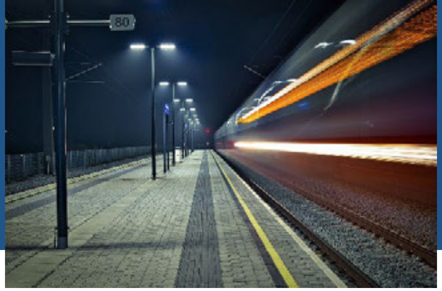
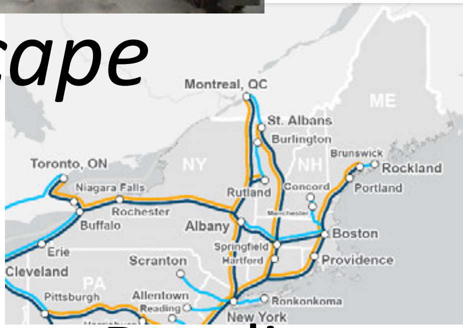
ScHARe Think-a-thon Preparing for AI 3: Computational Data Science Strategies 101

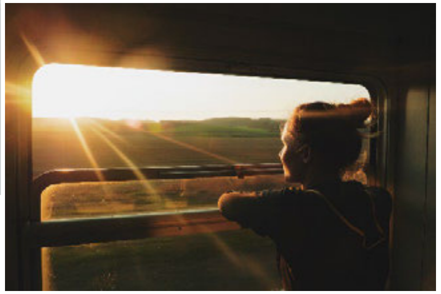
Ken Wilkins, PhD
Biostatistics Program, Office of Clinical Research
Data Science Working Group, Office of the Director
National Inst. of Diabetes & Digestive & Kidney Diseases, NIH

National Institute of Diabetes and Digestive and Kidney Diseases



Overview: *a whistlestop tour of a landscape*





- Understanding the Landscape
- Traditional Statistics & Epidemiologic Methods as Baseline
- Artificial Intelligence in Data Science as Broad New Horizon
- Machine Learning Unveiled as a Bridge-building Trailblazer
- Python Libraries for Data Science Computational Strategies
- Ongoing Resources and Decision-Making Tools to use as a Guide
- Q&A and Closing Remarks

Microsoft 365 Stock Photos; https://media.amtrak.com/wp-content/uploads/2021/05/Amtrak-2021-Corridor-Vision-May27_2021.pdf

Again a ‘whistlestop’ rather than a ‘whirlwind’ tour…

ScHARe

Science Collaborative for Health disparities and Artificial intelligence bias REduction

O'REILLY

A Whirlwind Tour of Python

Jake VanderPlas

Understanding the Landscape

A. Definitions and Differentiations

1) Preliminaries to get everyone on the same page
2) Context while getting our lay of the land: **health disparities**

B. Decision-Making Framework: early teaser… hard to decide *which* tools without a few things in toolbox

…will use above ‘alarm’ icon to trigger our need to “unpack” some ‘jargon’ terms

NIH National Institute of Diabetes and Digestive and Kidney Diseases

https://www.digitscotland.com/what-is-landscape-surveying-recording/

Understanding the Landscape: Preliminaries

- *Consider yourself as a data science practitioner: be* practical *on what to use!*
  - *“data science”: coin termed by a statistician, adopted by computer science/informatics*
  - *Most recently viewed as an ‘interdiscipline’ –interdisciplinary/metadisciplinary nature*
  - *‘practical’ means bringing the most effective tool(s) for the task(s) at hand*
  - *We cover computational strategies ranging from traditional to modern statistics and epidemiologic methods, and where these don’t meet needs: AI & machine learning*
  - We cover working definitions of above, ahead of diving in… but we also bear in mind…

- Context of ScHARe goals of working toward **health disparities** (*primal aim*)
  - *“The aim of the ScHARe program is to increase participation of people from underrepresented populations in data science and cloud computing so that everyone can benefit from the research opportunities afforded by Big Data.”*

Understanding the Landscape: Preliminaries

- *Consider yourself as a data science practitioner: be a* scientist in *what you do!*
  - *"data science": science as the practice of adding to 'generalizable knowledge'*
  - *Scientists ought to maintain awareness of their 'blind spots': tacit assumptions in data*
  - *Consider how you must check your assumptions… how did data come to be at hand?*
  - *This 'design behind the data' hearken back to 'Research Design' of prior TaT session*
  - We cover working definitions of above, ahead of diving in… but we also bear in mind…
- Context of ScHARe **aims**
  - *Increase participation of women and underrepresented populations with health disparities in data science through data science skills training, cross-discipline mentoring, and multi-career level collaborating on research.*
  - *Leverage population science, SDOH, and behavioral Big Data and cloud computing tools to foster a paradigm shift in health disparity, and health and healthcare delivery outcomes research.*
  - *Advance AI bias mitigation and ethical inquiry by developing innovative strategies and securing diverse perspectives.*

Getting our lay of the land: **health disparities**

the lay of the land noun phrase **(US idiom)**

: the arrangement of the different parts in an area of land **:** where things are located in a place - She knew the *lay of the land* from hiking through it daily.

—often used figuratively

It takes time for new employees to get *the lay of the land* in this department.

https://www.merriam-webster.com/dictionary/the%20lay%20of%20the%20land

- Context of ScHARe goals
- **Decreasing Health Disparities – 'dual' problem of mitigating extant biases**

  The primal and dual are two sides of the same coin, with the primal being the original problem and the dual being the derived problem.
- *Mitigating Bias: does it mean the same thing to all parties?*
  - *not necessarily: varied forms of each type of 'bias' ought to be considered*
    - *Bias in perspective/experience (confirmation bias), bias in data available (selection bias), &c.*
  - *Theoretical behavior of data methods: 'bias' if estimates differ from target*
    - *Often referred to as 'statistical bias' – follows from any quantity derived from data being a 'statistic'*
  - *Practical applications to data: inherent imbalances of data's sources →* ***algorithmic bias***
    - *One distinction as written by AI/ML researchers: "In contrast to human bias, algorithmic bias occurs when an AI model, trained on a given data set, produces results that may be completely unintended by the model creators." – Chen, Szolovits, & Ghassemi 2019, AMA Journal of Ethics*

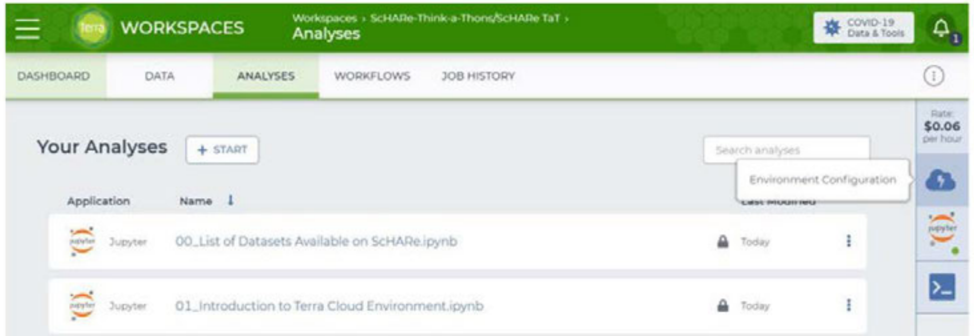
Getting our lay of the land: **health disparities**

- Context of ScHARe goals, while getting our lay of the land: health disparities
- *As a data scientist, you can have* ***agency*** *in some sources of bias*
  - If you lack individual-level features that 'explain' source of bias, use *supplements*
  - *Supplements easier to get with data linkage (e.g., ZIP code for area-level proxies)*
  - ***Ultimately:*** *some features need careful prep, others will be 'missing' (still recognize)*
    - ***Data prep:* numeric** *form of features used in algorithms, possible '***weighting***' for missed features*
    - ***Teaser of decision-making framework:*** *can't decide tools to use without actual toolbox… ScHARe@Terra*
    - ***NOTE:*** today will NOT involve live hands-on work
      - ***We have a lot to cover conceptually, prior to coding***
      - ***Concepts can be reinforced by experiential learning***

Experiencing
Reflecting
Thinking
Acting

bit.ly/schare-tat

If you have already created a Terra account and are logged in, you will see this:



WORKSPACES

Workspaces › ScHARe-Think-a-Thons/ScHARe TaT ›
Analyses

Your Analyses

Application | Name | Last Modified

Jupyter | 00_List of Datasets Available on ScHARe.ipynb | Today

Jupyter | 01_Introduction to Terra Cloud Environment.ipynb | Today