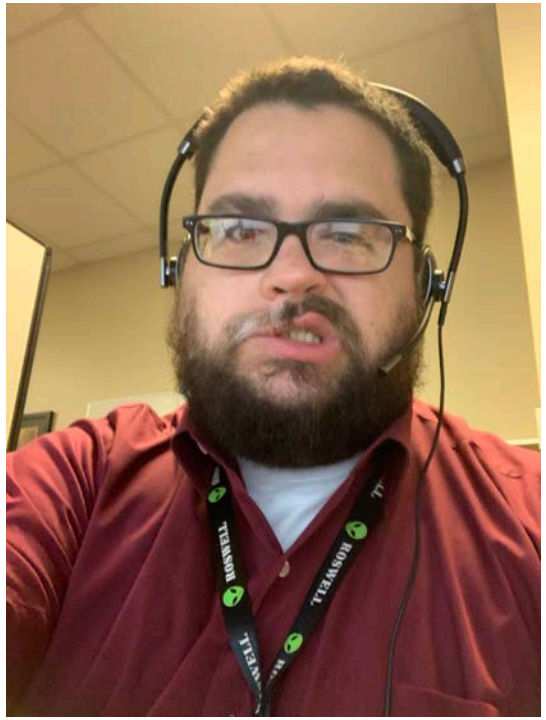
**Candidate: Neil S. Rose, GISP**

**Advisor: Ryan Baxter**

# Understanding Faulty Data

## Methods to Capture, Report, and Score Data Cleanliness

# Who am I?



## Neil S. Rose, GISP

GIS Manager, City of New Braunfels, TX

- BA Sociology
- BA Applied Geography
- Grad Cert, GIS
- Grad Cert, Geospatial Programming and Web Map Development
- 8 years in local government GIS
- Employs a high level of python scripting
- Automate, automate, automate
- Professional interest in data cleansing automation

# The Agenda

## Introduction and Background

What is data? Why is it important? How does it live? What makes data good or bad?

## Goals, Objectives, Methodologies

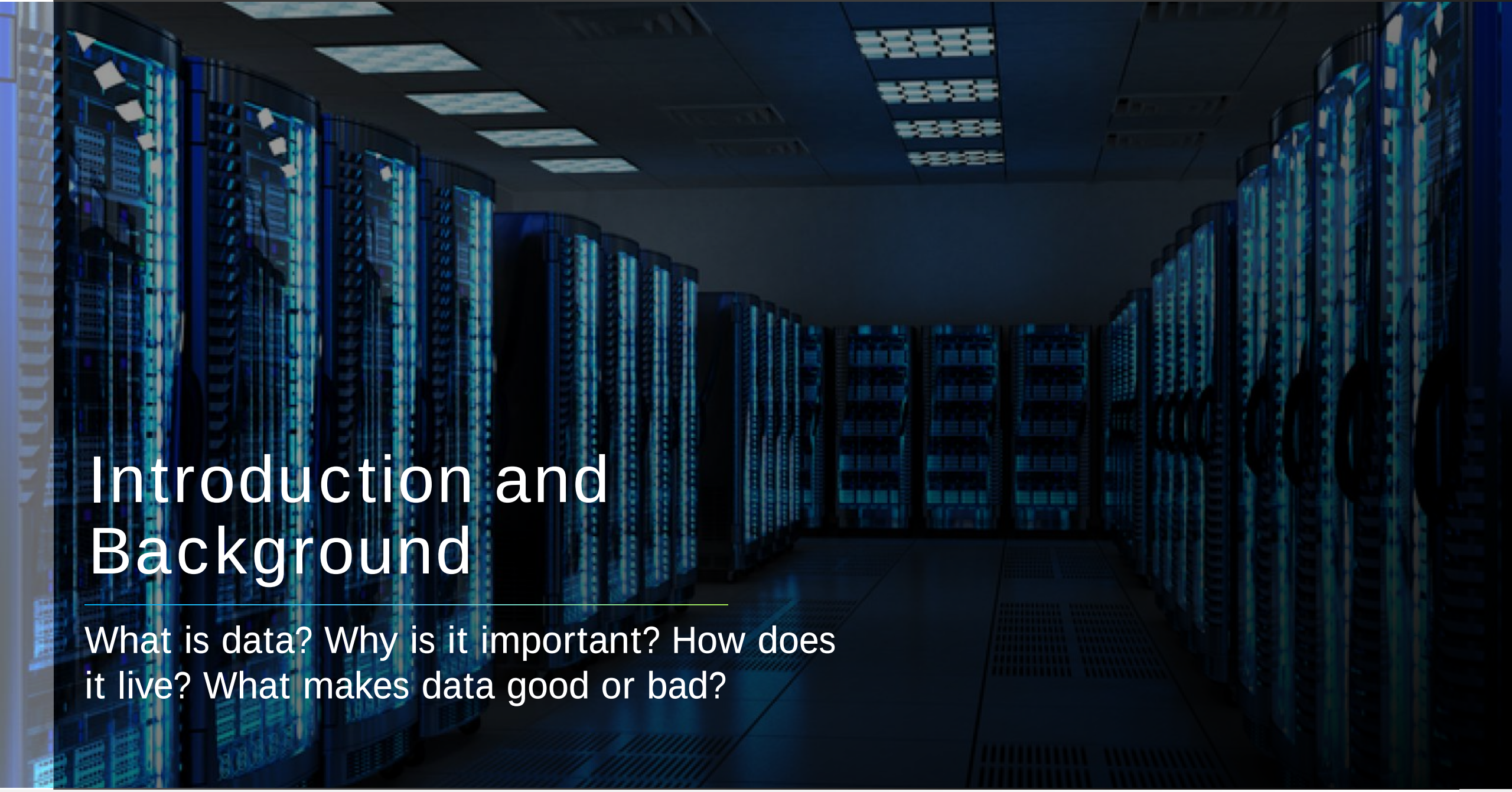
Identify, report, and score data cleanliness using python solutions

## Anticipated Results

What will work? What won't?

## Timeline

Project completion and presentation



# Introduction and Background

What is data? Why is it important? How does it live? What makes data good or bad?

# What is "data"?

| | |
|---|---|
| Data quantifies or qualifies a phenomena | • Temperature, location, cost, distance, color |
| Data != Information | • Information is derived from the analysis and interpretation of data |
| GIS data isn't special, only spatial | • Vector and raster data structures |
| Vector consists of points, lines, and polygons | • Acquired by heads up digitization or device capture |
| Raster consists of a grid of values | • Represents a continuous view of real-world phenomena |
| Attributes contextualize the spatial | • Non-spatial data that describes the spatial feature |

# Why is data important?

| | | |
|---|---|---|
| Improve people's lives | Make informed decisions | Stop molehills from turning into mountains |
| Get the results you want | Find solutions to problems | Back up your arguments |
| Stop the guessing game | Be strategic in your approaches | Know what you're doing well |
| Keep track of it all | Make the most of your money | Access the resources around you |

*Council on Quality and Leadership*

# From planning to destruction, the lifecycle of data

- Data Planning
- Data Generation
- Data Collection or Acquisition
- Data Processing
- Data Storage or Preservation
- Data Management
- Data Analysis
- Data Publishing or Sharing
- Data Visualization
- Data Interpretation
- Continual Actions
- Archiving/Destruction

USGS
science for a changing world

HDSR
HARVARD DATA SCIENCE REVIEW

sanborn
www.sanborn.com

UK • DATA ARCHIVE

# Data Planning

| |
|---|
| **Data Planning** |
| Data Generation |
| Data Collection or Acquisition |
| Data Processing |
| Data Storage or Preservation |
| Data Management |
| Data Analysis |
| Data Publishing or Sharing |
| Data Visualization |
| Data Interpretation |
| Continual Actions |
| Archiving/Destruction |

**The First Step**

- Why is the data important?
- How is it being collected?
- How will it be stored?
- Determine data parameters
  - Schema
  - Storage
  - Access
  - Maintenance

# Data Generation, Collection, or Acquisition

- Data Planning
- **Data Generation**
- **Data Collection or Acquisition**
- Data Processing
- Data Storage or Preservation
- Data Management
- Data Analysis
- Data Publishing or Sharing
- Data Visualization
- Data Interpretation
- Continual Actions
- Archiving/Destruction

## Generation

- Generating data is more applied to big data
- Large, unstructured datasets
- Doesn't work well for GIS without data wrangling
- *Not considered a core phase of the data lifecycle*

## Collection/Acquisition

- Data collected through field operations, digitization, or from an external source
- Field operations and digitization draw directly from **Data Planning**
- Acquisition may require data wrangling

# Data Processing

- Data Planning
- Data Generation
- Data Collection or Acquisition
- **Data Processing**
- Data Storage or Preservation
- Data Management
- Data Analysis
- Data Publishing or Sharing
- Data Visualization
- Data Interpretation
- Continual Actions
- Archiving/Destruction

**Also known as...**

- Data wrangling
- Data preparation
- Data munging

**How to process data...**

- Transforms raw data into desired data through:
  - Cleaning
  - Structuring
  - Enriching
- Differs from data cleansing

# Data Storage or Preservation

- Data Planning
- Data Generation
- Data Collection or Acquisition
- Data Processing
- **Data Storage or Preservation**
- Data Management
- Data Analysis
- Data Publishing or Sharing
- Data Visualization
- Data Interpretation
- Continual Actions
- Archiving/Destruction

**Outside of data planning...**

- Builds data storage based on generated and wrangled data
- Can result in schema, field naming, accuracy, precision, and naming convention errors
- *Not considered a core phase of the data lifecycle*

# Data Management

- Data Planning
- Data Generation
- Data Collection or Acquisition
- Data Processing
- Data Storage or Preservation
- **Data Management**
- Data Analysis
- Data Publishing or Sharing
- Data Visualization
- Data Interpretation
- Continual Actions
- Archiving/Destruction

| Jeanette Wing | Sanborn |
| --- | --- |
| • Optimized storage process<br>• Varies based on data generated<br>• *Not considered a core phase of the data lifecycle* | • Analysis<br>• Data updates<br>• *Not considered a core phase of the data lifecycle* |

# Data Analysis

- Data Planning
- Data Generation
- Data Collection or Acquisition
- Data Processing
- Data Storage or Preservation
- Data Management
- **Data Analysis**
- Data Publishing or Sharing
- Data Visualization
- Data Interpretation
- Continual Actions
- Archiving/Destruction

**Creates new data...**

- Through analysis, derivative data is created
- Initial data input is not changed
- *Not considered a core phase of the data lifecycle (but is closely related)*

# Data Publishing or Sharing

- Data Planning
- Data Generation
- Data Collection or Acquisition
- Data Processing
- Data Storage or Preservation
- Data Management
- Data Analysis
- **Data Publishing or Sharing**
- Data Visualization
- Data Interpretation
- Continual Actions
- Archiving/Destruction

**Providing access to data…**

- Whether by open data, ftp, REST, or other means of sharing
- Data can be viewed, downloaded, and accessed
- Data doesn't change
- *Not considered a core phase of the data lifecycle*

# Data Visualization

- Data Planning
- Data Generation
- Data Collection or Acquisition
- Data Processing
- Data Storage or Preservation
- Data Management
- Data Analysis
- Data Publishing or Sharing
- **Data Visualization**
- Data Interpretation
- Continual Actions
- Archiving/Destruction

**Seeing the data…**

- Symbolizes the data, whether on a map, chart, or graph
- The data doesn't change
- *Not considered a core phase of the data lifecycle*

# Data Interpretation

- Data Planning
- Data Generation
- Data Collection or Acquisition
- Data Processing
- Data Storage or Preservation
- Data Management
- Data Analysis
- Data Publishing or Sharing
- Data Visualization
- **Data Interpretation**
- Continual Actions
- Archiving/Destruction

**What does it all mean?**

- Provides explanation to data analysis and visualization
- The data doesn't change
- *Not considered a core phase of the data lifecycle*

# Continual Actions

- Data Planning
- Data Generation
- Data Collection or Acquisition
- Data Processing
- Data Storage or Preservation
- Data Management
- Data Analysis
- Data Publishing or Sharing
- Data Visualization
- Data Interpretation
- **Continual Actions**
- Archiving/Destruction

## Repeatable processes...

- Describe data
  - Metadata creation
  - Data dictionaries
  - Data discovery
- Data cleansing
  - “ensures that data are properly collected, handled, processed, used, and maintained at all stages” USGS

- Data security
  - Backups
  - Cyber security

# Archiving/Destruction

- Data Planning
- Data Generation
- Data Collection or Acquisition
- Data Processing
- Data Storage or Preservation
- Data Management
- Data Analysis
- Data Publishing or Sharing
- Data Visualization
- Data Interpretation
- Continual Actions
- **Archiving / Destruction**

## The Last Step

- What to do with data when it's being replaced or becomes irrelevant?
- None of the sources reviewed discussed this step
- Depends on organizational policies and practices

## How to...

- There are several ways to archive or destroy data:
  - Stored as a read-only
  - Transferred to external media
  - Server snapshot
  - Deleted
- Census vs Municipal

# The Core Data Lifecycle

- Data Planning
- Data Acquisition
- Data Wrangling
- Maintenance Operations
- Data Archival and Destruction

- Having bad data results in bad analyses, maps, interpretations, and results
- This can be prevented by utilizing data maintenance operations:
  - Backups
    - Prevents loss of data caused by degradation, accidents, corruption, or malicious actions
  - Documentation
    - Metadata creation and management
    - Data dictionaries
    - Keyword data discovery
  - Quality management
    - Data cleansing
    - In-place tools
    - Scheduled operations

# Garbage In, Garbage Out

Having data is good, having good data is better

# Existing Tools

### Topology

Rules set by the user to define a geospatial relationship

### Domains

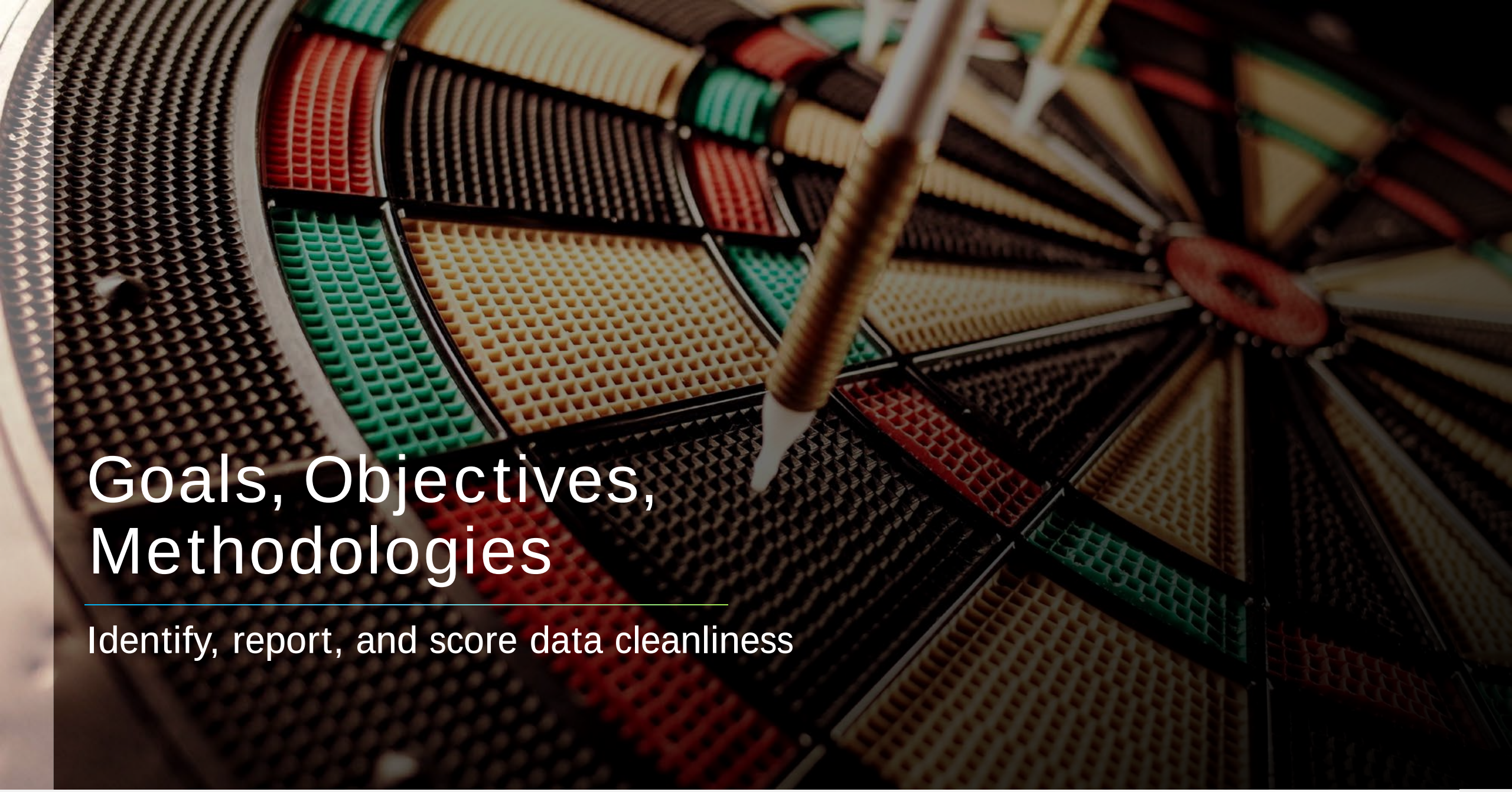
Set inputs for specified fields

### Data Reviewer

Semi-automated data review to identify errors and issues

### Attribute Assistant

Aids users in data entry based on intersecting features



# Goals, Objectives, Methodologies

Identify, report, and score data cleanliness

# Identify Known Issues

- Naming conventions
- Field name truncation
- Alias utilization
- Field attribute policies
  - NULL vs blank or 0
  - Empty data
  - Domain likeness, duplication, use
  - Field type vs data entered
- Metadata

- Bad/illegal characters
- Leading and ending spaces
- Double (or more) spaces
- Mixed coordinate systems
- Reserved words
- Hosted feature last edits
- Published feature metadata
  - Naming, summary, description, terms of use, tags, credits

**Your input needed!**

- Don’t see an issue you deal with on this list?
- Let me know about it!

# Determine Solutions for Reporting Known Issues

## Conceptual Solution Examples

### Reserved Words

- Using python and arcpy to compare a list of field names to the list of reserved words
- Using a looped if-then statement to flag uses of any reserved words

### Naming Conventions

- Types of cases…
  - flatcase
    UPPERFLATCASE
    lowerCamelCase
    UpperCamelCase
    snake_case
    SCREAMING_SNAKE_CASE
    Camel_Snake_Case
- Using python, regex, arcpy, and machine learning to identify whether the naming convention matches the user-selected convention

### Mixed Coordinate Systems

- Using python and arcpy to describe the spatialReference property and flag datasets that don't match the user-selected coordinate system

# Scoring Cleanliness, Summary Report and Appendix

## Flag and report the issues

- The tool will flag all known issues based on what the user wanted to run
- A summary report template will be filled in with the appropriate data about the issues observed

## Scoring cleanliness of data

- Using the reported issues, score the cleanliness of data
- Not yet determined formula for weighting and scoring issues, errors, and inconsistencies

## Provide appendix of issues

- To aid the user, an appendix of all issues will be generated
- The appendix will include feature dataset, feature, field name, and OID, grouped by issue

# Anticipated Results

What will work? What won't?

Unique Solutions for Unique Issues

Simple Solutions

- Reserved Word List Compare

Complex Solutions

- Naming Conventions

Philosophical Choice

- NULL vs Blank

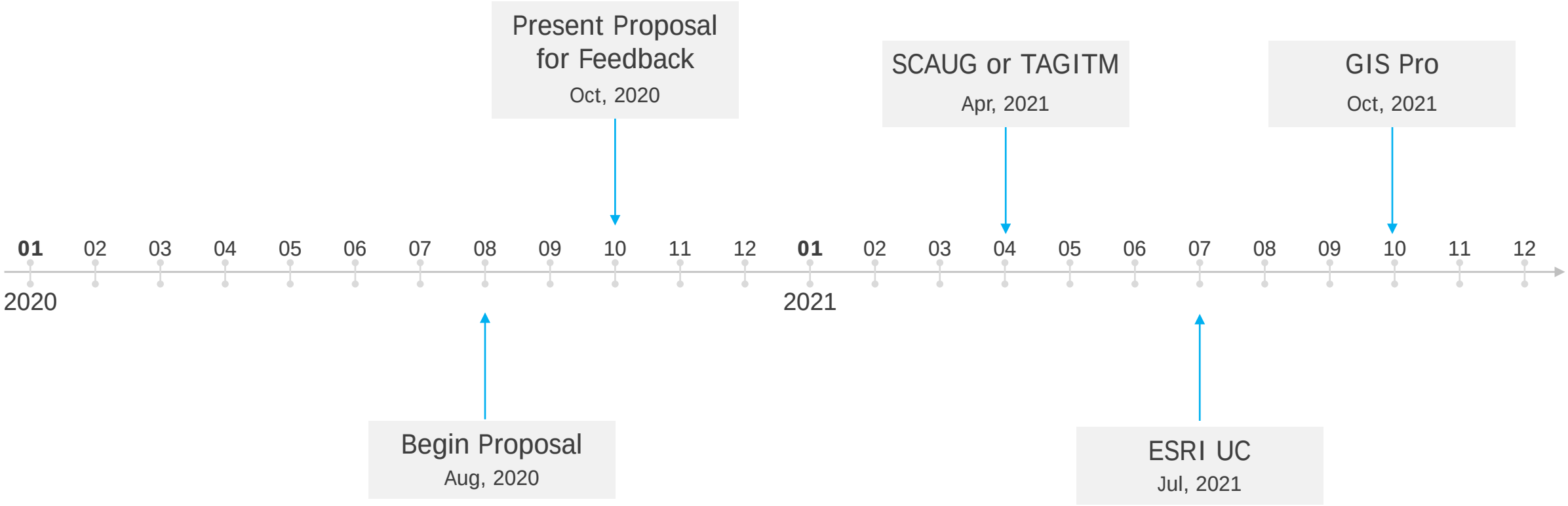
Data Cleanliness Score Formula

# Timeline

Project completion and presentation

# Timeline

From starting proposal to potential presentation venues



# Sources

The Big Geospatial Data Management Lifecycle. (2018, October 17). Retrieved August 21, 2020, from https://www.sanborn.com/the-big-geospatial-data-management-lifecycle/

CQL. (2020, July 27). 12 Reasons Why Data Is Important. Retrieved September 25, 2020, from https://www.c-q-l.org/resources/guides/12-reasons-why-data-is-important/

Data Management. (n.d.). Retrieved August 21, 2020, from https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle

Dempsey, C. (2017, May 1). Types of GIS Data Explored: Vector and Raster. Retrieved September 25, 2020, from https://www.gislounge.com/geodatabases-explored-vector-and-raster-data/

ESRI. (2010). ArcGIS Geodatabase Topology Rules. Retrieved October 08, 2020, from http://resources.arcgis.com/en/help/main/10.2/01mm/pdf/topology_rules_poster.pdf

ESRI. (n.d.). Attribute Assistant. Retrieved October 09, 2020, from https://solutions.arcgis.com/shared/help/attribute-assistant/documentation/

ESRI. (n.d.). Get started with Data Reviewer. Retrieved October 09, 2020, from https://pro.arcgis.com/en/pro-app/help/data/validating-data/get-started-with-data-reviewer.htm

ESRI. (n.d.). Introduction to attribute domains. Retrieved October 09, 2020, from https://pro.arcgis.com/en/pro-app/help/data/geodatabases/overview/an-overview-of-attribute-domains.htm

ESRI. (n.d.). What is raster data? Retrieved September 25, 2020, from https://desktop.arcgis.com/en/arcmap/latest/manage-data/raster-and-images/what-is-raster-data.htm

Eynden, V. V. (2011). *Managing and sharing data: Best practice for researchers*. Colchester, Essex: UK Data Archive.
Fagin, S., & Ommen, A. (2017, February). The Importance of Data Quality Within Your Organization [Scholarly project]. In ESRI Proceedings. Retrieved September 18, 2020, from https://proceedings.esri.com/library/userconf/fed17/papers/fed_29.pdf

Jasuja, N., T, K., Sehgal, P., & S, P. (n.d.). Data vs Information. Retrieved September 25, 2020, from https://www.diffen.com/difference/Data_vs_Information

Rouse, M. (2008, March 12). What is garbage in, garbage out (GIGO)? Retrieved September 25, 2020, from https://searchsoftwarequality.techtarget.com/definition/garbage-in-garbage-out

Sadler, C. E., Glass, A., & Doyle, A. C. (1981). Sir Arthur Conan Doyle's The Adventures of Sherlock Holmes. New York, NY: Avon Books.

Wing, J. M. (2019). The Data Life Cycle. *Issue 1 Harvard Data Science Review*. doi:10.1162/99608f92.e26845b4

# Questions?

# Thank You

Neil S. Rose, GISP

nsr8@psu.edu