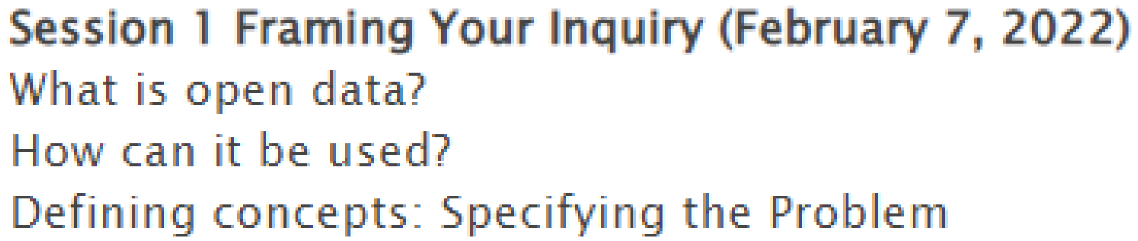
Open Data Boot Camp

SESSION 1

Wende A. Mix, Ph.D.

Introduction to Open Data

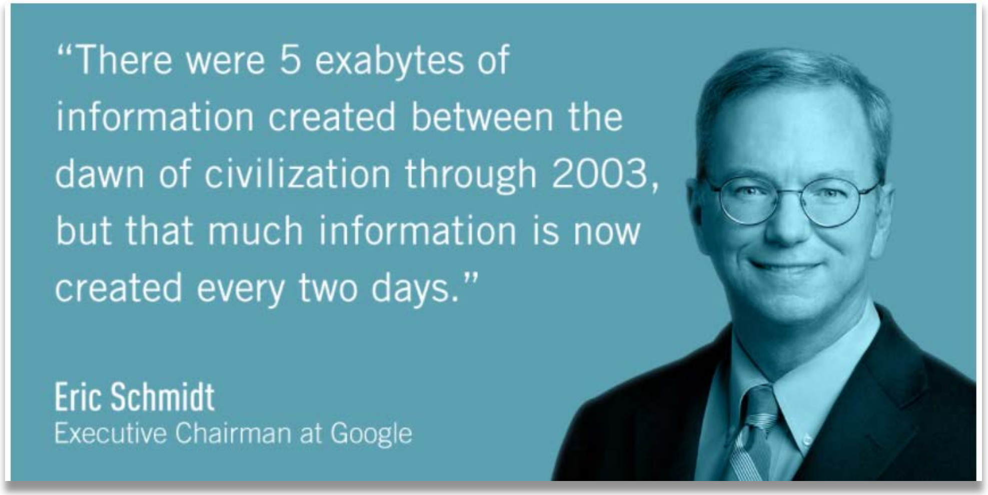


Session 1 Framing Your Inquiry (February 7, 2022)

What is open data?

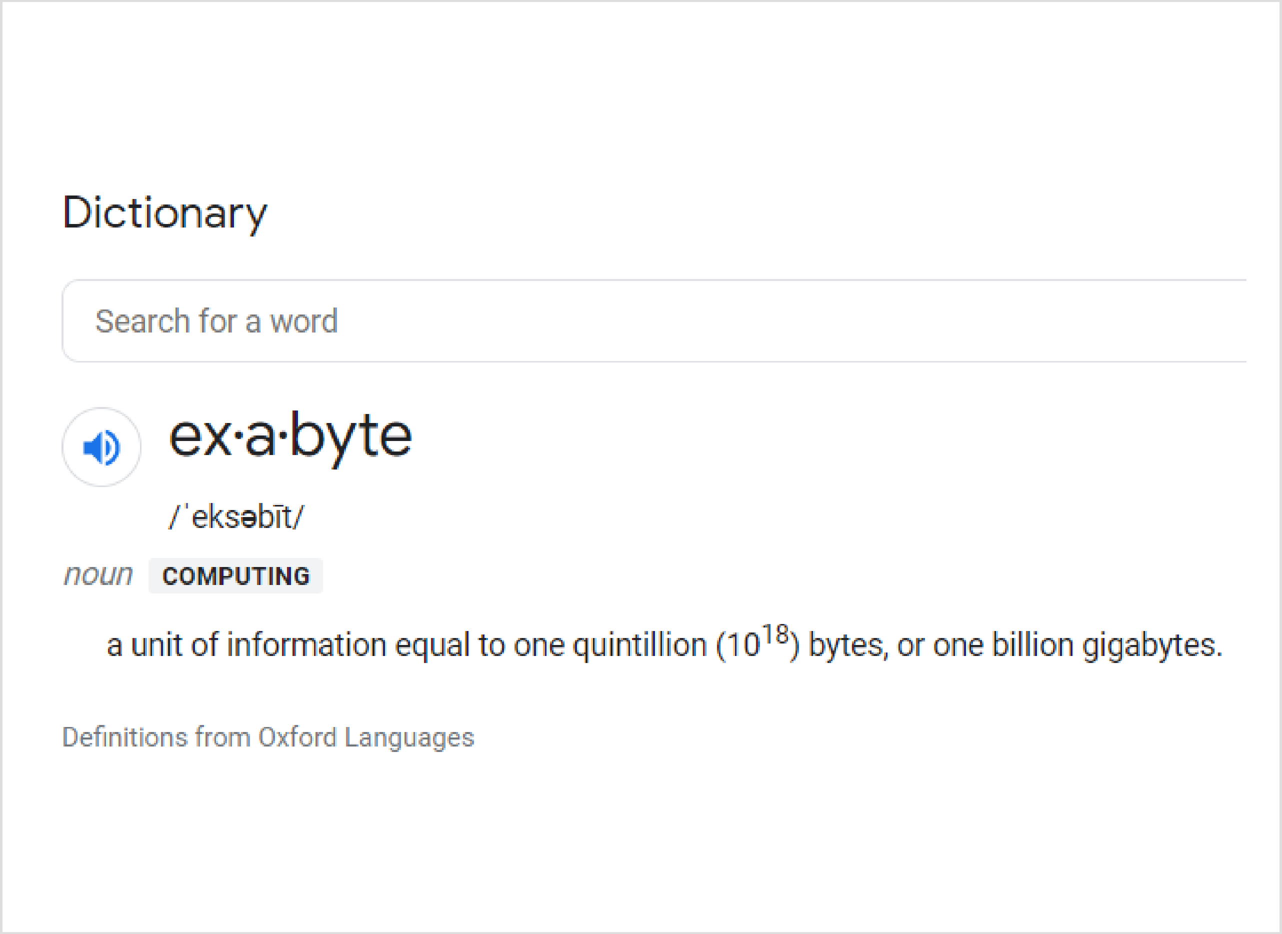
How can it be used?

Defining concepts: Specifying the Problem



“There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days.”

Eric Schmidt
Executive Chairman at Google



Dictionary

Search for a word

ex·a·byte

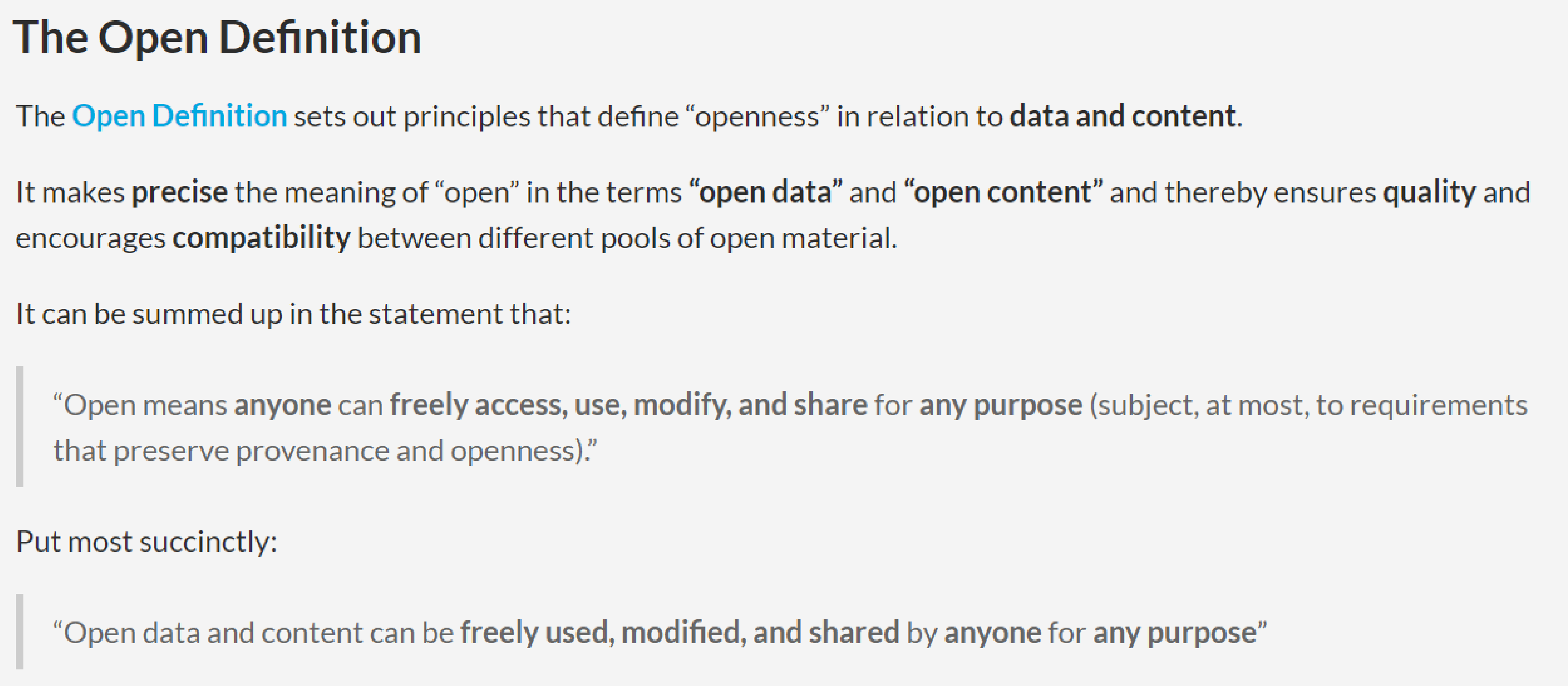
/ˈeksəbīt/

noun **COMPUTING**

a unit of information equal to one quintillion (10^{18}) bytes, or one billion gigabytes.

Definitions from Oxford Languages

OPEN DATA?



The Open Definition

The Open Definition sets out principles that define “openness” in relation to **data and content**.

It makes **precise** the meaning of “open” in the terms **“open data”** and **“open content”** and thereby ensures **quality** and encourages **compatibility** between different pools of open material.

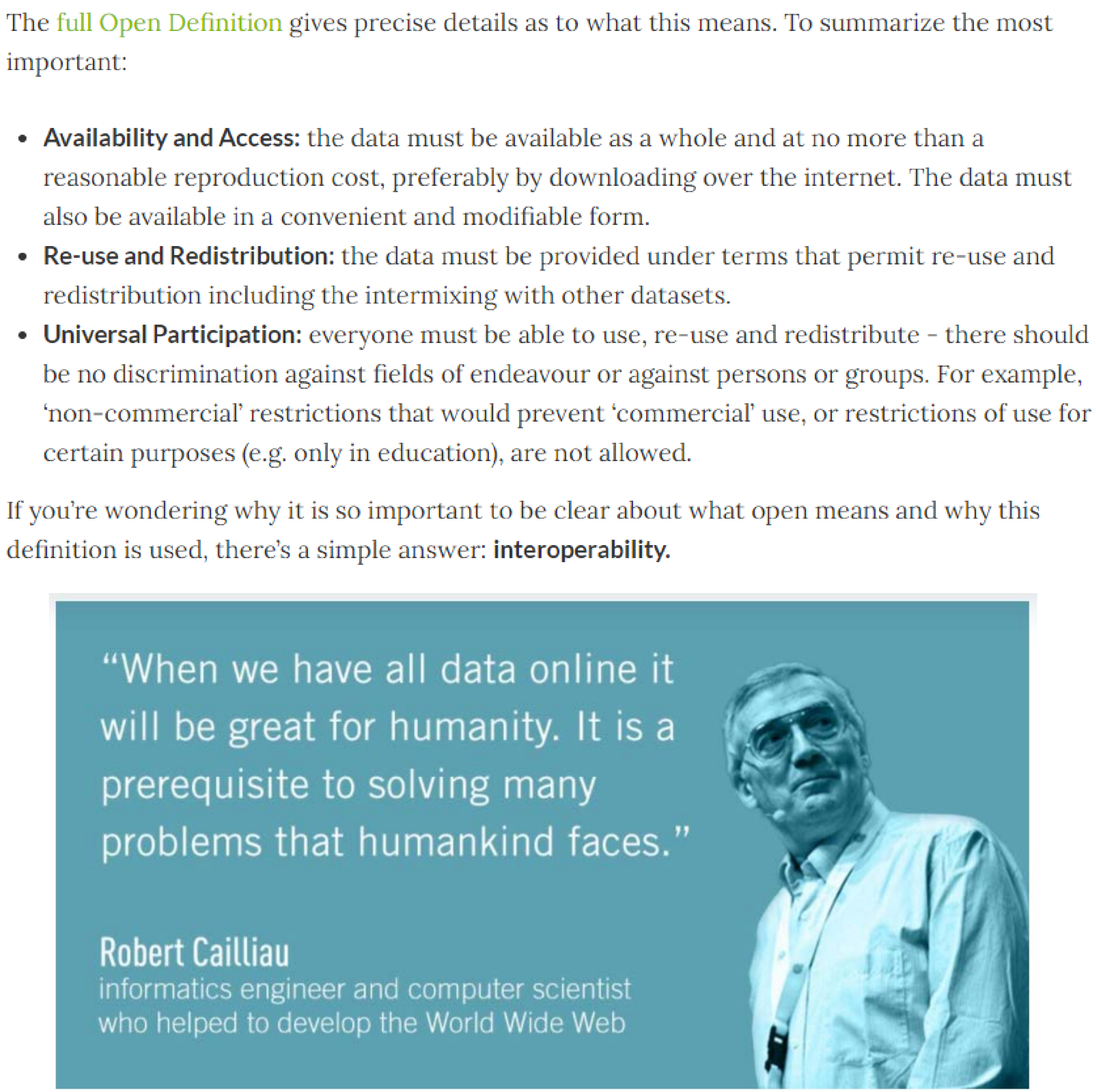
It can be summed up in the statement that:

> “Open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness).”

Put most succinctly:

> “Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**”

Open Data Handbook



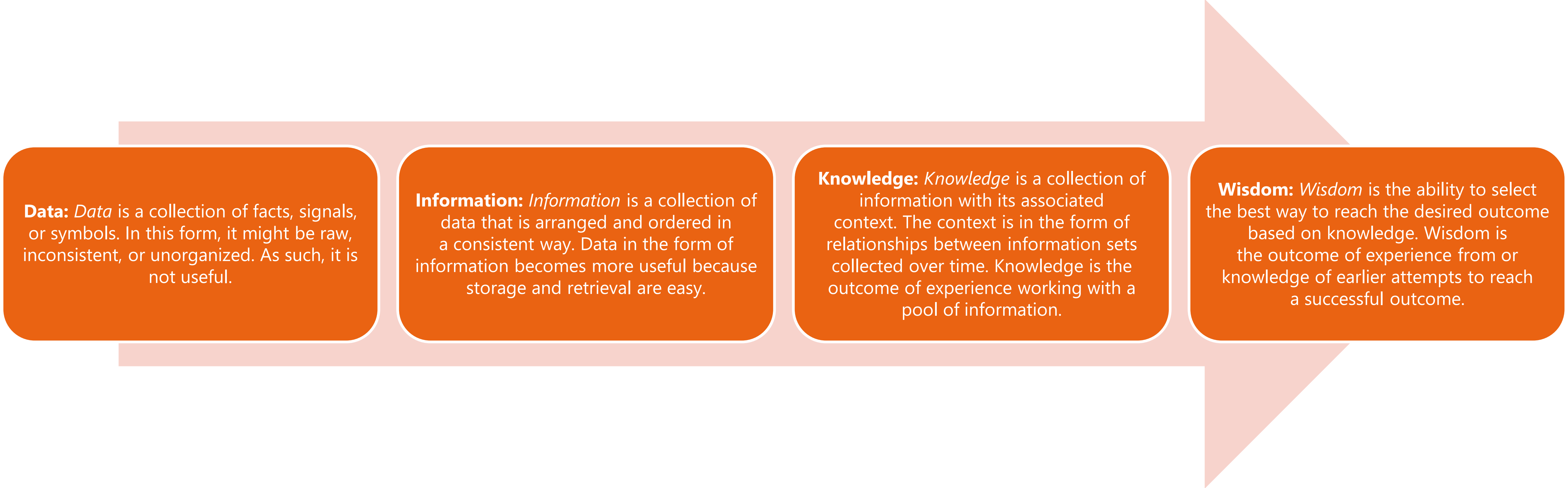
The full Open Definition gives precise details as to what this means. To summarize the most important:

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Re-use and Redistribution:** the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

If you're wondering why it is so important to be clear about what open means and why this definition is used, there's a simple answer: **interoperability.**

> "When we have all data online it will be great for humanity. It is a prerequisite to solving many problems that humankind faces."
>
> **Robert Cailliau**
> informatics engineer and computer scientist who helped to develop the World Wide Web

DIKW Model



Data: *Data* is a collection of facts, signals, or symbols. In this form, it might be raw, inconsistent, or unorganized. As such, it is not useful.

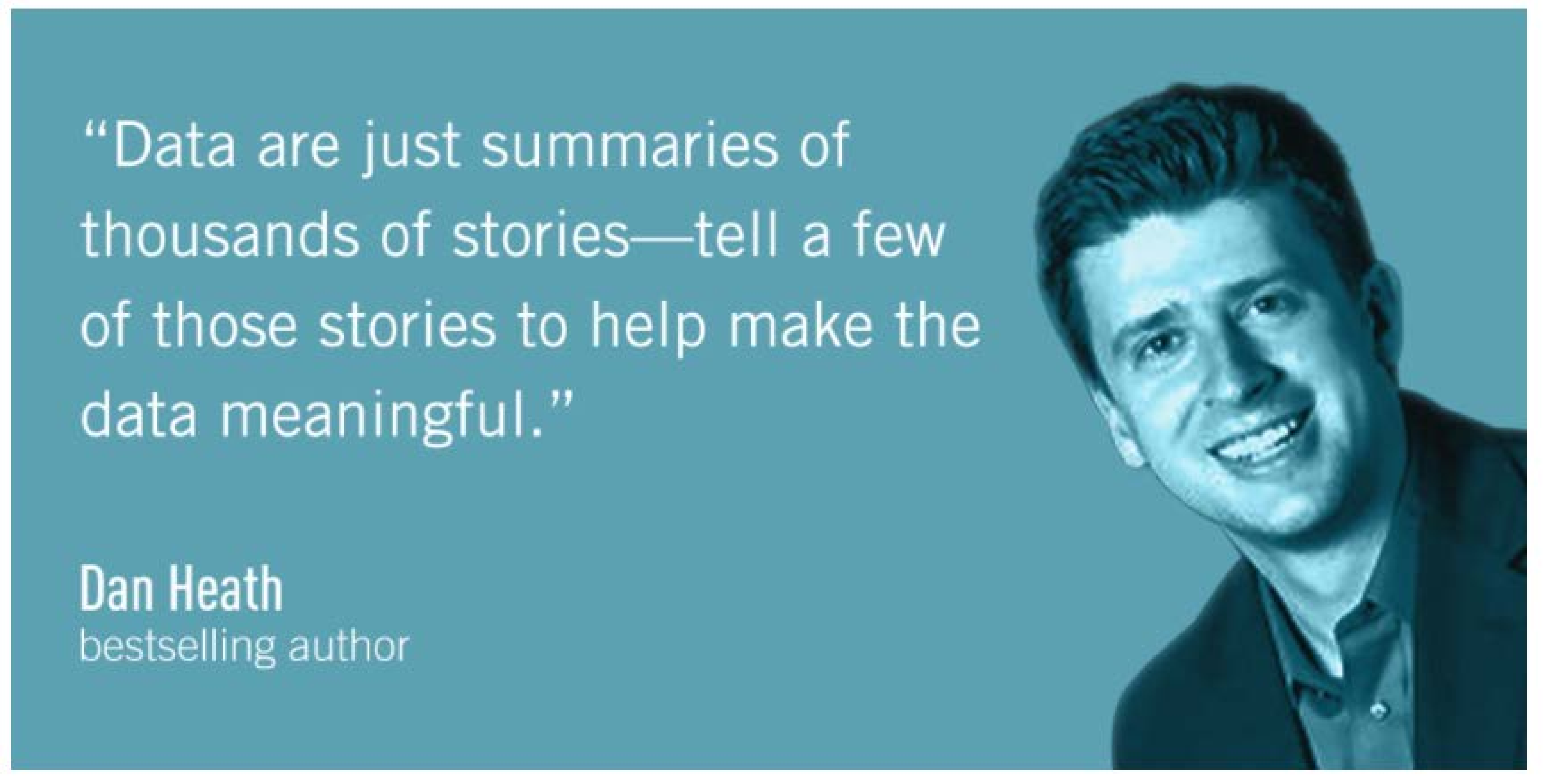
Information: *Information* is a collection of data that is arranged and ordered in a consistent way. Data in the form of information becomes more useful because storage and retrieval are easy.

Knowledge: *Knowledge* is a collection of information with its associated context. The context is in the form of relationships between information sets collected over time. Knowledge is the outcome of experience working with a pool of information.

Wisdom: *Wisdom* is the ability to select the best way to reach the desired outcome based on knowledge. Wisdom is the outcome of experience from or knowledge of earlier attempts to reach a successful outcome.

https://developer.ibm.com/articles/ba-data-becomes-knowledge-1/

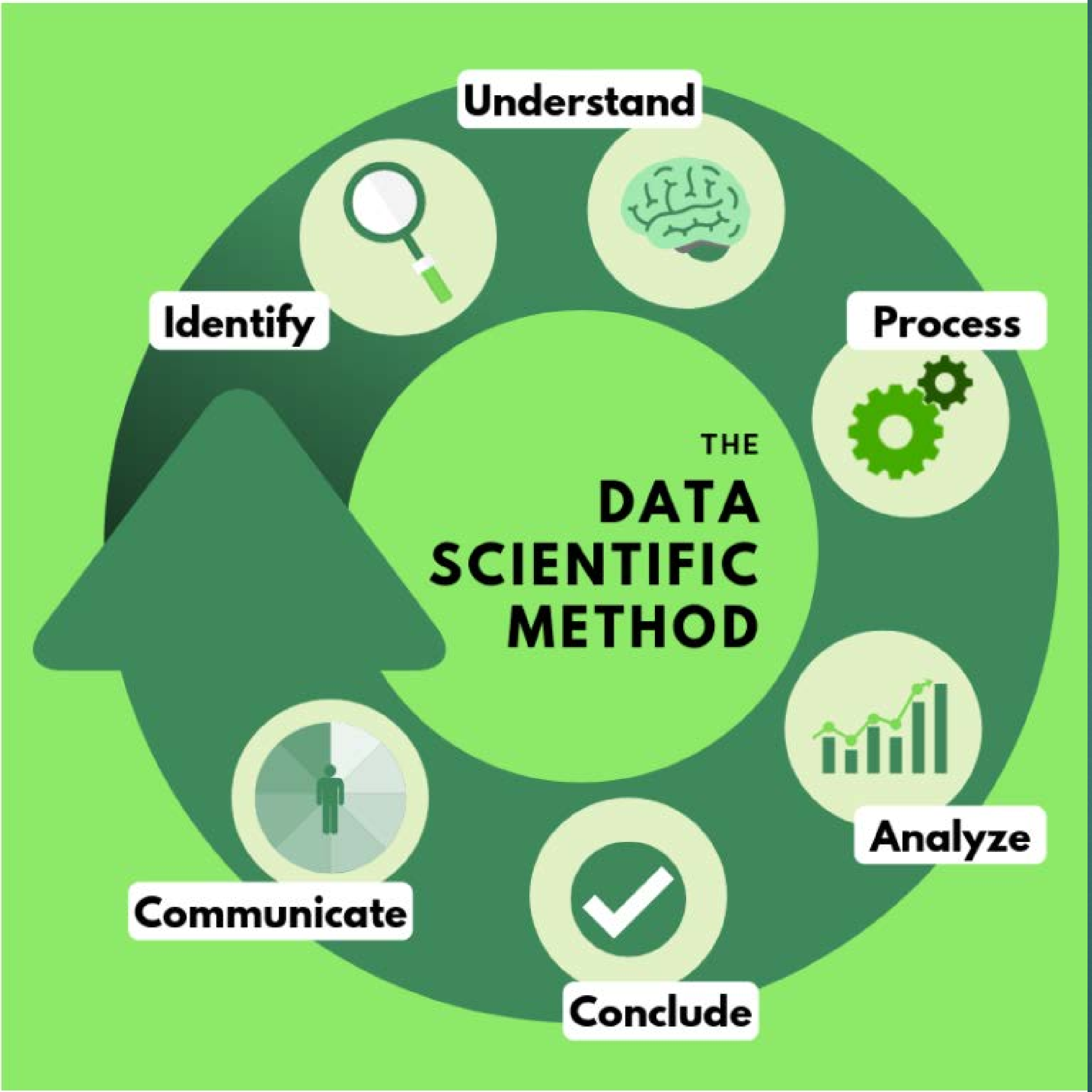
DIKW ?



> “Data are just summaries of thousands of stories—tell a few of those stories to help make the data meaningful.”

Dan Heath
bestselling author

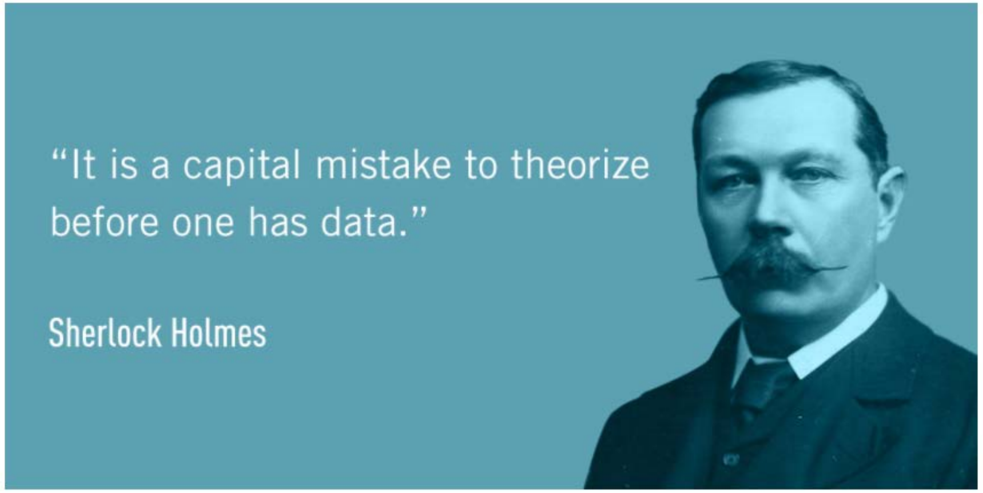
The Data Scientific Method



- Identify – questions?
- Understand – data content and structure, quality
- Process – get data into a state that is ready for analysis.
  - Cleaning, integrating, wrangling, 80% of effort
- Analyze – find patterns and relationships, statistics, modeling
- Conclude
- Communicate – storytelling, depends on audience

https://towardsdatascience.com/a-data-scientific-method-80caa190dbd4

Identify



"It is a capital mistake to theorize before one has data."

Sherlock Holmes

Questions posed

- What data do you need to answer those questions?
  - Is there more crime in neighborhoods with higher poverty rates?
  - Do rural areas or poor urban neighborhoods have worse access to the internet?

Data exploration

- Finding patterns in the data
  - Spatial and / or temporal patterns of violent crimes
  - Spatial and / or temporal patterns 311 calls for service

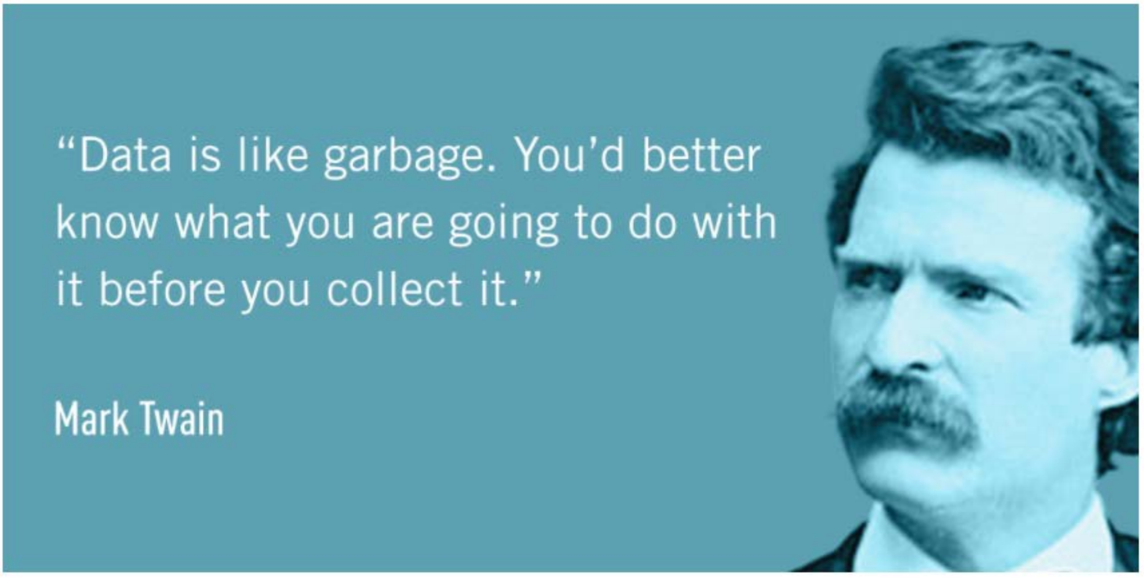
Identify - Define concepts

- Define Study Area
- Determine Spatial Scale
- Determine Temporal Scale
- Vague wording –What do these concepts mean?
  - Crime
  - Poverty
  - Neighborhood
  - Rural
  - Urban
  - Access to the internet
  - Violent crime

Understand

- Why is the data collected?
- How is the data collected?
  - Administrative records,
  - Crowdsourcing, citizen science,
  - Sensors
- How frequently is the data updated?

Metadata – data about the data!



“Data is like garbage. You’d better know what you are going to do with it before you collect it.”

Mark Twain

Understand

- Population versus Sample
  - All events versus (random) subset of events
- Data set structure
  - Tables – rows and columns (fields)
    - Each row = single observation
    - Each row represents sub event
  - Collection of text or images
  - Spatial / temporal characteristics
- Quantitative (numeric) / Qualitative (text/classes) fields
  - Field ranges,
  - missing/null values,
  - observed or derived