Introduction to Data Management

Prof. Carl Kesselman

Logistics

Logistics

- Blackboard
  - Lectures slides
  - Readings
  - Assignments
  - Discussion forums
- Laptop
  - For access to databases and tools for homework/labs
- Amazon Web Services

Grading structure

- Homework assignments: 55%
  - About 5, will involve programming
- Quizzes: 45%
  - Weekly
  - Will drop lowest quiz
  - No excuses

Schedule

- We will not be having class the week of Feb 23rd
- We will not have a final, but you are expected to attend all classes through the last week
  - We will have a quiz on the last day ☺

Office Hours

- If you have a question, email will often get a quick response
- Office Hours will be in Michelson Convergent Bioscience (MBC) Room 303
  - You have all been given access to the building using your USC ID.
- My office hours are on Tuesday 2-4PM
  - Best to let me know if you are coming
- TA Office Hours to be Friday 2-4PM

Policy

- Late homework
  - 20% deduction for late homework
  - No credit after 24 hours
- Makeups for quizzes
  - Normally permitted only for medical emergencies
  - Doctor notes needed as proof
  - No makeups for job interviews, job fairs, etc.
- Unless instructed otherwise, work is expected to be your own…. We will check.

Policy

- Quizzes
  - Based on last week's materials
  - Closed book
- Please: *NO TEXTING*

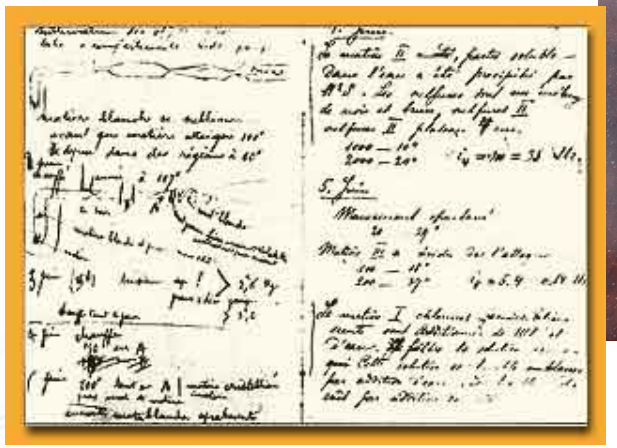
Course Outline

- We are going to start from the bottom and work our way up…
  - Introduction, what is data
  - Data storage
  - File systems, network file systems.
  - File formats, data encoding
  - Data Modeling
  - Relational databases and analytics
  - Data warehouses
  - NoSQL databases
  - MapReduce
  - Metadata
  - Identity, naming and persistence

Movie time….

What are data?

NASA Astronomy Picture of the Day



Marie Curie's notebook aip.org

hudsonalpha.org

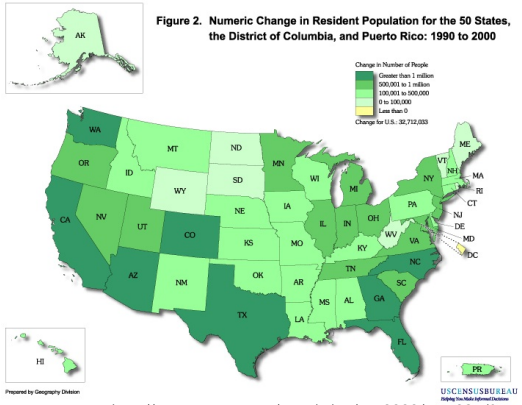
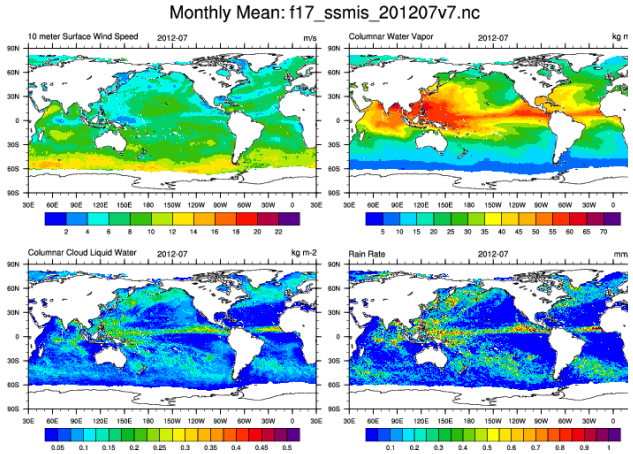


Figure 2. Numeric Change in Resident Population for the 50 States, the District of Columbia, and Puerto Rico: 1990 to 2000

http://www.census.gov/population/cen2000/map02.gif



ncl.ucar.edu

Date:1/2.07.75 Place:Sakaltutan
Zafor
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. eS came from Germany and did the plastering. He arranged the carpentry in Kayseri. Çok para gitti. {much money went} Has a tractor.

Date:July1980 Place:Sakaltutan
Zafor:
Household now Zafor and wife; Nazif Unal and wife and youngest son, still a boy. They run two dolmuß; one with a driver from Süleymanli. Goes in and out once a day. He gets 8,000 a month. Zafor then said, keskin de©il. { not sharp - i.e.? not profitable} I said he did very well on 8,000 TL with only two journeys a day. Nazif Unal has "bought" a Durak {dolmuß stop} from Beledive and works all day in Kayseri.

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php

Data

- **Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research, discovery or scholarship.**

What is Data Management? (DAMA Definition)

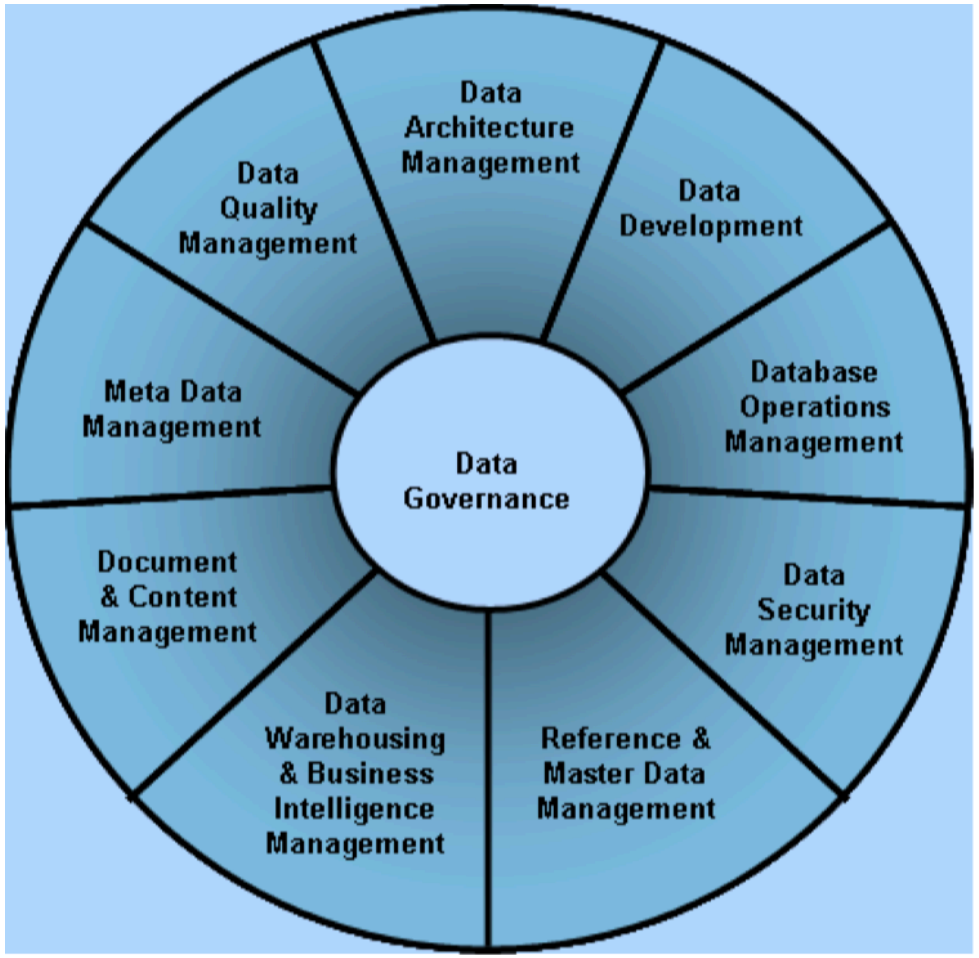
- Data Resource Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise.
- Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.

Data Management

- Enterprise perspective
  - Support business intelligence, decision making, business continuity
- Scientific perspective
  - Support knowledge extraction, discovery
- Convergence with big-data

Data Management Topics

(DAMA Data Management Body of Knowledge)



Data Management Functions

- Data Governance
  - planning, supervision and control over data management and use
- Data Architecture Management
  - as an integral part of the enterprise architecture
- Data Development
  - analysis, design, building, testing, deployment and maintenance
- Database Operations Management
  - support for structured physical data assets
- Data Security Management
  - ensuring privacy, confidentiality and appropriate access

Data Management Functions (cont).

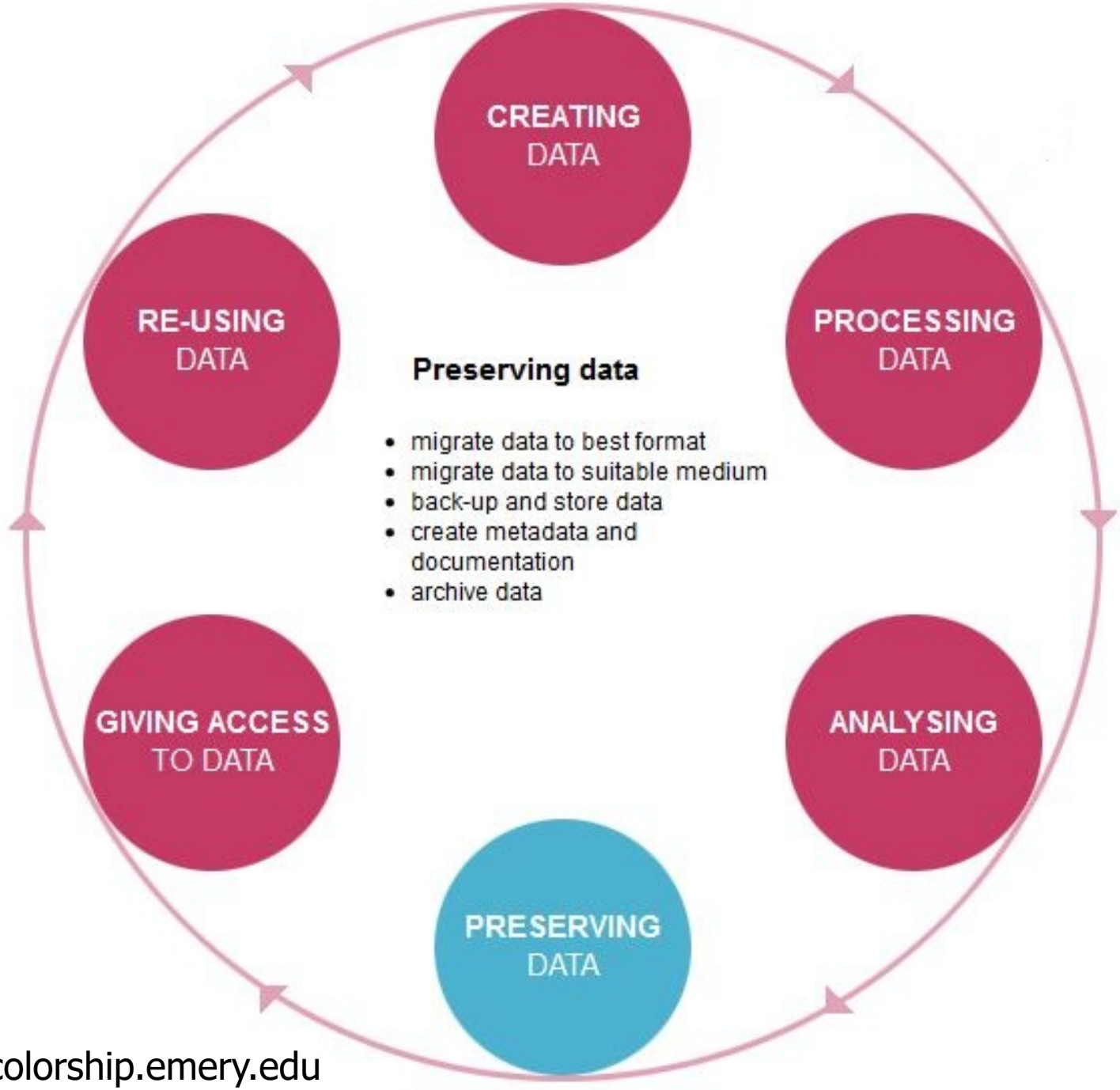
- Reference & Master Data Management
  - managing golden versions and replicas
- Data Warehousing & Business Intelligence Management
  - enabling access to decision support data for reporting and analysis
- Document & Content Management
  - storing, protecting, indexing and enabling access to data found in unstructured sources (electronic files and physical records)
- Meta Data Management
  - integrating, controlling and delivering meta data
- Data Quality Management
  - defining, monitoring and improving data quality

Discovery Paradigms (Gray)

- Thousand years ago
  - Empirical
- Few hundred years ago
  - Theoretical
- Fifty years ago
  - Computational
- Now: data exploration
  - Capture data, process, extract knowledge, analyze

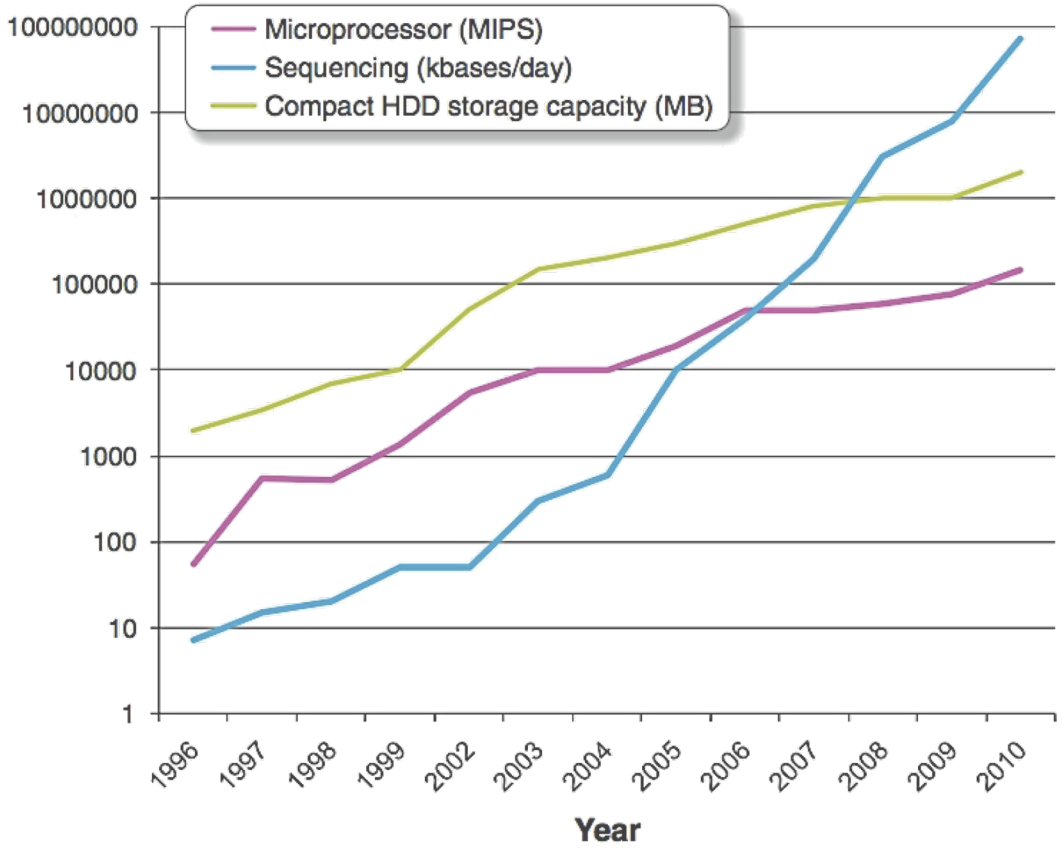
Data Challenges

- Ingest
- Manage large volumes
- Organize/reorganize
- Share it
- Query and visualize
- Integrate with other data and literature
- Document experiments
- Curation and long-term preservation



Digitalscolorship.emery.edu

Exploding data volumes

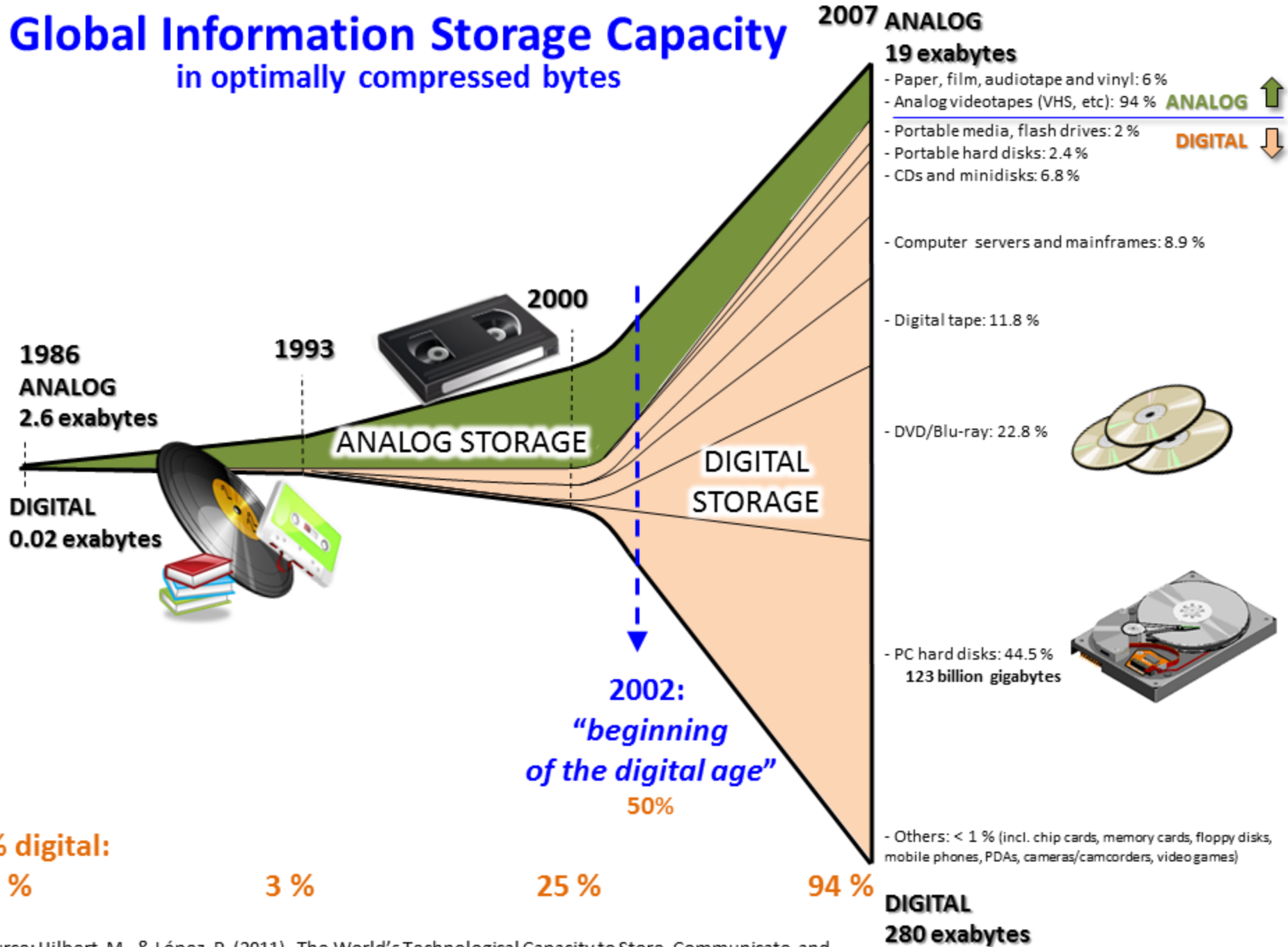


Data volumes are growing **much faster** than Moore's law

...(10,000x more over last 6 years for genome data)

Global Information Storage Capacity

in optimally compressed bytes



2007 **ANALOG**
19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % **ANALOG**

DIGITAL

- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %

- PC hard disks: 44.5 %
  123 billion gigabytes

- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

DIGITAL
280 exabytes

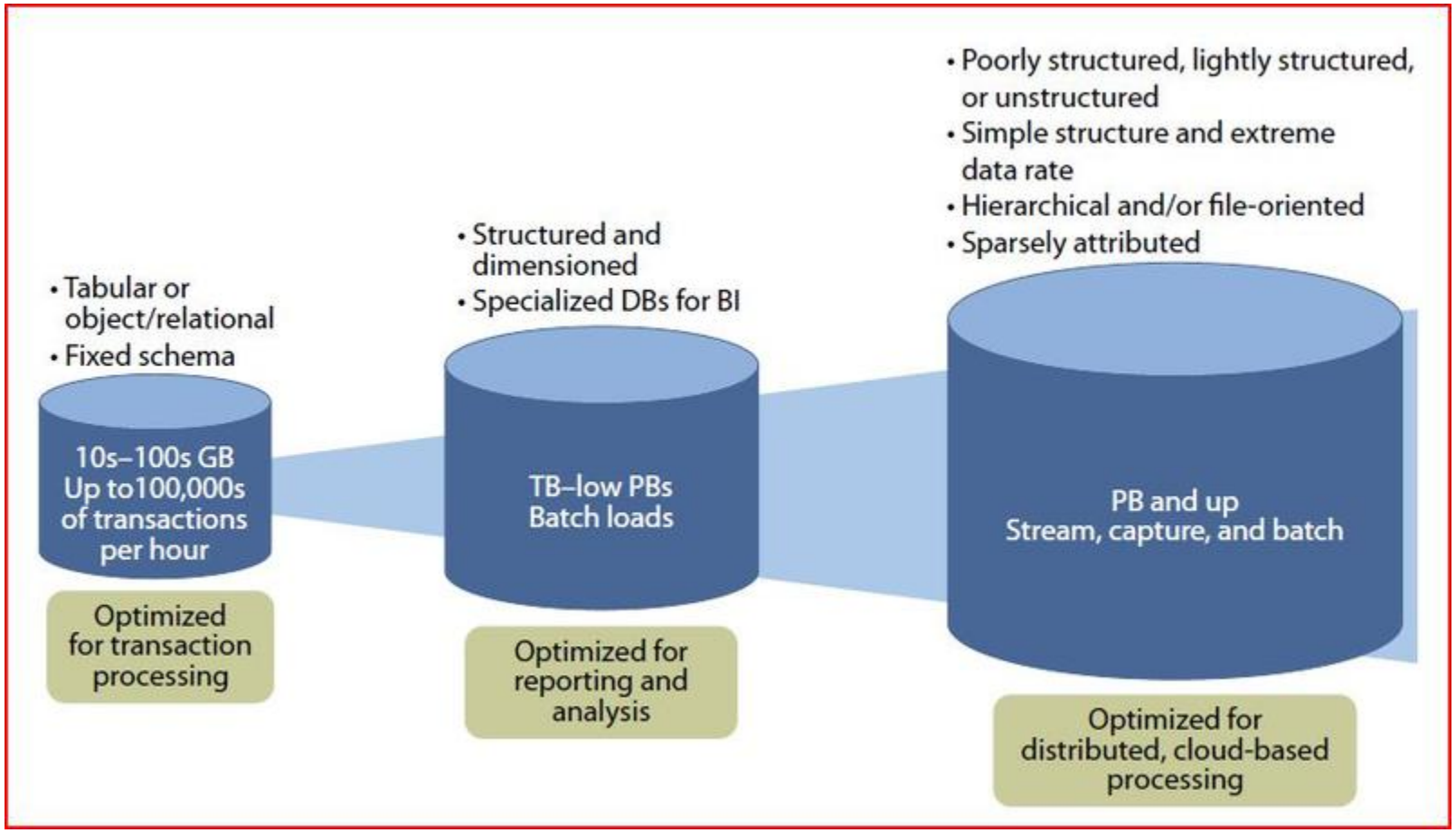
Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Big Data
(Gartner Group Definition)

"Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value”.

- Volume: big data doesn't sample. It just observes and tracks what happens
- Velocity: big data is often available in real-time
- Variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion

Data Management For Big Data



(Andreas Weigend)

Analysis Process

Sales
Text
Measurements
Usage
Images

Catalog
Classify
Visualize
Quantify
Summarize
Inference
Missing data
Forecast
Causal Inference
Annotations
Associations
Likelihoods

Learn about the whole from a part

Tell a story

Make a prediction

Explain

What is the problem….

"The first thing we've learned is the importance of having outstanding data to actually base your ML on. In our own shop, we've been working on a few big projects, and of having outstanding data to actually working on a few big projects, and we've had to spend most of the time just cleaning the data sets before you can even run the algorithm. That's taken us years just to clean the datasets. I think people underestimate how little clean data there is out there, and how hard it is to clean and link the data."

- Vasant Narasimhan, CEO Novartis

- "There is no point wasting good thoughts on bad data.”
  – Francis Crick

FAIR Data Principles

- Data should be Findable, Accessible, Interoperable, Reusable
- The principles refer to three types of entities:
  - data (or any digital object),
  - metadata (information about that digital object)
  - infrastructure.

What is “metadata”

- Its data about the data
- E.G
  - How big is it
  - When was it created
  - Who created it
  - What is the type of content

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers.

- **F1. (Meta)data are assigned a globally unique and persistent identifier**
- **F2. Data are described with rich metadata (defined by R1 below)**
- **F3. Metadata clearly and explicitly include the identifier of the data they describe**
- **F4. (Meta)data are registered or indexed in a searchable resource**

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorization.

- **A1. (Meta)data are retrievable by their identifier using a standardised communications protocol**
- **A1.1 The protocol is open, free, and universally implementable**
- **A1.2 The protocol allows for an authentication and authorisation procedure, where necessary**
- **A2. Metadata are accessible, even when the data are no longer available**

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- **I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.**
- **I2. (Meta)data use vocabularies that follow FAIR principles**
- **I3. (Meta)data include qualified references to other (meta)data**

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- **R1. Meta(data) are richly described with a plurality of accurate and relevant attributes**
- **R1.1. (Meta)data are released with a clear and accessible data usage license**
- **R1.2. (Meta)data are associated with detailed provenance**
- **R1.3. (Meta)data meet domain-relevant community standards**

Coming up….

- A brief introduction to computing systems…