

## Data Science and Data Engineering

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

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State UNIVERSITY

---

---

---

---

---

---

---

---

## 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

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State UNIVERSITY

---

---

---

---

---

---

---

---

## Structured vs. Unstructured Data

### Structured Data

Can be displayed in rows, spreadsheet, relational database

Highly-organized data that is easy to analyze

Examples: customer's phone numbers, names, zip codes

VS

### Unstructured Data

Cannot be displayed in rows, spreadsheet, relational database

Not-organized data that is hard to collect and analyze

Examples: emails, video files, images, data from social media as Facebook, LinkedIn

Pocatello | Idaho

intelispat.com

Idaho State UNIVERSITY

---

---

---

---

---

---

---

---

## Structured vs. Unstructured Data (cont'd)

- To date, we have focused on structured data
- Can we use unstructured data in GIS?
  - This is a task for Data Engineering tools and techniques

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

---

---

---

## First Step, Take out the Trash

- Some data is really is trash!
  - How to identify trash
- ROT
  - Redundant
  - Obsolete
  - Trivial



Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

---

---

---

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

- Understand, it is not as *unstructured* as you may think
  - A social media feed is not just a collection of words
- The digital file behind the scenes contains much more
  - Metadata (author and date)
  - Message Content (text, recipient, hyperlink, image, hashtag)
  - Interactions (reposts, favorites, replies)

The first day of the new semester is upon us. Welcome back everyone. We are back to our normal hours M-F 8a-5p. If you need any assistance feel free to stop by, call (ext. 4444), or email us. @IdahoStateU @IdahoStateUCoSE

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

---

---

---

## Posts are Machine-Readable

- Using **Esri's GeoEvent** server, a server can read X posts *and* filter them
  - **Filtering** is the second key to unlocking unstructured data
  - In other words, identify data of interest (e.g., Posts with a hashtag #GISDay)
  - When written to a database, these data become structured

Pocatello | Idaho Falls | Meridian | Twin Falls

**Idaho State**  
UNIVERSITY

## Structured Spatial Data from Email

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

Pocatello | Idaho Falls | Meridian | Twin Falls

**Idaho State**  
UNIVERSITY

## All Emails can be stored for a very long time

- For example, it is estimated just one data center can store 1 (or more) Yottabytes of data
  - Thus, a 12 EB data center could store all emails for 250 years, OR
  - Store all emails for 25 years AND
  - All social media posts for 25 years, etc.

Measure	Acronym	Characters	Relationship
Byte	B	1	8 bits
Kilobyte	KB	1,024	1,024 B
Megabyte	MB	1,048,576	1,024 KB
Gigabyte	GB	1,073,741,824	1,024 MB
Terabyte	TB	1,099,511,627,776	1,024 GB
Petabyte	PB	1,125,899,906,842,620	1,024 TB
Exabyte	EB	1,152,921,504,606,850,000	1,024 PB
Zetabyte	ZB	1,180,591,620,717,410,000,000	1,024 EB
Yottabyte	YB	1,208,925,819,614,630,000,000,000	1,024 ZB

Pocatello | Idaho Falls | Meridian | Twin Falls

**Idaho State**  
UNIVERSITY

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

# Spatial Data from Web Browsers and Smart Phones

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

The screenshot shows the 'Location History' settings on an Android phone. The 'Location History' toggle is turned on. Below the toggle, there is a list of location history items. The list has three columns: 'Name', 'Location History', and 'Date'. The items are as follows:

Name	Location History	Date
Android Device Configuration Service		
Cloud Print		
Currents Circles		
Currents Stream		
Google Account		
Google Play		
Google Photos		
Location History		
Maps		
My Activity		
Search		
Voice		
Semantic Location History		
Recorder.jam		2013
Settings.jam		2014
		2015
		2016
		2017
		2018
		2019

At the bottom of the screen, there is a link to 'archive\_browser.html'.

---

---

---

---

---

---

[illegible]

---

---

---

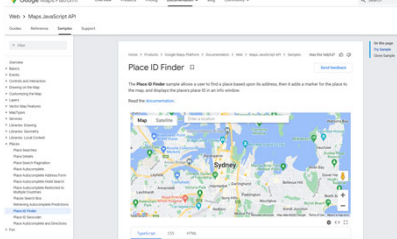
---

---

---

## What is PlaceID?

- <https://developers.google.com/maps/documentation/javascript/examples/places-placeid-finder>



Pocatello | Idaho Falls | Meridian | Twin Falls

**Idaho State**  
UNIVERSITY

---

---

---

---

---

---

---

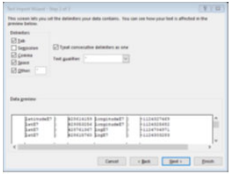
---

---

---

## How Accurate are these Locations?

- latitudeE7: 428614159, longitudeE7: -1124327469
- In other words:
  - 42.8761416°
  - 112.4327469°
- Now its easy!
  - Any programmers out there?
  - MS Excel enthusiasts?
  - How about ArcGIS Pro?



Pocatello | Idaho Falls | Meridian | Twin Falls

**Idaho State**  
UNIVERSITY

---

---

---

---

---

---

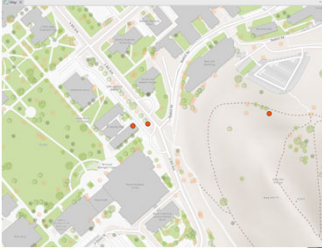
---

---

---

---

## Google Location Data in ArcGIS Pro



Pocatello | Idaho Falls | Meridian | Twin Falls

**Idaho State**  
UNIVERSITY

---

---

---

---

---

---

---

---

---

---

## Location Precision

- Everyday, everyone's location can be tracked (well, your smartphone's location is tracked)
- Limitations:
  - I did not find data stored for a full 24 hrs.
  - 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)

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

---

---

---

## Information from Surveillance Cameras



[Photo courtesy of Wikimedia Commons. Author is licensed under CC BY.](#)

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

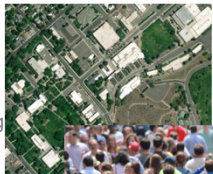
---

---

---

## 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
- Gait analysis



Pocatello | Idaho Falls | Meridian | Twin Falls

---

---

---

---

---

---

---

---

## Facial Recognition

- In 2016, Facebook boasted its facial recognition had 98% accuracy

98 percent of the time



Facebook, according to the company, is able to accurately identify a person **98 percent of the time**. Compare that with the FBI's facial recognition technology, Next Generation Identification, which according to the FBI, identifies the correct person in the list of the top 50 people only 85 percent of the time.

Facebook's Facial Recognition Software Is Different From The FBI's. [www.rpr.org/sections/alttech/considered/2016/05/18/477819617/facebook-fa...](https://www.rpr.org/sections/alttech/considered/2016/05/18/477819617/facebook-fa...)

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

---

---

---

## Advanced Data Engineering Heightens the Need for Professional Ethics

### Advantages of harvesting data

- Brainstorm

### Disadvantages of harvesting data

- Brainstorm

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

---

---

---



## 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

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

---

---

---



## Professional Hints and Tips

- Building a great resume
  - Promote your skillset
  - Introductory sentence
  - Don't misrepresent yourself
  - Aesthetics

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

---

---

## Questions/Discussion?



Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State  
UNIVERSITY

---

---

---

---

---

---

---