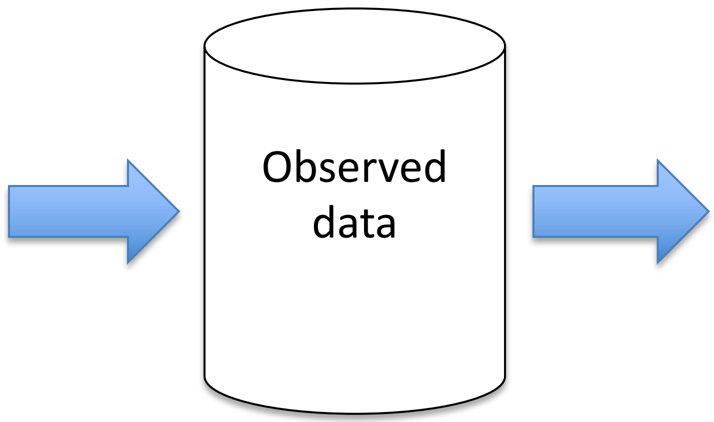
Causality 101 for Geoscientists & Strategies for Successful Collaboration

Imme Ebert-Uphoff

Research Faculty, Electrical and Computer Engineering,
Colorado State University



What/how can we learn about causality from observed data?

AGU Workshop on
Emerging Data Science and ML Opportunities in the Weather and Climate Science
Dec 13, 2018

Overview

1. **A few concepts from causality theory** – *to convince you that this is **solid science** and give you some **intuition.*** ***[Link to tutorial paper for climate scientists]***
2. **Applications in climate science** – *to convince you that this is **useful in climate science**.*
3. **Strategies for successful collaboration** between climate scientists and data scientists.

Causal Discovery Methods

- **Seek to identify cause-effect relationships from observed data.**
- **A few milestones:**
  - **Granger** (1969): **Granger causality** - Causality defined based on predictability.
  - **Pearl** (late 1980s):
    - Causal Calculus.
    - Graph language, probabilistic graphical models.
  - **Spirtes, Glymour, Scheines** (1990s): *[LINK TO FREE BOOK]*
    - Practical algorithms for causal search.
    - Dealing with hidden common causes.

Two Types of Causality Studies

1) Intervention Study: when interventions are possible.

Supports **necessary** and **sufficient** conditions for causality.

But: In climate science rarely possible!

2) Observational Study: purely from observations / model output.

Only supports **necessary** conditions for causality

→ Weaker statements possible, but still powerful.

→ Topic of **this talk**.

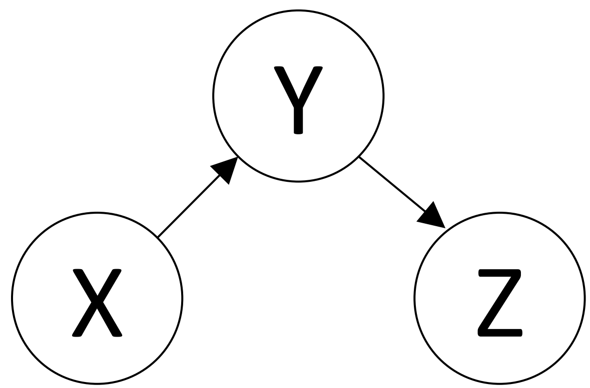
Concept 1: Graphs as Language for causal models

Express causal relationships as graph

- Variables are nodes of graph.
- Arrows indicate: **cause → effect**.

In this example:

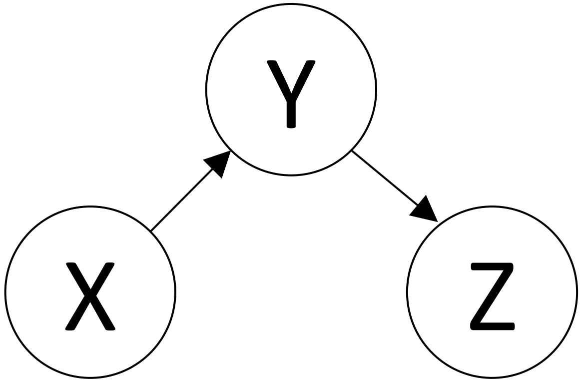
- Three variables.
- X is a cause of Y.
- Y is a cause of Z.



You should have a question here…

Concept 2: **Direct** vs. **indirect** connections

Arrows indicate **direct** causes only.



In this plot:

- X is a **direct** cause of Y.
- Y is a **direct** cause of Z.
- X is only an **indirect** cause of Z.

Goal of causal analysis: we want to identify only direct connections. Eliminate all others.

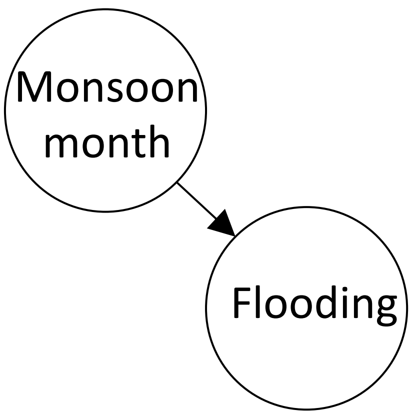
Why eliminate indirect connections?

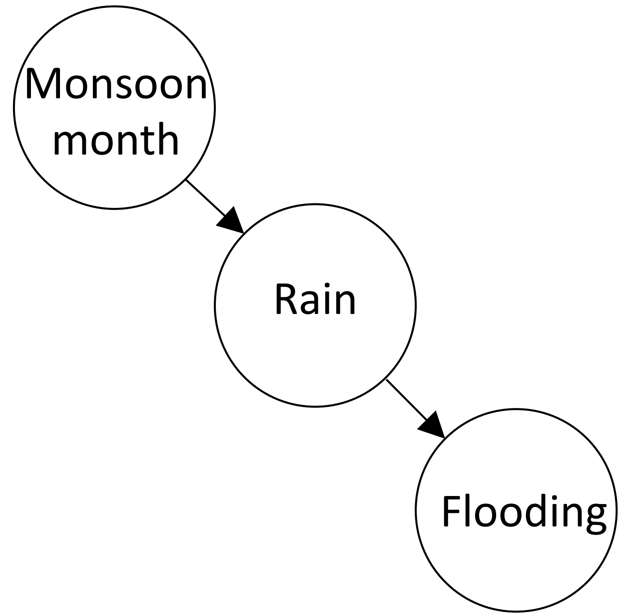
1) Sparsity, simplicity.
2) Only then can you understand effect of interventions!

Concept 3: **Directness** is **relative** property

One can always transform a direct connection into an indirect one by including an intermediate cause!

Toy example:





Monsoon month is **direct** cause of flooding in this model.

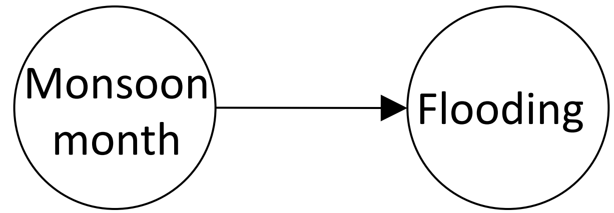
Monsoon month is only **indirect** cause of flooding in this model.

Both models are correct!

Directness is only defined *relative to variables included in model*.

Concept 4: Causality is probabilistic relationship

Example:



This graph implies:

1) Flooding is *more likely* in monsoon months, but *not* certain.
2) Flooding can also happen outside of monsoon months.

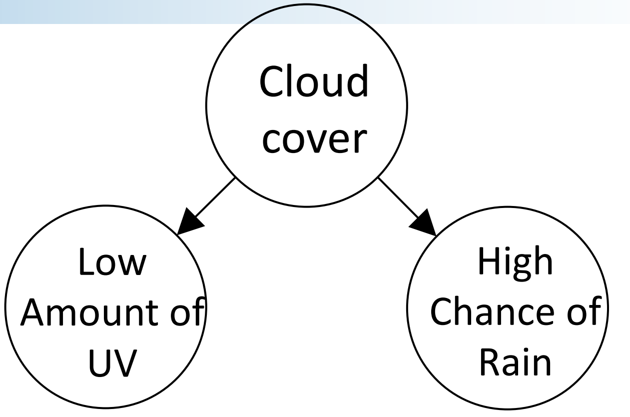
→ Supplement graph with probabilities.
→ **Probabilistic graphical model.**

When learning these models from data:
Step 1: Identify **graph structure** from data – hard!
Step 2: Determine probabilities afterwards – very easy!

Here: Care only about graph structure.

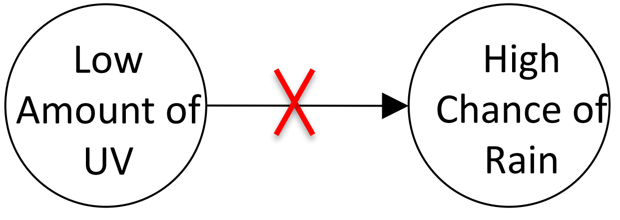
Concept 5: Hidden common causes (latent variables)

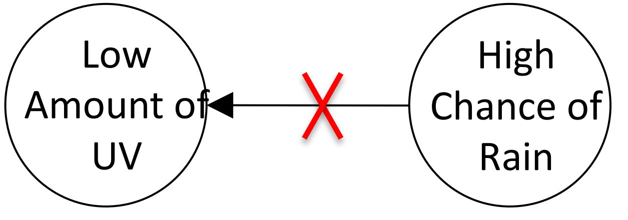
Ex.: Cloud cover is **common cause** of "low UV" and "high rain".



If we remove the common cause (Cloud cover) in model:

Can no longer get a correct causal model!





Conclusion:

1) We can never prove causal connections (w/o interventions).

2) But we can disprove causal connections (w/o interventions).

→ Tool for that: Conditional independence tests.

How can we remove connections based on data?

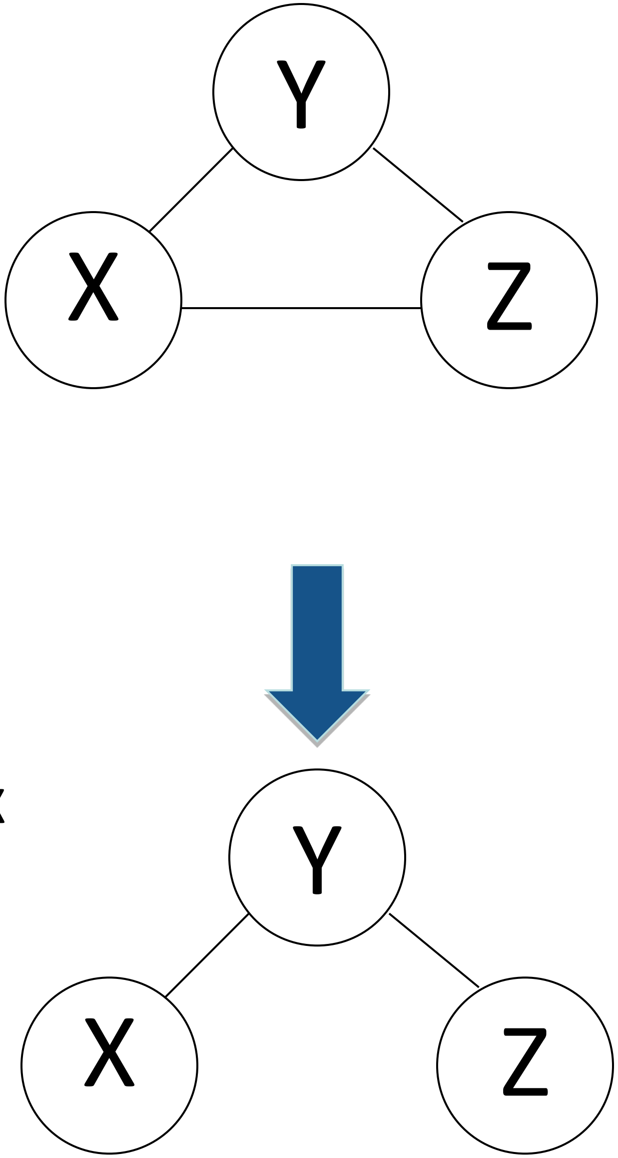
The following 4 questions are equivalent:

1) Can we eliminate edge between X and Z?

2) Is there *direct* connection between X and Z?

3) "Is X conditionally independent of Z *given Y*?"

4) Is $P(X \mid Y, Z) \approx P(X \mid Y)$?



If yes for any of the above: eliminate edge between X and Z.

→ Use conditional independence test:

Many statistical tests available to test for conditional independence.

An elimination algorithm: **PC algorithm**

Now we have: **Statistical test to detect and eliminate *indirect* connections** (graph edges).

Basic algorithm for learning <u>independence graph</u> from data:

1. Nodes of graph = observed variables.
2. Start with **fully connected graph** = assume that every variable is a cause of every other variable.
3. Eliminate as many edges as possible using conditional independence tests.
4. Establish arrow directions (using more statistical tests or temporal constraints).

Elimination procedure.

Whatever is left at end: **potential causal connections.**

Whatever is left at end: **potential causal connections**.

- In climate science there may always be a **hidden common cause**
  - that we are not aware of,
  - that cannot be measured,
  - or including them all may make model too complex.

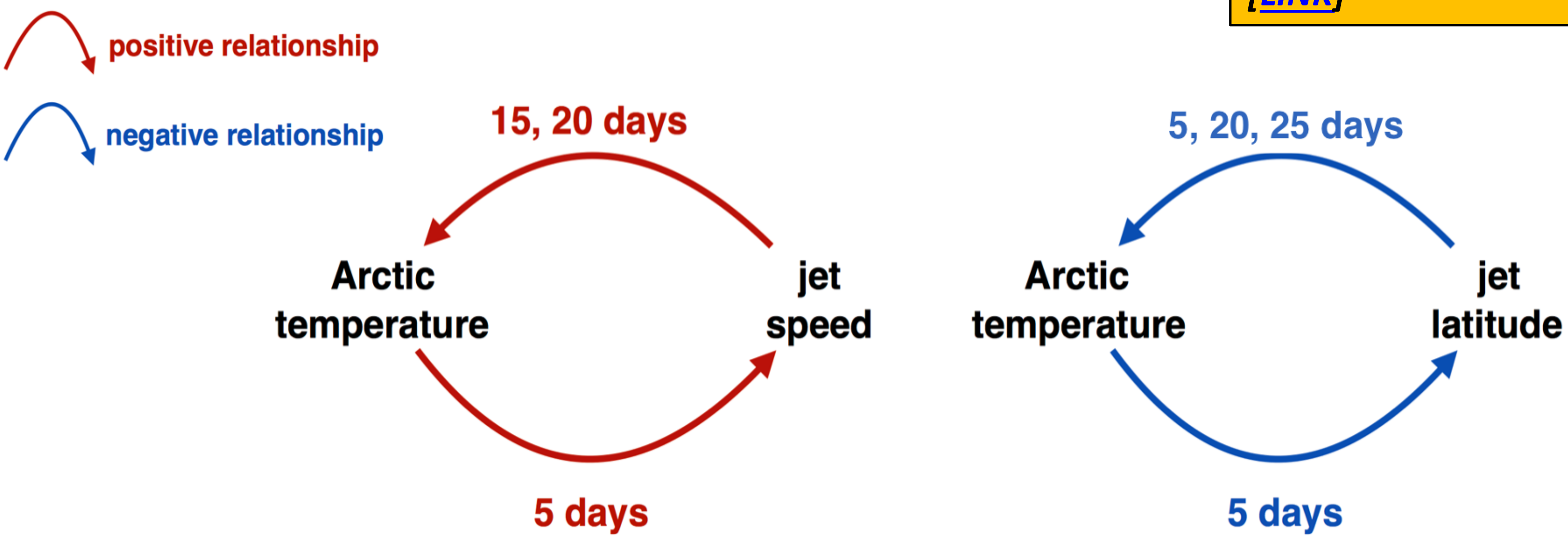
- Need to keep that possibility in mind when interpreting results
  → **results are only causal *hypotheses*.**

- **Each hypothesis could be direct connection, due to hidden common cause, or combination of both.**

Application 1: Arctic connection to jet stream

Science question: What is the effect of arctic temperature on speed / latitude of jet stream, and vice versa?

Samarasinghe, McGraw, Barnes, Ebert-Uphoff, Environmetrics, 2018. *[LINK]*



- Dominant relationships: Positive for jet speed, negative for jet latitude.
- Both are thus positive (reinforcing) feedback loops.

Application 2: Spatially-distributed systems

Ebert-Uphoff & Deng, GRL 2012. *[LINK]*

Nodes = grid points (each with associated time series)

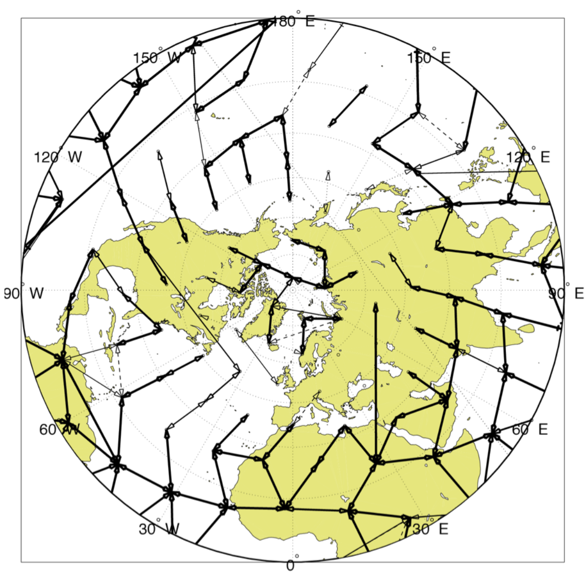
Input: Atmospheric field on global grid

Example: 500 mb geopotential height

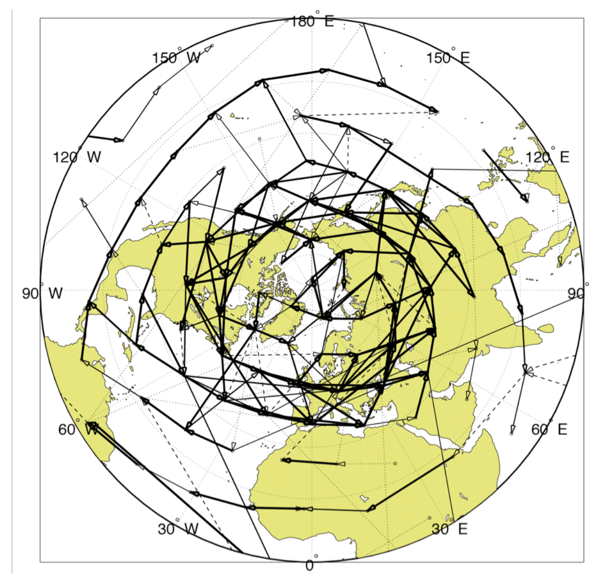
- NCEP/NCAR Reanalysis, 1948-2011, results for winter (DJF months)

Output:

- **Interactions between grid points.**



(a) 0-day-delay

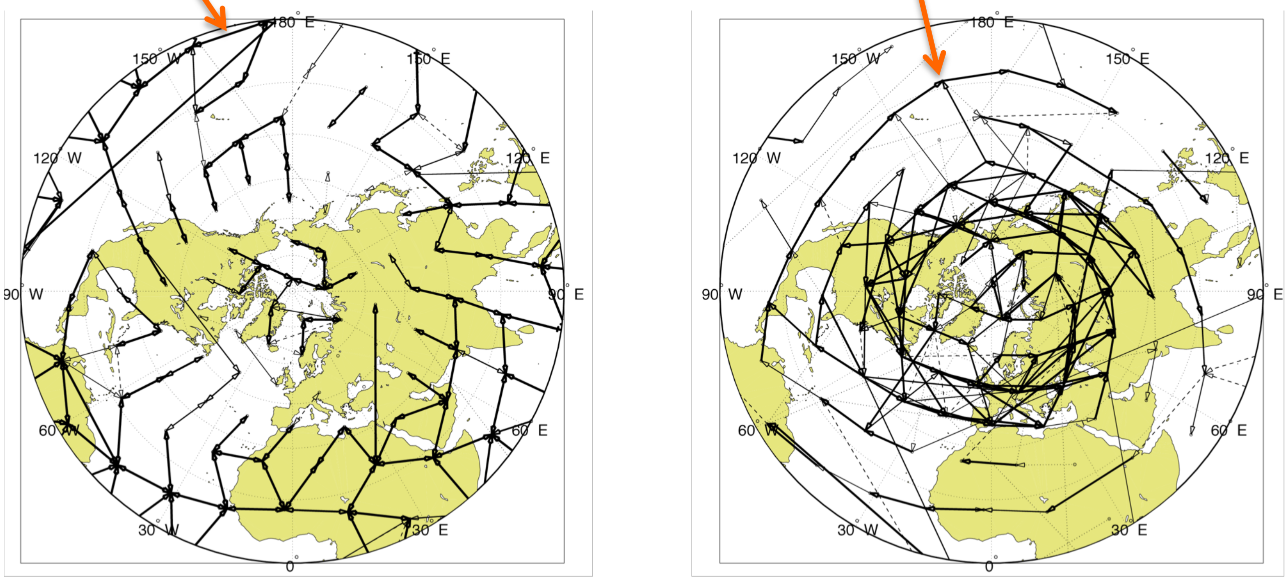


(b) 1-day-delay

Evaluation Step

Due to dominant **diffusion** processes near equator

Due to **advection** processes (storm tracks)



(a) 0-day-delay

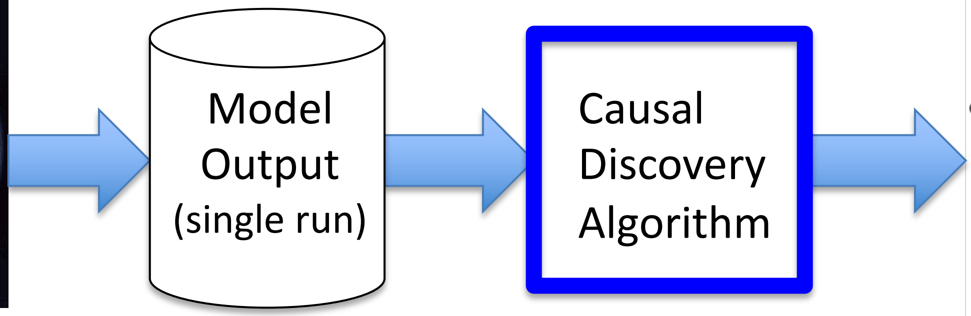
(b) 1-day-delay

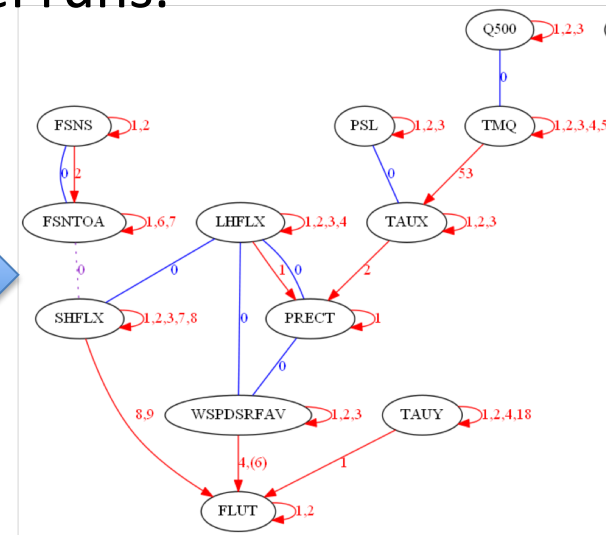
Application 3: Apply to Climate Model Runs

Baker et al., [*LINK*] Geoscentific Model Development, 2016

Determine "causal signatures" of climate model runs.

CESM Model





- Calculate "causal signature" for individual model outputs (e.g. different initial conditions), then compare their "signature".
- First experiments: use only 15 variables, use **global averages**.
- Applications: effect of compression, error check, understanding of differences between ensemble members or models.

Causal Discovery - Summary

Limitations:

- **Causal interpretation requires caution:** can only identify *potential* cause-effect relationships.
- **Further limitations discussed in:** *Jakob Runge, Causal network reconstruction from time series, Chaos, 2018.* *[LINK]*
- **Causal discovery is *not* a magic bullet.**
- But solid tool - **underutilized** in the geosciences.

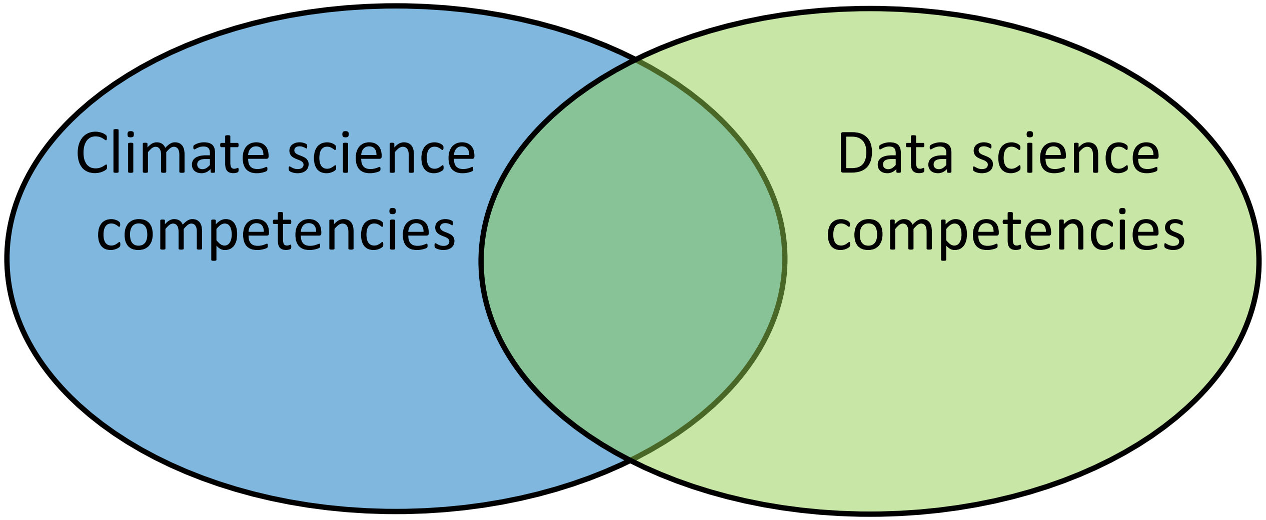
Proposed Use:

- Can help climate scientists **sift through increasing amounts of available data** to generate new hypotheses.
- **Primary purpose: Generate hypotheses.**

Tough Question

How many different disciplines
do you have to cover in your team
for efficient interdisciplinary collaboration
in climate science
and data science?

Areas of expertise



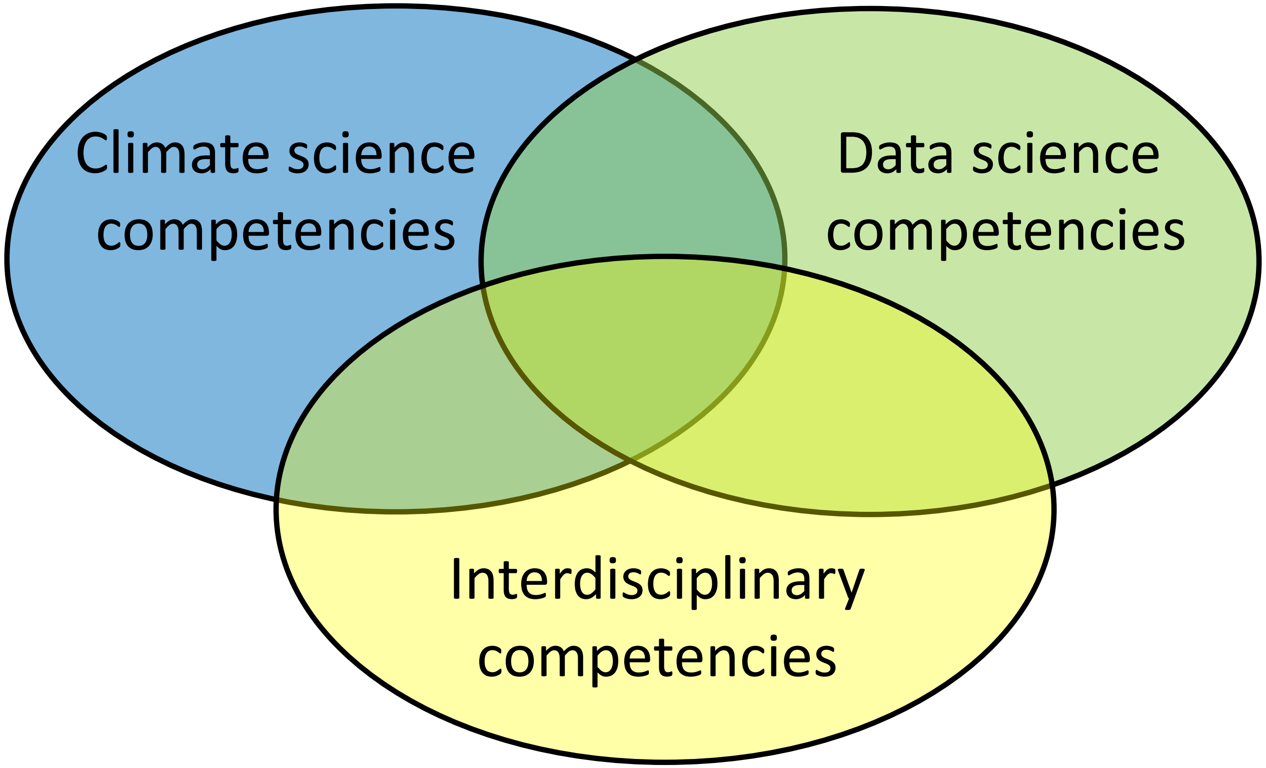
Here are two areas – but what else?

Question

Hint: the 3rd area is mentioned on this slide.

How many different disciplines
do you need to cover in your team
for efficient interdisciplinary collaboration
in climate science
and data science?

<u>Three</u> areas of expertise



Important: The three different areas do <u>not</u> have to be represented by 3 different people. For example, the idea is more to have the team members learn the interdisciplinary skills, rather than bringing in another person.

Interdisciplinary Studies

- **Is its own research area!**
- Goal: Draws on disciplines to *integrate* their insights.
- *Integration* literally means to make whole. […] integration is a process by which ideas, data and information, methods, tools, concepts and/or theories from two or more disciplines are **synthesized, connected, or blended**.

Source: Repko, Allen F., Interdisciplinary Research: Process and Theory. SAGE Publications, 2011. *[LINK]*

Perfect example of full integration:

- Theory-guided data science (TGDS – Karpatne et al.).

Interdisciplinary Habits of the Mind

Source: Newell and Luckie, *Pedagogy for Interdisciplinary Habits of the Mind*, Conference on Interdisciplinary Teaching and Learning, 2012. [*LINK*]

Subset of ID habits of the mind:

- Set aside personal convictions;
- ***Strive for a feel of each discipline's perspective;***
- Embrace contradictions (ask how it can be both);
- Look for unexamined linkages and unexpected effects;
- Strive for balance (among disciplinary perspectives)
- ***Don't fall in love with a solution until you understand the full complexity of the problem;***
- ***Value intellectual flexibility and playfulness.***

Helpful Personal Qualities and Skills

Foster these skills in yourself & Look for these skills in collaborators.

- **Communication skills, organizational skills;**
- **Broad interest**, **flexibility**, **creativity**, **openness**;
- Tolerance for ambiguity;
- Transcendence of disciplines;
- Respect toward people, perspectives, and cultures;
- Scientific skills for gathering, translating, analyzing, structuring, weighting and valuing, and synthesizing knowledge and information.

Source: Flinterman et al., *Transdisciplinarity: The New Challenge for Biomedical Research*, Bulletin of Science, Technology & Society, Vol. 21, No. 4, 2001. *[LINK]*

Resources

A) Literature on interdisciplinary studies:

Newell & Luckie, Repko, many others. See links above.

Our own work: Pennington et al., Bridging Sustainability Science, Earth Science, and Data Science through Interdisciplinary Education, submitted to *Sustainability science* (in review), 2018.

B) Examples of Interdisciplinary Graduate Programs (NSF NRTs) that teach interdisciplinary studies:

1. **Univ. of Chicago:**

   Data science for Energy and Environmental Research *[LINK]*

2. **UC Berkeley:**

   Environment and Society: Data Science for the 21st Century *[LINK]*

3. **Northwestern University:**

   Integrated Data-Driven Discovery in Earth and Astrophysical Sciences *[LINK]*

Meet Peter and Andrea – Two companions throughout this discussion

Peter

Andrea

Source: Cartoon guide: *[LINK]*
Ebert-Uphoff and Deng, Three Steps to Successful Collaboration with Data Scientists, **EOS**, **Aug 2017**.

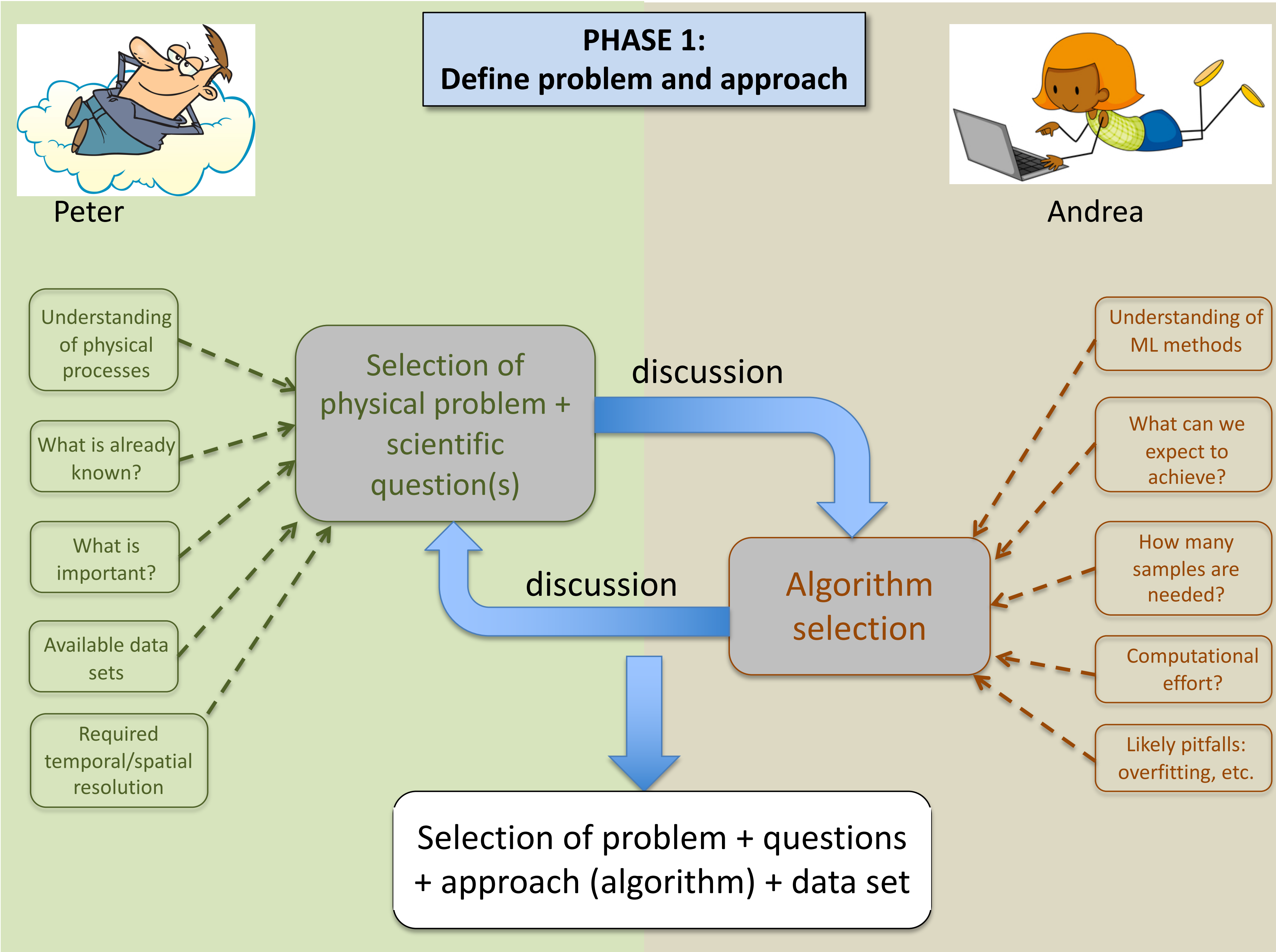
Meet Peter and Andrea – Two companions throughout this discussion

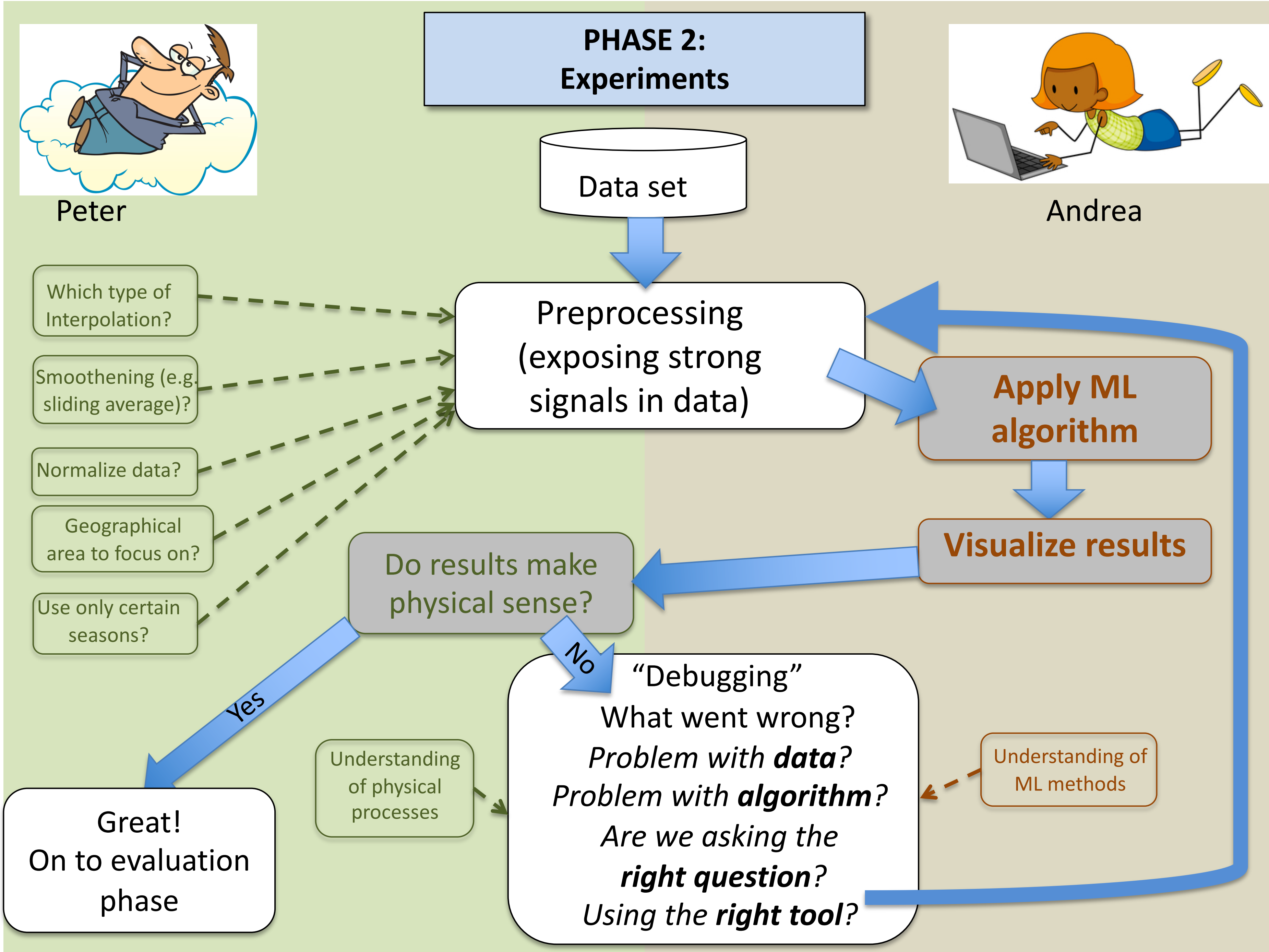
Peter

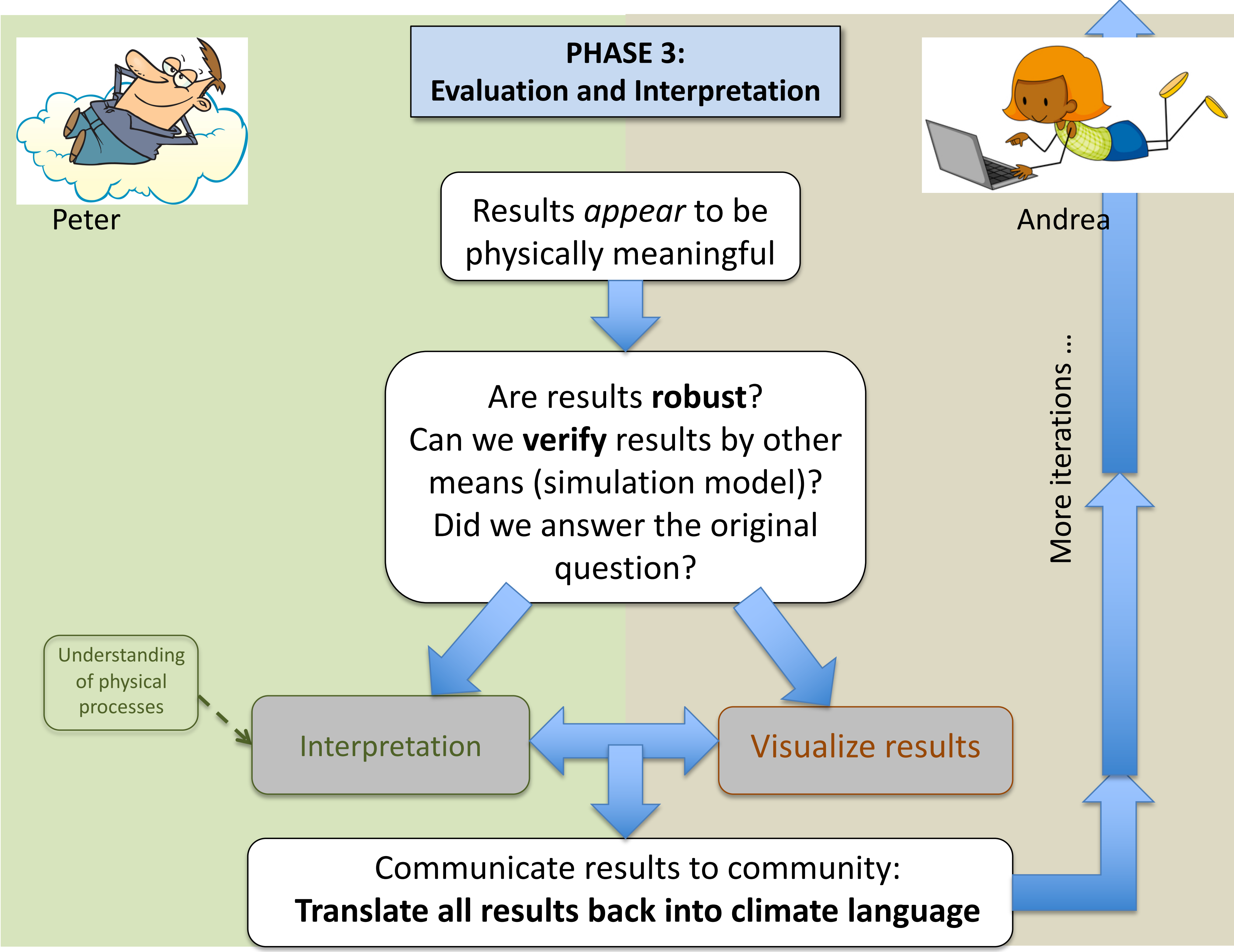
Climate scientist

Andrea

Data scientist


PHASE 1:
Define problem and approach
Peter
Andrea
Understanding of physical processes
What is already known?
What is important?
Available data sets
Required temporal/spatial resolution
Selection of physical problem + scientific question(s)
discussion
discussion
Algorithm selection
Understanding of ML methods
What can we expect to achieve?
How many samples are needed?
Computational effort?
Likely pitfalls: overfitting, etc.
Selection of problem + questions + approach (algorithm) + data set


PHASE 2:
Experiments
Peter
Andrea
Data set
Which type of Interpolation?
Smoothening (e.g. sliding average)?
Normalize data?
Geographical area to focus on?
Use only certain seasons?
Preprocessing (exposing strong signals in data)
Apply ML algorithm
Visualize results
Do results make physical sense?
Yes
No
"Debugging"
What went wrong?
Problem with data?
Problem with algorithm?
Are we asking the right question?
Using the right tool?
Understanding of physical processes
Understanding of ML methods
Great! On to evaluation phase


PHASE 3:
Evaluation and Interpretation
Peter
Andrea
Results *appear* to be physically meaningful
Are results **robust**?
Can we **verify** results by other means (simulation model)?
Did we answer the original question?
More iterations ...
Understanding of physical processes
Interpretation
Visualize results
Communicate results to community:
Translate all results back into climate language

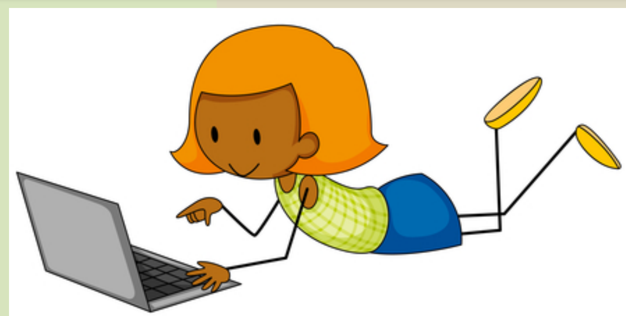
Observations:

1) Many tasks cannot be split into two separate parts that each person works on independently.
2) Many decisions must be made *together,* requiring both of their special knowledge.

Therefore:

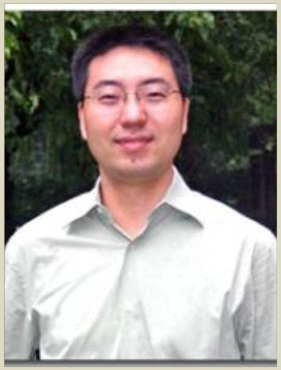
1) Peter and Andrea cannot stay completely on their own side.
2) Each person needs to have a basic understanding of the thinking process of the other person.
3) Each person must be willing to teach / learn some basic vocabulary and tools.
4) Constant feedback from both sides is essential. Talk to each other, talk, talk, then talk some more!

Climate science



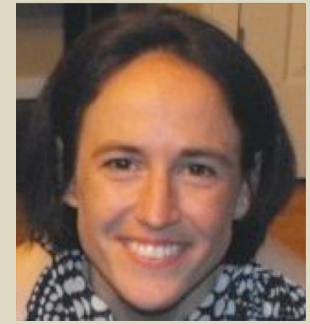
Data science

Causal Discovery Collaborators



Yi Deng
Earth and Atmospheric Sciences, Georgia Tech

Dorit Hammerling
NCAR

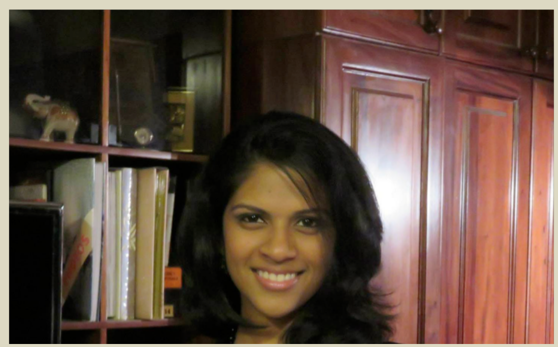


Elizabeth Barnes
Atmospheric Science, Colorado State Univ (CSU).

Allison Baker
NCAR

Collaborators on "Interdisciplinary Collaboration":

- Deana Pennington, UT El Paso.
- Jo Martin, Univ. Vermont.
- Natalie Freed, UT Austin.
- Suzanne Pierce, UT Austin.
- Yi Deng, Georgia Tech.



Savini Samarasinghe
Ph.D. student at CSU
(with Imme).

Marie McGraw
Ph.D. student at CSU
(with Libby Barnes).

Join Existing Communities

1) Climate Informatics

- Annual workshop – often at NCAR
- Go to: Climateinformatics.org

2) IS-GEO

- Intelligent Systems in the Geosciences
- NSF RCN (Research Coordination Network)
- Monthly telecons
- Go to: [Is-geo.org](Is-geo.org)