
AIX0008: Introduction to Data Science

Patrice KOEHL, UC Davis

*Department of Computer Science
Genome Center
University of California, Davis*



Introduction to Data Science

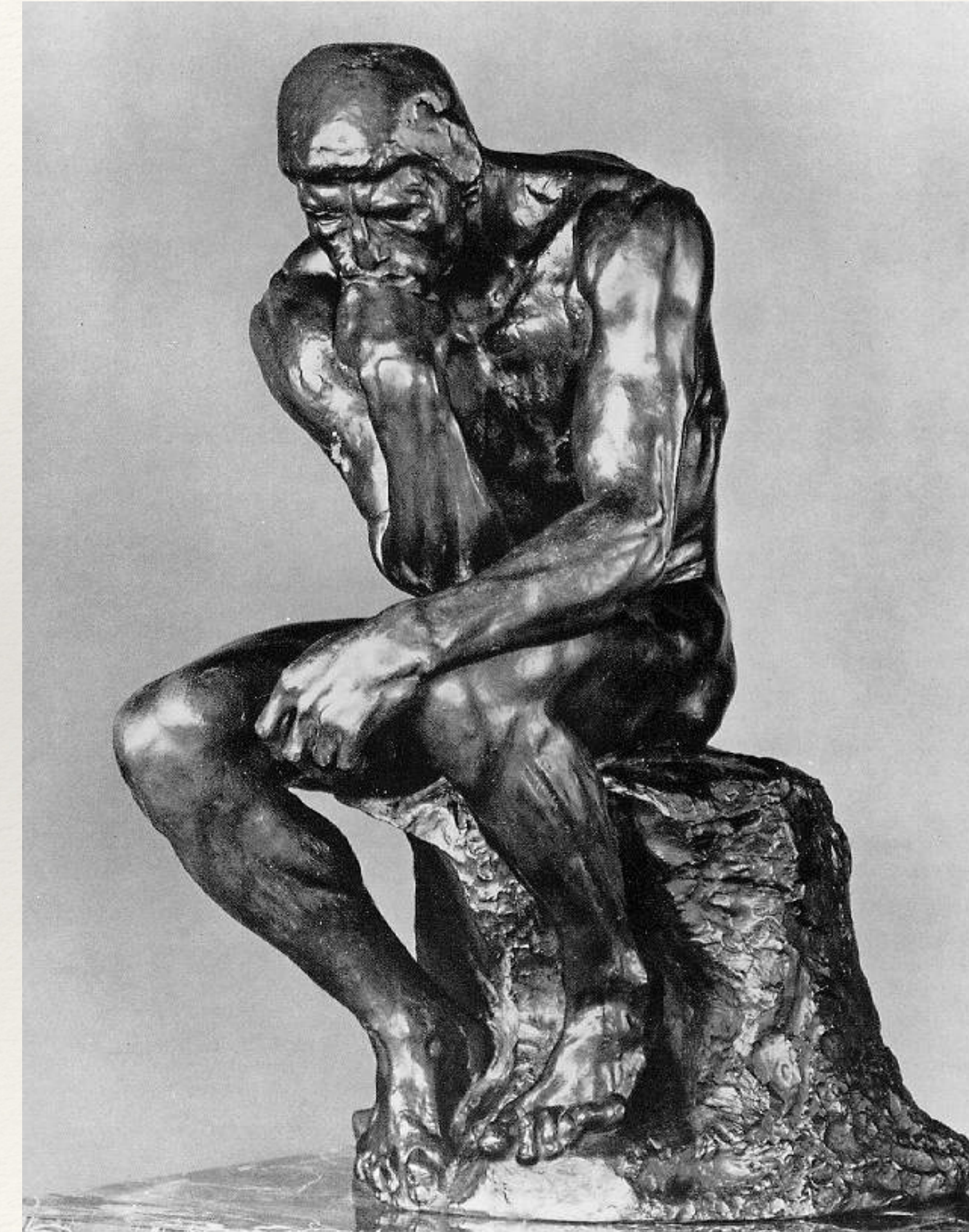
1. A paradigm shift in Science
2. What is “Big Data”?
3. Learning from Data / Data Science -Artificial Intelligence

Introduction to Data Science

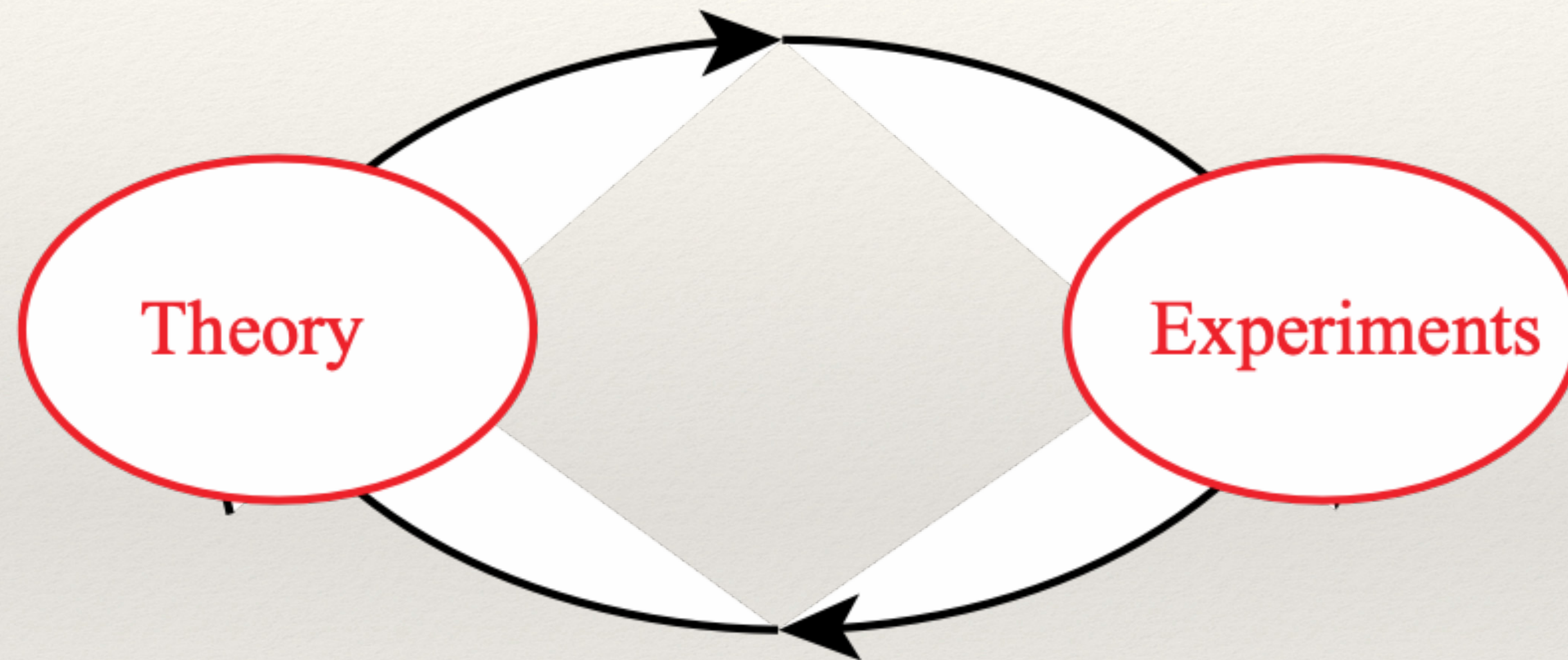
1. A paradigm shift in Science
2. What is “Big Data”?
3. Learning from Data / Data Science Artificial Intelligence

Science, then, and now

- For a long time, people thought that it would be enough to reason about the existing knowledge to explore everything there is to know.
- One single person could possess all knowledge in her cultural context.
(encyclopedia of Diderot and D'Alembert)
- Reasoning, and mostly passive observation were the main techniques in scientific research



Science, then, and now

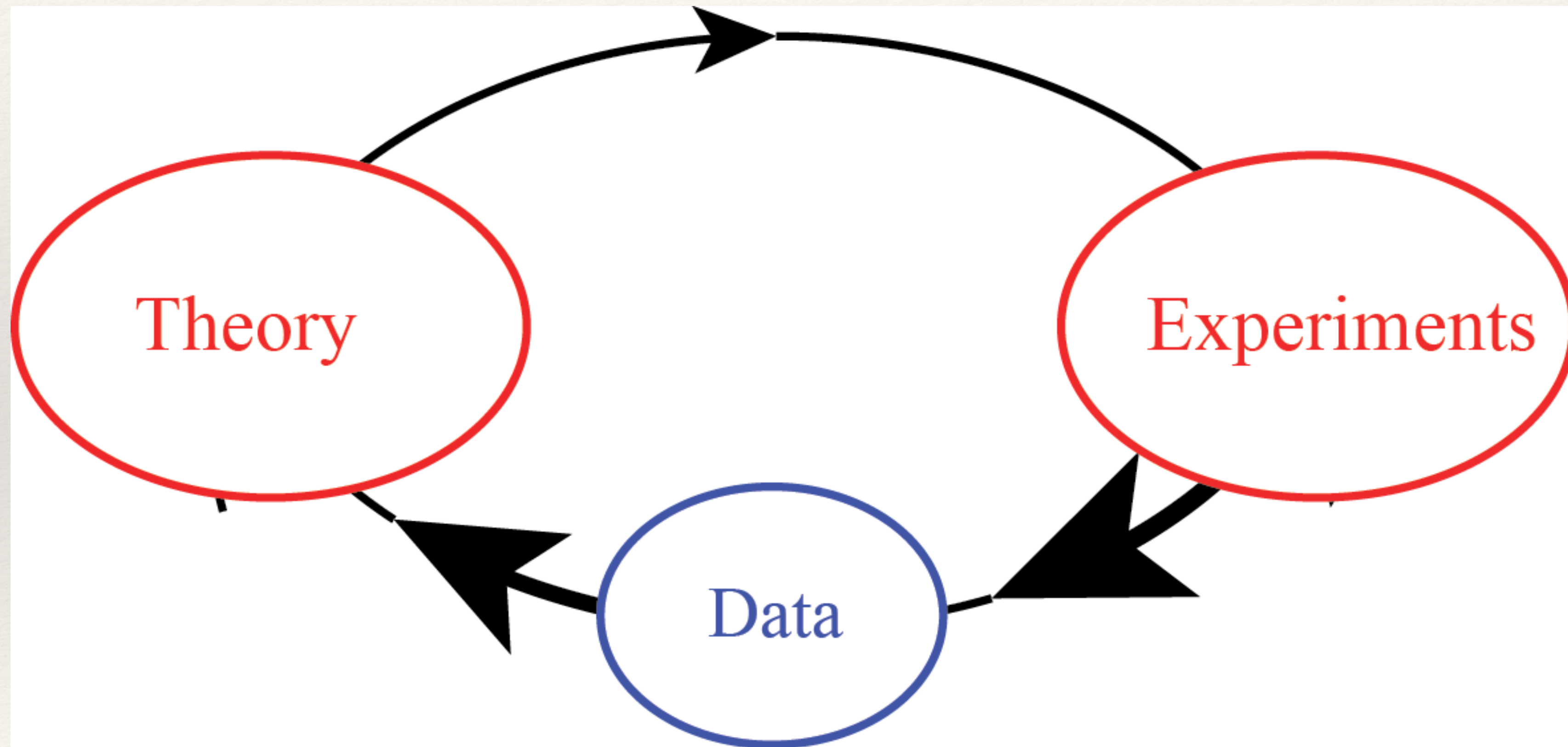


Science, then, and now

“All science is either physics, or stamp collecting”

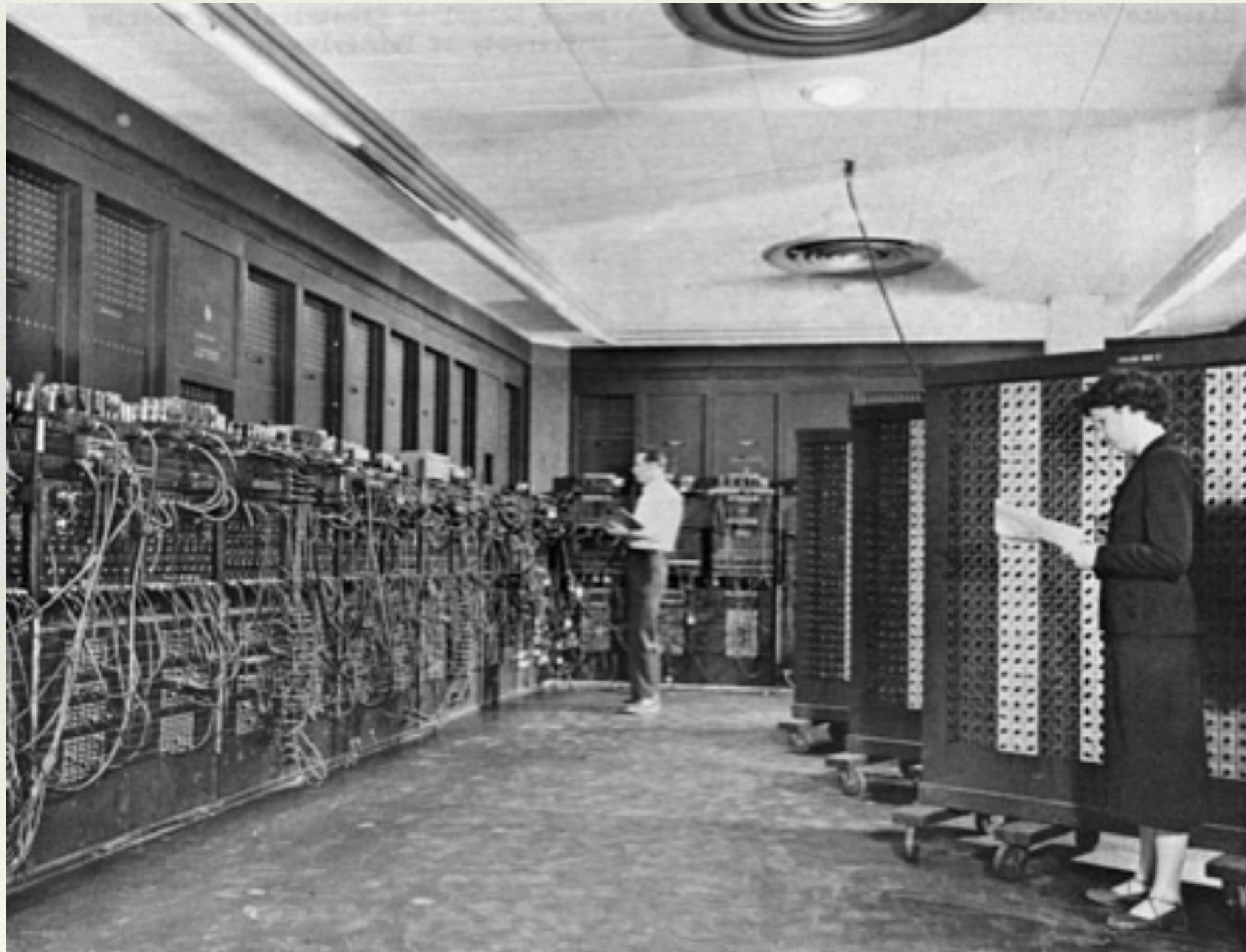
Rutherford, chemist and physicist, 1876-1937

Science, then, and now



Science, then, and now

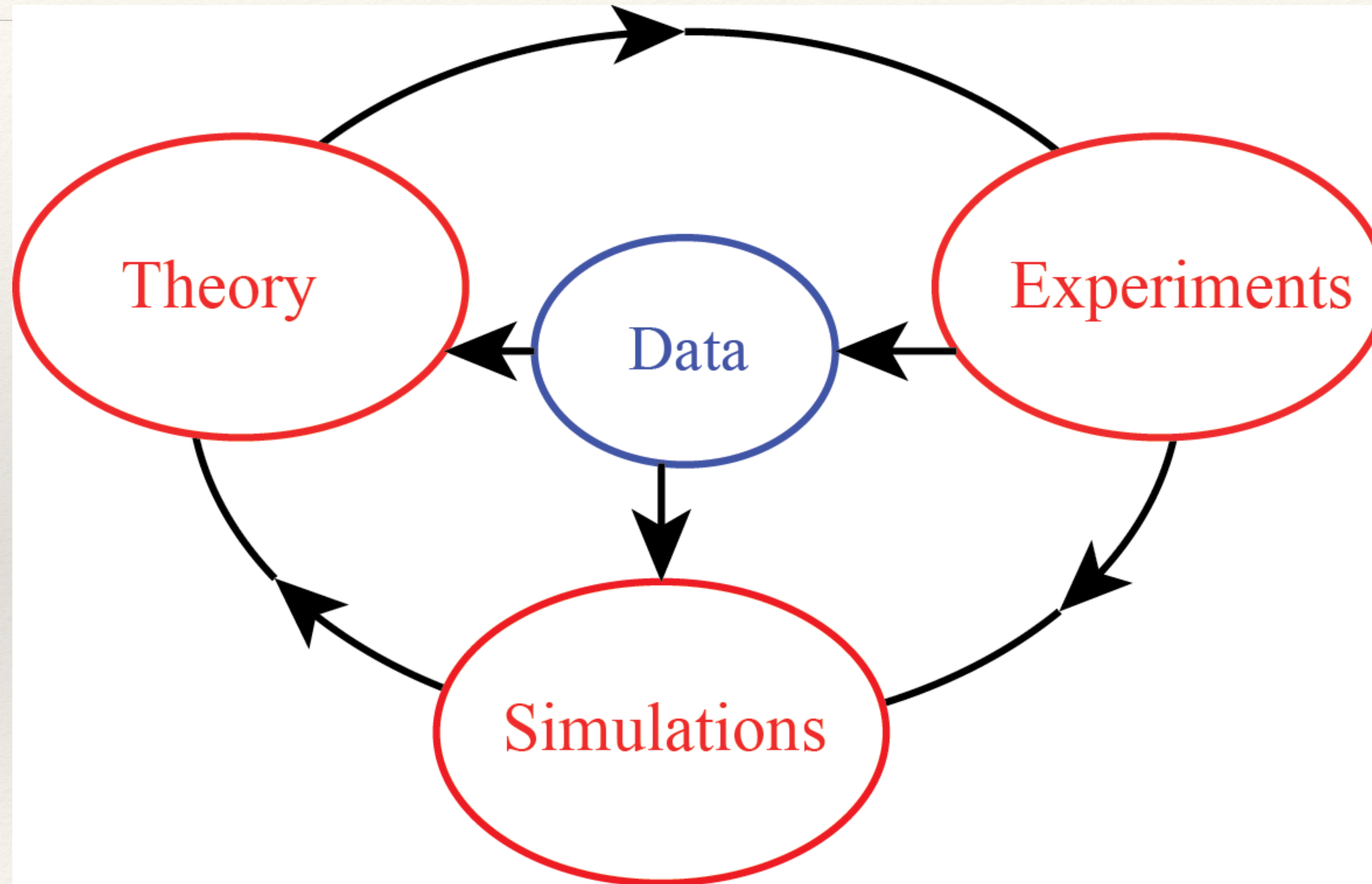
About 80 years ago: computers....



Science, then, and now

- Computer simulations developed hand-in-hand with the rapid growth of computers.
- A computer simulation is a computer program that attempts to simulate an abstract model of a particular system
- Computer simulations complement theory and experiments, and often integrate them
- They are becoming widespread in: Computational Physics, Chemistry, Mechanics, Materials, ..., Biology

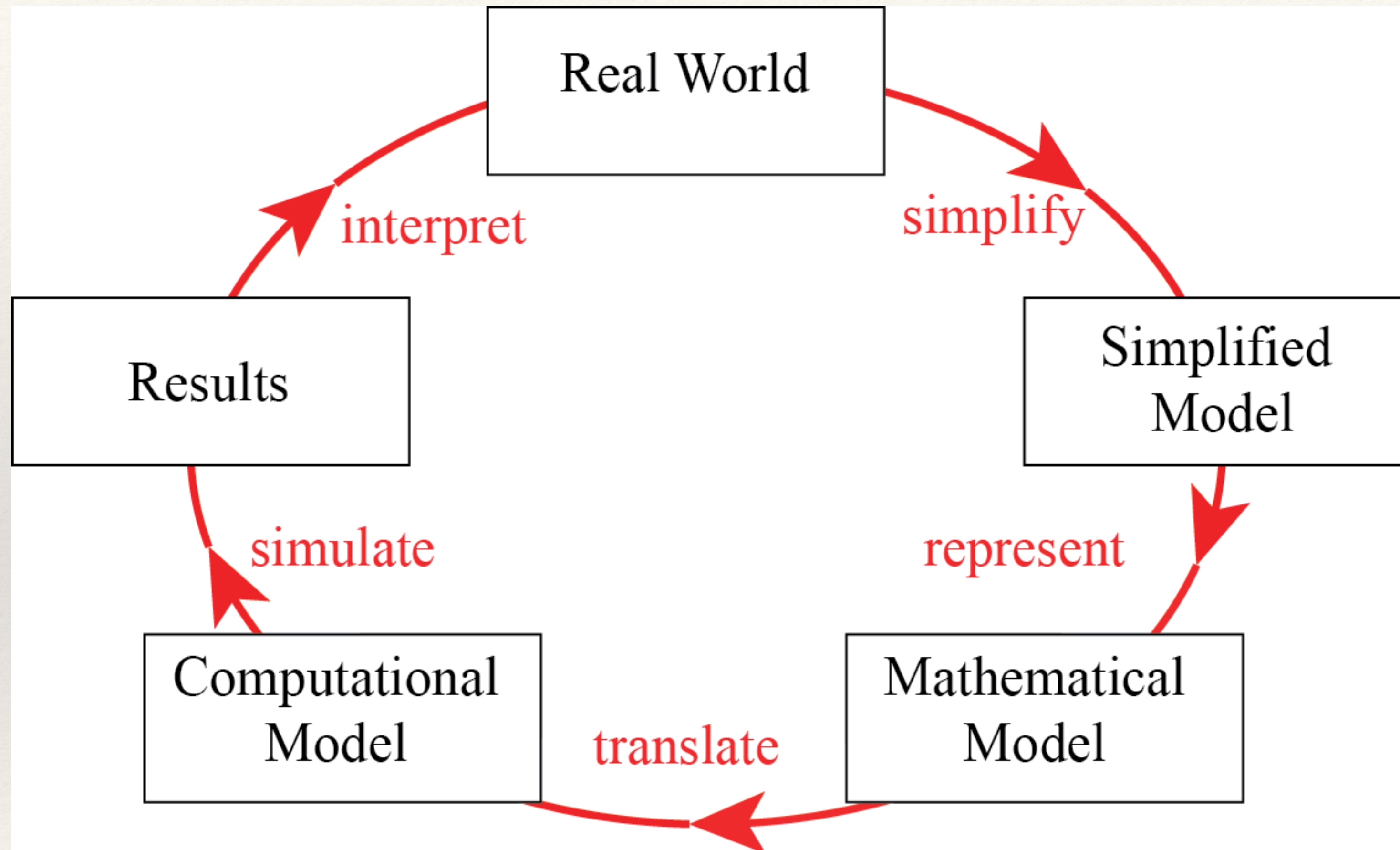
Science, then, and now



Mathematical Modeling

- Is often used in place of experiments when they are *too large, too expensive, too dangerous, or too time consuming*.
- Can be useful in “what if” studies; e.g. to investigate the use of *pathogens* (viruses, bacteria) to control an insect population.
- Is a modern tool for *scientific investigation*.

Mathematical Modeling



Science, then, and now

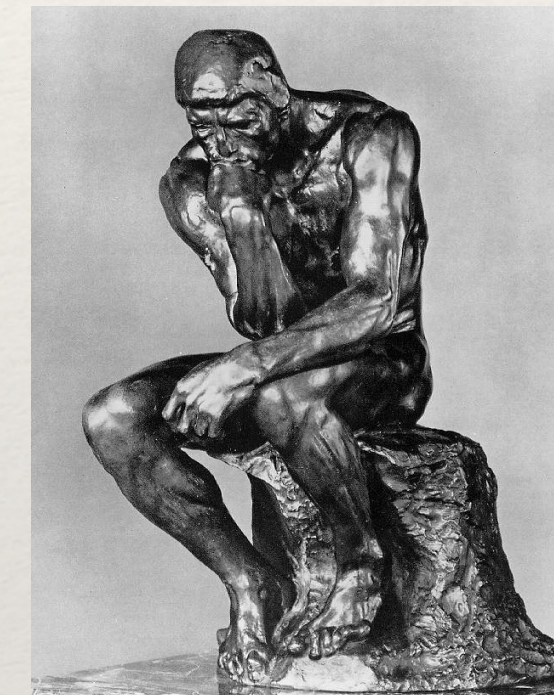
1. Thousand years ago – **Experimental Sciences**

Description of natural phenomena



2. Last few hundred years – **Theoretical Sciences**

Newton's law, Maxwell's equations...



3. Last few decades – **Computational Sciences**

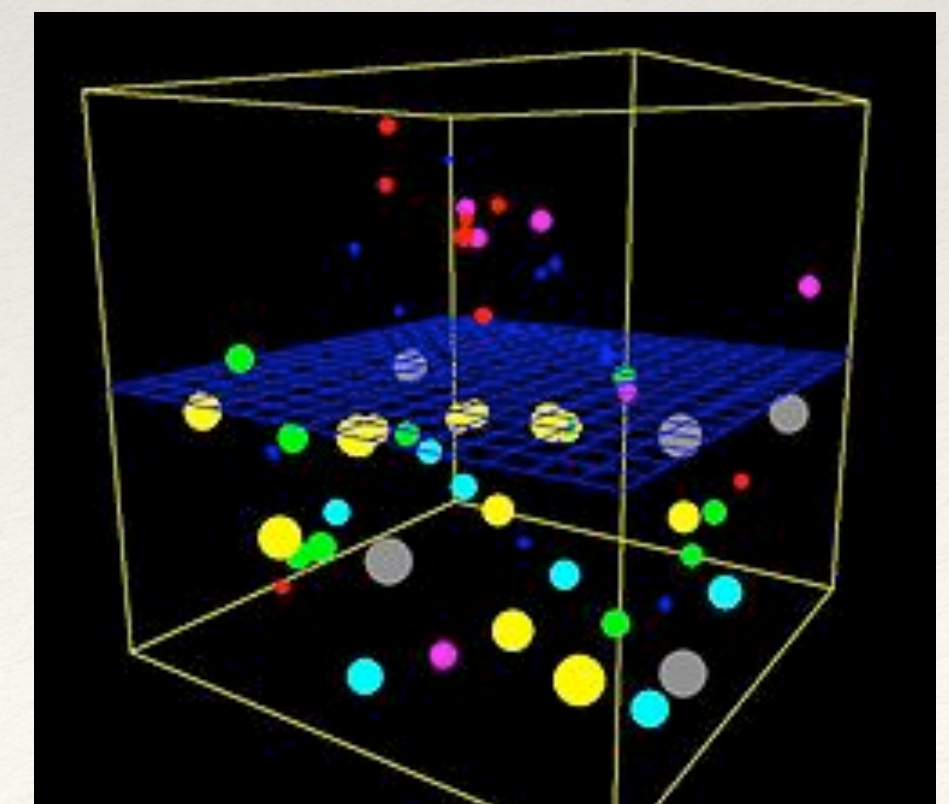
Simulation of complex phenomena

4. Today – **Data-Intensive Sciences**

Scientist overwhelmed with data sets from many different sources

Data captured by instruments

Data generated by simulations



Introduction to Data Science

1. A paradigm shift in Science
2. What is “Big Data”?
3. Learning from Data / Data Science Artificial Intelligence

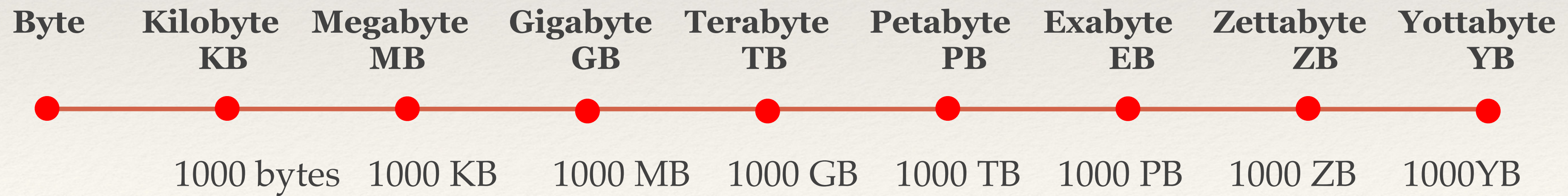
The three I's of Big Data

Big Data:

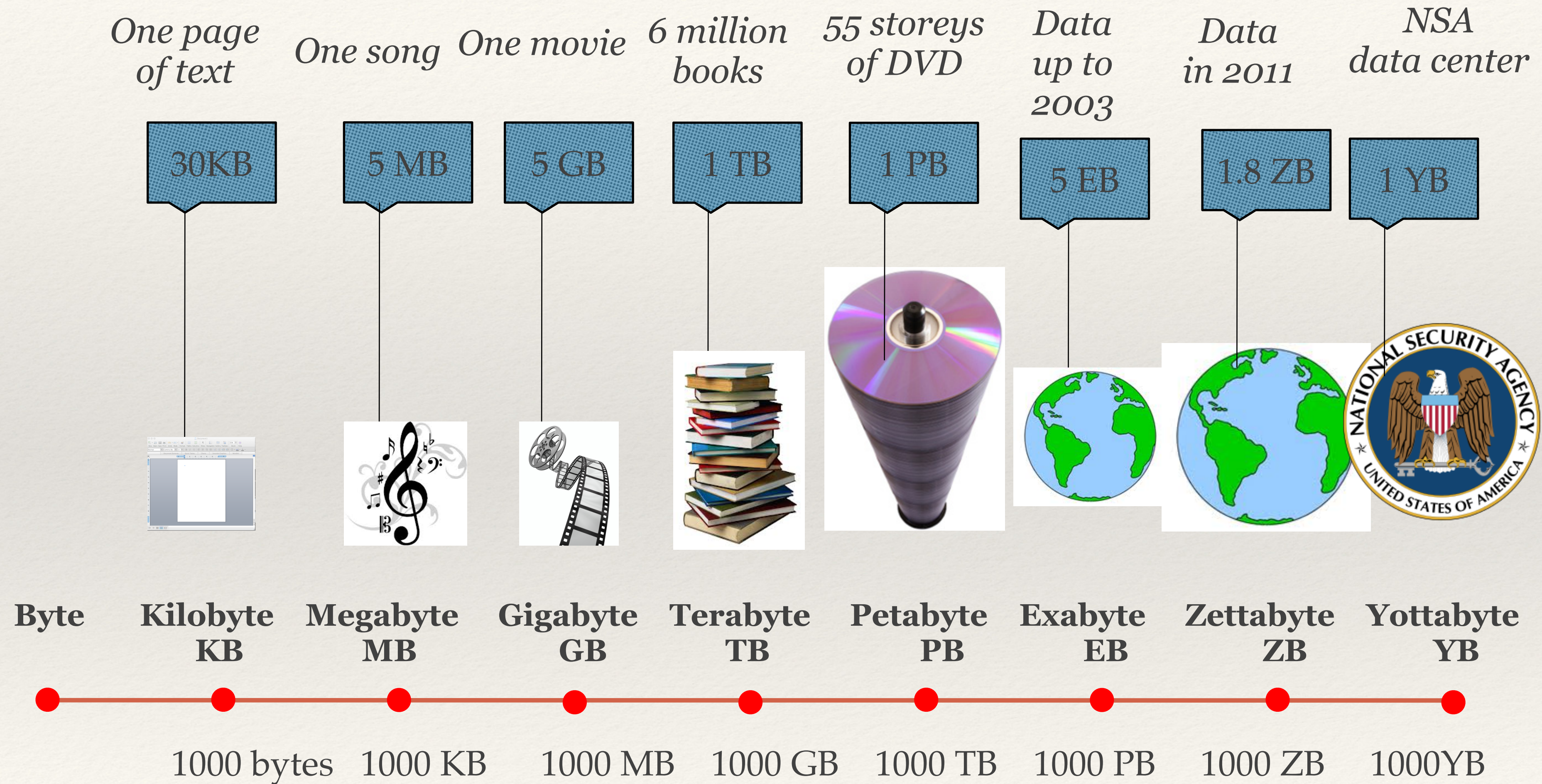
- **Immediate** (we need to do something about it now)
- **Intimidating** (what if we don't)
- **Ill-defined** (what is it?)

(loosely adapted from Forbes)

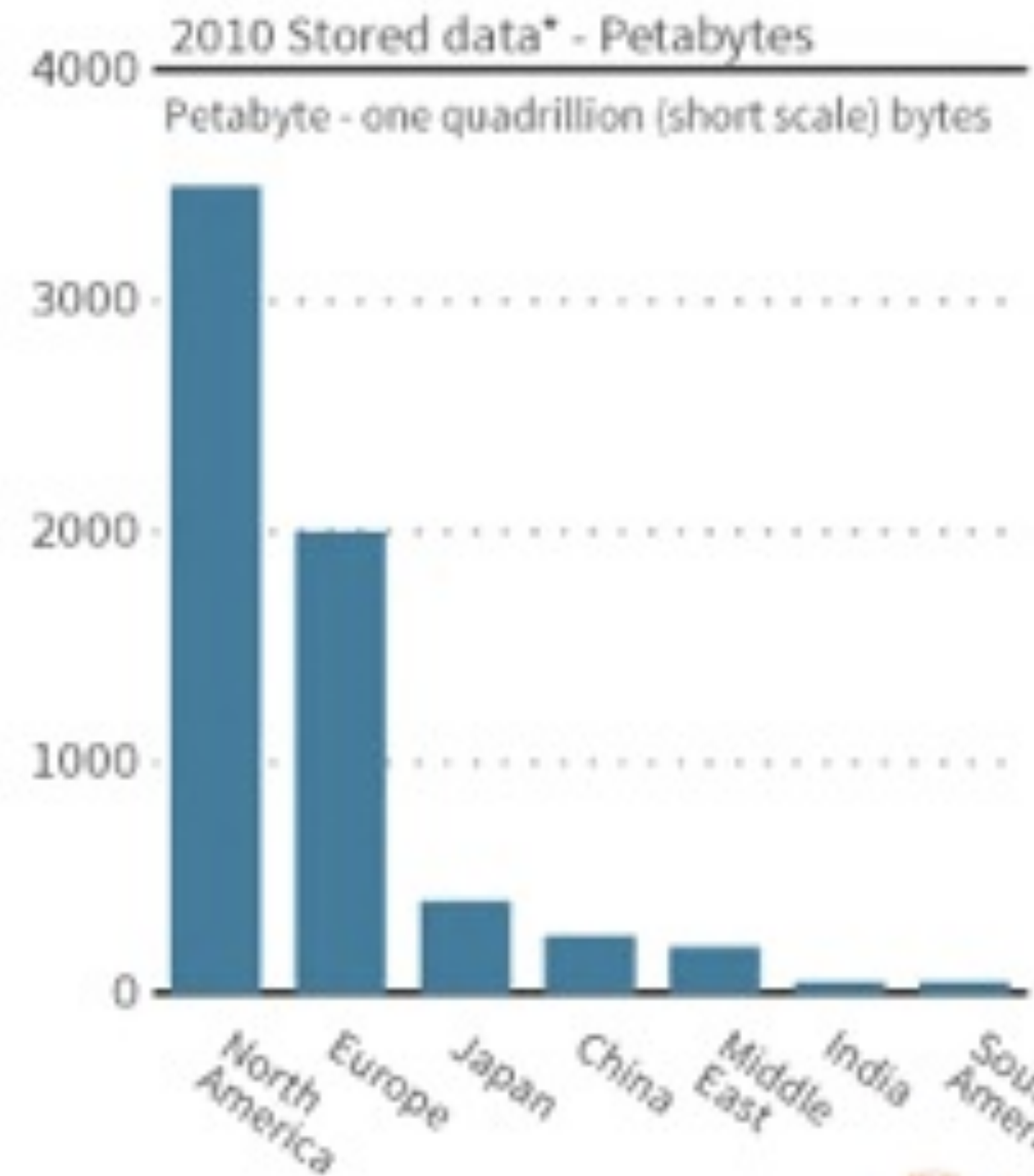
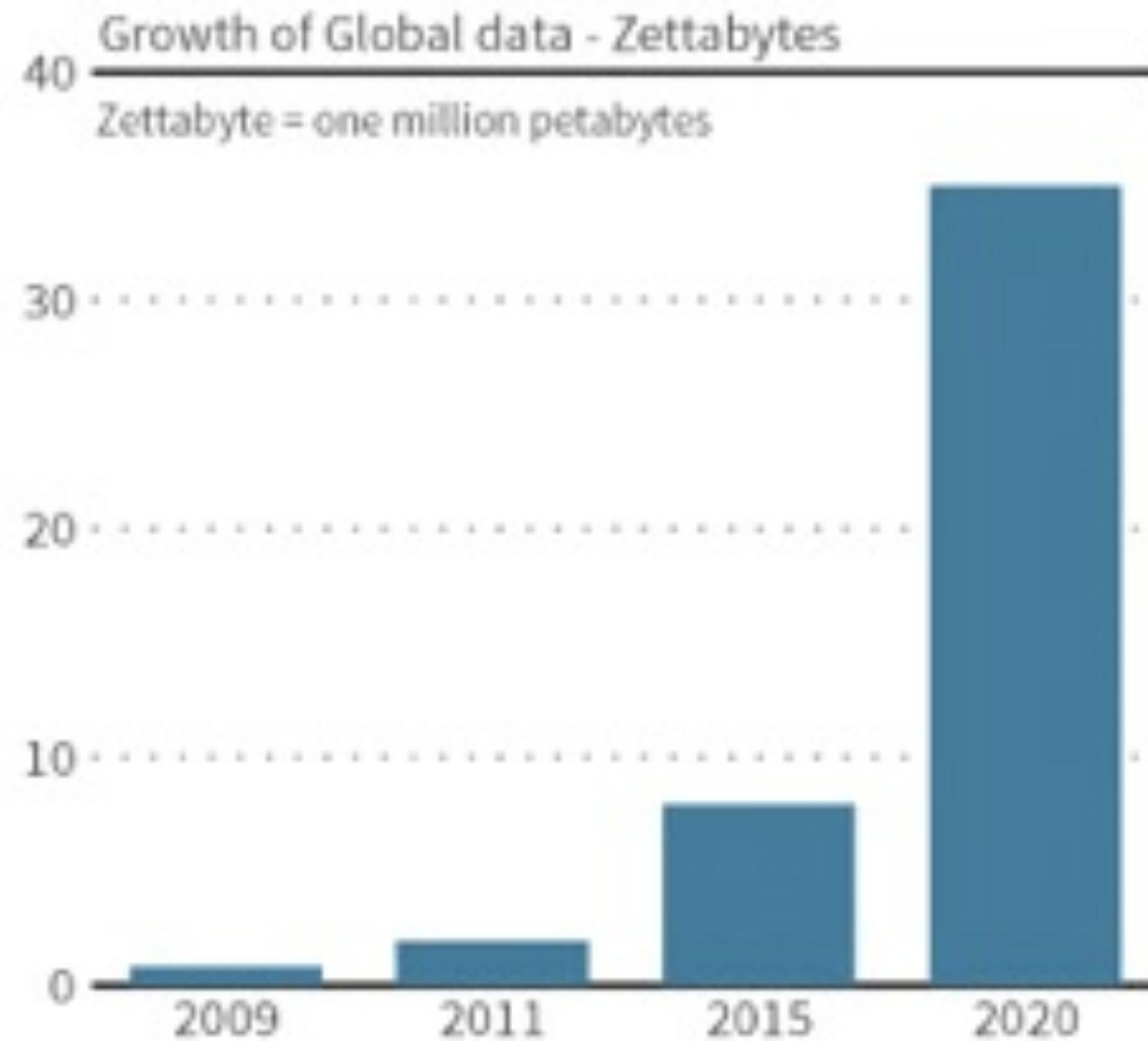
Big Data: Volume



Big Data: Volume



Big Data: Volume



*greater than

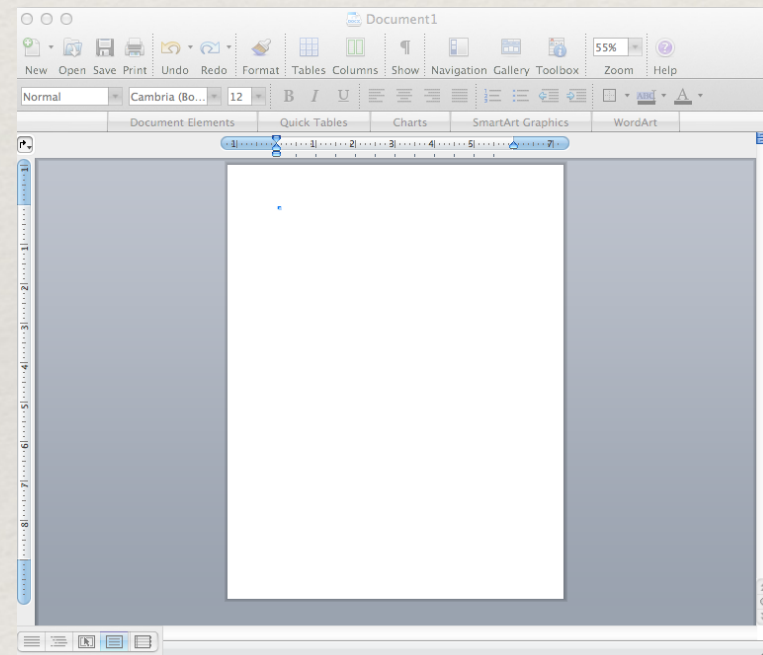
Sources: Nasscom -CRISIL GR&A analysis

Big Data: Volume, Velocity, Variety

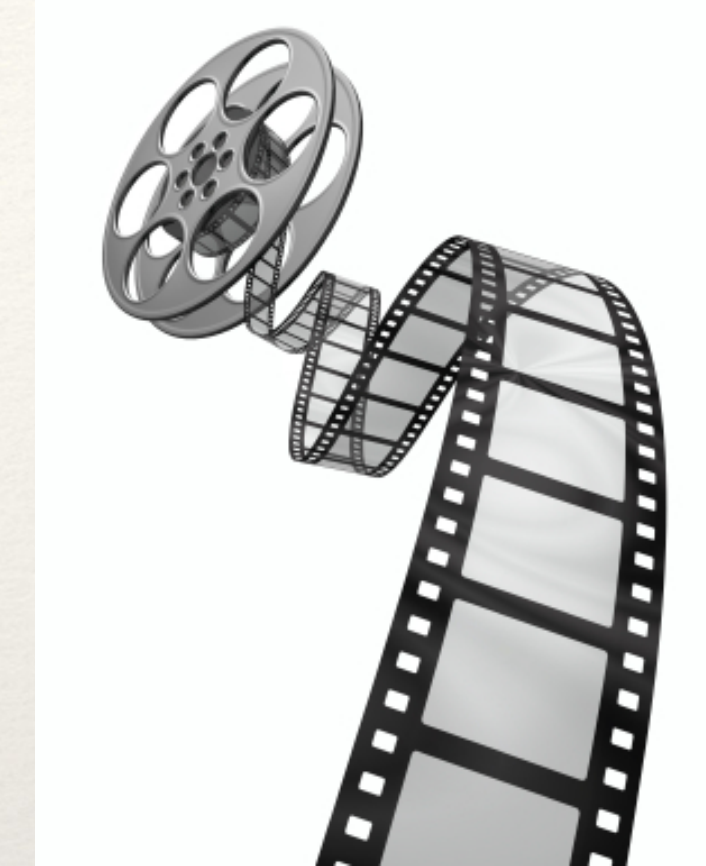
Numbers



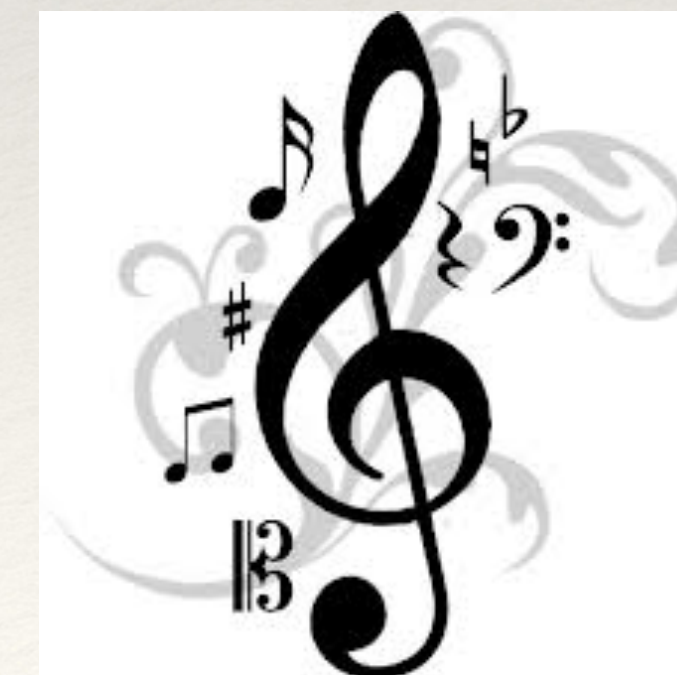
text



Images



sound



HealthCare Data

Patient records....

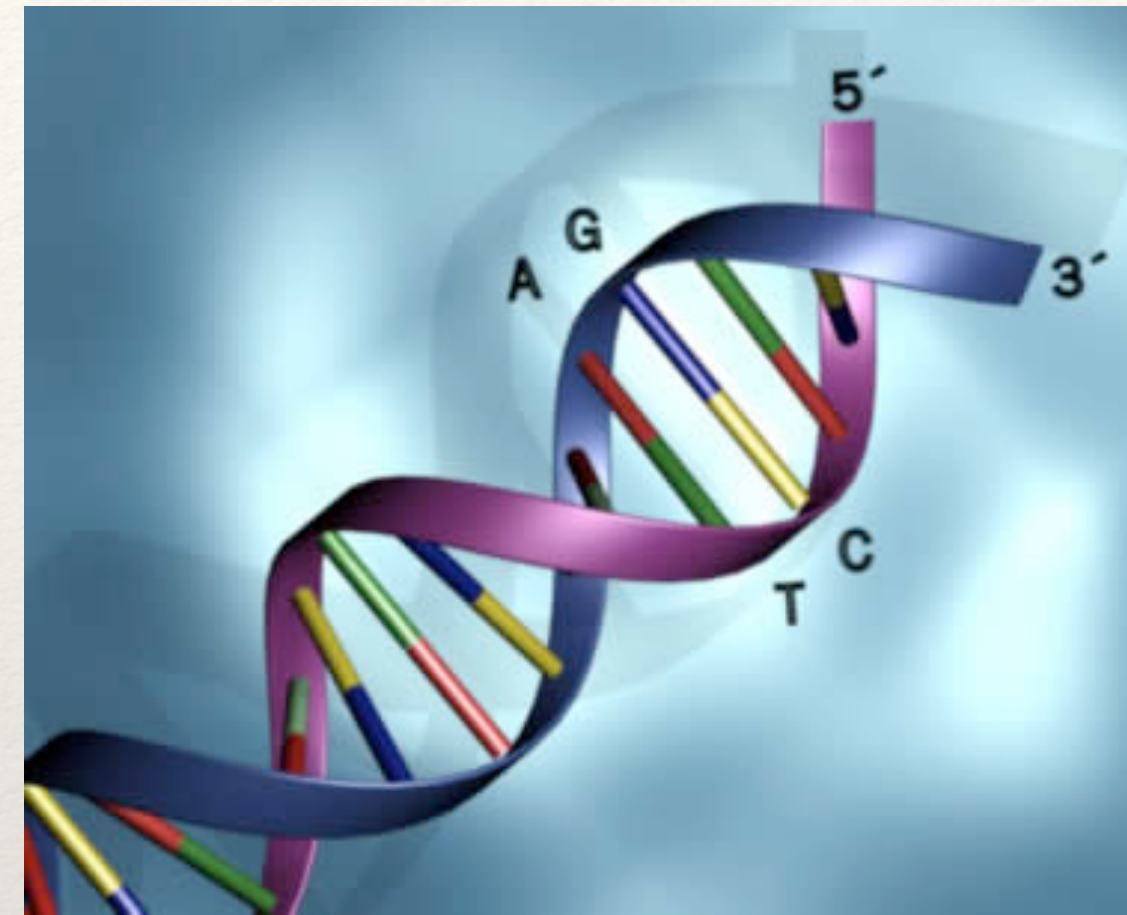


Test results....



HealthCare Data

Patient records....



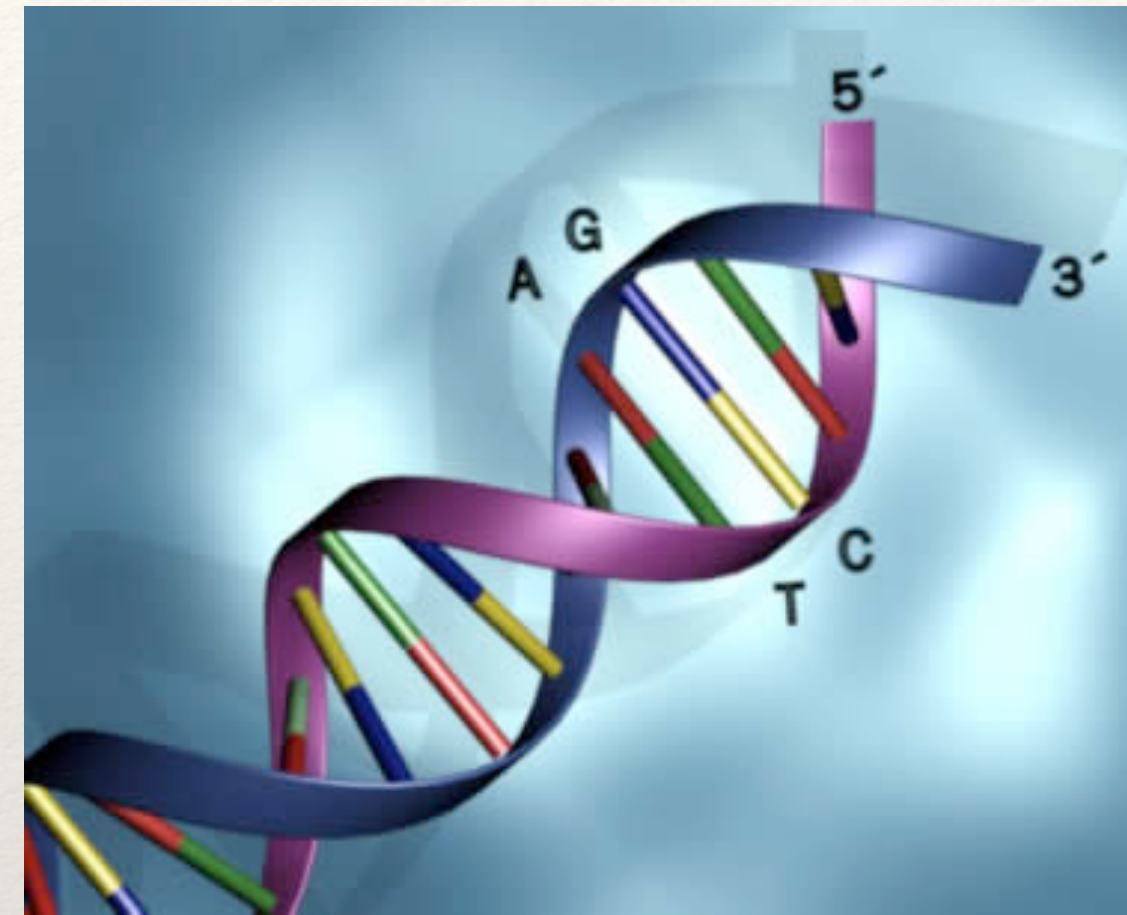
*Genomics
research*

Test results....



HealthCare Data

Patient records....

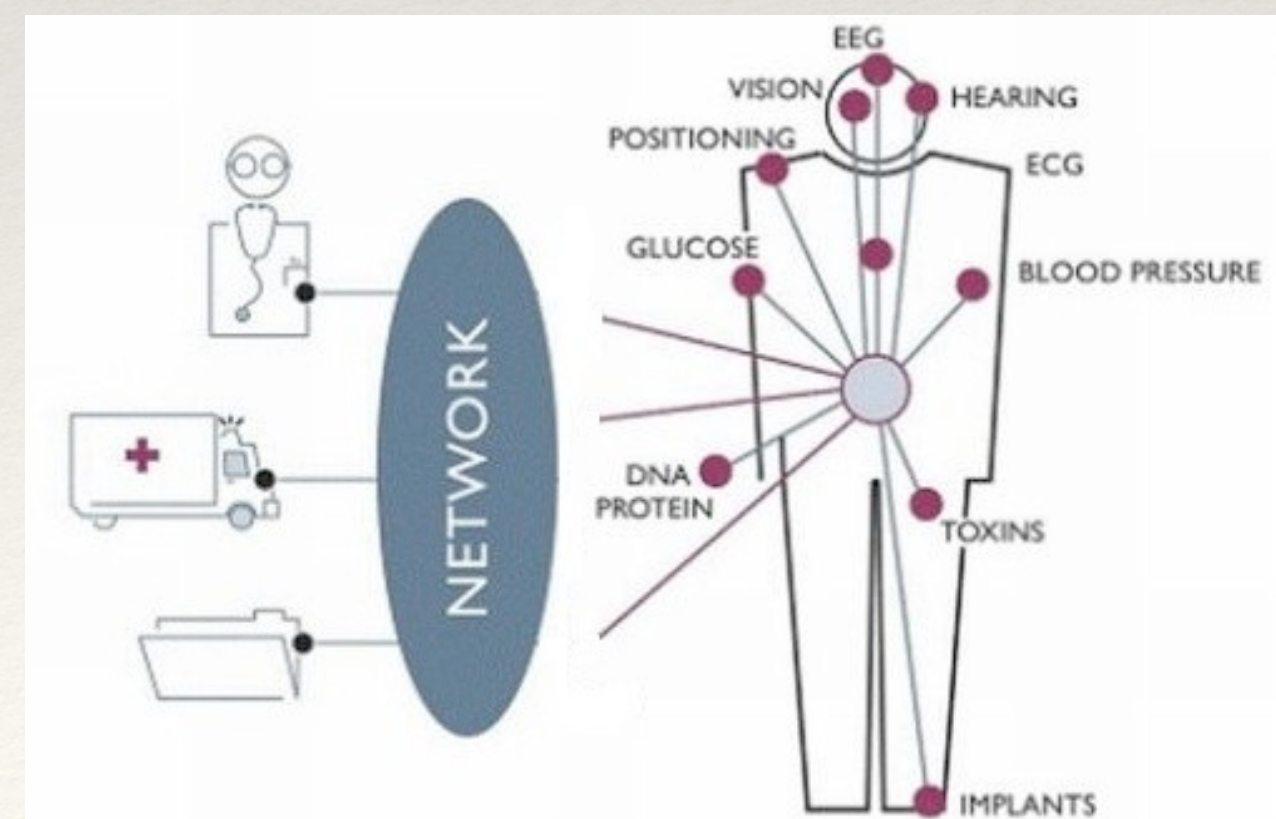


*Genomics
research*

Test results....



Wearable health monitoring...



Social medias...



Social Consequences of Commodity Sequencing

- ❖ The danger of misuse
predict sensitivities to various industrial or environmental agents → discrimination by employers?
- ❖ The impact of information that is likely to be incomplete
an indication of a 25 percent increase in the risk of cancer?
- ❖ Reversal of knowledge paradigm
- ❖ Are the "products" of the Human Genome Project to be patented and commercialized?
Myriad genetics and BRCA1/2
- ❖ How to educate about genetic research and its implications?

Social Consequences of Commodity Sequencing

Based on your genes, what is your
Sensitivity to Warfarin?
(or Coumadin®, a common blood thinning drug)



may require
typical dose



may require
decreased dose



23andMe will tell you:
Your drug sensitivity
What to tell your doctor

Social Consequences of Commodity Sequencing



Introduction to Data Science

1. A paradigm shift in Science
2. What is “Big Data”?
3. Learning from Data / Data Science Artificial Intelligence

What is Data Science?

“Data science is the study of extracting value from data”

Jeannette Wing

What is Data Science?

*“Data science is the study of extracting value from **data**”*

Jeannette Wing

What is Data Science?

*“Data science is the study of extracting **value** from data”*

Jeannette Wing

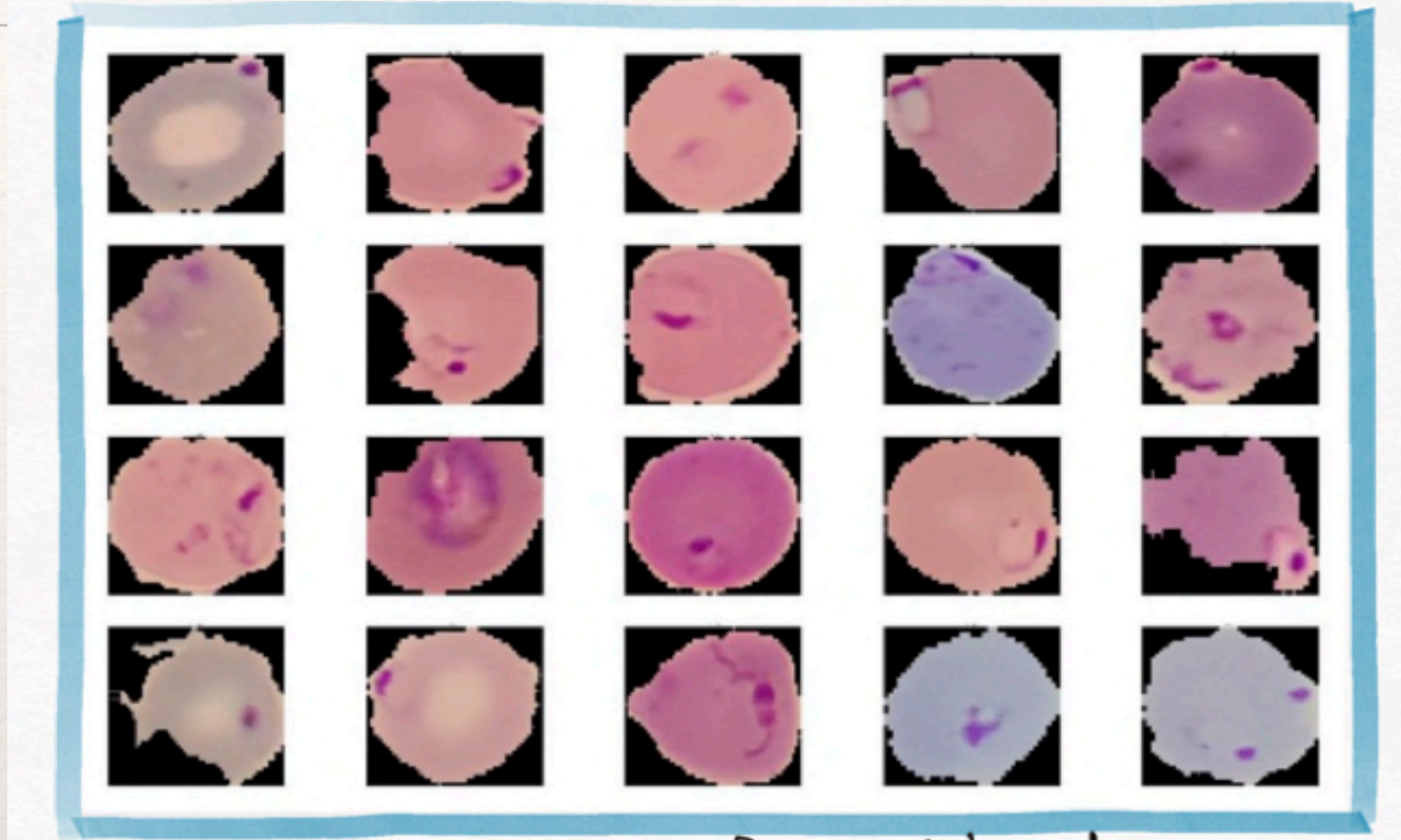
Data Science: Opportunities

- ❖ Fourth Paradigm: data driven science



Data Science: Opportunities

Disease Diagnosis



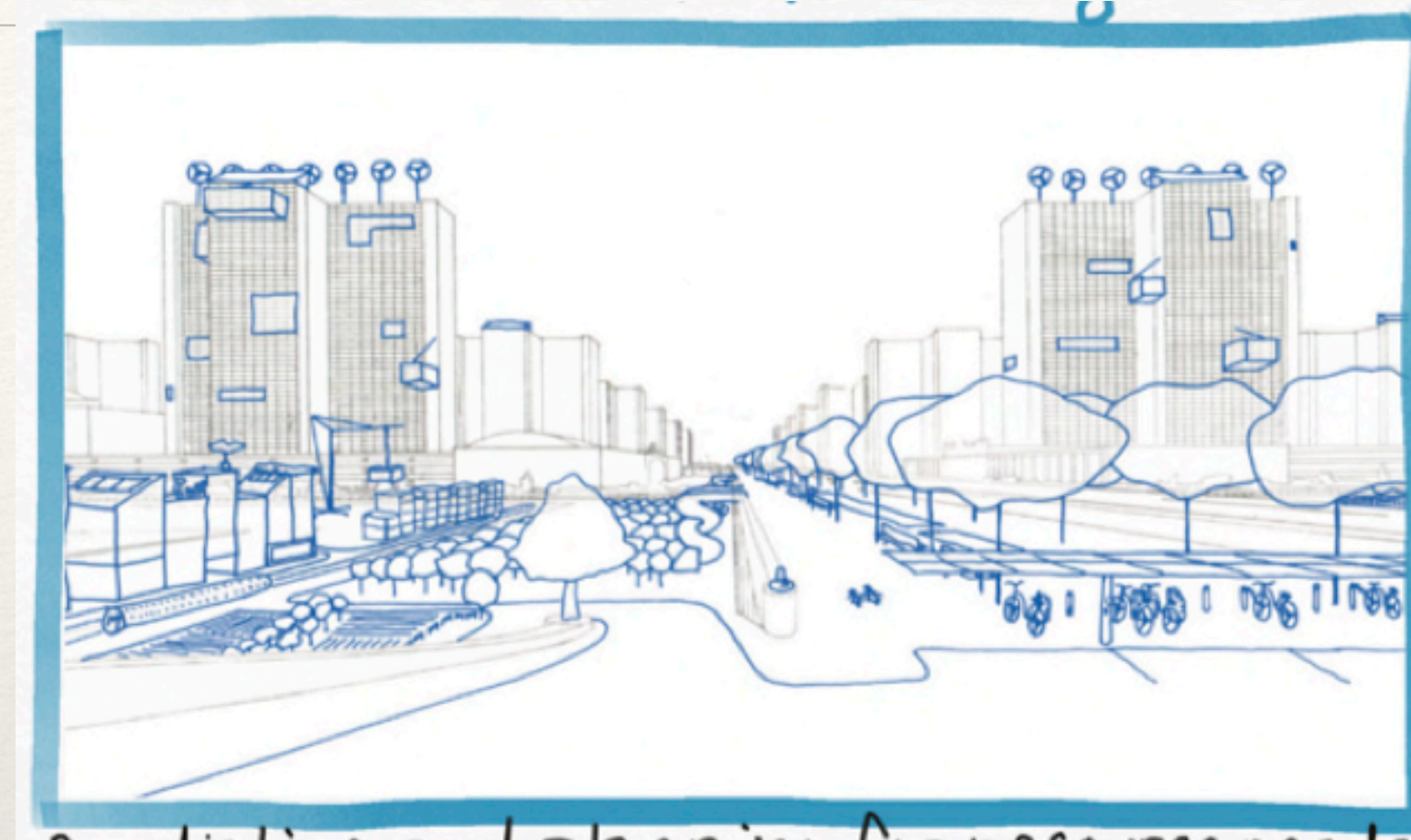
Detecting malaria from blood smears

Drug Discovery



Quickly discovering new drugs for COVID

Urban Planning



Predicting and planning for resource needs

Agriculture



Precision agriculture

What is Data Science?

*“Data science is the study of **extracting** value from data”*

Jeannette Wing

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

The Data Science Process

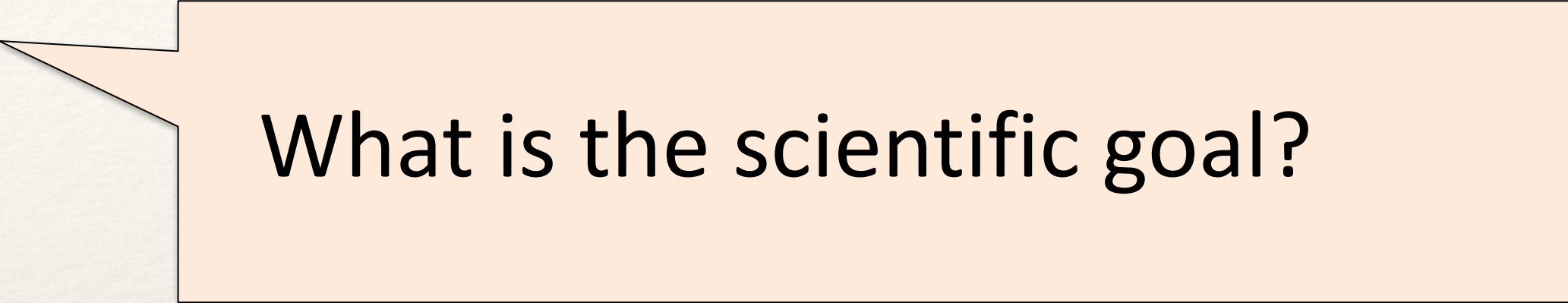
Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results



What is the scientific goal?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

How were the data obtained?

Which data are relevant?

Are there privacy issues?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Plot the data

Are there anomalies?

Are there obvious patterns?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Build a model

Fit the model

Validate the model

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

What did we learn?

Is it meaningful?

Does it have “value”?

Machine Learning

	<i>Supervised learning</i>	<i>Unsupervised learning</i>
<i>Discrete</i>	Classification of categorization	Clustering
<i>Continuous</i>	Regression	Dimensionality Reduction

What is Data Science?

*“Data science is the **study** of extracting value from data”*

Jeannette Wing

Big Data: Challenges

- ❖ Volume and Velocity
- ❖ Variety
 - ❖ Structured, Unstructured....
 - ❖ Images, Sound, Numbers, Tables,...
- ❖ Security
- ❖ Reliability, Integrity, Validity

Big Data: Challenges

Large N:

“Any dataset that is collected by a scientist whose data collection skills are far superior to the analysis tools available in her field”

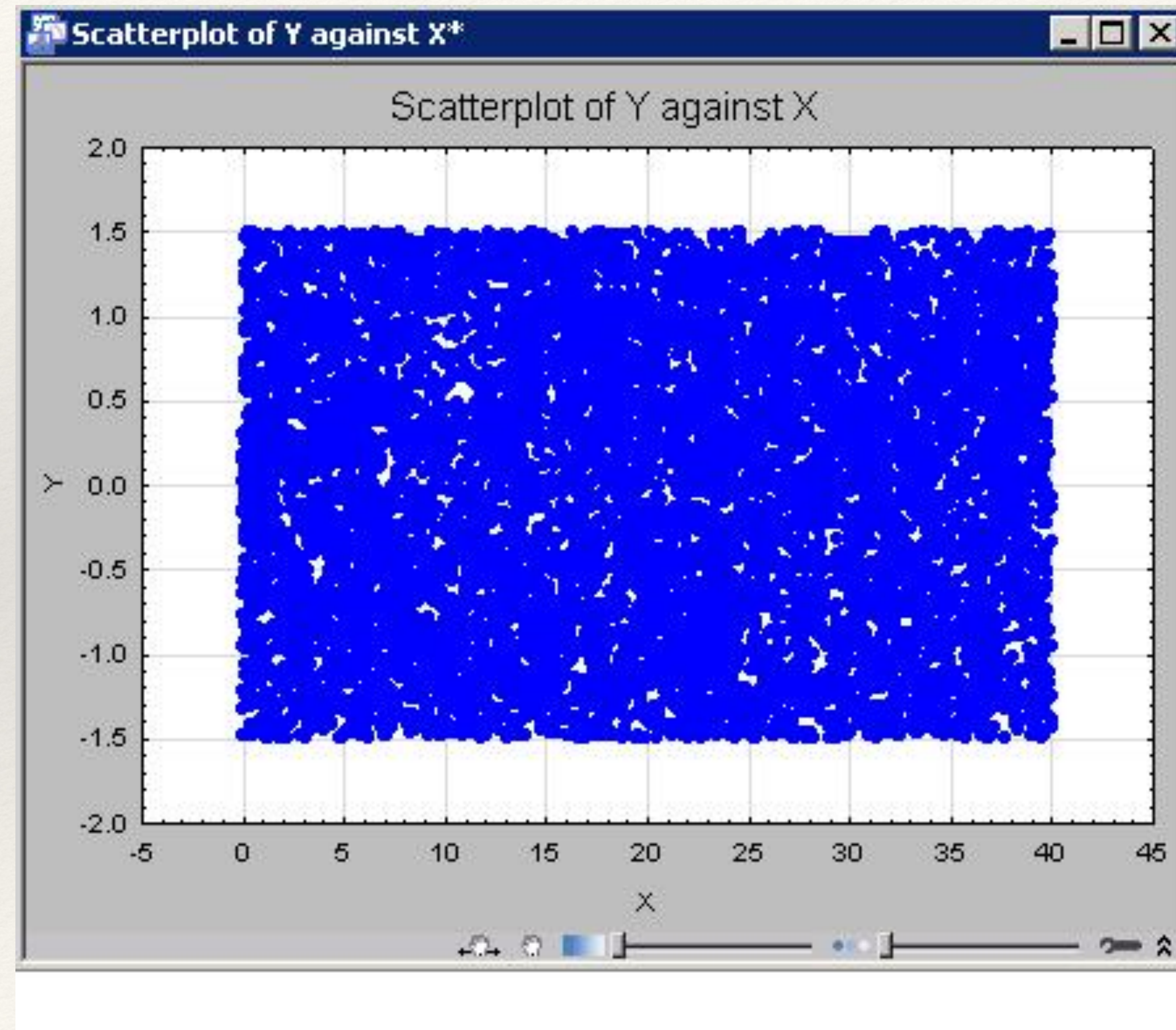
Computing issues:

- Data transfer
- Scalability of algorithms
- Memory limitations
- Distributed computing

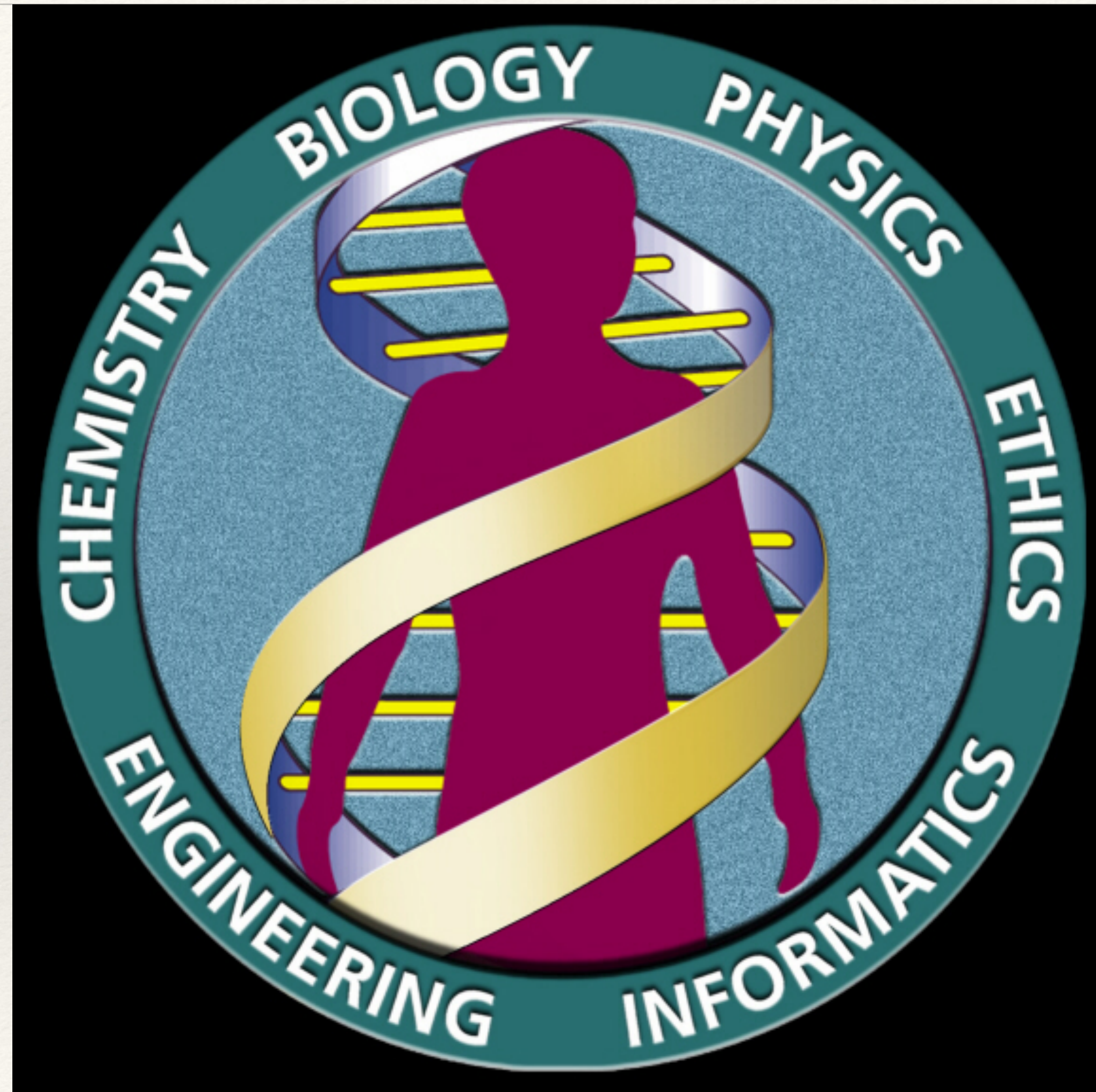
Big Data: Challenges

Vizualization issues:

The “black” screen
problem



How to Approach Data Science



How to approach Data Science

*Domain
Sciences*

Discover

Develop

Analytics

Training

Distribute

Infrastructure

