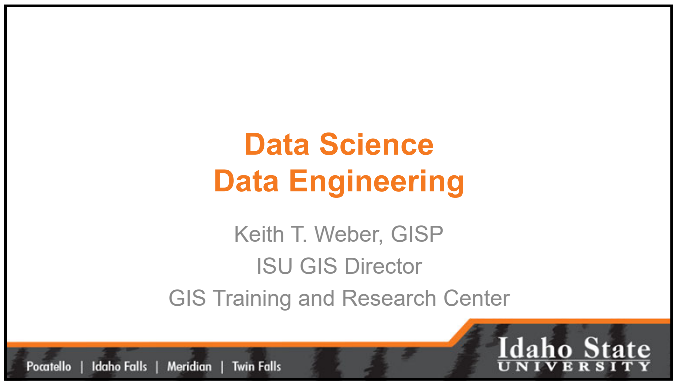
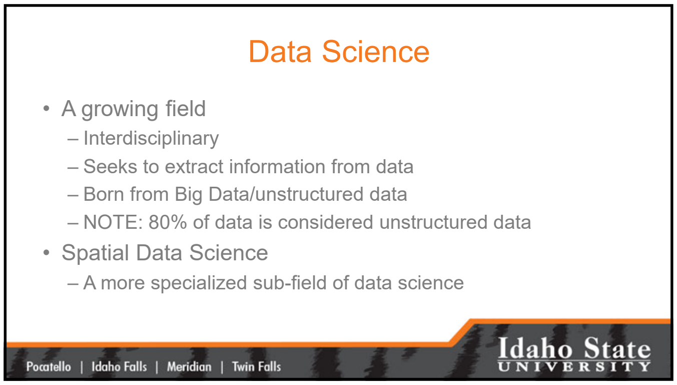


Data Science
Data Engineering

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

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State University

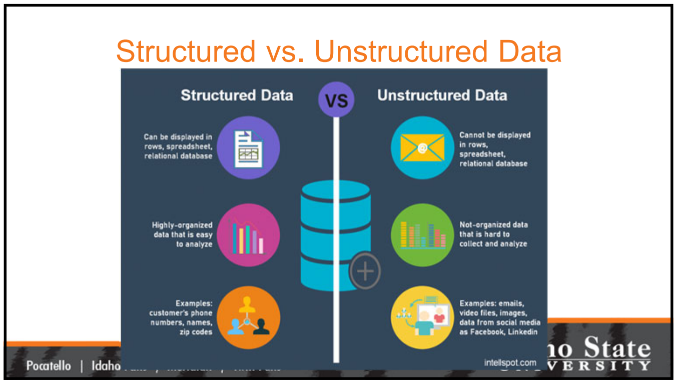


Data Science

- 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 VS Unstructured 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

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

intellspot.com

Pocatello | Idaho

Structured vs. Unstructured Data (cont’d)

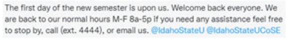
- This semester, we have focused on structured data
- How can we work with unstructured data, especially in GIS
  - This is task for Data Engineering tools and techniques

First, Take out the Trash

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

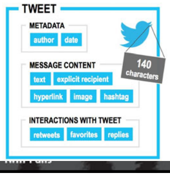
Second, Use these **Keys** to Unlocking Unstructured Data

- It is not as *unstructured* as you may think
  - A Twitter feed is not just a collection of words



The 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

  - The digital file behind the scenes contains much more



Tweets (etc.) **are** Machine-Readable

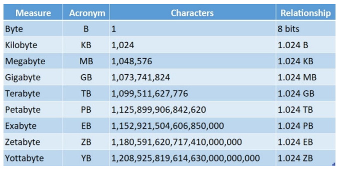
- Using **Esri's GeoEvent** server, our server can read Tweets and filter them
  - Filtering is the second key to unlocking unstructured data
  - When written to a database, these data become structured

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.

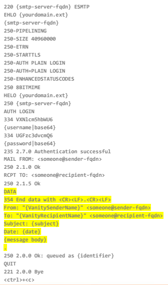


Measure	Acronym	Characters	Relationship
Byte	B	1	8 bits
Kilobyte	KB	1,024	1.024 B
Megabyte	MB	1,048,576	1.024 KB
Gigabyte	GB	1,073,741,824	1.024 MB
Terabyte	TB	1,099,511,627,776	1.024 GB
Petabyte	PB	1,125,899,906,842,620	1.024 TB
Exabyte	EB	1,152,921,504,606,850,000	1.024 PB
Zetabyte	ZB	1,180,591,620,717,410,000,000	1.024 EB
Yottabyte	YB	1,208,925,819,614,630,000,000,000	1.024 ZB

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

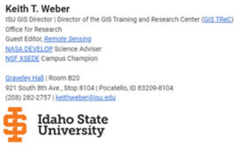


```
220 {smtp-server-fqdn} ESMTP
EHLO {yourdomain.ext}
{smtp-server-fqdn}
250-PIPELINING
250-SIZE 40960000
250-ETRN
250-STARTTLS
250-AUTH PLAIN LOGIN
250-AUTH=PLAIN LOGIN
250-ENHANCEDSTATUSCODES
250 8BITMIME
HELO {yourdomain.ext}
250 {smtp-server-fqdn}
AUTH LOGIN
334 VXNlcm5hbWU6
{username|base64}
334 UGFzc3dvcmQ6
{password|base64}
235 2.7.0 Authentication successful
MAIL FROM: <someone@sender-fqdn>
250 2.1.0 Ok
RCPT TO: <someone@recipient-fqdn>
250 2.1.5 Ok
DATA
354 End data with <CR><LF>.<CR><LF>
From: "{VanitySenderName}" <someone@sender-fqdn>
To: "{VanityRecipientName}" <someone@recipient-fqdn>
Subject: {subject}
Date: {date}
{message body}
.
250 2.0.0 Ok: queued as {identifier}
QUIT
221 2.0.0 Bye
<ctrl>+<c>
```

- 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 smart phone?



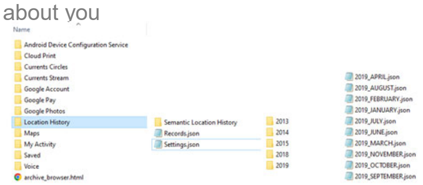
Keith T. Weber
ISU GIS Director | Director of the GIS Training and Research Center (GIS TReC)
Office for Research
Guest Editor, Remote Sensing
NASA DEVELOP Science Adviser
NSF XSEDE Campus Champion

Graveley Hall | Room B20
921 South 8th Ave., Stop 8104 | Pocatello, ID 83209-8104
(208) 282-2757 | keithweber@isu.edu

Idaho State University

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



Name

Android Device Configuration Service
Cloud Print
Currents Circles
Currents Stream
Google Account
Google Pay
Google Photos
Location History
Maps
My Activity
Saved
Voice
archive_browser.html

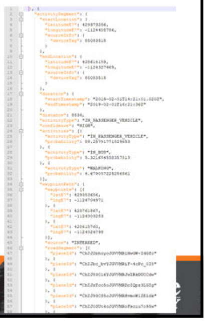
Semantic Location History
Records.json
Settings.json

2013
2014
2015
2018
2019

2019_APRIL.json
2019_AUGUST.json
2019_FEBRUARY.json
2019_JANUARY.json
2019_JULY.json
2019_JUNE.json
2019_MARCH.json
2019_NOVEMBER.json
2019_OCTOBER.json
2019_SEPTEMBER.json

Let's look at one of these JSON files

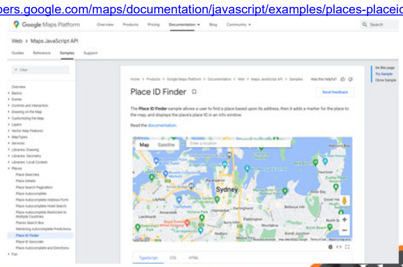
- Look for keywords or something that can be used to parse these data
- BRAINSTORM…



```
}, {
  "activitySegment": {
    "startLocation": {
      "latitudeE7": 429973256,
      "longitudeE7": -1124600786,
      "sourceInfo": {
        "deviceTag": 85083518
      }
    },
    "endLocation": {
      "latitudeE7": 428614159,
      "longitudeE7": -1124327469,
      "sourceInfo": {
        "deviceTag": 85083518
      }
    },
    "duration": {
      "startTimestamp": "2019-02-01T14:21:01.020Z",
      "endTimestamp": "2019-02-01T14:21:34Z"
    },
    "distance": 8694,
    "activityType": "IN_PASSENGER_VEHICLE",
    "confidence": "HIGH",
    "activities": [{
      "activityType": "IN_PASSENGER_VEHICLE",
      "probability": 89.25791771829653
    }, {
      "activityType": "IN_BUS",
      "probability": 5.321454550357513
    }, {
      "activityType": "WALKING",
      "probability": 4.479057225286861
    }],
    "waypointPath": {
      "waypoints": [{
        "latE7": 429353656,
        "lngE7": -1124704971
      }, {
        "latE7": 428761847,
        "lngE7": -1124303253
      }, {
        "latE7": 428615760,
        "lngE7": -1124324798
      }],
      "source": "INFERRED",
      "roadSegment": [{
        "placeId": "ChIJ2bbrypoJUVVRR1HwGW-I40fc"
      }, {
        "placeId": "ChIJbc_bvYJUVVRRiF-4cPc_0I8"
      }, {
        "placeId": "ChIJ93C14YJUVVRRJwIRkDOOOdw"
      }, {
        "placeId": "ChIJxToo5oJUVVRR2c0Qps3L0Zg"
      }, {
        "placeId": "ChIJ93C55oJUVVRRRwmcW12E1dx"
      }, {
        "placeId": "ChIJ0UUt4oJUVVRRoFaciu7c86w"
      }, {
```

What is PlaceID?

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



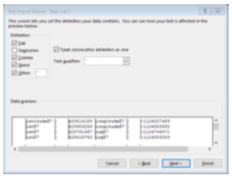
Place ID Finder

The **Place ID Finder** sample allows a user to find a place based upon its address, then it adds a marker for the place to the map, and displays the place's place ID in an info window.

Read the documentation.

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?



Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

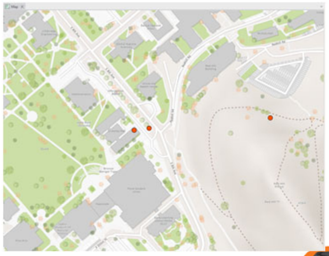
☑ Tab
□ Semicolon
☑ Comma
☑ Space
☑ Other:

☑ Treat consecutive delimiters as one

Text qualifier:

Data preview

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

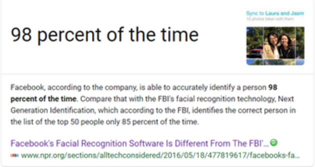
This Photo by Unknown Author is licensed under CC BY

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

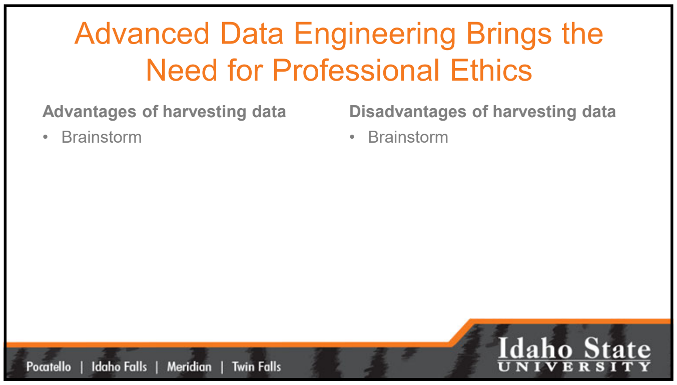
- In 2016, Facebook boasted its facial recognition had 98% accuracy



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'...
www.npr.org/sections/alltechconsidered/2016/05/18/477819617/facebooks-fa...



Advanced Data Engineering Brings the Need for Professional Ethics

Advantages of harvesting data

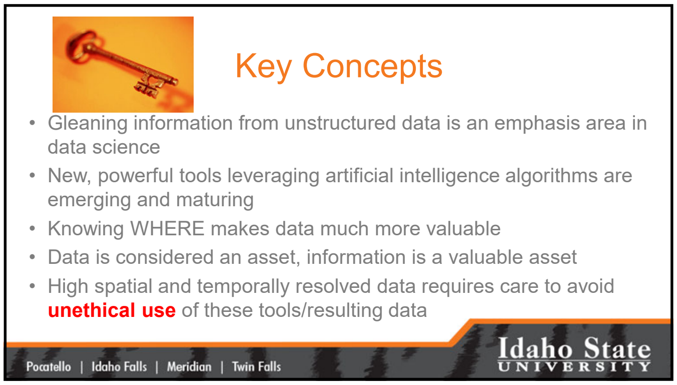
- Brainstorm

Disadvantages of harvesting data

- Brainstorm

Pocatello | Idaho Falls | Meridian | Twin Falls

Idaho State UNIVERSITY

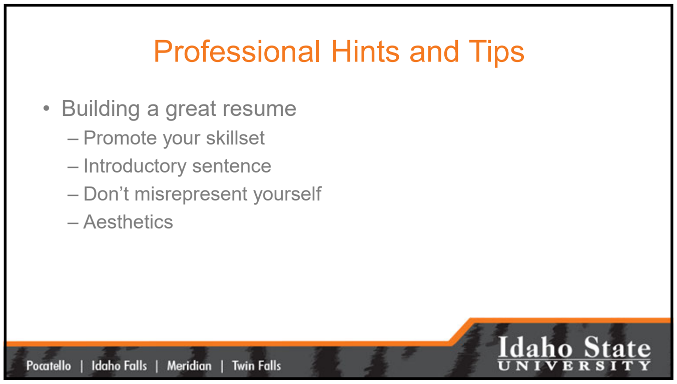


Key Concepts

- 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