Computation and Visualization for Analytics

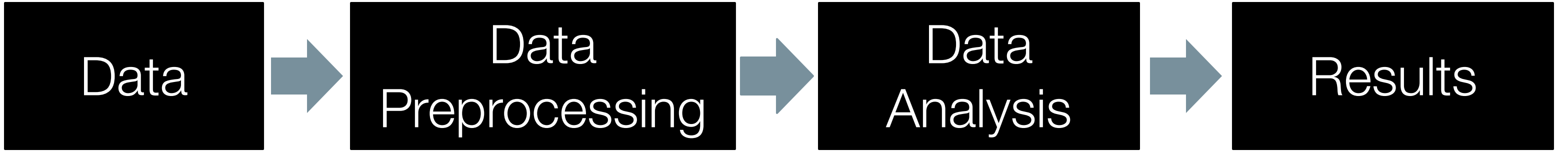
Spring 2021

Week 1.1

- Data analytics pipeline
- Practical data visualization
- Data vis jobs
- Course structure
- Expectations
- Contact

Data Analytics Pipeline

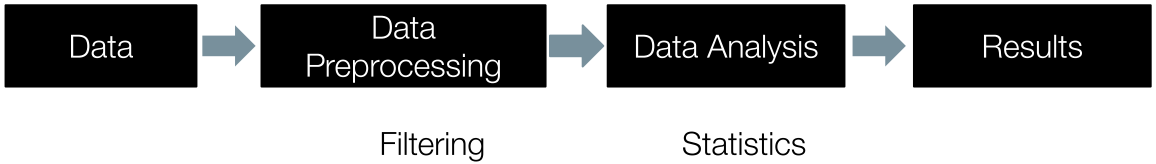
Data Analytics Pipeline



Data → Data Preprocessing → Data Analysis → Results

- Cleaning
- Filtering
- Pivoting
- Aggregating

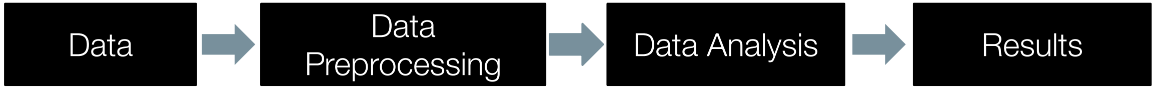
- Statistics
- Mathematics
- Machine Learning
- **Visualization**



Name	Program	GPA
Joe	UG	3.8
Alice	G	3.9
Xu	G	3.85
Amal	UG	3.83
Amit	G	3.75

Name	Program	GPA
Alice	G	3.9
Xu	G	3.85
Amit	G	3.75

Average GPA



Data

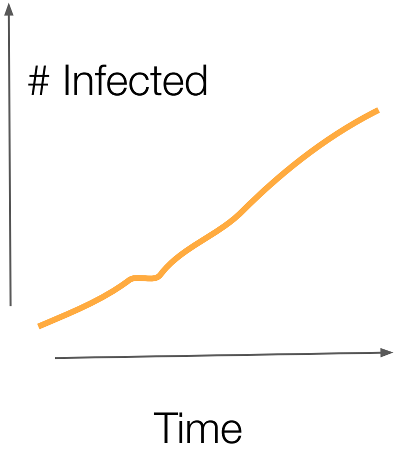
Data Preprocessing

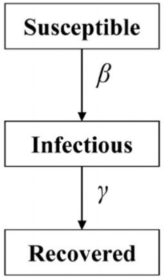
Data Analysis

Results

Filtering

Math Model

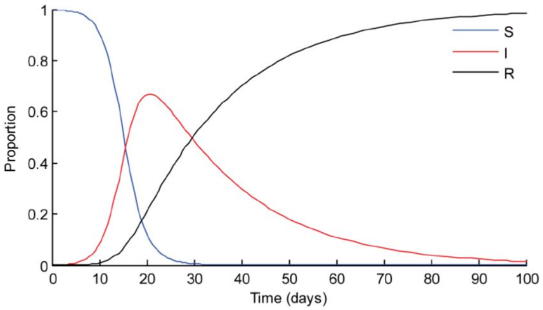


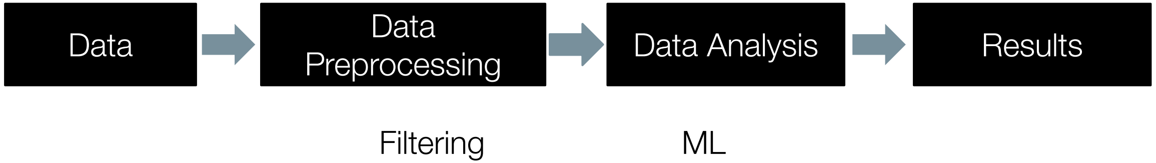


$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

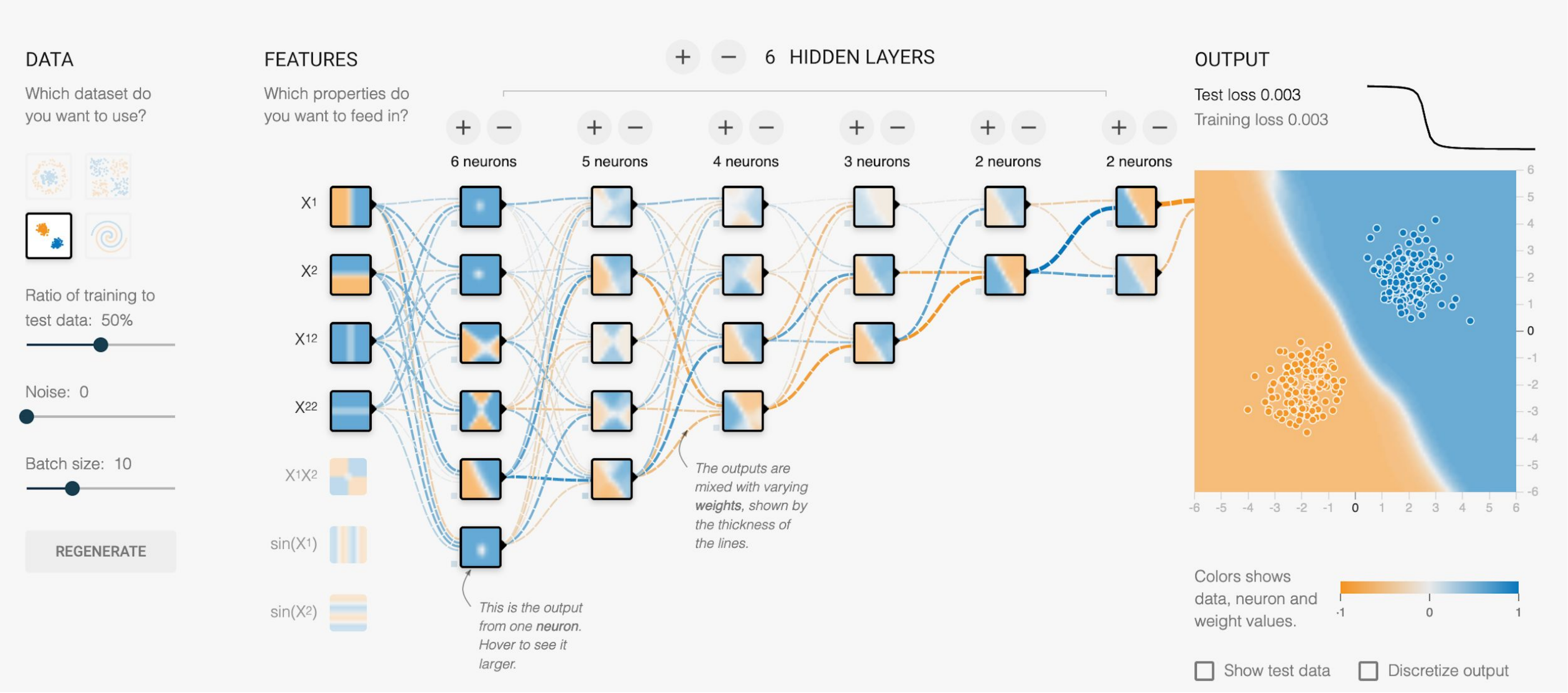




Data → Data Preprocessing → Data Analysis → Results

Filtering

ML



DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

REGENERATE

FEATURES

Which properties do you want to feed in?

+ − 6 HIDDEN LAYERS

OUTPUT

Test loss 0.003

Training loss 0.003

This is the output from one ***neuron****. Hover to see it larger.*

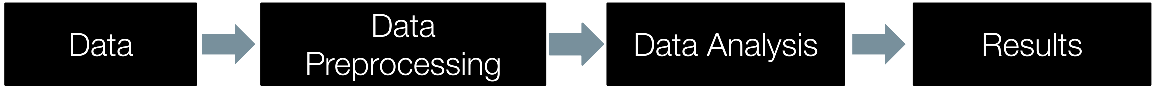
The outputs are mixed with varying ***weights****, shown by the thickness of the lines.*

Colors shows data, neuron and weight values.

□ Show test data

□ Discretize output

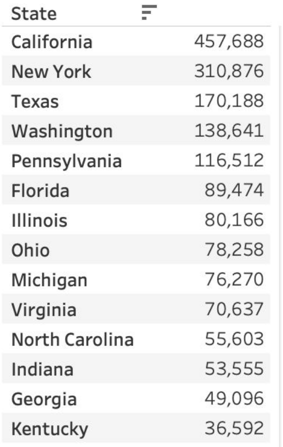
Tensorflow



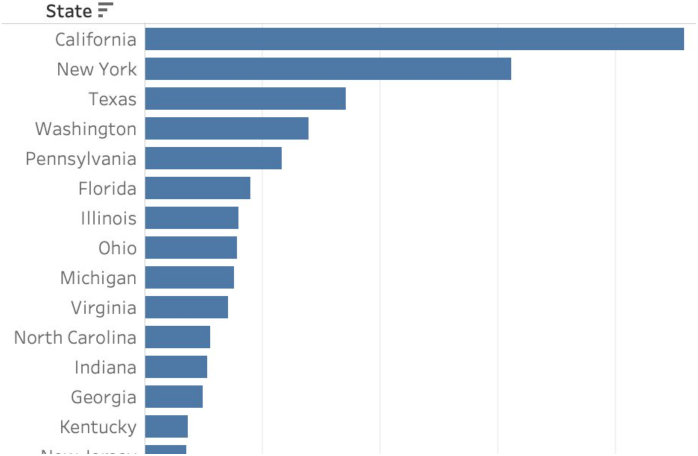
Data → Data Preprocessing → Data Analysis → Results

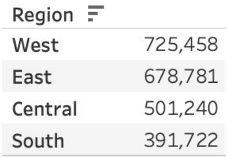
Aggregation

Visualization

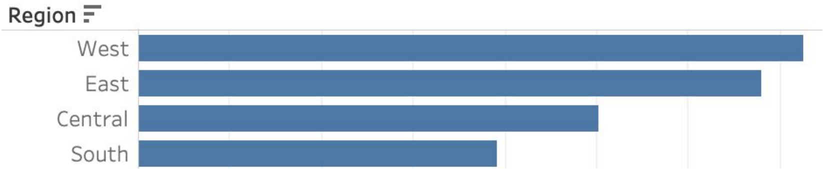


State	
California	457,688
New York	310,876
Texas	170,188
Washington	138,641
Pennsylvania	116,512
Florida	89,474
Illinois	80,166
Ohio	78,258
Michigan	76,270
Virginia	70,637
North Carolina	55,603
Indiana	53,555
Georgia	49,096
Kentucky	36,592

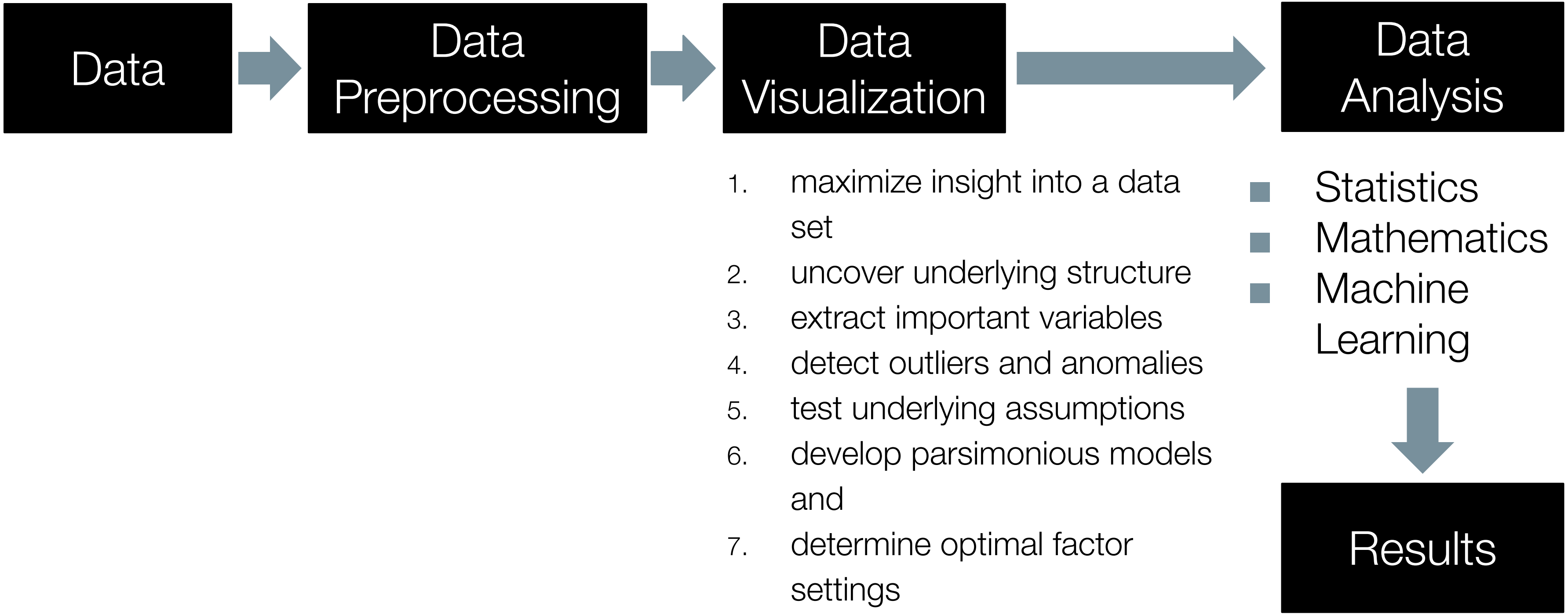




Region	
West	725,458
East	678,781
Central	501,240
South	391,722



Data Visualization Pipeline for EDA



Data

Data Preprocessing

Data Visualization

Data Analysis

1. maximize insight into a data set
2. uncover underlying structure
3. extract important variables
4. detect outliers and anomalies
5. test underlying assumptions
6. develop parsimonious models and
7. determine optimal factor settings

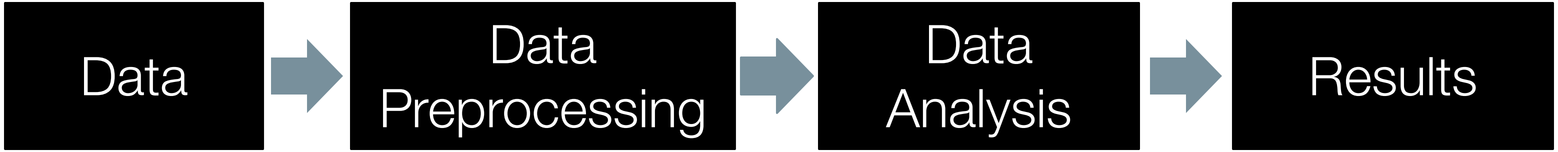
- Statistics
- Mathematics
- Machine Learning

Results

NIST

Practical Data Visualization

Data Vis Pipeline (Tools)



Data → Data Preprocessing → Data Analysis → Results

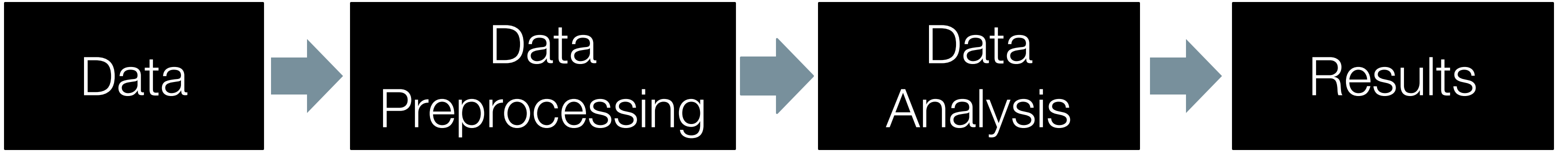
- SQL
- Cloud SQL
- Excel

- R
- Python
- C++
- Java
- Tableau prep
- Google cloud dataprep

- R
- Python
- Tableau
- Power Bi
- D3
- Google data studio

Jobs

Expectations


Data

Data Preprocessing

Data Analysis

Results

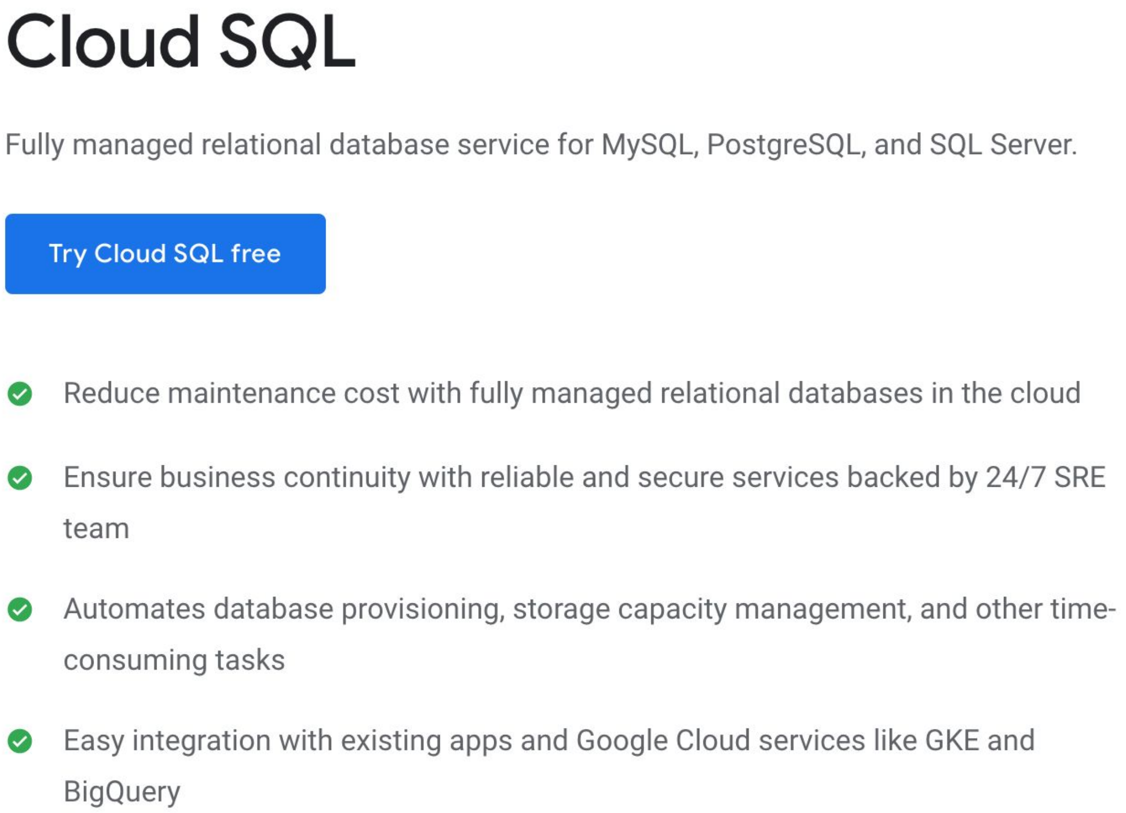
Skills + Tools
Data analysis = skill
Data vis = skill
Machine learning = skill

Skills ≠ Tools

Data analyst
Data scientist
Data vis specialist
Data vis engineer
Data vis in journalism

Data

- SQL
- Cloud SQL



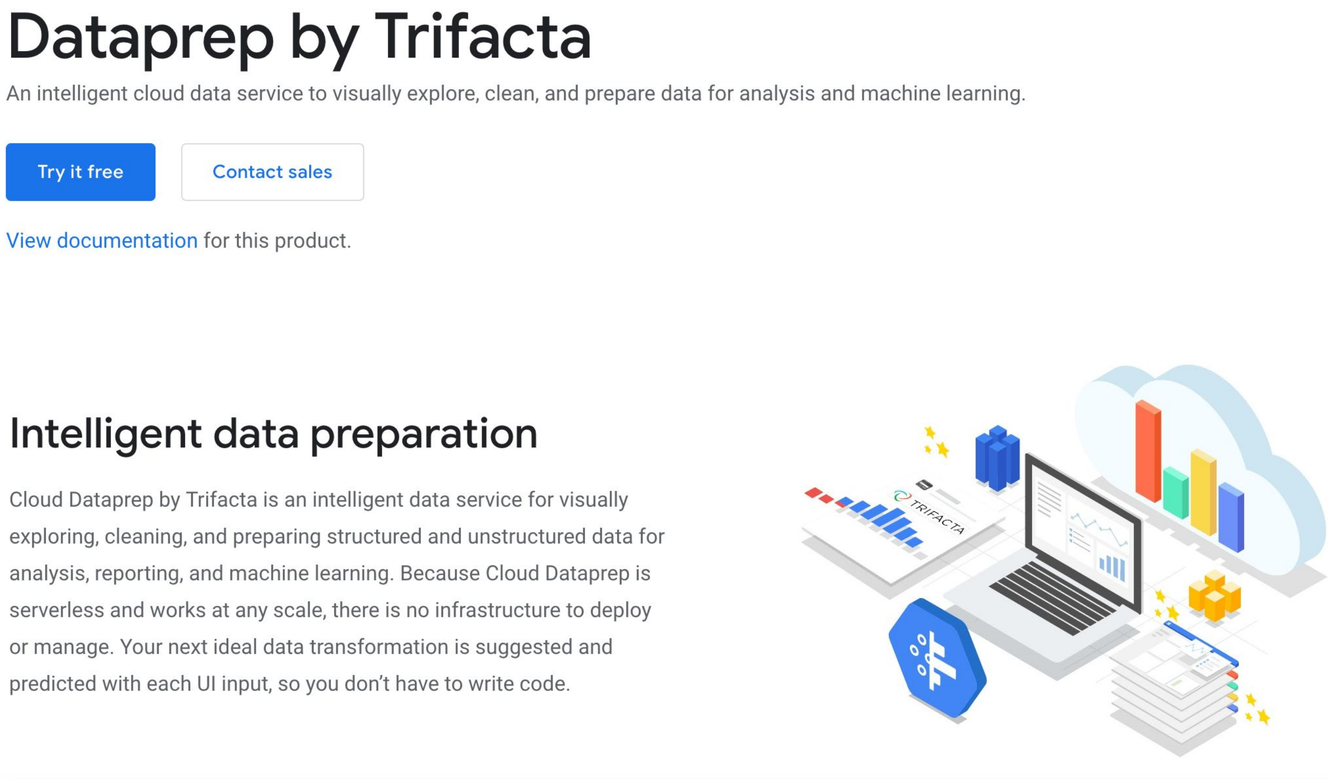
Cloud SQL

Fully managed relational database service for MySQL, PostgreSQL, and SQL Server.

- ✓ Reduce maintenance cost with fully managed relational databases in the cloud
- ✓ Ensure business continuity with reliable and secure services backed by 24/7 SRE team
- ✓ Automates database provisioning, storage capacity management, and other time-consuming tasks
- ✓ Easy integration with existing apps and Google Cloud services like GKE and BigQuery

Data Preprocessing

- Cleaning
- Filtering
- Pivoting
- Aggregating



Dataprep by Trifacta

An intelligent cloud data service to visually explore, clean, and prepare data for analysis and machine learning.

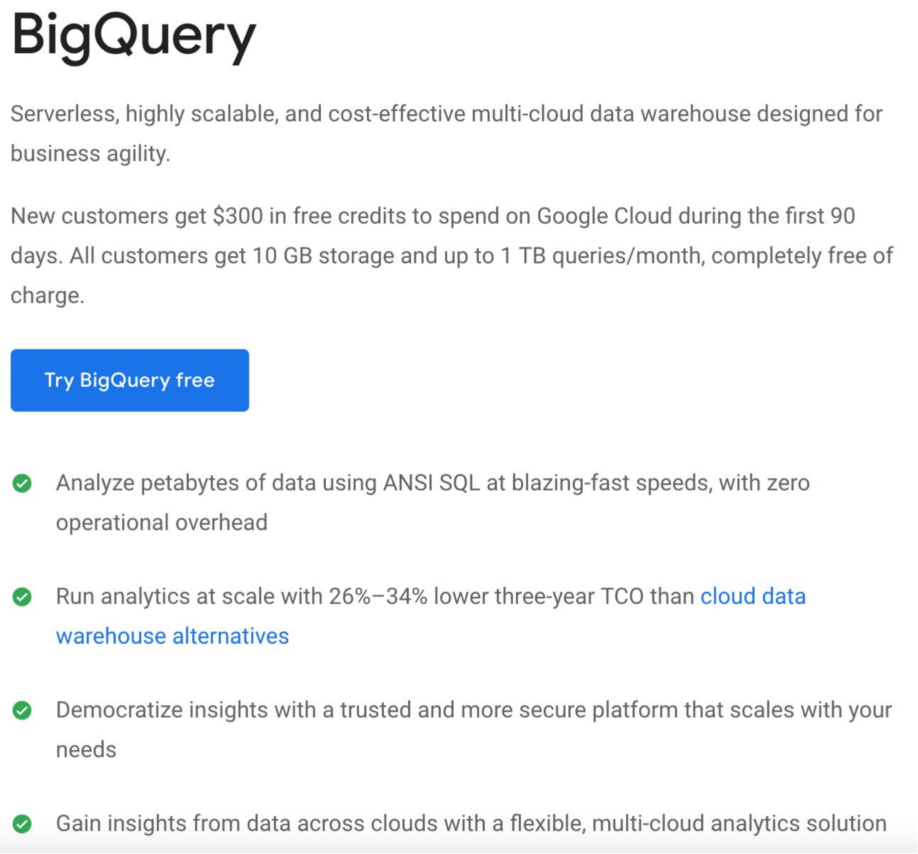
View documentation for this product.

Intelligent data preparation

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis, reporting, and machine learning. Because Cloud Dataprep is serverless and works at any scale, there is no infrastructure to deploy or manage. Your next ideal data transformation is suggested and predicted with each UI input, so you don't have to write code.

Data Analysis

- Statistics
- ML
- Visualization



BigQuery

Serverless, highly scalable, and cost-effective multi-cloud data warehouse designed for business agility.

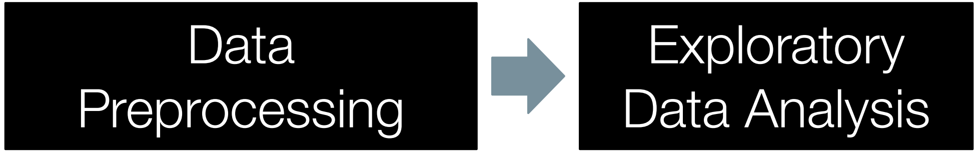
New customers get $300 in free credits to spend on Google Cloud during the first 90 days. All customers get 10 GB storage and up to 1 TB queries/month, completely free of charge.

Try BigQuery free

- Analyze petabytes of data using ANSI SQL at blazing-fast speeds, with zero operational overhead
- Run analytics at scale with 26%–34% lower three-year TCO than cloud data warehouse alternatives
- Democratize insights with a trusted and more secure platform that scales with your needs
- Gain insights from data across clouds with a flexible, multi-cloud analytics solution

BigQuery for ML   BigQuery for Data Vis

Trends

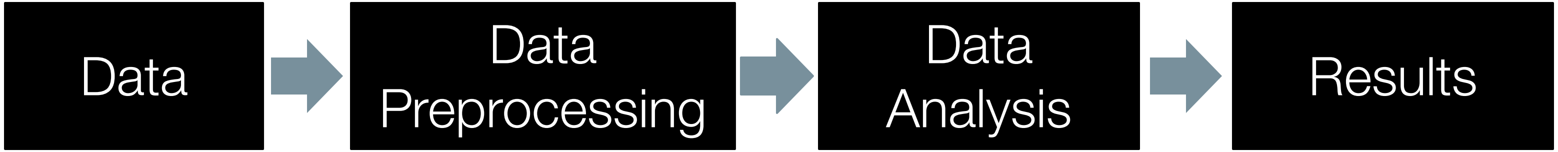


Data Preprocessing

Exploratory Data Analysis

[Automating Data Wrangling]

Resiliency



Skills + Tools

Data analysis = skill

Data vis = skill

Machine learning = skill

Skills ≠ Tools

Domain Expertise

Domain Knowledge

- Healthcare
- Manufacturing
- Engineering
- Design
- Business
- Finance
- Supply chain
- Socio technical systems

Problem: Do Helmets Increase Head Injuries?

At the beginning of the first World War, the uniform of the British soldiers included a brown cloth cap. They were not provided with metal helmets. As the war went on, the army authorities and the War Office became alarmed at the high proportion of men suffering head injuries. They therefore decided to replace the cloth headgear with metal helmets. From then on, all soldiers wore the metal helmets. However, the War Office was amazed to discover that the incidence of head injuries then increased. It can be assumed that the intensity of fighting was the same before and after this change. So why should the recorded number of head injuries per battalion increase when men wore metal helmets rather than cloth caps?

Solution: Do Helmets Increase Head Injuries?

The number of recorded head injuries increased, but the number of deaths decreased. Previously, if a soldier had been hit on the head by a piece of shrapnel, it would have pierced his cap and probably killed him. This would have been recorded as a death, not a head injury. After helmets were issued it was more likely that a fragment of shrapnel would cause an injury rather than death. Thus, the incidence of head injuries increased, while the incidence of deaths decreased.

Source in notes

Data Visualization

- Statistics
- Mathematics
- Graphic design
- Cartography
- Computer science
- Design
- Art
- Psychology

Data Vis Jobs

Data analyst
Data scientist
Data vis specialist
Data vis engineer
Data vis in journalism Link

Types of Data Visualizations

Static	https://www.economist.com/graphic-detail/2020/01/23/how-is-netflix-faring-in-the-streaming-wars
Dynamic	https://www.nytimes.com/interactive/2018/03/19/upshot/race-class-white-and-black-men.html
Interactive	https://public.tableau.com/en-us/gallery/coffee-calculator?tab=viz-of-the-day&type=viz-of-the-day

Practical Data Visualization

Columns / Variables / Fields / Features

Rows / Records / Instances

State	Postal Code	Median Income
MA	02115	50000
NH	03087	45000
MA	02116	48000

2 factors 3 factors

State	Date	Median Income	Expenditure
MA	02/10/2019	50000	10000
NH	02/10/2019	45000	8000
ma	01/20/2020	48000	9100
MA	02/10/2019	50500	11000
NH	02/10/2019	42$00	7000
MA	02/10/2019	49000	8300
NH	02/10/2019	44000	6800

Data Cleaning

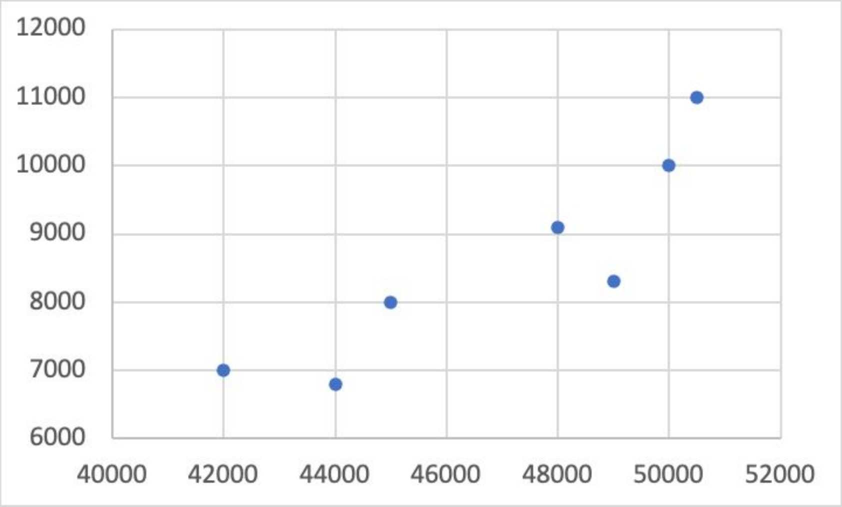
- Data type inconsistencies
- Case
- Clean special Char
- Handling NA

Data Wrangling

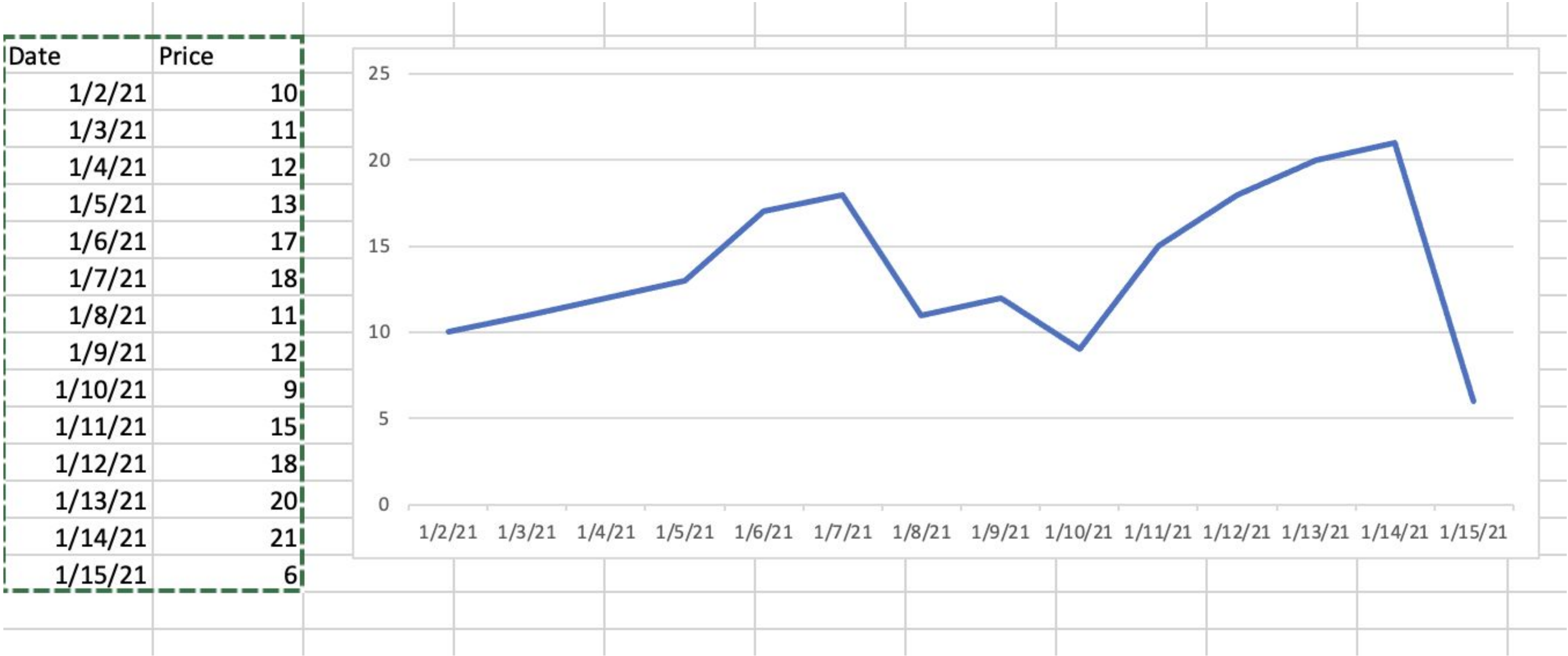
- Add remove cols/rows
- Subset data
- Pivot
- Aggregation

State	Date	Median Income	Expenditure
MA	02/10/2019	50000	10000
NH	02/10/2019	45000	8000
MA	01/20/2020	48000	9100
MA	02/10/2019	50500	11000
NH	02/10/2019	42000	7000
MA	02/10/2019	49000	8300
NH	02/10/2019	44000	6800

Non-Aggregate Visualization



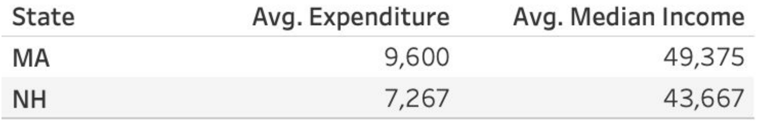
Non-Aggregate Visualization



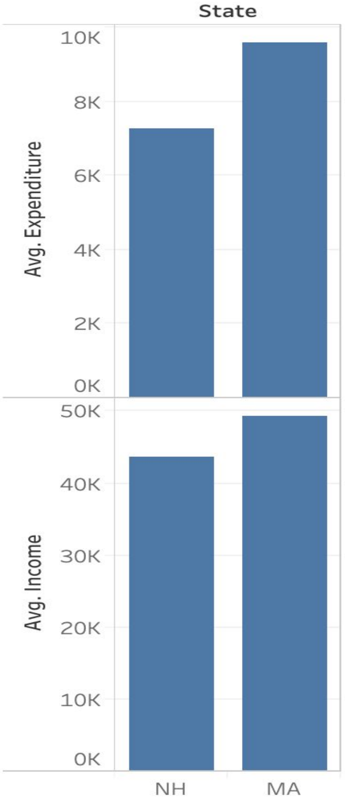
Date	Price
1/2/21	10
1/3/21	11
1/4/21	12
1/5/21	13
1/6/21	17
1/7/21	18
1/8/21	11
1/9/21	12
1/10/21	9
1/11/21	15
1/12/21	18
1/13/21	20
1/14/21	21
1/15/21	6

Aggregate Visualization

State	Date	Income	Expenditure
MA	02/10/2019	50000	10000
NH	02/10/2019	45000	8000
MA	01/20/2020	48000	9100
MA	02/10/2019	50500	11000
NH	02/10/2019	42000	7000
MA	02/10/2019	49000	8300
NH	02/10/2019	44000	6800



State	Avg. Expenditure	Avg. Median Income
MA	9,600	49,375
NH	7,267	43,667



Course Objectives

- To understand the principles and methodologies of visualization
- To learn how to explore data using visualization tools
- To design, validate and critique visualizations
- To implement interactive data visualizations

Topics

- Modern data structures
- Data cleaning and data wrangling
- Introduction to data visualization
- Visual Encoding
- Visualizing amounts
- Visualizing distributions and relationships
- Visualizing trends
- Map based visualizations
- Working with colors
- Network visualizations
- Text visualizations
- Interactive visualizations
- Dashboards and storyboards

Software

- R (R Studio)
- Tableau

Course Evaluation

- Assignments (**30%**) (6 assignments)
- Quiz (**20%**) (5 quizzes)
- Hackathon (**10%**)
- Final Project (**30%**)
  - project proposal = 30%
  - project progress presentation = 30%
  - final project presentation = 30%
  - project documentation = 10%
- Class Participation (**10%**)

- **Assignments** focus on improving the technical skills of the students
- The assignments will help the students gain expertise in using the visualization tools
- Late submission of homework will receive a penalty. For two-day delay, grades will be cut by 10%
- Beyond two days past the deadline, the submission will not be accepted

- **Quizzes** will test the understanding of data visualization concepts
- The first 15 minutes of the class will be devoted to answering the quiz questions

- **Hackathon** will test the ability of students to extract visual insights from real world datasets.
- The dataset and goals will be provided to the students on the day of hackathon.
- The data may need preprocessing and the insights from the data must be presented in the form of visualizations.

- **Final Project** will require the students to create an interactive dashboard.
- The students are expected to select a topic of interest and collect the relevant data.
- The dashboard should contain the visual elements that allow the users to interact and gain insights from the data.

Final Project Example

Contact

- Office hours will be notified by end of week
- Office hours will be conducted via Teams
- Join Slack group (Link)
- TA contact and TA hours will be updated end of week