Data Visualization

Data Analytics, Visualization, and Storytelling Webinar
Data Science Workflow

- Geothermal Design Challenge 2019
- What is a Data Scientist?
- Storytelling
- Data Acquisition and Cleaning
- Data Analytics
- Visualization
- Data Analytics Resources
- FORGE Data

"Statistical modeling and machine learning are now becoming table stakes in order to become a data scientist. The differentiator is how well those working in the field can communicate their findings in a simple, but actionable way."

Source: Prashutdesai, Manager of Statistical Analysis, Charles Schwab
Geothermal Design Challenge 2019

- **Who**
  - High school and university (undergraduates & graduates) teams of 2-3 members

- **What**
  - Visualize the world of geothermal energy through compelling analytics, visualization & storytelling
  - Use provided data (or in combination with other open source data) to recommend a location within the FORGE (Frontier Observatory for Research in Geothermal Energy) footprint for an enhanced geothermal reservoir

- **When**
  - January 7, 2019 – April 10, 2019 (10 PM ET)

- **How**
  - There are standard methodologies to cite geothermal wells, such as geological interpretation, geospatial analysis, etc. However, other methods may offer opportunities and promise like advanced data analytics and machine learning. This data visualization challenge could provide opportunities that haven’t been explored yet.

- **But... I am not a geothermal expert**
  - EXCELLENT!!! We are looking for unique perspectives. Ask questions, engage us. Your lack of experience (which can be biased) is a strength in this competition
What is a Data Scientist?

Data scientists extract knowledge, insights, or solutions from big data. Here's a look at the life of a data scientist, based on two surveys of data scientists around the globe.
Storytelling

- Some Resources
  - Storytelling with Data (Cole Knaflic)
  - Tableau - Storytelling with Data
  - FORGE Earth Model
  - Data Visualization: How To Tell A Story With Data

- How are stories different from visualization?
  - Maybe stories are just data with a soul. — Brené Brown
  - In school, we learn a lot about language and math. But no one teaches us to tell stories with numbers – C. Knaflic
  - Tableau Gallery - Rainforests
Storytelling Basics

- Understand the Context
  - Who is your audience? What do you need them to know? What is the best communication mechanism and desired tone?
  - STORYBOARDING - use a whiteboard or stickie notes to craft your structure

- Choose an Appropriate Visual Display
  - What is the best way to show data you want to communicate? What are the best graphs to use and those to not use?

- Clutter is your Enemy
  - Think Cognitive Load
  - How does each single element you add convey a message?
  - Gestalt Principles of Visual Perception - how your audience perceives order in the world around them

The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
Storytelling Basics

- **Focus Attention Where You Want It**
  - How people see and how to use that
  - Pre-attentive attributes (size, color, position)

- **Think Like a Designer**
  - Form follows Function – YOU are a storyteller!
  - Affordances (highlight important stuff, eliminate distractions), Accessibility (know your audience, don’t overcomplicate), Aesthetics (be smart with color, alignment, leverage white space)

- **Tell a Story**
  - Think about how visuals will be interpreted; dissect them
  - Stories stick with us, data does not
  - Use Big Idea, storyboarding to sequence and tell your story

Infographic Examples
THE BARRIERS TO INVENTION

Not every creative mind actually gets to create, and where or how you live can make a big difference. Respondents saw a lot of potential obstacles in the path of inventors.

INVENTIONS IN NEED OF REINVENTION

- **13%** COMPUTER
- **11%** CAR
- **8%** TELEPHONE

- **20%** among Indian consumers
- **33%** among South African consumers
- **31%** among Indonesian consumers
- **2%** NOTHING LEFT TO INVENT
- **5%** NO URGENCY TO INVENT
- **9%** NO INCENTIVE TO INVENT
- **12%** FAST PACE OF LIFE STIFLES CREATIVITY
- **14%** A WEAK ECONOMY
- **16%** LACK OF GOVERNMENT PROTECTION FOR INVENTORS’ RIGHTS
- **17%** LACK OF RESOURCES
- **18%** A POOR EDUCATION SYSTEM
- **9%** POLITICAL INSTABILITY
DATA SCIENCE LIFECYCLE

01 BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING
Gather and scrape the data necessary for the project.

03 DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.

06 PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
Data Acquisition & Cleaning

- Data Mining / Acquisition
  - Data has been provided to you although you may choose to combine with other open source data

- Data Cleaning
  - Fix inconsistencies, handle missing values

- Exploratory Data Analysis & Feature Engineering
  - Simple analysis to understand patterns and biases in data to form a hypothesis
  - Selection (or creation) of important features from raw data (or combination of raw data) - increase interpretability, maintain accuracy of your models
  - Includes dimensionality reduction
  - Example: Food Pyramid
Data Acquisition & Cleaning (50-80% of the task)

- **Acquisition (at Geothermal Design Repository)**
  - Files are provided: csv and las file formats
  - Python liblas library; R package rlas; ArcGIS
  - Next webinar (Feb 19) goes into details on data

- **Data Cleaning**
  - Identify incomplete, incorrect, inaccurate or irrelevant parts of the data and then replace, modify, and/or delete dirty data
  - Example: Data & data types can be inconsistent (e.g., 0 / 1 vs no / yes; integers or strings; case sensitive)
  - Common packages
    - R - dplyr
    - Python - pandas
    - Tableau best practices
Data Acquisition & Cleaning (50-80% of the task)

- **Exploratory Data Analysis**
  - Brainstorming of data analysis; understand patterns and bias in the data; simple visualizations, subset evaluation, explore story behind outliers
  - Beginning of a hypothesis and your story

- **Feature Engineering (Selection & Construction)**
  - Dimensionality reduction – increase clarity, reduce complexity
  - Using domain knowledge to transform your raw data into informative features that represent the problem.
  - Common packages
    - Python: [Scikit-learn](https://scikit-learn.org), featuretools
    - SAS automated feature engineering
Data Analytics
Everything goes in a box or on a curve

- **Clustering** (unsupervised learning)
  - Grouping similar objects together
- **Classification** (supervised learning)
  - Putting objects in a specific class
  - Maps input-output pairs
- **Regression**
  - Predicting an input-output quantity
- Reinforcement Learning (not covered)
- Example: rock types
Machine Learning Algorithms Cheat Sheet

Unsupervised Learning: Clustering
- **k-means**
  - Gaussian Mixture Model: **YES**
  - Prefer Probability: **NO**
  - Categorical Variables: **YES**
  - Need to Specify k: **YES**
  - DBSCAN: **NO**
- **k-modes**
  - YES

Unsupervised Learning: Dimension Reduction
- **Dimension Reduction**
  - YES
  - Principal Component Analysis: **YES**
  - Singular Value Decomposition: **NO**
- **Topic Modeling**
  - Probabilistic: **YES**
  - **Latent Dirichlet Analysis**

Supervised Learning: Classification
- **Linear SVM**
  - Data Is Too Large: **NO**
  - Explainable: **NO**
  - Speed or Accuracy: **SPEED**
  - Naive Bayes: **YES**
  - Decision Tree: **YES**
  - Kernel SVM: **NO**
- **Naive Bayes**
  - Linear Regression: **NO**
  - Random Forest: **NO**
  - Neural Network: **NO**
  - Gradient Boosting Tree: **NO**

Supervised Learning: Regression
- **Predicting Numeric**
  - Speed or Accuracy: **ACCUACY**
  - Decision Tree: **SPEED**
  - Random Forest: **ACCUACY**
  - Neural Network: **ACCUACY**
  - Gradient Boosting Tree: **ACCUACY**
Data Analytics – Common Software
Everything goes in a box or on a curve

- **SAS overview, algorithms**
Data Analytics - Validation

- **Validation**
  - Critical to evaluate success
  - Check your work!
Visualization

- Introduction
  - Ted Talk - Beauty of Data Visualization
  - Tableau Gallery
- Chart types
  - Tableau chart types
  - Tableau getting started with visualization
  - Tableau training (free)
  - Examples: Bar, Line, Scatter plot, Gantt, Bubble, Histogram, Bullet, Heat map, Table, Tree map, Bow & whisker
- Be creative! Tell your Story

Most of us need to listen to the music to understand how beautiful it is. But often that’s how we present statistics: we just show the notes, we don’t play the music – H. Rosling

My best advice is to not start in PowerPoint. Presentation tools force you to think through information linearly, and you really need to start by thinking of the whole instead of the individual lines – N. Duarte
Analytics Resources

- **The Art of Data Visualization**
  - https://www.youtube.com/watch?v=Ad5ZJzb-aX8

- **Python - Real World Examples**
  - https://www.youtube.com/watch?v=VI9R-QnDwNA

- **ArcGIS Basics**
  - https://www.youtube.com/watch?v=N-5FCICaMyM

- **Tableau - Data Visualization Beginners Guide**
  - https://www.tableau.com/learn/articles/data-visualization

- **Overview of SAS Visual Analytics**
Analytics Resources

- Top 74 Data Visualization Software
- A 5-Step Guide to Data Visualization
- 10 Useful Ways to Visualize Your Data (With Examples)
- 25 Tips to Instantly Improve Your Data Visualization Design
- Tableau for Students
- Tableau for Students FAQs
- Tableau Student Resource Page
- Python
- Python Guides and Resources
- R Studio
- SAS University Edition (Free)
- Learn SAS
- Matlab
- Datacam
FORGE Data

- Geothermal Design Repository

Utah FORGE Roosevelt Hot Springs Well Data for Student Competition

Abstract
Well 58-32 (previously labeled MU-ESW1) was drilled near Milford Utah during Phase 2B of the FORGE Project to confirm geothermal reservoir characteristics met requirements for the final FORGE site.

Well Accord-1 was drilled decades ago for geothermal exploration purposes. While the conditions encountered in the well were not suitable for developing a conventional hydrothermal system, the information obtained suggested the region may be suitable for an enhanced geothermal system.

Geophysical well logs were collected in both wells to obtain useful information regarding the nature of the subsurface materials. For the recent testing of 58-32, the Utah FORGE Project contracted with the well services company Schlumberger to collect the well logs.

12 Resources

- 58-32 Thermal Conductivity Data.csv
  - Manual measurements of thermal conductivity drill cuttings
  - Download (1.47 MB)

- 58-32 X-ray Diffraction Data.csv
  - Manual measurements of mineral abundances from drill cuttings
  - Download (4.08 MB)

- Well Survey from Earth Model.csv
  - Data exported from the site earth model
  - Download (4.03 MB)

- Well Data ReadMe.txt
  - An overview of the data contained in this submission
  - Download (27.3 MB)

- Well Locations from Earth Model.csv
  - Data exported from the site earth model
  - Download (24.5 MB)