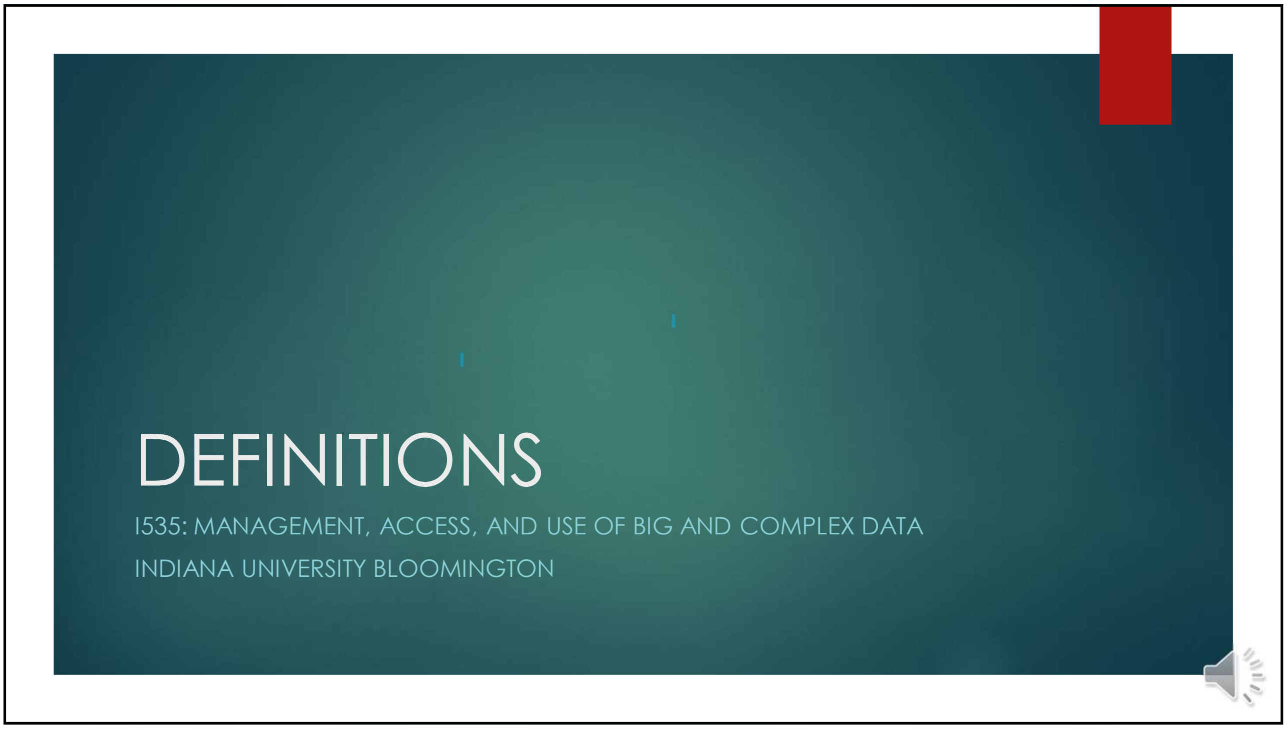
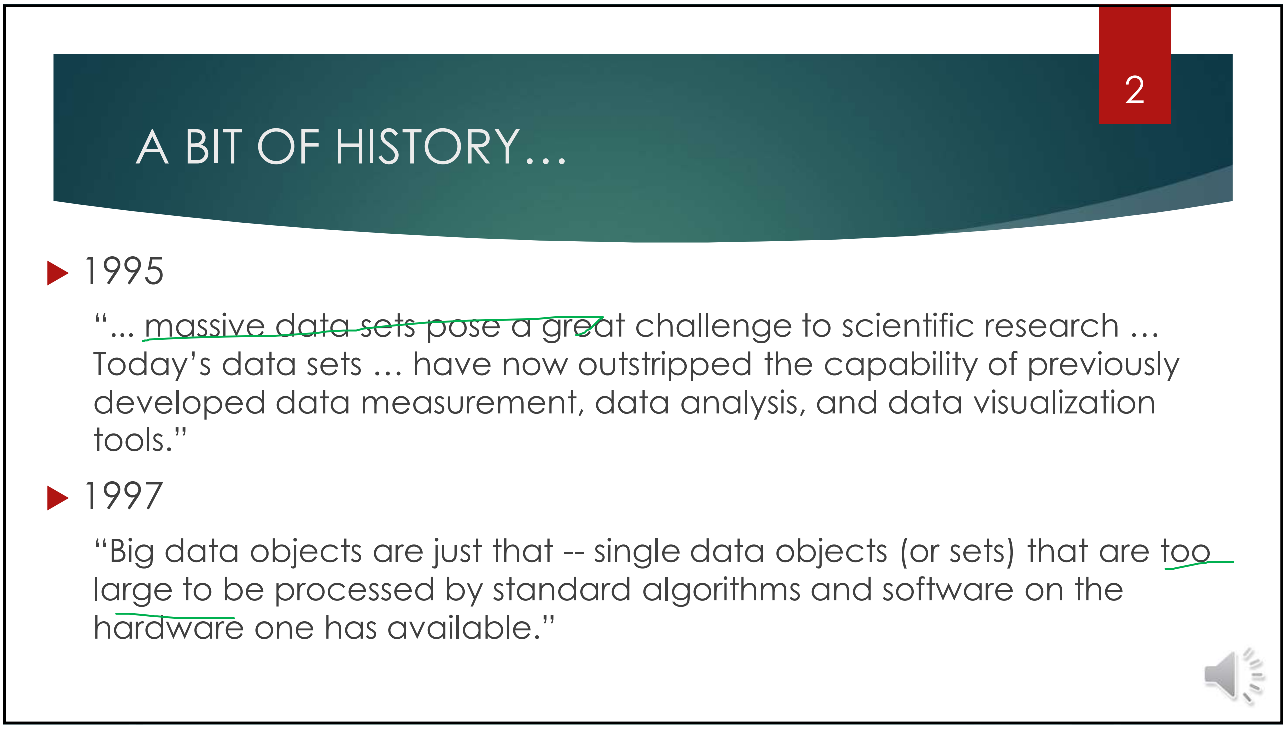


DEFINITIONS

I535: MANAGEMENT, ACCESS, AND USE OF BIG AND COMPLEX DATA

INDIANA UNIVERSITY BLOOMINGTON

1



A BIT OF HISTORY…

2

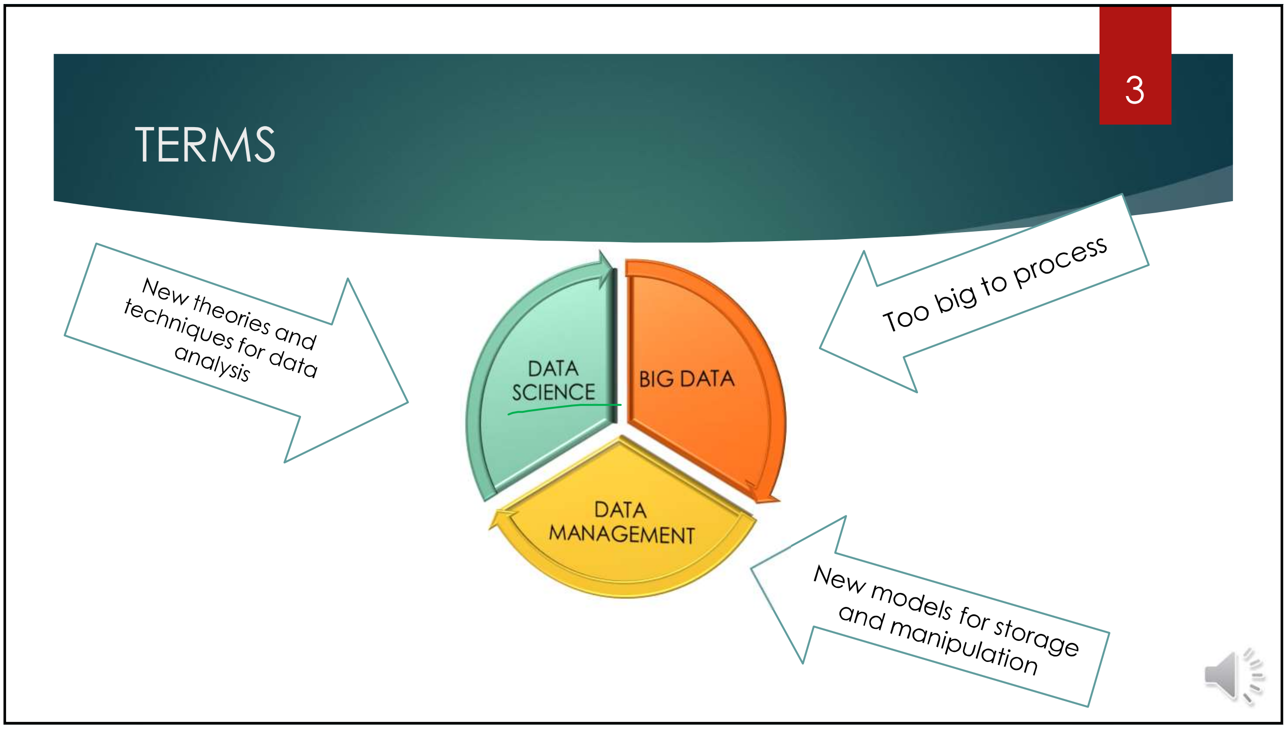
- 1995

  "… massive data sets pose a great challenge to scientific research … Today's data sets … have now outstripped the capability of previously developed data measurement, data analysis, and data visualization tools."

- 1997

  "Big data objects are just that -- single data objects (or sets) that are too large to be processed by standard algorithms and software on the hardware one has available."

2

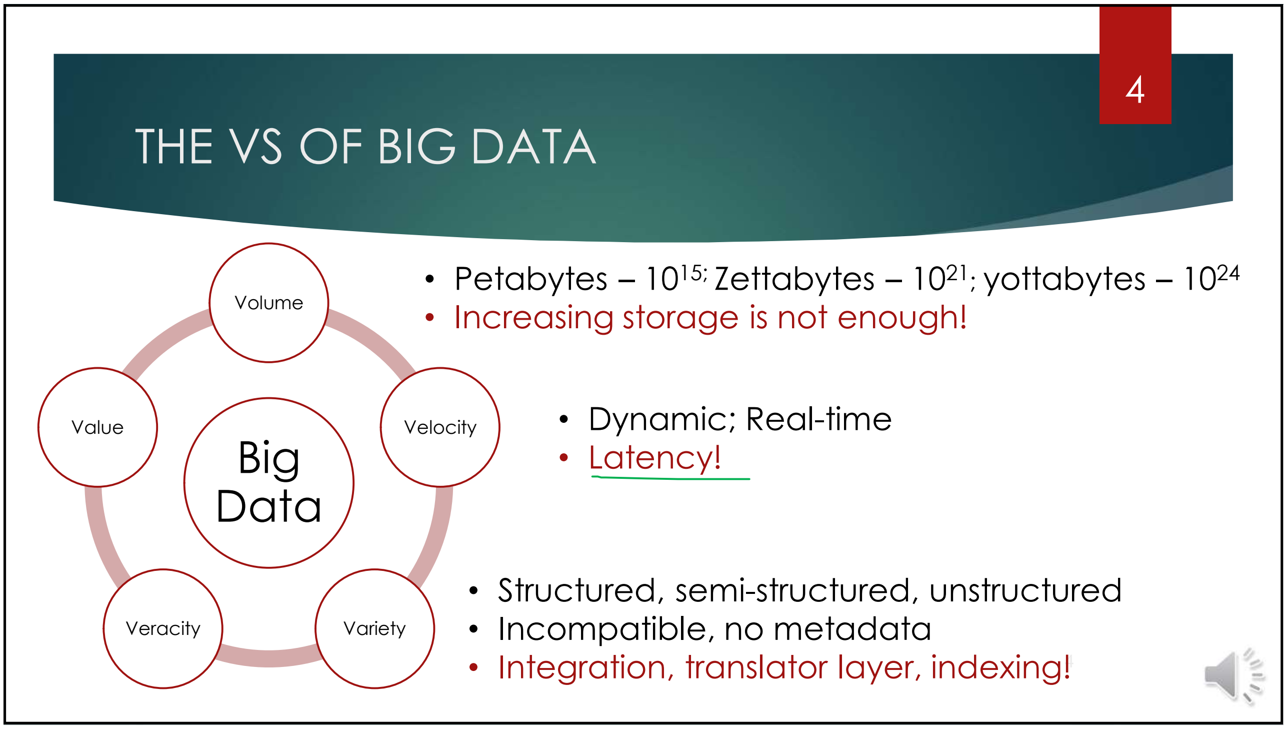


TERMS

3

DATA SCIENCE
BIG DATA
DATA MANAGEMENT
New theories and techniques for data analysis
Too big to process
New models for storage and manipulation

3



THE VS OF BIG DATA

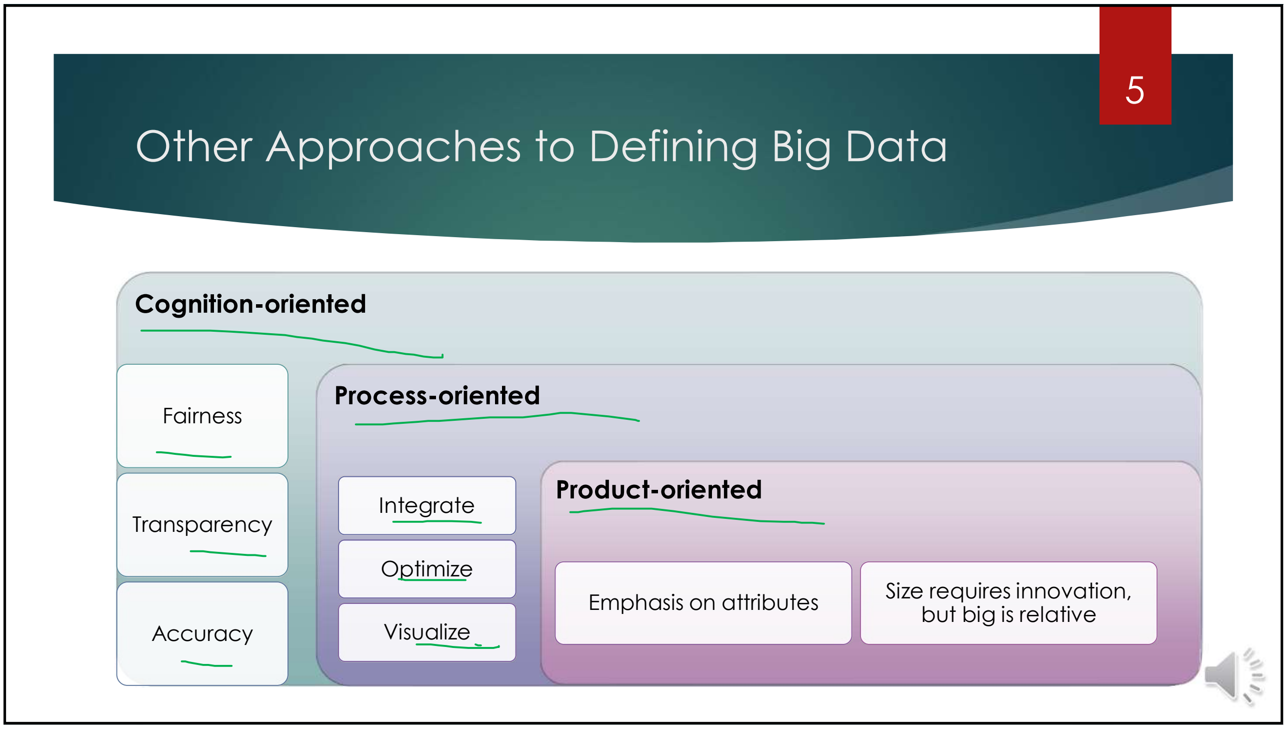
4

- Petabytes – $10^{15;}$ Zettabytes – 10^{21}; yottabytes – 10^{24}
- Increasing storage is not enough!

- Dynamic; Real-time
- Latency!

- Structured, semi-structured, unstructured
- Incompatible, no metadata
- Integration, translator layer, indexing!

4



Other Approaches to Defining Big Data

5

Cognition-oriented

- Fairness
- Transparency
- Accuracy

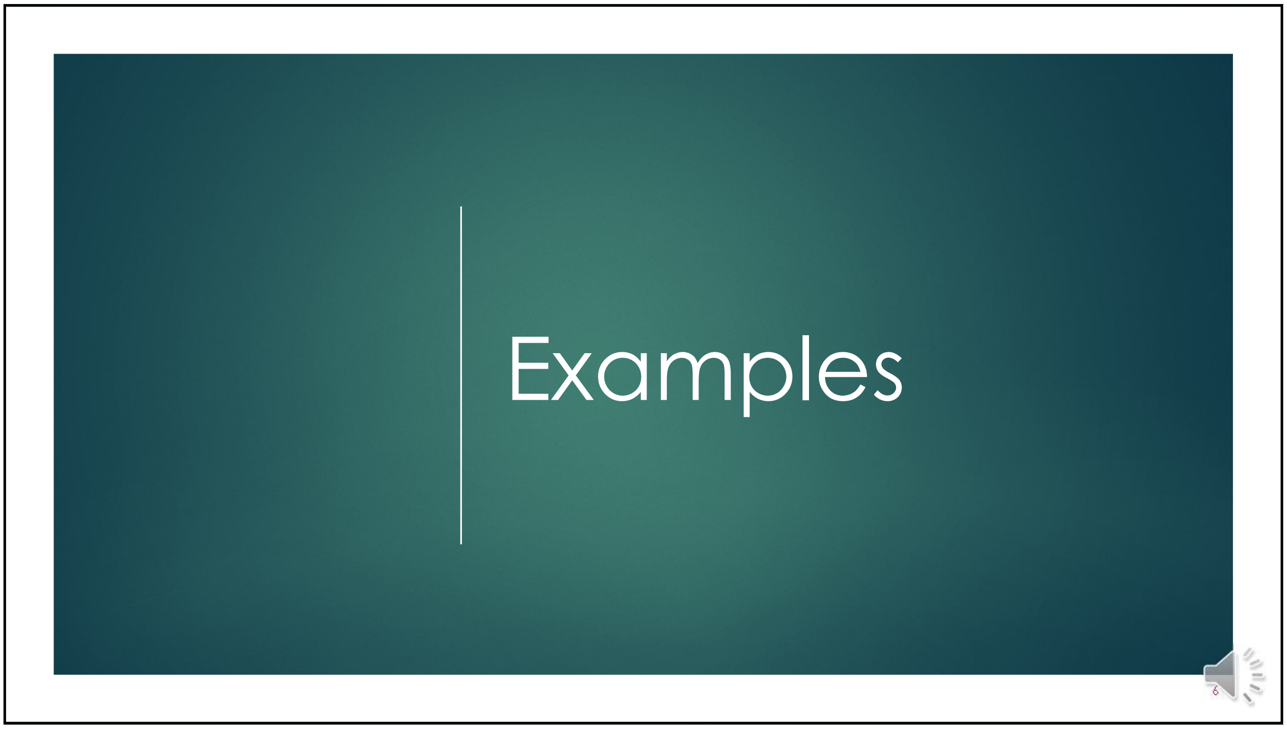
Process-oriented

- Integrate
- Optimize
- Visualize

Product-oriented

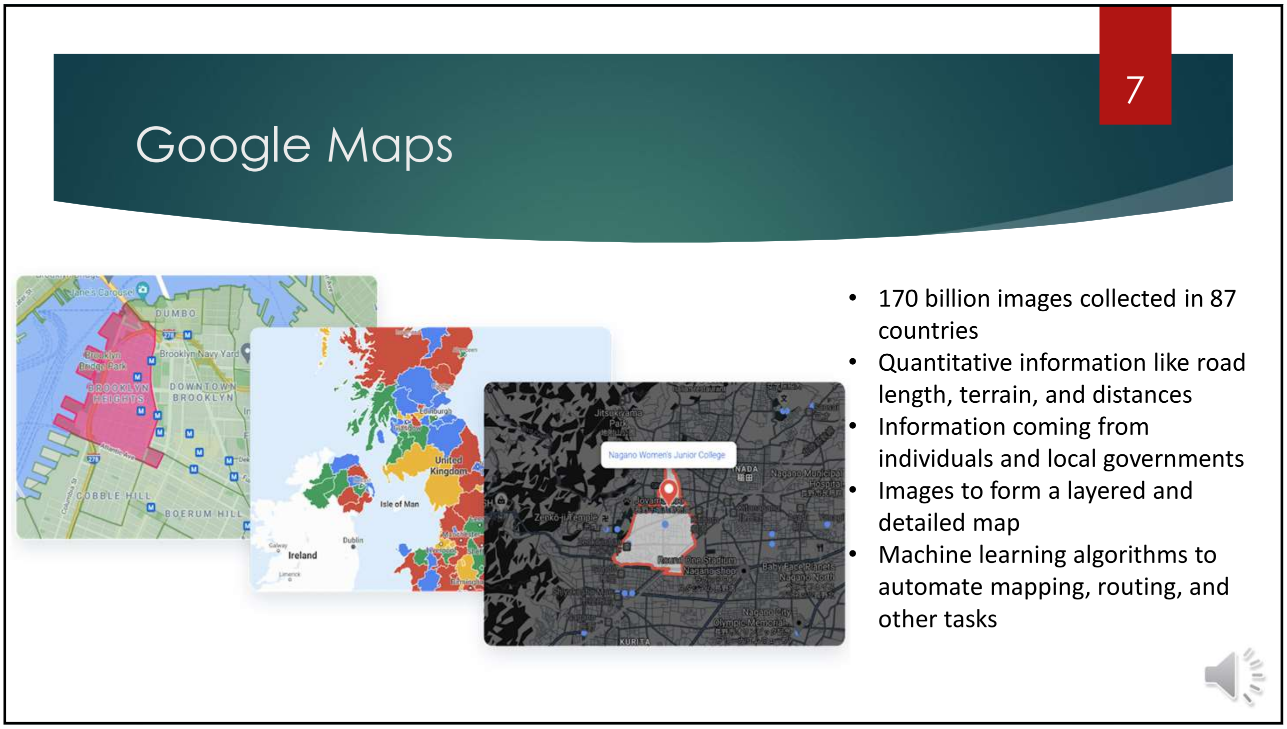
- Emphasis on attributes
- Size requires innovation, but big is relative

5



Examples

6

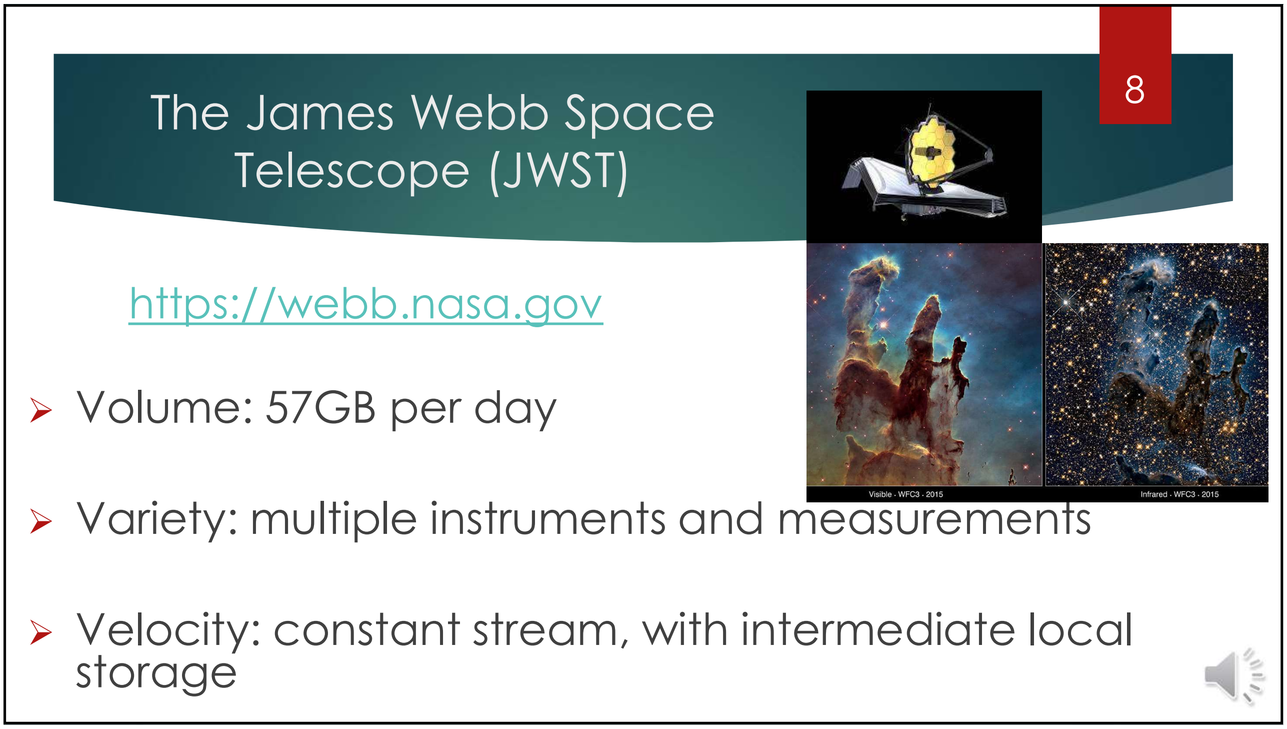


7

Google Maps

- 170 billion images collected in 87 countries
- Quantitative information like road length, terrain, and distances
- Information coming from individuals and local governments
- Images to form a layered and detailed map
- Machine learning algorithms to automate mapping, routing, and other tasks

7



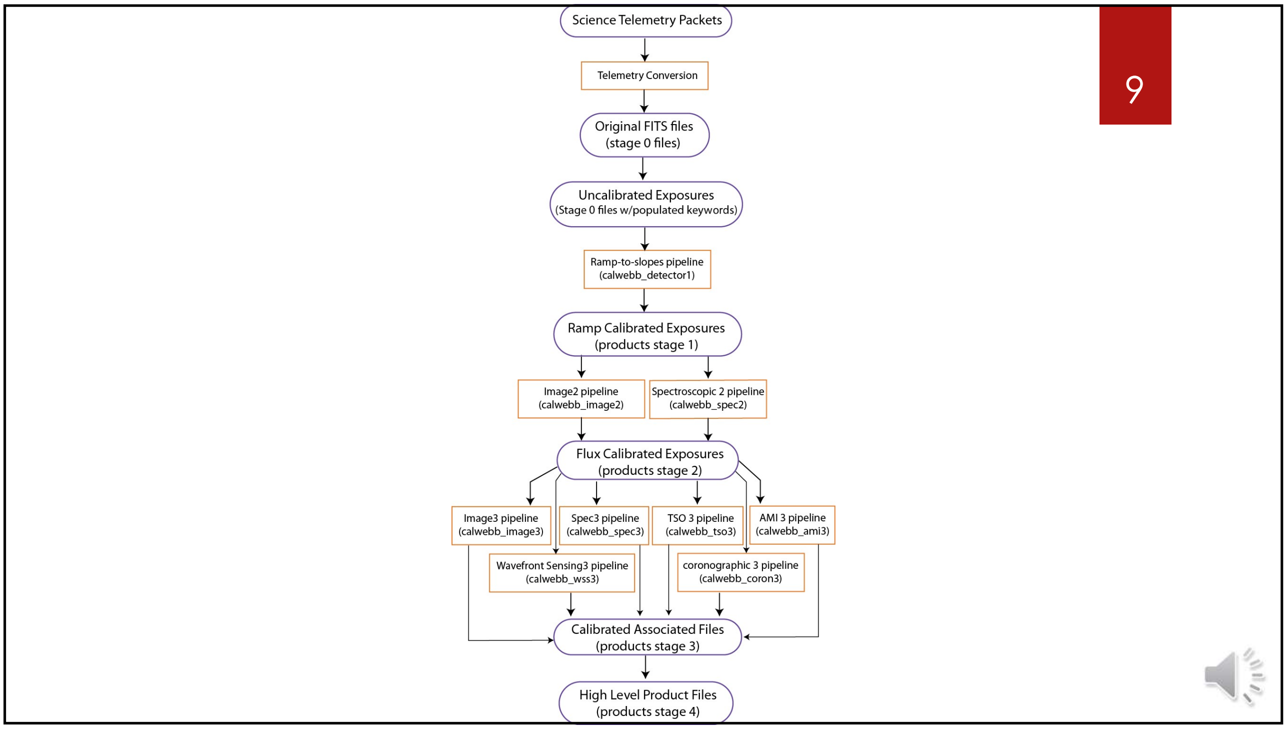
8

The James Webb Space Telescope (JWST)

https://webb.nasa.gov

- Volume: 57GB per day
- Variety: multiple instruments and measurements
- Velocity: constant stream, with intermediate local storage

8



9

Science Telemetry Packets

Telemetry Conversion

Original FITS files
(stage 0 files)

Uncalibrated Exposures
(Stage 0 files w/populated keywords)

Ramp-to-slopes pipeline
(calwebb_detector1)

Ramp Calibrated Exposures
(products stage 1)

Image2 pipeline
(calwebb_image2)

Spectroscopic 2 pipeline
(calwebb_spec2)

Flux Calibrated Exposures
(products stage 2)

Image3 pipeline
(calwebb_image3)

Spec3 pipeline
(calwebb_spec3)

TSO 3 pipeline
(calwebb_tso3)

AMI 3 pipeline
(calwebb_ami3)

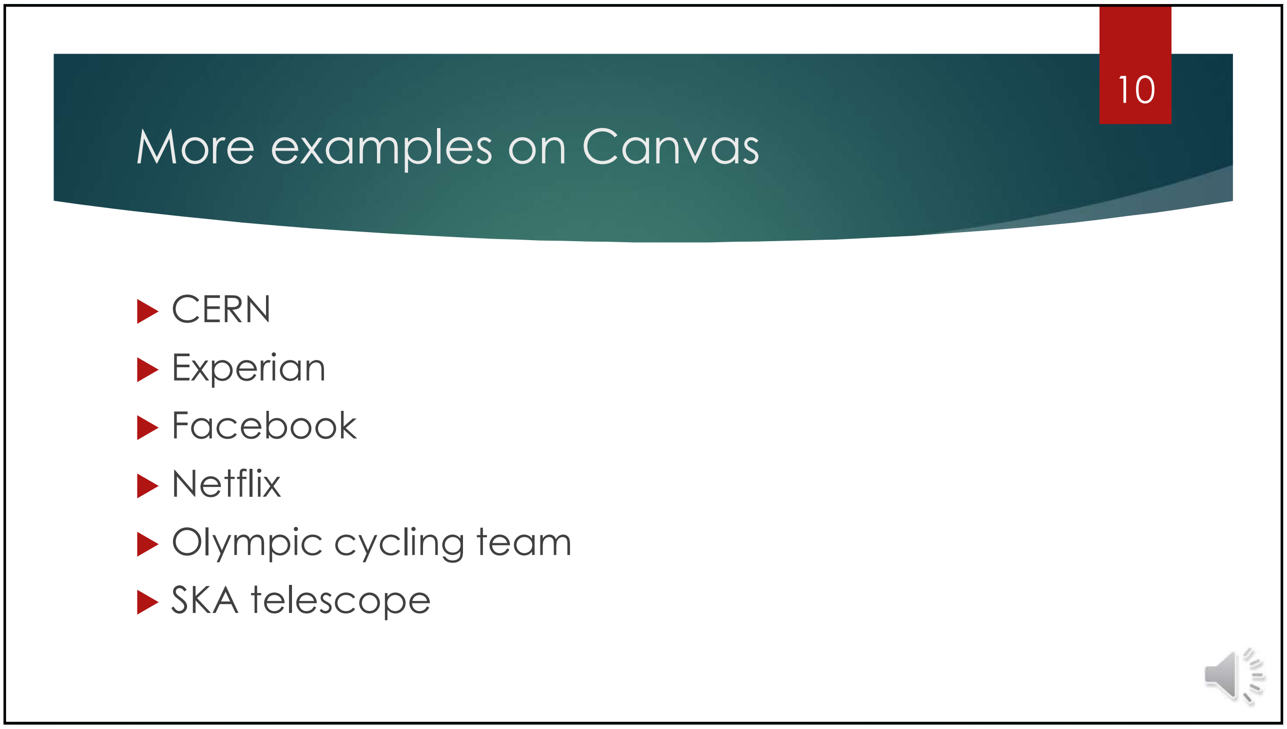
Wavefront Sensing3 pipeline
(calwebb_wss3)

coronographic 3 pipeline
(calwebb_coron3)

Calibrated Associated Files
(products stage 3)

High Level Product Files
(products stage 4)

9

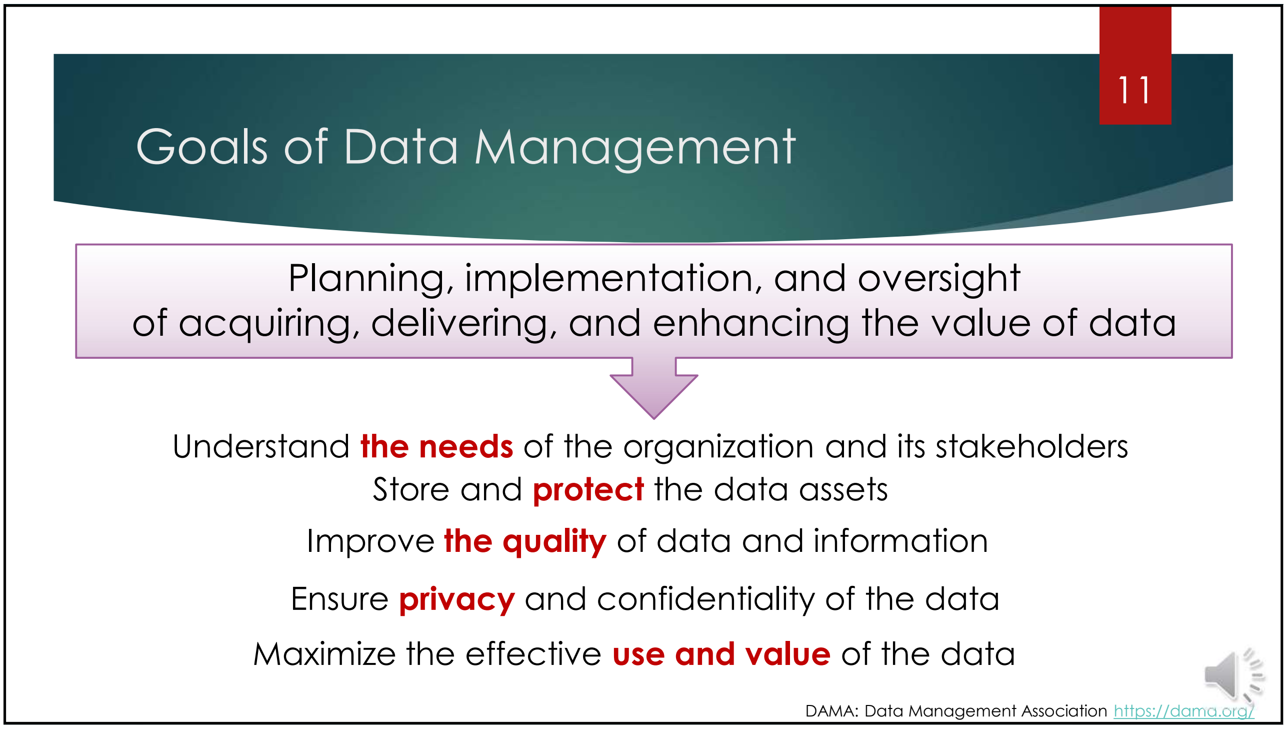


10

More examples on Canvas

- CERN
- Experian
- Facebook
- Netflix
- Olympic cycling team
- SKA telescope

10



11

Goals of Data Management

Planning, implementation, and oversight
of acquiring, delivering, and enhancing the value of data

Understand **the needs** of the organization and its stakeholders

Store and **protect** the data assets

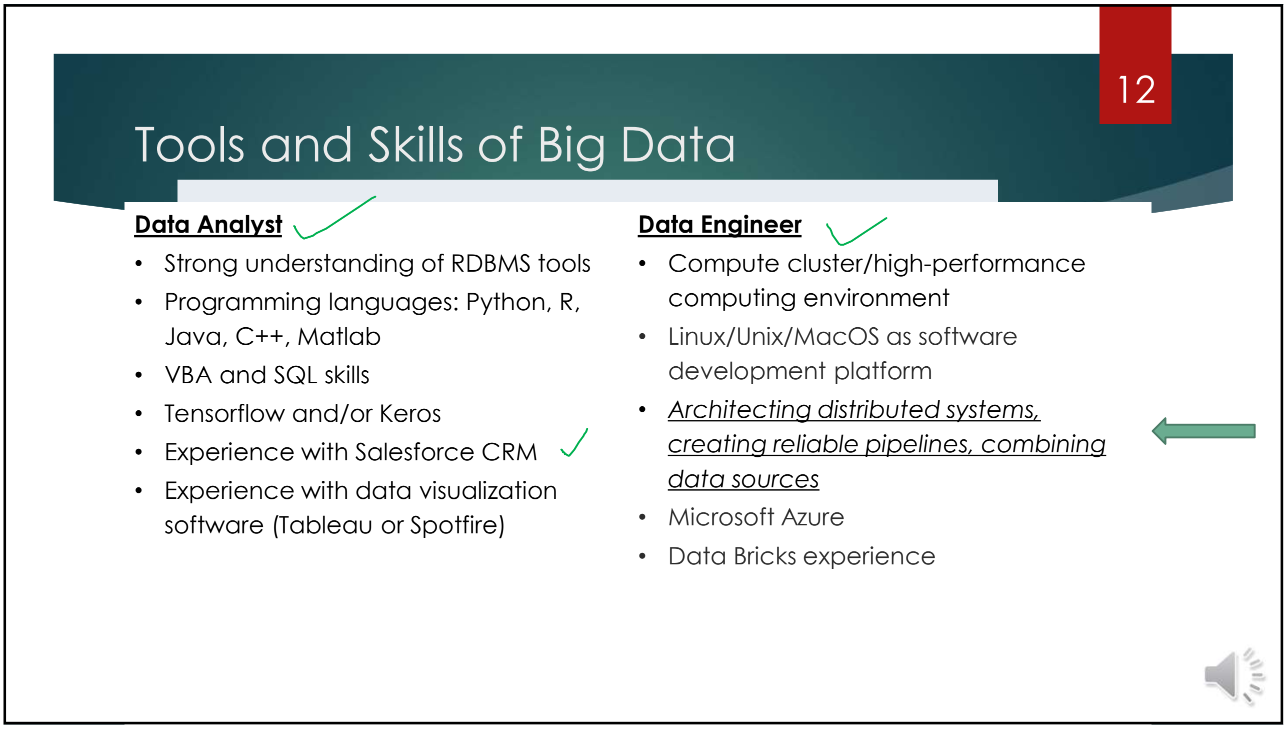
Improve **the quality** of data and information

Ensure **privacy** and confidentiality of the data

Maximize the effective **use and value** of the data

DAMA: Data Management Association https://dama.org/

11



12

Tools and Skills of Big Data

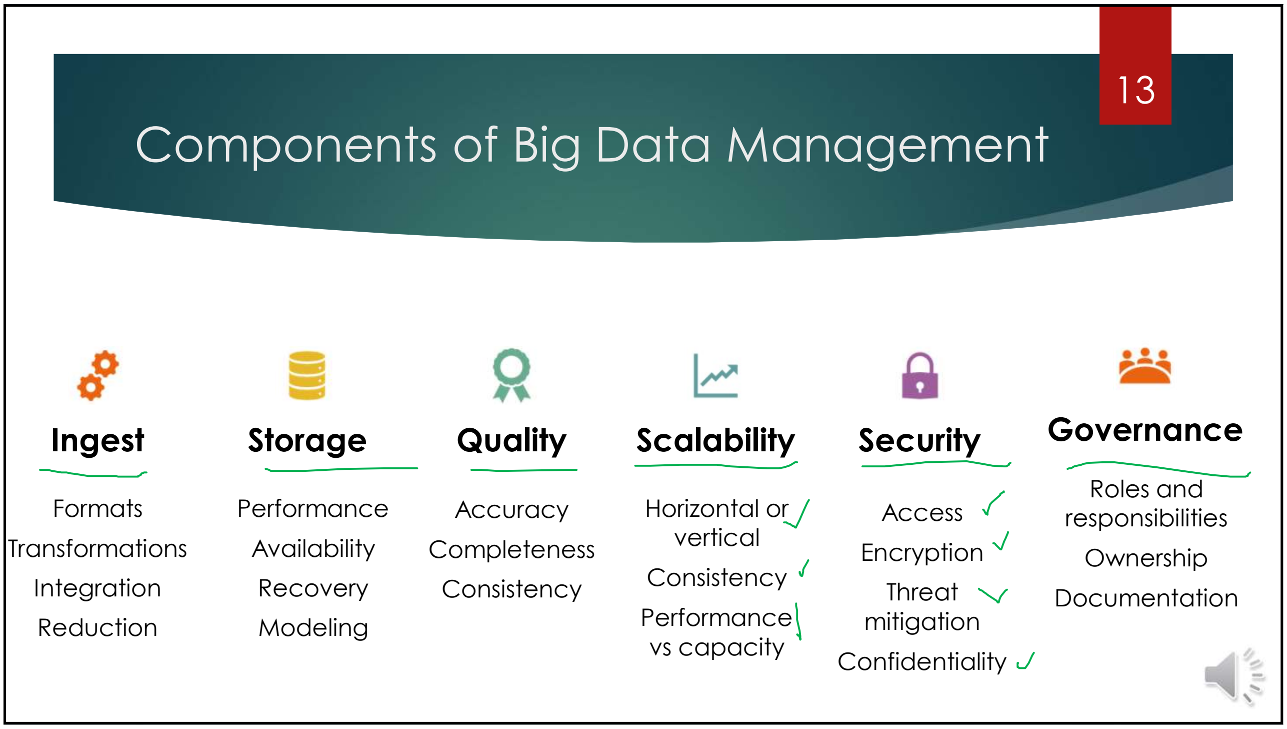
Data Analyst ✓

- Strong understanding of RDBMS tools
- Programming languages: Python, R, Java, C++, Matlab
- VBA and SQL skills
- Tensorflow and/or Keros
- Experience with Salesforce CRM ✓
- Experience with data visualization software (Tableau or Spotfire)

Data Engineer ✓

- Compute cluster/high-performance computing environment
- Linux/Unix/MacOS as software development platform
- *Architecting distributed systems, creating reliable pipelines, combining data sources*
- Microsoft Azure
- Data Bricks experience

12

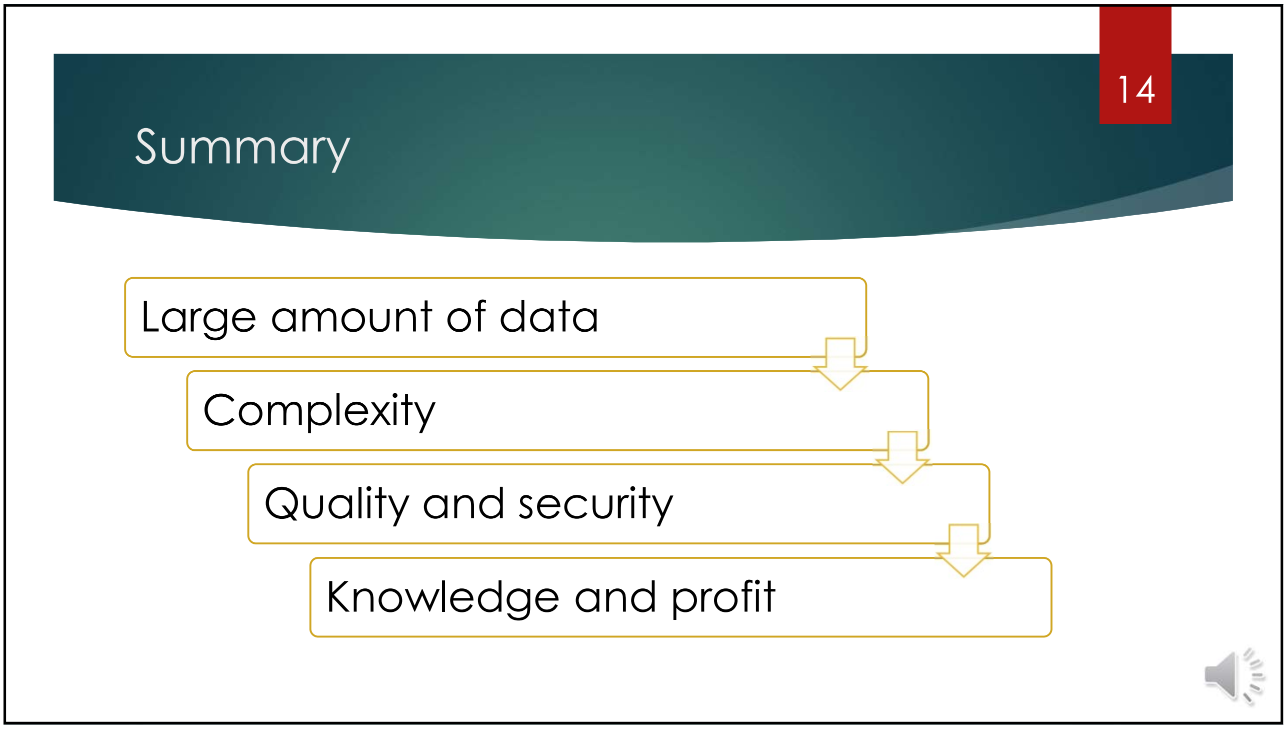


Components of Big Data Management

13

Ingest	Storage	Quality	Scalability	Security	Governance
Formats	Performance	Accuracy	Horizontal or vertical ✓	Access ✓	Roles and responsibilities
Transformations	Availability	Completeness	Consistency ✓	Encryption ✓	Ownership
Integration	Recovery	Consistency	Performance vs capacity ✓	Threat mitigation ✓	Documentation
Reduction	Modeling			Confidentiality ✓	

13

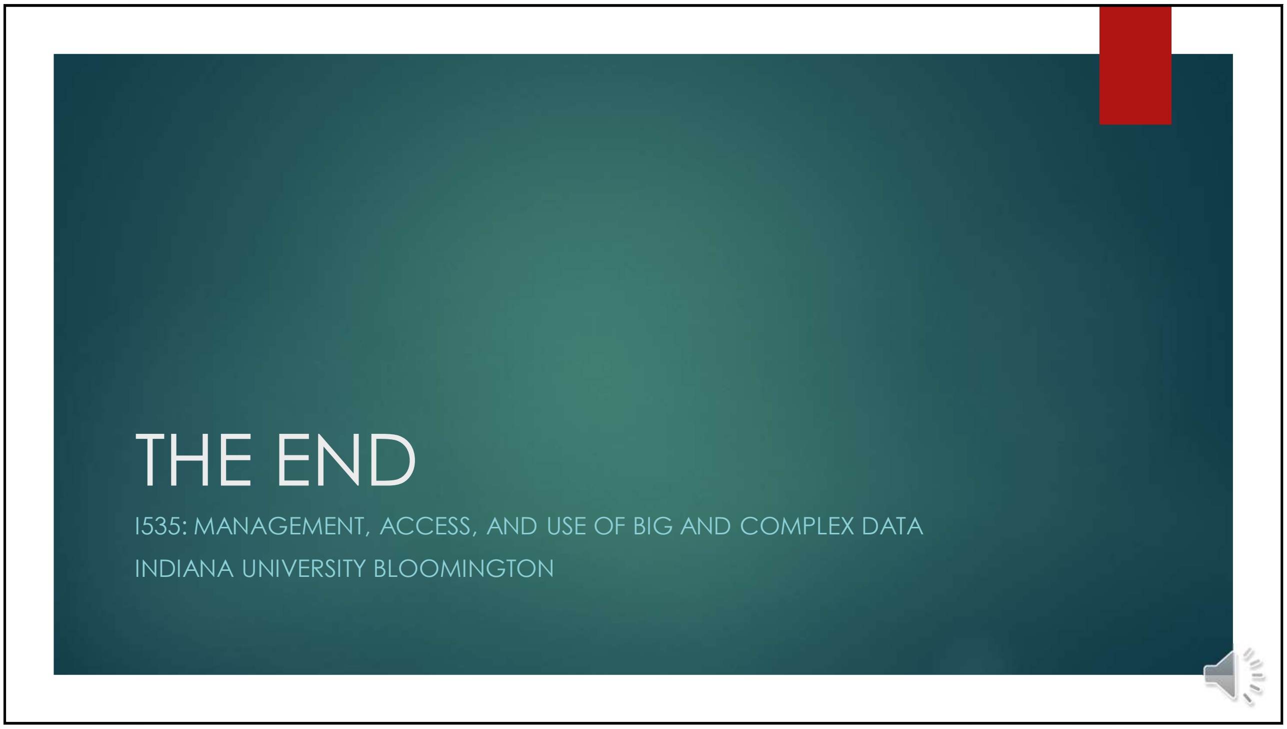


Summary

14

- Large amount of data
- Complexity
- Quality and security
- Knowledge and profit

14



THE END

I535: MANAGEMENT, ACCESS, AND USE OF BIG AND COMPLEX DATA

INDIANA UNIVERSITY BLOOMINGTON

15