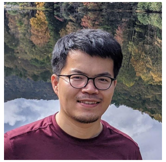
Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 3: Speakers

Amy McGovern (OU)

Christopher Wirz (NCAR)

Andrea Schumacher (CIRA)


Douglas Rao (NOAA)

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

NSF

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 3: Goals

- Understand the importance of trustworthy data and workflows
- Learn why case studies are useful and how to select them

NSF

ARTIFICIAL INTELLIGENCE · ENVIRONMENTAL SCIENCE · RISK COMMUNICATION

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR
ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 3: Agenda

- 9:00 Trustworthiness of data and workflows
- 10:30 *Brain & bio break*
- 10:45 The importance of case studies and tips for using them effectively

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do
and use the
code TAI4ES

NSF

ARTIFICIAL INTELLIGENCE · ENVIRONMENTAL SCIENCE · RISK COMMUNICATION

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR
ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Warm-up and refresher from yesterday

Let’s do couple quick questions to get us back in the trustworthy AI mindset:

1. What words/phrases would you use to describe “**explainable** AI?”

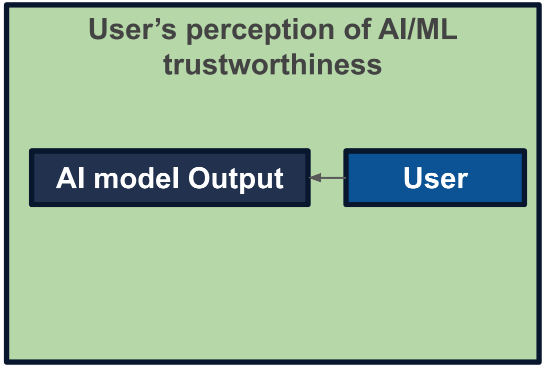
2. What was your favorite part of yesterday’s lectures?

3.1. & 3.2. Go to sli.do and use the code TAI4ES

Data sets and workflows

Why do we care about trustworthy data and workflows?

Reminder of the AI2ES Definition: Trustworthiness is a trustor's evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted.



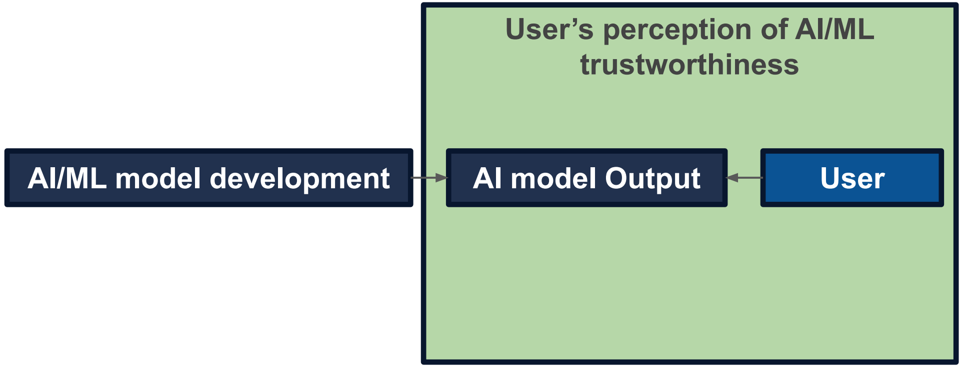
User's perception of AI/ML trustworthiness

AI model Output

User

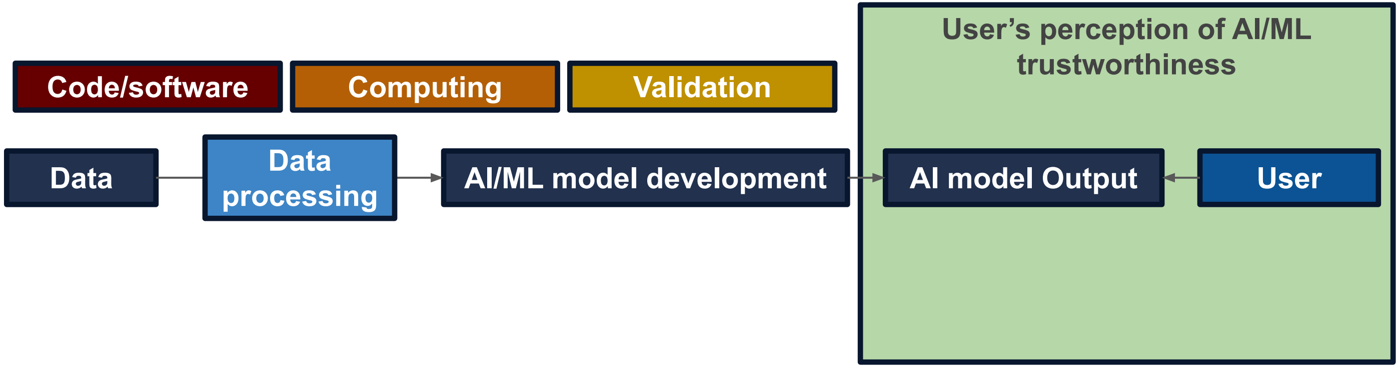
Why do we care about trustworthy data and workflows?

Reminder of the AI2ES Definition: Trustworthiness is a trustor's evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted.



Why do we care about trustworthy data and workflows?

Reminder of the AI2ES Definition: Trustworthiness is a trustor's evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted.



User's perception of AI/ML trustworthiness

Code/software

Computing

Validation

Data

Data processing

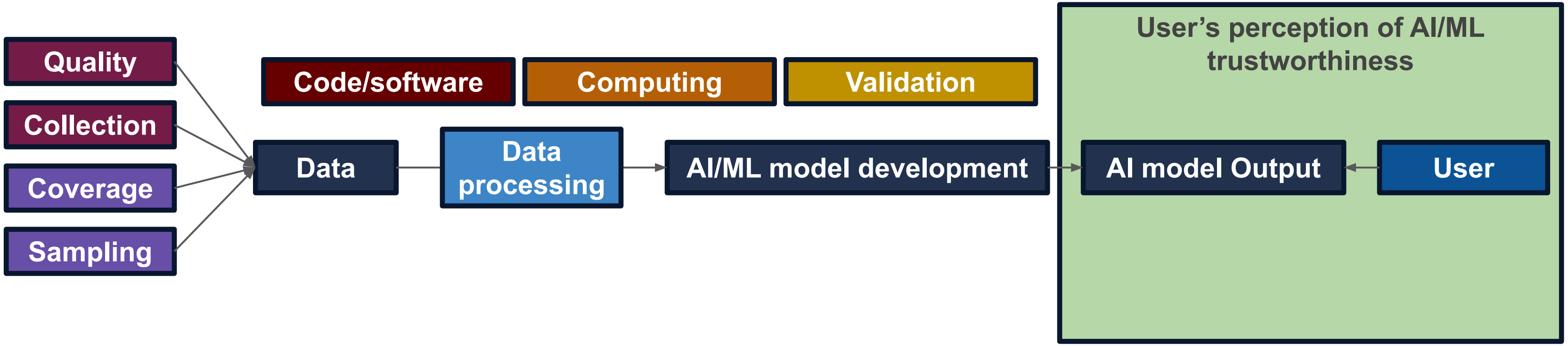
AI/ML model development

AI model Output

User

Why do we care about trustworthy data and workflows?

Reminder of the AI2ES Definition: Trustworthiness is a trustor’s evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted.



Quality

Collection

Coverage

Sampling

Code/software

Computing

Validation

Data

Data processing

AI/ML model development

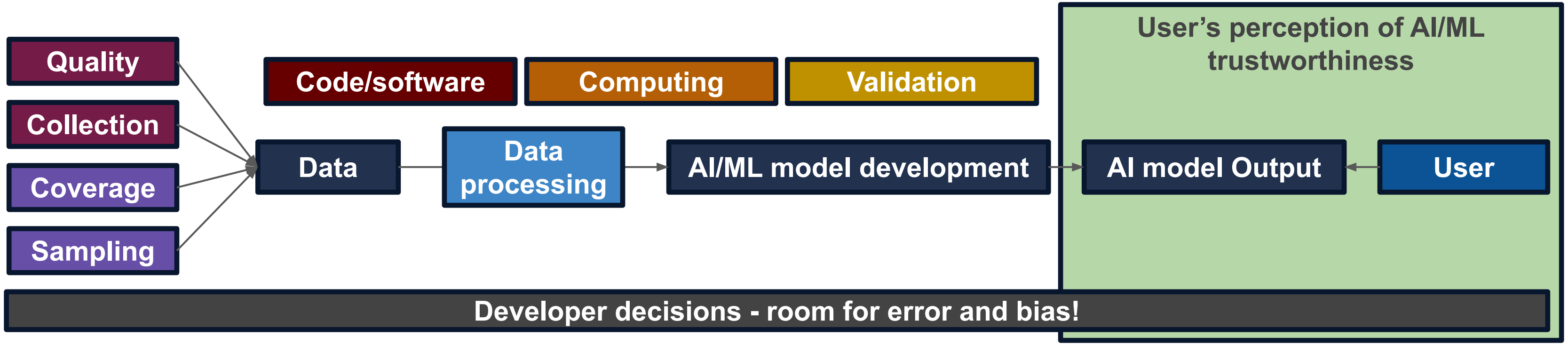
User’s perception of AI/ML trustworthiness

AI model Output

User

Why do we care about trustworthy data and workflows?

Reminder of the AI2ES Definition: Trustworthiness is a trustor’s evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted.



Quality

Collection

Coverage

Sampling

Code/software

Computing

Validation

Data

Data processing

AI/ML model development

User’s perception of AI/ML trustworthiness

AI model Output

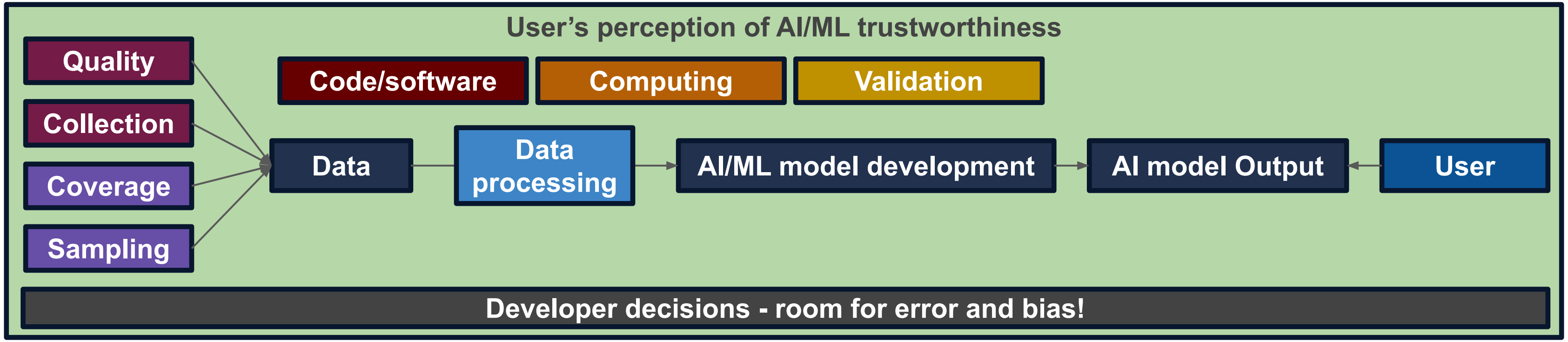
User

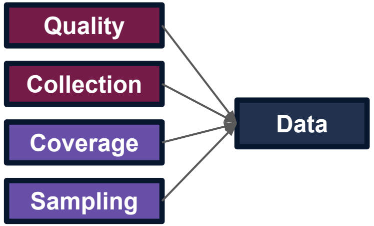
Developer decisions - room for error and bias!

Why do we care about trustworthy data and workflows?

Reminder of the AI2ES Definition: Trustworthiness is a trustor’s evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted.

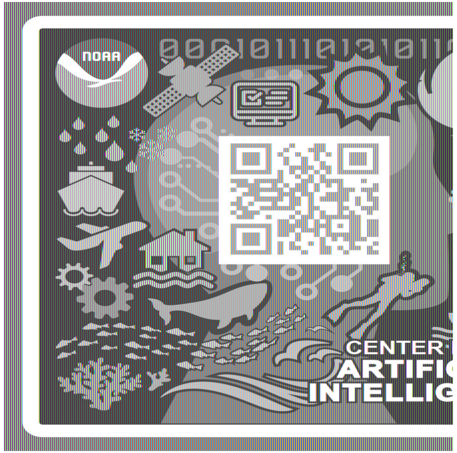
Because the whole process can impact trustworthiness



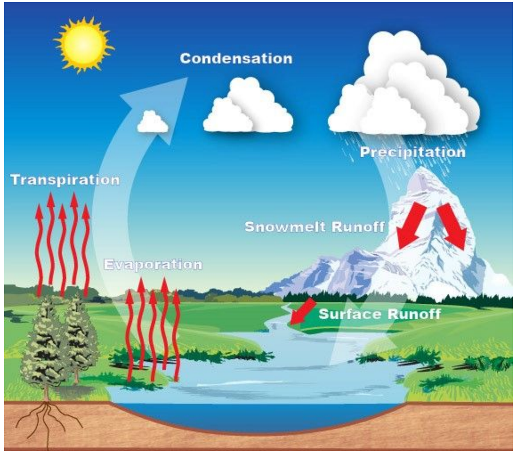


Understand your data source

Observational data

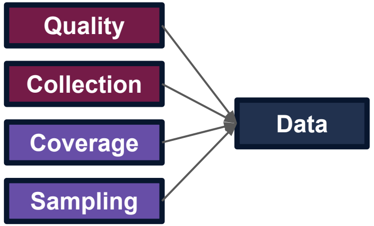


Simulated/synthetic data



Crowd-sourced data

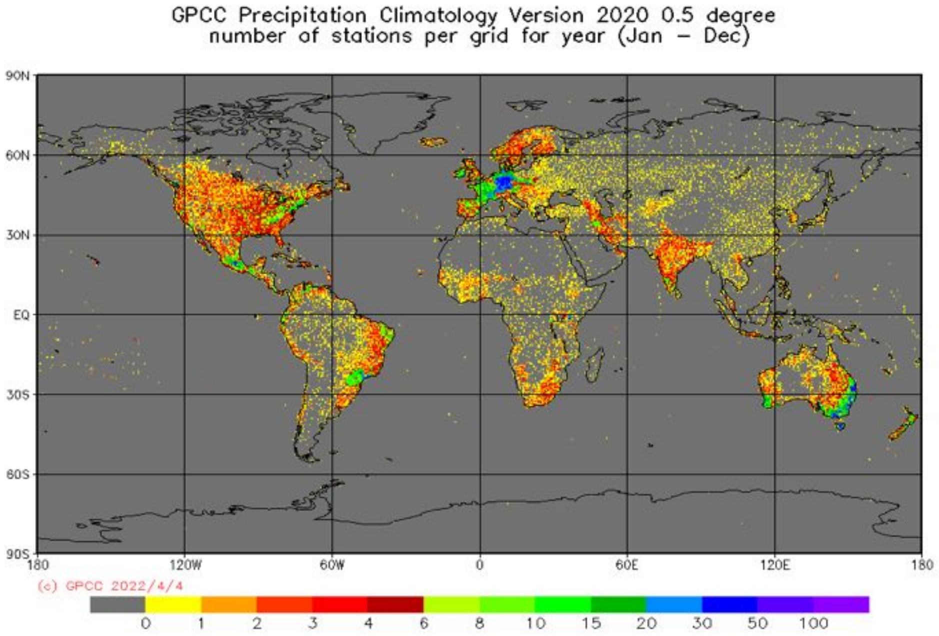
AI in environmental applications takes a diverse set of data in the development and evaluation process. Building trust should start from accounting for the quality and limitations of these data sources.



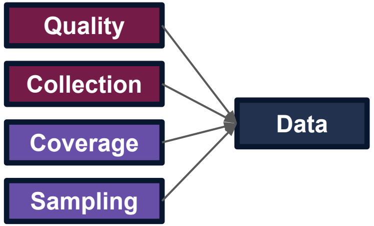
Observational Data: *in situ*

Environmental AI applications often rely on in situ observation data to train or evaluate the model.

1) How representative (spatially and/or temporally) are these data?
2) Are there any systematic bias/error of these data?
3) What is the uncertainty of these data?


GPCC Precipitation Climatology Version 2020 0.5 degree
number of stations per grid for year (Jan – Dec)
90N
60N
30N
EQ
30S
60S
90S
180
120W
60W
0
60E
120E
180
(c) GPCC 2022/4/4
0
1
2
3
4
6
8
10
15
20
30
50
100

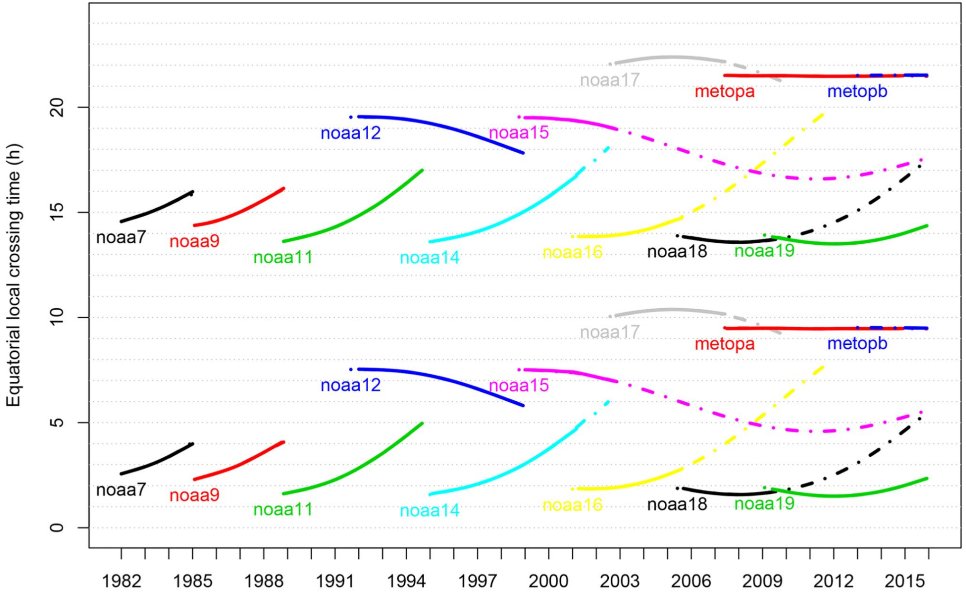
Source: Global Precipitation Climatology Centre



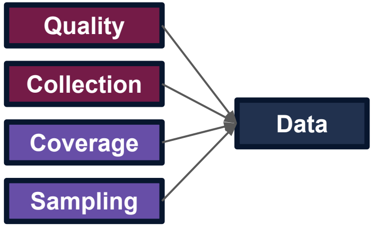
Observational Data: satellite

Satellite data are another set of observational data often used. You should consider:

1) What is the quality of single-sensor data?
2) How consistent are data from multiple sensors? (E.g., satellite orbital drift for climate applications, see right figure)?
3) Is the satellite observation the same as what you want?



Bojanowski et al (2022) https://doi.org/10.5194/amt-13-6771-2020



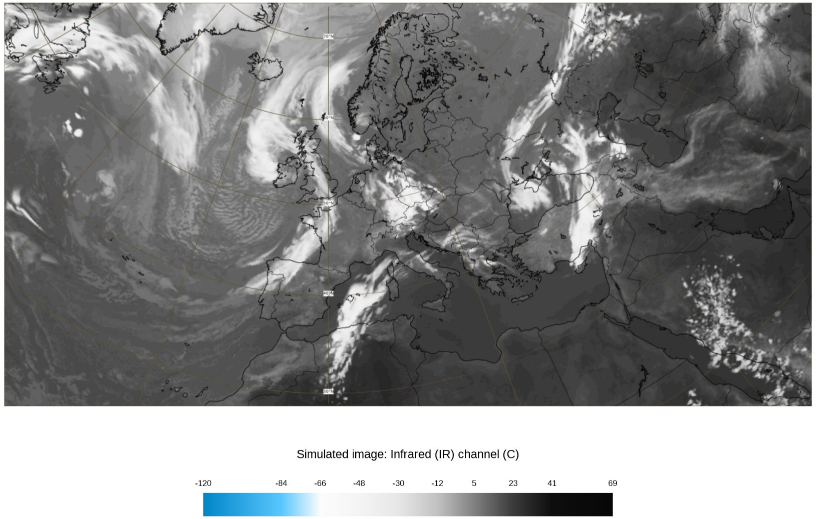
Simulated/Synthetic Data

Why use simulated/synthetic data?

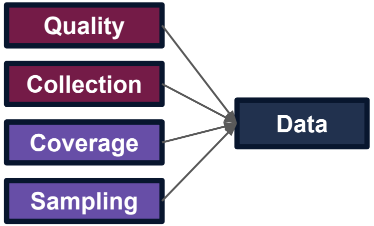
- Real data can be incomplete or inaccessible
- Real data cannot be directly used (due to restrictions such as privacy)
- Common example:
  - Radiative transfer model simulation to simulate satellite data for retrieval algorithm development
  - Reanalysis data (e.g., ERA-5)
  - Climate model simulations (e.g., CMIP6)
  - Large eddy simulation (LES)

Simulated images - infrared

Base time: Thu 23 Jun 2022 00 UTC Valid time: Fri 24 Jun 2022 21 UTC (+45h) Area : Europe



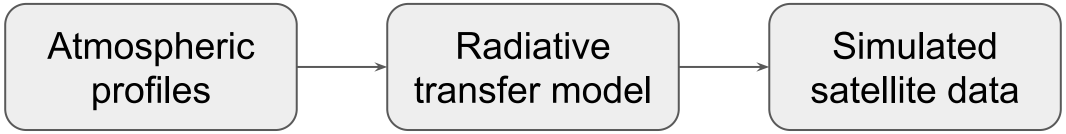
ECMWF



Simulated/Synthetic Data

Things to consider re: simulated data

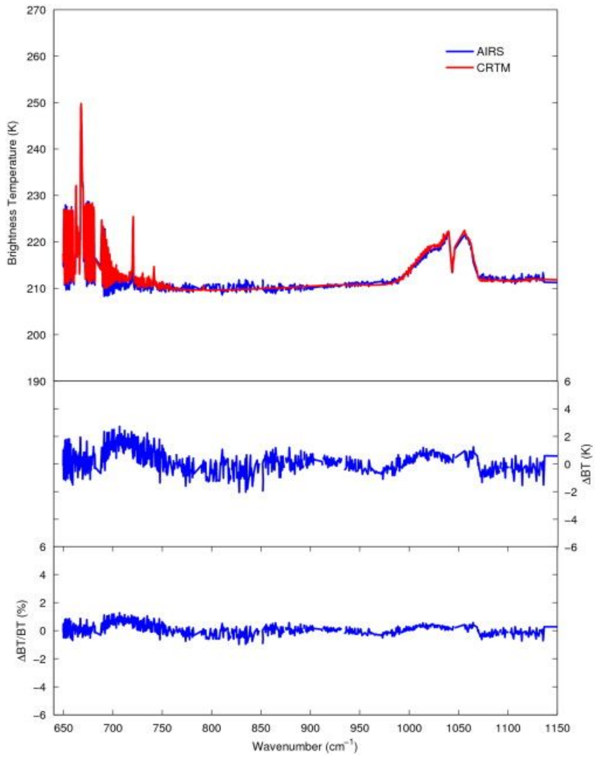
- There is an algorithm behind the simulated/synthetic data (including input and output)



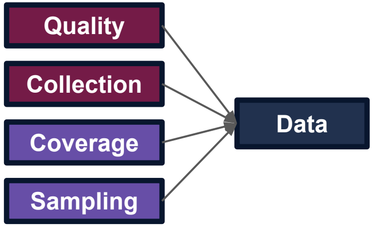
Input quality

Model accuracy

Final quality



Ding et al. (2011) Validation of the community radiative transfer model. https://doi.org/10.1016/j.jqsrt.2010.11.009



Quality

Collection

Coverage

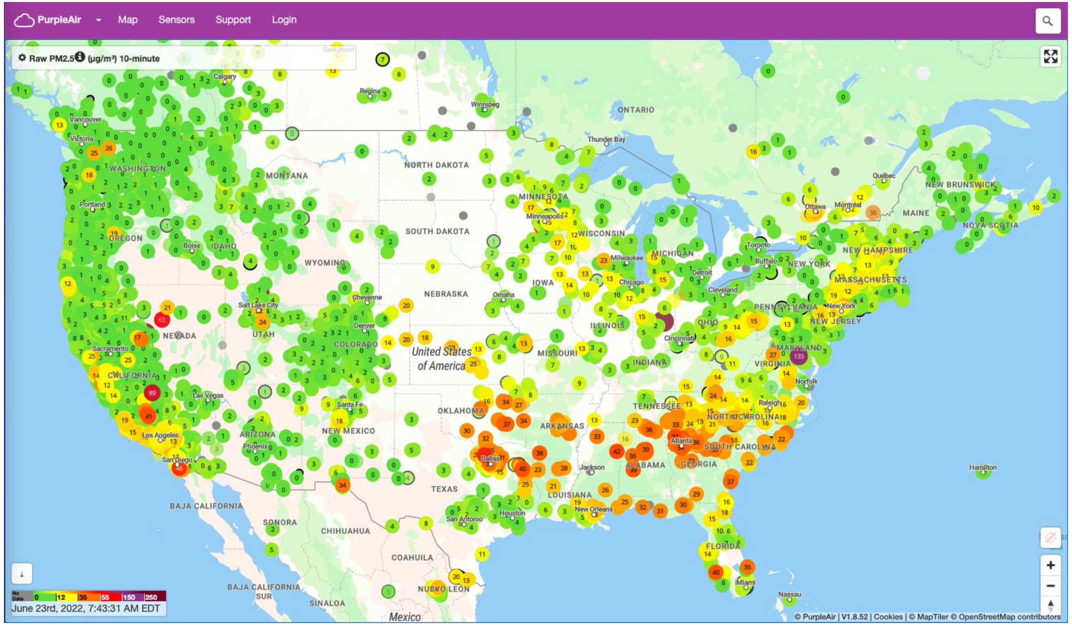
Sampling

Data

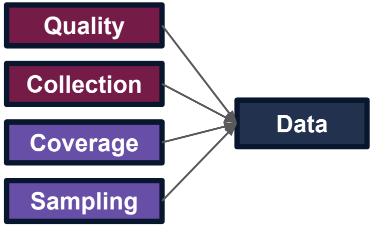
Crowd-Sourced Data

Crowd-sourced data provides unique opportunity to fill the data gap in traditional data collection methods, e.g.,

- PurpleAir (air quality)
- CoCoRaHS (precipitation)
- NASA GLOBE
- NOAA Urban Heat Mapping



Real time PM2.5 data from the PurpleAir sensor network.



Quality

Collection

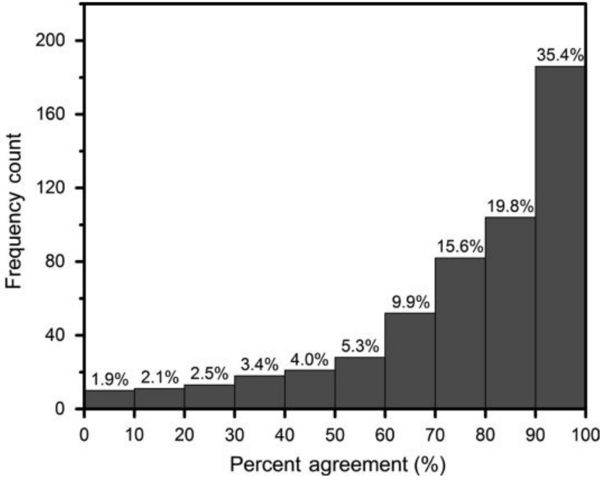
Coverage

Sampling

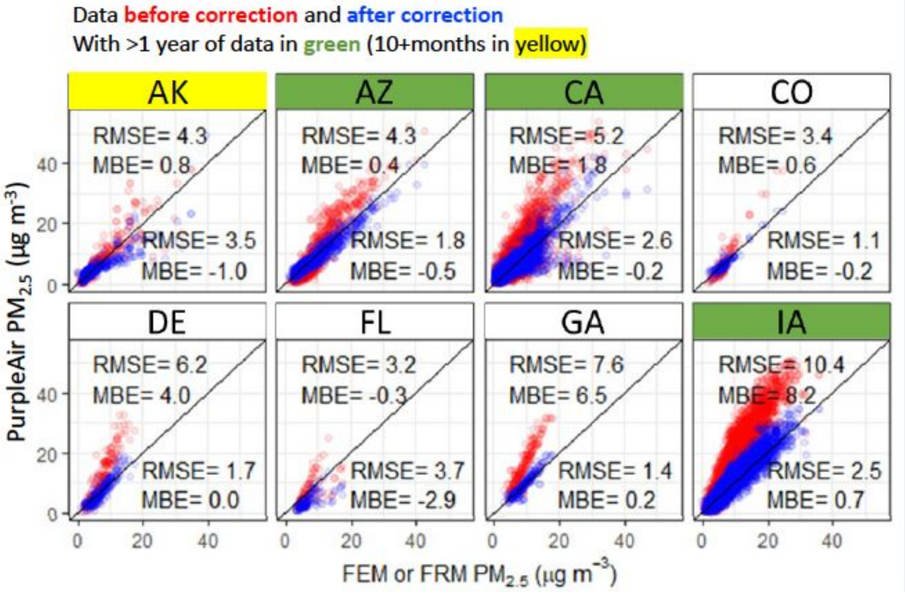
Data

Crowd-Sourced Data

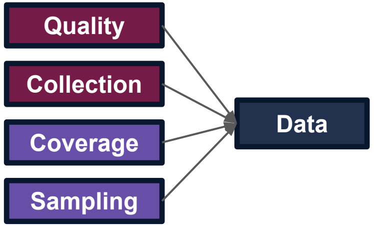
Challenging to establish consistent data quality for crowd-sourced data.



Aceves-Bueno et al. (2017) The Accuracy of Citizen Science Data: A Quantitative Review. doi: 10.1002/bes2.1336

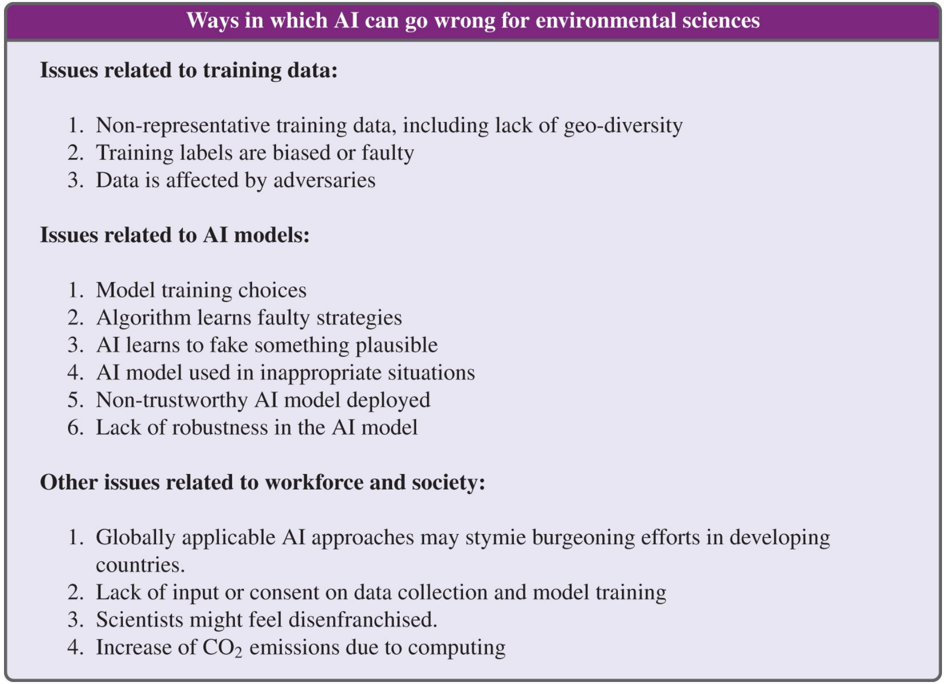


Credit: Karoline Barkjohn (https://cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=CEMM&dirEntryId=350379)



Data, Trust, and Ethical AI

- AI needs to be used and created in an ethical manner for all applications
- Poor training data can create biased or faulty models
    - Garbage in -> Garbage out
- Unethical and biased models should not be trusted!



Ways in which AI can go wrong for environmental sciences

Issues related to training data:

1. Non-representative training data, including lack of geo-diversity
2. Training labels are biased or faulty
3. Data is affected by adversaries

Issues related to AI models:

1. Model training choices
2. Algorithm learns faulty strategies
3. AI learns to fake something plausible
4. AI model used in inappropriate situations
5. Non-trustworthy AI model deployed
6. Lack of robustness in the AI model

Other issues related to workforce and society:

1. Globally applicable AI approaches may stymie burgeoning efforts in developing countries.
2. Lack of input or consent on data collection and model training
3. Scientists might feel disenfranchised.
4. Increase of CO_2 emissions due to computing

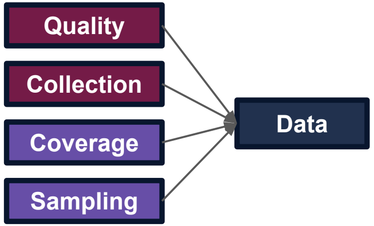
McGovern, A., Ebert-Uphoff, I., Gagne, D., & Bostrom, A. (2022). Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environmental Data Science, 1*, E6. doi:10.1017/eds.2022.5

Non-representative training data

- Rare events:
  - Tornadoes, turbulence, hail, many extreme phenomena
  - Data collection is challenging
- Non-uniform sensors
  - Air pollution more prevalent in affluent areas / countries
- Remote mountain areas, and oceans, might not be well represented
  - Many sensors use visible light (cameras, visible bands of satellites)
- Phenomena not well represented at night
  - E.g. Tropical cyclones

By Justin1569 at English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=5943918

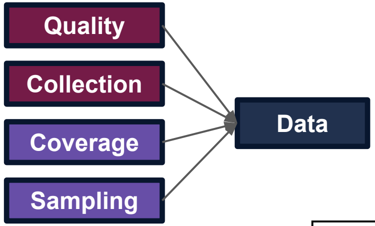
https://www.jetphotos.com/photo/10352183


Quality
Collection
Coverage
Sampling
Data

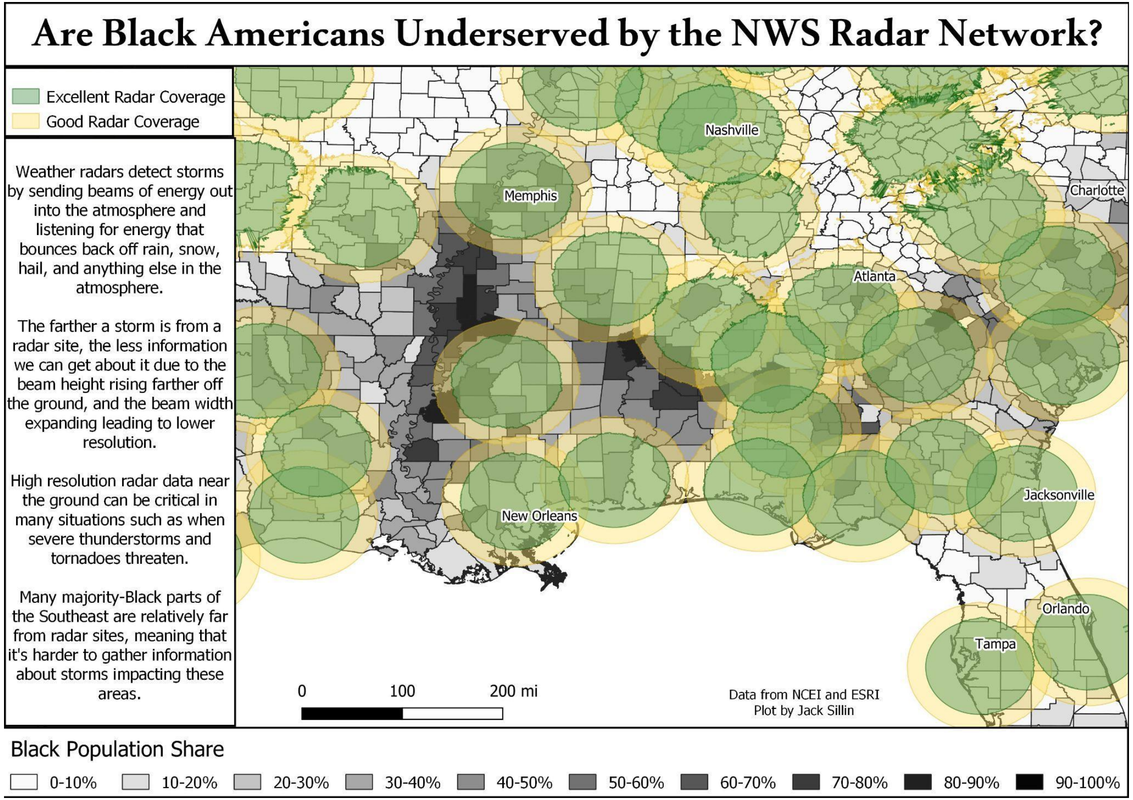
Non-representative training data

- Climate change is both rare & unprecedented and non-uniformly sensed
    - Climate change is altering the frequency of extreme phenomena
- ML model learns from data
    - Those scenarios are then under-represented in ML model as well
- Could result in unintentional biases in climate models
- Can result in unintentional environmental injustice.

https://www.science.org/content/article/europe-s-deadly-floods-leave-scientists-stunned



Biased data:
Geographic Under-sampling



Are Black Americans Underserved by the NWS Radar Network?

Weather radars detect storms by sending beams of energy out into the atmosphere and listening for energy that bounces back off rain, snow, hail, and anything else in the atmosphere.

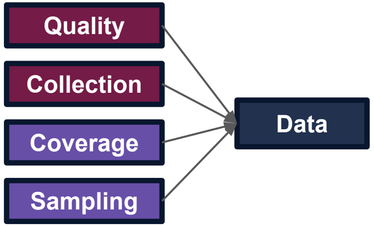
The farther a storm is from a radar site, the less information we can get about it due to the beam height rising farther off the ground, and the beam width expanding leading to lower resolution.

High resolution radar data near the ground can be critical in many situations such as when severe thunderstorms and tornadoes threaten.

Many majority-Black parts of the Southeast are relatively far from radar sites, meaning that it's harder to gather information about storms impacting these areas.

Data from NCEI and ESRI
Plot by Jack Sillin

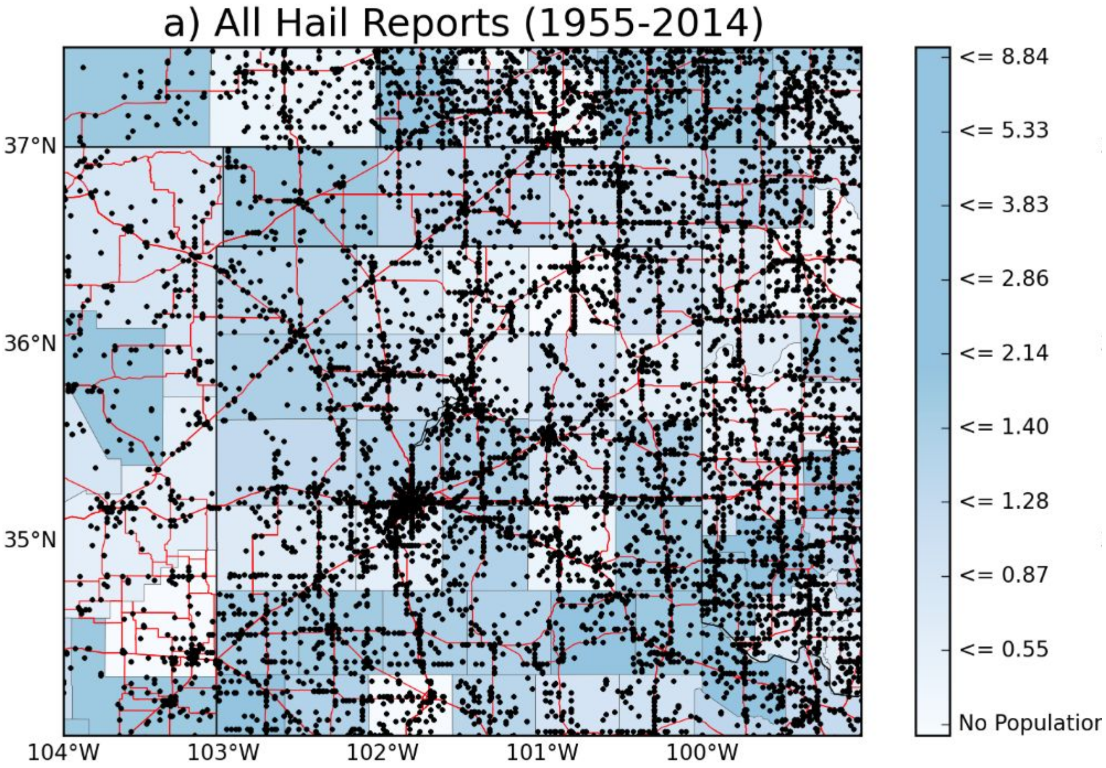
From Jack Sillin @JackSillin:
https://twitter.com/JackSillin/status/1372957704138981378?s=20



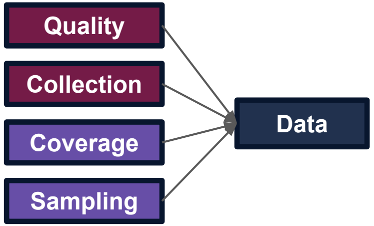
Biased data:
Human-created biases

Relying on human labeled data can create unintentional biases

- It only hails where there are people apparently!


a) All Hail Reports (1955-2014)

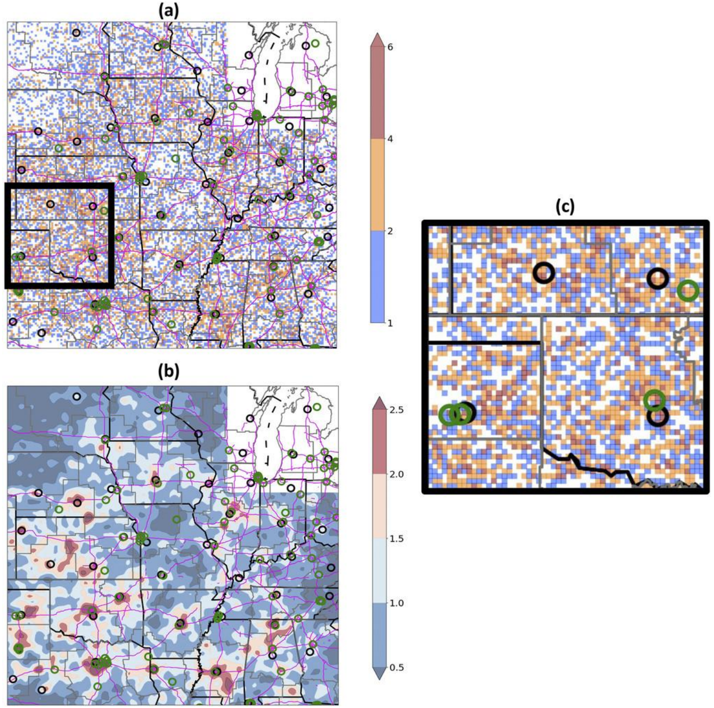
Allen, J. T., and M. K. Tippett, 2015: The characteristics of United States hail reports: 1955–2014. Electronic J. Severe Storms Meteor., 10 (3), 1–31.



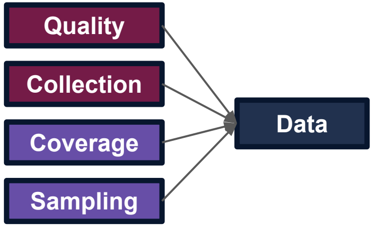
Biased data: Human-created biases

Relying on human labeled data can create unintentional biases

- It only has tornadoes where there are people also?



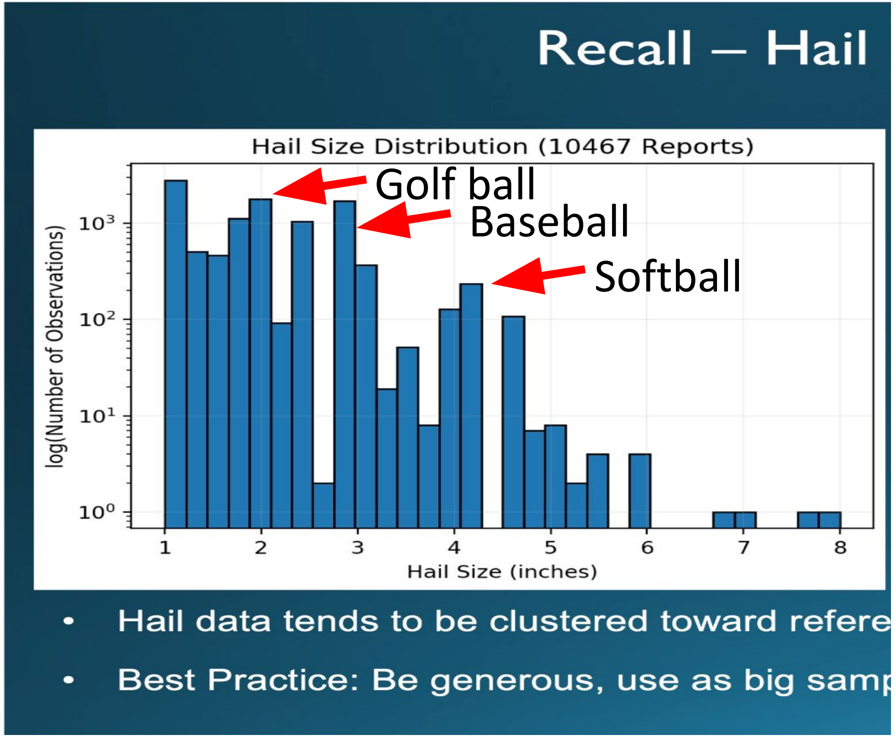
Potvin, C. K., Broyles, C., Skinner, P. S., Brooks, H. E., & Rasmussen, E. (2019). A Bayesian Hierarchical Modeling Framework for Correcting Reporting Bias in the U.S. Tornado Database, *Weather and Forecasting*, *34*(1), 15-30. Retrieved Jul 24, 2021



Biased data:
Human-created biases

Relying on human labeled data can create unintentional biases

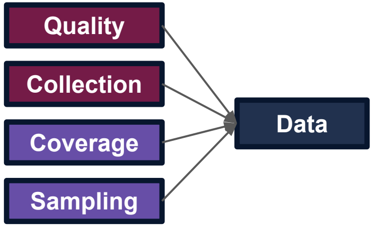
- Human labels can be wrong and the distribution discrete rather than continuous
- Hail size is continuous yet people cluster labels to common objects



Recall — Hail

- Hail data tends to be clustered toward refere
- Best Practice: Be generous, use as big samp

Allen, J. T., M. R. Kumjian, C. J. Nixon, R. E. D. Jewell, B. T. Smith, and R. L. Thompson, 2020: Forecast Parameters for U.S. Hail Occurrence and Size. *30th Conference on Weather Analysis and Forecasting (WAF)/26th Conference on Numerical Weather Prediction (NWP), AMS 100th Annual Meeting, Boston, MA.*



Quality

Collection

Coverage

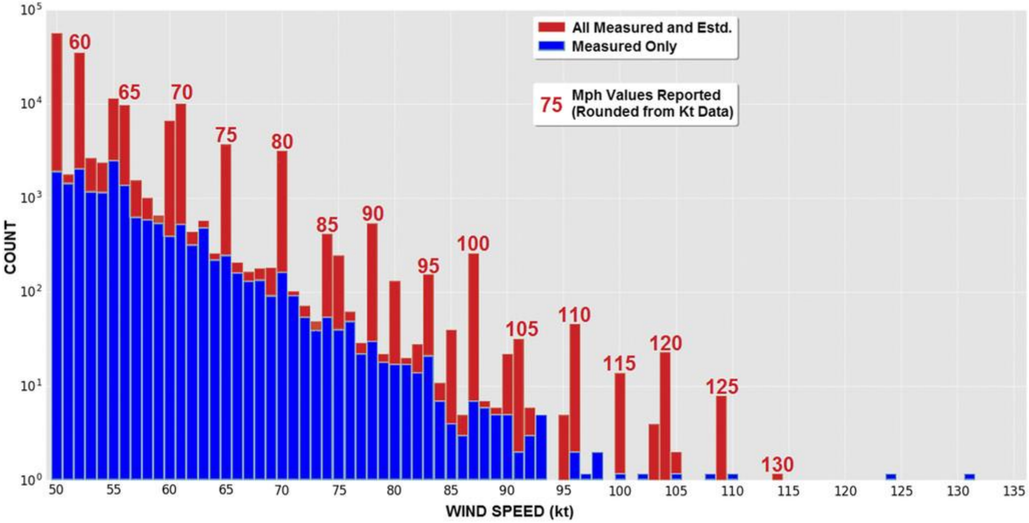
Sampling

Data

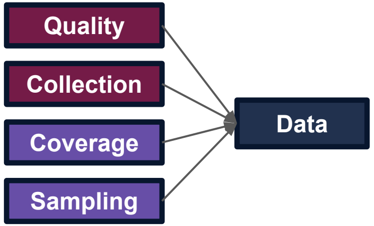
Biased data:
Human-created biases

Relying on human labeled data can create unintentional biases

- Human labels can be wrong and the distribution discrete rather than continuous
- Wind measured in convenient 5 mph bins



Edwards, R., Allen, J. T., & Carbin, G. W. (2018). Reliability and Climatological Impacts of Convective Wind Estimations, *Journal of Applied Meteorology and Climatology*, *57*(8), 1825-1845



Quality

Collection

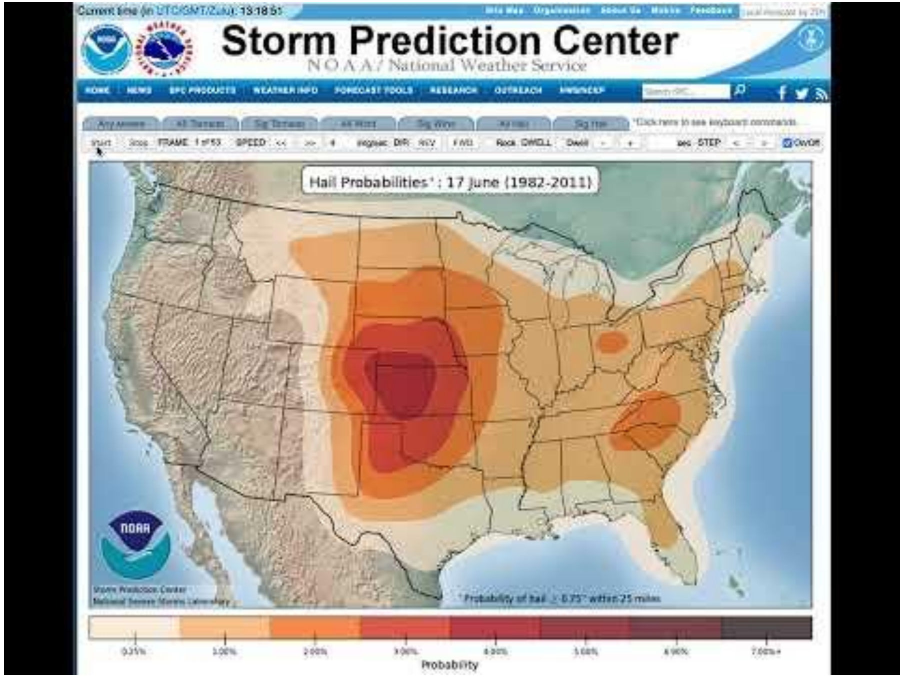
Coverage

Sampling

Data

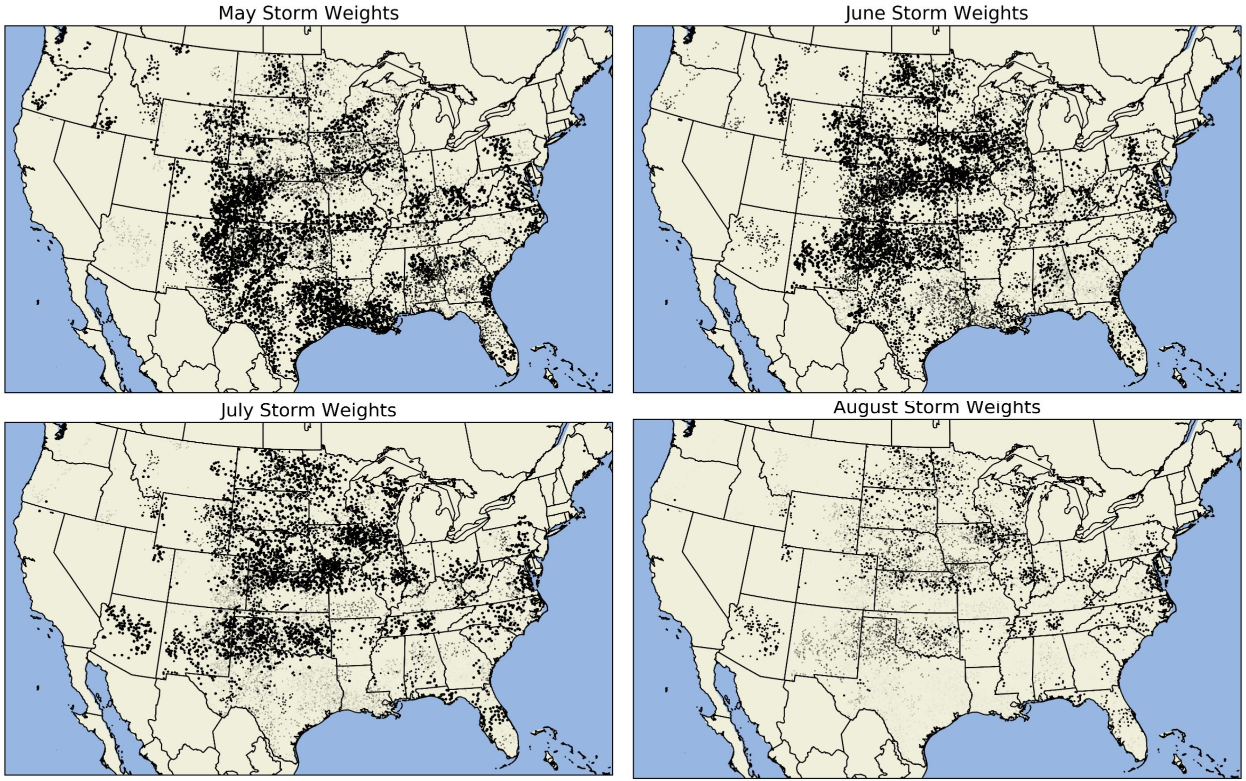
Biased data:
Temporal or seasonal biases

- Often target phenomena is seasonal
  - If you train on all the data, you miss the seasonal biases that your model should produce
- Example:
  - Hail risk areas move around the US by season



Quick example of adjusting for seasonal biases

- Idea: weight the examples by their season
  - Train model on all examples (rare phenomena/limited data) but weight by season of interest



May Storm Weights

June Storm Weights

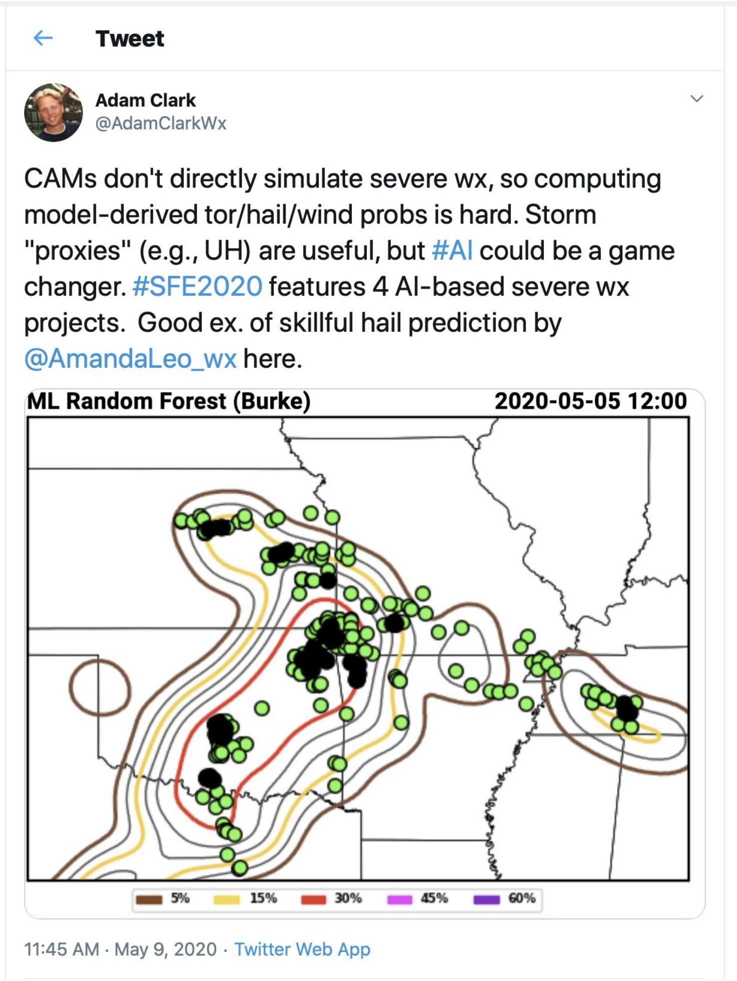
July Storm Weights

August Storm Weights

Burke et al (under review) and Burke, Amanda; McGovern, Amy; Gagne II, David John; Snook, Nathan (2020) Temporally Weighting Machine Learning Models for High-Impact Severe Hail Prediction AI for Earth Sciences Workshop at NeurIPS 2020.

Result of seasonal adjustment

- New spatial weighted tested in NOAA's Hazardous Weather Testbed in Spring 2020 (all virtual)
- "AI could be a game changer" – Adam Clark of the Storm Prediction Center
- Result: Improved trust and a better model!



Tweet

Adam Clark
@AdamClarkWx

CAMs don't directly simulate severe wx, so computing model-derived tor/hail/wind probs is hard. Storm "proxies" (e.g., UH) are useful, but #AI could be a game changer. #SFE2020 features 4 AI-based severe wx projects. Good ex. of skillful hail prediction by @AmandaLeo_wx here.

11:45 AM · May 9, 2020 · Twitter Web App

Adversarial Data

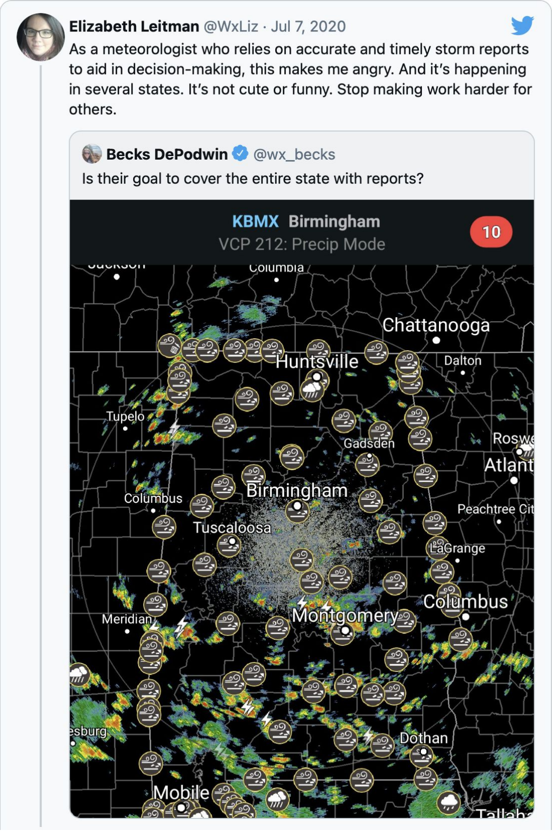
Adversarial data can affect ML models

- **Crowd-sourced data can be hacked or deliberately falsified**
- **False damage/storm reports**
- Insurance fraud

"IT HAS REAL LIFE REPERCUSSIONS. IN THIS CASE IT DID NOT RESULT IN SOMEBODY GETTING HURT, WHICH IS GREAT. BUT WHEN YOU GET TOO MANY OF THESE FALSE REPORTS, IT RESULTS IN THE DEGRADATION OF OUR WARNINGS. WE HAVE TO MAKE A WARNING DECISION IN SECONDS."

Dennis Cavanaugh, warning and coordination meteorologist, NWS in North Little Rock, AR

https://discovertornadoes.com/2022/04/20/ohio-woman-files-5-false-arkansas-storm-reports-tornado-report-included/



Elizabeth Leitman @WxLiz · Jul 7, 2020

As a meteorologist who relies on accurate and timely storm reports to aid in decision-making, this makes me angry. And it's happening in several states. It's not cute or funny. Stop making work harder for others.

Becks DePodwin @wx_becks

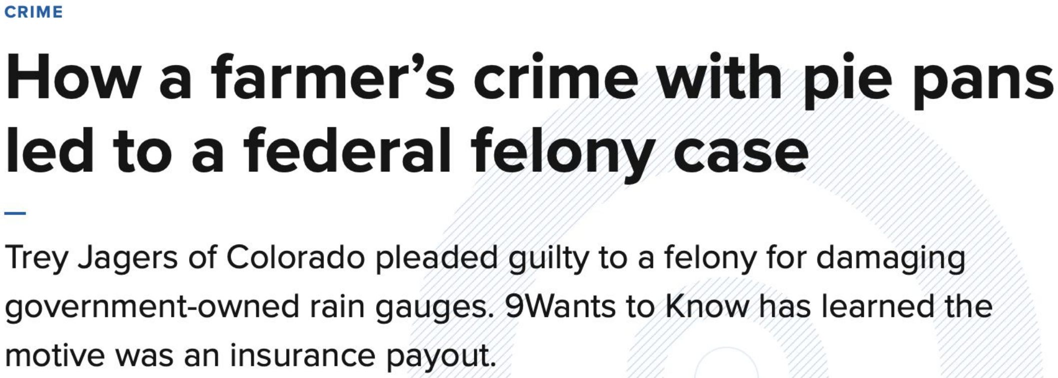
Is their goal to cover the entire state with reports?

https://www.washingtonpost.com/weather/2020/07/14/noaa-app-mping-suspended/

Adversarial Data

Adversarial data can affect ML models

- Crowd-sourced data can be hacked or deliberately falsified
- False damage/storm reports
- **Insurance fraud**



CRIME

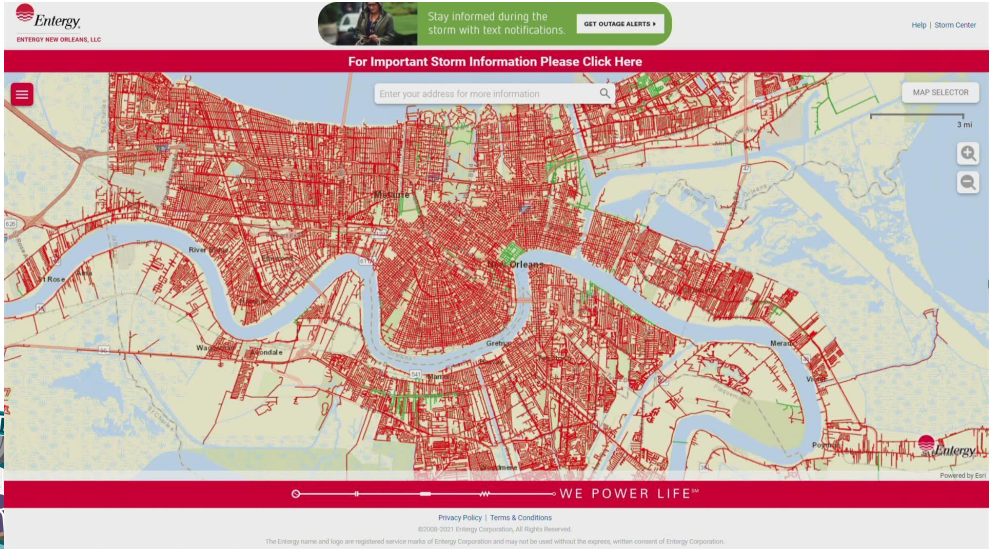
How a farmer’s crime with pie pans led to a federal felony case

Trey Jagers of Colorado pleaded guilty to a felony for damaging government-owned rain gauges. 9Wants to Know has learned the motive was an insurance payout.

https://www.9news.com/article/news/crime/colorado-farmer-pie-tins-rain-gauges-federal-felony/73-1638f8a4-967a-4f12-af11-d6903e8b5d0d

Adversarial Data

- Weather is also an adversary!
  - Power outages
  - Destroying sensors



Entergy

ENTERGY NEW ORLEANS, LLC

Stay informed during the storm with text notifications.

GET OUTAGE ALERTS ›

Help | Storm Center

For Important Storm Information Please Click Here

Enter your address for more information

MAP SELECTOR

WE POWER LIFE℠

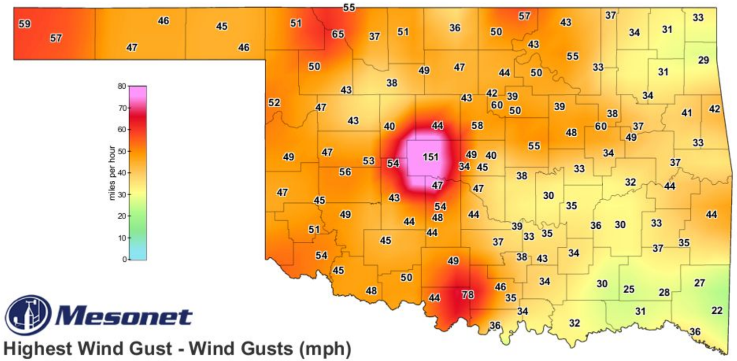
Powered by Esri

Privacy Policy | Terms & Conditions

©2008-2021 Entergy Corporation, All Rights Reserved.

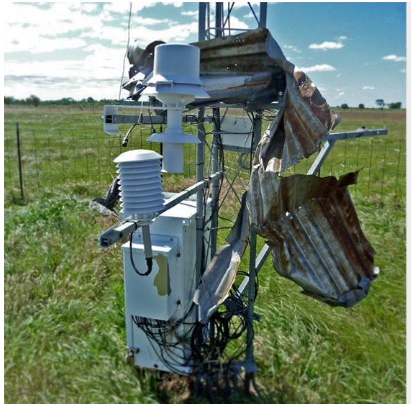
The Entergy name and logo are registered service marks of Entergy Corporation and may not be used without the express, written consent of Entergy Corporation.

https://www.wwltv.com/article/weather/hurricane/widespread-power-outages-reported-9400-in-the-dark/289-d8a78748-9a37-4937-90af-0d2d7cb3fbd6



Highest Wind Gust - Wind Gusts (mph)

May 24, 2011 - El Reno (151 mph)



El Reno Mesonet Station after recording a wind gust of 151mph on May 24, 2011

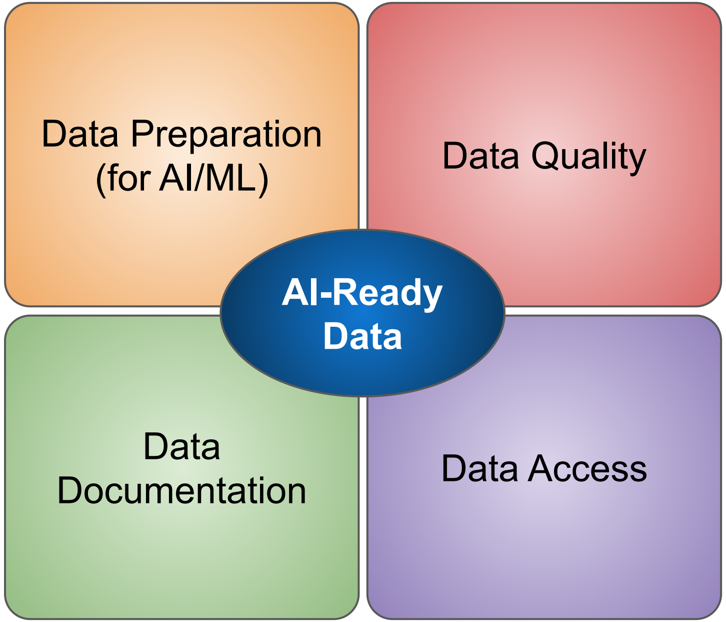
https://www.mesonet.org/20th/

Implications for workflow:
What to do when you’re working with secondary data?

There is community-driven effort to define “AI-ready data” that provides accessible, well-documented, and reusable open environmental data for AI-applications.

Data quality factors to consider for ***YOUR*** use cases:

- Bias/accuracy
- Completeness/coverage
- Resolution/frequency
- Consistency
- Timeliness



Data Preparation (for AI/ML)

Data Quality

AI-Ready Data

Data Documentation

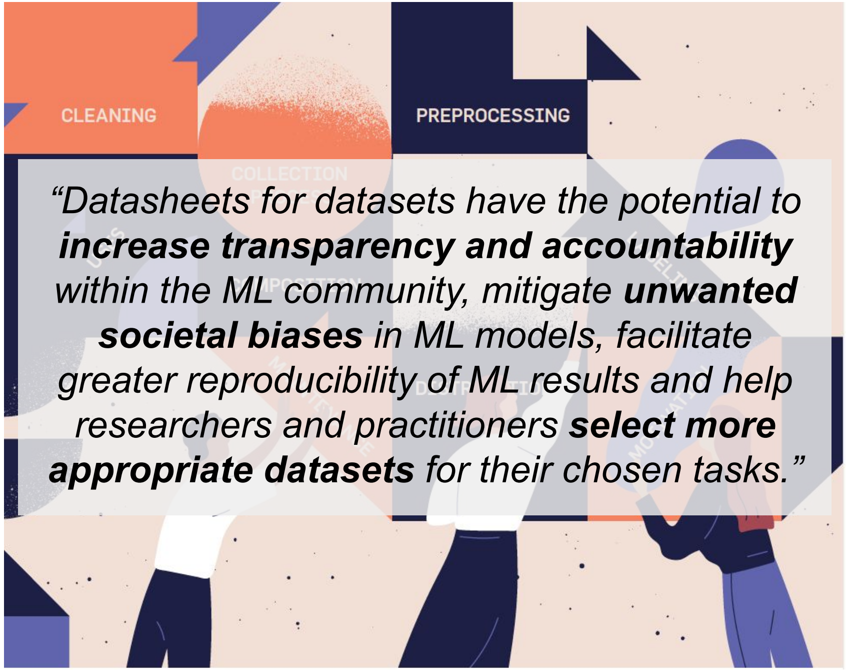
Data Access

ESIP Data Readiness Cluster (2022): Checklist to Examine AI-readiness for Open Environmental Datasets. https://doi.org/10.6084/m9.figshare.19983722.v1

Datasheets for datasets

“Documentation to facilitate communication between dataset creators and consumers.”

- Motivation (***why & who*** created it)
- Composition (***what*** is in it)
- Collection process (***how*** it is created)
- Preprocessing/cleaning/labeling (***provenance***)
- Uses (***intended & not-suitable***)
- Distribution (***who*** can use it)
- Maintenance (***dataset sustainability***)

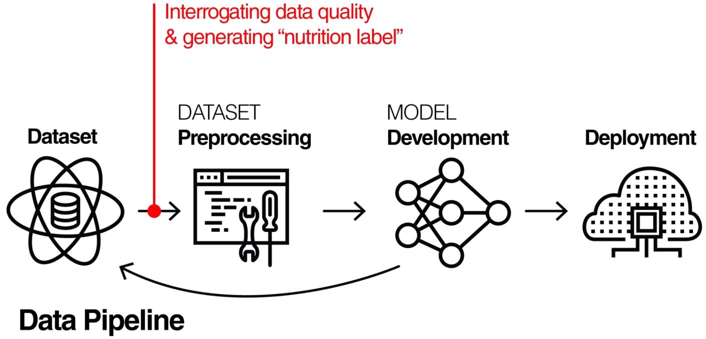


“Datasheets for datasets have the potential to ***increase transparency and accountability*** *within the ML community, mitigate* ***unwanted societal biases*** *in ML models, facilitate greater reproducibility of ML results and help researchers and practitioners* ***select more appropriate datasets*** *for their chosen tasks.”*

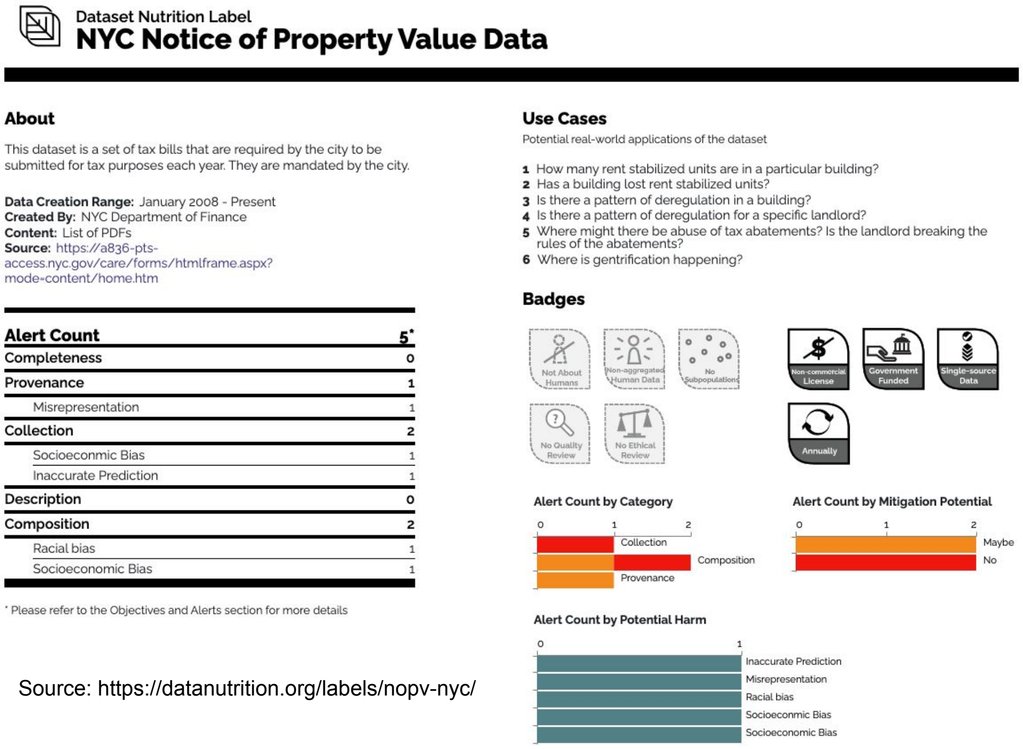
Gebru et al. (2018) Datasheets for Datasets. https://arxiv.org/abs/1803.09010

Dataset Nutrition Project

Clear communication of the intended use cases, data coverage, contents, and potential harm/risks is crucial to facilitate proper & responsible AI development.



Interrogating data quality & generating “nutrition label”



Dataset Nutrition Label

NYC Notice of Property Value Data

About

This dataset is a set of tax bills that are required by the city to be submitted for tax purposes each year. They are mandated by the city.

Data Creation Range: January 2008 - Present
Created By: NYC Department of Finance
Content: List of PDFs
Source: https://a836-pts-access.nyc.gov/care/forms/htmlframe.aspx?mode=content/home.htm

Alert Count	**5***
Completeness	0
Provenance	1
Misrepresentation	1
Collection	2
Socioecnomic Bias	1
Inaccurate Prediction	1
Description	0
Composition	2
Racial bias	1
Socioeconomic Bias	1

* Please refer to the Objectives and Alerts section for more details

Use Cases

Potential real-world applications of the dataset

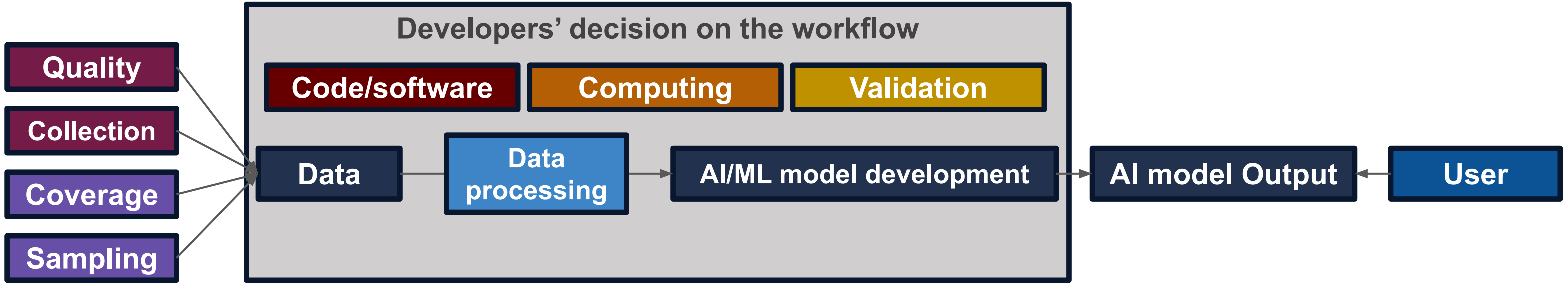
1 How many rent stabilized units are in a particular building?
2 Has a building lost rent stabilized units?
3 Is there a pattern of deregulation in a building?
4 Is there a pattern of deregulation for a specific landlord?
5 Where might there be abuse of tax abatements? Is the landlord breaking the rules of the abatements?
6 Where is gentrification happening?

Badges

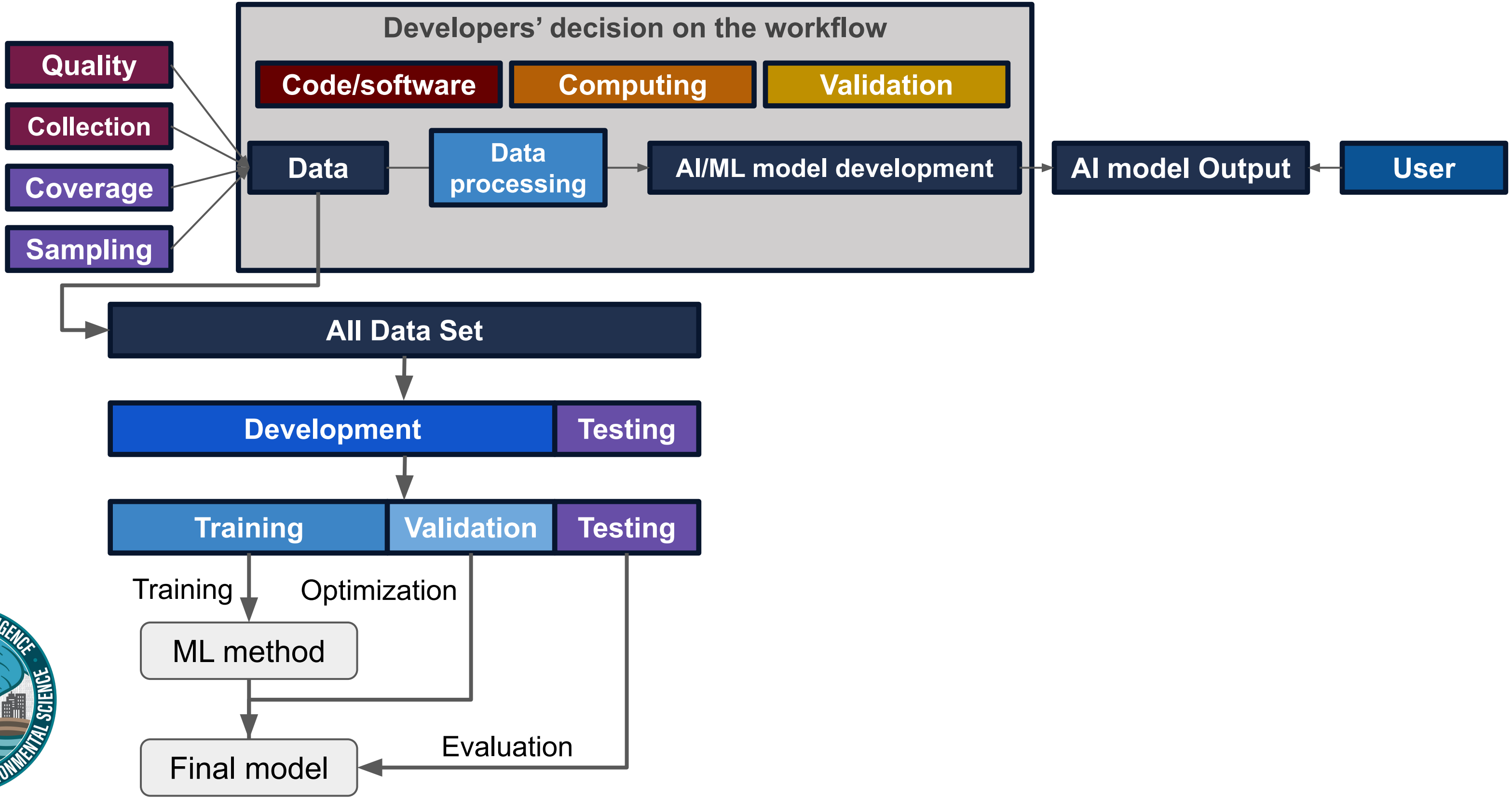
Source: https://datanutrition.org/labels/nopv-nyc/

Chmielinski, et al. "The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence." *arXiv preprint arXiv:2201.03954* (2022).

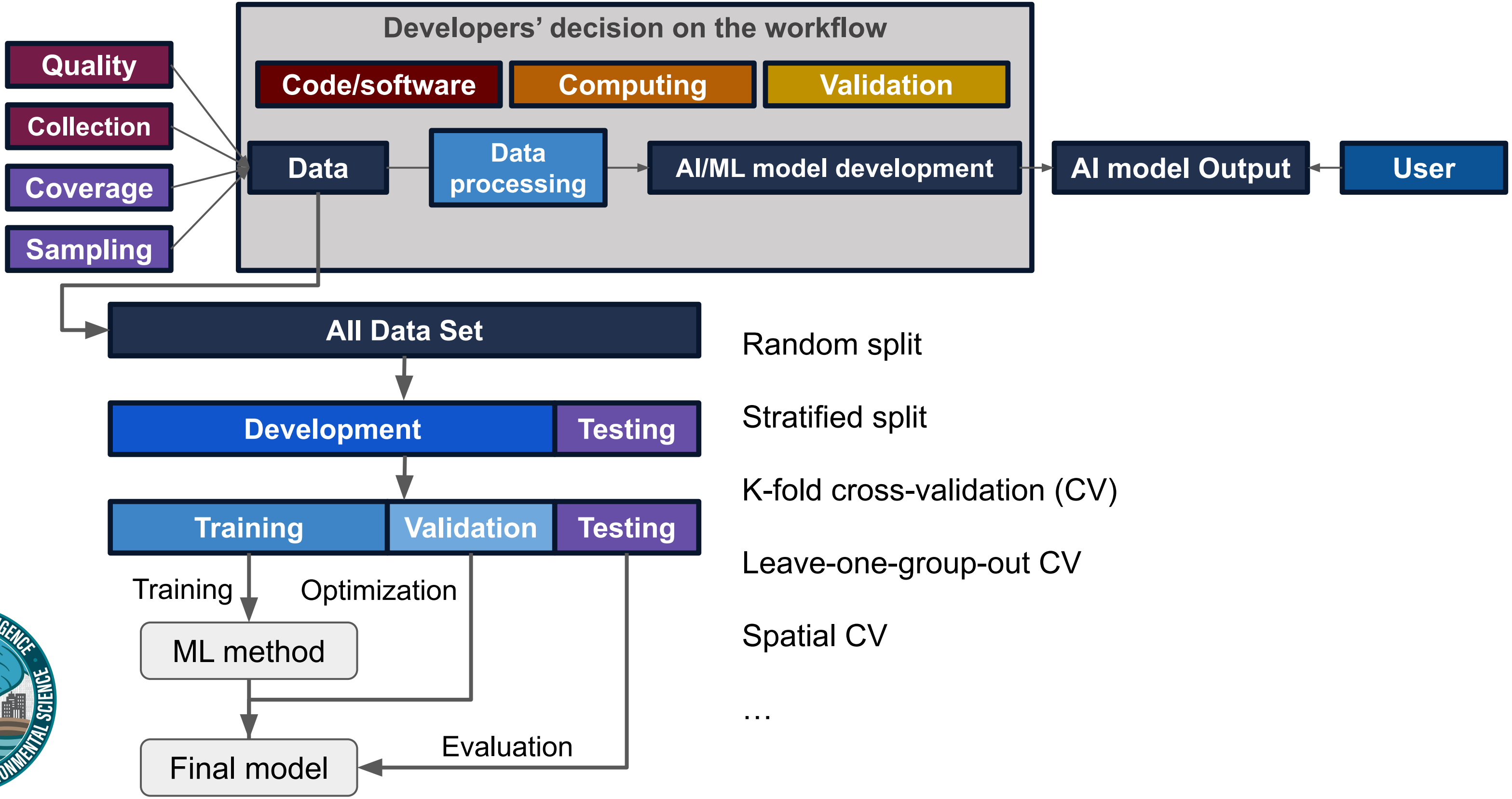
Data processing for model development & evaluation



Data processing for model development & evaluation



Data processing for model development & evaluation



Random split

Stratified split

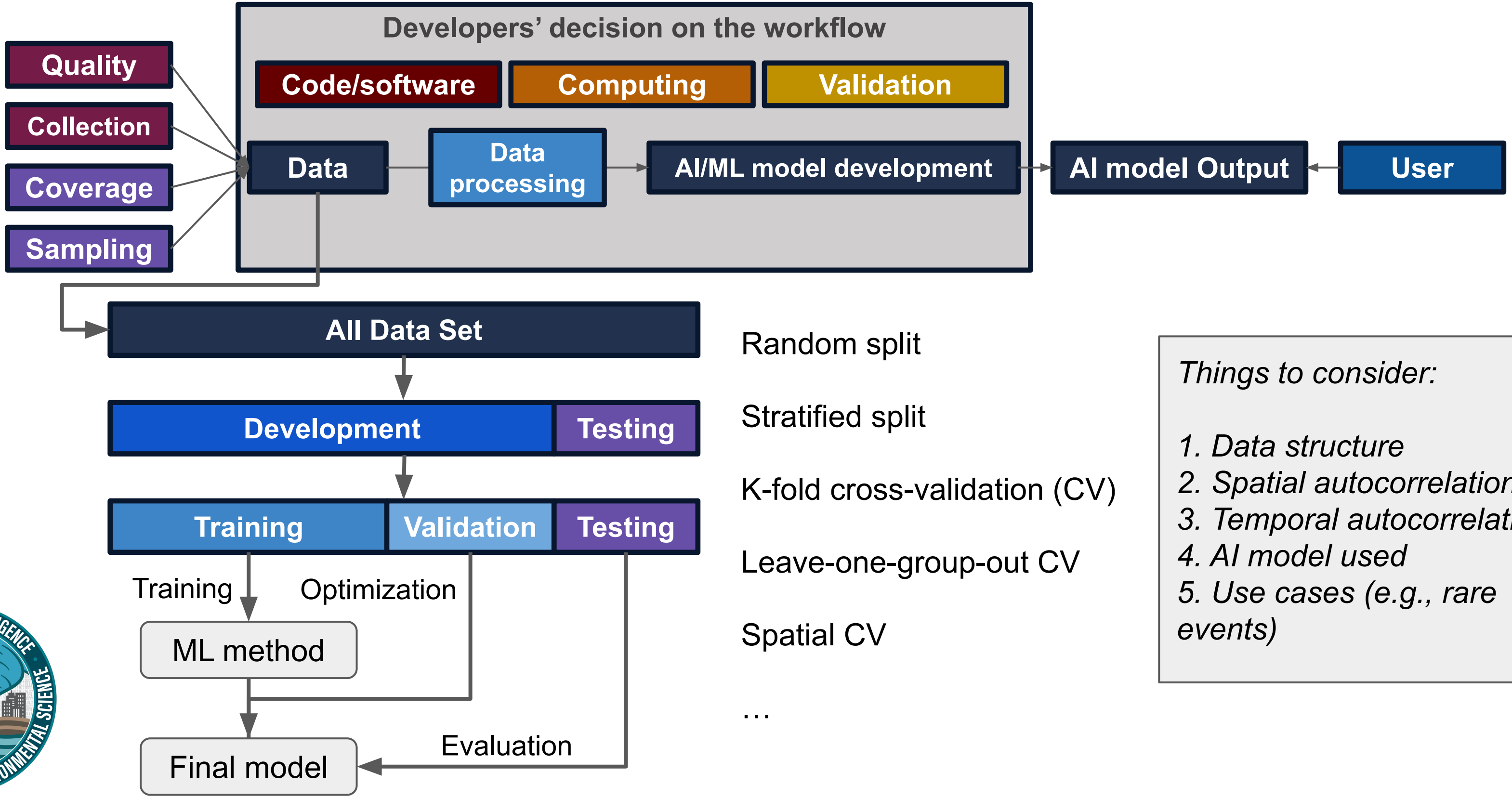
K-fold cross-validation (CV)

Leave-one-group-out CV

Spatial CV

…

Data processing for model development & evaluation



Random split

Stratified split

K-fold cross-validation (CV)

Leave-one-group-out CV

Spatial CV

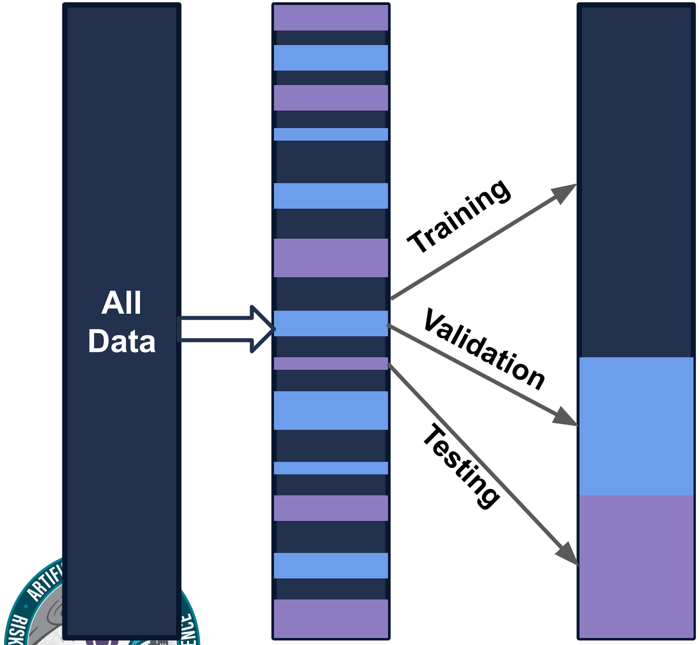
…

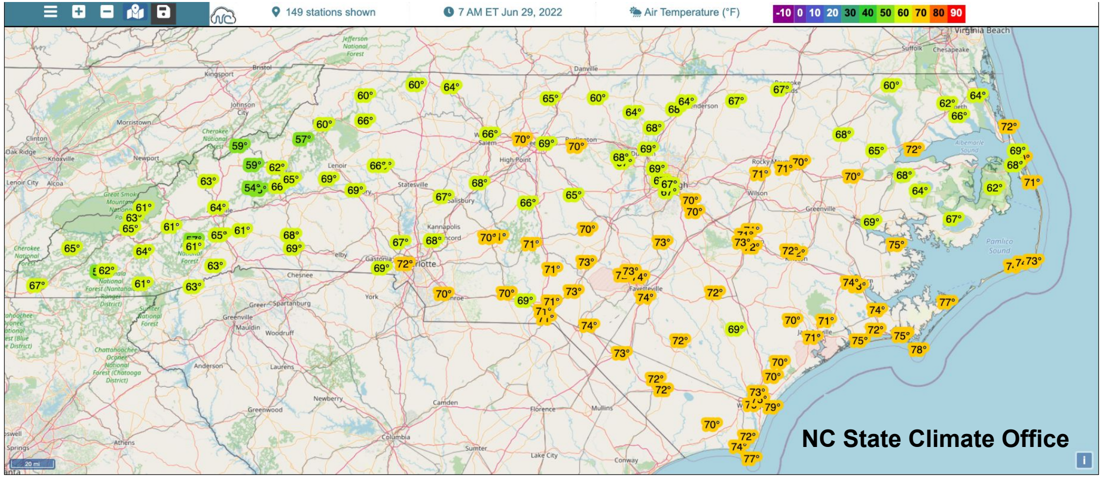
Things to consider:

1. *Data structure*
2. *Spatial autocorrelation*
3. *Temporal autocorrelation*
4. *AI model used*
5. *Use cases (e.g., rare events)*

Data splitting – Spatial considerations

Random data splitting





First Law of Geography (Tobler, 1970): everything is related to everything else, but near things are more related than distant things.

Is random splitting the best way for an objective model assessment?

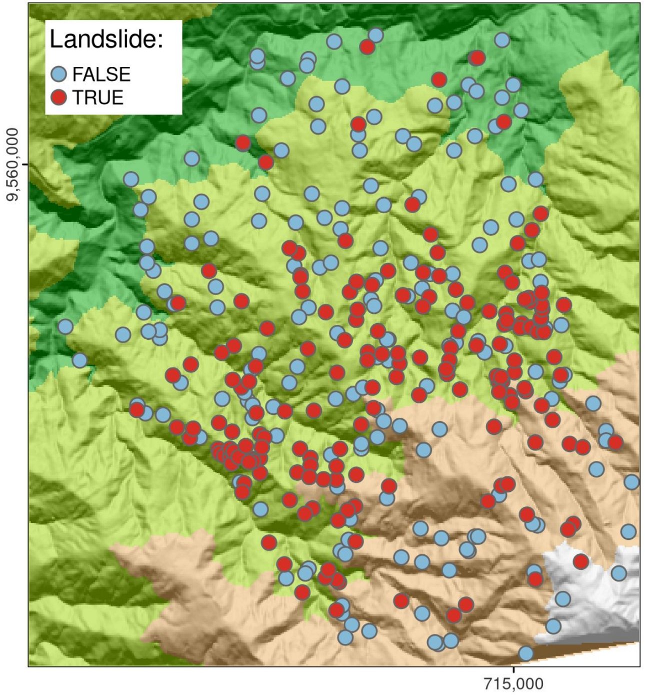
Considerations for spatial autocorrelation in data

Environmental data often contains spatial autocorrelation (or Tobler's first law of geography).

Random split often ignore this aspect of the geospatial data and cause spillover effect.

The result can lead to decreased model performance in unseen situations.

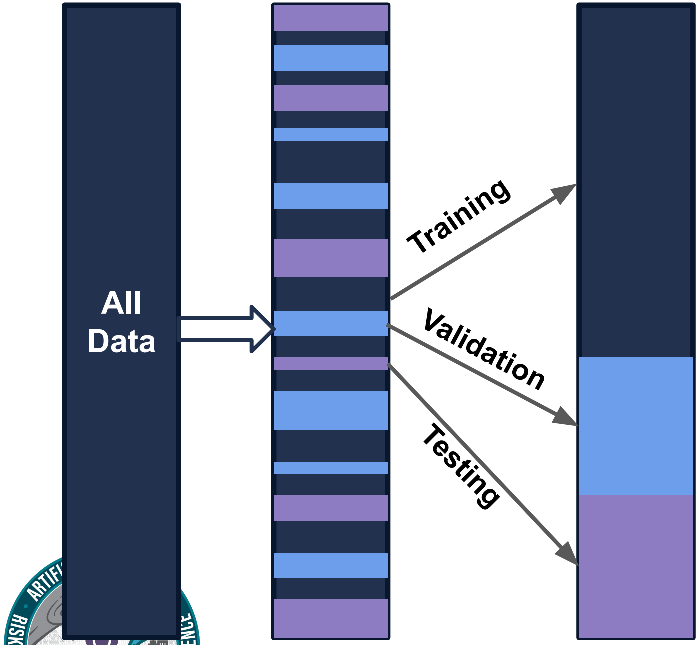
(Right: observed landslide occurrence imposed over topography data in Ecuador)

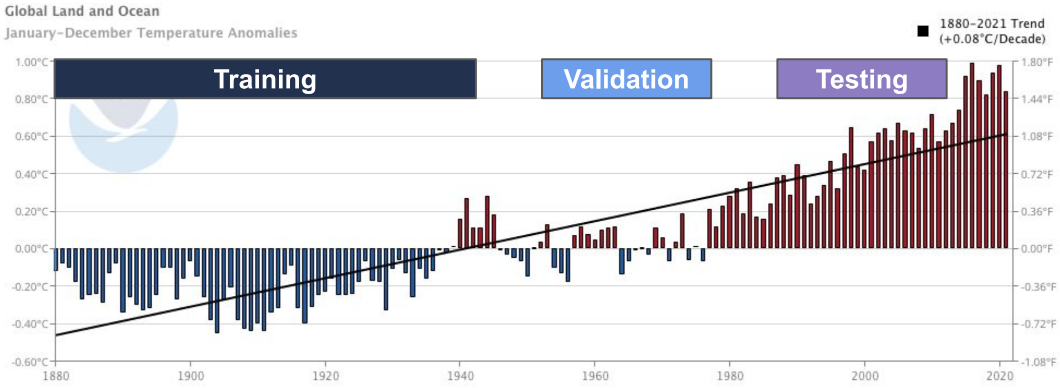


Source: https://geocompr.robinlovelace.net/spatial-cv.html

Data splitting – Temporal considerations

Random data splitting





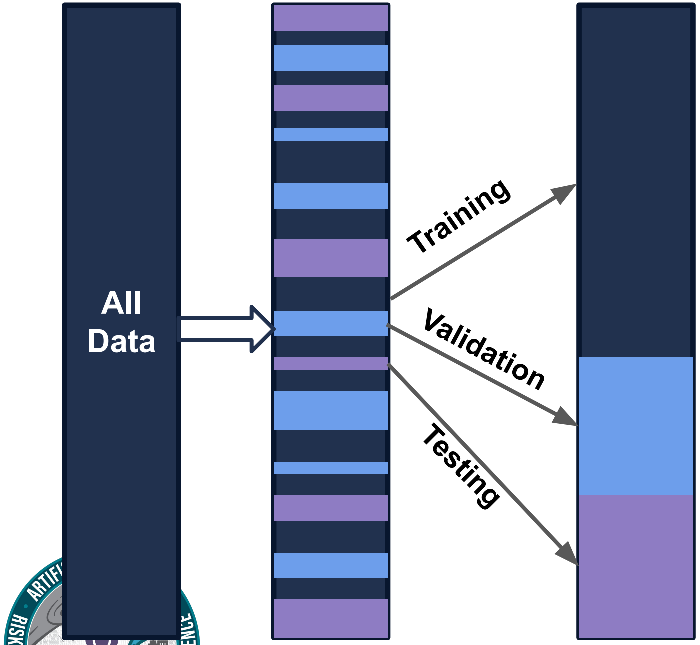
NOAA Climate at a Glance

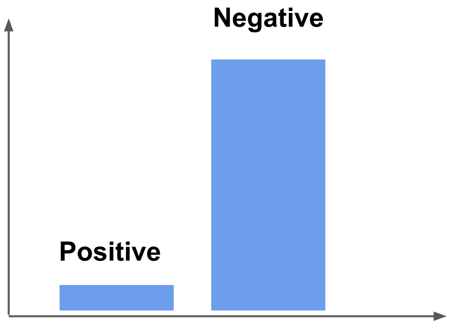
Environmental data often have temporal autocorrelation – the data that from previous time periods are related to the current time – random splitting may not account for this information.

Sometimes, a chronological splitting can be more appropriate.

Data splitting – Imbalance samples

Random data splitting





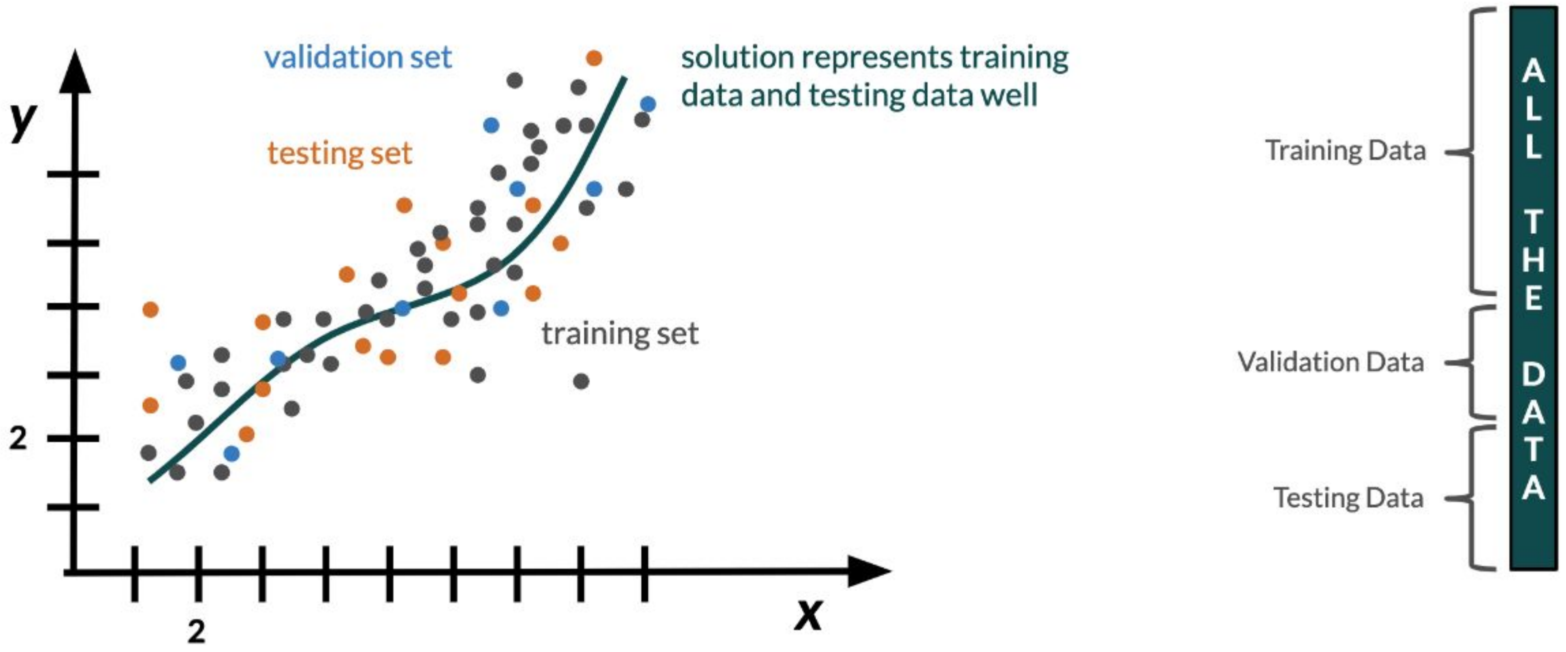
	Positive	Negative
Positive	10	0
Negative	90	900

In certain applications, we are dealing with imbalanced samples (some cases extremely imbalanced). These imbalanced samples need to be dealt with caution to avoid the artificial impact on the model performance.

There are a variety of methods dealing with imbalanced samples (see the review from Krawczyk, 2016).

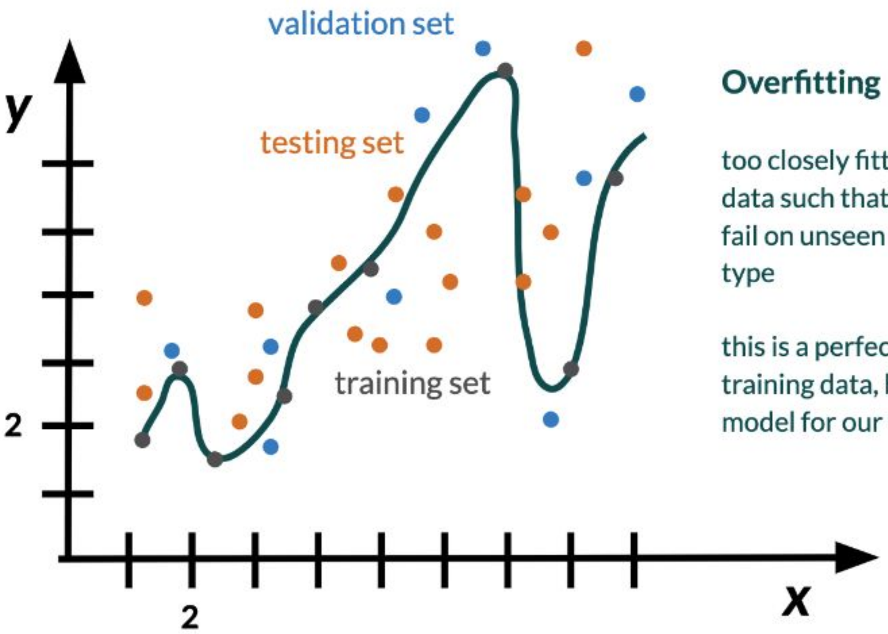
Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232.

Avoid overfitting



Barnes (2022). https://doi.org/10.5281/zenodo.6686879

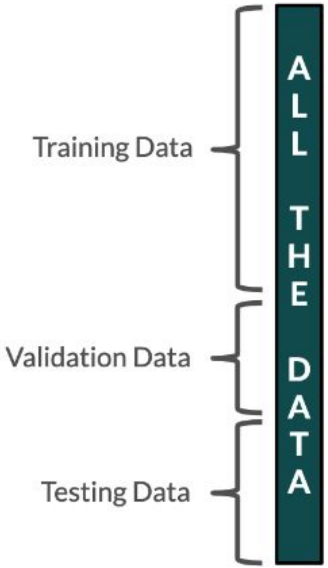
Avoid overfitting



Overfitting

too closely fitting the training data such that the model will fail on unseen data of the same type

this is a perfect model for the training data, but a very poor model for our testing data


Training Data

Validation Data

Testing Data

ALL THE DATA

Barnes (2022). https://doi.org/10.5281/zenodo.6686879

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Breathing break -
Any open questions?
Add them in Slido!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do
and use the
code TAI4ES

ARTIFICIAL INTELLIGENCE · ENVIRONMENTAL SCIENCE · RISK COMMUNICATION

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR
ARTIFICIAL INTELLIGENCE

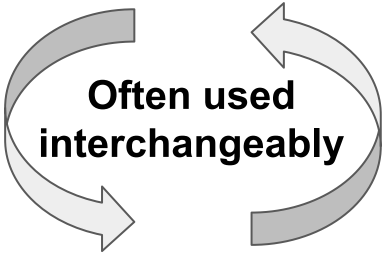
Radiant Earth
Foundation
EARTH IMAGERY FOR IMPACT

Reproducibility and replicability

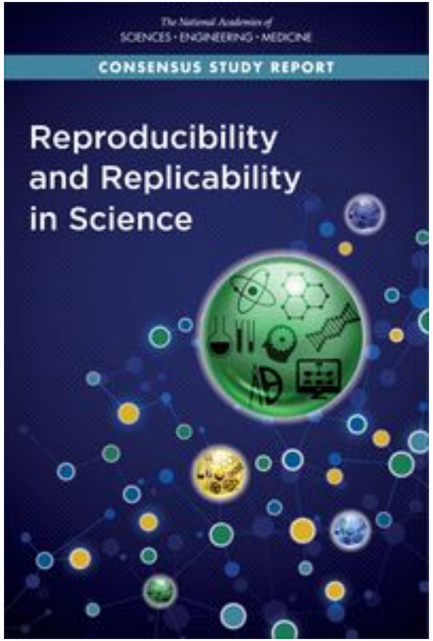
NASEM (2019) definitions:

Reproducibility: obtaining consistent computational results **using the same input** data, computational steps, methods, code, and conditions of analysis

Replicability: obtaining consistent results **across studies** aimed at answering the same scientific question, each of which has obtained its own data



Often used interchangeably


The National Academies of
SCIENCES · ENGINEERING · MEDICINE
CONSENSUS STUDY REPORT
Reproducibility and Replicability in Science

In part a response to the "replication crisis"

https://www.nationalacademies.org/our-work/reproducibility-and-replicability-in-science

Reproducibility

= obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis

- Why is it important?
  - reproduce own results in future (e.g., prior to publication)
  - other researchers can build on your work
  - validity/confidence in results
  - allows others to review code/methods and help find errors, biases, etc

What are the implications for trustworthiness?

Assessing reproducibility

- Assessing success
    - Choose metric/attribute of interest
    - Thoughtfully characterized uncertainties
    - Computational reproducibility usually lead to near-identical results (exceptions on next slide)
- Success does not guarantee correctness of results
- Reproducibility study types:
    - Direct (rare) - replaying the computations to obtain consistent results
    - Indirect - assessments of transparency of data and methods
- Systematic efforts to computationally reproduce results across various have **failed more than half the time** due to insufficient detail

Sources of non-reproducibility

- Non-public data and code
- Inadequate record keeping*
- Nontransparent reporting*
- Obsolescence of the digital artifacts (lack of continued curation)
- Flawed attempts to reproduce others' research (lack of expertise, didn't follow protocols)
- Barriers in the culture of research - lack of resources and incentives, publication bias (against confirmatory research)

Replicability

= obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data

- Why is it important?
    - One of the key ways scientists build confidence in the scientific merit of results
    - Identifying biases, outliers
- Thoughtful identification of sources of uncertainty, quantification of uncertainty, and documentation of uncertainty crucial to replicability

What are the implications for trustworthiness?

Assessing replicability

- Assessing success - can be done in a number of ways
  - Choose metric/attribute of interest
  - Thoughtfully characterized uncertainties
  - Exact replication not possible - can define consistent results via a “proximity-uncertainty” framework (next slide)
- Success does not guarantee correctness of results
- Systematic efforts to computationally reproduce results across various have **failed more than half the time** due to insufficient detail - more nuanced interpretation than reproducibility
  - Even when a study was rigorously conducted according to best practices, correctly analyzed, and transparently reported, it may fail to be replicated
  - A single inability to replicate does not mean original results are not correct

Uncertainty and Replicability

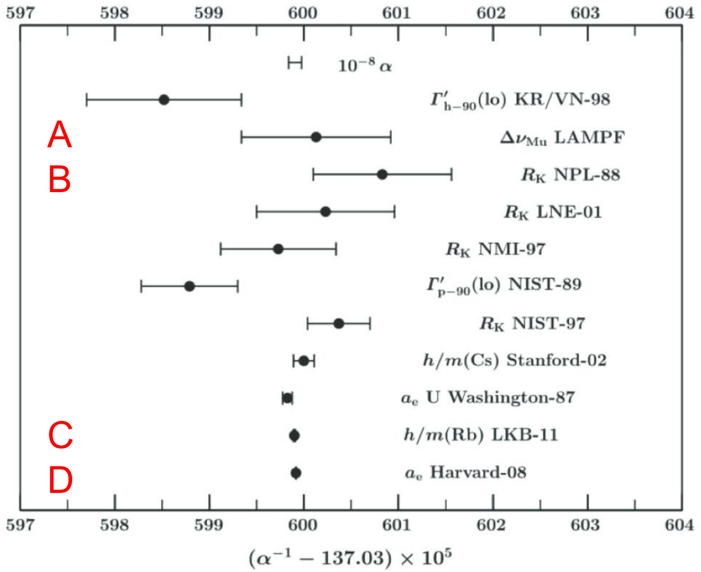


FIGURE 3-2 Evolution of scientific understanding of the fine structure constant over time.
NOTES: Error bars indicate the experimental uncertainty of each measurement. See text for discussion.
SOURCE: Reprinted figure with permission from Peter J. Mohr, David B. Newell, and Barry N. Taylor (2016). Reviews of Modern Physics, 88, 035009. CODATA recommended values of the fundamental physical constants: 2014. Copyright 2016 by the American Physical Society.

In comparing replication attempt results, we want to look at proximity (points) and uncertainty (error bars)

Do the results from study B successfully replicate study A? What about studies C & D?

What about studies A and C?

Sources of Non-Replicability in Research

Helpful

- Discovery of an unknown effect, interrelation, or interaction
- Previously unknown sources of uncertainty
- Nature of the problem under study and the prior likelihoods of possible results
- Novelty of the area of study and therefore lack of established methods of inquiry

Unhelpful

- Human error or poor research choices
- Bias toward a particular outcome
- Misconduct or fraud
- Publication bias
- Misaligned incentives
- Inappropriate statistical inference
- Incomplete reporting of a study

Non-replication is a normal consequence of studying complex systems

- Too much = conservative research, missing important novel discoveries
- Too little = lack of confidence in results, lack of consensus-building

Steps Necessary for Reproducibility and Replicability

1. **the input data** used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;

2. a **detailed description of the study methods** (ideally in executable form) together with its computational steps and associated parameters; and

3. information about the **computational environment where the study was originally executed**, such as operating system, hardware architecture, and library dependencies. (Library dependency,2 in the context of research software as used here, is the relationship of pieces of software that are needed for another software to run. Problems often occur when installed software has dependencies on specific versions of other software.)

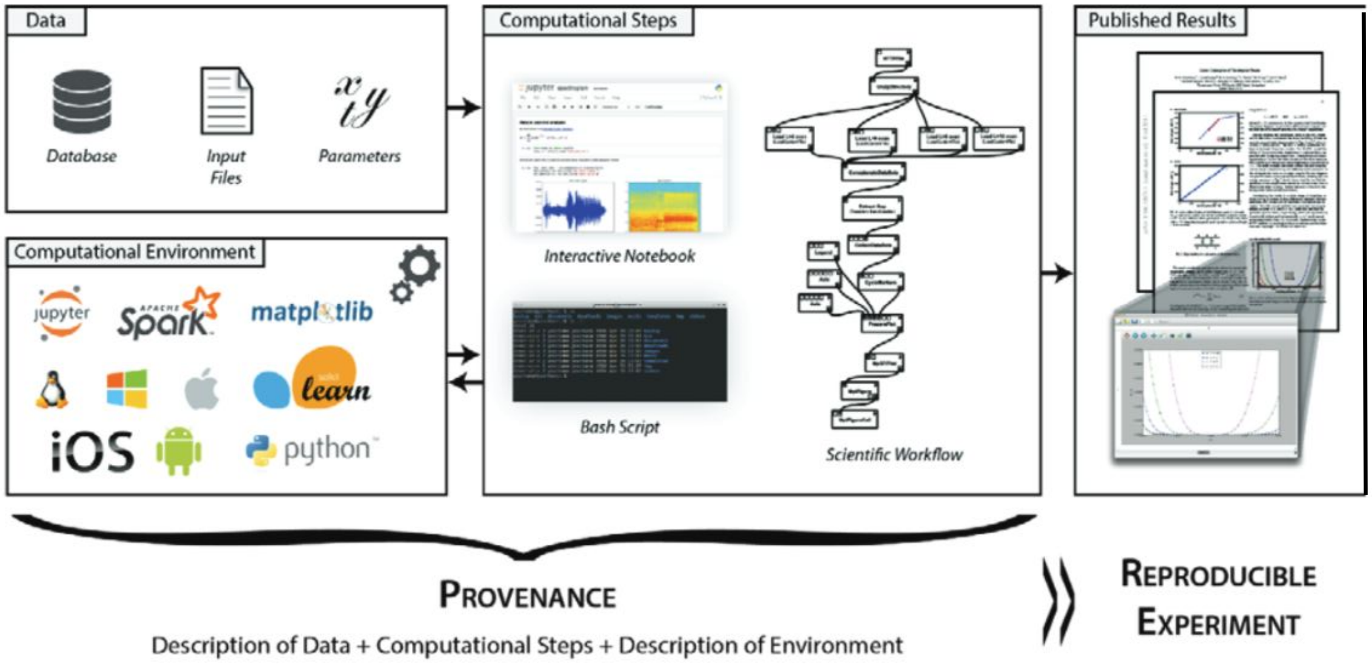
Pause for questions and a quick check in poll

Quick poll break to give you time to soak information in and ask questions!

In your own words, why should we care about **reproducibility** and **replicability?**

3.4. Go to sli.do and use the code TAI4ES

Provenance: importance of effective documentation



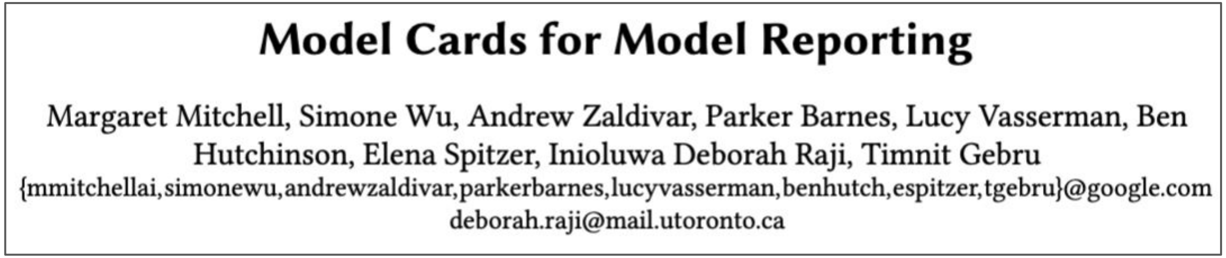
NASEM 2019, Fig. 6-2

Components of Provenance

- Data
  - Use version control
  - document cleaning and QC practices
  - providing original datasets and intermediate datasets
- Code/Software
  - Use version control (e.g., Git)
- Computational Environment
  - Virtual machines + source code and documentation
  - Using containers (e.g., Docker) and computational notebooks (e.g., Jupyter notebooks)

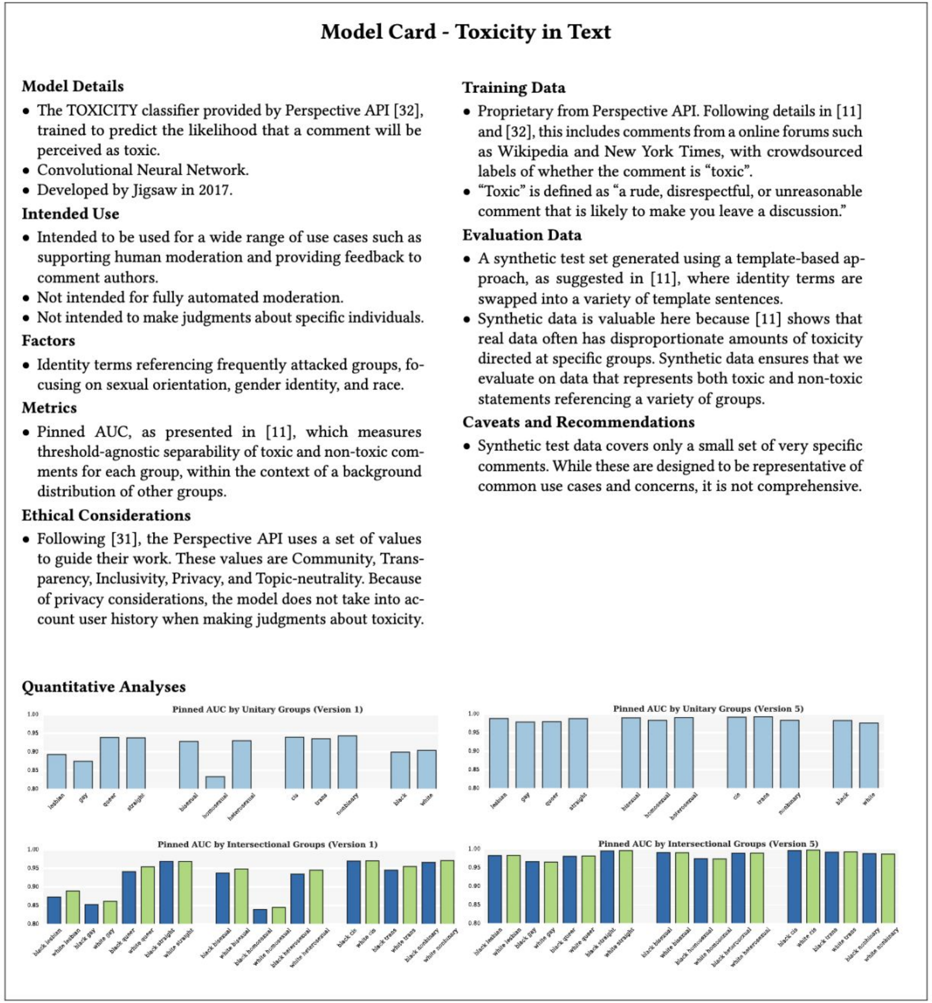
Example framework for model reporting - model cards

- Currently there is no standardized provenance for ML/AI models
- Mitchell et al. 2019 (https://arxiv.org/abs/1810.03993) proposed a framework for short documents accompanying trained machine learning models, called **model cards**
- Encourage transparent model reporting



Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca



Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses

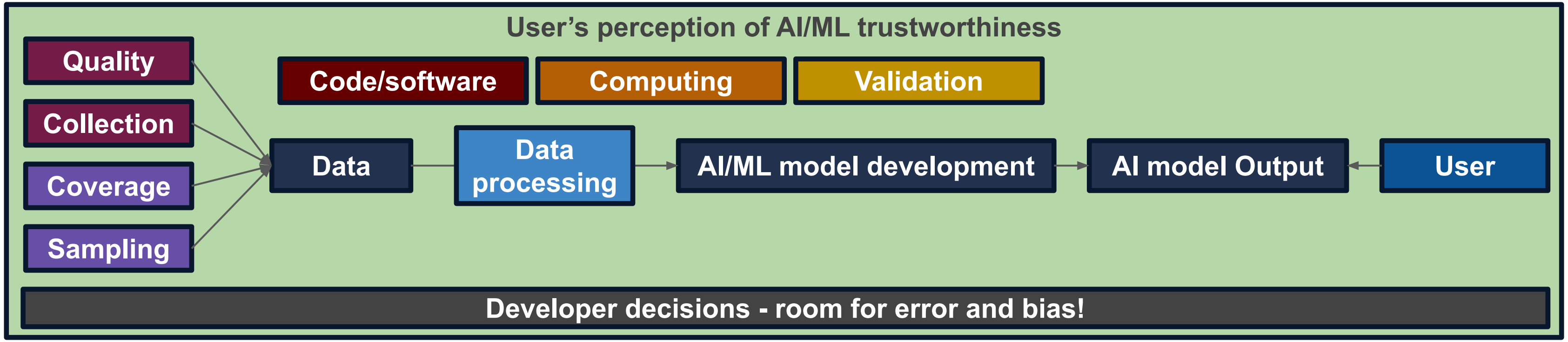
Figure 3: Example Model Card for two versions of Perspective API's toxicity detector.

- Key elements
  - Benchmarked evaluation in a variety of conditions
  - Context in which models are intended to be used
  - Details of the performance evaluation procedures
  - Other relevant information
- Can be used to document any trained machine learning model
- A step towards the responsible democratization of machine learning and related artificial intelligence technology,

Our decisions also may have (unintended) consequences

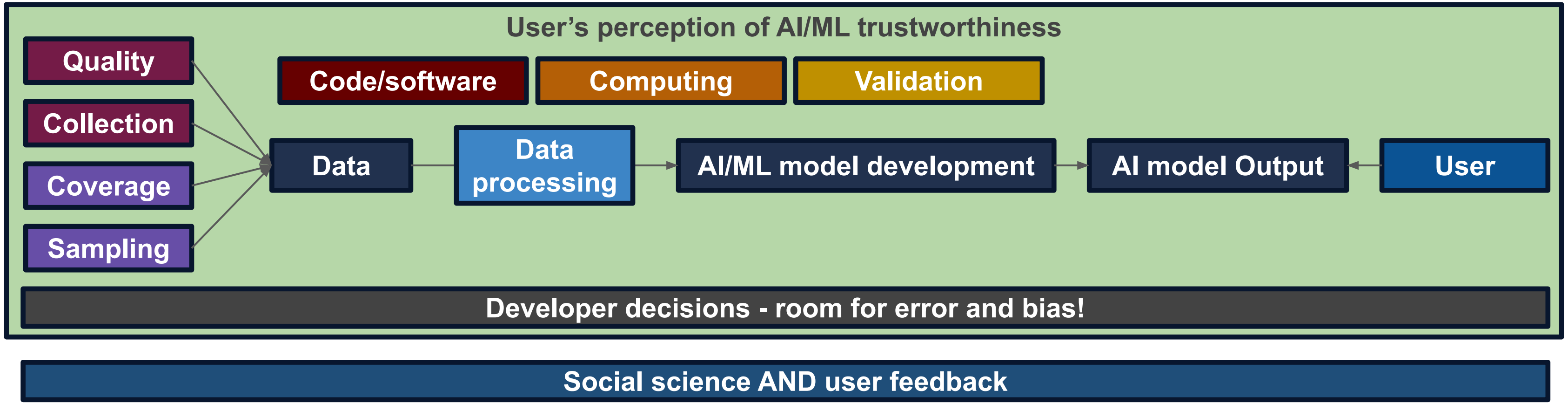
Should ALL science be open?

What about privacy? Protection of rights? Preserving trust?



Why do we care about trustworthy data and workflows?

We need <u>social science research with users</u> throughout this entire process!



Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Breathing break - Any open questions? Add them in Slido!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do
and use the
code TAI4ES

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 3: Agenda

- 9:00 Trustworthiness of data and workflows
- **10:30 *Brain & bio break***
- 10:45 The importance of case studies and tips for using them effectively

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do
and use the
code TAI4ES

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

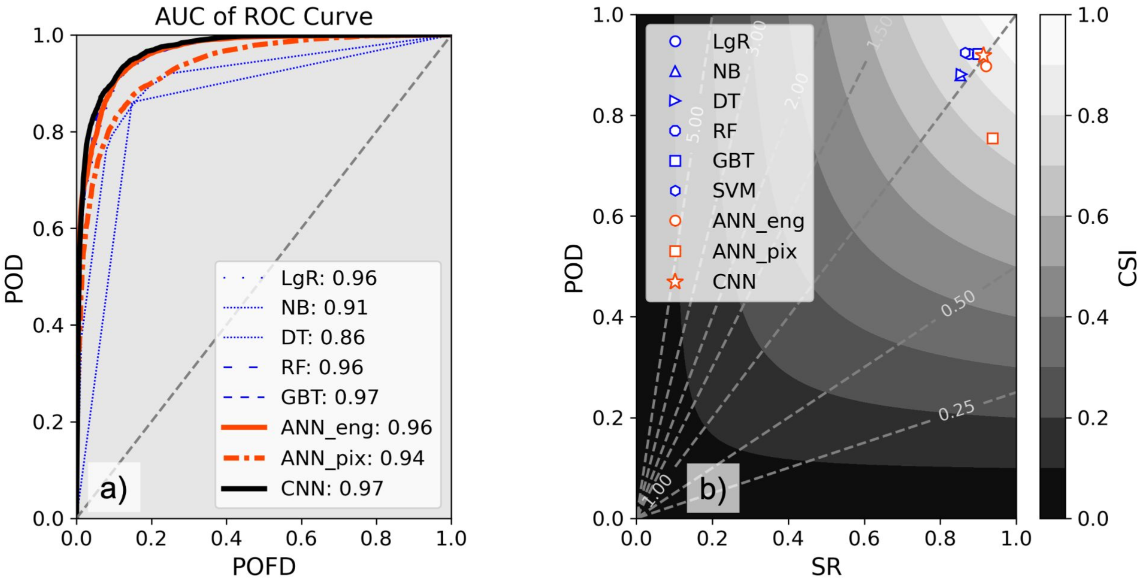
NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Integration of case studies

Trust and Model Verification

- Why can’t we just verify everything objectively?
- Is using multiple metrics to explain model performance enough to build users’ trust?



Chase, et al. (in prep.) *A Machine Learning Tutorial for Operational Meteorology, Part II: Neural Networks*, To be submitted to Weather and Forecasting.

3.5. & 3.6. Go to sli.do and use the code TAI4ES

Trust and case studies - examples from NWS Forecasters

Forecaster 10: Getting to use it more and **see how it does in different kinds of convective situations**. I really can’t think of a better way to raise the trustworthiness. And if it’s not perfect that’s okay but **knowing how it performs in certain situations and see that**, getting that kind of baseline for how it performs **really, really increases the trustworthiness**.

Forecaster 4: “We tend to not be classic supercell land - [storm mode is] very messy. So what may work in an area of Oklahoma or Kansas – discrete, very pretty supercells, we tend not to get those as often here. So I would say **I would need some time to make sure that [the guidance is] encompassing the sort of weather we see** in [the state].

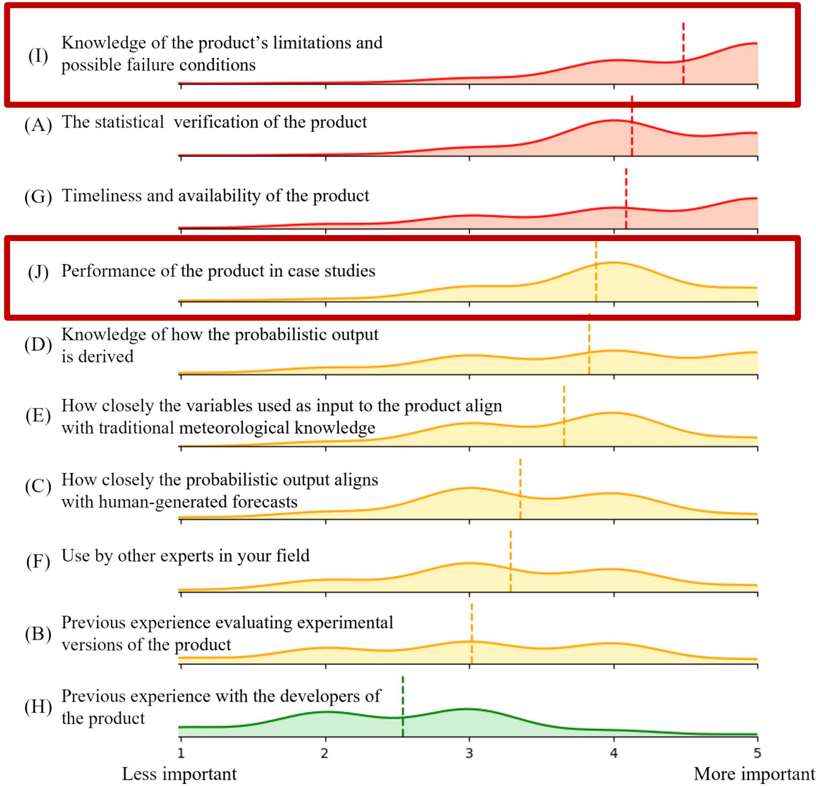
From AI2ES interviews introduced Day 1 and Day 2 of the Summer School - From the AI2ES Risk Communication team interviews with NWS Forecasters

We’ve seen this general idea across multiple studies

NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE) hosted by NOAA’s Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL)

- 2021 Experiment:
  - 133 participants over 5 weeks (virtual)
- Participants from:
  - Local NWS Weather Forecast Offices
  - NWS Environmental Modeling Center
  - OAR Laboratories
  - Cooperative Institutes
  - NCAR
  - Academia
  - International research and forecasting agencies

(Q7) Relative Importance When Evaluating a New Machine Learning Product



Harrison, David (2022) Machine Learning Co-Production in Operational Meteorology, PhD Thesis, School of Meteorology, University of Oklahoma.
Harrison, D., McGovern, A., Karstens, C., Demuth, J. L., Bostrom, A., Jirak, I. L., Marsh, P. T. (2022) Challenges and Benefits of Machine Learning in an Operational Environment: Survey Results from the 2021 Hazardous Weather Testbed Spring Forecasting Experiment. Presented at the 21st Conference on Artificial Intelligence for Environmental Science at the 2022 American Meteorological Society Annual Meeting.

Not just for AI - case studies are important across guidance

Forecaster: "The big thing for me, especially with anything new, **is have they tested it out west**? **Does it work here?** Is it something that's being developed across the Plains? Of course it's going to work great across the Plains. **But how does it work out here where you have sparsity of observation data?** You do not get to have two or three days of upstream data going into it. I have to gain confidence that it's going to be useful for me here." (No. 3-rain)

Forecaster: "But I also want to see, **if you look at [model verification] over the whole winter season**, this particular model may do the best. **But is it doing the best when it really matters?** [Like] when there are high winds? Because that's when the impacts are going to be greatest. Maybe a model does best at snowfall amounts over the whole season, **but is it catching our higher amounts** or does [under forecast]?" (No. 24-winter)

Demuth, J. L., Morss, R. E., Jankov, I., Alcott, T. I., Alexander, C. R., Nietfeld, D., Jensen, T. L., Novak, D. R., & Benjamin, S. G. (2020). Recommendations for Developing Useful and Usable Convection-Allowing Model Ensemble Information for NWS Forecasters. *Weather and Forecasting*, 35(4), 1381–1406. https://doi.org/10.1175/WAF-D-19-0108.1

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Breathing break - Any open questions? Add them in Slido!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do
and use the
code TAI4ES

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR
ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

So how do we communicate this type of information to users?

An example from AI2ES

ARTIFICIAL INTELLIGENCE · ENVIRONMENTAL SCIENCE · RISK COMMUNICATION

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR
ARTIFICIAL INTELLIGENCE

Radiant Earth
Foundation
EARTH IMAGERY FOR IMPACT

NSF

Interdisciplinary Risk Communication Research Approach

Hazard Use Cases

Interdisciplinary Research Team

Risk Communication Scientists
Environmental & Atmospheric Scientists
AI/ML Scientists and Developers

Social Science Data Collection Methods

Formative Research

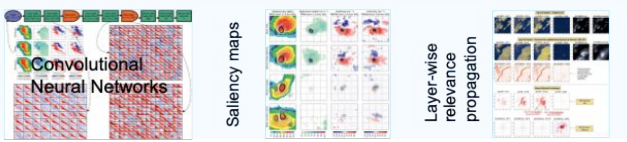
Semi-structured interviews

Randomized Experiments

Evaluative Research

Surveys in naturalistic settings

AI/ML & XAI Techniques



Users & Decision Makers

US Army Corps of Engineers.

AAPA

USGS science for a changing world

Trustworthy & Use-Driven AI and ML Products

Example using FogNet - Coastal Fog Predictions

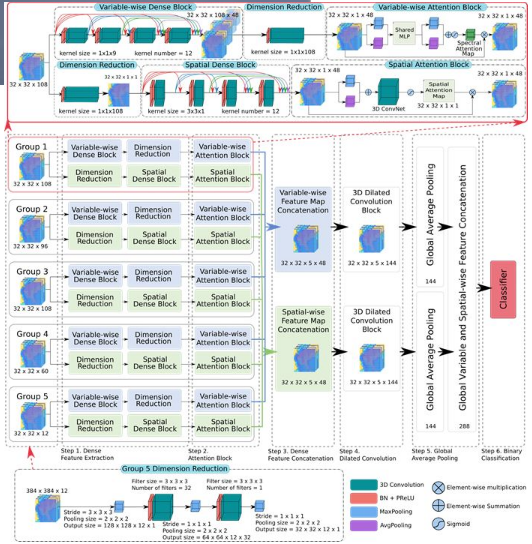
Project goals:

- Develop a new method to predict coastal fog using a 3D-CNN approach with operational potential.
- Apply XAI methods to gather insights into the relative importance of dynamic of coastal fog formation and trust for approach.
- Develop a DL architecture that ***captures air-sea-land interactions*** while accommodating predictor fields including satellite imagery with different spatial resolutions.
- Expand predictors field to Gulf of Mexico (long term).

Publications & Insights so far:

- Kamangir et al. (2021) “FogNet: A Multiscale 3D CNN with Double-Branch Dense Block and Attention Mechanism for Fog Prediction” in Machine Learning with Applications, 5, 100038. https://doi.org/10.1016/j.mlwa.2021.100038.
- Kamangir et al. (2022) Importance of 3D Convolution and Physics-based Feature Grouping in Atmospheric Predictions. Environmental Modelling & Software, 154, 105424, https://doi.org/10.1016/j.envsoft.2022.105424.
- The performance of FogNet was compared to operational models, HREF, SREF showing substantial performance improvement & XAI results indicate a spatial focus on the predicted area and a few key input variable maps.

FogNet: 3D CNN with Physics based grouping of ~ 300 input maps based on NWPs and satellite imagery



Example using AI2ES AI/ML fog guidance

- Guidance is being developed using new techniques (3D convolutional neural net or 3D-CNN) to predict fog and mist, this does not include the presence of rain or air pollution (e.g., smog).
- Below is a real-world example of the guidance output.

What else would you want to know about the model before using it?

3.7. Go to sli.do and use the code TAI4ES

FogNet: Fog prediction

Mustang Beach Airport *Port Aransas, Texas, USA*	**Init.** 0555z Jan 15 2020 **Valid** 150555z – 160555z

	≤1600 m		**≤3200 m**		**≤6400 m**	
	Pred.	*Prob.*	**Pred.**	*Prob.*	**Pred.**	*Prob.*
151155z (6h)	**Fog**	*0.92*	**Fog**	*0.58*	**Fog**	*0.73*
151755z (12h)	**Fog**	*0.64*	**Fog**	*0.60*	**Fog**	*0.72*
160555z (24h)	**Fog**	*0.77*	**Fog**	*0.78*	**Fog**	*0.60*

FogNet: Fog prediction

Mustang Beach Airport *Port Aransas, Texas, USA*	**Init.** 1155z Mar 11 2020 **Valid** 251155z – 261155z

	≤1600 m		**≤3200 m**		**≤6400 m**	
	Pred.	*Prob.*	**Pred.**	*Prob.*	**Pred.**	*Prob.*
251755z (6h)	**Fog**	*0.80*	**Fog**	*0.90*	**Fog**	*0.76*
252355z (12h)	**No Fog**	0.14	**No Fog**	0.27	**Fog**	0.60
261155z (24h)	**No Fog**	*0.10*	**Fog**	*0.57*	**No Fog**	*0.27*

Users may want to see how the model does in action

So how do I pick my case study to show them how the model does?

3.8. Go to sli.do and use the code TAI4ES

Here’s one example of how we’ve used a case study

Below is an overview of case study of FogNet conducted by one of the developers, who is an operational NWS forecaster. The case study examined how the guidance performed compared to observations using data from the training set (data from 2009-2020). Three results are shown below.

1. Fog cases were much less common in the training data than no fog cases

FogNet: Training set observations

Observation	# of cases	% of total
Negative (no fog)	8627	98.2
Positive (fog)	160	1.8
Total	**8,787**	**100**

2. Advection fog cases were dominant in the data set

FogNet: Training set case types

Fog Type	# of cases	% of total
Advection	128	80.0
Radiation	6	3.8
Advection-Radiation	10	6.3
Frontal	3	1.9
Cloud-base Lowering	8	5.0
Hybrid	21	13.1
Unknown	17	10.7
Total	**160**	**100**

3. FogNet performed much better on advection cases (accurately predicting about 67% of cases) than on radiation fog cases (accurately predicting ~10% of cases). This is likely because there were many more advection cases in the training set.

Example of Verification of AI/ML Guidance (1 of 2)

POD: Probability of Detection

Interpretation: POD gives the proportion of observed “yes” events that were correctly forecasted

Scale: Range from 0 (never correct) to 1 (always correct), higher numbers are better

FAR: False Alarm Ratio

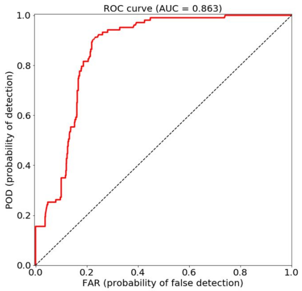
Interpretation: The FAR tells you the proportion of forecast “yes” events that were wrong or “misses”

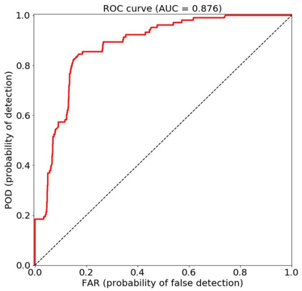
Scale: Range from 0 (always correct) to 1 (never), lower numbers are better

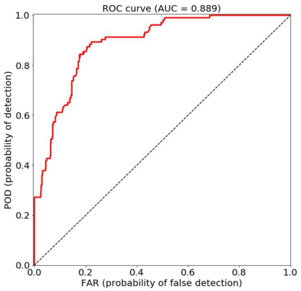
ROC Curve (AUC): Receiver Operating Characteristic (Area Under the Curve)

Interpretation: The ROC curve plots the POD vs. FAR for all possible forecast thresholds to visually represent how well the guidance could perform, independent of calibration. A perfect score falls in the upper left corner.

Scale: AUC has a range from 0 (always wrong) to 1 (always correct); higher numbers are better. Above 0.5 (reflected visually by the diagonal dotted line) is considered better than chance.







FogNet: Verification for 6h lead time

Mustang Beach Airport *Port Aransas, Texas, USA*	**6 hour lead time predictions**		
Visibility	**≤1600 m**	**≤3200 m**	**≤6400 m**
POD	0.61	0.56	0.77
FAR	0.40	0.36	0.42
AUC	0.86	0.80	0.76

FogNet: Verification for 12h lead time

Mustang Beach Airport *Port Aransas, Texas, USA*	**12 hour lead time predictions**		
Visibility	**≤1600 m**	**≤3200 m**	**≤6400 m**
POD	0.60	0.48	0.70
FAR	0.56	0.56	0.43
AUC	0.87	0.72	0.77

FogNet: Verification for 24h lead time

Mustang Beach Airport *Port Aransas, Texas, USA*	**24 hour lead time predictions**		
Visibility	**≤1600 m**	**≤3200 m**	**≤6400 m**
POD	0.54	0.57	0.67
FAR	0.50	0.58	0.41
AUC	0.89	0.77	0.79

Example of Verification of AI/ML Guidance (1 of 2)

This set of verification metrics include those that measure skill (PSS) and those that are useful for assessing performance during rare events (CSI and SEDI). In this context, “skill” refers to the accuracy of a forecast *relative* to some reference forecast.

CSI: Critical Success Index

Interpretation: The CSI measures the proportion of forecast and/or observed events that were correctly forecast (the # of hits divided by the sum of hits, misses, and false alarms). The CSI is considered a good performance measure to evaluate the forecasts of rare events because correct rejections are excluded.

Scale: Range from 0 (no skill) to 1 (perfect score)

PSS: Peirce Skill Score

Interpretation: The PSS is a skill score with POD as the accuracy measure and the POD for a unbiased random forecast as the reference. The PSS measures how well a forecast system can separate the ‘yes’ events from the ‘no’ events and can be written as POD minus the probability of false detection (POFD).

Scale: Range [-1,1]. 1: perfect score; 0 (negative): performance equal to (worse than) the reference forecast.

SEDI: Symmetric Extremal Dependence Index

Interpretation: The SEDI is a verification metric that measures the correspondence between forecasts and observations for rare binary events

Scale: Range from -1 (no skill) to 1 (perfect score), scores less than zero are not considered skillful; higher positive numbers are better

FogNet: Verification for <u>6h</u> lead time

Mustang Beach Airport *Port Aransas, Texas, USA*	**6 hour lead time predictions**		
Visibility	**≤1600 m**	**≤3200 m**	**≤6400 m**
CSI	0.38	0.43	0.50
PSS	0.59	0.58	0.63
SEDI	0.84	0.76	0.83

FogNet: Verification for <u>12h</u> lead time

Mustang Beach Airport *Port Aransas, Texas, USA*	**12 hour lead time predictions**		
Visibility	**≤1600 m**	**≤3200 m**	**≤6400 m**
CSI	0.34	0.30	0.46
PSS	0.49	0.43	0.60
SEDI	0.79	0.74	0.84

FogNet: Verification for <u>24h</u> lead time

Mustang Beach Airport *Port Aransas, Texas, USA*	**24 hour lead time predictions**		
Visibility	**≤1600 m**	**≤3200 m**	**≤6400 m**
CSI	0.35	0.32	0.45
PSS	0.50	0.46	0.59
SEDI	0.78	0.72	0.81

Comparing AI/ML (FogNet) to other guidance

Guidance comparisons

Mustang Beach Airport
Port Aransas, Texas, USA

BSS: Brier Skill Score

Interpretation: A skill score that uses the Brier score (mean square error of probability forecasts for a binary event) as the accuracy measure and climatological forecasts as the reference

Scale: [-∞,1]; 1: perfect score, 0 (negative): performance equal to (worse than) the reference

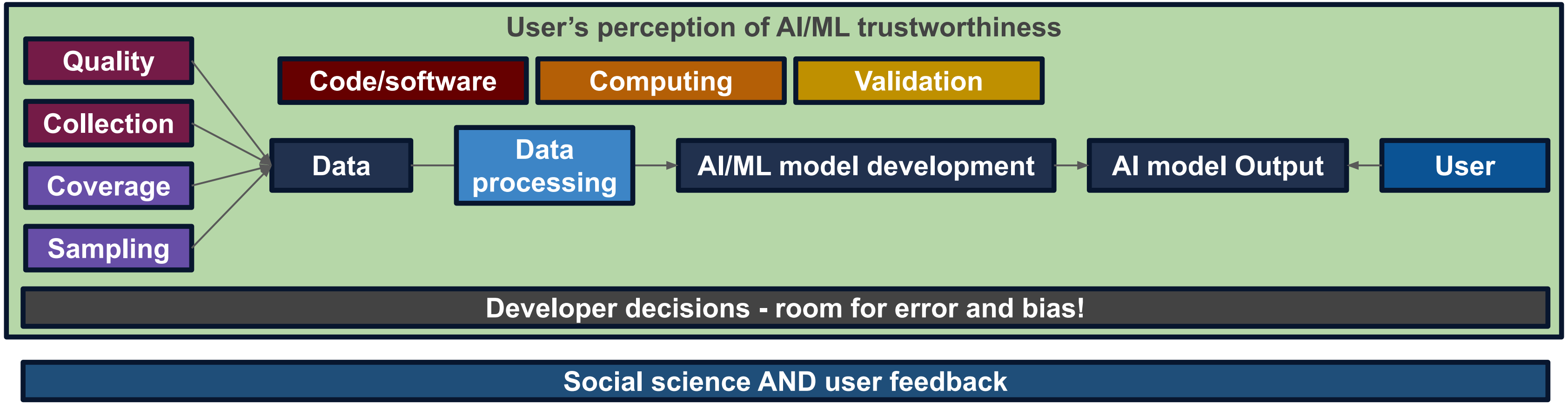
	≤1600 m		**≤3200 m**		**≤6400 m**	
Lead time	*FogNet*	*HREF*	*FogNet*	*HREF*	*FogNet*	*HREF*
6h	0.25	0.09	0.21	0.20	0.28	0.25
12h	0.14	0.08	0.16	0.12	0.16	0.16
24h	-0.14	-0.19	-0.08	-0.05	-0.22	-0.01

These are verification results using an independent dataset from 2018 to 2020.

But what do users care about?

We need social science research with users throughout this entire process!

We'll be testing this soon in an upcoming data collection.



User's perception of AI/ML trustworthiness

Quality

Collection

Coverage

Sampling

Code/software

Computing

Validation

Data

Data processing

AI/ML model development

AI model Output

User

Developer decisions - room for error and bias!

Social science AND user feedback

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Breathing break -
Any open questions?
Add them in Slido!

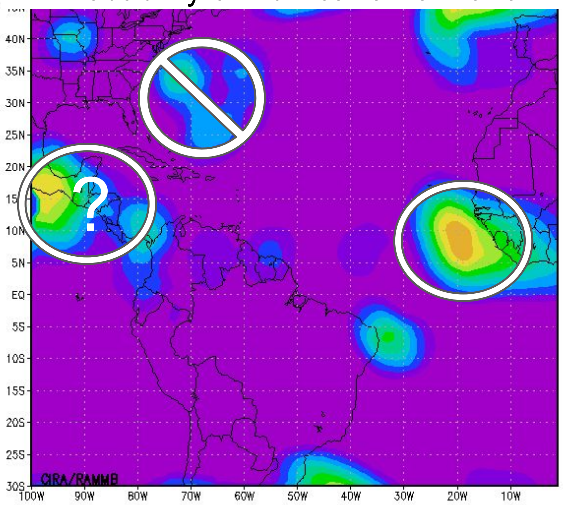
Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do
and use the
code TAI4ES

Using cases in a robust way - How to pick cases?

New Product!
Probability of Hurricane Formation



Important Questions to Ask Yourself:

What is the purpose of this case study?

Why am I picking these cases?

What am I trying to show?

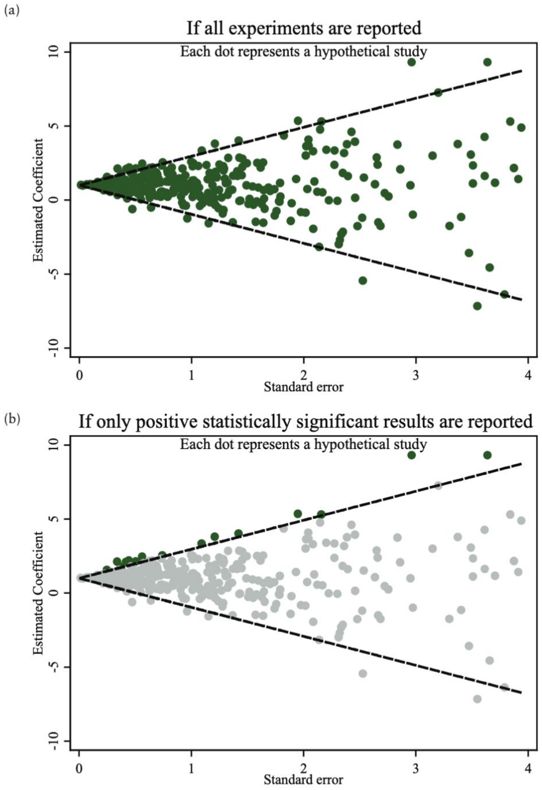
What do users want to see?

Are they representative? Or rare?

How does my model perform on them?

Considerations when picking cases - Model performance

- Important to communicate where and when model performs well and poorly
    - Fognet example - advection vs. radiation fog
    - What does model claim to predict? Are there exceptions (e.g., based on model training set limitations)
- Be cautious when using objective criteria for choosing cases!
    - Objective selection criteria do not guarantee unbiased results
- Avoid cherry picking



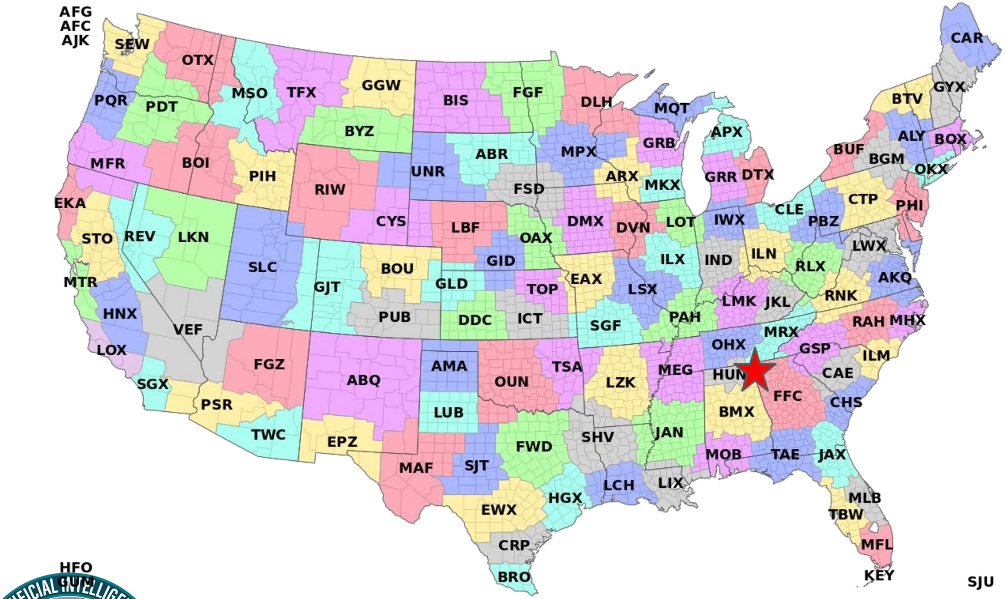
(a) If all experiments are reported

Each dot represents a hypothetical study

(b) If only positive statistically significant results are reported

Each dot represents a hypothetical study

Considerations when picking cases - Relevance to user

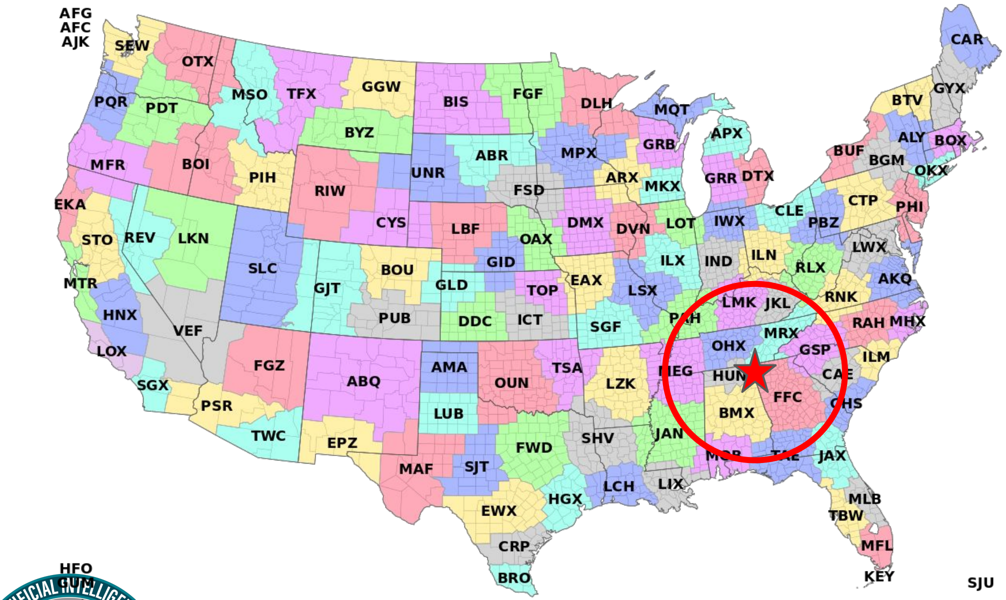


Spatial Relevance

Scenario: A new AI tool for predicting tornadoes is being demonstrated to forecasters at the Huntsville, AL WFO.

Question to ask: What spatial area is most relevant to the user?

Considerations when picking cases - Relevance to user

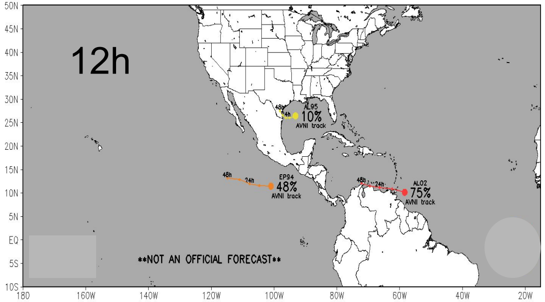


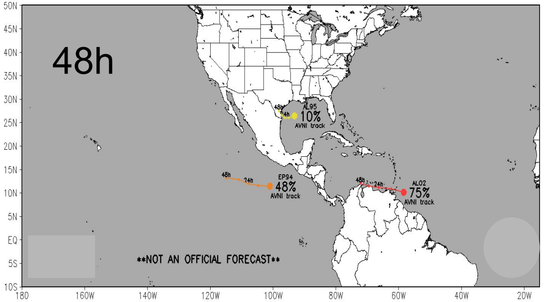
Spatial Relevance

Scenario: A new AI tool for predicting tornadoes is being demonstrated to forecasters at the Huntsville, AL WFO.

Question to ask: What spatial area is most relevant to the user?

Considerations when picking cases - Relevance to user



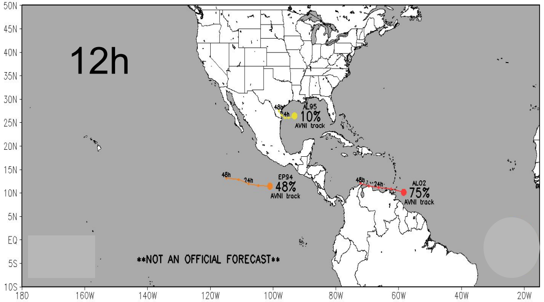


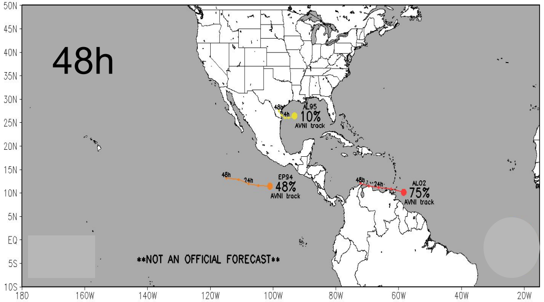
Temporal Relevance

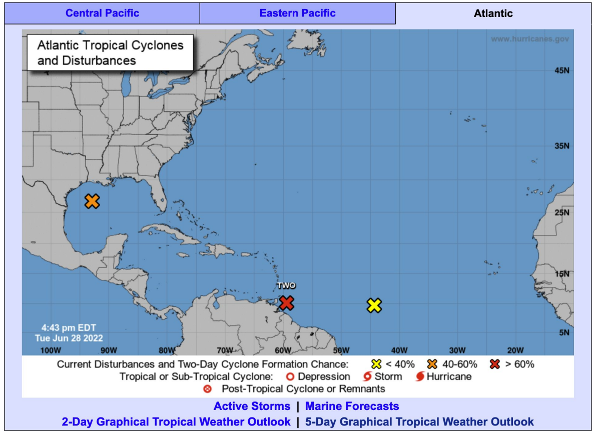
Scenario: A new AI tool predicts likelihood of TC genesis within next 12, 24, and 48 hours. What types of cases would be most relevant to an NHC forecaster?

Question to ask: What time period is most relevant to the user?

Considerations when picking cases - Relevance to user







4:43 pm EDT
Tue Jun 28 2022

Current Disturbances and Two-Day Cyclone Formation Chance: ☒ < 40% ☒ 40-60% ☒ > 60%
Tropical or Sub-Tropical Cyclone: ○ Depression ⚲ Storm ⚲ Hurricane
⊗ Post-Tropical Cyclone or Remnants

Active Storms | Marine Forecasts
2-Day Graphical Tropical Weather Outlook | 5-Day Graphical Tropical Weather Outlook

Temporal Relevance

Scenario: A new AI tool predicts likelihood of TC genesis within next 12, 24, and 48 hours. What types of cases would be most relevant to an NHC forecaster?

Question to ask: What time period is most relevant to the user?

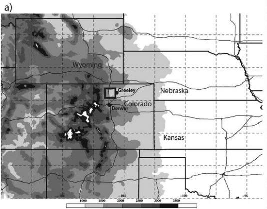
NHC makes 48-hr and 120-hr forecasts for genesis.

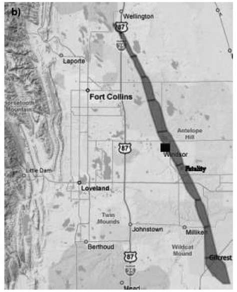
Considerations when picking cases - Is it interesting?

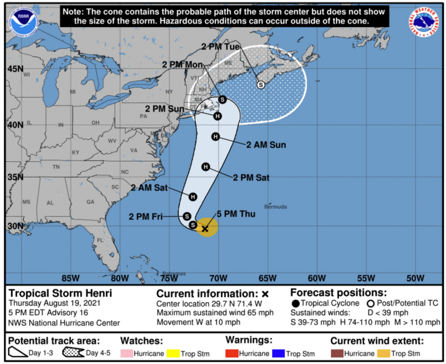
Especially important when developing AI/ML models and tools for forecasters

What makes a case interesting to users (e.g., forecasters)

- High impact events
- Rare events with significant impacts
- Forecast challenges

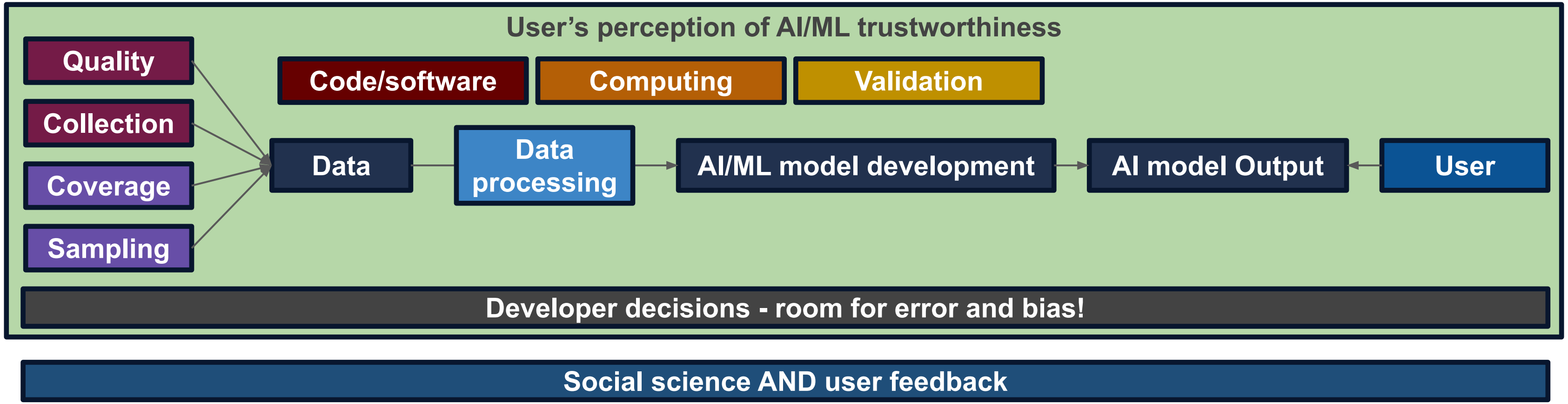






How do we find out what types of cases are most relevant and interesting to users?

- We ask them! And we work to understand their needs, timelines, limitations, workflow, etc.
- To reiterate what Chris said: We need social science research with users throughout this entire process!



User's perception of AI/ML trustworthiness

Quality

Collection

Coverage

Sampling

Code/software

Computing

Validation

Data

Data processing

AI/ML model development

AI model Output

User

Developer decisions - room for error and bias!

Social science AND user feedback

Case studies and Testbeds

Testbeds are great opportunities to demonstrate and collect feedback on an algorithm or product in real time

However, many testbeds occur over a short period of time and may not capture the types of cases you want to demonstrate

Plan ahead and bring the case studies you want to show with you!

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Breathing break - Any open questions? Add them in Slido!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do and use the code TAI4ES

ARTIFICIAL INTELLIGENCE · ENVIRONMENTAL SCIENCE · RISK COMMUNICATION

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

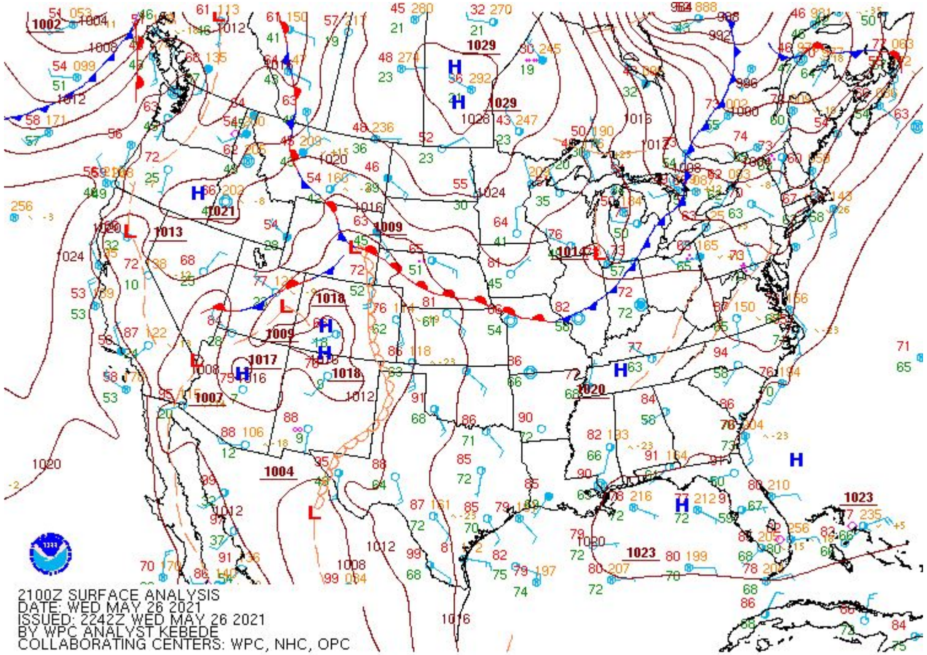
NSF

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Second example - Fronts

Goal of our work: develop a first-guess system to reduce time needed for forecasters to generate surface analyses.

- Identifying fronts is critical to many weather forecasting tasks
- Approach: Build on Lagerquist et al (2019) and develop a deep learning system to identify cold, warm, occluded, and stationary fronts



2100Z SURFACE ANALYSIS
DATE: WED MAY 26 2021
ISSUED: 2242Z WED MAY 26 2021
BY WPC ANALYST KEBEDE
COLLABORATING CENTERS: WPC, NHC, OPC

Justin, Andrew D.; Willingham, Colin; McGovern, Amy; Allen, John T. (in preparation) T*oward Operational Real-time Identification of Frontal Boundaries Using Machine Learning*. To be submitted to Artificial Intelligence for the Earth Systems

Building trust with the end-users

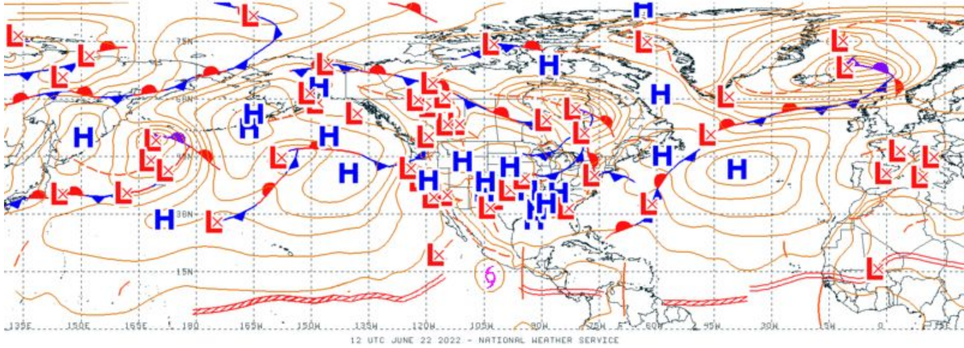
- How can we build trust with our targeted end-users?
    - Forecasters who will use our product to help in real-time with their task of drawing fronts in their domain

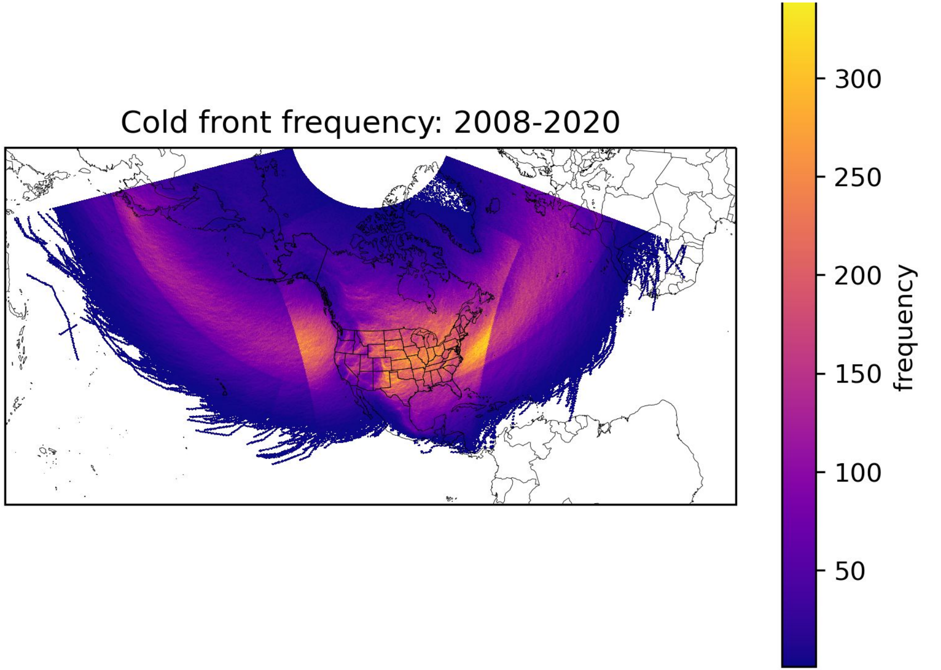
End-user needs

The Unified Surface Analysis is jointly Produced by:

- National Centers for Environmental Prediction (NCEP)
- Weather Prediction Center (WPC)
- Ocean Prediction Center (OPC)
- National Hurricane Center (NHC)
- Pacific Region
- Honolulu Forecast Office (HFO)

Their needs differ by region!



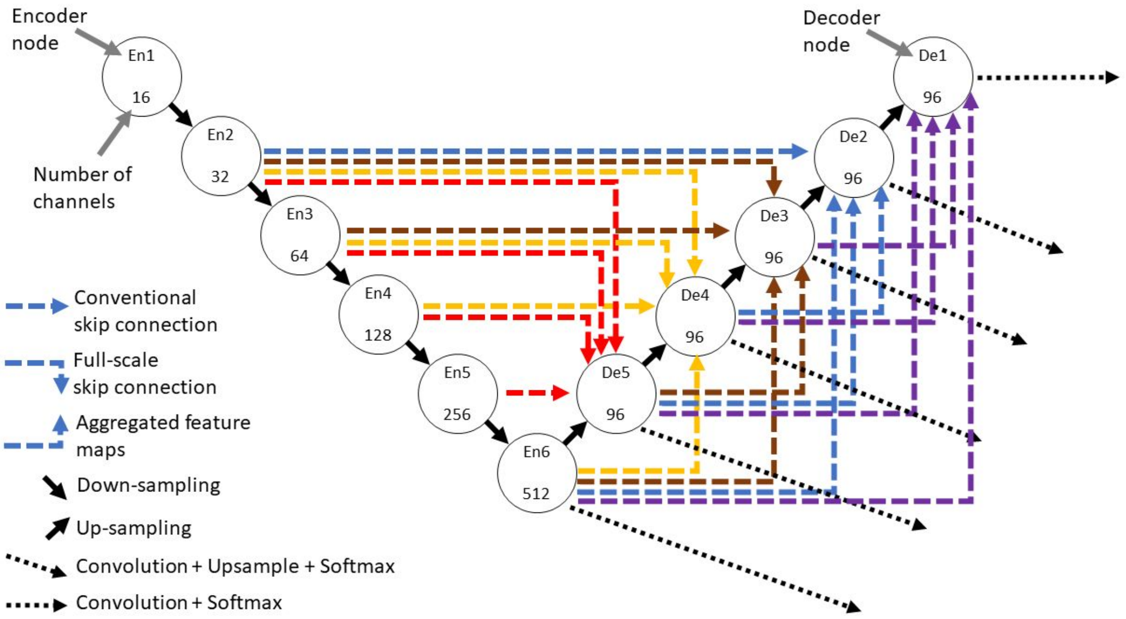


Building trust with the end-users

- How can we build trust with our targeted end-users?
  - Forecasters who will use our product to help in real-time with their task of drawing fronts in their domain
- Ideas:
  - **Explain the underlying ML model**

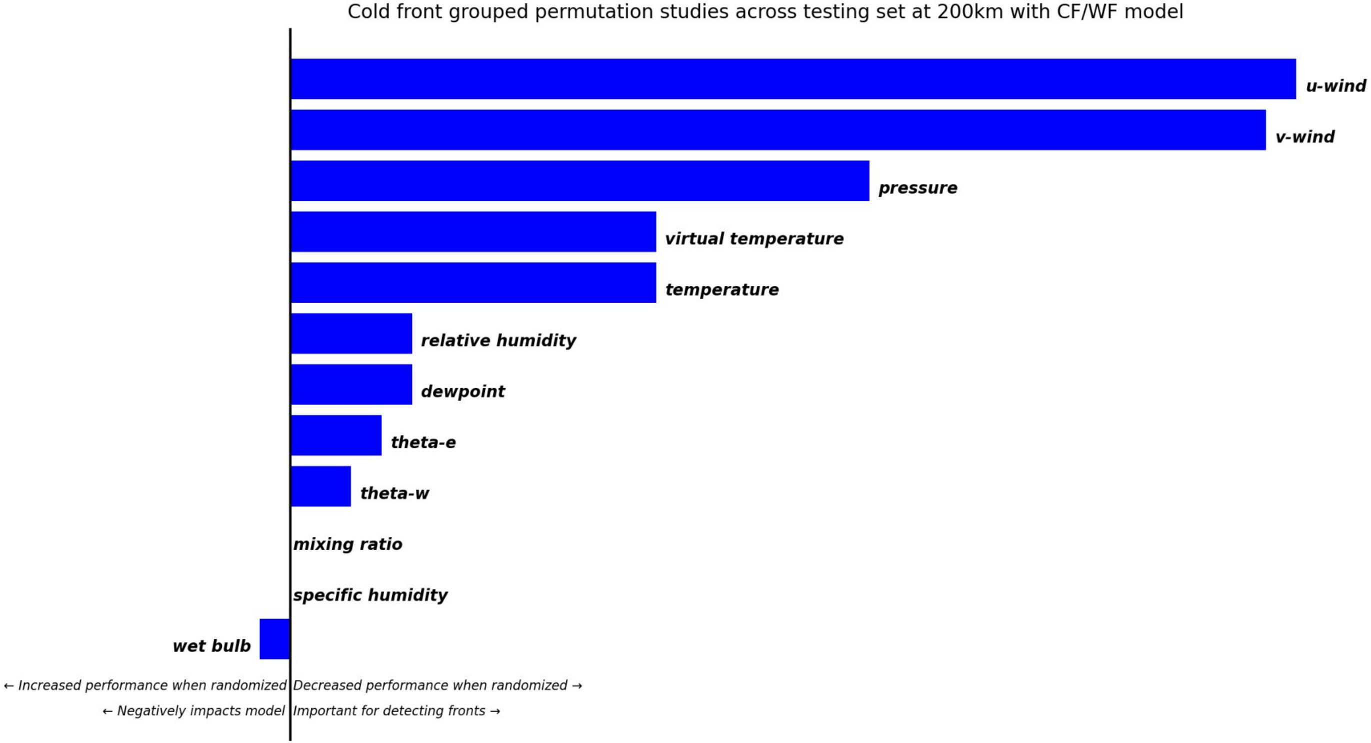
Explaining the ML model

- U-net 3+ model
- 3D inputs of meteorological variables used to distinguish fronts by hand



XAI to peer inside what the model learned

Does the model learn what we expect?

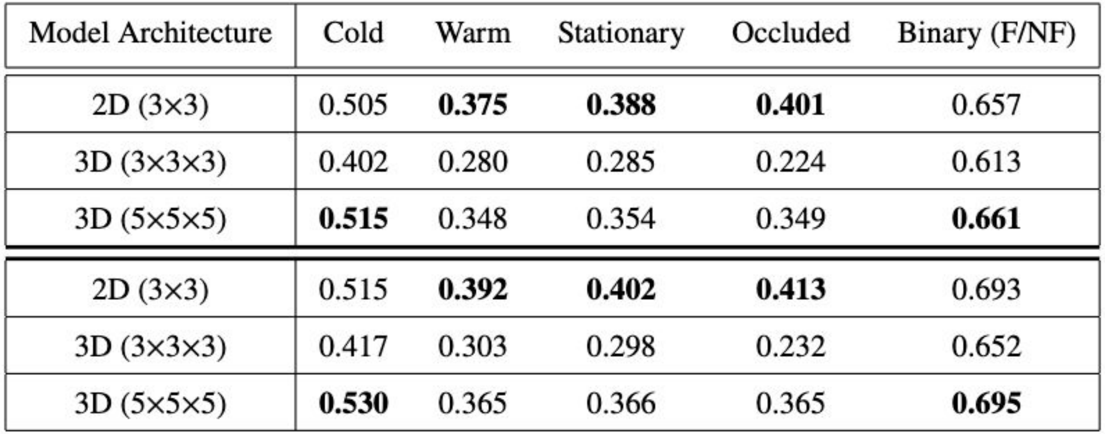


Building trust with the end-users

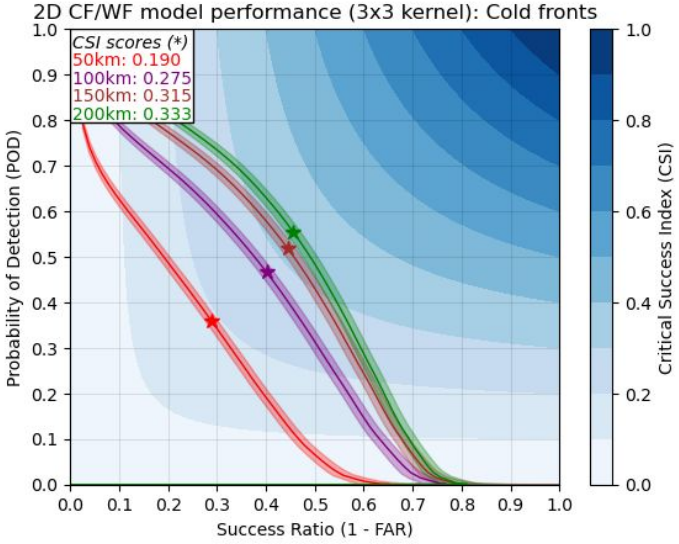
- How can we build trust with our targeted end-users?
    - Forecasters who will use our product to help in real-time with their task of drawing fronts in their domain
- Ideas:
    - Explain the underlying ML model
    - **Objective verification**

Objective verification

Objective evaluation is easy and doable



Model Architecture	Cold	Warm	Stationary	Occluded	Binary (F/NF)
2D (3×3)	0.505	**0.375**	**0.388**	**0.401**	0.657
3D (3×3×3)	0.402	0.280	0.285	0.224	0.613
3D (5×5×5)	**0.515**	0.348	0.354	0.349	**0.661**
2D (3×3)	0.515	**0.392**	**0.402**	**0.413**	0.693
3D (3×3×3)	0.417	0.303	0.298	0.232	0.652
3D (5×5×5)	**0.530**	0.365	0.366	0.365	**0.695**



If you were a forecaster, what else would you want to know about the model before using it?

Building trust with the end-users

- How can we build trust with our targeted end-users?
  - Forecasters who will use our product to help in real-time with their task of drawing fronts in their domain
- Ideas:
  - Explain the underlying ML model
  - Objective verification
  - **Case studies**

Revisiting our earlier questions

Given what you now know about this domain, how do I pick my case study to best improve trust and show forecasters how the model does?

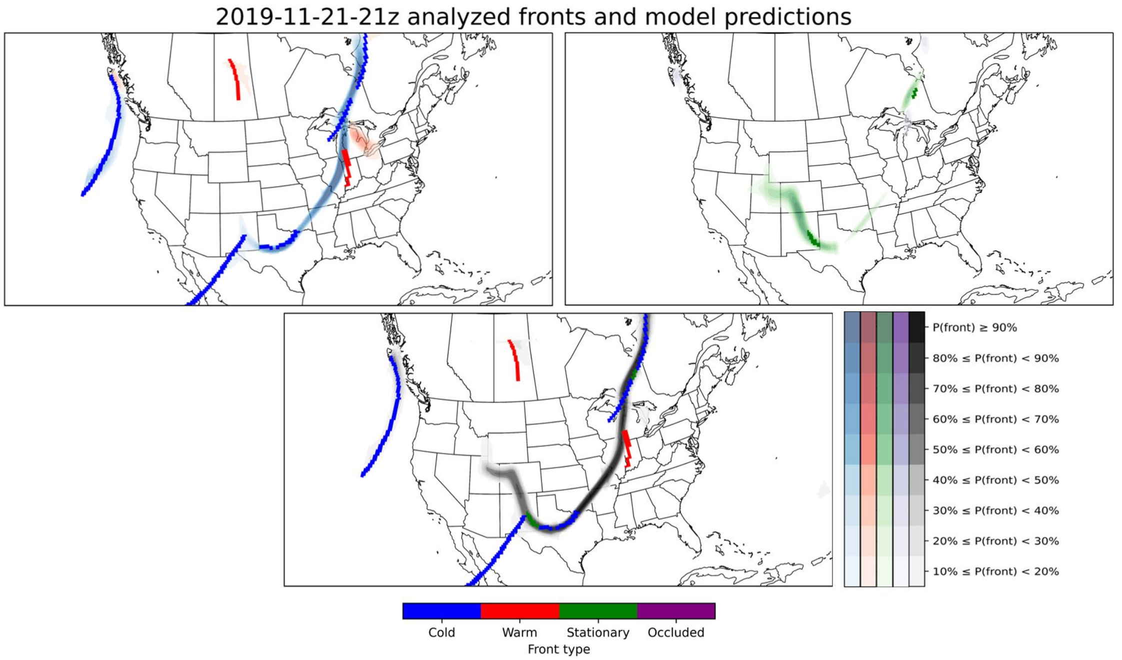
3.9. Go to sli.do and use the code TAI4ES

Building trust with the end-users

- How can we build trust with our targeted end-users?
  - Forecasters who will use our product to help in real-time with their task of drawing fronts in their domain
- Ideas:
  - Explain the underlying ML model
  - Objective verification
  - **Case studies**

Case studies Part 1

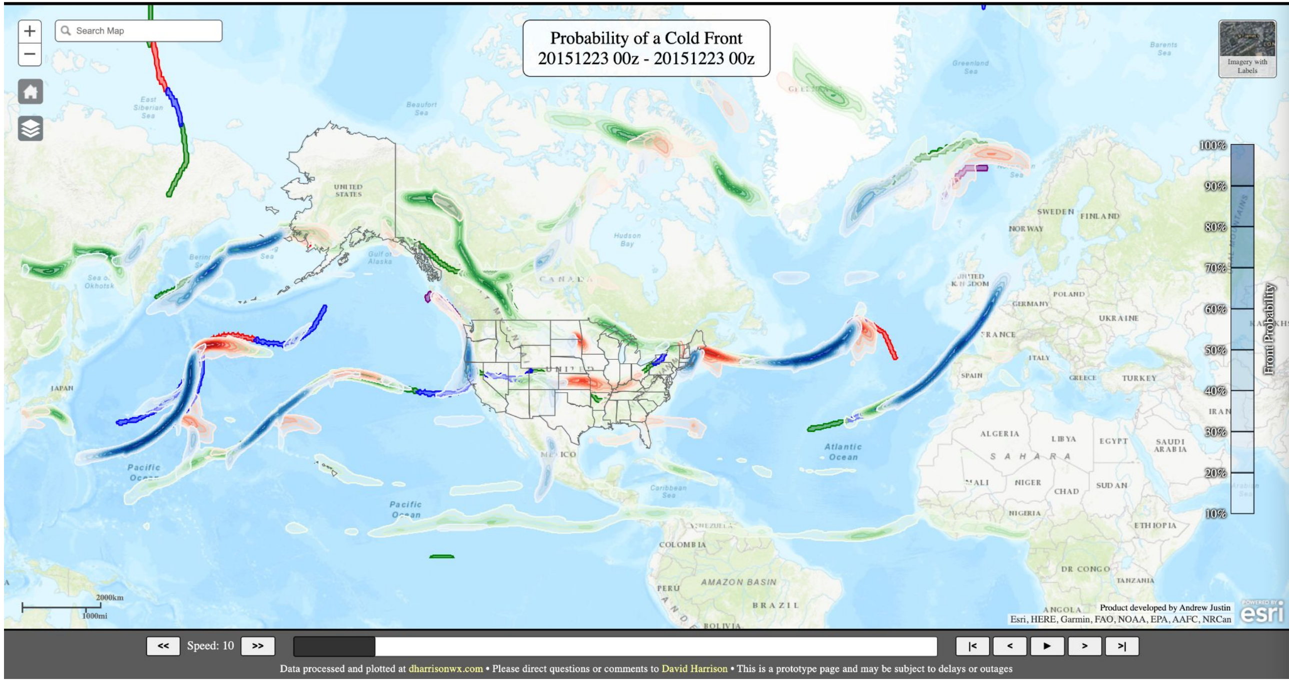
- First pass on case studies:
  - Pick interesting cases with good/bad performance
  - Analyze what the model did well/poorly
- Prototyped the case studies with a forecaster
  - Feedback: Need more data and interaction



2019-11-21-21z analyzed fronts and model predictions

Case studies part 2

- Second pass on case studies:
    - Pick interesting cases
    - Put into interactive web interface
    - Ask forecasters for feedback
- Interviewed forecasters at WPC, OPC, and TAFB
    - N = 9 forecasters



Probability of a Cold Front
20151223 00z - 20151223 00z

Product developed by Andrew Justin
Esri, HERE, Garmin, FAO, NOAA, EPA, AAFC, NRCan

Data processed and plotted at dharrisonwx.com • Please direct questions or comments to David Harrison • This is a prototype page and may be subject to delays or outages

Building trust with the end-users

- How can we build trust with our targeted end-users?
  - Forecasters who will use our product to help in real-time with their task of drawing fronts in their domain
- Ideas:
  - Explain the underlying ML model
  - Objective verification
  - Case studies
  - **Co-production and working with our end-users to understand their needs!**

Co-production and case studies

- Initial results
  - All said they would incorporate the model into their workflow (preferred in NAWIPS)
  - All preferred the labeled fronts over the fronts/no-fronts
  - All preferred probabilistic guidance over deterministic
  - **Forecasters want to see case studies over specific domains where they know the human forecasters struggle - and these domains differ by NWS office!**
  - In discussing areas of disagreement between ML and the forecasters, the forecasters often said they preferred the ML answer
- Key issue: when doing forecasts in real-time, the forecasters have a tremendous time pressure!

Case study choices

Our forecasters checked what happened in the main part of their domain

- And then they focused on the hard parts of their domains!

Examples:

- TAFB wanted a closer look at fronts coming over the Mexican mountains
- WPC wanted a closer look at the upper part of Canada and where fronts hit the Pacific coast
- OPC wanted to look at the poorly sensed parts of the Pacific & Atlantic

General case study advice

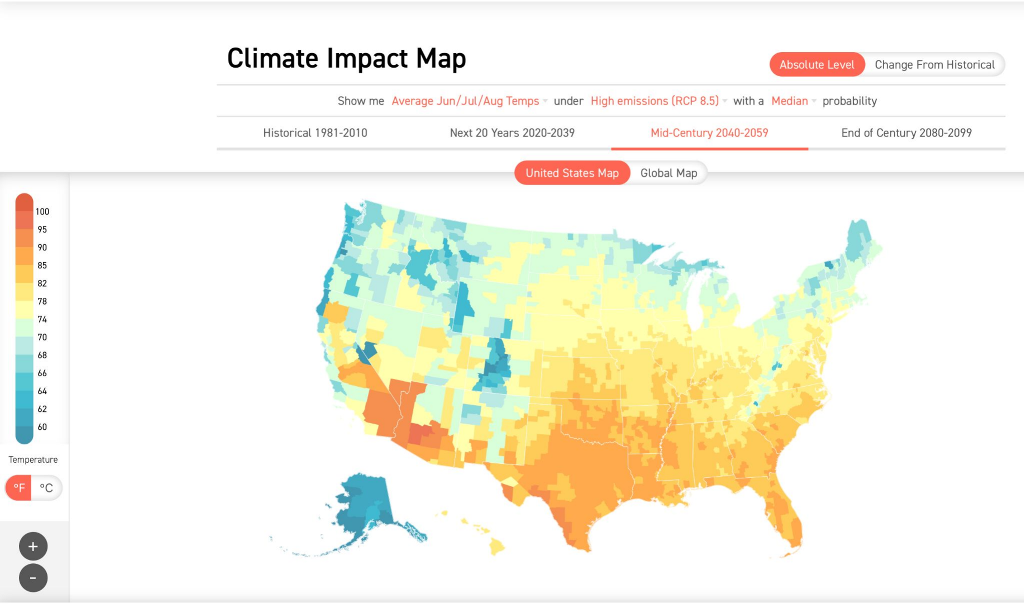
- Provide case studies in the main area of use
  - It is important to know that the AI/ML product works for the majority of the cases
- Trust requires understanding where the model will break also and how it performs under stress!
  - ***Forecaster 5:*** *“If you're saying trustworthiness of a certain product, it would be - how does it **perform on a consistent basis**? No **model is ever going to be perfect**.”*
  - Work with your end-users to identify the challenging scenarios
- Focus on the challenges for the end-users
  - Boundary or edge-cases
  - Rare events
  - Well-known trouble-spots

From AI2ES interviews introduced Day 1 and Day 2 of the Summer School - From the AI2ES Risk Communication team interviews with NWS Forecasters

Case study open questions

How do we do a case study if we don’t know the right answer?

- Does verifying in the past provide enough evidence?
- What do you do when there really are no comparisons?
- Can a case study with no known truth improve trust?



Climate Impact Map

Show me Average Jun/Jul/Aug Temps under High emissions (RCP 8.5) with a Median probability

https://impactlab.org/map/#usmeas=absolute&usyear=2040-2059&gmeas=absolute&gyear=1986-2005

3.10. Go to sli.do and use the code TAI4ES

Case study open questions

How can we use the idea of case studies to improve understanding of the risk and trust in future predictions?

- Are there ethical challenges involved in presenting hypothetical future extreme weather?
- How does hypothetical AI generated scenarios impact trust?

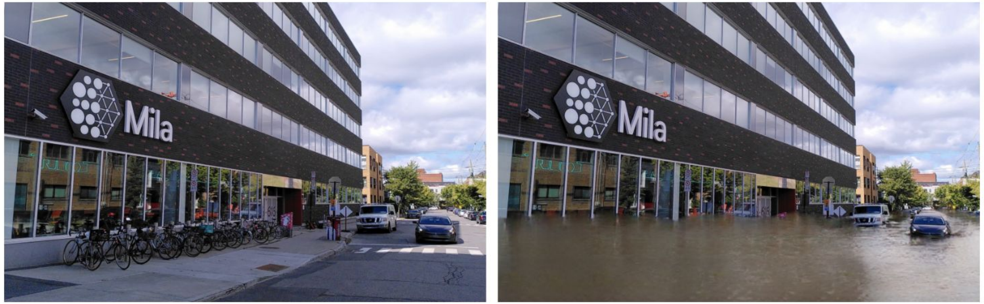


Figure 1: *We present ClimateGAN, a model that simulates extreme floods (right) on real scene images (left).*

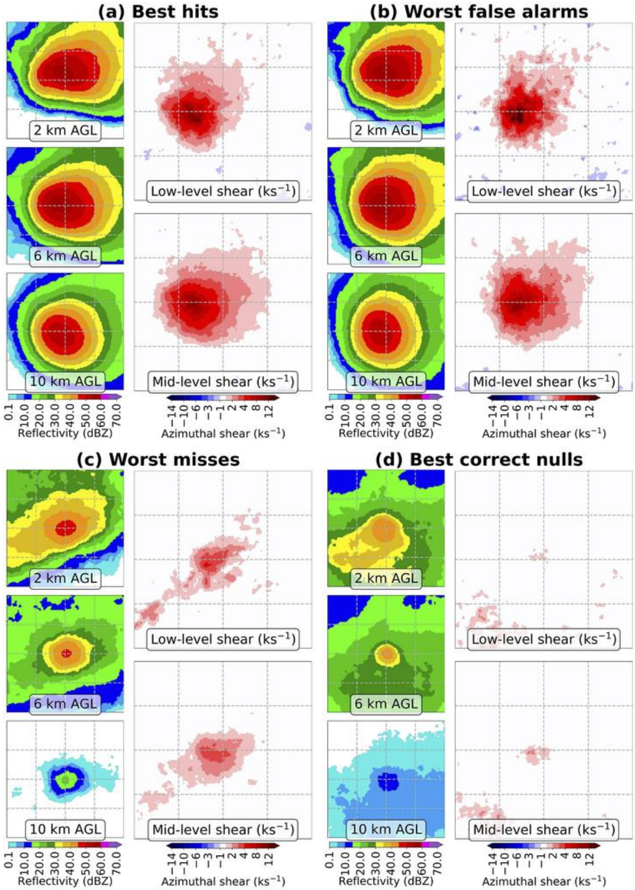
Schmidt, Victor, Alexandra Sasha Luccioni, Mélisande Teng, Tianyu Zhang, Alexia Reynaud, Sunand Raghupathi, Gautier Cosne, et al. 2021. "ClimateGAN: Raising Climate Change Awareness by Generating Images of Floods." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2110.02871

3.11. Go to sli.do and use the code TAI4ES

Case study open questions

How do we connect case studies with XAI and interpretability?

- Connect explanations to a more global view
- Pick specific cases for XAI to see what was important
  - Best hits
  - Worst misses
  - Other challenging or user-requested specific studies

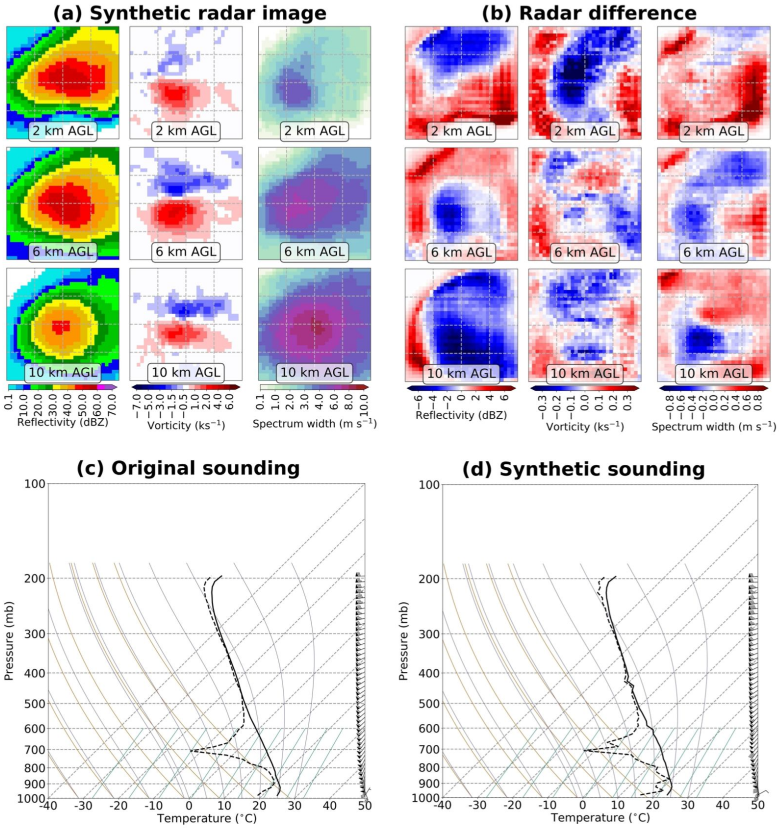


Lagerquist, R., McGovern, A., Homeyer, C. R., Gagne II, D. J., & Smith, T. (2020). Deep Learning on Three-Dimensional Multiscale Data for Next-Hour Tornado Prediction, *Monthly Weather Review*, *148*(7), 2837-2861. Retrieved Jun 23, 2022, from https://journals.ametsoc.org/view/journals/mwre/148/7/mwrD190372.xml

Case study open questions

How do we connect case studies with XAI and interpretability?

- Generate synthetic examples
  - Example uses backwards optimization, an XAI technique not yet discussed



[Ryan Lagerquist](#) (2020): *Using Deep Learning to Improve Prediction and Understanding of High-Impact Weather*. PhD Thesis, School of Meteorology, University of Oklahoma.

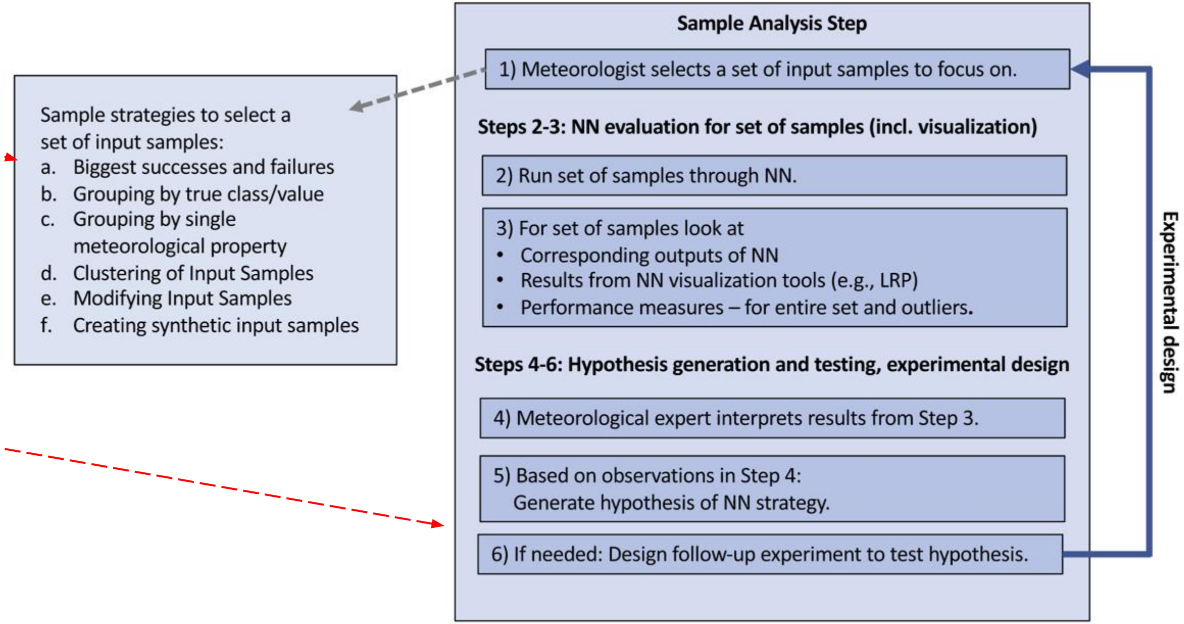
Sample workflow using case studies and XAI

Sample workflow to construct sets of sample, generate hypothesis and test hypothesis.

Ideas on how to select or construct sets of samples for initial XAI experiments

Ideas on how to **test hypothesis**:

- Select or construct a ***new*** set of samples and predict what you ***should*** see for them in NN output, XAI, etc..
- Check whether your predictions are correct: back to Step 1!



Sample strategies to select a set of input samples:

a. Biggest successes and failures
b. Grouping by true class/value
c. Grouping by single meteorological property
d. Clustering of Input Samples
e. Modifying Input Samples
f. Creating synthetic input samples

Sample Analysis Step

1) Meteorologist selects a set of input samples to focus on.

Steps 2-3: NN evaluation for set of samples (incl. visualization)

2) Run set of samples through NN.

3) For set of samples look at
- Corresponding outputs of NN
- Results from NN visualization tools (e.g., LRP)
- Performance measures – for entire set and outliers.

Steps 4-6: Hypothesis generation and testing, experimental design

4) Meteorological expert interprets results from Step 3.

5) Based on observations in Step 4:
Generate hypothesis of NN strategy.

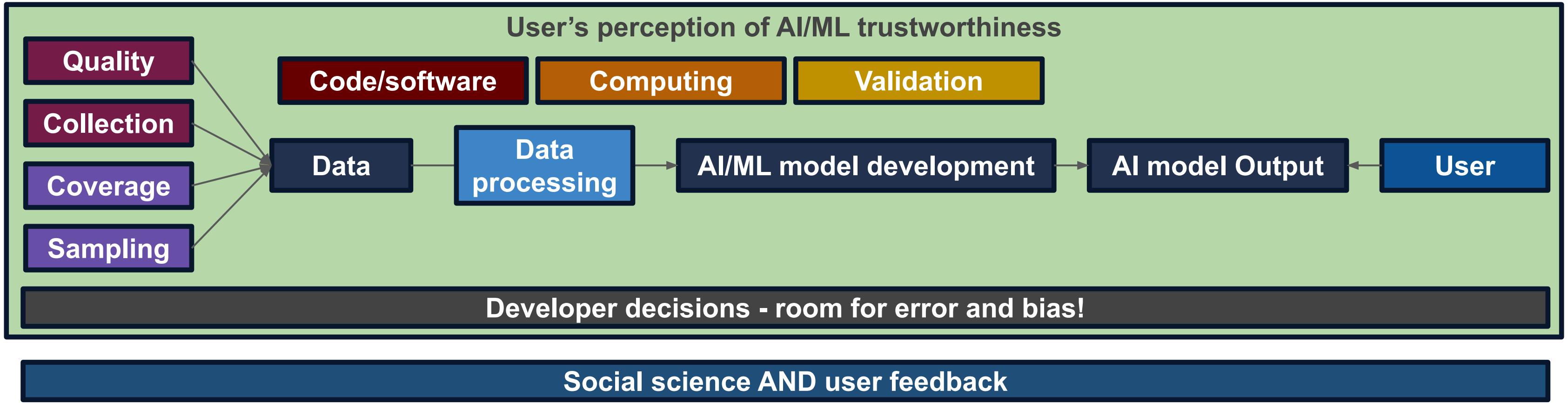
6) If needed: Design follow-up experiment to test hypothesis.

For detailed description of this workflow, see the paper below.

Ebert-Uphoff, I., & Hilburn, K. (2020). Evaluation, Tuning, and Interpretation of Neural Networks for Working with Images in Meteorological Applications, Bulletin of the American Meteorological Society, 101(12), E2149-E2170. Retrieved Jun 28, 2022, from https://journals.ametsoc.org/view/journals/bams/101/12/BAMS-D-20-0097.1.xml

Conclusion - where do we go from here?

- Work to increase the trustworthiness of our data and workflows
- Clearly communicate any limitations in the data and workflow
- Co-produce with end-users and social scientists throughout
- Understand and communicate constraints and potential fail safes



User's perception of AI/ML trustworthiness

Quality

Collection

Coverage

Sampling

Code/software

Computing

Validation

Data

Data processing

AI/ML model development

AI model Output

User

Developer decisions - room for error and bias!

Social science AND user feedback

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Breathing break - Any open questions? Add them in Slido!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do
and use the
code TAI4ES

NSF

ARTIFICIAL INTELLIGENCE · ENVIRONMENTAL SCIENCE · RISK COMMUNICATION

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR
ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 3: Agenda

- 9:00 Trustworthiness of data and workflows
- 10:30 *Brain & bio break*
- 10:45 The importance of case studies and tips for using them effectively

Time for any open questions!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do and use the code TAI4ES

ARTIFICIAL INTELLIGENCE · RISK COMMUNICATION · ENVIRONMENTAL SCIENCE

NSF

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT

Thank you!

- This material is based upon work supported by the National Science Foundation under Grant No. ICER-2019758.
- This summer school is being supported by NCAR/UCAR
- Thank you to:
  - Taysia Peterson and the multi-media team @ NCAR
  - Susan Dubbs @ OU
  - Our sponsors! NCAR/UCAR, Google cloud, LEAP, Radiant Earth
  - All of our guest speakers
  - All of you for coming and participating!

NSF

ARTIFICIAL INTELLIGENCE · ENVIRONMENTAL SCIENCE · RISK COMMUNICATION

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR
ARTIFICIAL INTELLIGENCE

Radiant Earth
Foundation
EARTH IMAGERY FOR IMPACT

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Looking ahead to Day 4: Agenda

- 9:00 Uncertainty quantification methods (Part 1)
- 9:55 *Short brain & bio break*
- 10:05 Uncertainty quantification methods (Part 2)
- 10:45 *Short brain & bio break*
- 10:55 Communicating uncertainty
- 11:55 Lecture series wrap up!

Questions?

https://app.sli.do/event/1zumy91n

Or go to sli.do
and use the
code TAI4ES

ARTIFICIAL INTELLIGENCE · ENVIRONMENTAL SCIENCE · RISK COMMUNICATION

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

LEAP

NOAA CENTER FOR
ARTIFICIAL INTELLIGENCE

Radiant Earth Foundation
EARTH IMAGERY FOR IMPACT