

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

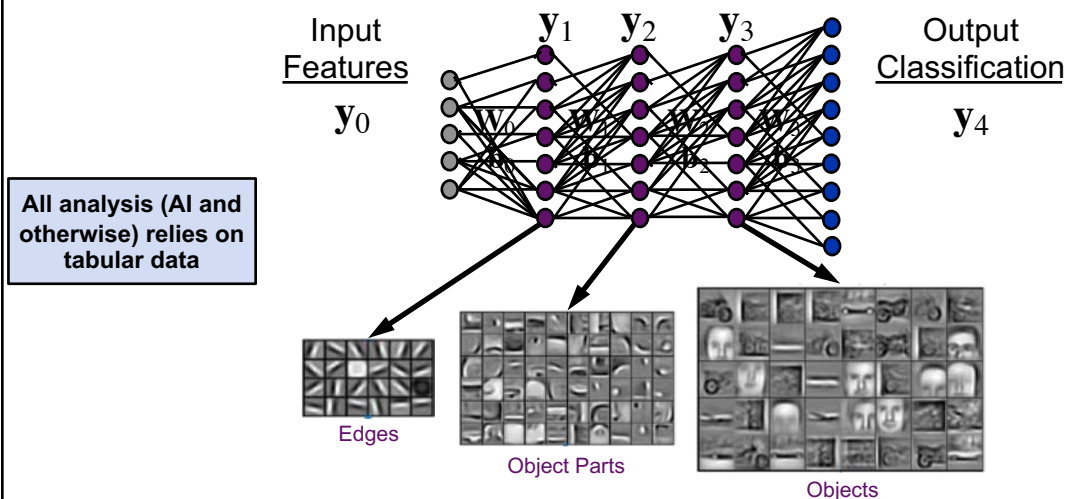
- 80% of AI effort can be data wrangling/architecture -

Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level read-speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogLeNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolutional neural network algorithm (1989)
2015	Google's Deepmind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		3 years	18 years

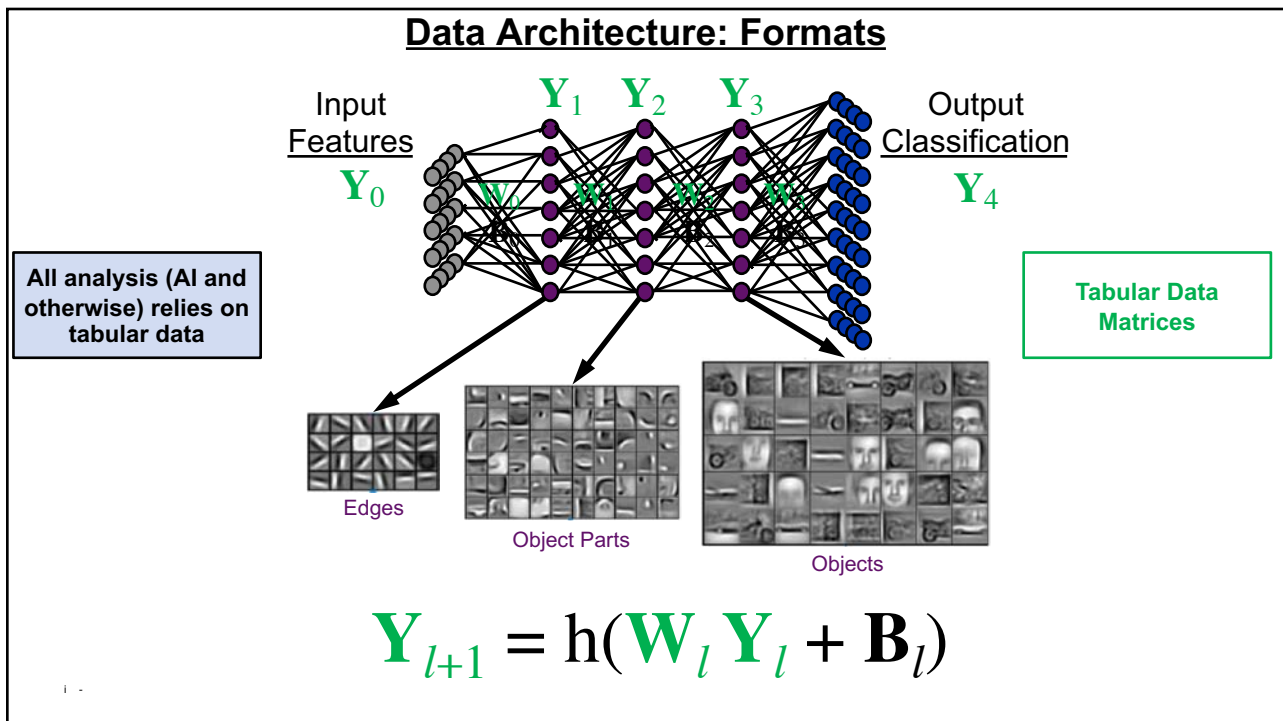
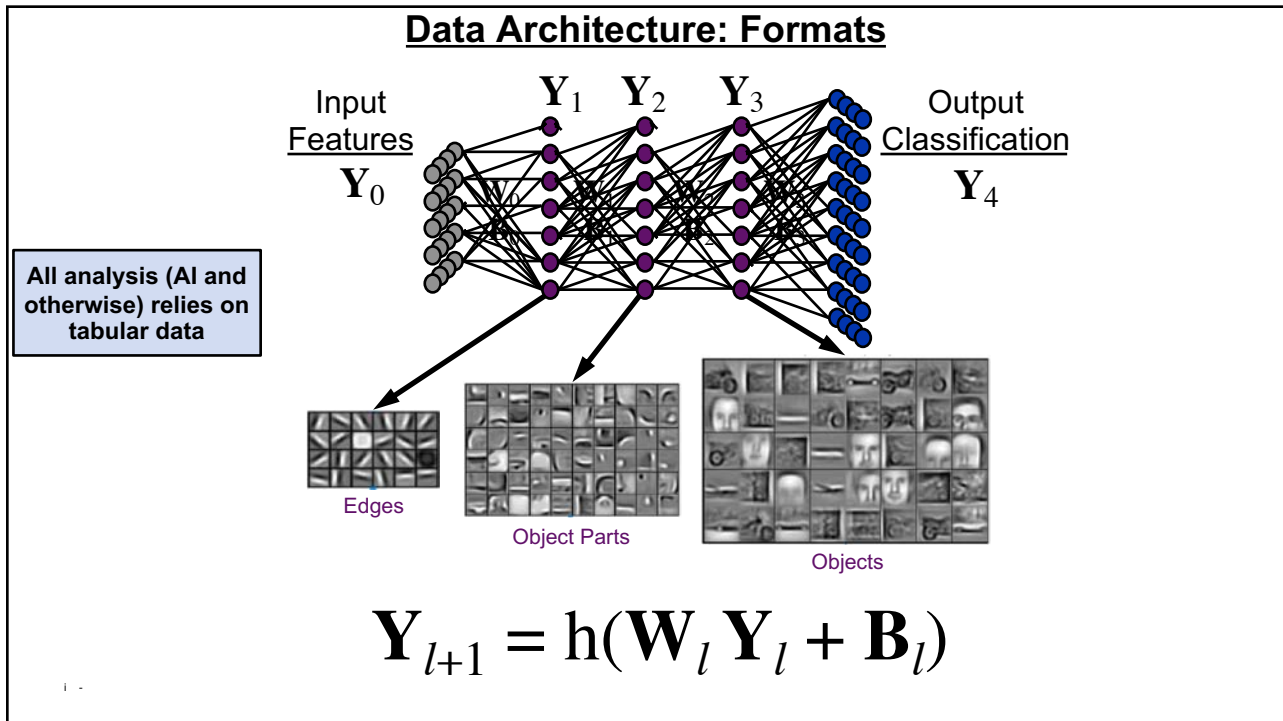
Slide - 3

Source: Train AI 2017, <https://www.crowdfunder.com/train-ai/>

Data Architecture: Formats

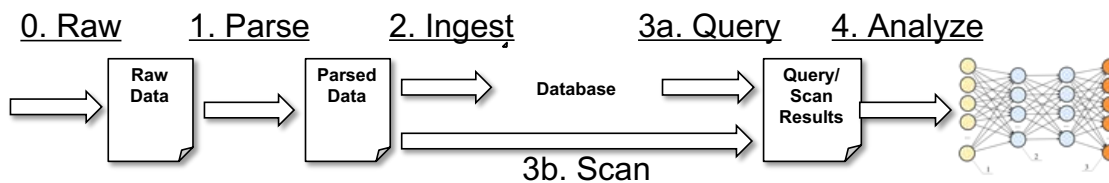


$$y_{l+1} = h(W_l y_l + b_l)$$





Data Architecture: Organization

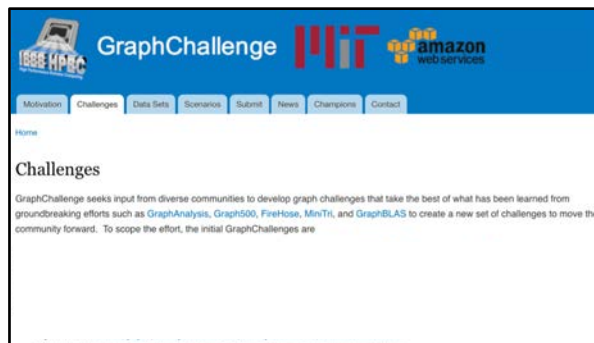
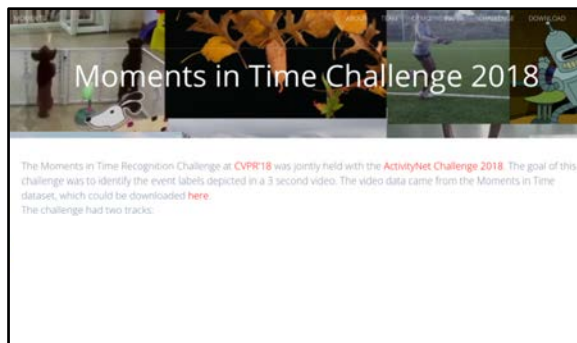


- **Easy-to-use**
 - Requires a minimal number of software tools implement
- **Easy-to-understand**
 - Processing steps and data flow are apparent from inspection
- **Easy-to-maintain**
 - Minimize dependencies (technical debt) by relying on features that are built into operating systems

Slide - 7



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**

Slide - 8



Outline

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

Slide - 9



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**

© St John's College, Oxford. Figure 6: Computus Table, (MS. 17, fol. 30r; St. John's College, Oxford) in F.T. Marchese, Pace University, "Exploring the Origins of Tables for Information Visualization." All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>

Table of Dionysius Exiguus produced in the 12th century in Cambridgeshire, England.¹

- **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**

Slide - 10

¹F. T. Marchese, "Exploring the origins of tables for information visualization," in 15th International Conference on Information Visualisation (IV), 2011, pp. 395–402, IEEE, 2011.



Spreadsheets

	A	B	C	D	E	F	G	H	I	J	K
1	1	2	2	1		Code	Name	Job		Date	\$ in bank
2	2	3	3	2		A0001	Alice	scientist		2000 Jan 01	\$11,700
3	2	3	3	2		B0002	Bob	engineer		2001 Jan 01	\$10,600
4	1	2	2	1		C0003	Carl	mathematician		2002 Jan 01	\$10,200
5										2003 Jan 01	\$8,600
6										2004 Jan 01	\$10,400
7										2005 Jan 01	\$10,600
8					y=10	6	12	18		2006 Jan 01	\$10,900
9					y=8	5	10	15		2007 Jan 01	\$12,300
10					y=6	4	8	12		2008 Jan 01	\$12,600
11					y=4	3	6	9		2009 Jan 01	\$9,000
12					y=2	2	4	6		2010 Jan 01	\$10,600
13					y=0	1	2	3		2011 Jan 01	\$11,700
14						x=0	x=5	x=10			



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

Slide - 11



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

Slide - 12



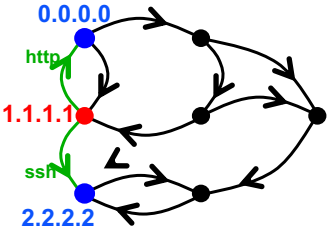
Databases

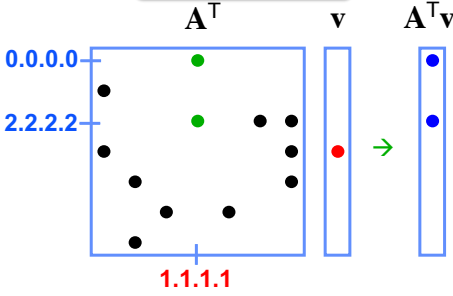
SQL

NoSQL

NewSQL

	src	link	dest
001	1.1.1.1	http	0.0.0.0
002	0.0.0.0	udp	1.1.1.1
003	1.1.1.1	ssh	2.2.2.2







Operation: finding 1.1.1.1's nearest neighbors


- 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


Slide - 13 D4M = Dynamic Distributed Dimensional Data Model (d4m.mit.edu)
Mathematics of Big Data, Kepner & Jananathan, MIT Press 2018





Languages, Libraries, AI Packages











- Lots of machine learning software
- Designed for tabular data
- Jupyter interactive portal interface
 - Similar to Mathematica notebooks

Slide - 14



Hierarchical Data: XML, JSON, ...

Hierarchical JSON

```
{
  "conformsTo": "https://project-open-
  data.cio.gov/v1.1/schema",
  "describedBy": "https://project-
  open-data.cio.gov/v1.1/schema/catalog.json",
  "@context":
  "https://project-open-
  data.cio.gov/v1.1/schema/catalog.jsonld",
  "@type":
  "dcat:Catalog",
  "dataset": [
    {
      "@type": "dcat:Dataset",
      "title":
      "SBA IT Policy Archive",
      "description": "A list of all public
      documents relating to SBA IT policy.",
      "modified": "2015-08-
      31",
      "accessLevel": "public",
      "identifier": "SBA-OCIO-2015-
      08-002",
      "landingPage": "https://www.sba.gov/about-
      sba/sba-performance/open-government/digital-sba/digital-
      strategy/it-policy-archive",
      "license":
      "http://www.usa.gov/publicdomain/label/1.0",
      "publisher":
      {
        "@type": "org:Organization",
        "name": "U.S. Small Business
        Administration",
        "distribution": [
          {
            "@type":
            "dcat:Distribution",
            "accessURL": ...
          }
        ]
      }
    }
  ]
}
```

Sparse Table

	dataset/accessLevel	dataset/keyword	dataset/modified	dataset/description	dataset/distribution/title	dataset/identifier
1001/0001	public	Agency IT Policy Archi	8/31/15	A list of all public documents relating to SBA IT policy.	SBA-OCIO-2015-08-002	SBA IT Policy Archive
1001/0001/0001					SBA IT Policy Archive Policy Zip	
1001/0002	public	Bureau IT Leadership	8/15/15	A directory of all SBA officials with the title of CIO or d/s	SBA-OCIO-2015-08-001	Bureau IT Leadership Directory
1001/0002/0001					Bureau IT Leadership Directory List	
1001/0002/0002	public	CIO role on program	8/30/15	Information about SBA CIO AAs membership in govern	SBA-OCIO-2015-08-003	CIO Governance Board
1001/0003/0001					CIO Governance Board Membership List	
1001/0004	non-public	Loan	6/30/17	Not available to the public. Additional information is av	SBA-OCA-2017-11-001	Orig@pass
1001/0004/0001					ExtractOrigination	
1001/0005	non-public	Loan	6/30/17	Not available to the public. Additional information is av	SBA-OCA-2017-11-002	ExtractServicing
1001/0005/0001					Originate3	
1001/0006	non-public	Loan	6/30/17	Not available to the public. Additional information is av	SBA-OCA-2017-11-003	OriginateStatus
1001/0006/0001					Orig@pass	
1001/0007	non-public	Loan	6/30/17	Not available to the public. Additional information is av	SBA-OCA-2017-11-004	OrigScore
1001/0007/0001					OrigUpdate	
1001/0008	non-public	Loan	6/30/17	Not available to the public. Additional information is av	SBA-OCA-2017-11-005	OrigUpdate
1001/0008/0001					OrigUpdate	
1001/0009	non-public	Loan	6/30/17	Not available to the public. Additional information is av	SBA-OCA-2017-11-006	OrigUpdate
1001/0009/0001					OrigUpdate	
1001/0010	non-public	Loan	6/30/17	Not available to the public. Additional information is av	SBA-OCA-2017-11-007	OrigUpdate
1001/0010/0001					OrigUpdate	
1001/0011	non-public	Loan	6/30/17	Not available to the public. Additional information is av	SBA-OCA-2017-11-008	OrigUpdate

Source: public domain

- Hierarchical data is an increasingly important form of “big data”
- Can be converted to sparse tables

Slide - 15

<https://catalog.data.gov/dataset/sba-public-datasets>



Tabular Terminology

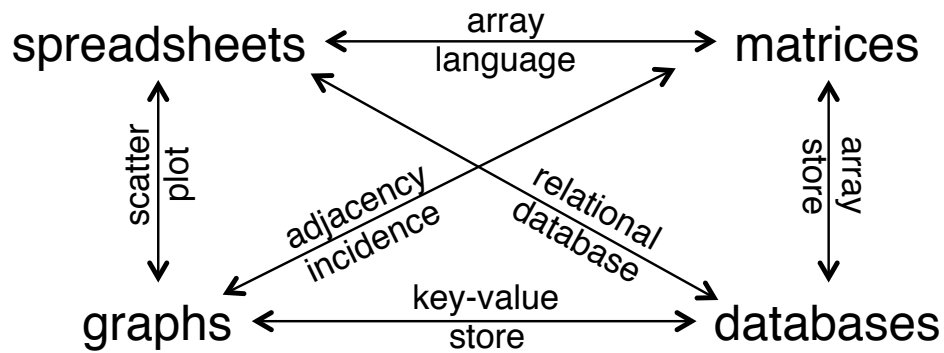
Context	Table	Row Label	Row	Column Label	Column	Value	Math
Spreadsheet	sheet	row name; number	row	column name; letter	column	cell	addition-multiplication
SQL	table	sequence ID	record	field	column	value	union-intersection
Relational Algebra	set	ordinal	relation; tuple	index	set	entry	union-intersection
NoSQL	table	key	tuple	column name		value	or-and
GraphBLAS	graph	source vertex	out edges	destination vertex	in edges	count; weight	semiring
Linear Algebra	matrix	row index	row vector	column index	column vector	value	addition-multiplication
Neural Network	network	feature; input neuron	feature vector; forward connections	category; output neuron	category score; backward connections	weight	addition-multiplication; max-addition
Data Frame	frame	index	row	index	column	value	various
Associative Array	array	row label; row key	array	column label; column key	array	value	semiring

- Tables have been implemented many ways
- Lots of different names for the same concepts

Slide - 16



Tabular Data is a Natural Interchange Format



Slide - 17


Mathematics of Big Data, Kepner & Jananathan, MIT Press 2018



Outline

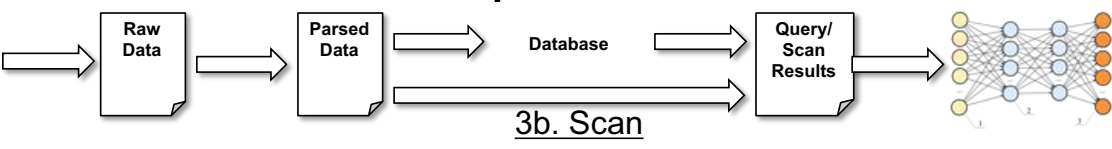
- Introduction
- Tabular Data
- **Files and Folders**
- Using and Sharing
- Summary

Slide - 18



Files and Folders


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



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
- <https://vijayg.mit.edu/sites/default/files/documents/DataforAIReadiness.pdf>

Slide - 19



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

Slide - 20



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**

Slide - 21



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**

```

Pipeline_Template/
  README.txt # This file.
  Step0_Raw/
    README.txt # Step 0 instructions for users.
    code/ # Code for getting raw data from source.
    data/ # Raw data
    dataFileList.txt # List of all raw data filenames.
  Step1_Parse/
    README.txt # Step 1 instructions for users.
    code/ # Code for parsing raw data.
    data/ # Parsed data.
    dataFileList.txt # List of all parsed data filenames.
  Step2_Ingest/
    README.txt # Step 2 instructions for users.
    code/ # Code for ingesting parsed data into a database.
    data/ # Database ingest logs.
    dataFileList.txt # List of all log filenames.
  Step3_Query/
    README.txt # Step 3 instructions for users.
    code/ # Code for querying data from a database.
    data/ # Results of database queries.
    dataFileList.txt # List of all query data filenames.
  Step4_Analysis/
    README.txt # Step 4 instructions for users.
    code/ # Code for analyzing data.
    data/ # Results of analysis.
    dataFileList.txt # List of all analysis data filenames.
  Step5_Viz/
    README.txt # Step 5 instructions for users.
    code/ # Code for visualizing data.
    data/ # Results of visualizations.
    dataFileList.txt # List of all visualizing data filenames.
  
```

Slide - 22



Outline

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

Slide - 23



Co-Design Using and Sharing

- 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

Slide - 24



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

Slide - 25

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

Slide - 26



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**

Slide - 27



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 accidentally 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.

Slide - 28



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.

Slide - 29



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.

Slide - 30



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

Slide - 31

MIT OpenCourseWare
<https://ocw.mit.edu/>

RES.LL-005 Mathematics of Big Data and Machine Learning
IAP 2020

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.