GOALS & INTERPRETABLE VARIABLES IN NEUROSCIENCE

DAVID DANKS

DATA SCIENCE // PHILOSOPHY
UNIVERSITY OF CALIFORNIA, SAN DIEGO

GOALS & INTERPRETABLE VARIABLES (IN NEUROSCIENCE)

DAVID DANKS

DATA SCIENCE // PHILOSOPHY
UNIVERSITY OF CALIFORNIA, SAN DIEGO

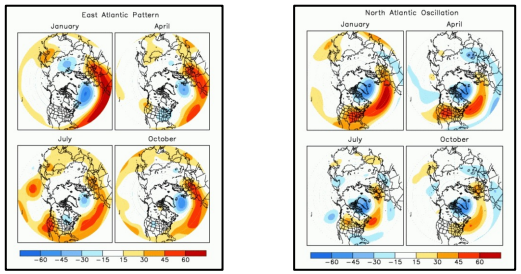
INTERPRETABILITY & MACHINE LEARNING

- *Most common focus*: Interpretable models
  - Restrict (or bias) search space towards "simpler" models
  - *Post hoc* generation of (local) "as if" models

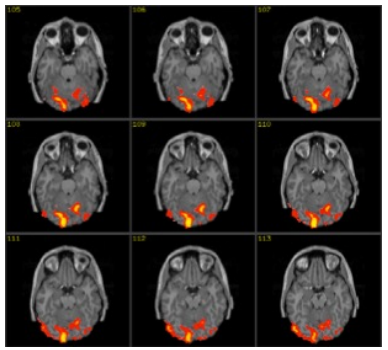
- *Different issue*: What if the variables do not "make sense"?
  - *Related*: Can we interpret intermediate layers of a DNN?

INTERPRETABILITY & MACHINE LEARNING

- **Measured** variables don't necessarily "make sense"
  - Climate science



  - Educational psychology
  - Neuroimaging



Interpretable variables must be constructed / discovered from measures

CONSTRUCTING VARIABLES

- Multiple methods to construct variables
  - High inter-correlation (clustering)
  - Shared motifs (higher-order clustering)
  - High correlation with a target (≈ supervised variable construction)
  - Suitability for a model (e.g., causal feature learning)

CONSTRUCTING VARIABLES

- Multiple methods to construct variables
  - High inter-correlation
  - Shared motifs
  - High correlation with a target
  - Suitability for a model

Correspond to different loss functions (= values)

⇒ Different methods will be appropriate for different goals

CONSTRUCTING VARIABLES: IMPACT OF GOALS

- *Claim*: Different methods can lead to different "best" variables ***even in the large sample limit***
  - (theorems, etc. omitted for time…)

- ⇒ Variable construction can depend on our goals

- ⇒ Interpretable ML can depend on our goals
  - And not only for which model-type or *post hoc* construction

CONSTRUCTING VARIABLES IN NEUROSCIENCE

- Neuroimaging variables often not interpretable (space or time)
- Different criteria for variable construction (many methods):
  - High local inter-correlation
  - High predictive power
  - Coherent causal dynamics (static *or* dynamic)

- Sometimes, different variable sets are constructed
  - But choice of criterion depends on goals
  - ⇒ Usable neuroscientific variables can depend on our goals

BROADER IMPLICATIONS (FOR SOCIETY)

- Different variables may be needed for different goals / groups
  - "Incommensurable interpretability"?
- Need to think about *why* interpretability, not just *for whom*
- Causally interpretable ≠ Predictively interpretable
  - May need to choose which is "more" societally important

SUMMARY / CONCLUSIONS

- Interpretability involves variables, not only models
- Interpretable variables can depend on our goals
  - Variables often must be constructed from measurements
  - Constructed variables can depend on criteria / loss functions
  - Choice of criterion depends on our (scientific & other) goals
- And we see goal-dependence in present-day neuroimaging

THANKS!

www.daviddanks.org
ddanks@ucsd.edu // david@danks.org

Key conversationalists:
- Steve Fancsali
- Clark Glymour
- Tae Wan Kim
- DK Lee
- Joy Lu
- Sergey Plis
- Richard Scheines
- Jim Woodward
- Corey Zhou

(and many others)