Data Science
Data Engineering

Keith T. Weber, GISP
ISU GIS Director
GIS Training and Research Center

Data Science

• A growing field
  – Interdisciplinary
  – Seeks to extract information from data
  – Born from Big Data/unstructured data
  – NOTE: 80% of data is considered unstructured data

• Spatial Data Science
  – A more specialized sub-field of data science

Structured vs. Unstructured Data
Structured vs. Unstructured Data (cont’d)

• This semester, we have focused on structured data
• How can we work with unstructured data, especially in GIS
  – This is task for Data Engineering tools and techniques

First, Take out the Trash

• Some data is trash!
  – How to identify trash
• ROT
  – Redundant
  – Obsolete
  – Trivial

Second, Use these **Keys** to Unlocking Unstructured Data

• It is not as *unstructured* as you may think
  – A Twitter feed is not just a collection of words
  – The digital file behind the scenes contains much more
Tweets (etc.) are Machine-Readable

- Using Esri's GeoEvent server, our server can read Tweets and filter them
  - Filtering is the second key to unlocking unstructured data
  - When written to a database, these data become structured

Spatial Data from Email

- There are approximately 300 billion emails sent each day
- Each email (without attachments) is approximately 500 bytes in size
  - 150 trillion bytes of data storage
  - 146 billion KB
  - 136 TB
- In one year, harvesting and storing all emails would require:
  - 0.05 EB

All Emails can be stored for a very long time

- It is estimated the NSA data centers can store 1 (or more) Yottabytes of data
  - Thus, a 12 EB data center can store all emails for 250 years, OR
  - Store all emails for 25 years AND
  - All social media posts for 25 years, etc.
Finding the Needle in the Haystack

- Finding the needle of information in a haystack of data
  - HINT: go back to KEY #1
  - Learn the structure of the data
  - Develop and further mature/evolve AI search tools

Structure of an Email

- DATA is a keyword
  - Extract all text following DATA
  - FROM field (text data type)
  - TO field (text data type)
  - SUBJECT (text data type)
  - DATE (date data type)
  - Next is the body of text (text data type or CLOB data type)

I Don't See Any Spatial Data!

- Do you use a signature line?
- Do you mention places in your emails?
  - Subject line of email body
  - “Boise visit”
  - “Vacation photos from Italy”
  - Classify Text Using Deep Learning
- Do you send emails from your smartphone?
Spatial Data from Web Browsers and Smart Phones

- For Example, Google (Android smart phones and our email/browser system) collects and stores a lot of data about you.

Let's look at one of these JSON files

- Look for keywords or something that can be used to parse these data.

BRAINSTORM...

What is PlaceID?

- [https://developers.google.com/maps/documentation/javascript/examples/places-placeid-finder](https://developers.google.com/maps/documentation/javascript/examples/places-placeid-finder)
How Accurate are these Locations?

- latitudeE7: 428614159, longitudeE7: -1124327469
- In other words:
  - 42.8614159°
  - -112.4327469°
- Now its easy!
  - Any programmers out there?
  - MS Excel enthusiasts?
  - How about ArcGIS Pro?

Google Location Data in ArcGIS Pro

Location Precision

- Everyday, everyone’s location is tracked to +/- 0.02 meters with their smartphone (2 cm)
- Limitations:
  - I did not find data stored for all 24 hrs. in a day
  - BUT only when I had LOCATIONS turned on
- PS- These data are stored for a long time (I got my first Android smart phone in 2013)
Information from Imagery

Remote Sensing and The IoT
- IoT is the Internet of Things
- Object/Feature Extraction with ArcGIS Pro
  - Detect Objects Using Deep Learning
- Map buildings
- Identify people

Facial Recognition
- In 2016, Facebook boasted its facial recognition had 98% accuracy
Advanced Data Engineering Brings the Need for Professional Ethics

Advantages of harvesting data
• Brainstorm

Disadvantages of harvesting data
• Brainstorm

Key Concepts

• Gleaning information from unstructured data is an emphasis area in data science
• New, powerful tools leveraging artificial intelligence algorithms are emerging and maturing
• Knowing WHERE makes data much more valuable
• Data is considered an asset, information is a valuable asset
• High spatial and temporally resolved data requires care to avoid unethical use of these tools/resulting data

Professional Hints and Tips

• Building a great resume
  – Promote your skillset
  – Introductory sentence
  – Don’t misrepresent yourself
  – Aesthetics