Andrew Plum

CS 479

1/11/24

Reminders

Questions

Notes

1/11/24

* Working with big data
  + Big data
    - If it can not all be stored on a regular computer
      * Then it is big data
* Definitions
  + Data
    - Representation of facts that can be operated upon by a computer
      * Typically the results of measurements, computations, or observations and can be the basis of graphs images and models
      * Example
        + Classes offered in Spring 2024 and the students taking them
    - Encoding of variables
    - Plural of “datum”
    - Information
      * The meaning of the data as interpreted by human beings
      * The representation (of facts? data?) in a form that lends itself to human use
      * The message of the dataset
    - Knowledge is the expertise or familiarity of a subject
    - Metadata
      * Data about data
      * Label to a data package
      * Example
        + He said the label on a cleaning bottle is metadata and I think the data would be whats in the bottle itself
    - Data documentation
    - Data life-cycle elements
      * Acquisition
      * Curation
      * Preservation
      * Look at MIT DDI Alliance Life Cycle
* Examples
  + Rock sample:
    - Data – weight, composition, hardness, shape, size
    - Information – type of the rock as interpreted from the data
    - Knowledge – evidence of geologic activity
    - Metadata – location, time, collector of data collection
    - Documentation – published lab report …
  + Weather
    - Data – wind speed and direction, temperature, ..
    - Information – weather map with contours and features
    - Knowledge – high pressure system, stable weather
    - Metadata – type of radar, sensor, use of model
* How do we get data
  + Analog vs digital ‘data’
    - Analog attempts to be continuous
    - Digital data is discrete
  + Example
    - Thermometer, watch, calculator
    - ChatGPT
* Producers and Consumers of data
  + Producers
    - Quality Control
    - Fitness for Purpose
    - Trustee
  + Consumers
    - Quality Assessment
    - Fitness for Use
    - Trustor
* Kinds of database models
  + Data Modeling in the databases
  + 3 kinds
    - Conceptual
      * What do you want
        + Initial concepts
        + Questions and answers
        + Grant info
    - Logical
      * How the data is structured
    - Physical
      * The data itself
* How to find good data
  + Dataset search engine

1/16/2024

* Dynamic datasets
  + It changes
  + You can come back to check your model to see how accurate it is
    - Readjust if its off
    - And good if its correct
* Mineral species
  + The prediction curve
  + The curve could be off in the future but so far it is correct
* Management of data
  + Creation of logical collections
    - From the physical entity to the logical collection of attributes
  + Physical data handling
    - Think of it like spreadsheet like excel
  + Interoperability
  + Security support
  + Data ownership
  + Metadata collection, management, access
  + Persistence
  + Discovery
  + Data dissemination and publication

1/18/2024

* Physical Data Handling
  + Data formats
    - .txt file
      * Used especially to store words for natural language processing
  + Naming conventions
    - Can exist to help organize the data files
* Interoperability Support
  + Interoperability meaning
    - Smallest number of things to agree on so that you do not need to agree on anything else
* Security
* Data ownership
  + Rights and policies
* Metadata
  + If you download the data from a data portal, there may be metadata to records
    - Record the metadata and the standards
* Persistence
* Discovery
* Dissemination
  + GitHub can host your dataset for people to download
    - Should list collaborators
* Data formats
  + Relevant for assignment 1
  + ASCII, UTF-8, ISO 8859-1
  + Self-describing formats
  + Table-driven
  + Markup languages and other web-based
  + Database
    - MongoDB
    - MySQL
    - Posgres
    - Many more
  + Graphs
    - Entity and entity relationships
  + Unstructured
    - Media, imaging, audio, sensor data, text data
  + Discussion… because this is part of your assignment
* Curation
  + Consider the organization and presentation of data
  + Documenting
* Preservation
  + Archive data
  + May involve unconventional steps

1/23/2023

* Analyzing the Earth at night
  + Some places are more developed than others
  + USA
    - USA has a large oil field in South Dakota
    - Can see developed areas of the US
  + Australia has large wildfires
* Data vs Metadata
  + Metadata is just the description about the data
  + A screenshot of a document

    Description automatically generated
* Dublin Core Metadata Elements
  + Don’t need to choose all of them if you are creating the metadata yourself
  + Format
    - This is the file format
  + Rights
    - CC
      * CC 0
        + No requirements
      * CC by
        + Requires a citation
      * CC as
        + Can create a commercial license?
    - NC
      * Noncommercial
* Data Formats
  + ASCII uses spaces for data column categorization
    - A number of numbers on a white background

      Description automatically generated with medium confidence
  + Self describing
    - A screenshot of a data

      Description automatically generated
  + Ipynb
    - Is actually JSON
  + CSV
    - Comma separated variables

1/25/2024

* Schema.org
  + Initiative by Google, Yahoo, and Bing to classify metadata of anything on the internet
* Metadata
  + Dublin Core metadata elements
    - Relationship
      * A related resource
    - If you are unsure about any of those metadata elements, you can go to this website:
      * [www.dublincore.org/specifications](http://www.dublincore.org/specifications)
* Analog computer
  + Computing by physical means rather than digital
* What is common about data and metadata
  + Both are data and can be stored in any digital formats

2/6/2023

* Internet vs Web
  + Internet is the fundamental framework
    - Internet is a network of computers
  + One of the many applications built on top of the internet
    - Web is the software that lets use access the internet
  + Internet uses TCP/IP
    - Transmission control protocol

2/8/2024

* Dbpedia
* SPARQL
* Ontology
  + The study of the nature of existence (being)
    - Not phenomonology
  + For semantic web purpose
  + An ontology spectrum
    - Catalog
      * Alphanumerical list
    - Glossary
      * Alphanumerical list
    - Taxonomy
      * Supergroup-subgroup
    - Thesaurus
      * Hierarchy
    - Conceptual Schema
      * Formal superclass-subclass
    - Formal assertions
      * Disjoint subclass
        + Ex: three classes – people, men , and women where men and woman are subclasses of people but also disjoint to one another
      * Transitive properties
      * Symmetric properties
    - The further down is more enriching semantic expressions
    - Interesting field at the cross section of artificial intelligence and philosophy:
      * Knowledge representation and knowledge reasoning
* Semantic web vision

3/26/2024

* Google collab
* Data lake
* Data warehouse
* Interesting data mining techniques
  + Outlier analysis
  + Sequence and trend and evolution analysis
  + Mining data streams
  + Graph mining
    - Useful for finding trends in things that are in some kind of network
      * Like a social network
  + Information network analysis
  + Web mining
* Quick summary of applications of data mining
  + Web page analysis: classification, clustering, ranking
  + Collaborative analysis & recommendersystems
  + Basket data analysis to targeted marketing
  + Biological and medical data analysis
  + Data mining and software engineering
  + Data mining and text analysis
  + Data mining and social and information network analysis
  + Built-in (invisible data mining) functions in Google, MS, Yahoo!, LinkedIn, Facebook, …
  + Major dedicated data mining systems/tools
  + SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools)