

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Amy McGovern

Lloyd G. and Joyce Austin Presidential Professor
School of Computer Science and School of Meteorology
University of Oklahoma

Director, NSF AI Institute for Research on Trustworthy AI in
Weather, Climate, and Coastal Oceanography (AI2ES)
@profamymcgovern



NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)

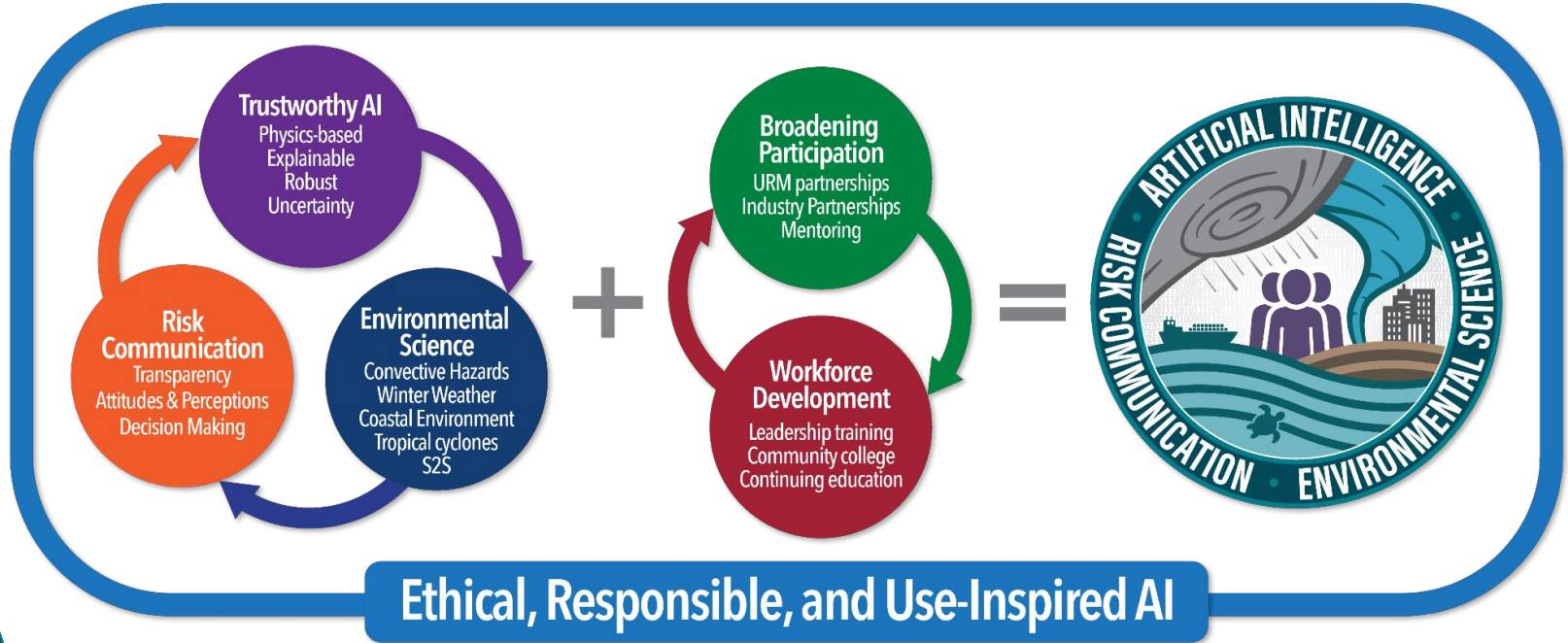
AI2ES will uniquely benefit humanity by developing *novel, physically based* AI techniques that are demonstrated to be *trustworthy*, and will directly improve *prediction, understanding, and communication* of high-impact environmental hazards.



ai2es.org



AI2ES

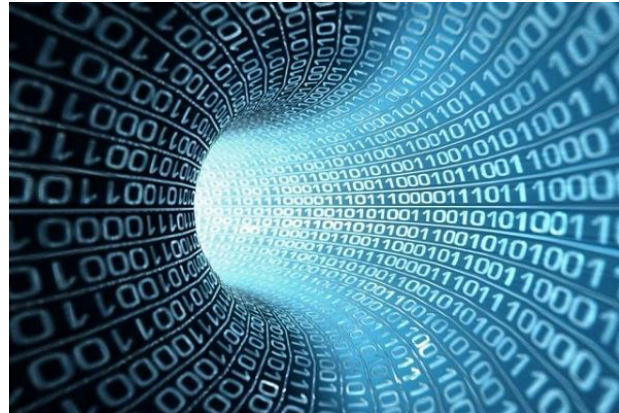


ai2es.org



Summer School Goals

- Share AI2ES work on trustworthy AI with researchers and students worldwide
- Engage researchers outside of AI2ES normal collaborators in work on trustworthy AI
- Build the global community of trustworthy AI researchers
- Serve as a nexus for trustworthy AI



[lucky_sun/Flickr/CC BY-SA 2.0](https://www.flickr.com/photos/lucky_sun/)

Summer School & Trust-a-thon

- Intertwined lectures and hands-on activities
 - Morning foundational topics on trust and communication
 - Afternoon trust-a-thon hands-on topics on trust in AI for ES models
- Each day has a theme
 - Day 1: Trust, Interdisciplinary research, XAI Part 1
 - Day 2: Explainability, Interpretability, XAI Part 2
 - Day 3: Trust and Data
 - Day 4: Uncertainty Quantification
- Morning topics will help with the afternoon goals in the trust-a-thon



Trust-a-thon

David John Gagne



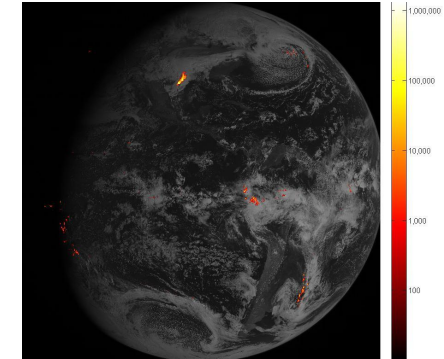
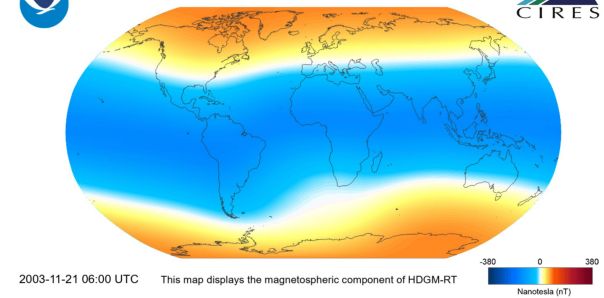
Trust-a-thon's unique idea

Key idea: Let's APPLY the AI2ES approach and see how we can learn to think about our end users and building their trust in an ML method for earth sciences!

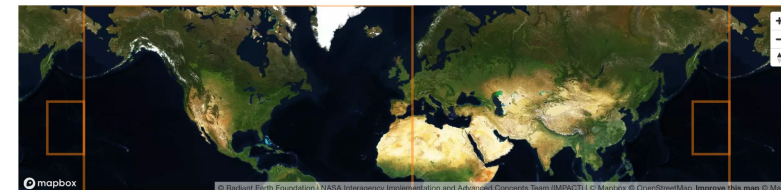
- Three datasets
 - Severe weather, space weather, tropical cyclones
- Each dataset has a set of “personas”
 - These are short descriptions of an end user that you will be developing your model for and thinking about how best to communicate about it with



High Definition Geomagnetic Model - Real Time



Tropical Cyclone Wind Estimation Competition



Trust-A-Thon Logistics

- Participants are randomly assigned to a team, challenge problem, and user persona with groupings by time zone
- Each challenge problem has a slack channel
 - 1-4pm MT will be the main time that each slack channel is monitored though there may be additional monitoring in other hours
- Each day has specific assignments for each challenge problem and user persona
 - The assignments are just there to give you ideas and get you going. You do **not** have to follow the assignments.
 - Each team will write a blog post for each day
- All teams invited to a final zoom Thursday 6/30 3-4pm MT to discuss in break-out rooms with organizers (not a formal presentation, just a discussion!)



Trustathon Organizers

Overall

- David John Gagne
- Chris Wirz
- Taysia Peterson
- Jennifer Warrilow

Space Weather

- Rob Redmon
- Manoj Nair
- LiYin Young

Severe

- Monte Flora
- Randy Chase

Tropical

- Jason Stock
- Marie McGraw
- Akansha Singh Bansal
- Imme Ebert-Uphoff
- Hamed Alemohamad
- Renee Pieschke



Summer School Logistics

Amy McGovern



Summer School Code of Conduct

UCAR and NCAR are committed to providing a safe, productive, and welcoming environment for all participants in any conference, workshop, field project or project hosted or managed by UCAR, no matter what role they play or their background. This includes respectful treatment of everyone regardless of gender, gender identity or expression, sexual orientation, disability, physical appearance, age, body size, race, religion, national origin, ethnicity, level of experience, political affiliation, veteran status, pregnancy, genetic information, as well as any other characteristic protected under state or federal law.

All participants (and guests) are required to abide by this Code of Conduct. This Code of Conduct is adapted from the one adopted by AGU, complies with the new directive from the National Science Foundation (NSF) and applies to all UCAR-related events, including those sponsored by organizations other than UCAR but held in conjunction with UCAR events, in any location throughout the world.

The full Code of Conduct document (also linked on the summer school site)

<https://drive.google.com/file/d/102qyd0YNnA-7EN19HZ2KEuQhxxk9f6lr/view?usp=sharing>



Schedule and logistics

- All talks are broadcast live and recorded (to be posted ASAP)
 - <https://www2.cisl.ucar.edu/events/tai4es-2022-summer-school>
- Questions should be submitted through slido
 - The chairs for that day will moderate questions and send them to the speaker
- Morning schedule is 9-12 MT with brain breaks
- Afternoon trust-a-thon is 1-4 MT

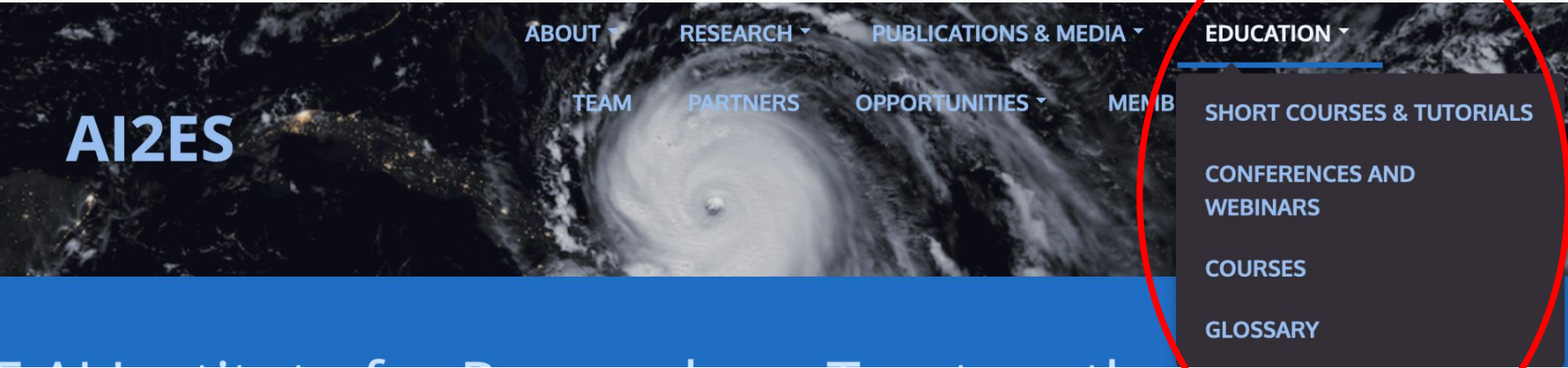


<https://app.sli.do/event/1zumy91n>

Or go to [sli.do](https://app.sli.do) and use the code
TAI4ES



Resources for learning more



ai2es.org



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!



Thanks to all the lecture series speakers!



Amy
McGovern
(OU)



David John
Gagne
(NCAR)



Imme
Ebert-Uphoff
(CSU)



Ann
Bostrom
(UW)



Christopher
Wirz
(NCAR)



Douglas
Rao
(NOAA)



Andrea
Schumacher
(CIRA)



Montgomery
Flora
(CIWRO)



Mariana
Cains
(NCAR)



Randy
Chase
(OU)



Antonios
Mamalakis
(CSU)



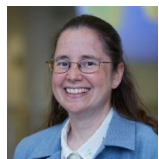
Marie
McGraw
(CSU)



Ryan
Lagerquist
(CSU)

Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 1: Speakers



Amy McGovern
(OU)



David John Gagne
(NCAR)



Ann Bostrom
(UW)



Christopher Wirz
(NCAR)



Andrea Schumacher
(CIRA)



Montgomery Flora
(CIWRO)



Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 1: Goals & Objectives

- Learn about the nature of trust and trustworthiness in AI
- Learn about doing meaningful interdisciplinary work
- Learn about evaluation metrics and preliminary XAI methods for evaluation trust
- Apply these to your trust-a-thon data



Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 1: Agenda

- 9:00 Welcome and Overview
- **9:10 What does it mean to trust?**
- 9:40 *Short brain & bio break*
- 9:45 Meaningful interdisciplinary work
- 10:25 *Short brain & bio break*
- 10:30 Evaluation metrics
- 11:00 XAI for traditional ML



What does it mean to trust?

From a social science and interdisciplinary
perspective

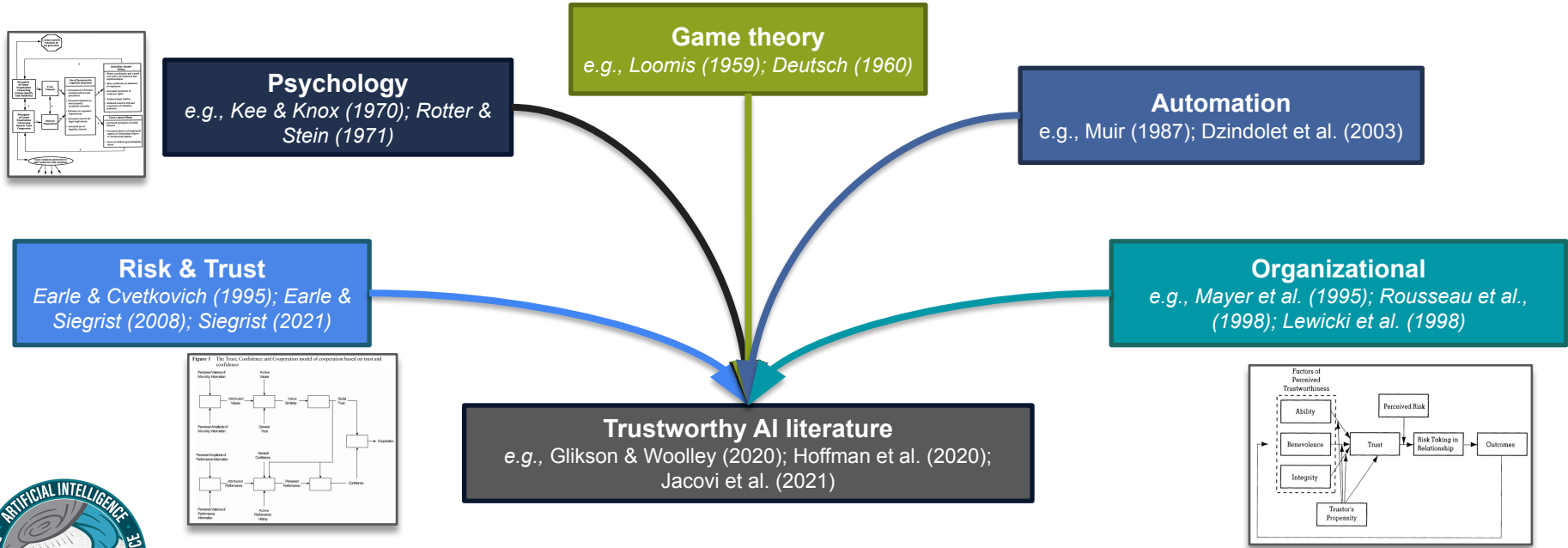


In your own words, what does
“trustworthy AI”
mean to you?

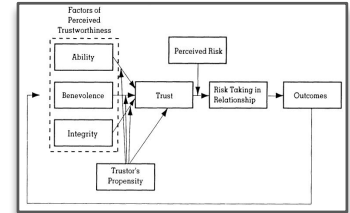
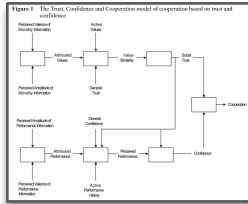
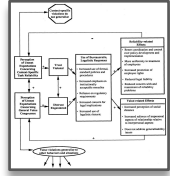


1.1. Go to sli.do and use the code TAI4ES

The trust family tree



Rich, variable, and context-dependent



Definition of Trust

- Trust is the **willingness of a party to be vulnerable to the actions of another** party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. (e.g., Mayer et al 1995)
- Trust: In the presence of uncertainty, **the degree to which someone does or does not rely on**, or put faith in, someone or something (Wirz et al.)
 - Definition is purposefully broad, so as to capture the many different definitions and related dimensions of trust. Our definition of trust is designed to capture trust in all forms.
- Trust is the **relationship between a trustor and a trustee**: the trustor trusts the trustee. Trust is dynamic, evolves with interactions, and is easier to lose than gain.

AI2ES Definition: Trust is the willingness to assume risk by relying on or believing in the actions of another party.

Trust is contextually dependent - always think about...

- **Actors:** Who is being expected to trust?
- **Targets:** What are they being expected to trust?
- **Purpose:** What should they trust something/someone for?
- **Reason:** Why should they trust someone/something?
- **Setting:** In what place or role are they being asked to trust?



Trust is subjective.

- Trustworthiness is in the eye of the beholder – it's **subjective**. In our case the user.
- So we need to be careful about making **normative, potentially biased**, statements
- We decide what's trustworthy for ourselves - **not** for others



So what does this mean for trustworthy AI?

So we know trust is relational, context-dependent, and subjective.

In the next section we review how AI2ES has defined trustworthiness and put it into the context of some of our empirical work.



Definition of Trustworthiness

- Trust is relational - there is an actor (trustor) and target (trustee)
 - Who or what am I trusting, and what am I trusting it for?
- Trustworthiness is evaluative - why should I trust you?

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.



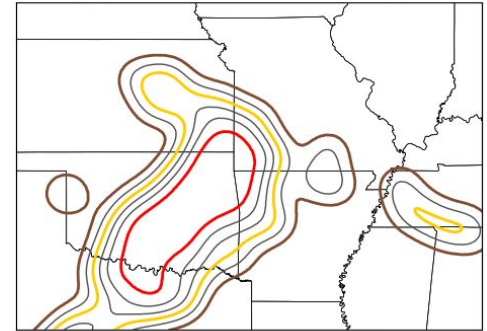
Overview of Interview Process

Interviewed National Weather Service Forecasters:

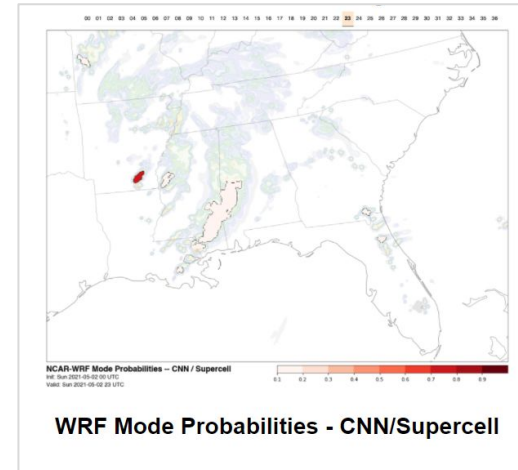
- 16 forecasters from Central, Southern, and Eastern Regions
- 7 GS 5-12 meteorologists, 4 lead meteorologist, and 5 science & operations officers

Topics covered in the interviews:

- Perceptions of and attitudes toward AI and AI trustworthiness
- Perceptions of and feedback about AI/ML convective forecast guidance for two products



Probability of hail: Burke *et al.*, 2020



Storm mode: Sobash *et al.*, in prep.



Results: Trustworthiness Theme 1 – Overall Guidance Performance

Forecaster 4: “I would say for trustworthiness in this scenario would be **how did this technique or output verify** [for] previous events?”

Forecaster 5: “If you're saying trustworthiness of like a certain product, it would be **how does it perform on a consistent basis**. You know, no model is ever going to be perfect.”

Forecaster 9: “Obviously **we don't expect the models to be 100% perfect** because otherwise I would have no job.”

Trustworthiness:

1. Performance overall (consistency, verification) and contextually (in certain/different situations)
2. Relevance and utility of guidance
3. Familiarity and experience with guidance



****Note: Themes are intersectional, not mutually exclusive***



Results: Trustworthiness 2 – Relevance and Utility of Guidance

Forecaster 1: “But examples of how it could have **improved the forecast specifically in terms of timing, location and intensity**. Because, ultimately, that's what our end-user cares the most about.”

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]**.”

Trustworthiness:

1. Performance overall (consistency, verification) and contextually (in certain/different situations)
2. **Relevance and utility of guidance**
3. Familiarity and experience with guidance



Results: Trustworthiness 3 – Familiarity and Experience with Guidance

Forecaster 2: “As I experience the new guidance with multiple events, my trust would go up as it's doing well. And if, like, a new guidance misses something or leads me astray, my trust would go way down.”

Forecaster 8: “So, if we know about the [guidance’s] nuances and the biases upfront, perhaps we can develop a better tool set and approach to those tools, given what they're about.”

Forecaster 10: “When I think of trustworthy, I think of [the guidance] being predictable in its ways that you know that’s not going to work out so you can make mental adjustments.”

Trustworthiness:

1. Performance overall (consistency, verification) and contextually (in certain/different situations)
2. Relevance and utility of guidance
3. **Familiarity and experience with guidance**



Activity!

Answer the questions in the survey with this prompt in mind:

To what extent do you agree or disagree that trustworthy AI models for environmental sciences means that...



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

The AI2ES Risk Communication team is working on a (re)conceptualization and argue:

Trustworthiness is...

1. about more than performance
2. is a subjective evaluation made by the user
3. context-dependent



Ethics, and bias

“one reason to desire trust is an ‘almost necessary’ condition on ethical action: that the user has a reasonable belief that the system (whether human or machine) will behave approximately as intended.” (Danks, AIES’19)

- Both **bias** and **uncertainty** (including error, or noise) can cause a system to behave in unintended ways.



Ethics, and bias

“one reason to desire trust is an ‘almost necessary’ condition on ethical action: that the user has a reasonable belief that the system (whether human or machine) will behave approximately as intended.” (Danks, AIES’19)

- Both **bias** and **uncertainty** (including error, or noise) can cause a system to behave in unintended ways.
- More broadly, whether an action is ethical may depend on either the process or the outcomes of the action:
 - utility/benefits (consequentialism),
 - whether it is virtuous/the right thing to do (virtue ethics), or
 - whether it is required by moral principles or duties (deontological ethics)
- Honesty is a deontological imperative, to respect others’ rights and dignity, and the autonomy of their will. “Be honest” is also a virtue rule.



Ethics, and bias

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 CO2 emissions due to computing



Ethics, and bias

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 CO2 emissions due to computing



Ethics, and bias

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:

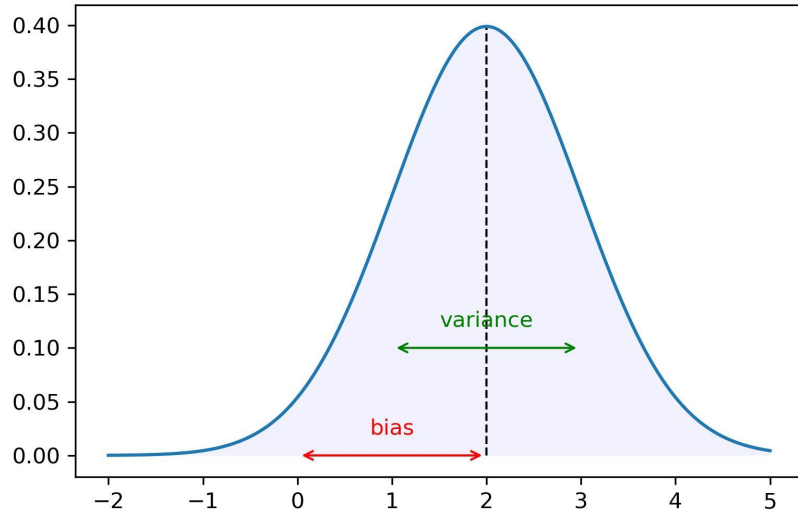
1. Model training choices
2. Algorithm learns faulty strategies
3. AI learns to fake something plausible
4. AI model used in inappropriate situations
5. Non-trustworthy AI model deployed
6. Lack of robustness in the AI model

Other issues related to workforce and society:

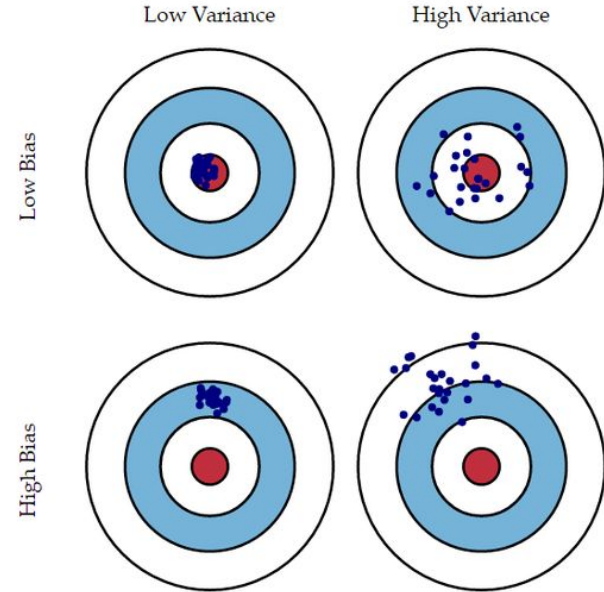
1. Globally applicable AI approaches may stymie burgeoning efforts in developing countries.
2. Lack of input or consent on data collection and model training
3. Scientists might feel disenfranchised.
4. Increase of CO2 emissions due to computing



Bullseye's diagram for bias/variance tradeoff



Lopez de Prado, 2020



<https://towardsdatascience.com/tradeoffs-how-to-aim-for-the-sweet-spot-c20b40d5e6b6>

Bias Breakdown

- Different biases can be introduced during model development and deployment
- In the following slides, we will discuss 3 forms of bias, which can affect AI for ES
 - Computational/Model Bias
 - Data Bias
 - Decision-Making Bias
- This is not an exhaustive treatment of bias as additional biases may exist!



Computational/Model Bias

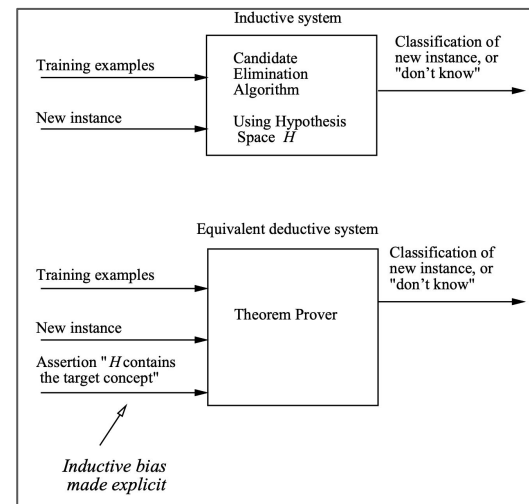
- **Estimation/Mathematical bias** (0 is perfect) → bias-variance tradeoff
 - **Decomposition of Error** = **Approximation Error** + **Estimation Error**
 - **Approximation Error (bias)**: what's the best we can do w/ our hypothesis space (given what model we've chosen)
 - **Estimation Error (variance)**: error on top of the approximation error for our solution; a consequence of the search method we use, essentially.

- **Frequency Bias:**

- Does the frequency of forecast “yes” match the frequency of observed “yes” (1 is perfect)

- **Inductive bias (B):**

- assumptions that facilitate generalization to unseen data
- (for all x_i in X)
[(B and D_c and x_i) implies deductively $L(x_i, D_c)$] for an inductive learner L



Computational/Model Bias

734 Pierre Geurts

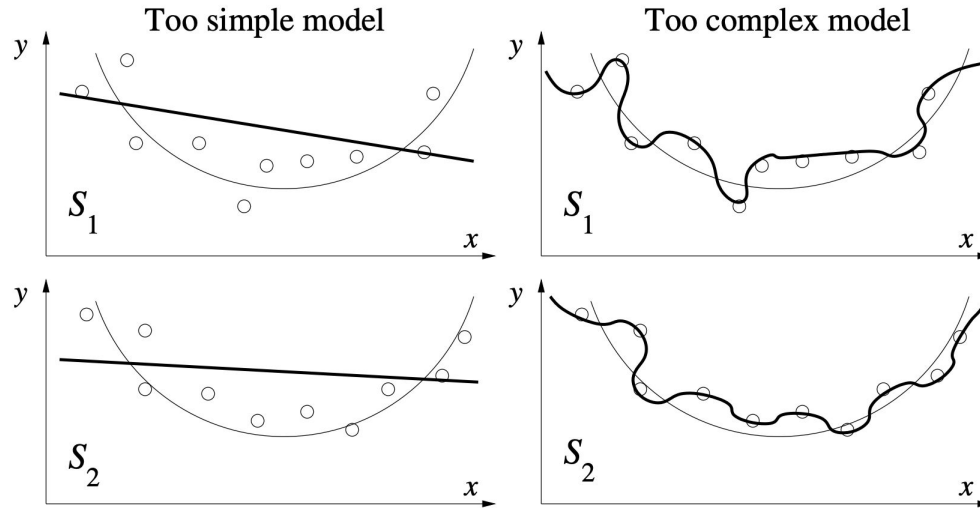
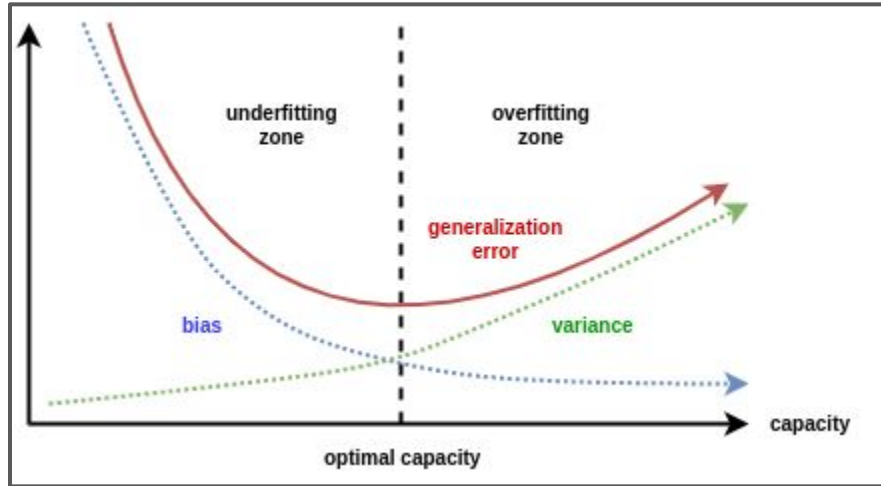


Fig. 37.1. Left, a linear model fitted to two learning samples. Right, a neural network fitted to the same samples

Bias vs variance decomposition

Bias-Variance Tradeoff

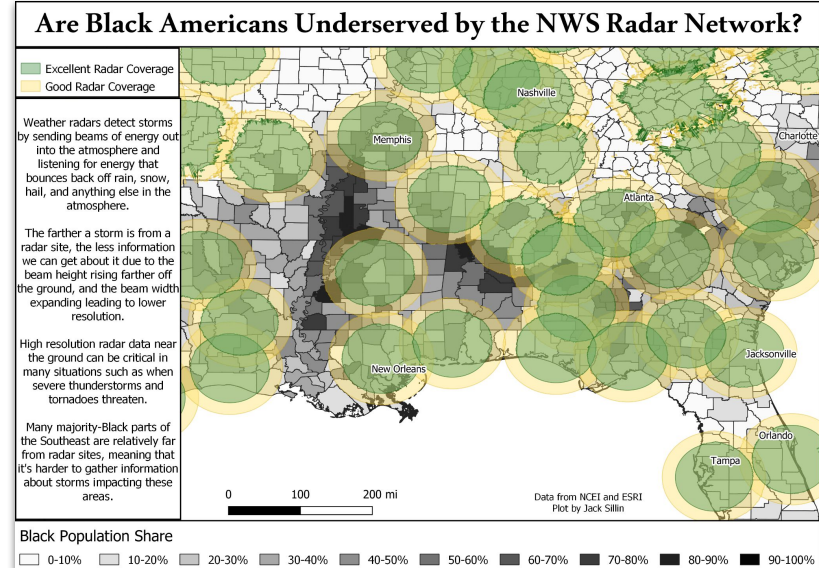


<https://towardsdatascience.com/tradeoffs-how-to-aim-for-the-sweet-spot-c20b40d5e6b6>

- To capture regularities in a dataset often requires a low-bias prediction.
- Lowering the prediction bias comes at the cost of increasing the variance of the model prediction.
- Increasing the prediction variance tends to increase overfitting (i.e., lack of generalization to unseen data).

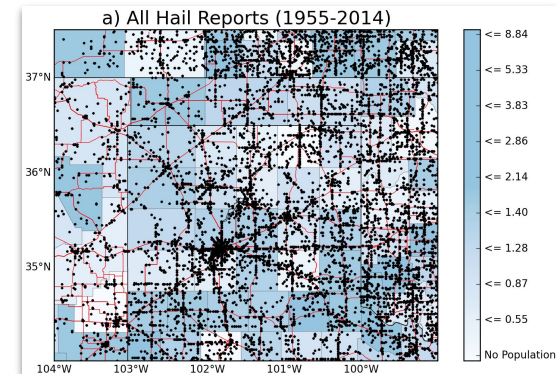
Data Bias

- The data itself can contain biases, which affect the AI/ML model
 - Biases could be caused by underlying human biases (e.g. unintentional or intentional)
 - Biases can be caused by sampling and selection of data
- Potential definition:
 - A class imbalance or distortion in the data from what we know is true based on meteorological and other knowledge about parameters of interest



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

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



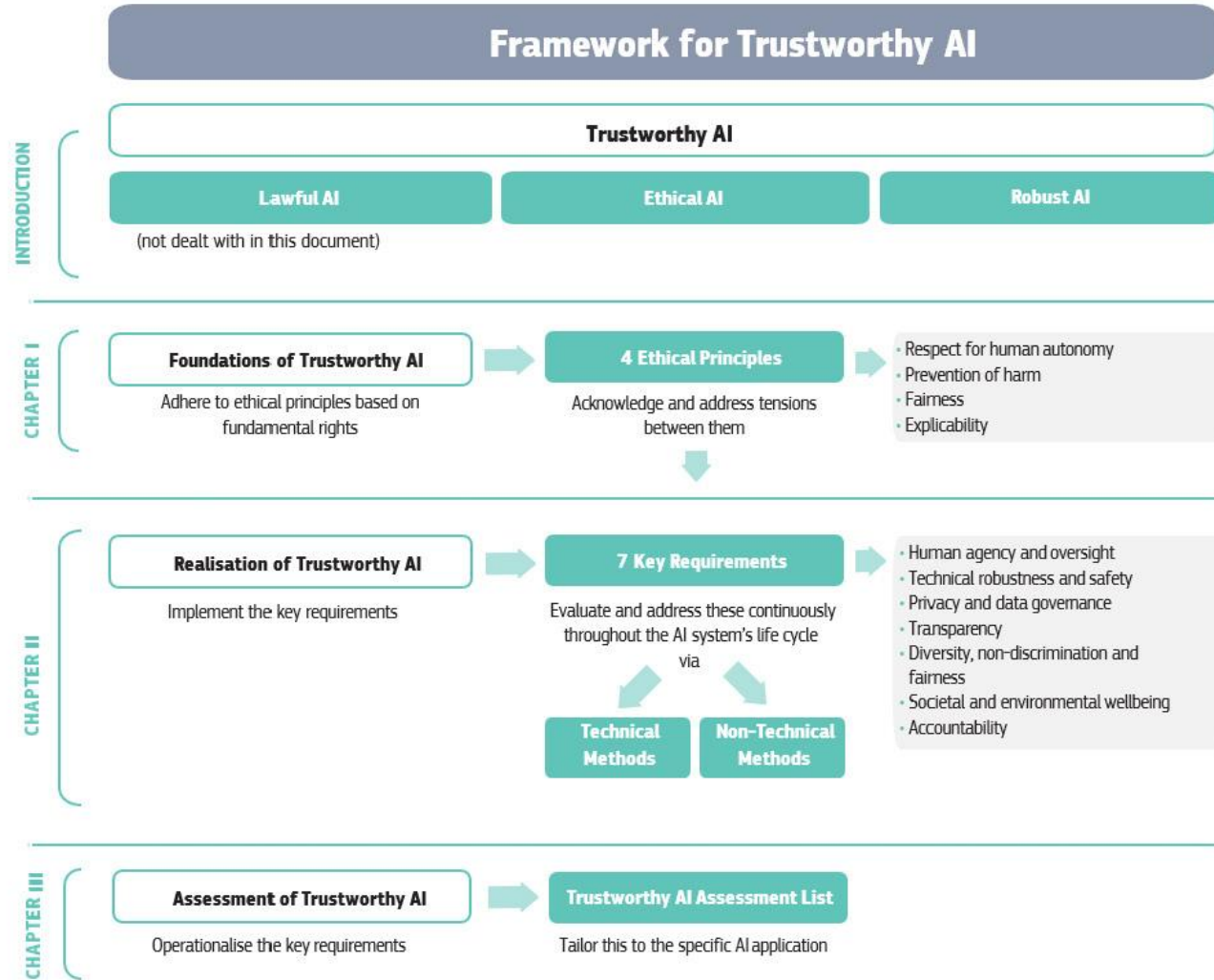
Decision-Making Bias

- Heuristics in human decision making
 - Perceptual biases - motion, color, orientation (Wolfe, Psych Bull & Rev 28[4], 2021)
 - Memory biases -
 - working memory (Miller's "*magical number seven plus or minus two*")
 - categorization biases, determined in part by expertise
 - Attribute substitution (Kahneman & Frederick, 2002), such as -
 - Representativeness heuristic
 - Affect heuristic
 - Anchoring and adjustment, for example -
 - familiarity, salience
 - Example: preference to use models and tools that are familiar to you
- Systemic biases stemming from social norms and institutions also affect decision making.

Ethics and Bias

Ethical Principles provide a foundation for trustworthy AI, as illustrated by the European Commission HLEG on AI 2019 Ethics Guidelines for Trustworthy AI.

Figure 1. The Guidelines as a framework for Trustworthy AI



Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 1: Agenda

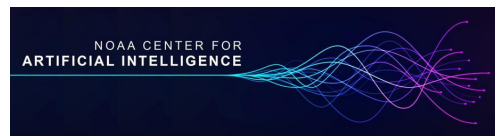
- 9:00 Welcome and Overview
- 9:10 What does it mean to trust?
- **9:50 *Short brain & bio break***
- **9:55** Meaningful interdisciplinary work
- 10:25 *Short brain & bio break*
- 10:30 Evaluation metrics
- 11:00 XAI for traditional ML

Questions?



<https://app.sli.do/event/1zummy91n>

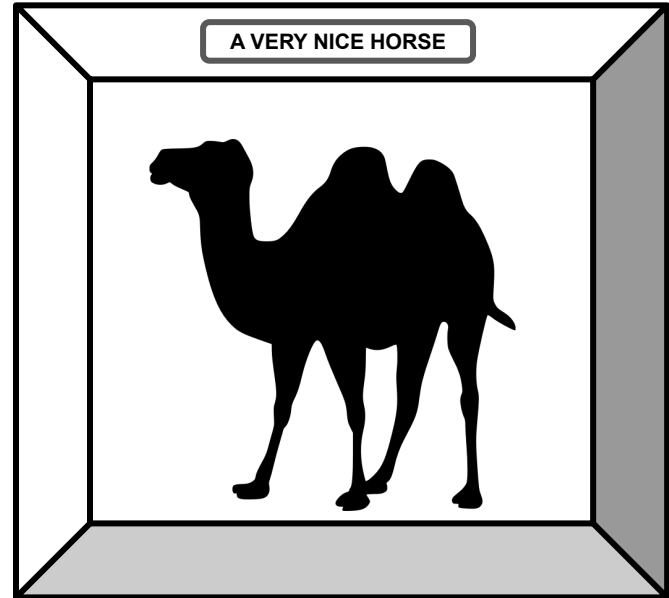
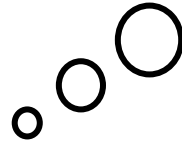
Or go to sli.do
and use the
code TAI4ES



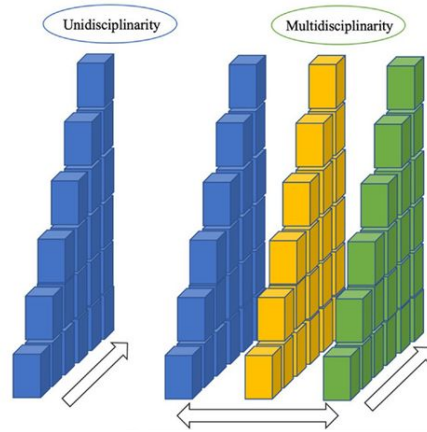
Insights on (meaningful) interdisciplinary work in the AI/ML development process



A camel is a horse designed by an **committee**
interdisciplinary research team

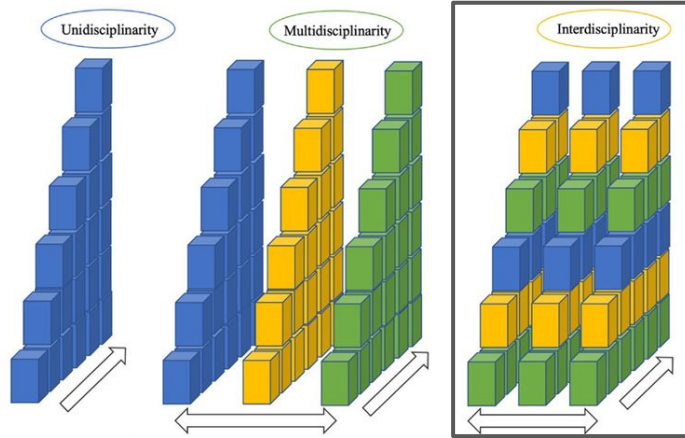


What does it mean to do “interdisciplinary” work?



Adapted from: Peek, L., & Guikema, S. (2021). Interdisciplinary Theory, Methods, and Approaches for Hazards and Disaster Research: An Introduction to the Special Issue. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, 41(7), 1047–1058. <https://doi.org/10.1111/risa.13777>

What does it mean to do “interdisciplinary” work?

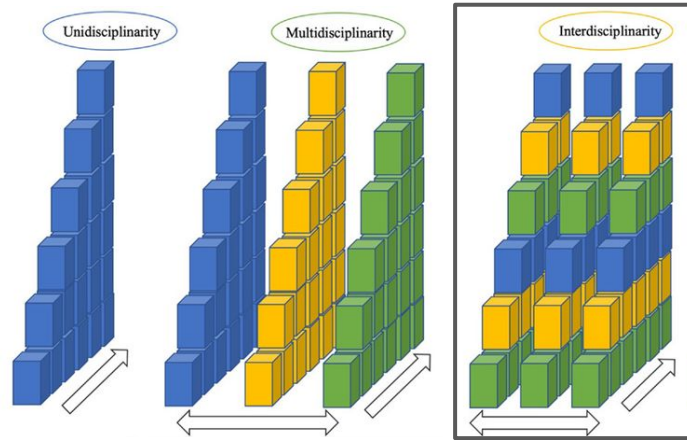


Interdisciplinary: “Integrates information, data, methods, tools, concepts, and/or theories from two or more disciplines focused on a complex question, problem, topic, or theme.”

The key defining concept of interdisciplinarity is **integration, a blending of diverse inputs that differs from and is more than the sum of the parts.**” (Peek and Guikema, 2021, p. 1049)

Adapted from: Peek, L., & Guikema, S. (2021). Interdisciplinary Theory, Methods, and Approaches for Hazards and Disaster Research: An Introduction to the Special Issue. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, 41(7), 1047–1058. <https://doi.org/10.1111/risa.13777>

What does it mean to do “interdisciplinary” work?



Interdisciplinary: “Integrates information, data, methods, tools, concepts, and/or theories from two or more disciplines focused on a complex question, problem, topic, or theme.”

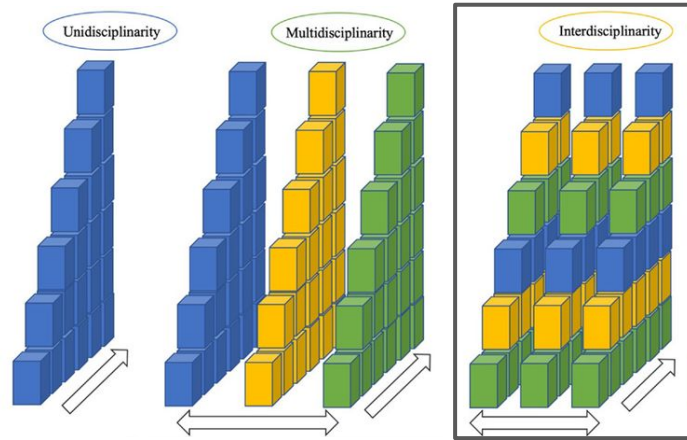
The key defining concept of interdisciplinarity is **integration, a blending of diverse inputs that differs from and is more than the sum of the parts.**” (Peek and Guikema, 2021, p. 1049)

Adapted from: Peek, L., & Guikema, S. (2021). Interdisciplinary Theory, Methods, and Approaches for Hazards and Disaster Research: An Introduction to the Special Issue. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, 41(7), 1047–1058. <https://doi.org/10.1111/risa.13777>

When planning this work ask:

- Would the project totally fall apart if one discipline dropped out?
- How do we make the leap from multidisciplinary to interdisciplinary?

What does it mean to do “interdisciplinary” work?



Interdisciplinary: “Integrates information, data, methods, tools, concepts, and/or theories from two or more disciplines focused on a complex question, problem, topic, or theme.”

The key defining concept of interdisciplinarity is **integration, a blending of diverse inputs that differs from and is more than the sum of the parts.**” (Peek and Guikema, 2021, p. 1049)

Adapted from: Peek, L., & Guikema, S. (2021). Interdisciplinary Theory, Methods, and Approaches for Hazards and Disaster Research: An Introduction to the Special Issue. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, 41(7), 1047–1058. <https://doi.org/10.1111/risa.13777>

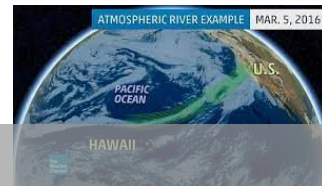
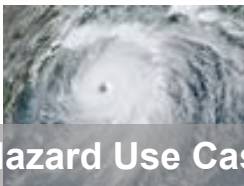
When planning this work ask:

- Would the project totally fall apart if one discipline dropped out?
- How do we make the leap from multidisciplinary to interdisciplinary?

For the trust-a-thon this week:

- Are you really listening to and meaningfully integrating **each team member’s** perspective?
- Is your work really **the sum of all parts**?

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

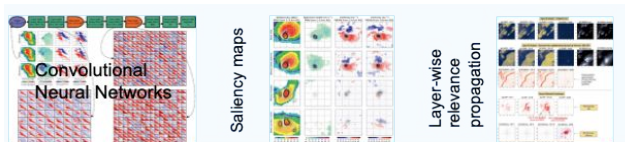
Evaluative Research

Semi-structured interviews

Surveys in naturalistic settings

Randomized Experiments

AI/ML & XAI Techniques



Users & Decision Makers



Trustworthy & Use-Driven AI and ML Products



Insights on (meaningful) interdisciplinary work in the AI/ML development process

1. Incorporates **full intellectual participation** by each contributing area of expertise, forming a **multiway partnership**
2. Generates **novel** research questions, approaches, and interpretations
3. Provides **rigorous, useful new insights** about a **complex** intellectual and/or societal problem



Adapted from: Morss, R. E., Lazrus, H., & Demuth, J. L. (2021). The “inter” within interdisciplinary research: Strategies for building integration across fields. *Risk Analysis: An Official Publication of the Society for Risk Analysis*. <https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.13246>

Communicating values: For the space and the research

Goal: Incorporating **full intellectual participation** by each contributing area of expertise, **forming a multiway partnership** (Morss et al., 2021)

Example from AI2ES: Tropical Cyclones Working group

Communicating values - both for the space and the research

- Transparency, respect, enthusiasm
- Giving all time to speak and think
- Our different disciplines (interdisciplinarity) and experiences/expertise
- Really communicate *to* and *with* each other



1. Introductions and set-up

1. Follow-up to the sitewide last fall with Kaye Husbands Fealing
 - a. Recap: Transparency, respect, enthusiasm, giving all time to speak and think, inclusion
 - b. NEW: What are our **research** values for this space?
 - i. Vulnerability to learning new things
 - ii. intentionally thinking about users and meeting their needs,
 - iii. explicitly talking about what our DVs are
 - iv. Our different disciplines (interdisciplinarity) and experiences/expertise
 - v. Communicating results and expertise back to larger AI2ES
 - vi. Interdisciplinarity with social science
 - vii. Really communicate "to" and "with" each other
 - viii. mapping the AI back to the atmospheric science, i.e., to something that's physically realistic

Risk Communication Researcher Overview

goals, methods, timelines,
examples of past relevant work

Tropical Cyclone Researcher Overview

goals, methods, timelines,
examples of past relevant work

Define mutual goals for current product(s)
Terminology & vocabulary (e.g., "product")
Existing relationships with users
Limitations

TC/RC AI Research

common goals, timelines
mutual benefit

time



Collecting and integrating the different expositions and mental models

Goal: Generate **novel research** questions, approaches, and interpretations (Morss et al., 2021)

Example from AI2ES: Winter Working group

Integrating communication research methods into supervised machine learning for precipitation detection in mesonet images.

Research led by Vanessa Przybylo with help of Carly Sutter, Mariana Cains, and Chris Wirz

NYSM Images - Labeling Codebook
Finalized: 05/03/2022

Overview: This codebook was designed for the labeling of night-time images from the New York State Mesonet (NYSM) for the presence/absence of precipitation. These labeled images are intended to then be used for training AI/ML models to classify these same types of images for the presence/absence of precipitation. This codebook serves as the 'calibration' instrument for the human coders. All coders will work from this document when making decisions, both for the initial reliability trials (when the human coders ensure they are consistent and reliable in their codings on a training set of shared images) and during their independent coding (when they code independently over unique images).

Labels/Classes: The coders can apply the following mutually exclusive codes (meaning only one may be applied per image) when assessing the presence/absence of precipitation:
(1) *Precipitation*, (2) *No Precipitation*, (3) *Obstructed*, (4) *Unsure*.

NOTE: Below are examples to help clarify precip-like streaks vs. other streaks

Precip-like streaks:



Other lines/streaks (NOT precip-like)



Contextual factors and constraints

Goal: Provides rigorous, useful new insights about a complex intellectual and/or societal problem (Morss et al., 2021)

Example from AI2ES: Severe Working group

AI, meteorology, and risk communication experts coming together to help address needs and challenges faced by National Weather Service forecasters.

This work requires deep integration of expertise and experience to better understand a complex and time constrained decision making context.



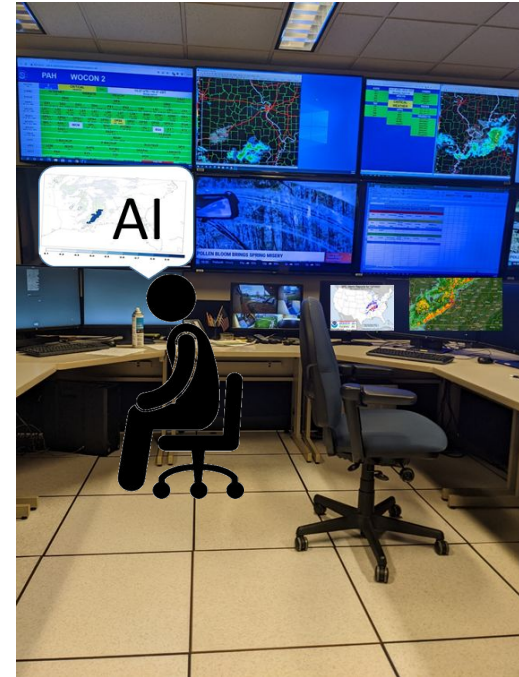
Contextual factors and constraints

Goal: Provides rigorous, useful new insights about a complex intellectual and/or societal problem (Morss et al., 2021)

Example from AI2ES: Severe Working group

AI, meteorology, and risk communication experts coming together to help address needs and challenges faced by National Weather Service forecasters.

This work requires deep integration of expertise and experience to better understand a complex and time constrained decision making context.



Contextual factors and constraints

Goal: Provides rigorous, useful new insights about a complex intellectual and/or societal problem (Morss et al., 2021)

Example from AI2ES: Severe Working group

AI, meteorology, and risk communication experts coming together to help address needs and challenges faced by National Weather Service forecasters.

This work requires deep integration of expertise and experience to better understand a complex and time constrained decision making context.



Contextual factors and constraints

Goal: Provides rigorous, useful new insights about a complex intellectual and/or societal problem (Morss et al., 2021)

Example from AI2ES: Severe Working group

AI, meteorology, and risk communication experts coming together to help address needs and challenges faced by National Weather Service forecasters.

This work requires deep integration of expertise and experience to better understand a complex and time constrained decision making context.



Some best practices for interdisciplinary research

Start with the Problem – The problem is something which no single discipline has been able to resolve, and which seems to interface across multiple fields. Make sure each collaborator understands how the problem involves their area of research.

Build your Team – It is good to have people who can play key roles such as Facilitator, Visionary, Mediator, and members who are technically strong in the component areas.

Take time to learn the Lingo – Don't assume your collaborators understand the language of your field. This even applies to terms that two or more fields use that on the surface mean the same thing. They may carry layers of meaning that differ depending on who uses them and how.

Be Flexible & don't assume anything – Take time early on to have everyone (1) describe the project for the team as they would explain it to an outsider and (2) say what they hope to get out of the project. This is a good way to uncover conceptual disconnects.

Publications – Since different disciplines may have different standards or disclosures, be sure to discuss how publications will be handled by the group.

History of working together – Even brief periods of collaboration prior to submitting proposals can help you figure out whether and how you might be able to work together, and makes funding agencies more comfortable with the idea that you can deliver on what you promise.



Interdisciplinary work section activity and wrap up

- Go to slido and add in terms that represent good interdisciplinary work!
- Any other questions or other things you'd like us to talk about?



1.3. Go to [sli.do](#) and use the code TAI4ES

Standards



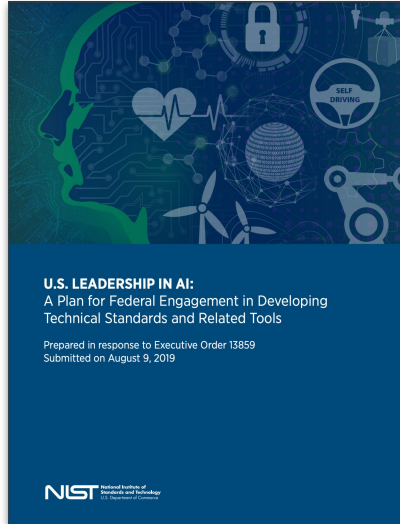
Standards for Trustworthy AI



“Ensure that technical standards...reflect Federal priorities for innovation, public trust, and public confidence in systems that use AI technologies...and develop international standards to promote and protect those priorities.” EO 13859, 2019



ai2es.org

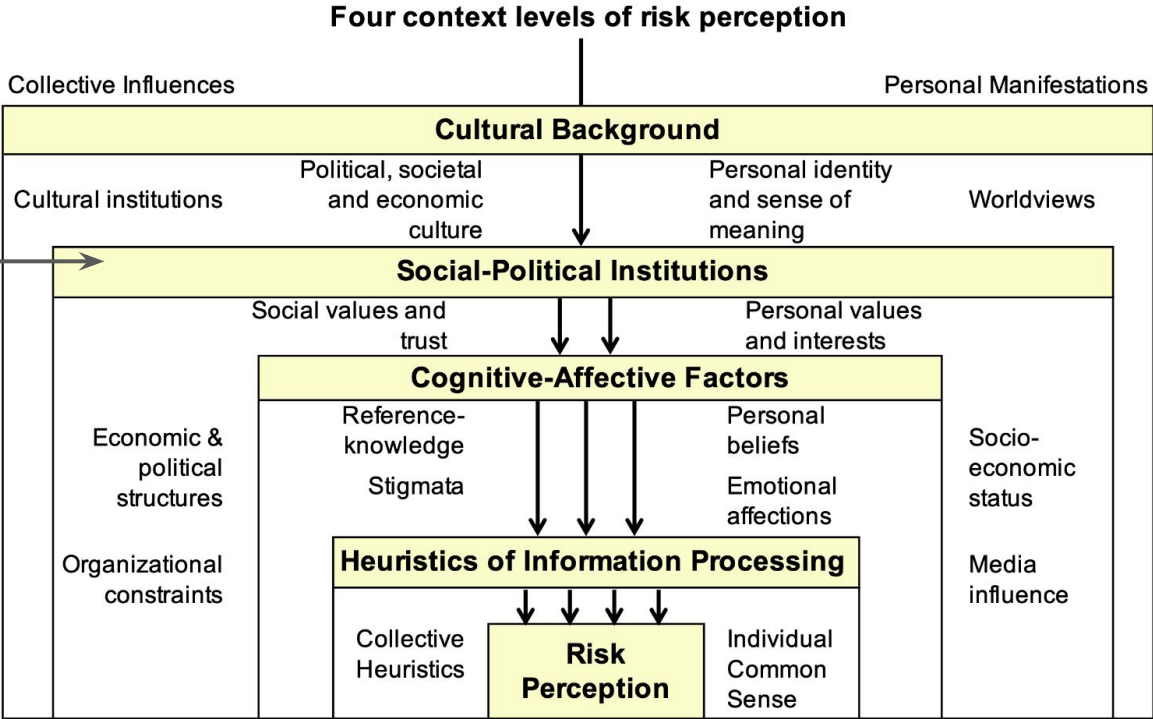
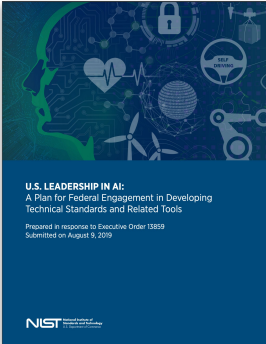


“Trustworthiness standards include guidance and requirements for accuracy, explainability, resiliency, safety, reliability, objectivity, and security.” NIST, 2019

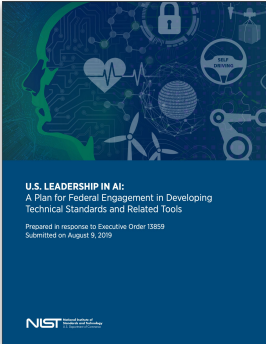


“Characteristics of trustworthiness include, for instance, reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality, usability.” ISO, 2020

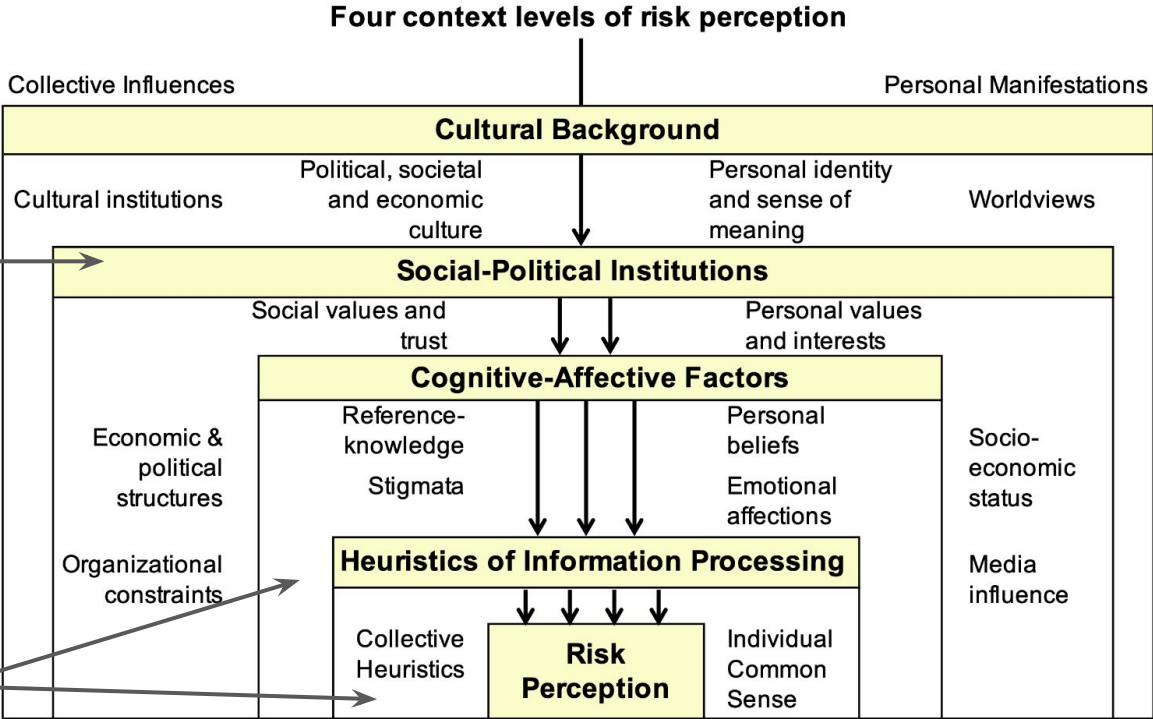
Standards for Trustworthy AI



Standards for Trustworthy AI



But how do standards at the institutional level accommodate differences at the individual level?

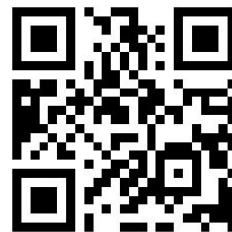


Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 1: Agenda

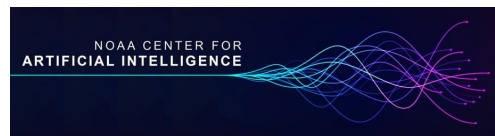
- 9:00 Welcome and Overview
- 9:10 What does it mean to trust?
- 9:40 *Short brain & bio break*
- 9:45 Meaningful interdisciplinary work
- **10:25 *Short brain & bio break***
- 10:30 Evaluation metrics
- 11:00 XAI for traditional ML

Questions?



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

Or go to sli.do
and use the
code TAI4ES



Model verification

Montgomery Flora



What makes an ML model “good”?

A key component of ML development is evaluating the model’s performance, but what makes a model “good”?

Murphy (1993) defined three types of “goodness” (in terms of weather forecasting, but applicable to any model evaluation task):

Consistency - the degree to which the forecast corresponds to the forecaster's best judgement about the situation, based upon their knowledge base

Quality - the degree to which the forecast corresponds to what actually happened

Value - the degree to which the forecast helps a decision maker to realize some incremental economic and/or other benefit



What makes an ML model “good”?

A key component of ML development is evaluating the model’s performance, but what makes a model “good”?

Murphy (1993) defined three types of “goodness” (in terms of weather forecasting, but applicable to any model evaluation task):

Consistency - the degree to which the forecast corresponds to the forecaster's best judgement about the situation, based upon their knowledge base

Quality - the degree to which the forecast corresponds to what actually happened

Value - the degree to which the forecast helps a decision maker to realize some incremental economic and/or other benefit



Can we use just one statistic to describe prediction quality?

Nope. Prediction error, like other random variables, is a distribution and requires multiple statistics to describe it.

For example, we may want information about the following:

- What is average diff. b/t the prediction and target?
- Was the predicted magnitude correct?
- Was the prediction biased?
- How often did the prediction make an unacceptably large error?
- If it is a forecast, was the timing correct?



Slide based information from <https://www.cawcr.gov.au/projects/verification/BestStatistic.php>

Verification Diagrams

Go to sli.do and answer a quick poll on your familiarity with the following kinds of verification diagrams (not being familiar with them is a perfectly valid answer!)

Classification Task

- ROC Diagram (The “AUC” diagram)
- Attributes (Reliability) Diagram
- Performance (Precision-Recall) Diagram

Regression Task

- Taylor Diagram



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

Summarizing Multiple Statistics with Verification Diagrams

Goal: Summarize multiple statistics in a single visualization

Classification Task

- ROC Diagram (The “AUC” diagram)
- Attributes (Reliability) Diagram
- Performance (Precision-Recall) Diagram

Regression Task

- Taylor Diagram



Classification probabilities and deterministic outcomes

For classification models, we can issue deterministic and/or probabilistic predictions.

- E.g., for binary outcomes (e.g., yes/no; event vs. no event), we can issue binary predictions or frequencies between 0-1 that we can interpret as probabilities.

For the former, we can build a contingency table:

| | | | |
|---------|-----|---|---|
| Outcome | No | False Alarm/ False Positive/ Type I Error | Correct Negatives/ True Negative |
| | Yes | Hit/ True Positive | Missed/ False Negative/ Type II Error |
| | | Yes | No |
| | | Predicted | |

Unfortunately, naming conventions are inconsistent across different domains

Model Performance and the End User

A perfect prediction only produces hits and correct negatives with no misses or false alarms.

For most tasks, predictions will have misses and false alarms. The main goal of model development is to limit both or, in most cases, strike a fair balance between them.

The decision on how to balance misses and false alarms ought to be in conjunction with the end user

- E.g., a forecaster might be more concerned about limiting misses



Contingency Table Statistics

We can compute several statistics from the contingency table values (non-exhaustive list)

Table 2.1: Common verification metrics associated with the components of the contingency table (non-exhaustive list). The terms h , m , f , c refer to hits, misses, false alarms, and correct negatives, respectively.

| Metrics | Formulas |
|---------------------------------------|-------------------|
| Probability of Detection (POD) | $\frac{h}{h+m}$ |
| Probability of False Detection (POFD) | $\frac{f}{f+c}$ |
| Success Ratio (SR) | $\frac{h}{h+f}$ |
| Critical Success Index (CSI) | $\frac{h}{h+m+f}$ |
| False Alarm Ratio (FAR) | $\frac{f}{h+f}$ |
| Frequency Bias (BIAS) | $\frac{h+f}{h+m}$ |

From Flora (2020, dissertation)



Nomenclature Issues

The same statistic can have different names depending on the discipline

Table 2.2: Aliases for the contingency metrics in Table 2.1.

| Metric | Aliases |
|---------------------------------------|---|
| Probability of Detection (POD) | Sensitivity, Recall, Hit Rate, True Positive Rate |
| Probability of False Detection (POFD) | Fall-out or False Positive Rate |
| Success Ratio (SR) | Precision |
| Critical Success Index (CSI) | Threat Score |
| False Alarm Ratio | False Discovery Rate |

Converting probabilities to binary predictions

Probabilities (p) can be converted to yes/no predictions using a threshold (t)

$$p_b = \begin{cases} 1, & \text{if } p \geq t \\ 0, & \text{if } p < t \end{cases}$$

Using the binary predictions, we can compute the contingency table statistics (see annotated regions)

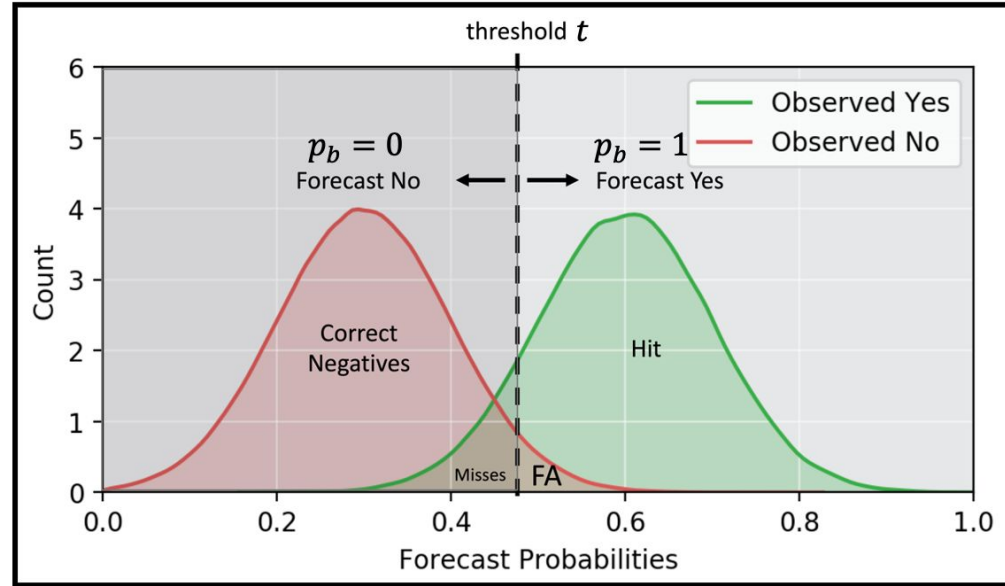
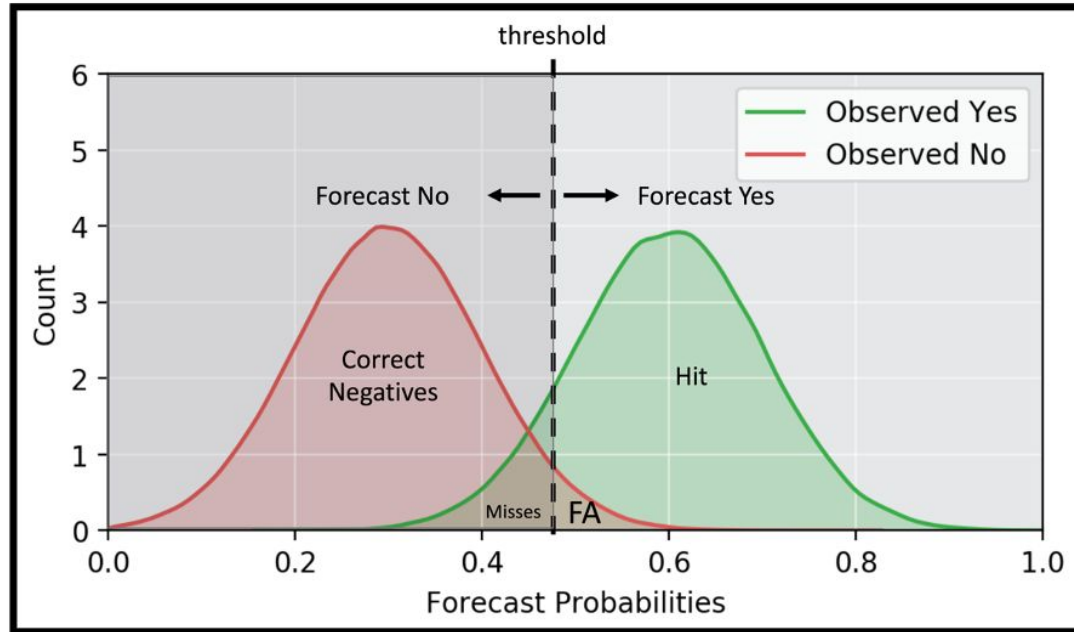


Figure 2.1: Distribution of forecast probabilities conditioned on being matched to an observed yes (green) or observed no (red). Forecast probabilities are converted to yes/no forecasts based on some threshold (e.g., 45% in this example). The regions of the two distributions are annotated by their corresponding contingency table term. FA is short for false alarms.

ROC Diagram

The receiver operating characteristic (ROC) diagram measures how well probabilities discriminate between event and non-event (e.g., the separation between the red and green regions)

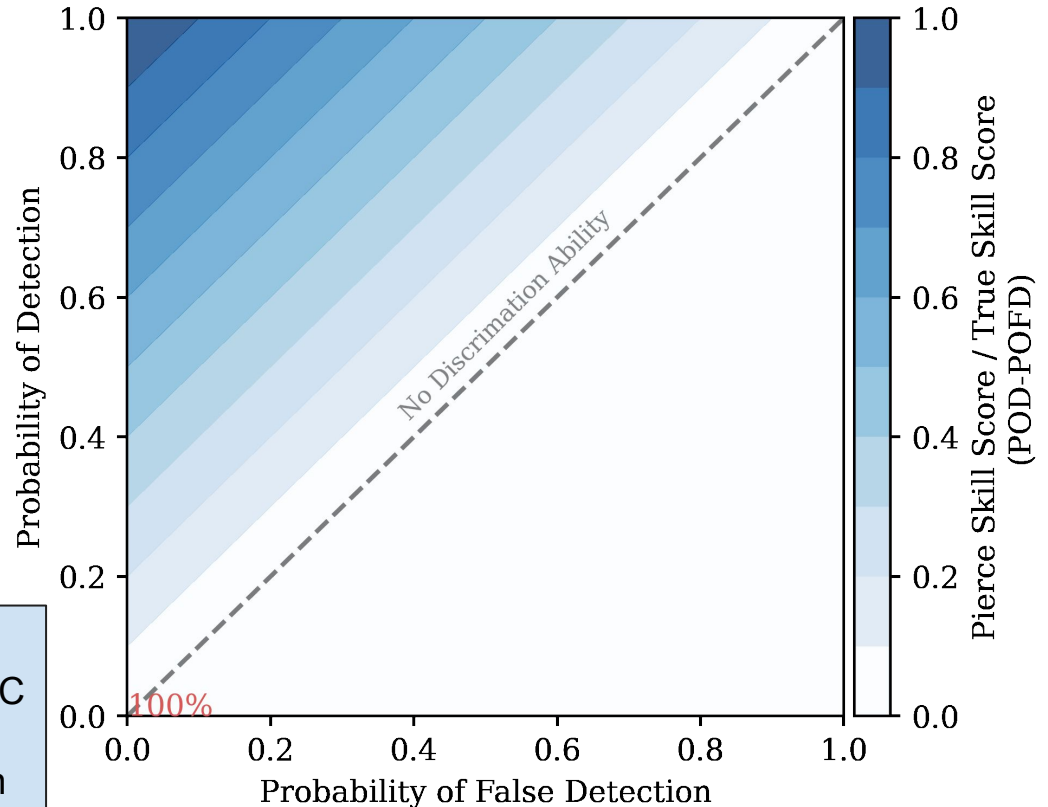


ROC Diagram

Using a series of probability thresholds, we can compute POD (*what proportion of the observed yes region was hit?*) and POFD (*what proportion of the observed no region has false alarms?*) to produce a curve.

If the predicted probabilities discriminate well, the POD ought to increase faster than POFD as the threshold increases.

If $POD=POFD$ for all thresholds, then the predicted probabilities have not discrimination ability.



We can summarize the diagram by the area under the curve (AUC). AUC varies from 0.5 to 1.0 where higher values indicate better discrimination



Performance Diagram (PD)

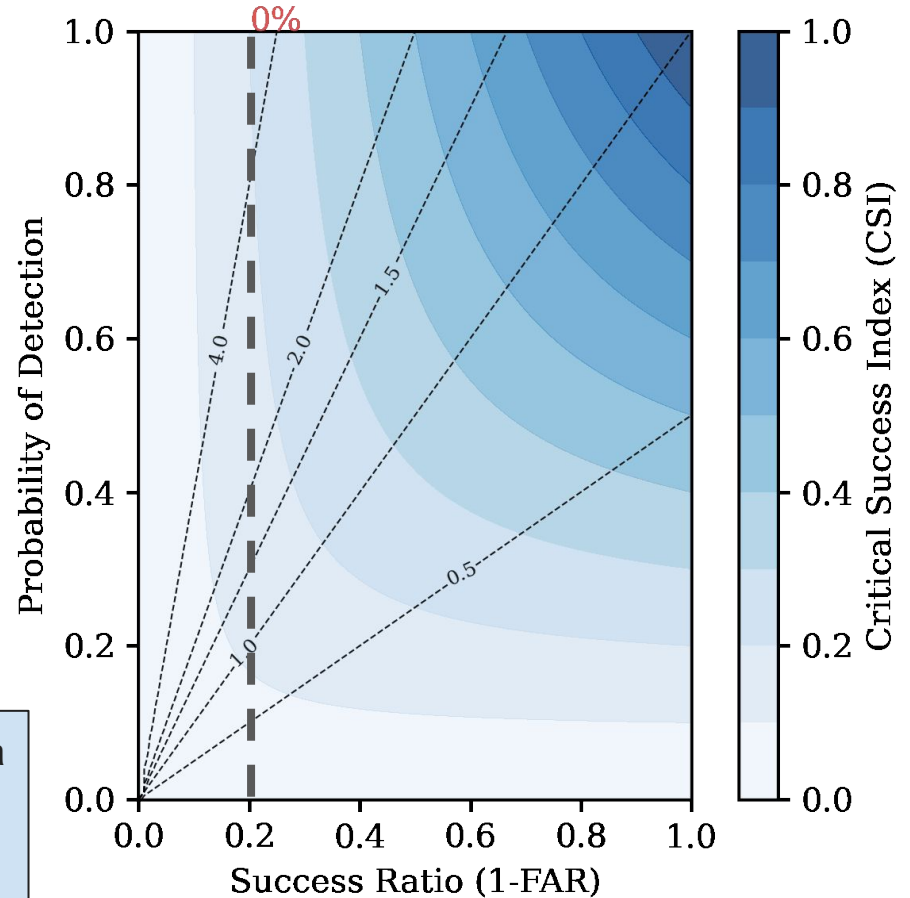
The PD measures how well the model correctly predicts events and if it balances hits (POD) and false alarms (SR).

A model that balances these hits and false alarms will have a maximum CSI (colored contours) associated with a frequency bias near 1.0 (ratio of hits and false alarms; dashed black lines)

A random predictor will produce a PD curve along the vertical dash line

- Equal to the event rate of the dataset
- Event rate = number of positive examples / total number of examples

We can summarize the diagram by the area under the curve (AUPDC). To improve AUPDC, we can normalize out the no-skill area (Flora et al. 2021, Miller et al. 2022)

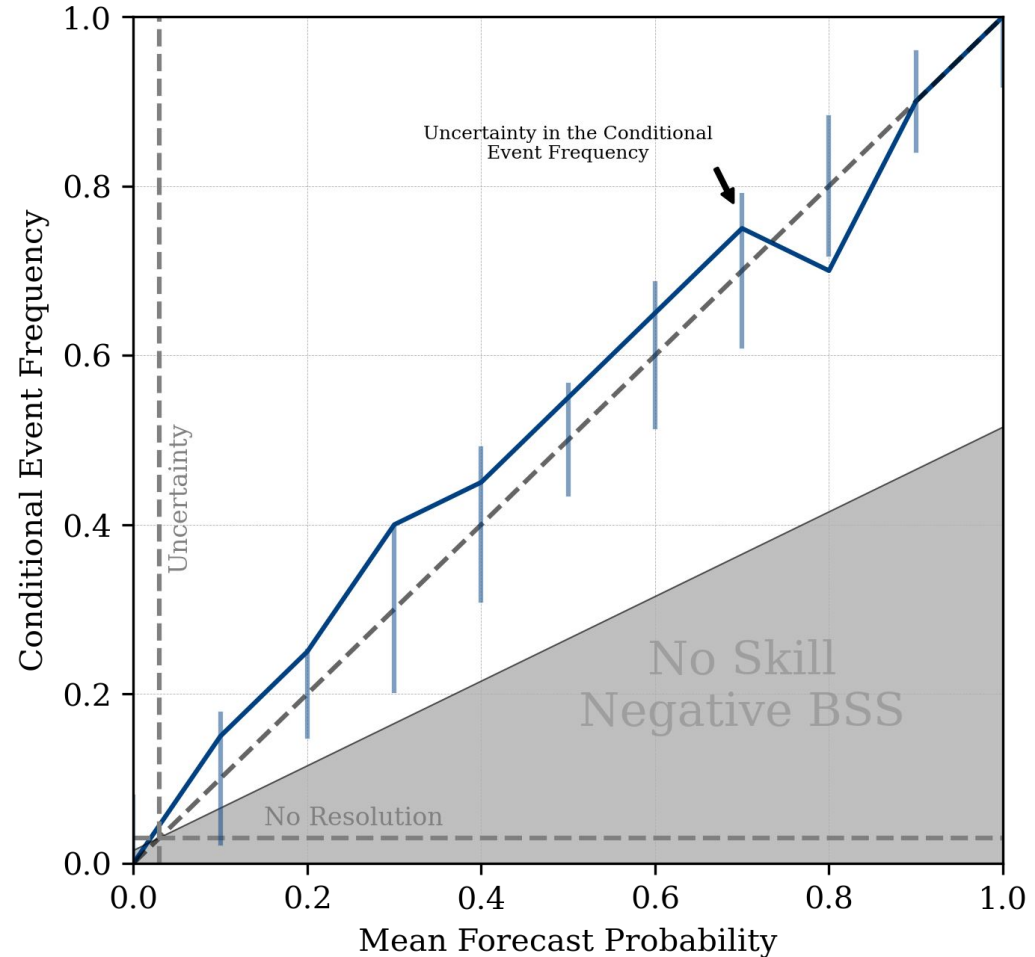


How reliable are the probabilities?

Do the probabilities correspond to long-term event frequencies?

Bin forecast probabilities and the binary outcomes (e.g., every 10%) and compute the mean forecast probability and conditional event frequencies per bin.

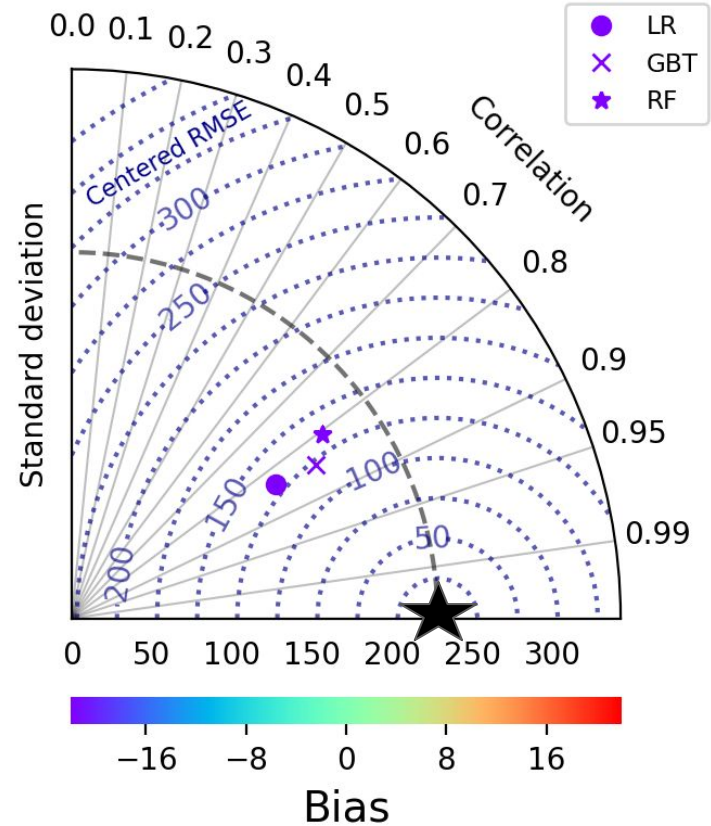
If a forecast system is reliable, then the mean forecast probability == conditional event frequency for all bins (i.e., line along the dashed diagonal).



Taylor Diagram

The Taylor diagram is one of the few, if not the only, verification diagrams for regression tasks.

The Taylor diagram shows the following metrics:

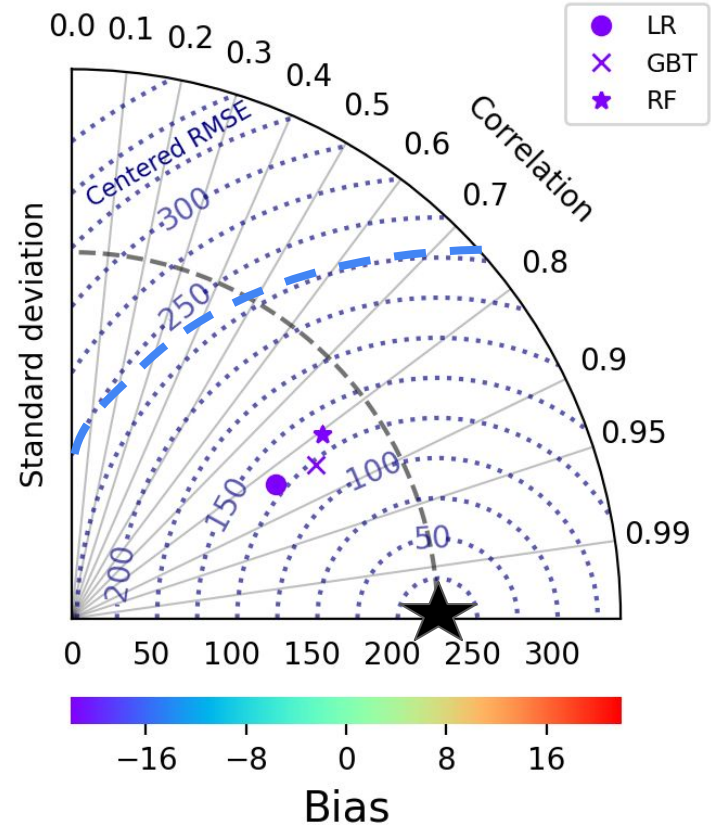


Taylor Diagram

The Taylor diagram is one of the few, if not the only, verification diagram for regression tasks.

The Taylor diagram shows the following metrics:

- Bias-corrected (Centered) RMSE
 - Radial distance from the star
 - Closer to the star is better

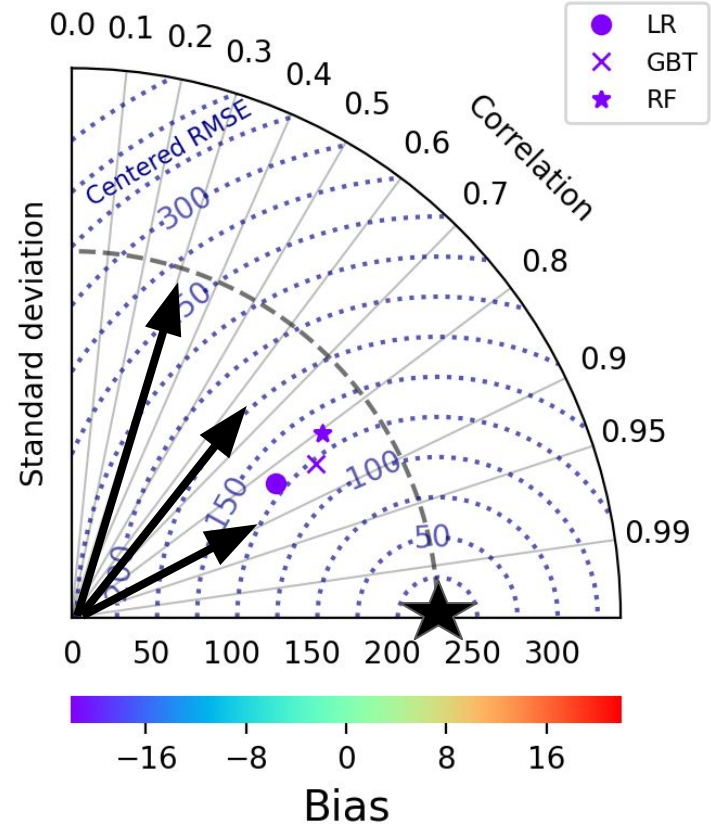


Taylor Diagram

The Taylor diagram is one of the few, if not the only, verification diagram for regression tasks.

The Taylor diagram shows the following metrics:

- Bias-corrected (Centered) RMSE
 - Radial distance from the origin
 - Closer to the gray dashed line is better
- Standard deviation

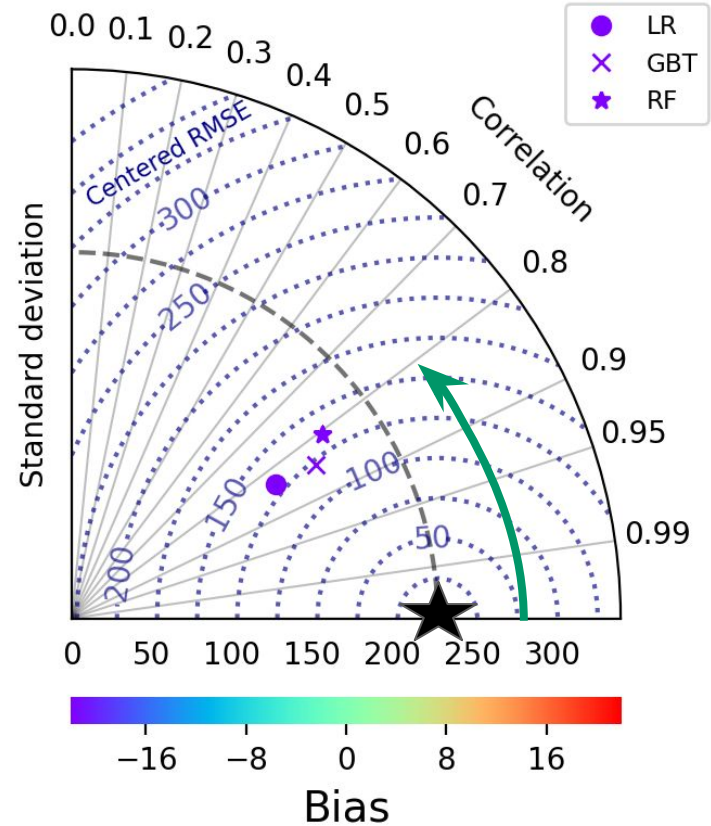


Taylor Diagram

The Taylor diagram is one of the few, if not the only, verification diagram for regression tasks.

The Taylor diagram shows the following metrics:

- Bias-corrected (Centered) RMSE
- Standard deviation
- Correlation Coefficient
 - Angle from the original
 - Point closer to the origin is better

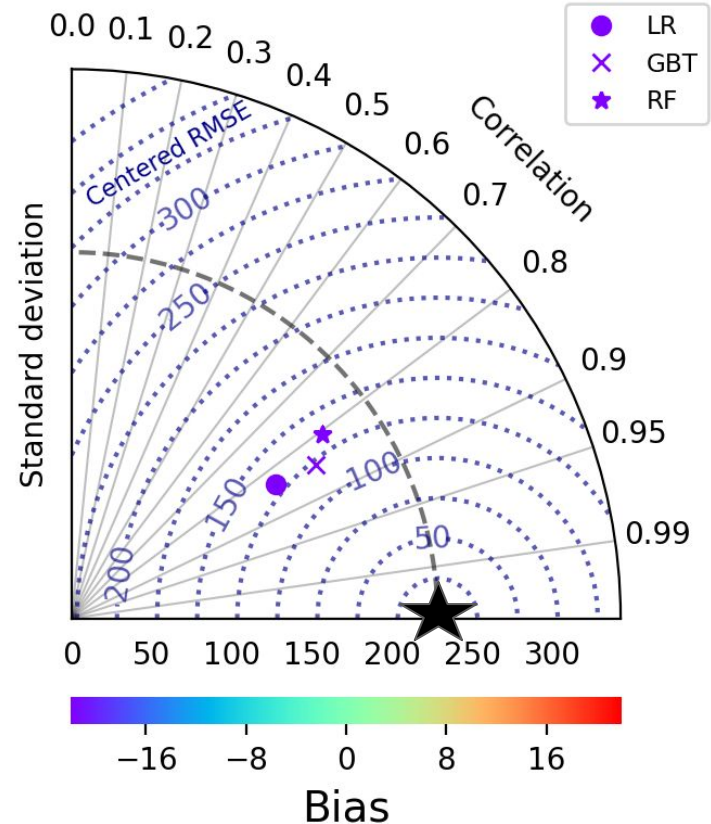


Taylor Diagram

The Taylor diagram is one of the few, if not the only, verification diagram for regression tasks.

The Taylor diagram shows the following metrics:

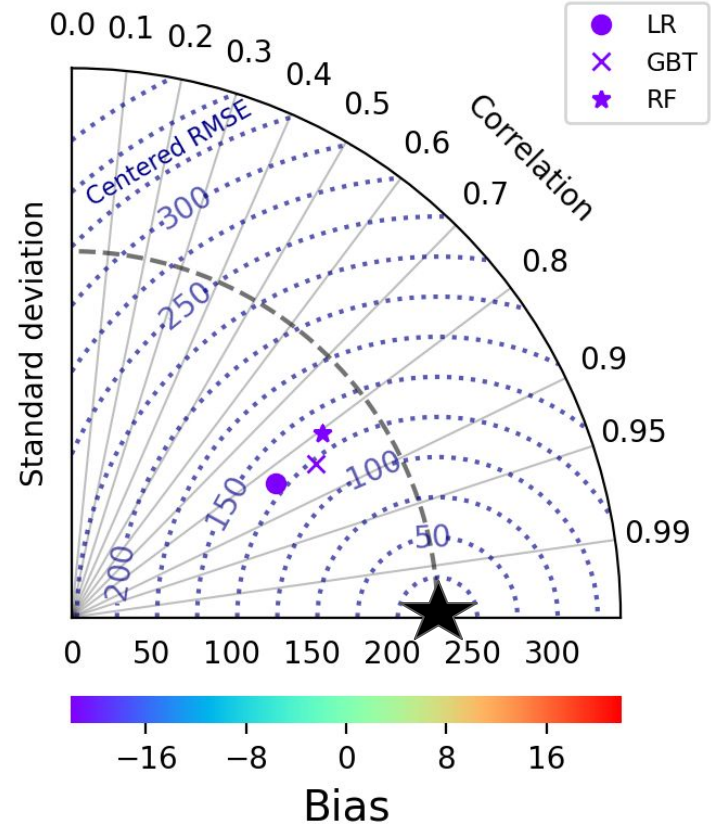
- Bias-corrected (Centered) RMSE
- Standard deviation
- Correlation Coefficient
- Bias
 - Color-coding of the dot
 - Closer to zero is better



Taylor Diagram

A perfect model will have a point lining on the star with a bias of zero.

- Correlation coefficient is 1
- Prediction variance matches the target variance
- Centered RMSE is zero

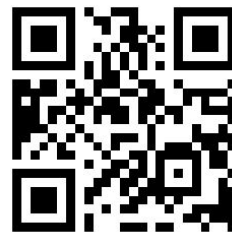


Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 1: Agenda

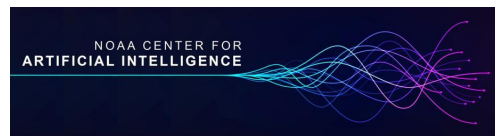
- 9:00 Welcome and Overview
- 9:10 What does it mean to trust?
- 9:40 *Short brain & bio break*
- 9:45 Meaningful interdisciplinary work
- 10:25 *Short brain & bio break*
- 10:30 Evaluation metrics
- **11:00 XAI for traditional ML**

Questions?



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

Or go to sli.do
and use the
code TAI4ES



Traditional ML Explainability

Montgomery Flora
Amy McGovern



Activity!

Head over to slido and answer the following:

How well can we currently explain traditional ML models in a way that will increase trust?



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

Topic Overview

- Global vs. Local Explainability
- Global Explainability
 - Feature Importance
 - Feature Effects
- Local Explainability
 - Feature Attributions



Global vs. local explainability

Global explanations attempt to describe the model as a whole

- What are the important features?
- What relationship has been learned for this feature?

Local explanations attempt to describe individual predictions

- Which feature is making the biggest impact on the prediction for this example?
- If this feature value was slightly different, how would it change the prediction?



Global Explainability

Global explainability products can be divided into 3 categories:

1. Feature Importance/Relevance

- a. *Importance*: How does this feature contribute to the model's performance?
- b. *Relevance*: How does this feature contribute to the model's prediction?

2. Feature Effects

- a. What is the relationship between this feature's values (or these set of features) and the model's prediction?

3. Feature Interactions

- a. How is a feature's effect impacted by the effects of other features?



Local explainability

The most common local explanation methods are known as **feature attribution** methods where we assume that a model's prediction (P) can be interpreted as a linear combination of contributions from each feature.

$$P = \phi_0 + \sum_{j=1}^M \phi_j$$




Local explainability

The most common local explanation methods are known as **feature attribution** methods where we assume that a model's prediction (P) can be interpreted as a linear combination of contributions from each feature.

$$P = \phi_0 + \sum_{j=1}^M \phi_j$$

$\sum_{i=1}^N P_j$ “Average prediction”



Local explainability

The most common local explanation methods are known as **feature attribution** methods where we assume that a model's prediction (P) can be interpreted as a linear combination of contributions from each feature.

$$P = \phi_0 + \sum_{j=1}^M \phi_j$$

“Sum of contribution from each feature”
(M number of features)



Feature Importance

What features contribute most to model performance?



Establishing the important features helps inform the explainability downstream

Given their greedy nature, ML models will tend to favor only a subset of the total features they are trained on

Thus, explaining an ML model largely comes down to explaining the top features. By knowing the top features we can ask the following questions:

- How much more important are they than the less important features?
- What are the learned relationships for these top features?
- What features are interacting with them?



Permutation importance

One of the first explainability method was permutation importance (Breiman 2001).

An intuitive way to determine the importance of a feature is to remove it from the model and evaluate how the model performance suffers.

Explicitly removing a feature requires re-training

- *Problem:* By re-training the model, we would no longer know how important the feature was to the original model

How can we remove a feature without having to retrain the model? We could shuffle its values!

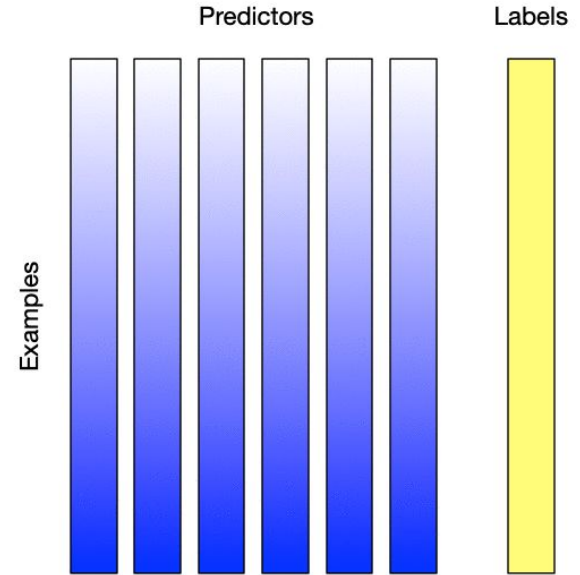


Illustration of single-pass permutation importance. Each predictor is permuted one at a time (blue boxes) and ranked by the difference in score from the original model and the model with permuted data (red shaded values at end).



Permutation importance

By shuffling the values, we maintain the marginal distribution of the feature (not introducing bias!) while breaking its connection to the target variable.

For traditional permutation importance, we shuffle each feature once, compute the loss in performance, and rank accordingly.

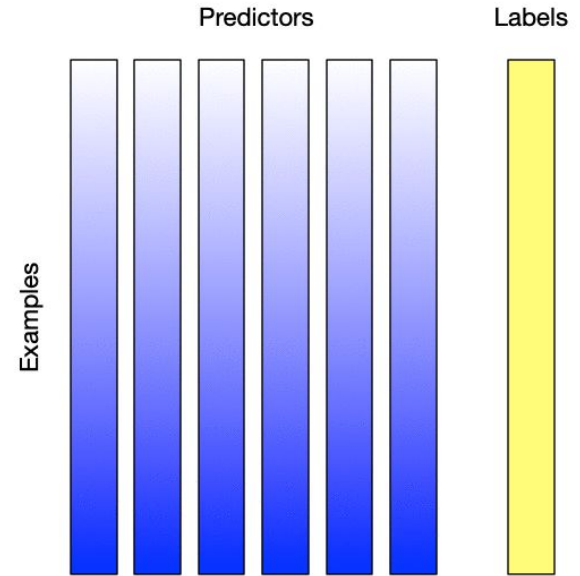
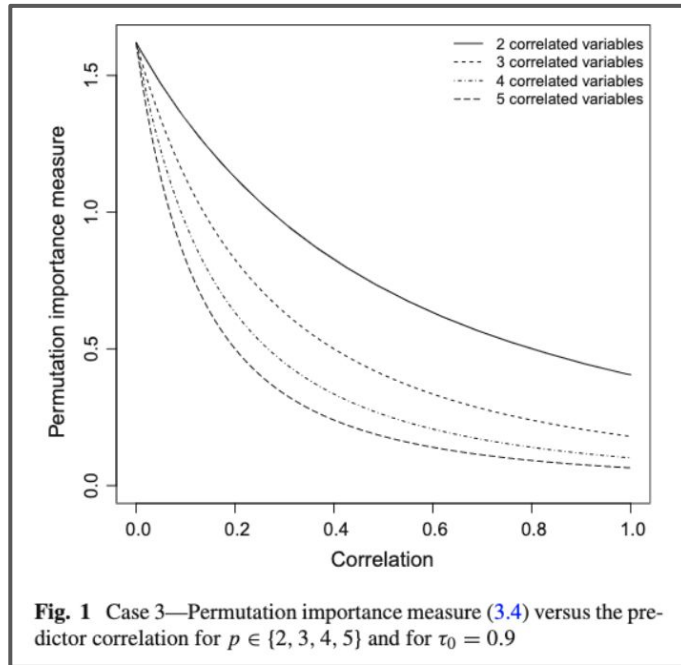


Illustration of single-pass permutation importance. Each predictor is permuted one at a time (blue boxes) and ranked by the difference in score from the original model and the model with permuted data (red shaded values at end).

Limitation of Traditional Permutation Importance

The traditional permutation importance does not consider hierarchical structures in the data (e.g., correlations, multicollinearities, etc).



Gregorutti et al. (2017) found that the higher the number of correlated variables, the faster the permutation importance of different variables decrease to zero (see figure left)

Real-world Example: Surface temp. and 2-m temp. are highly correlated. If we “removed” surface temp. from the model, the model can still rely on 2-m temp. Vice versa would be true for 2-m temp as well. Thus, both features would have reduced *individual* importance.

How can we estimate feature importance in a way that maintains feature hierarchies?

Lakshmanan et al. (2015) introduced a method where multiple features are permuted. When permuting more than one feature, the total importance is not only equal to the sum of the individual importances from each feature, but also includes a term based on the covariance between the features (Gregorutti et al 2015).

Pseudo-algorithm for the this new method:

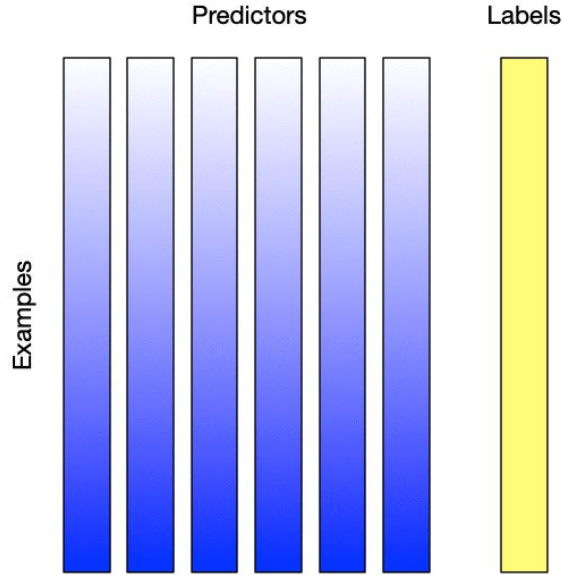
1. The most important feature is left permuted
2. Permutation importance is re-computed to determine the second-most important feature.
3. The top two features are left jointly permuted
4. Permutation importance is re-computed to determine the third-most important features.
5. Repeat until the top N features are computed for.

In the literature, the traditional permutation importance is known as the “single-pass” while this new method is known as the “multi-pass”



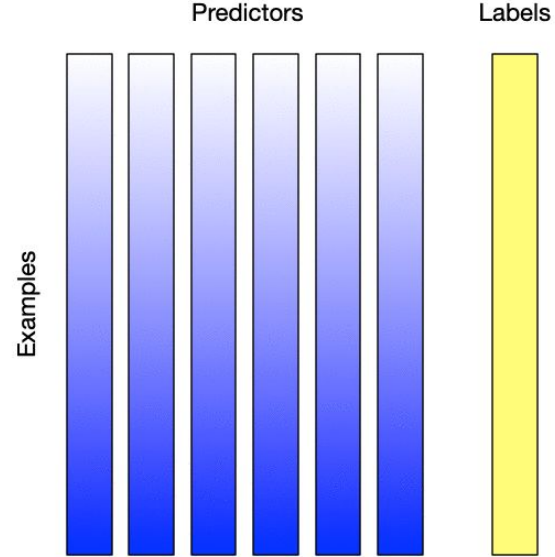
Single-pass vs. Multi-pass Permutation Importance

Single-Pass



Measuring unique, individual importance

Multi-Pass



Measuring multivariate importance in the context of other features



Limitations of Multi-pass Permutation Importance

- As more features are permuted, the importance score relies on higher-order covariances, which are often poorly sampled.
- Later iterations inherit the faults of previous iterations
 - The first pass can fail to identify the most important feature, which is inherited by the next iteration
- It is a greedy algorithm like other sequential feature selection methods and does not consider all possible combinations of features
 - There is no theoretical guarantee that multipass produces the most important top N features
- Computationally demanding for large datasets with a larger number of features



Backward vs. Forward Permutation Importance

Up to this point, we have discussed the **backward** version of the permutation importance methods. An alternative is the **forward**-based methods.

- **Backward**

- How does removing this one feature or these set of features reduce model performance?

- **Forward**

- How well do these features perform by themselves (i.e., if all other features were removed)?

Forward Single-Pass

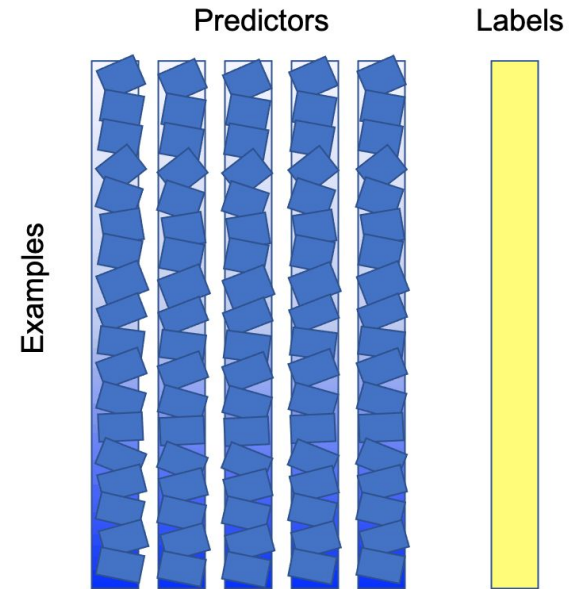


Illustration of forward single-pass permutation importance. All predictors start jointly permuted (blue boxes) and then each predictor is unpermuted one at a time. Predictors are then ranked by difference in score from model with unpermuted data and model with all predictors permuted (red shaded values at end)



Final thoughts on Permutation Importance

- For robust results, multiple permutations are required which is computationally demanding for large datasets
 - A single permutation is insufficient to measure importance. It is possible that a given permutation only slightly alters a particular feature's values, which will render it less important.
- Choice of model verification/evaluation
 - As discussed above, different verification metrics measure different aspects of model performance. Therefore, we must remember that *importance* is measured in the context of the verification metric chosen.
- An alternative to single-pass and multi-pass permutation importance is *grouped* importance (Au et al. 2021) where the grouping of permuted (or unpermuted) features is manually selected.



Activity!

Head over to slido and answer the following:

How do you think you could use permutation importance methods to improve trust in your AI method?



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

Feature Effects

What is the learned relationship for a feature?



Why is this feature important?

Permutation importance informs us about the top features, but it does not explain *why they are important*.

To better understand the top features, we can explore the model sensitivity to a given feature. For example, how does the model prediction change, on average, when we increase or decrease the value of a single feature (or of a set of features)?

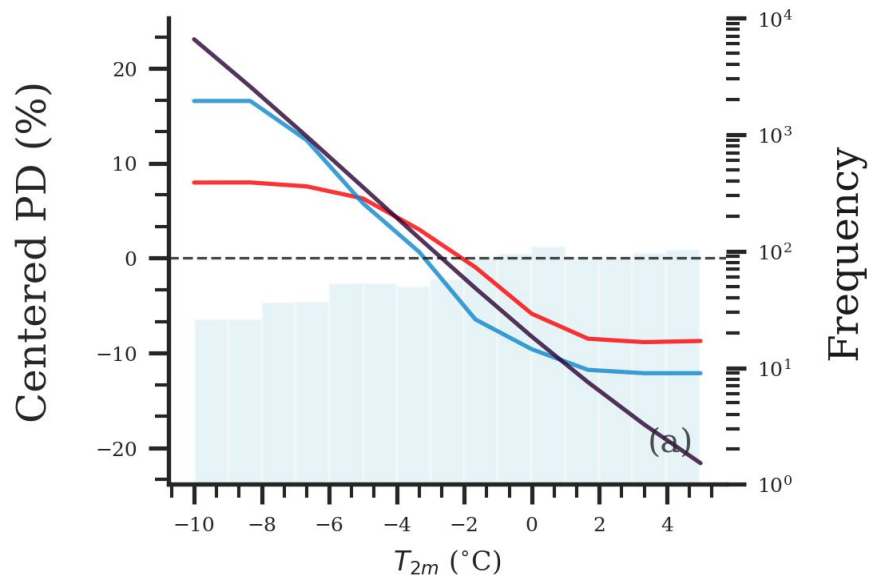


Partial Dependence (PD)

PD is a simple way to evaluate the sensitivity of a model's prediction to changes in the value for a particular feature.

To compute PD for feature x_i , we replace each example with a single value of x_i and evaluate the average model prediction. We then repeat this process for multiple values to get a curve. To center the curve, we subtract out the average PD value (so the mean effect is zero).

Sub-freezing road surface temperature prediction (Handler et al. 2021)



In this example, for the different ML models, lower 2-m temperatures tend to increasing ML probability of a sub-freezing road surface.

Limitations of Partial Dependence

- Assumes the features are independent.
 - When features are dependent, then replacing values from the marginal distribution can lead to extrapolation errors (Fig. 1)

- Only shows the average effect, which can be misleading if feature interaction effects are strong
 - In Fig. 2, the average effect is zero, but based on the [ICE](#) curves the feature is clearly having an impact on the model prediction.

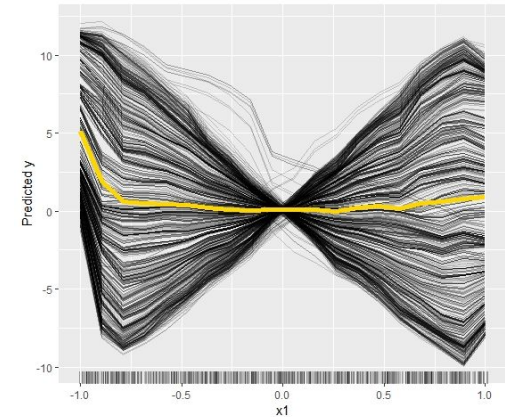
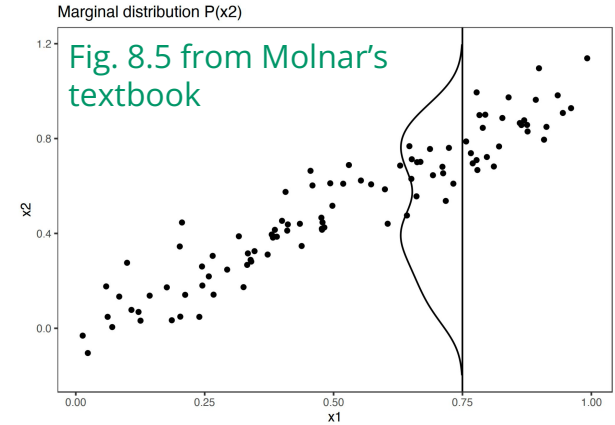
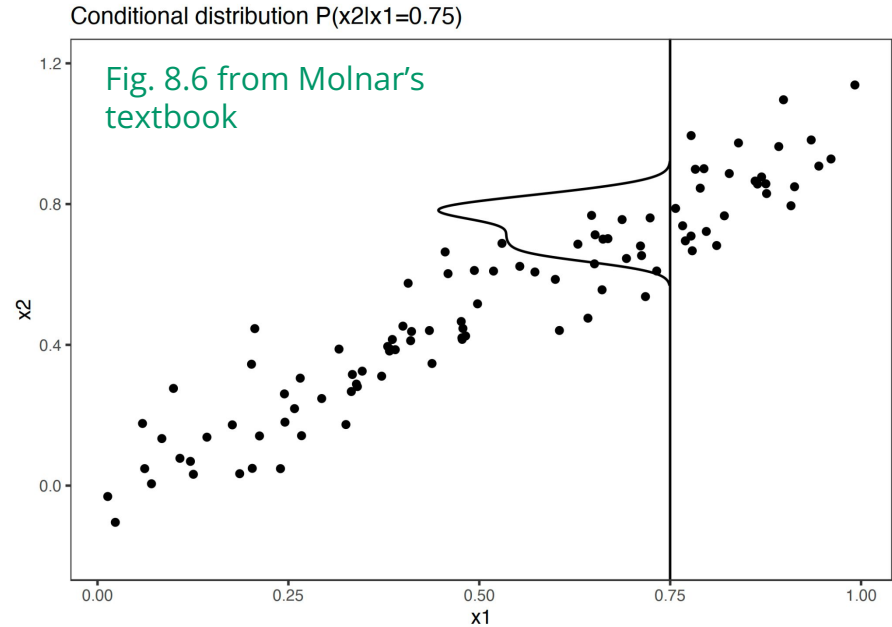


Fig. 4.2 from Limitations of Interpretable ML

Accumulated Local Effects (ALE)

An alternative to PD is ALE.

Unlike PD, ALE computes change in model prediction over conditional rather than marginal distributions → *more immune to correlated features*



Accumulated Local Effects (ALE)

ALE computes the average change in prediction over a series of small windows.

For a given bin, we set the values of a given feature to left side of bin, compute the prediction

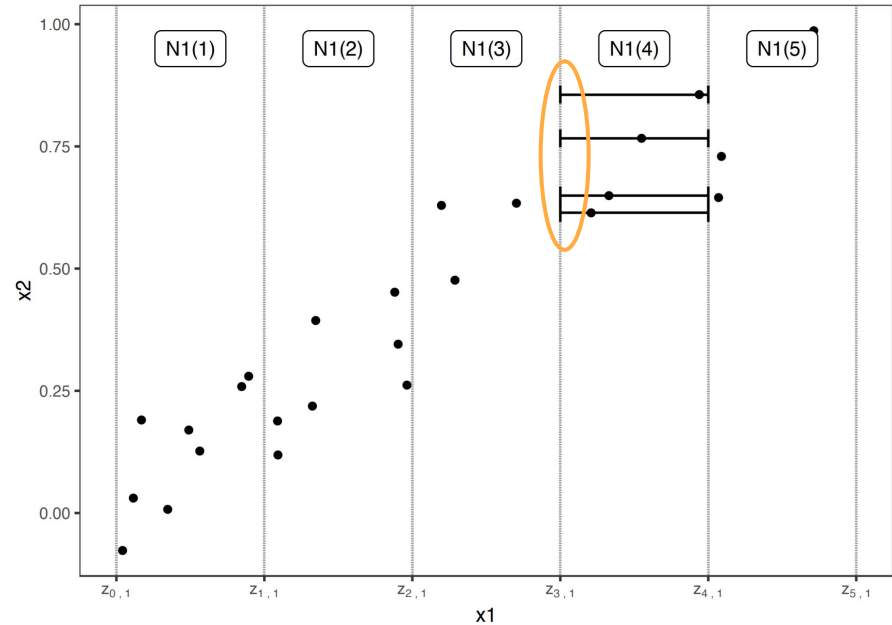
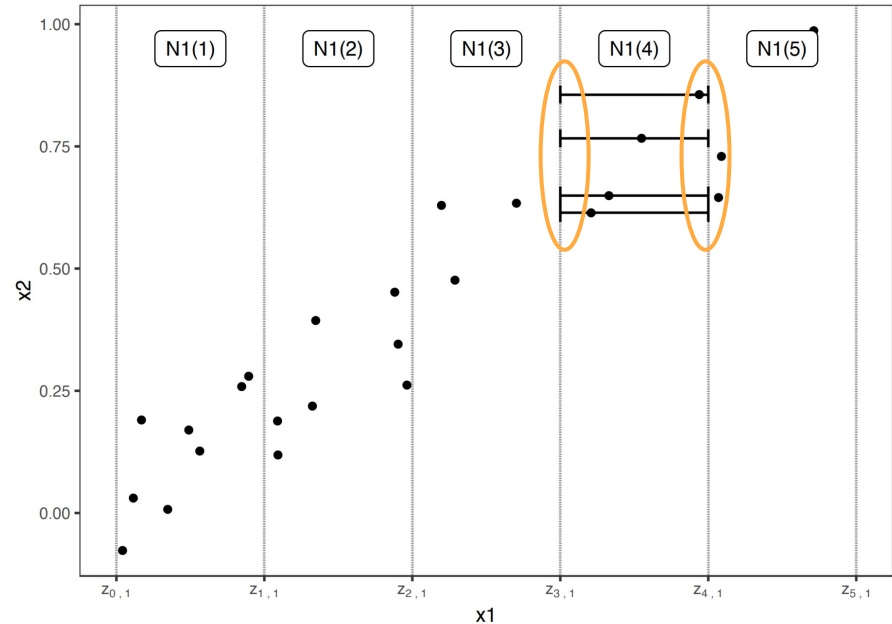


Fig. 8.7 from Molnar's textbook

Accumulated Local Effects (ALE)

ALE computes the average change in prediction over a series of small windows.

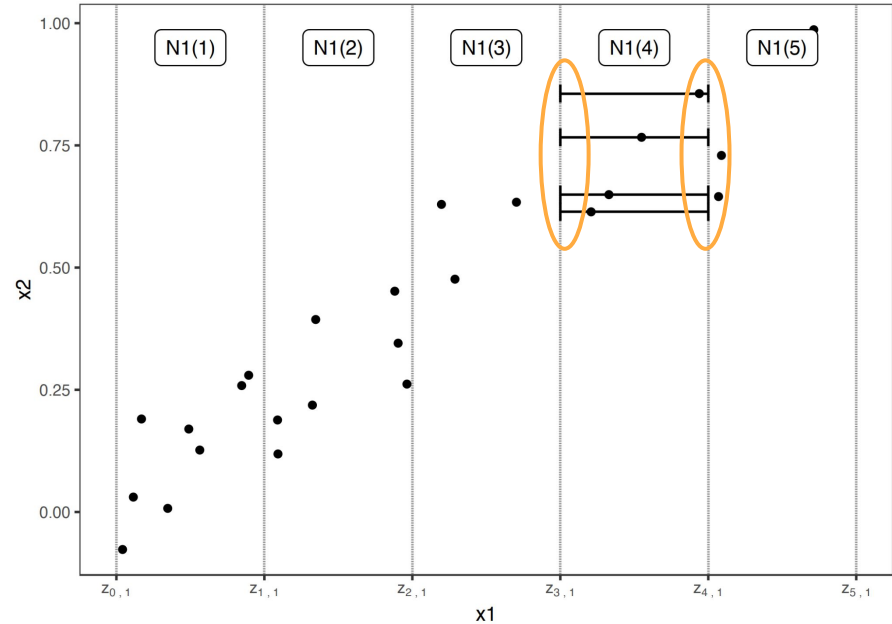
For a given bin, we set the values of a given feature to left side of bin, compute the prediction, and repeat for the right side of the bin.



Accumulated Local Effects (ALE)

ALE computes the average change in prediction over a series of small windows.

For a given bin, we set the values of a given feature to left side of bin, compute the prediction, and repeat for the right side of the bin. We then compute the average change in prediction. Lastly, we compute the accumulated sum over the different bins.



Comparison of PD and ALE

Imagine the data with the following relationship:

$$y = x_1 x_2 x_3$$

Let's assume that x_2 and x_3 are correlated with each other. If fit a model to this data, we may get the following ALE/PD curves shown right.

Takeaways:

- ALE for x_1 is shaky and not showing the clear linear relationship like PD
- PD is failing to identify the quadratic relationship for x_2 and x_3 due to its insensitive to correlations.

From Limitations of Interpretable Machine Learning

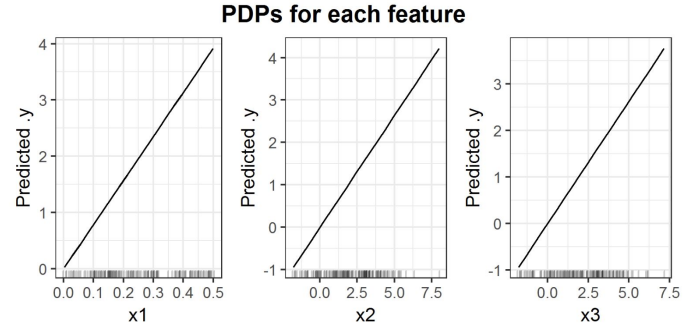


FIGURE 6.1: PDPs for prediction function $f(x_1, x_2, x_3) = x_1 x_2 x_3$.

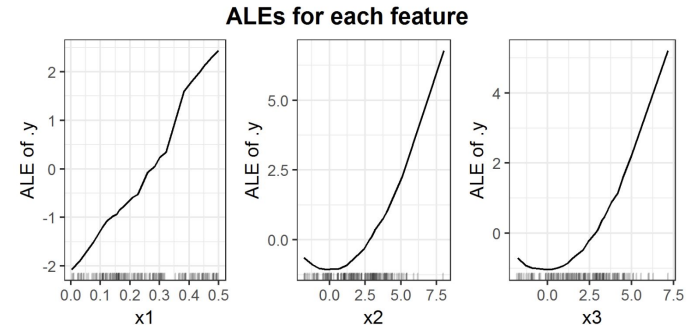


FIGURE 6.2: ALEs for prediction function $f(x_1, x_2, x_3) = x_1 x_2 x_3$.



ALE and PD can be interpreted as additive descriptions of an ML model's prediction

ALE and PD can be computed for more than one feature and we can use them to approximate model predictions.

$$f(x) = \underbrace{f_0}_{\text{Intercept}} + \underbrace{\sum_{j=1}^P f_j(x_j)}_{\text{1st order effects}} + \underbrace{\sum_{j < k}^P f_{jk}(x_j, x_k)}_{\text{2nd order effects}} + \dots + \underbrace{f_{1, \dots, P}(x_1, \dots, x_P)}_{\text{P-th order effects}}.$$



Limitations of ALE

- Due to binning, the curve can be “shaky” and give false impressions of the learned relationship.
- Possible bias near the edge of the distribution due to limited sample size
- Interpretation is not as straightforward as partial dependence and more difficult to implement.
- ALE is not completely devoid of the impact of strongly correlated features (a limitation of nearly all explainability methods).



Feature Attributions

How do the features individually contribute to a model's prediction?



When using ML models, we want the story behind the prediction



When using ML models, we want the story behind the prediction

Let's say we have an ML model that predicts the likelihood of a tornado in the next hour.

$P(\text{tornado}) = 60\% \rightarrow$ Why?

When using ML models, we want the story behind the prediction

Let's say we have an ML model that predicts the likelihood of a tornado in the next hour.

$P(\text{tornado}) = 60\% \rightarrow$ Why?

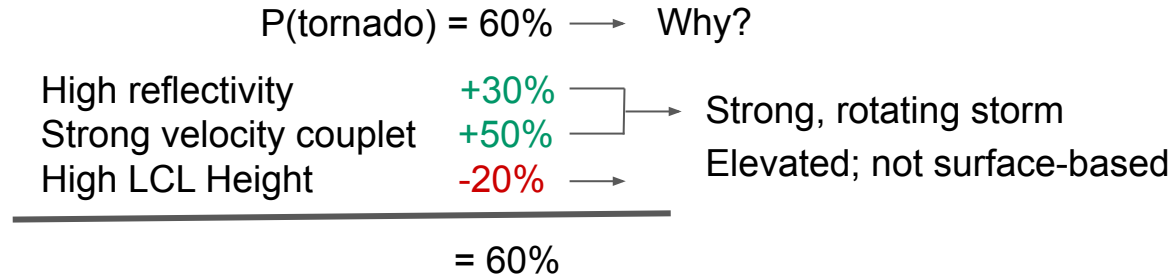
| | |
|-------------------------|------|
| High reflectivity | +30% |
| Strong velocity couplet | +50% |
| High LCL Height | -20% |

= 60%



When using ML models, we want the story behind the prediction

Let's say we have an ML model that predicts the likelihood of a tornado in the next hour.



Shapely Value and Game Theory

How can we divide money (model's prediction) between players (features) in a fair way?

Fairness properties (in terms of ML):

1. **Additivity:** Sum of features contributions must equal the prediction
2. **Consistency:** If we change the ML model such that the feature has a stronger effect, then its contribution must likewise increase
3. **Missingness:** If a feature is missing, then its contribution must be zero.

Shapely values are the only solution that satisfies all 3 of these properties!



Shapely Values

The Shapely value (Φ_i) for feature x_i is the weighted average difference in model prediction when it is included and not included in some subset of features for all possible features subsets.

$$\phi_i = \text{Average over all Features' subset } S \subseteq M/\{i\} \left(\begin{array}{c} \text{Marginal Contribution} \\ f(S \cup \{i\}) - f(S) \end{array} \right)$$

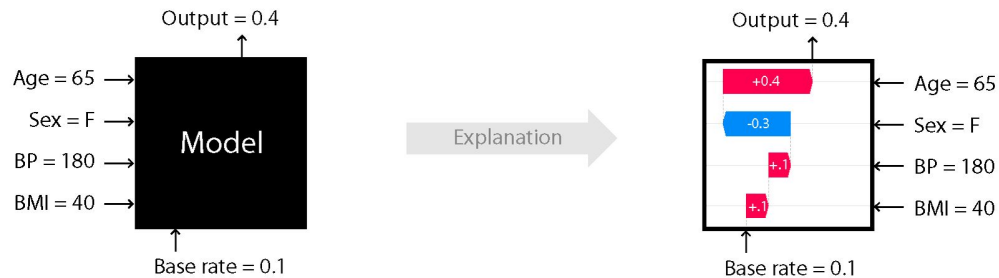
Prediction difference with and without a feature

SHAP, Shapely values, and Feature attributions

The SHapely Additive Explanation (SHAP) python package combines Shapely values with the concept of feature attribution i.e., expressing model predictions as a linear sum of Shapely values from each feature.



SHAP



Computing exact Shapely values for larger datasets is practically impossible

Requires creating $P!$ possible subsets (P = number of features) and being able to have account for “missing” features in the ML model.

The SHAP python package can compute the exact Shapely values, but this is only viable for datasets with a small number of features.

To approximate the Shapely values, the SHAP python package uses two methods:

1. Only use a small set of feature subsets, but do it in a smart way
2. For the “missing” features, replace their values with values from a background dataset (usually 100-1000 samples for the training dataset) and compute the sample-average Shapely value



Smart feature subsetting: SHAP's Permutation-based Method

| X ₁ | X ₂ | X ₃ |
|----------------|----------------|----------------|
| Missing | Missing | Missing |
| Present | Missing | Missing |
| Present | Present | Missing |
| Present | Present | Present |
| Missing | Present | Present |
| ... | ... | ... |

Diagram illustrating the SHAP's Permutation-based Method. The table shows feature values (X₁, X₂, X₃) for various instances. To the right, arrows indicate the permutation of features, resulting in Shapely values $\phi_{1,1}$, $\phi_{2,1}$, $\phi_{3,1}$, $\phi_{1,2}$, and $\phi_{2,2}$.

Uses the simplest possible feature subsets to compute the Shapely values.

- Forward and backward permutations (see diagram right)

For “missing” features replace their value with a reference value

- For robust results, use multiple reference values and compute the average Shapely value.

Owen values (Shapely values based on feature cohorts)

Rather than treat features independently, we can define feature hierarchies

- Defined manually or through clustering techniques (e.g., based on correlations)

The feature subsetting is then determined based on these feature coalitions

- Instead of removing a single feature, we remove multiple features based on the coalition
- Dramatically reduces the computation run time.

Shapely values based on these feature hierarchies are known as *Owen Values*



Model-specific versions of SHAP

The permutation-based method is a model-agnostic approach, but there are model-specific alternatives, which are either more computationally efficient or offer higher accuracy

- **Neural networks**

- Based on the DeepLIFT (*Deep Learning Importance FeaTures*) algorithm, which is an additive feature attribution method for neural networks.
- The algorithm works by replacing values for missing features with a reference values and then using backpropagation to evaluate how the model prediction changes

- **Decision Trees, Random Forests, and Gradient-Boosted Trees**

- Uses the structure of the tree to derive the Shapely values
- Longer run time than the permutation-based method, but does return the exact Shapely values



Limitations of SHAP

- **Assumes a prediction can be represented by a linear combination**
 - The fundamental assumption of SHAP is that an ML prediction can be represented by a sum of contributions from each feature. For a highly nonlinear models, this assumption may not hold true.
 - Can be difficult for the end-user to think of a model prediction in this way.
- **Difficult to interpret when features are correlated.**
 - Shapely values, like other explainability methods, assume the features are independent. The alternative is to compute the Owen values, but the user has to declare the feature hierarchies.
- **Computational inefficiency**
 - To efficiency compute SHAP values, multiple approximation are made, which reduces accuracy. Despite those efforts computing SHAP for multiple examples can take awhile, especially for models with a large number of features.
- **Limited information about feature interactions**
 - When strong feature interactions are present, slight changes to a single feature's value can lead to a dramatic change in the SHAP values.
- **Constant requirement of a reference dataset.**
 - Computing SHAP values requires a background dataset, which may be limiting in operational settings.



What's ahead

Later today:

Trust-a-thon!

Tomorrow:

Explainability, Interpretability, and XAI for Deep Learning

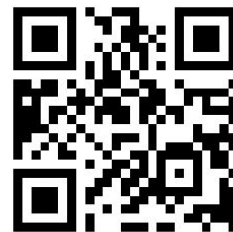


Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Day 1: Agenda

- 9:00 Welcome and Overview
- 9:10 What does it mean to trust?
- 9:40 *Short brain & bio break*
- 9:45 Meaningful interdisciplinary work
- 10:25 *Short brain & bio break*
- 10:30 Evaluation metrics
- 11:00 XAI for traditional ML

Time for any open questions!



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

Or go to sli.do
and use the
code TAI4ES



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!

