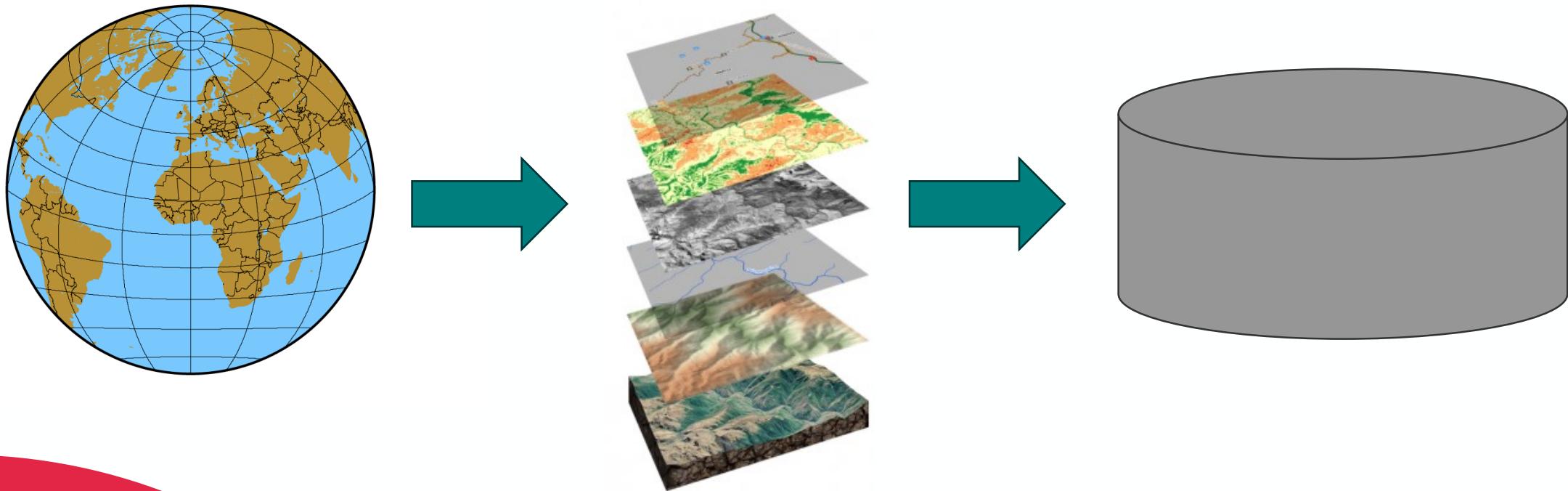
- Store data in tables
- Variables are grouped in logical units:
 - By data source
 - By visit or interaction with system
 - By type of data (i.e. laboratory test)
- Key variables
- It implies logical or organised design of a database with minimal redundancy of data

TABLES AND RELATIONSHIPS

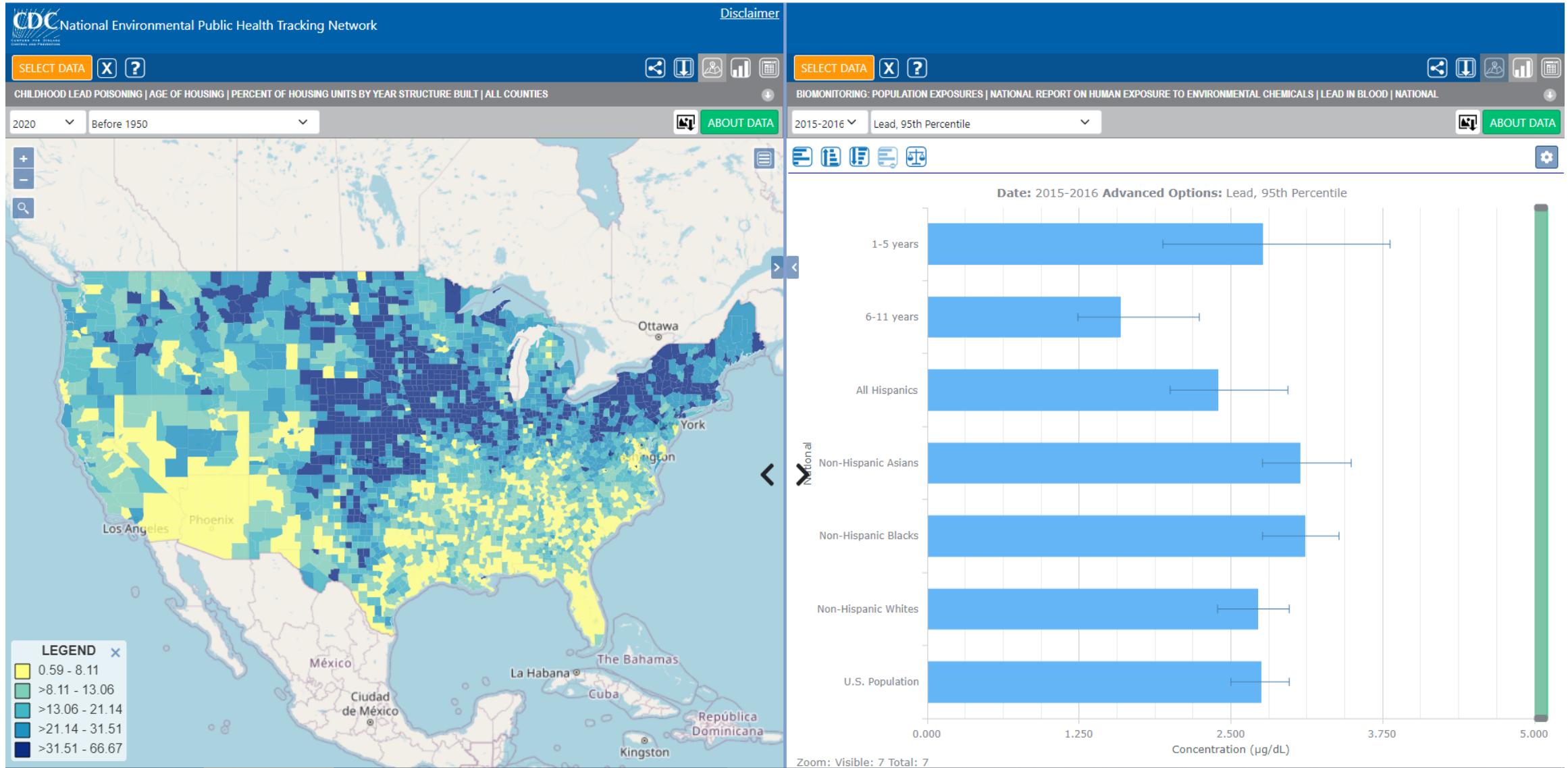


SPATIAL DATA

- Data that are referenced to specific geographic locations
- Geodatabases manage both spatial data and tabular attribute data



SPATIAL DATA



<https://epntracking.cdc.gov/DataExplorer/>

HUMAN-READABLE VS NON HUMAN-READABLE

- Human-readable data:
 - Intended for humans to interpret
 - Line lists, exposure tables, case reports
- Non human-readable data:
 - xml
 - json

DATA MANAGEMENT IN PUBLIC HEALTH

WHAT IS DATA MANAGEMENT?

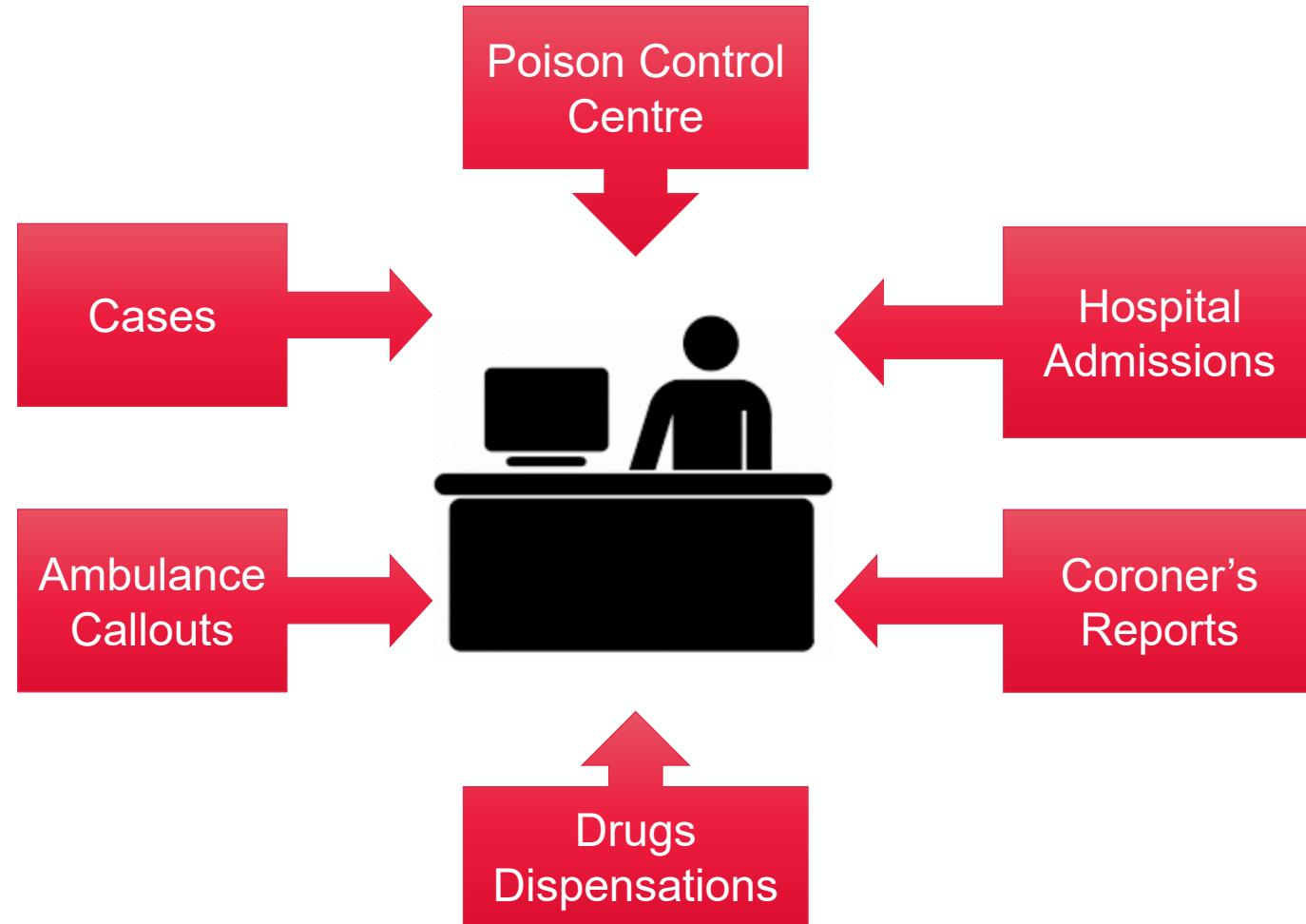
- Data management is a practice that includes methods and strategies for **capturing** (collecting, entering), **storing**, **maintaining** (integrating, validating, cleaning), and **utilising** (analysis, reporting) data, including information **security**.
- A data management **plan** aides in the efficient production of valid results upon which to base conclusions and recommendations.

DATA MANAGEMENT FOR PUBLIC HEALTH



Chiolero and Buckeridge, 2020

DATA MANAGEMENT FOR PUBLIC HEALTH



HOW TO MANAGE ALL THAT DATA?

Tools

Plan Elements

HOW TO MANAGE ALL THAT DATA?

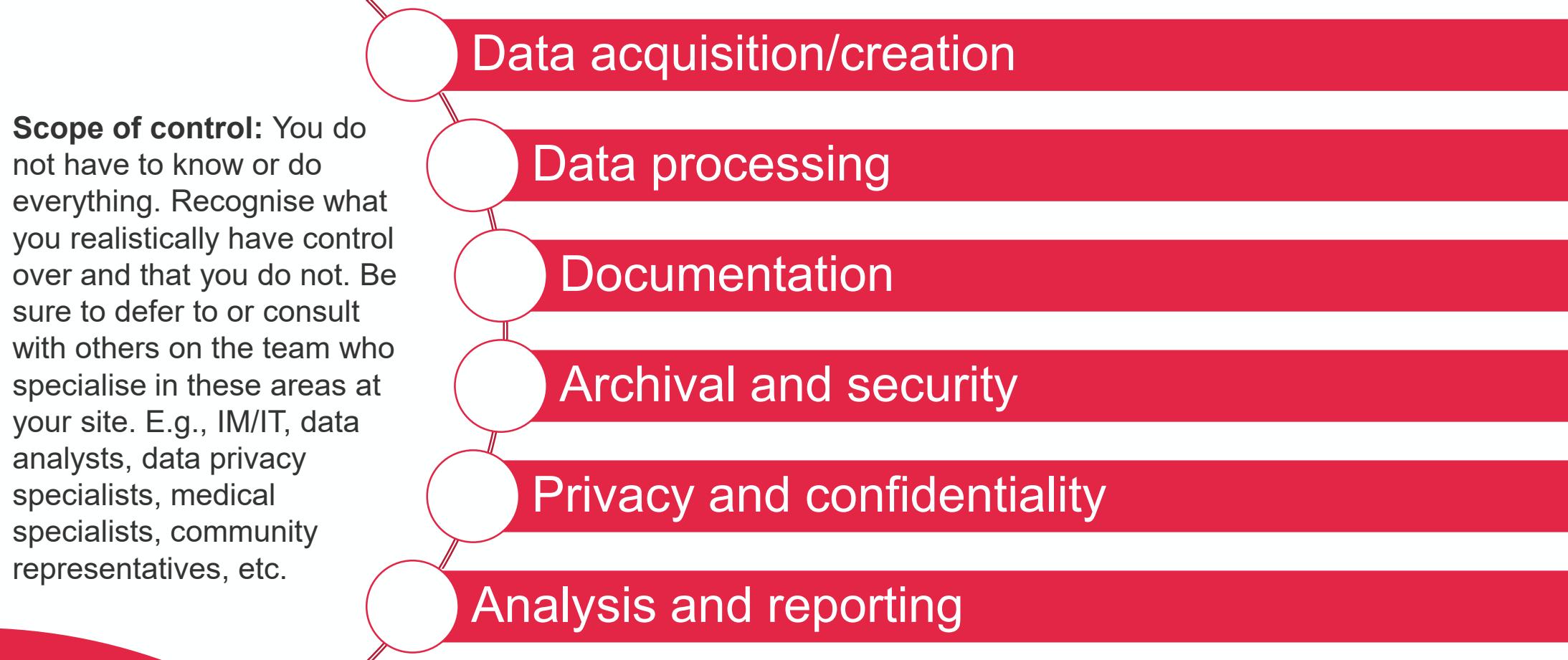
- Data dictionaries
- Data standards
- Data summaries
- Quality checks
- Databases
- Spreadsheets
- Statistical software / pre-programming data processing and analyses
- Data entry forms
- Data sharing / access agreements
- Analysis plans / technical appendices
- Archival plans
- Legislation / F/P/T privacy acts
- Unique identifiers / key variables
- Etc.

DATA MANAGEMENT PLANNING

- Adapted from DMPTool.org

- What data will you collect or create?
What existing data will you access?
- How will the data be collected, created and/or accessed?
- What is your plan for quality assessment, cleaning, collation, and analysis?
- What documentation will accompany the data?
- Are there any ethical issues to be considered, and if so how will they be managed?
- Where will the data be stored?
- How is access and security managed?
- Do you need to make any provisions for retention and long-term storage?
- Are any restrictions on data sharing?
- Who will be responsible for data management?
- How will information be shared?
- What resources will you require to deliver your plan?
- Who needs to be informed of findings, and how must they be formulated to facilitate decision making?

DATA MANAGEMENT AND THE ROLE OF THE EPIDEMIOLOGIST IN THE FIELD



CONSIDERATIONS FOR DATA MANAGEMENT

Implementation

Data Cleaning

All data need cleaning. It is now your job to validate the data and assess quality. Find and fix any errors, improve variable names, and re-format the raw data to suit your desired analysis plan.

Data Analysis Plan

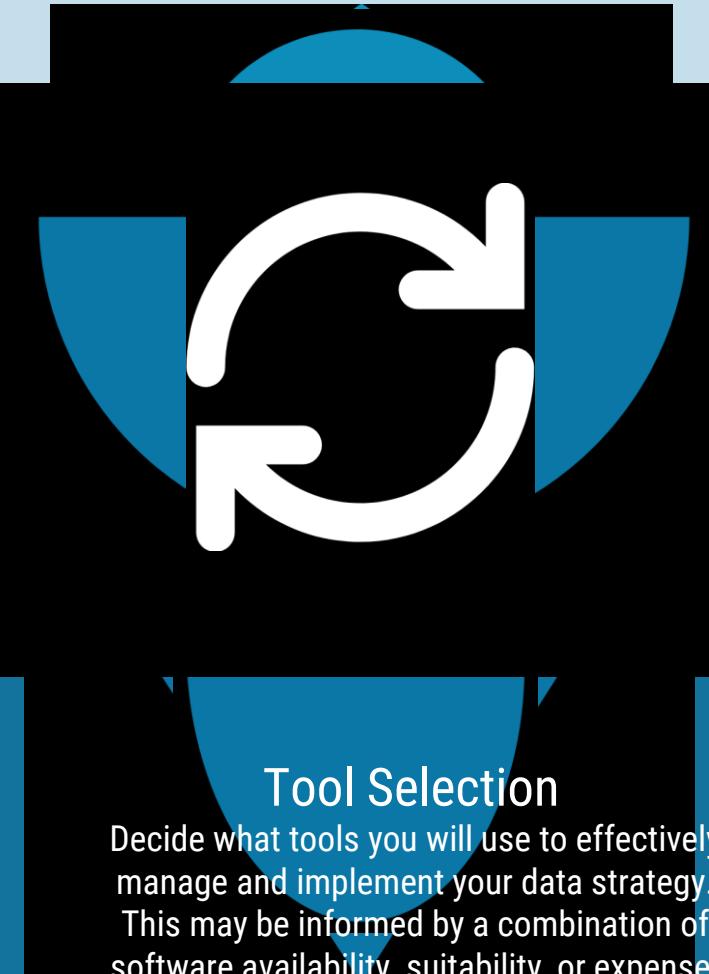
What analyses do you intend to do? Have you considered SGBA+ applications to your work? Will you need multiple datasets? Do all the required data exist already, or will you need to create new variables? When in doubt refer back to your original question(s) and purpose.

Data Collection Plan

Consider what variables you do (and do not) need for your analysis. How will your data be collected? Surveys? Questionnaires? Linking to existing databases?

Data Analysis & Interpretation

The fun part! Time to apply your analysis plan and generate some results. Interpret with caution, applying any appropriate caveats discovered during planning and validation.



Recommendations

Present your work back to your team/community/public. Disseminate your results, interpretations, etc. Don't forget data-based caveats and limitations to interpretation, or recommendations for improving data management and quality in the future.

Community Liaising

Consider engaging affected members in the community as early as possible in the planning stage. Their lived experience and insight into the situation is often instrumental to data collection, interpretation of findings, and informing realistic public health interventions.

Data Governance

Consider what you need for approved access to data. Who will be using or seeing the data? Where will data be housed? How long will access be required for? Who will be the end-users or stakeholders?

Planning

COMMUNITY LIAISON

- Trust
 - Justice
 - Respect for persons, communities and human rights
 - Promoting well-being
 - Minimising harm
 - Working together
-
- Procedural considerations:
 - Accountability
 - Openness and transparency
 - Inclusiveness
 - Responsiveness
 - Intersectionality
 - Cultural humility

[Public Health Ethics Framework: A Guide for Use in Response to the COVID-19 Pandemic, PHAC](#)

DATA GOVERNANCE

Ensuring that throughout the data lifecycle, data are:

- Available for use
- High quality
- Usable
- Consistent
- Safe and secure



PRIVACY AND CONFIDENTIALITY

- Core function of public health requires use of personal information
- Used for numerous public health purposes
- Highly sensitive
- Electronic collection and transmission of information
- Public perception of personal privacy

TERMINOLOGY: PRIVACY AND CONFIDENTIALITY

- Personal information
 - Ethnicity, age, religion, marital status, education, finances, medical results, criminal and employment history, ID numbers, address, blood type, etc.
- Reporting
 - Nominal, non-nominal, anonymous
- Confidentiality
 - Security from inappropriate disclosure
- Residual disclosure
 - Inadvertent disclosure of confidential information through cross reference between sources of publicly accessible information
- Privacy
 - Autonomous individual control over personal information
- Consent
 - Expressed, implied, assumed implied

PRIVACY AND CONFIDENTIALITY

BEST PRACTICES FOR PRIVACY MANAGEMENT:

- Collecting personal information
 - Purpose of collection
 - Possibility of information being useful later does not make its collection appropriate
 - Where possible, collect data that cannot identify an individual (e.g., anonymized, de-identified)
- Informing individuals
 - Purpose of data collection
 - Collection required by law (when relevant)
 - Identity of agency/people to whom information will be disclosed
 - Can access own information

ARCHIVAL AND SECURITY

- Always ensure data are archived
- Location
- Backups
- Accessibility
- Security of backup
- Store data on secure server if necessary
- Raw data, working copy, backups, and archived data require equal security

TOOL SELECTION

Availability
of software

Site
capacity to
use
software

Complexity
of data
processing

Analytical
capacity of
software

Number
and
complexity
of data
sources

Time
available to
complete
tasks

Availability
of learning
resources
for trouble-
shooting

Number of
users –
data entry

Number of
users –
data access

Privacy
approvals

Cost

TOOL SELECTION - SPREADSHEET OR DATABASE?

	Spreadsheet	Database
Easy to learn	✓	
Data analysis	✓	
Data management		✓
Data storage		✓
Calculations	✓	
Data entry		✓
Data cleaning		✓
Ability to combine different datasets		✓
Speed to set up	✓	
Multiple users		✓

Software	Advantages	Disadvantages	Appropriate Uses
Microsoft Excel ^a	User-friendly	Spreadsheets are not a substitute for a database!	Single user, small dataset; quick & dirty graphs/tables
Microsoft Access ^b	Form view; data validation; queries	Complex; requires programming knowledge	Multiple users & sources of data
Epi-Data ^{b,c}	Free; stats/analytical capabilities	Lack of advanced statistical capacity	Designing questionnaires; multiple users
Epi Info ^{b,c}	Free; dashboard (visual summary/analysis)	Some training required; may become slow and buggy	Designing questionnaires; simple analysis
SPSS/SAS/ Stata/R ^c	User support; runs on multiple operating systems	Cost (except R); training required (programming)	Advanced statistical analysis, processing and matching/linkage; sophisticated data visualisations
Go.Data ^d	Free, control over data entry, large datasets, facilitates case and contact management, questionnaires, laboratory data, data visualization, oriented to specific situations (I.e., outbreaks)	Training required, limited availability, lack of advanced statistical capacity, oriented to specific situations (I.e., outbreaks)	To manage and utilise data during outbreak investigations involving pathogens transmissible from person-to-person and includes activities such as contact tracing

^a spreadsheet application software

^b relational database management software

^c statistical analysis software

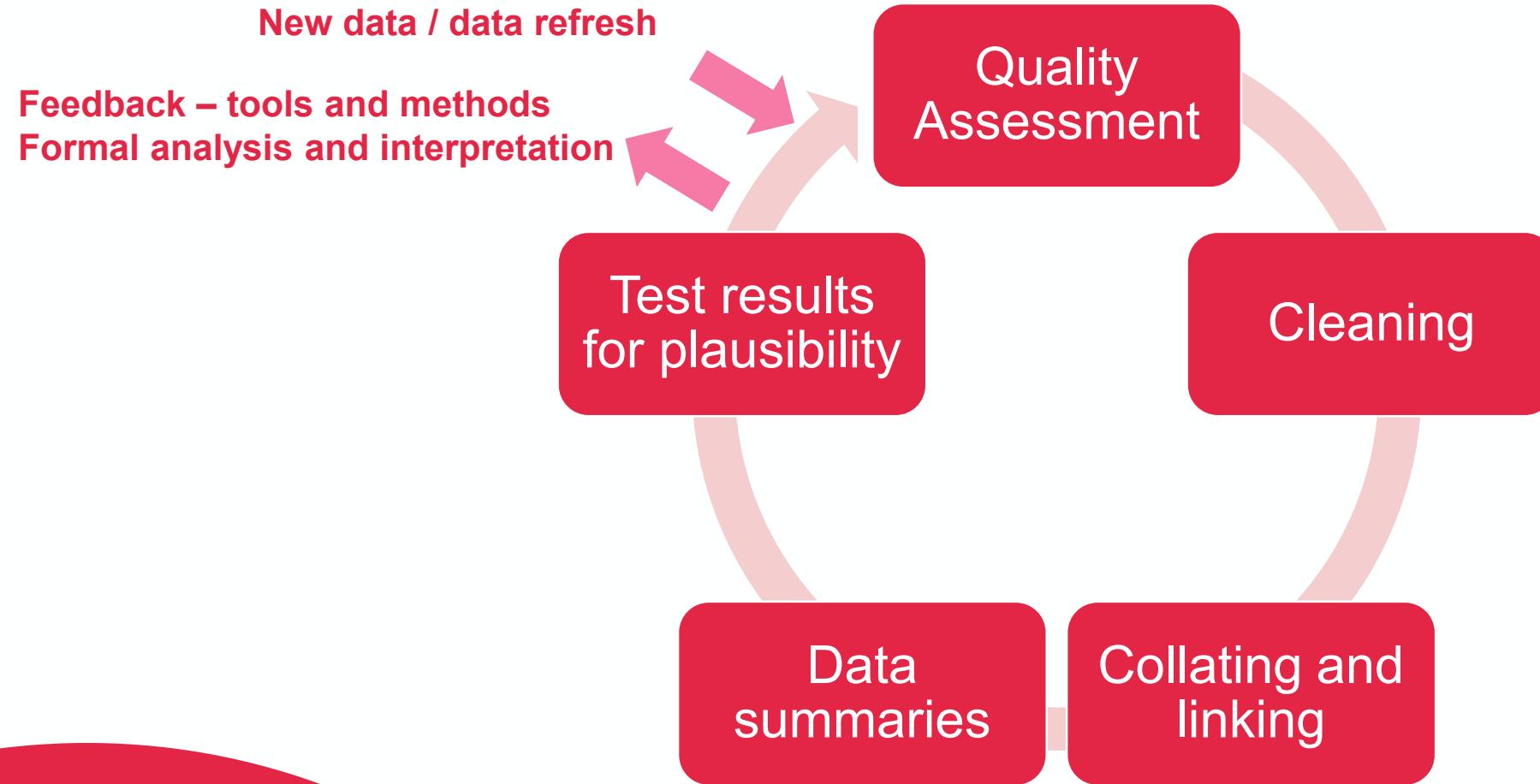
^d outbreak management software

DATA GENERATION AND COLLECTION

- Review of existing data
- Establish data sharing/access agreements
- Data collection and capture
- Database design and construction
- Data entry

Seek only the data you need!

DATA PROCESSING



VALIDATION

- Data validation refers to how you restrict or control the input of data into certain variables in your table or database
- E.g.,
 - Date variables should only contain dates
 - Text variables should contain only text
 - Numeric variables should only contain numbers
 - Etc.
- Depending on the software you're using, you may have different strategies for controlling what data are permissible under each variable

1. UNIQUE IDENTIFIERS

- Unique identifiers are *absolutely critical* and an important consideration when collecting your data.
- A unique identifier should be given to each individual present in your dataset to avoid confusion

FirstName	LastName
Michael	Smith
M	Smith

E.g., How do you know if these are the same or different person?

1. UNIQUE IDENTIFIERS

- Unique identifiers are *absolutely critical* and an important consideration when collecting your data.
- A unique identifier should be given to each individual present in your dataset to avoid confusion

Because the unique ID is different, you know that these records are distinct

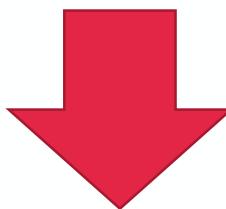
ID	FirstName	LastName
CAN0001	Michael	Smith
CAN0002	M	Smith

1. UNIQUE IDENTIFIERS

- Unique identifiers are also required for combining data from multiple tables in a database

ID	FirstName	LastName
CAN0001	Michael	Smith
CAN0002	M	Smith

ID	Sex	Age
CAN0001	M	32
CAN0002	F	45



ID	FirstName	LastName	Sex	Age
CAN0001	Michael	Smith	M	32
CAN0002	M	Smith	F	45

2. DATES

- Many date formats exist, and can change the interpretation of your data
- E.g: “11-11-02”
 - November 11, 2002?
 - November 02, 2011?
- Be wary of dates in programs like Excel:
 - $11-11-02 = 40849$?

PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27½-13 2013.158904109
MMXIII-II-XXVII MMXIII ^{LVII}_{CCCLXV} 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013 Mississ 229
10/11011/1101 02/27/20/13 01237 234

2. DATES

Recommendations:

- Maintain a consistent format for all dates in your dataset
- Be explicit about the format in the header (e.g., “Date (yyyy-mm-dd)”)
- Always double-check the format being applied to your dates in the software you’re using

ANALYSIS AND REPORTING

- Stakeholders and partners (information for action)
- Production and display of indicators
- Dissemination
- Formulation of recommendations

INFORMATION PRODUCTION

- Caveats for interpretation
 - In essence, things related to the data which place boundaries or limits on how they are interpreted once analysed
 - A "hole" in the data? I.e., a critical group is not represented --> engage with the affected community ; collaboratively identify and acquire more data or information
 - Other missing pieces? I.e., you have enough for a sketch of the situation, but not enough for a fulsome picture --> identify and acquire more data or information from a variety of sources
 - Quality issues? I.e., there is something to address during data collection, processing, analysis, etc. that makes certain aspects of the data potentially unreliable --> identify what needs to change and how
 - Interpretation
 - Providing a snap-shot of the situation based on the data you've analysed
 - It is important to think critically about the data upon which your recommendations are made – how they may limit our capacity to utilise the information we've just produced
 - Used to inform feasible and realistic recommendations, which are then used to make decisions

DISSEMINATION AND DECISION MAKING

- Stakeholders and partners (information for action)
 - Know from the outset who needs to receive the information you will produce
 - Understand why they need to know and how your message can be tailored to that audience and purpose
 - Report back to the public
 - Report to Chief Medical Officer of Health
- Formulation of recommendations
 - When we look at and analyse our data, it implies a duty to act upon what we find
- The point of why we do what we do with our data!

SUMMARY

- In the practice of public health and applied or field epidemiology, information is used to inform meaningful action
- In acquiring/creating, managing, and using data we have a responsibility to ensure that:
 - Data are kept safe and secure
 - Privacy and confidentiality are maintained
 - Our work is transparent and transferrable through the documentation that accompanies our datasets/databases
 - Data are of sufficient quality to produce information that will inform appropriate recommendations and decisions
 - Ethical considerations for the populations being served

SUMMARY

- A data management plan will assist in ensuring all obligations and responsibilities are met by providing a clear path forward from the outset whether you are working on surveillance, outbreaks, or research
- Ultimately, data must be accurate, secure, and timely to support meaningful public health interventions and actions

TRIVIA! NAME THE CLASSIC MOVIE MONSTER

1



[This Photo](#) by Unknown author is licensed under [CC BY-NC-ND](#).

2



[This Photo](#) by Unknown author is licensed under [CC BY-SA](#).

3



[This Photo](#) by Unknown author is licensed under [CC BY-NC-ND](#).

4



[This Photo](#) by Unknown author is licensed under [CC BY-SA](#).

5



[This Photo](#) by Unknown author is licensed under [CC BY-SA-NC](#).

6



[This Photo](#) by Unknown author is licensed under [CC BY-NC-ND](#).

DATA MANAGEMENT HORROR STORIES

Turn your lemons into super lemons!

DATA MANAGEMENT HORROR STORIES

- Part 1
 - Describe the issues with data management presented in your scenario.
 - What problems are occurring?
 - How are they impacting, or are likely to impact, the quality of work being performed by the epidemiologist and public health team in the scenario?
 - Provide suggestions for how the management of public health data could be improved in your scenario.
 - For each of the issues that you've identified:
 - Provide an alternate suggestion,
 - Describe how this would lead to improving the data management in the scenario.

DATA MANAGEMENT HORROR STORIES

- Part 2
- 5 minute presentation to the large group
- Prepare to share your scenario back to the large group.
- Tell everyone your spooky horror story, the scary violations in data management presented, and the resolutions that you and your teammates brainstormed to arrive at a safe ending.
- Feel free to assign speaking roles, use creative voices, and create illustrations on your jam-board to share back to the group.

DATA MANAGEMENT HORROR STORIES



Example:

- "Curse of the Mummy"
 - You've recently started a new job where your tasks pertain mainly to community disease surveillance data and reporting
 - The database software storing and maintaining surveillance data, MUMMY, is woefully antiquated
 - You cannot find a manual or other helpful information and you are pretty sure that the software vendor no longer exists!
 - You've been asked to extract all the data from each of the tables, and work with them in software outside of MUMMY as an interim solution while a plan to modernize the surveillance system is underway
 - You become concerned very quickly that the counts you produce are widely different from counts published in reports over the last few years. You can't find any details on how data tables were combined during the extraction process from MUMMY before your arrival...

DATA MANAGEMENT HORROR STORIES

Data Management Issues Identified:

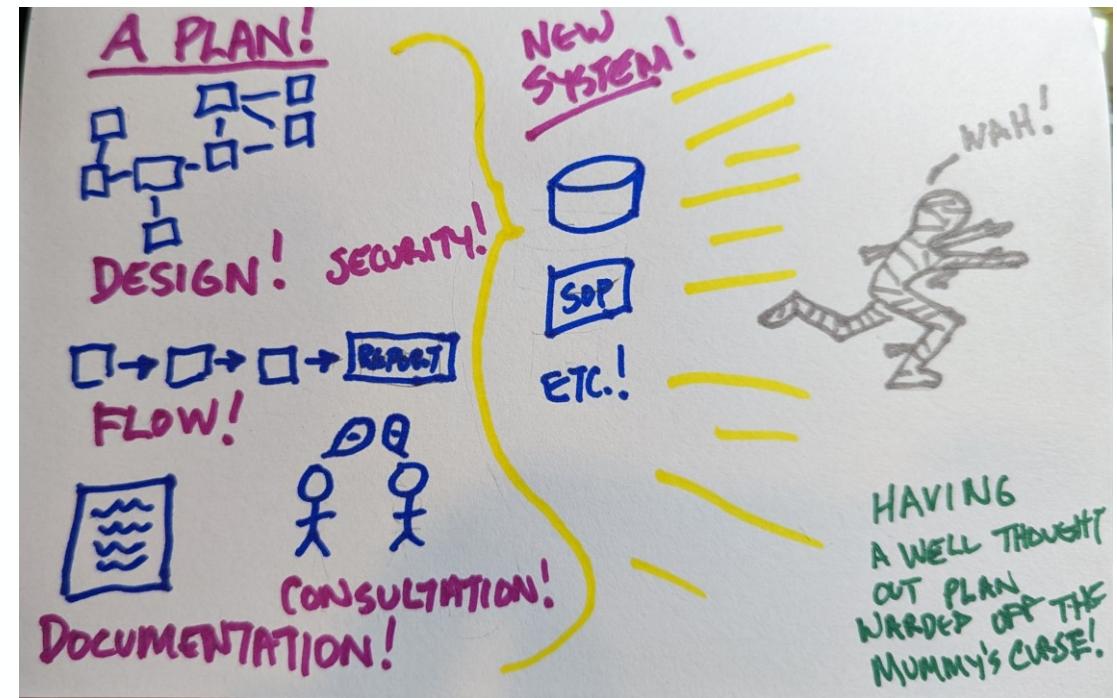
- The database software is complex and difficult to work with
- It is unclear how the team used the MUMMY database to extract data for reporting, or if there is anyone on the team who can fill you in
- In the absence of documentation, it is unclear if your approach to assembling, processing, and analyzing data outside the system is accurate

Suggested Solutions for Improvement:

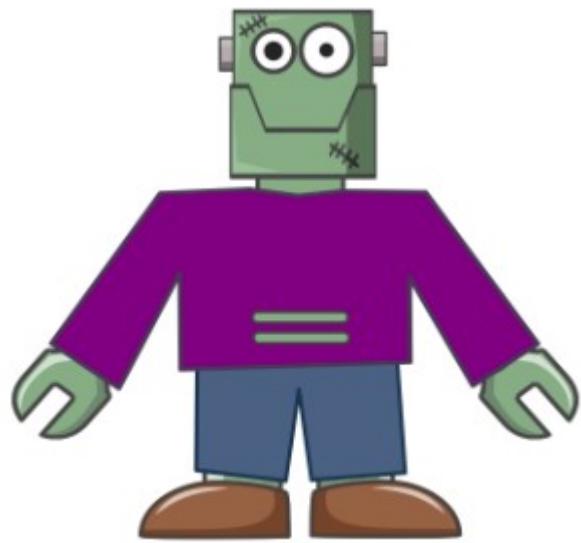
- If possible, recreate the same approach to data assembly outside the system
- Confirm the process of data capture
- Confirm that the team did not capture data elsewhere in parallel or in lieu of the existing system (e.g., Excel spreadsheets)
- Dataflow should be reviewed, standardized, and documented
- A new database software solution that is simpler and easier to use should be implemented
- Continue to discuss with teammates about what may exist in the team's file share network or corporate memory

DATA MANAGEMENT HORROR STORIES

"There once was a database named MUMMY
It produced information that was funny
So, we designed a new system with care
And made sure all the data were in there
And now my reports aren't so crummy"



DM HORROR STORIES



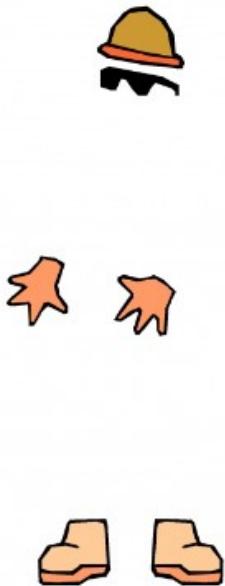
Group 1



Group 2



Group 3



Group 4

INTERSESSION ACTIVITY

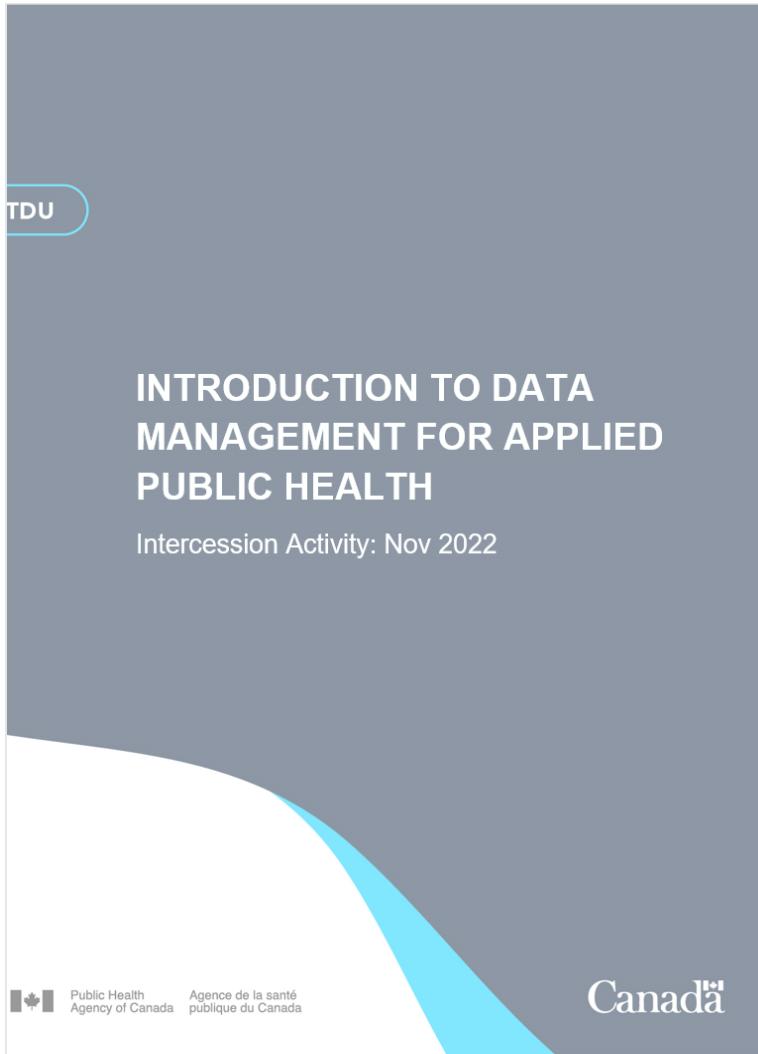
INDEPENDENT ACTIVITY – REQUIRED DOCUMENTS

Filename	Description
DM.Intersession.Activity.docx	Heat-Dome Scenario – Instructions (independent activity)
DM.Intersession.Activity.KEY.docx	Heat-Dome Scenario – Answer Key
DM.Nov2022.activity.linelist.xlsx	Heat-Dome Scenario - Excel file
Excel101 (folder)	Excel 101 Resources

INDEPENDENT ACTIVITY - SCHEDULE

- Feel free to turn off your cameras, go have a break, work through the activity independently
- Facilitators will remain in the virtual classroom – please pop in and ask questions if you need help!
- Let us know if you and a partner would like to work together, we can put you in a break-out room!
- If you leave the zoom session, **please return by 3:10PM ET**

SMALL GROUP DISCUSSION



SMALL GROUP DISCUSSION

- Describe the caveats and limitations of your analysis (e.g., what assumptions did you make when cleaning the data? Are there any data you omitted? Are there any gaps in the dataset you were provided? What else would you need to know to provide a more complete picture of the scenario?)
- Based on your interpretation, and given the caveats and limitations discussed above – provide a list of next steps and recommendations for improving data quality and facilitating interpretation. What would be your recommendations for ensuring that you or your team have all the information required to make well-informed decisions?

WRAP-UP

LEARNING OBJECTIVES

- By the end of the training participants will be able to:
 - Explain principles of data management in the context of applied public health practice;
 - Identify common ways in which data are stored for use in public health investigations;
 - Compose caveats for interpretation and formulate recommendations to improve data quality; and
 - Execute descriptive epidemiological analyses to summarize and interpret data in the context of a public health investigation.

EVALUATION

- Please complete the course evaluation!

www.slido.com
#DMNov2022



REMINDER

- Cohort 48 and Public Health Officers
 - Intro to R – 11:30 AM ET start time
 - Feel free to join us after 11:00 AM ET to settle in and chat