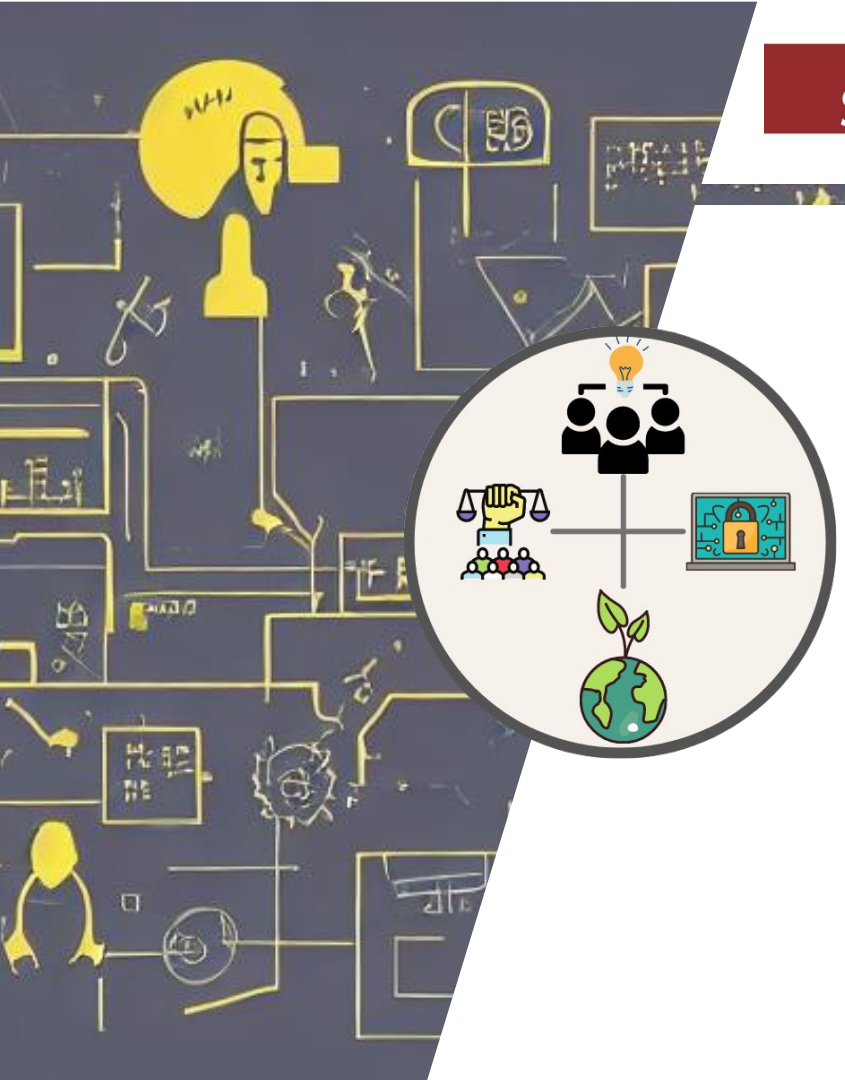


SFU SIMON FRASER UNIVERSITY

TACC TEXAS ADVANCED COMPUTING CENTER

aws

HACKHPC.ORG

HACK THE THREAT 23:

Mentor Pitches

March 2, 2023

HTTPS://HACKHPC.GITHUB.IO/HACKTHETHREAT23

Diverse Camps

Dawn Hunter - University of Texas at Austin, *Texas Advanced Computing Center (TACC)*

Create a program that would score summer camp applications. This program will be used to select attendees for one of our four free UT TACC one-week residential computer science and cybersecurity camps that target underserved students but applications are open for everyone to apply.

Needs:

- Provide equitable scoring regardless of writing skills (Understanding students from schools write differently)
- Score ~1000 short answers per camp
- Provide a score of 1 - 5 for each short answer

Suggested Skills / Resource Links

- Sentiment Analysis
- https://www.tacc.utexas.edu/education/k-12-outreach

Protect our sensitive data!

Ricardo Carvalho - *Simon Fraser University (SFU)*

Survey: 92% of Canadians expressed some level of concern about the protection of their privacy [source].

- Privacy regulations restrict the use of personal information. For example: GDPR, PIPEDA.
- However, sensitive data can be useful in, e.g., policy making, research, and drug trials.

How can we enable the use of sensitive data, while protecting the privacy of individuals?

Goal: Create a tool/app that "cleans up" a dataset w.r.t. sensitive information, making it safer to release.

Photo by Jason Dent on Unsplash

Suggested Skills / Resource Links

- Python (pandas, numpy, re)
- Dataset 1: Census, Dataset 2: HR
- One simple technique: k-anonymity

Bad Data Makes Bad AI

Lydia Fletcher, University of Texas-TACC

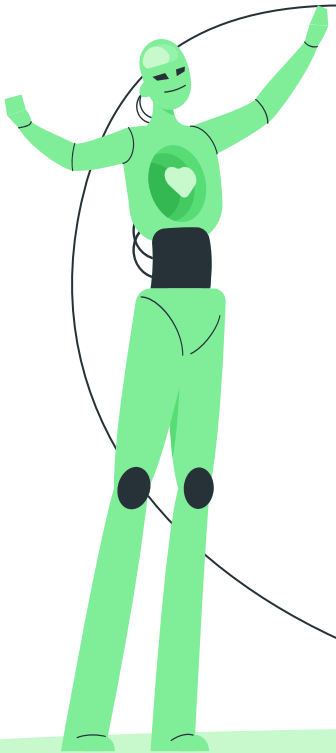
Responsible AI has FATE: fairness, accountability, transparency, and explainability.

To establish FATE and create responsible AI, we need **data provenance**.

A key component of provenance is using **metadata** to record details about a digital object.

So...

Good Data Makes Good(er) AI



... but how do we make it easier for us humans? Can we automate metadata extraction from a given dataset to reduce human processing time requirements?

GOAL: Create a tool/app that makes data curation easier.

Example:
https://github.com/unsw-cse-soc/Data-curation-API

Sample data:
https://ms2-dev.tacc.utexas.edu/texas-damage-plain/static/hackathon.zip

Choose Your Own Adventure Projects

Project ideas for ad-hoc teams

Combined impact of Hurricane Maria on different socially vulnerable areas to look at impact of storms on different populations

Information and Dataset:
https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-3525

My Hurricane Preparedness Scrapbook

Information and Dataset:
https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published//PRJ-3692/MV-HarveyNET

Spam Classification while preserving privacy of emails

Dataset:
https://www.kaggle.com/datasets/xiaoyuwan/spam-mail

Machine learning of textual data using privacy-preserving methods with open source libraries.

- Possible tasks:
  - Data cleaning
  - Feature engineering
  - Training classification model with privacy-preserving methods
  - Building a simple web application to enter free text and return spam or ham