

Using Data Literacy to Drive Insight

Live Webcast

September 17, 2020

11:00 am PT

Today's Presenters



Jim Fister
Principal
The Decision Place



Glyn Bowden
Chief Architect, AI & Data Science Practice
HPE

SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
 - Any slide or slides used must be reproduced in their entirety without modification
 - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

SNIA-At-A-Glance



185

industry leading
organizations



2,000

active contributing
members



50,000

IT end users & storage
pros worldwide

What We Do



Educate vendors and users on cloud storage, data services and orchestration



Support & promote business models and architectures: OpenStack, Software Defined Storage, Kubernetes, Object Storage



Understand Hyperscaler requirements
Incorporate them into standards and programs



Collaborate with other industry associations

Agenda

- What is data literacy?
- The data of the pandemic
- Understanding data provenance
- The power of data aggregation
- Cleaned vs Raw data
- Critical Analysis
- Summary

What is data literacy?

...and who needs it?

What is Data Literacy?



The ability to create, read, understand and communicate data as information



Assessing the information by leveraging multiple data sources



Applying external context to the data set in an appropriate manner



Asking the right questions of that data

Who Needs to Have Data Literacy Skills?



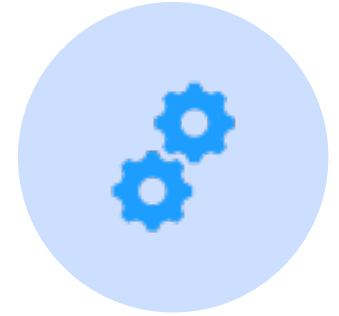
DATA SCIENTISTS AND
DATA ENGINEERS



INFORMATION
ARCHITECTS



OPERATIONS
ENGINEERS



TECHNICAL DECISION
MAKERS

..in fact

EVERYONE

We all need to interpret the information offered to us by people, press, journals, educators, colleagues, friends

The data of the pandemic

More data, more opinions!

The Data of the Pandemic



COVID-19 has bombarded the public with more “data sources” than any event in history



We see statistics on infection rates, deaths, R0 numbers



We see clinical data comparing COVID-19 with pandemics of the past



We see medical data on pre-existing conditions and risk



We see cultural data on which communities might be impacted more



We see economic data of how that impact has manifested



We see political data on why we should ignore other data

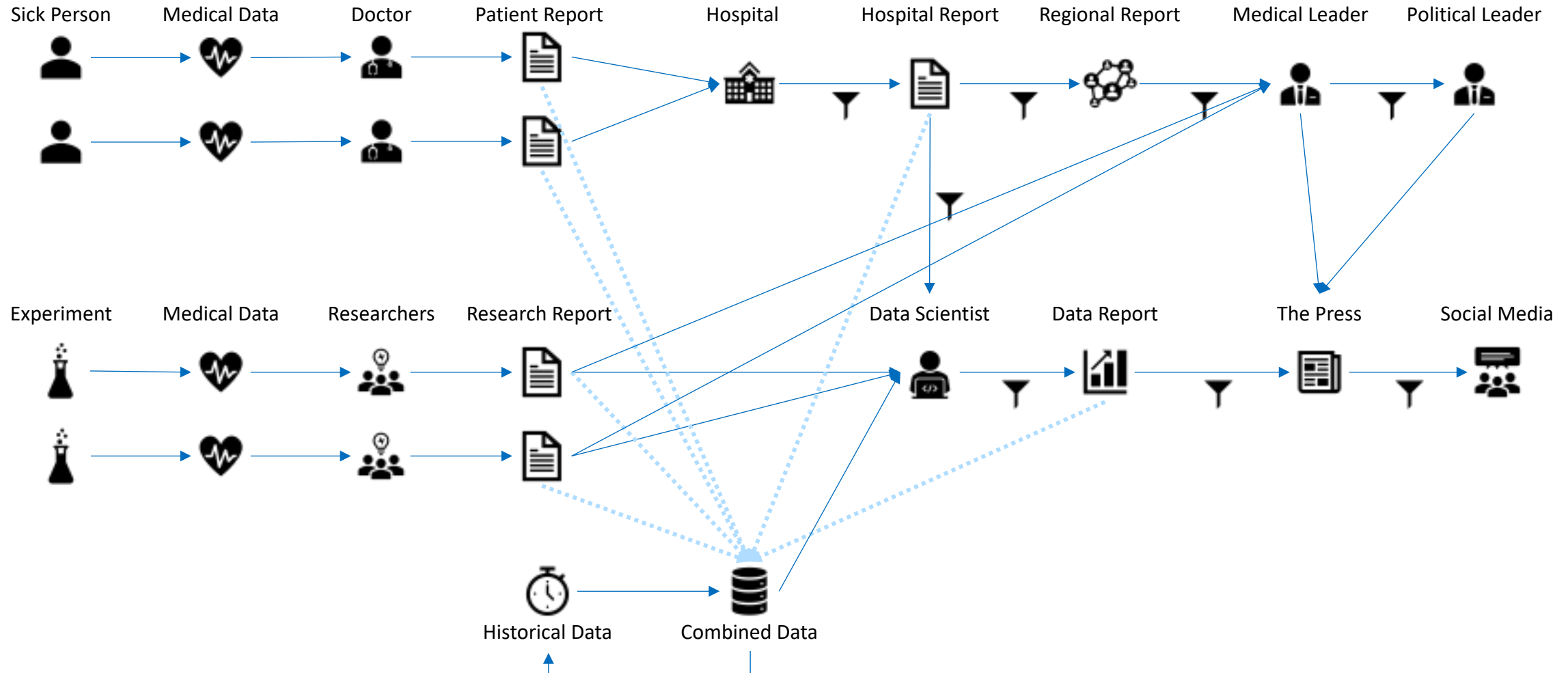


How much of this data is INFORMATION, and how much OPINION?

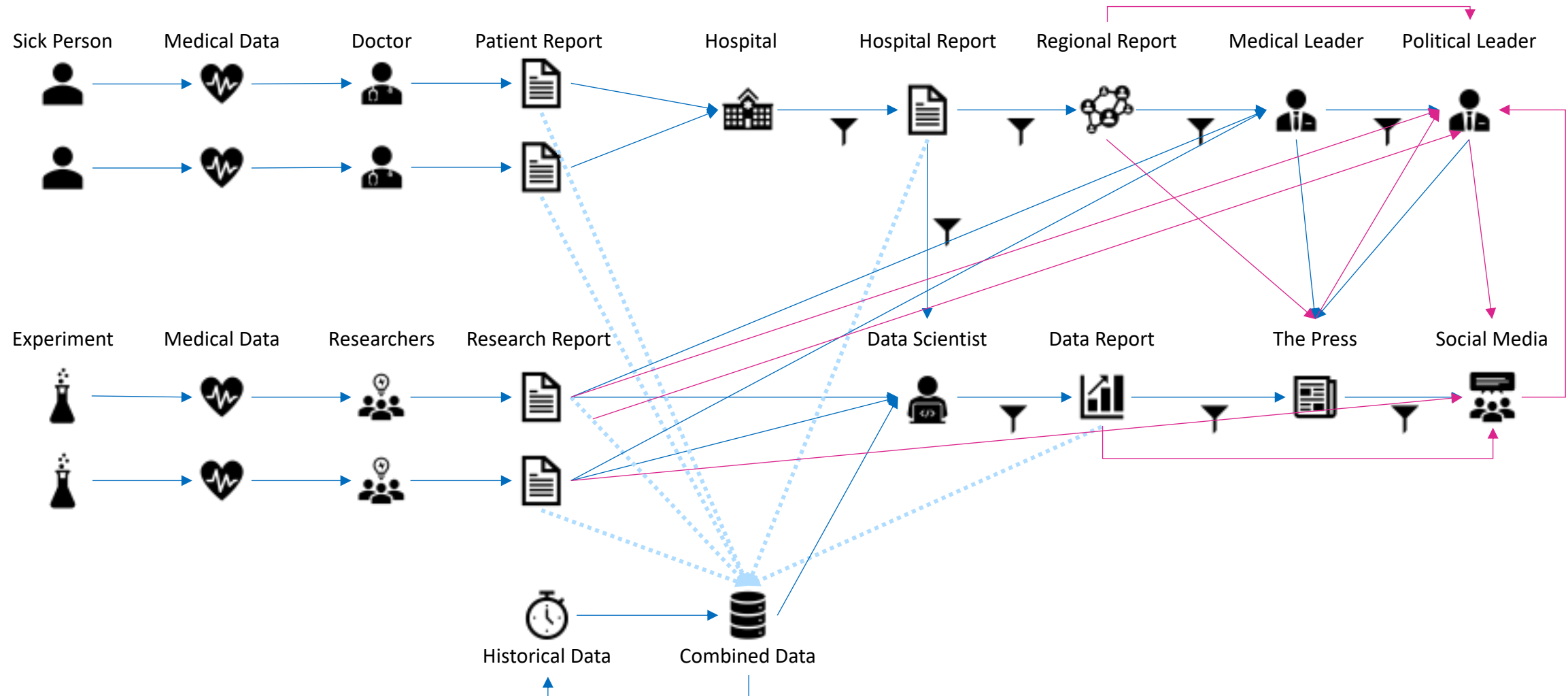
Understanding data provenance

The history of data

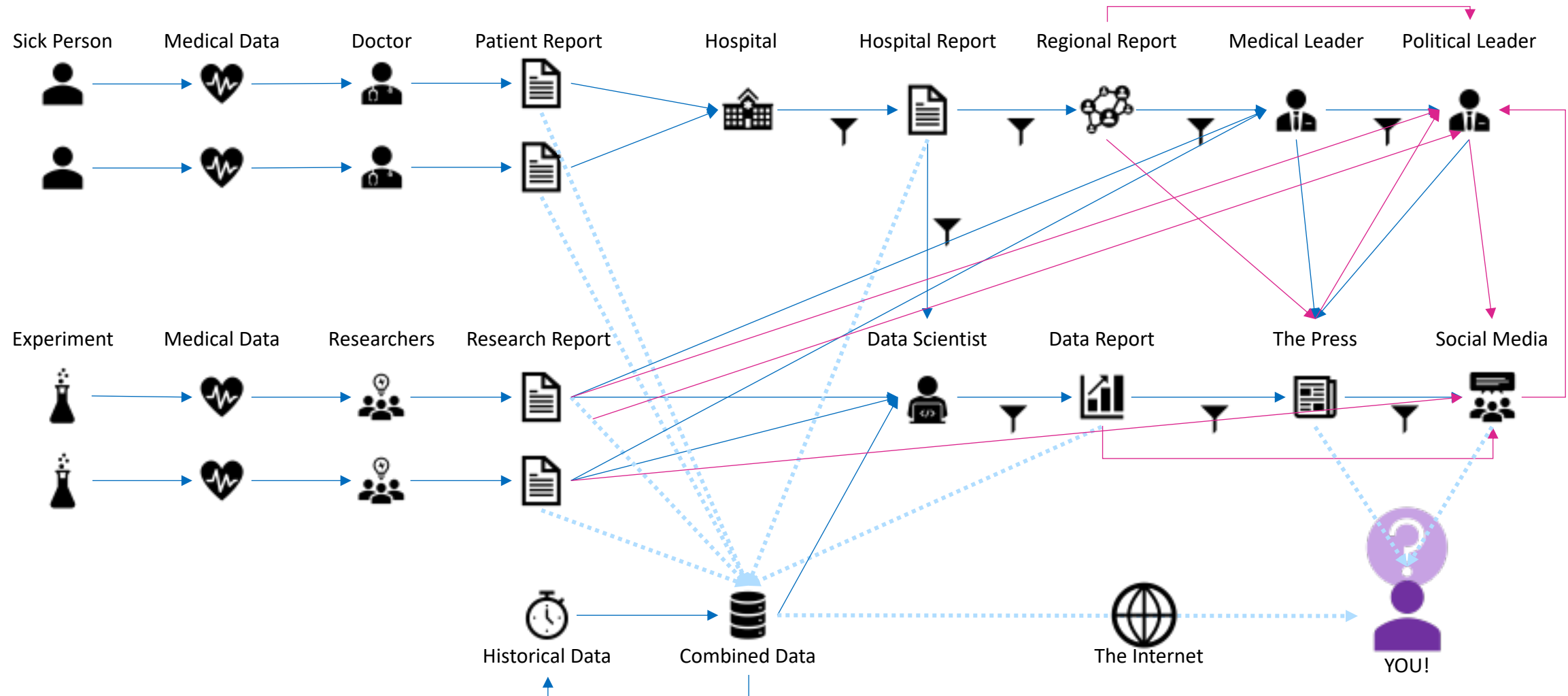
Understanding Data Provenance (standard)



Understanding Data Provenance (reality)



Understanding Data Provenance (reality)



The power of data aggregation

The sum of the parts

The Power of Data Aggregation

UNDERSTANDING

Historical Data



What happened?

Sick Person



Medical Data



What is happening?



Experiment

Medical Data



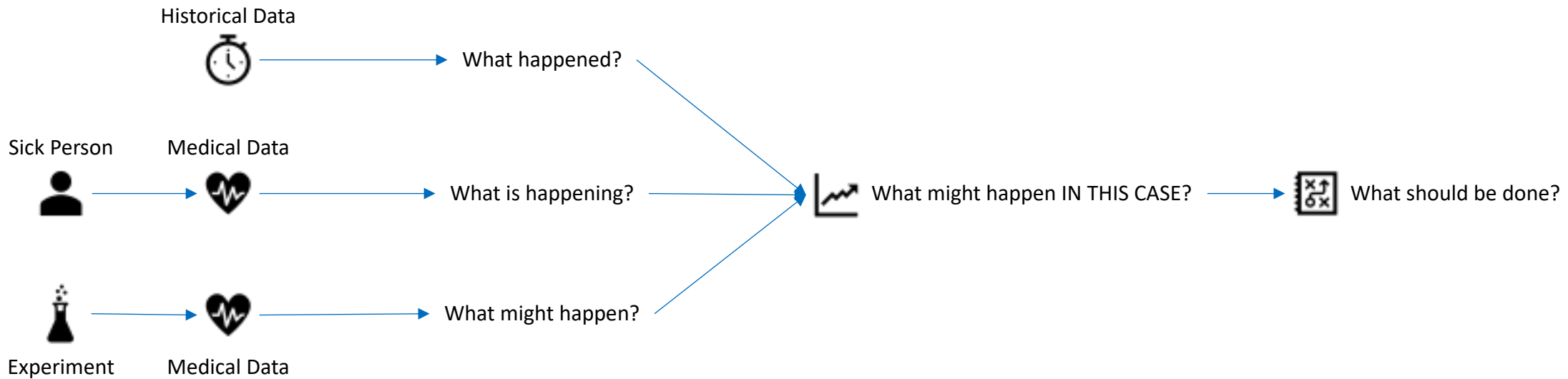
What might happen?

The Power of Data Aggregation

UNDERSTANDING

PREDICTION

PRESCRIPTION



The Power of Data Aggregation

- **Seek out supporting data**
 - Generally only summary data is provided for public consumption
 - Ask what has been left out? Why?
 - Does more data exist that could support or challenge the conclusions?
 - Look for data that particularly clarifies supposition and opinion
- **Additional data can refine the context or drastically change it!**
 - All data is presented with a context in mind.
 - This might be different than the context it was collected in.
 - Ensure the data is validated under any new context

Cleaned vs Raw data

When to cook the books

Cleaned vs Raw Data

Raw data is that which is gathered directly from the source

- Sensors
- Software produced (logs etc)
- Raw survey results

Raw data isn't perfect

- Contains gaps, outliers deliberately incorrect entries, errors!

Cleaned data removes the rough edges

- Gaps are either removed completely or “smoothed” with aggregation to ensure it does not impact final results
- Some corrections of outliers and “errors” are human judgement

Aggregated data usually relies on cleaned data rather than raw

- Reports assume outliers and gaps have been resolved
- As the aggregation layers increase the accuracy resolution decreases

Summary

- Data literacy is something that would benefit anyone
- Although pandemic used as example, this is of course transferrable to any data
- These are the skills being used by data scientists in most organizations, these demands will translate to impact on storage and data platforms.
- Understanding data means understanding its meta-data too.
 - Where is it from?
 - Who created it and for what purpose?
 - What data is related to it that can support it?
 - When was it created?

After This Webcast

- Please rate this webcast and provide us with feedback
- This webcast and a copy of the slides will be available at the SNIA Educational Library <https://www.snia.org/educational-library>
- A Q&A from this webcast will be posted to the SNIA Cloud blog: www.sniacloud.com/
- Follow us on Twitter @SNIACloud



Thank you!