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

- Theory
- Experiments
- Data
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

1. Real World
2. Simplified Model
3. Mathematical Model
4. Computational Model
5. Results

- Interpret
- Simplify
- Simulate
- Represent
- Translate
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

<table>
<thead>
<tr>
<th></th>
<th>Byte</th>
<th>Kilobyte (KB)</th>
<th>Megabyte (MB)</th>
<th>Gigabyte (GB)</th>
<th>Terabyte (TB)</th>
<th>Petabyte (PB)</th>
<th>Exabyte (EB)</th>
<th>Zettabyte (ZB)</th>
<th>Yottabyte (YB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 bytes</td>
<td>1000 KB</td>
<td>1000 MB</td>
<td>1000 GB</td>
<td>1000 TB</td>
<td>1000 PB</td>
<td>1000 ZB</td>
<td>1000 YB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Big Data: Volume

Growth of Global data - Zettabytes
Zettabyte = one million petabytes

2010 Stored data* - Petabytes
Petabyte - one quadrillion (short scale) bytes

Sources: Nasscom - CRISIL GR&A analysis
Big Data: Volume, Velocity

A new style of IT emerging

Every 60 seconds

- 98,000+ tweets
- 695,000 status updates
- 11 million instant messages
- 698,445 Google searches
- 168 million+ emails sent
- 1,820TB of data created
- 217 new mobile web users
Big Data: Volume, Velocity, Variety

Numbers

Images

sound

text
HealthCare Data

*Patient records....*

*Test results....*
HealthCare Data

Patient records....

Test results....

Genomics research
HealthCare Data

Patient records....

Test results....

Wearable health monitoring...

Genomics research

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 ➔ Basic ➔ Knowledge ➔ Translational ➔ Societal Benefit
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?

“What 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

Questions:

- 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

1. Ask an interesting question
2. Get the Data
3. Explore the Data
4. Model the Data
5. Communicate/Visualize the Results

What did we learn?
Is it meaningful?
Does it have “value”?
## Machine Learning

<table>
<thead>
<tr>
<th>Supervised learning</th>
<th>Unsupervised learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification of categorization</td>
<td>Clustering</td>
</tr>
<tr>
<td>Regression</td>
<td>Dimensionality Reduction</td>
</tr>
</tbody>
</table>

- **Classification of categorization** is a form of supervised learning where the output is a discrete category.
- **Clustering** is an example of unsupervised learning where the goal is to group similar data points together.
- **Regression** is another form of supervised learning where the output is a continuous value.
- **Dimensionality Reduction** is also a form of unsupervised learning that aims to reduce the number of variables under consideration.
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

Visualization issues:

The “black” screen problem
How to Approach Data Science
How to approach Data Science

Domain Sciences

Discover

Develop

Training

Distribute

Analytics

Infrastructure