The 80/20 rule: can and should we break it using efficient data management tools?

Ghaleb Abdulla, Senior Data Scientist
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

DATAWorks 2019
The 80/20 Rule in Data Science

- Data scientists spend most of their time on data management and **wrangling** data.

- Is it 50/50, ...., 80/20, 90/10?

- Important tasks:
  - Data wrangling and management
    - Capture the user and domain experience
  - Systems and tools
  - Software engineering

"Yes, I am impressed at how fast you got here, but where's my pizza?"
“Learning is any process by which a system improves performance from experience.”

Herbert Simon, Department of Psychology, Carnegie Mellon University

“Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.”

Arthur Samuel (1959), Computer Science (AI), IBM

“Machine Learning is the study of algorithms that — improve their performance P — at some task T — with experience E. A well-defined learning task is given by <P, T, E> “

Tom Mitchell (1998), Computer Science Professor, Carnegie Mellon University
Traditional programming versus Machine Learning

- **Experience** → **Program** → **Data** → **Computer Programming** (expert systems) → **Output**

- **Experience** → **Data** → **Output** → **Computer ML** → **Program (model)**

**Rules are derived from the data**
Data Scientist: Where do we get the needed experience?
Data Scientist: Where do we get the needed experience?

Domain knowledge

Software eng., data intensive computing & systems

Algorithms and Tools
Use the domain knowledge to define the problem and the goal of the project—collect good data

- Data is
  - messy
  - Missing values
  - Hard to interpret
  - Contains aggregated or summarized values

- Study or experiments
  - Based on unknown assumptions
  - Not fully planned, data is collected after the system is built
    - True for most data science problems
National Ignition Facility (NIF), a complex human engineered system with a lot of data

- NIF is a large laser-based inertial confinement fusion research device at LLNL
- NIF uses lasers to heat and compress a small amounts of hydrogen fuel with the goal of inducing nuclear fusion
- Many instruments-detectors, oscilloscopes, interferometers, streak cameras, and other diagnostics-surround the target chamber to:
  - Measure the system’s performance
  - Record experimental results
Damage inspection and optics recycling are critical to NIF operation

- Sites initiate small but grow
- If a site grows larger than ~300 microns it cannot be repaired
- FODI is imaging ~40 cm optics from 7+ m away
  - This results in noisy, low-resolution images
- OI cleans images and infers site size from light scatter\(^1\)

Damages’ sizes over time for one of the NIF optics

We can use one of the many available ML algorithms to build a model
Did we ask all the relevant questions?

- How many of these Optics do we have?
- Are all the same type?
- Where are they physically installed in the system?
- Which side of the optic has the damage?
- After maintenance, do the optics get installed in the same place?
- Are there any other factors that could affect the site size?
  - Impurities or damages in the mirrors can play a part in the damage initiation

Some of these questions were not discussed before we started the work. It came across while we were examining the data, building models and testing it. “Oh, I forgot to mention....”
Privacy and security can introduce data challenges, the Norwegian cervical cancer data set

When to screen for cervical cancer

<table>
<thead>
<tr>
<th>Age</th>
<th>Screening Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 y</td>
<td>Begin screening for cervical cancer</td>
</tr>
<tr>
<td>21-30 y</td>
<td>Pap test every 3 years if results normal</td>
</tr>
<tr>
<td>31-64 y</td>
<td>Pap test every 3 years or Pap test + HPV test every 5 years</td>
</tr>
<tr>
<td>65 y and older</td>
<td>Stop routine screening if results normal for the previous 10 years</td>
</tr>
</tbody>
</table>

Goal: Move from one size fits-all (in this case age group) to a more precise (personalized) screening approach
Datasets

Screening data

<table>
<thead>
<tr>
<th>ID</th>
<th>Birthdate</th>
<th>Diagnosis Date</th>
<th>Test Type</th>
<th>Diagnosis</th>
<th>Region</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12/15/1978</td>
<td>2/15/1998</td>
<td>Cyt</td>
<td>11</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>8/15/2004</td>
<td>12/15/1978</td>
<td>Cyt</td>
<td>11</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>10/15/2007</td>
<td>15/15/1978</td>
<td>Cyt</td>
<td>11</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>4/15/2010</td>
<td>22/15/1978</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>5/15/1956</td>
<td>7/15/1999</td>
<td>Cyt</td>
<td>12</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8/15/1999</td>
<td>15/15/1956</td>
<td>Cyt</td>
<td>11</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10/15/1999</td>
<td>20/15/1956</td>
<td>Hist</td>
<td>20</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Survey data

<table>
<thead>
<tr>
<th>Question</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marital Status</td>
<td>0 married, 1, single, 2 divorced</td>
</tr>
<tr>
<td>Do you smoke?</td>
<td>1-2 times/week, 3-4 times/week, 3 daily, 4 quit smoking, 5 never</td>
</tr>
<tr>
<td>When did you start smoking?</td>
<td>Age</td>
</tr>
<tr>
<td>When did you quit smoking?</td>
<td>Age</td>
</tr>
<tr>
<td>How many cigarettes on average did you smoke per week at age __?</td>
<td>Number</td>
</tr>
<tr>
<td>. . . when you were 14-17 years old?</td>
<td>Number</td>
</tr>
<tr>
<td>. . . when you were 18-21 years old?</td>
<td>Number</td>
</tr>
<tr>
<td>. . . when you were 22-24 years old?</td>
<td>Number</td>
</tr>
<tr>
<td>Have you been diagnosed with chlamydia before?</td>
<td>1 yes, 0 no</td>
</tr>
<tr>
<td>When were you diagnosed with chlamydia?</td>
<td>Age</td>
</tr>
<tr>
<td>Have you ever had sex intercourse?</td>
<td>1 yes, 0 no</td>
</tr>
<tr>
<td>Age at first sex</td>
<td>Age</td>
</tr>
<tr>
<td>How many sexual partners have you had until now?</td>
<td>Number</td>
</tr>
<tr>
<td>(Date of response, other STD history, etc.)</td>
<td>...</td>
</tr>
</tbody>
</table>

• Data anonymization
• Test dates were shifted for extra anonymization
• HPV tests were done after Cytology
• Test results are coded with multiple codes, needed to aggregate or summarize

Static in time, can we extract feature values over the lifetime of the patient?
Filling the gaps - fusing health register with survey data to create a rich temporal dataset

- This work required 3 months of a PhD student collaborating with three scientists.
ESGF an Integrated data ecosystem, a system that provides tools for efficient data wrangling and management

- Enabling integrated research
- Data Center and interoperable services
- Virtual Laboratory infrastructure

Critical Complex Data Generation Systems

Data Collection and Management
- Sensors, field and lab experiments
- Data models
- Transport and communications
- Data quality and uncertainty

Data-Intensive Computing
- Architectures – persistent data to streams
- Programming environments
- Human computer interface

Decision and Control
- Design Optimization
- Policy Making (Humans)
  - Understanding and predicting use

Security

Data Analytics (local & remote)
- Descriptive statistics
- Graph analytics
- Machine learning
- Signal and image processing
- Pattern discovery
- Visualization
- Exploratory analysis

Provenance Capture

Network
ESGF software system and science integration

- NetCDF Climate and Forecast (CF) metadata convention, CDMS
- Climate Model Output Rewriter 3 (CMOR-3), controlled vocabulary
- Publishing
- Search and discovery
- Replication and transport
- Data Reference Syntax (DRS)
- Earth System Documentation (ES-DOC)
- Quality Control (QC)
- Websites and Web portal development
- Security
- Notifications, monitoring, metrics
- Product services
Use Demographic

Statistics Overview

- ESGF total number of datasets: 701,244
- ESGF total data volume: 4,635.828 TB
- CMIP5 total number of datasets: 150,824
- CMIP5 total data volume: 4,261.921 TB
- OBS454 total number of datasets: 365
- OBS454 total data volume: 0.285 TB
- CORDEX total number of datasets: 68,709
- CORDEX total data volume: 59,813 TB

Downloads by Continent and Country

Registered Users by Continent and Country
Back to our original question:
The 80/20 rule, can and should we break it using efficient data management tools?

- We should provide the tools that makes data wrangling more efficient and less error prone:
  - Formalize the process of learning from the domain (interacting with scientists or domain experts)
  - Standard set of questions that need to be asked
  - Data transformation steps that need to be made
  - ............

- Capture data provenance and data manipulation history (code that touched the data)
  - Keep track of all the data manipulation work for reproducibility, verification, and testing.
    - Save the code, decisions, rationale behind the decisions, etc.
  - Workflow
    - Use case/domain dependent