AI Data Architecture

Jeremy Kepner, Vijay Gadepally, Lauren Milechin, Sid Samsi

Outline

• Introduction
  • Tabular Data
  • Files and Folders
  • Using and Sharing
  • Summary
### Data and Challenges Drive Breakthroughs in AI

- 80% of AI effort can be data wrangling/architecture

<table>
<thead>
<tr>
<th>Year</th>
<th>Breakthroughs in AI</th>
<th>Datasets (First Available)</th>
<th>Algorithms (First Proposed)</th>
</tr>
</thead>
</table>

**Average No. of Years to Breakthrough:**

- **3 years**
- **18 years**


---

**Data Architecture: Formats**

- **Input Features**
  - $y_0$
  - $y_1$
  - $y_2$
  - $y_3$
- **Output Classification**
  - $y_4$

- All analysis (AI and otherwise) relies on tabular data

- $y_{l+1} = h(W_l y_l + b_l)$
All analysis (AI and otherwise) relies on tabular data.

\[ Y_{l+1} = h(W_l Y_l + B_l) \]
Data Architecture: Organization

0. Raw  
1. Parse  
2. Ingest  
3a. Query  
4. Analyze

- Easy-to-use  
  - Requires a minimal number of software tools implement

- Easy-to-understand  
  - Processing steps and data flow are apparent from inspection

- Easy-to-maintain  
  - Minimize dependencies (technical debt) by relying on features that are built into operating systems

Data Architecture: Sharing

- Creating and sharing challenge quality data accelerates external and internal AI progress
- Requires strong collaboration amongst stakeholders
- Key: co-design AI application and sharing protocol
Tables are the Natural Format of Data Analysis

- Used by humans for thousands of years
  - Allows data to be visually inspected
- Tabular data is compatible with nearly all analysis software
  - Spreadsheets: Microsoft Excel, Google Sheets, Apple Numbers, CSV, TSV
  - Databases: MySQL, PostgreSQL, Oracle, SQL Server, Accumulo, SciDB
  - Neural Networks: TensorFlow, Torch, MxNet
  - Languages/Libraries: Python/NumPy/Spark, Julia, R, Matlab/Octave/D4M/GraphBLAS
  - Hierarchical File Formats: JSON, XML
Spreadsheets

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>Code</td>
<td>Name</td>
<td>Job</td>
<td>Date</td>
<td>$ in bank</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>A0001</td>
<td>Alice</td>
<td>scientist</td>
<td>2000 Jan 01</td>
<td>$11,700</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>B0002</td>
<td>Bob</td>
<td>engineer</td>
<td>2001 Jan 01</td>
<td>$10,600</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>C0003</td>
<td>Carl</td>
<td>mathematician</td>
<td>2002 Jan 01</td>
<td>$10,200</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2003 Jan 01</td>
<td>$8,600</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2004 Jan 01</td>
<td>$10,400</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2005 Jan 01</td>
<td>$10,600</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>y=10</td>
<td>6</td>
<td>12</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>2006 Jan 01</td>
<td>$10,900</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>y=8</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>2007 Jan 01</td>
<td>$12,500</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>y=6</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td>2008 Jan 01</td>
<td>$12,600</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>y=4</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td>2009 Jan 01</td>
<td>$9,000</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>y=2</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>2010 Jan 01</td>
<td>$10,600</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>y=0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>2011 Jan 01</td>
<td>$11,700</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>x=0</td>
<td>x=5</td>
<td>x=10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Flexible representation of diverse data
- Used by 100M+ people daily

CSV and TSV

**CSV (Comma Separated Values)**

- `<filename>.csv`
- `<filename>.CSV`

**Code,Name,Job**

A0001,Alice,scientist
B0002,Bob,engineer
C0003,Carl,mathematician

**TSV (Tab Separated Values)**

- `<filename>.tsv`
- `<filename>.TSV`

**Code	Name	Job**

A0001 Alice scientist
B0002 Bob engineer
C0003 Carl mathematician

- Every analysis application should read and write CSV and TSV
- Easily readable/viewable/writable (Excel, Sheets, Numbers, …)
- TSV preferred (if available) as it enables faster reading and writing
Databases

- SQL database are good for medium size datasets that require ACID (atomicity, consistency, isolation, and durability) guarantees
- NoSQL databases are good for large datasets where ACID guarantees aren’t required
- NewSQL database are good when you have a need for scalability + ACID compliance

Operation: finding 1.1.1.1’s nearest neighbors

D4M = Dynamic Distributed Dimensional Data Model (d4m.mit.edu)
Mathematics of Big Data, Kepner & Jananthan, MIT Press 2018

Languages, Libraries, AI Packages

- Lots of machine learning software
- Designed for tabular data
- Jupyter interactive portal interface
  – Similar to Mathematica notebooks
Hierarchical data is an increasingly important form of "big data"

- Can be converted to sparse tables

Tabular Terminology

- Tables have been implemented many ways
- Lots of different names for the same concepts
Tabular Data is a Natural Interchange Format

Outline

- Introduction
- Tabular Data
  - Files and Folders
  - Using and Sharing
- Summary
Files and Folders

0. Raw  1. Parse  2. Ingest  3a. Query  4. Analyze

- Raw Data → Parsed Data → Database → Query/Scan Results

3b. Scan

Standard data processing pipeline

- Easy-to-use
  - Use built-in tools that incur minimal technical debt (simple files and folders)
- Easy-to-understand
  - Tabular file formats; source and time naming scheme
- Easy-to-maintain
  - Pipeline folder structure


Tabular File Formats

- Parse as much data as is practical into tabular files
  - Try to avoid proprietary formats
- Use .csv (comma separated values)
  - Even better .tsv (tab separated values) file formats
- Column labels
  - Each column label should be unique within the file
- Row labels in the first column
  - Row number or record number (each row label should be unique within the file)
- When lots of entries are empty (the data is sparse)
  - Use triples format: row, column, value
File Naming

- Avoid lots of tiny files (compress with zip when necessary)
  - Better to have fewer larger files (1MB to 100MB per file is a common practice)
- Use hierarchical directories keep of items in a directory <1000
- Use easy-to-understand names (source and time) are easy to share
  - source/YYYY/MM/DD/hh/mm/source/YYYY-MM-DD-hh-mm-ss.tsv
  - YYYY/MM/DD/hh/mm/source/YYYY-MM-DD-hh-mm-ss-source.tsv
- Databases and files can often be used together
  - Databases are good for quickly finding a particular data items (i.e., a small number of records when compared to the entire dataset)
  - Scanning files in the file system can be best when reading in a majority of a datasets
- SQL database are good for medium size datasets that require ACID (atomicity, consistency, isolation, and durability) guarantees
- NoSQL databases are good for large datasets where ACID guarantees aren't required.
- NewSQL database are good when you have a need for scalability + ACID compliance

Folder Structure

- Folders are built into all operating systems
  - Easy-to-share
  - Easy-to-maintain
- Data processing pipeline is common
  - Easy-to-understand
- Allows team members to contribute quickly
- `dataFileList.txt` enables scalable processing
  - Avoids many processors listing files
Most data starts out unusable and unsharable
Understanding data & purpose are necessary for usage and sharing
Identify keeper of data
  - Have one person get a copy of data (on behalf of team); Fewer accounts is good OpSec
Convert a sample to tabular form
  - Identify useful columns (features) and excise the rest
  - Minimize/anonymize/simulate/surrogate data
  - Obtain preliminary approval
  - Test with AI users
  - [repeat]
Create file naming and folder structure and apply to required data
Automate at data owner site so they have AI ready data that they can use themselves and share with others
Data Owner SME Research Engagement

- Nearly all effective AI data products are collected and curated by AI researchers
- Data owner SME engagement with the AI research community is essential for the construction of effective AI data products
- Data owner SMEs should be encouraged to engage with these communities
- Data owner SMEs should be resourced to publish and fully participate in academic research conferences

SME = Subject Matter Expert

Limiting Data Sharing Concerns

- Confusion on data sharing liability limits willingness to share data with researchers
- Data owners aim for the common denominator of international requirements (US, EU, ...)
- Standard practices exist that meet these requirements
  - Data available in curated repositories
  - Use standard anonymization methods where needed: hashing, sampling, simulation, ...
  - Access requires registration with repository and legitimate research need
  - Recipients agree to not repost corpus and not deanonymize data
  - Recipients can publish analysis and data examples necessary to review research
  - Recipients agree to cite the repository and provide publications back to repository
  - Repository can curate enriched products developed by researchers
- Funding agencies, journals, conferences and professional societies should encourage research conducted performed under these conditions
SMEs learn ISO terminology

- Sharing AI Data often requires Information Security Officer (ISO) sign-off
- ISOs and Subject Matter Experts (SMEs) have different terminology
- ISOs sign off requires confidence in SME data handling practices
- ISOs need basic information to allow data sharing
  - project, need, location, personnel, duration, ...
- SMEs often provide research descriptions that limit ISO security surety in SMEs data handling practices and results in ISOs limiting of data sharing requests

Example ISO Question and Answer (#1)

- What is the data you’re seeking to share?
  - Describe the data to be shared, focusing on its risk to the organization if it were to be accidently released to the public or otherwise misused.

- Example
  - The data was collected on <<date range>> at <<location(s)>> in accordance with our mission. The risk has been assessed and addressed by an appropriate combination of excision, anonymization, and/or agreements. The release to appropriate legitimate researchers will further our mission and is endorsed by leadership.

- Explanation
  - Sentence 1 establishes the identity, finite scope, and proper collection of the data. Sentence 2 establishes that risk was assessed and that mitigations were taken. Sentence 3 establishes the finite scope of the recipients, an appropriate reason for release, and mission approval.
Example ISO Question and Answer (#2)

• Where / to whom is the data going?
  – Please describe the intended recipients of the data, the systems they will use to receive / process the data.

• Example
  – The data will be shared with researchers at <<institution>>. The data will be processed on <<institution>> owned systems meeting their institution security policies, which include password controlled access, regular application of system updates, and encryption of mobile devices such as laptops. Authorized access to the data will be limited to personnel working as part of this effort.

• Explanation
  – Sentence 1 establishes the legal entity trusted with the data and with whom any agreements are ultimately made on behalf of. Sentence 2 establishes that basic technical safeguards are in place, without getting too specific, and that personally-owned computers will not be used as the institution has no legal control over them. Sentence 3 establishes that the data will not be used for other purposes than the agreed-upon research project.

Example ISO Question and Answer (#3)

• What controls are there on further release (policy/legal & technical)?
  – Is a non-disclosure or data usage agreement in place?
  – Is the data anonymized? If so, is there an agreement in place to prohibit de-anonymization attempts?
  – What technical controls are in place on the systems that will receive / process the data to prevent misuse?
  – Is there an agreement in place on publication of results from this effort?
  – Is there an agreement in place for the retention or deletion of the original data, intermediate products, and/or the results at the end of the effort?

• Example
  – An acceptable use guidelines that prohibit attempting to de-anonymize the data and will be provided to all personnel working on the data. Publication guidelines have been agreed to that allow for high-level statistical findings to be published, but prohibit including any individual data records. A set of notional records has been provided that can be published as an example of the data format, but is not part of the actual data set. The research agreement requires all data to be deleted at the end of the engagement except those items retained for publication.

• Explanation
  – Sentence 1 establishes that there is an agreement in place prohibiting de-anonymizing the data and clearly defining it as “misuse” of the data. Sentence 2 and 3 establish that it is known to all parties what may and may not be published. Sentence 4 establishes that data retention beyond the term of the agreement has been addressed and cleanup is planned as part of project closeout.
Summary

• Data and Challenges Drive Breakthroughs in AI
  – 80% of AI effort is data wrangling/architecture

• All analysis (AI and otherwise) relies on tabular data

• File and Folder Organization
  – Easy-to-use, easy-to-understand, easy-to-maintain

• Co-Design Using and Sharing
  – What is good for others is great for yourself