Beyond Black Box Models in Sensitive Environments

Ketemwabi Yves Shamavu, Ph.D.

Link to recording: https://skyel.net/pages/blog/post/1648526906.html

Collaborative Artificial Intelligence

- Human cognition + Artificial intelligence

- Goal:
  - Trust
  - Social acceptance
  - Better performance
  - Transparency

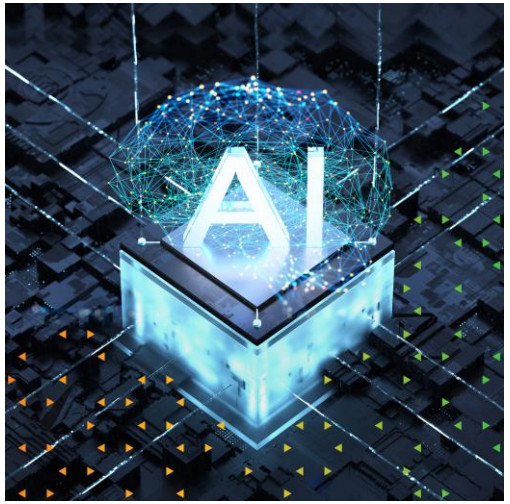
Black Box Models v. Alternatives

- Black box examples:
  - Deep neural networks
  - Random forests
  - Other models with stacks of simpler computations

- Alternatives:
  - Causal networks
  - Transparent models; i.e., Bayesian nets
  - Explainability

Ubiquitous Black Box Models

- Automated interviewing systems
- Clinical decisions
- Financial decisions
- Legal and court decisions

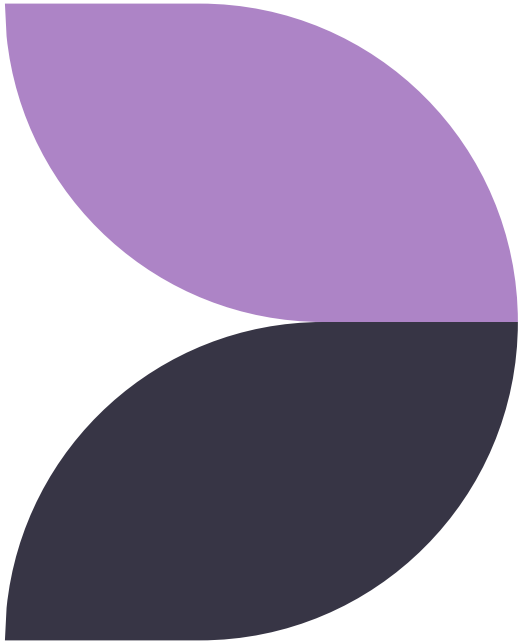


Agenda

- Illustrative Example (Using PyAgrum + Microsoft Azure)
- Inference With Forward Probabilities
- The Case For Inverse Probabilities
- Black Boxes And Explainable Techniques
- Considerations For The Future

PART I

Illustrative Example
(Using PyAgrum + Microsoft Azure)



Bayes Net Structure Inspired By Kedir N Turi, Ph.D. Et Al.

Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors

ORIGINAL RESEARCH — Volume 14 — March 9, 2017

Kedir N Turi, PhD[1]; David M. Buchner, MD, MPH[2]; Diana S. Grigsby-Toussaint, PhD, MPH[2] (View author affiliations)

Suggested citation for this article: Turi KN, Buchner DM, Grigsby-Toussaint DS. Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors. Prev Chronic Dis 2017;14:160244. DOI: http://dx.doi.org/10.5888/pcd14.160244 .

PEER REVIEWED

Abstract

Introduction

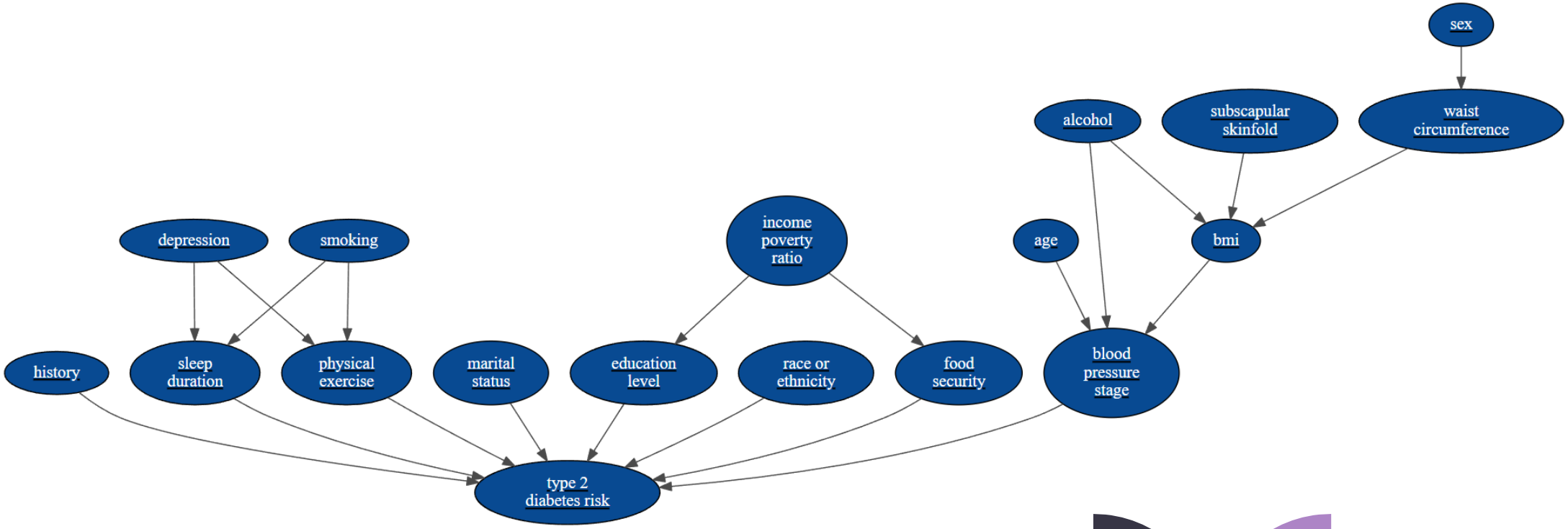
Statistical models for assessing risk of type 2 diabetes are usually additive with linear terms that use non-nationally representative data. The objective of this study was to use nationally representative data on diabetes risk factors and spline regression models to determine the ability of models with nonlinear and interaction terms to assess the risk of type 2 diabetes.

Methods

We used 4 waves of data (2005–2006 to 2011–2012) on adults aged 20 or older from the National Health and Nutrition Examination Survey (n = 5,471) and multivariate adaptive regression splines (MARS) to build risk models in 2015. MARS allowed for interactions among 17 noninvasively measured risk factors for type 2 diabetes.

On This Page

Bayes Net Structure Inspired By Kedir N Turi, Ph.D. Et Al.



*This structure is based on literature but wasn't reviewed by experts

Demo-1

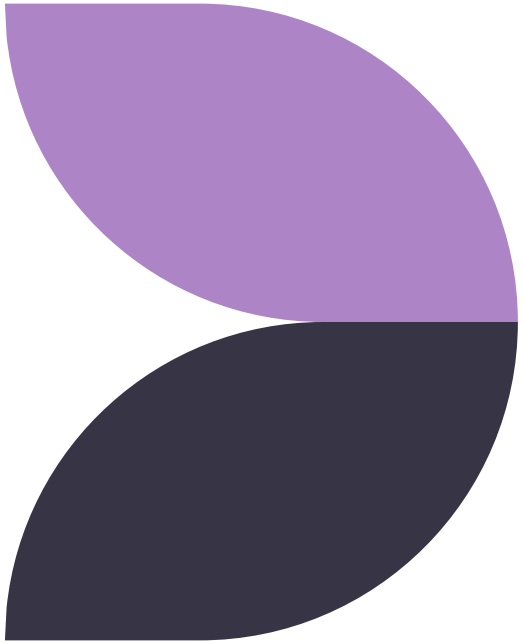
Microsoft Azure Notebooks

+

pyAgrum

PART II

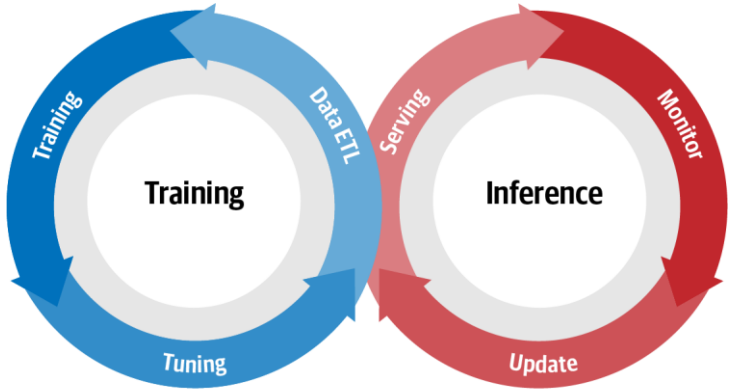
Inference With Forward Probabilities



Training And Inferencing Algorithms

- Training
  - Maximum likelihood estimation: a frequentist approach

- Inferencing
  - Exact
  - Approximate



Approximate Inferencing

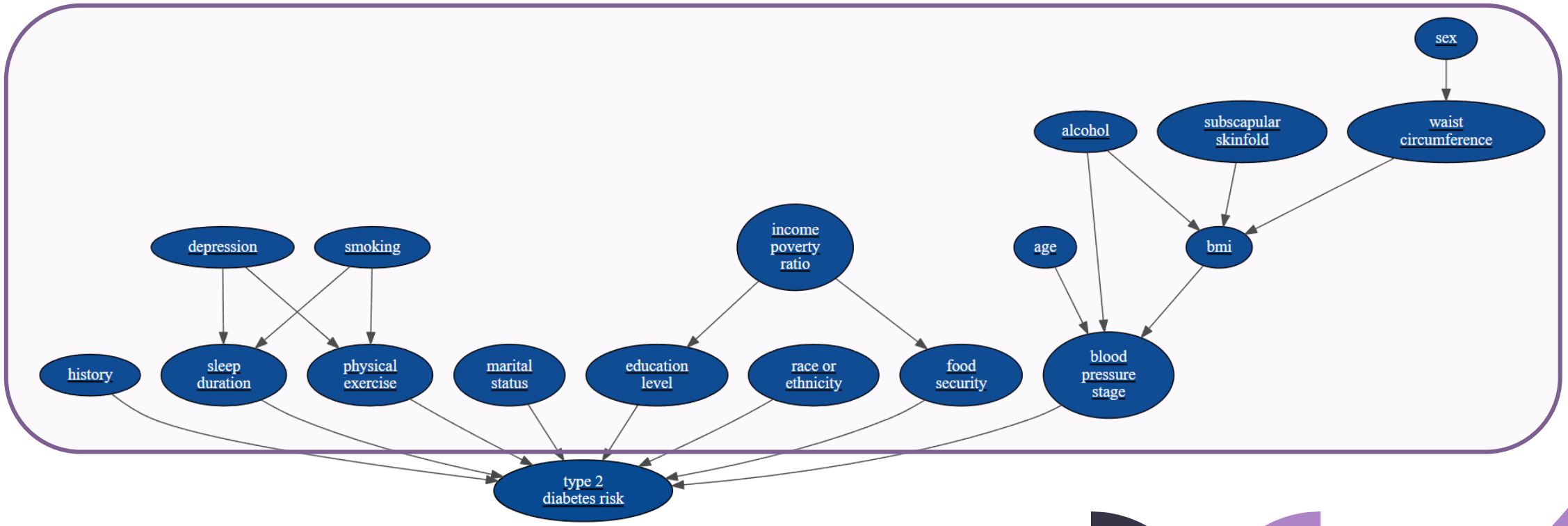
- Stochastic simulation
- Model simplification
- Loopy propagation

Deductive Reasoning

$P(\text{risk} \mid \text{BMI} \geq 27.5) = ?$

From cause to effect

From (Potential) Causes To Effect



*This structure is based on literature but wasn't reviewed by experts

Exact Inferencing

- Lazy propagation
- Variable elimination
- Message passing

Inferencing Tricks

- Transformation into a singly-connected graph
- D-separation
- Markov blanket

Demo-2

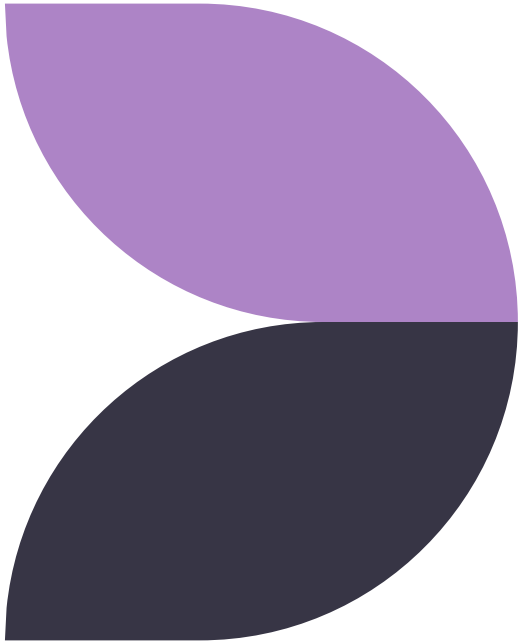
Microsoft Azure Notebooks

+

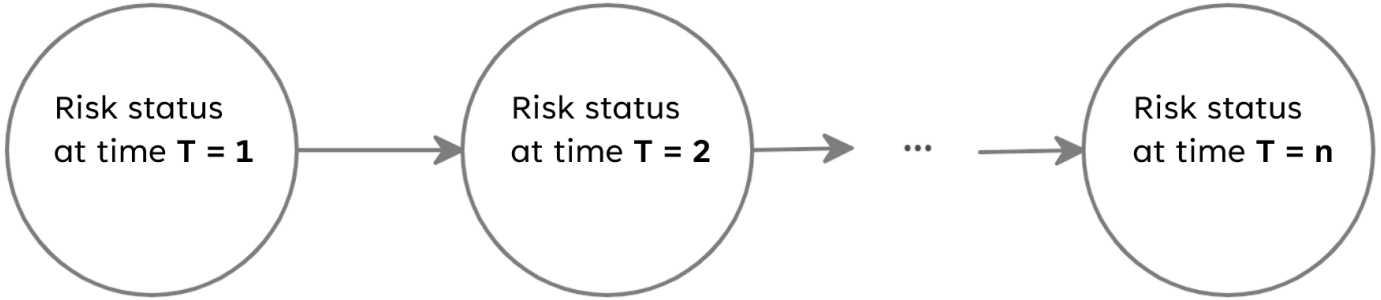
pyAgrum

PART III

The Case For Inverse Probabilities



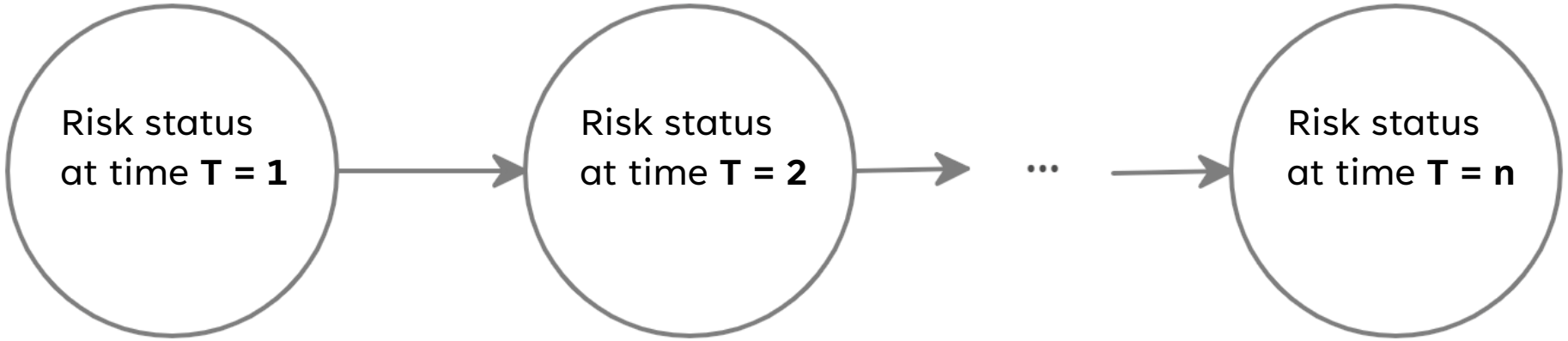
Time And Causality



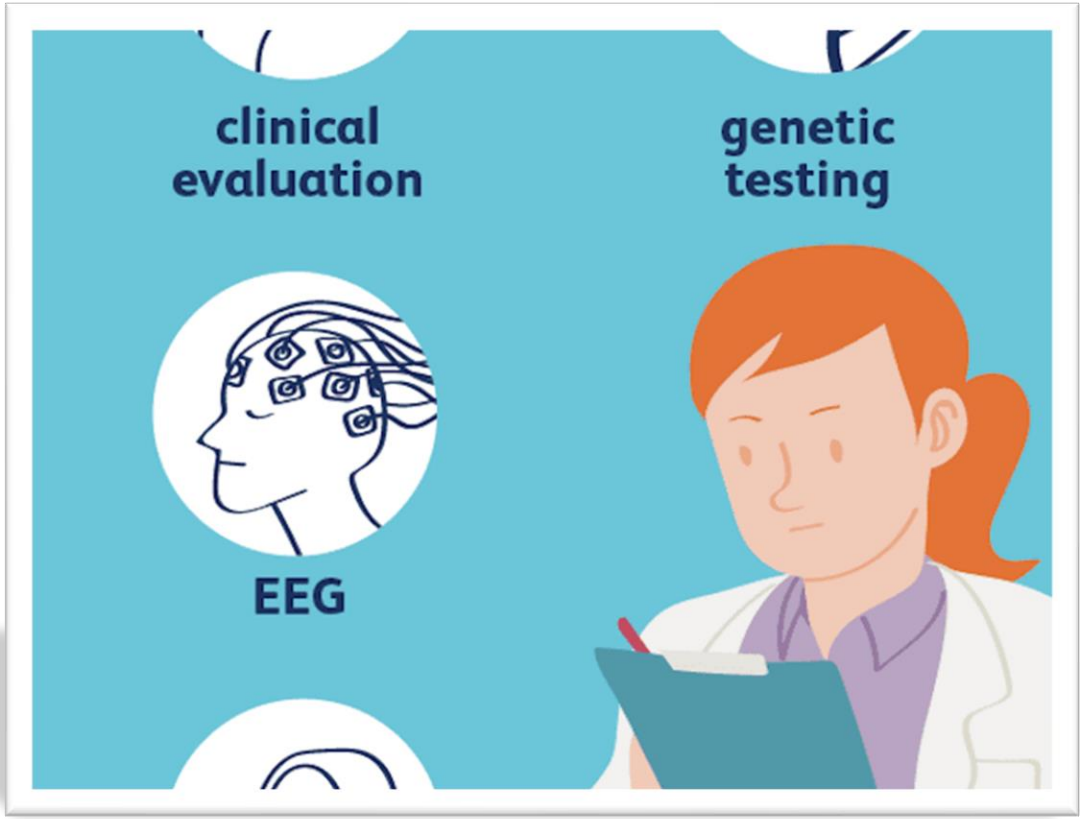
- Time is a precondition for causation

- Effect comes after the cause

From Effect To Cause



Inductive Reasoning



Ladder Of Causation

Imagining

Envision counterfactuals

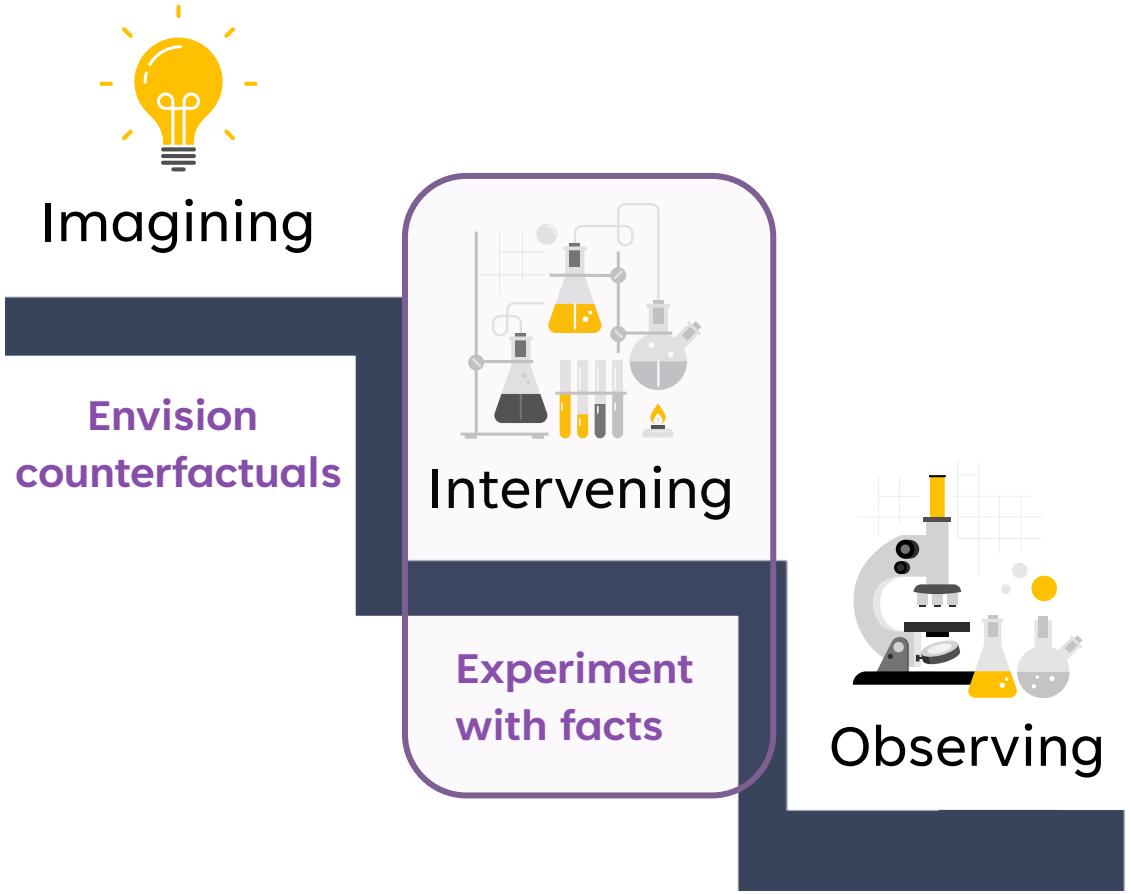
Intervening

Experiment with facts

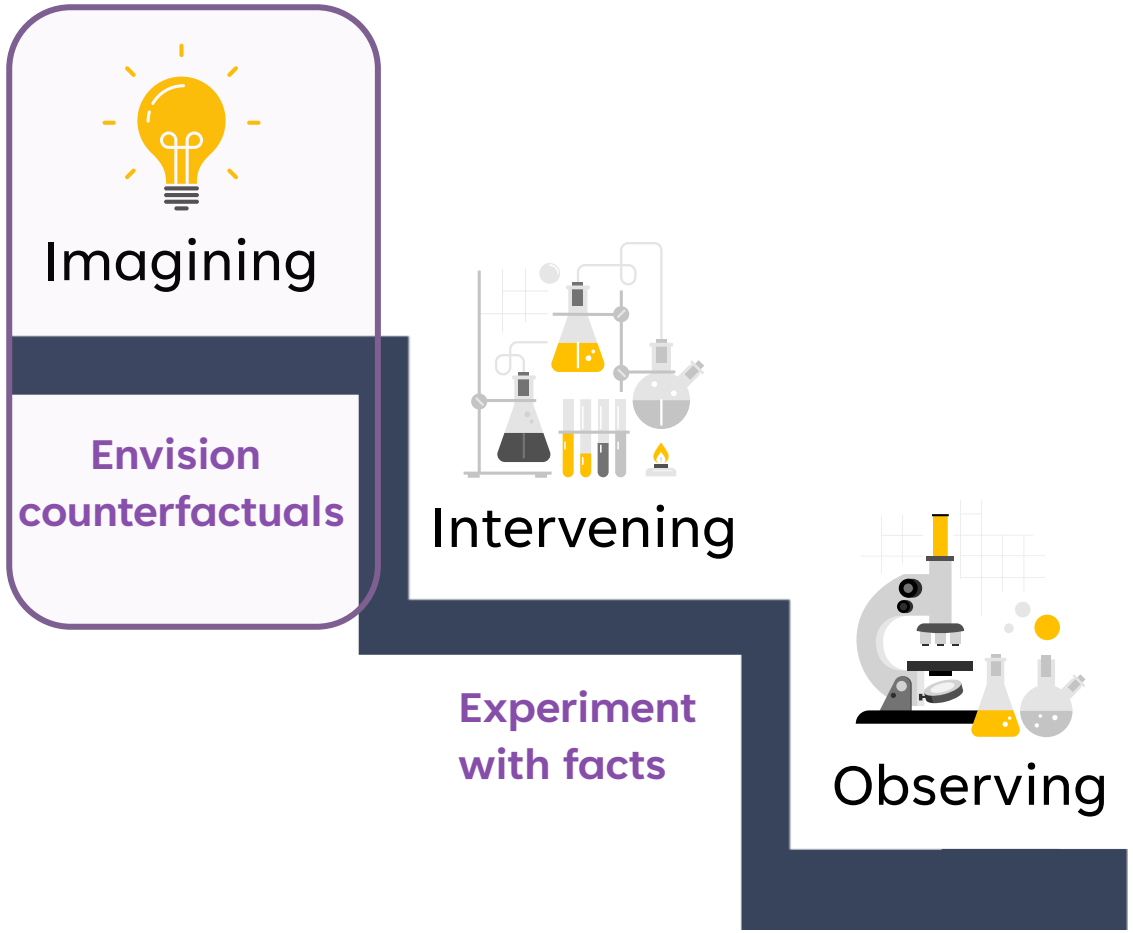
Observing

Establish associations

Experimentation



Counterfactuals



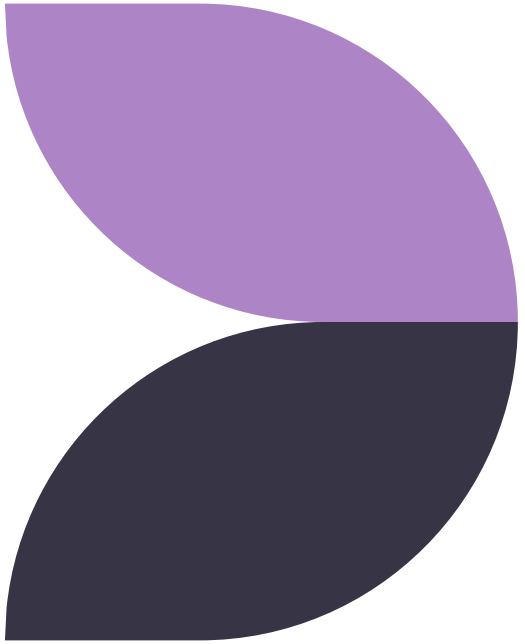
Counterfactuals: Imagination, Abstract Simulations

Inverse Probabilities And Counterfactuals

Beyond feature importance and acausal Shapley values

PART IV

Black Boxes And Explainable Techniques

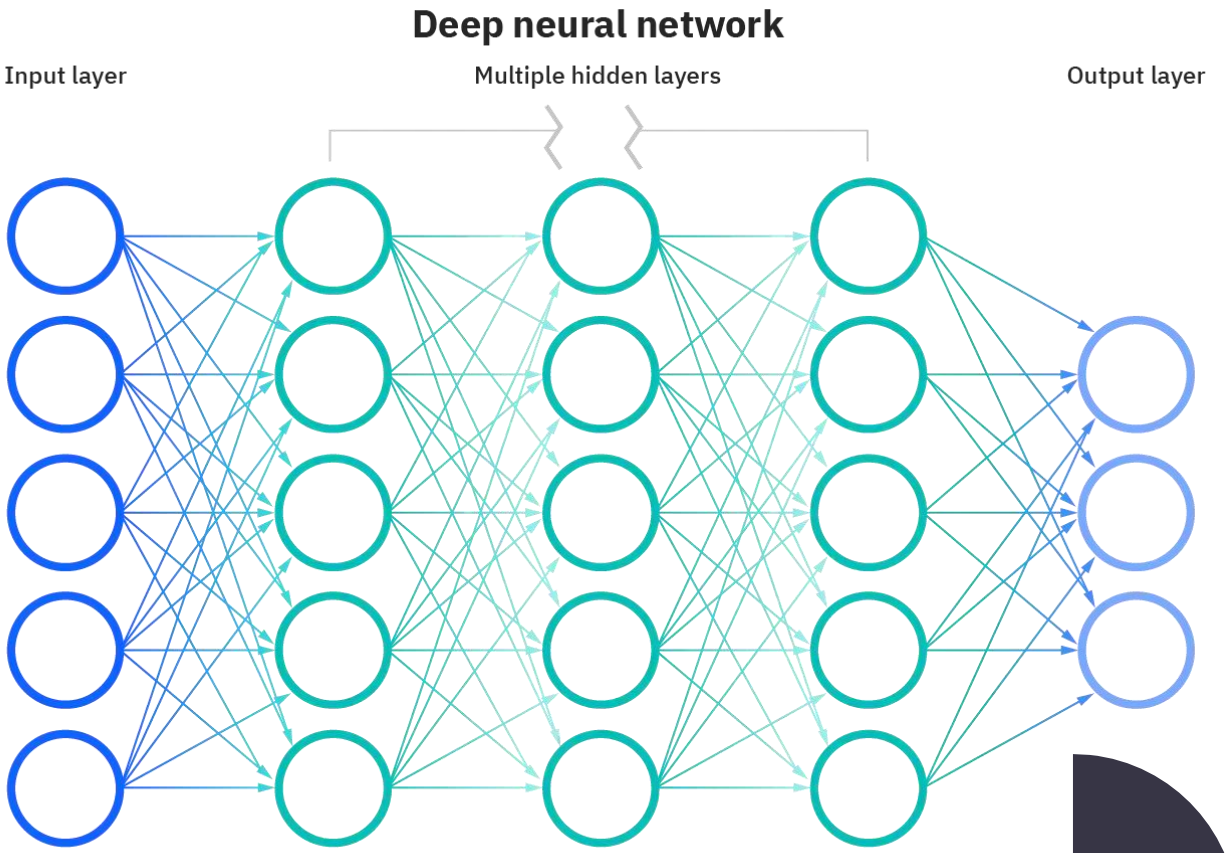


Strengths Of Black Box Models

Black box models shine in complex applications:

- Object recognition
- Speech translation
- Textual inferencing (NLI)
- Etc.

Connectionism And Layer-Stacking



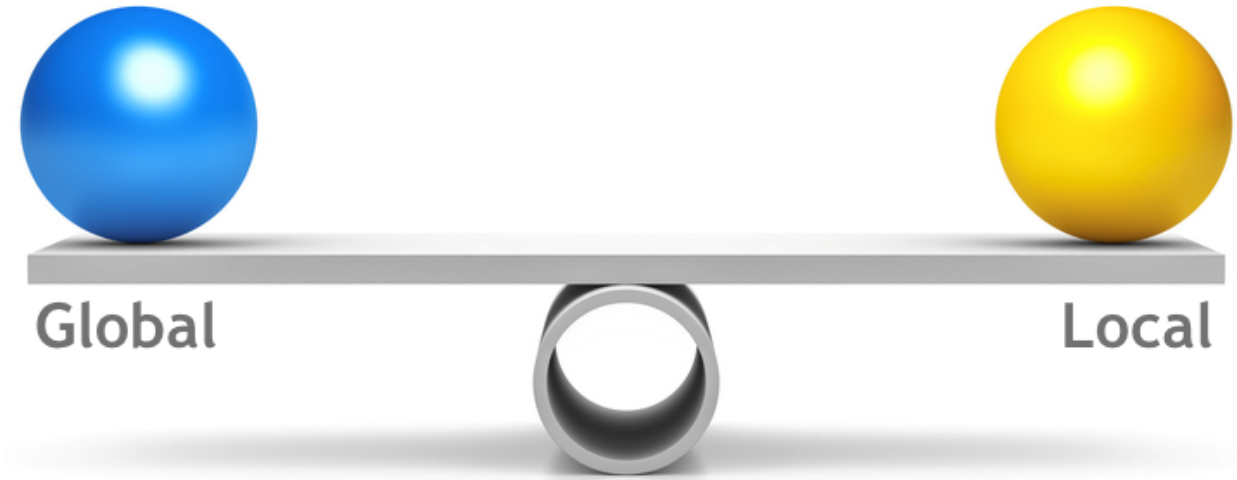
Issues With Layer-Stacking (Multi-Layered Models)

- Complexity through a chain of simpler computations
- Input features transformations
- Output could be unanticipated by end-users and developers alike

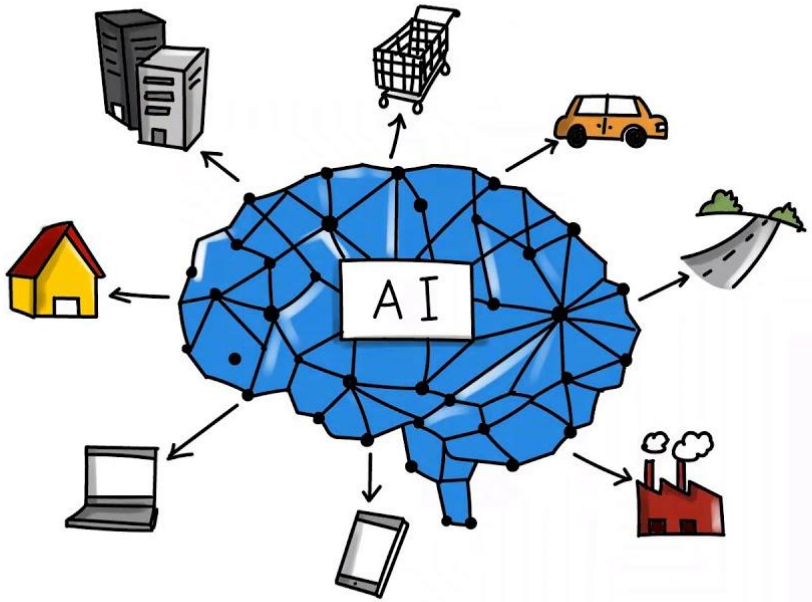
Fixing Black Box Opacity

- Local versus global
- Model-agnostic versus model-specific

Local Versus Global



Model-Agnostic Versus Model-Specific



Example: Shapley Values

- Coalition of factors
- Distributing fairly the outcome probability amongst the input features

Demo-3

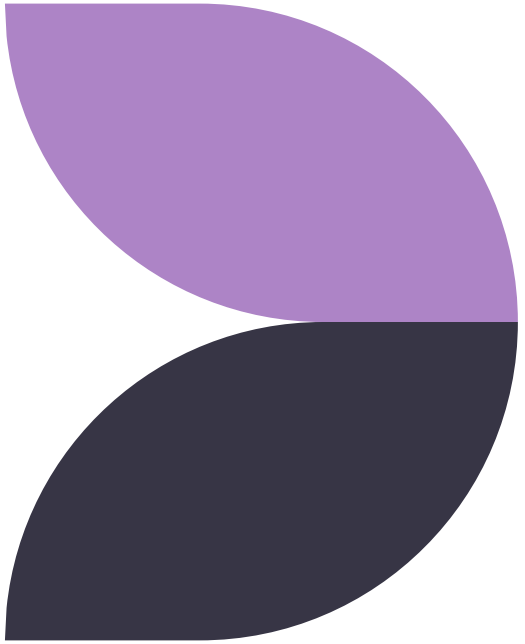
Microsoft Azure Notebooks

+

pyAgrum

PART V

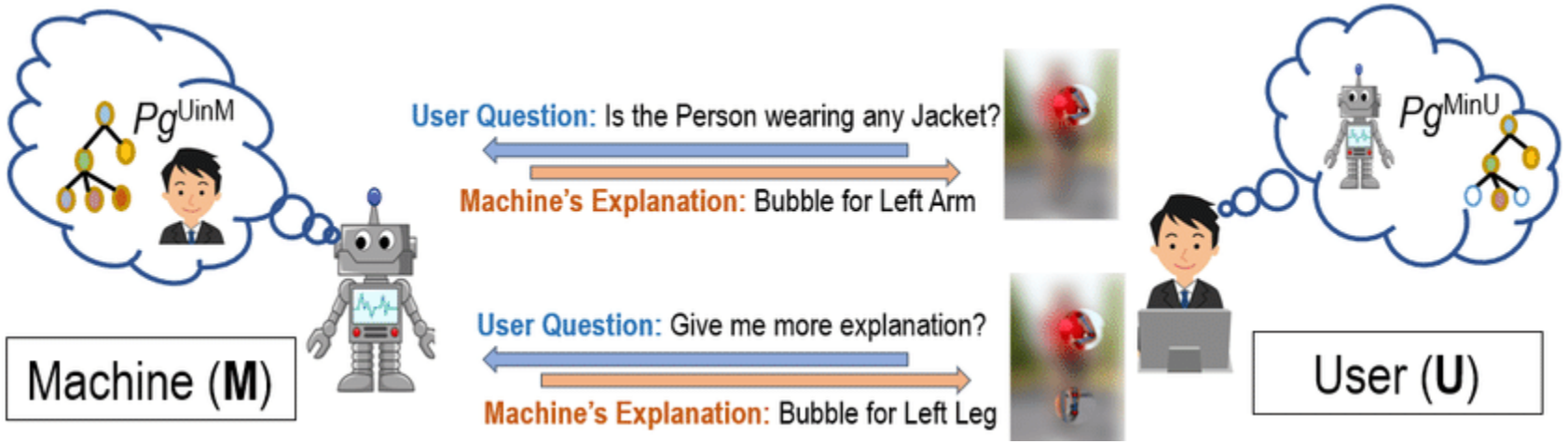
Considerations For The Future



Collaborative AI As An Integrative Pipeline

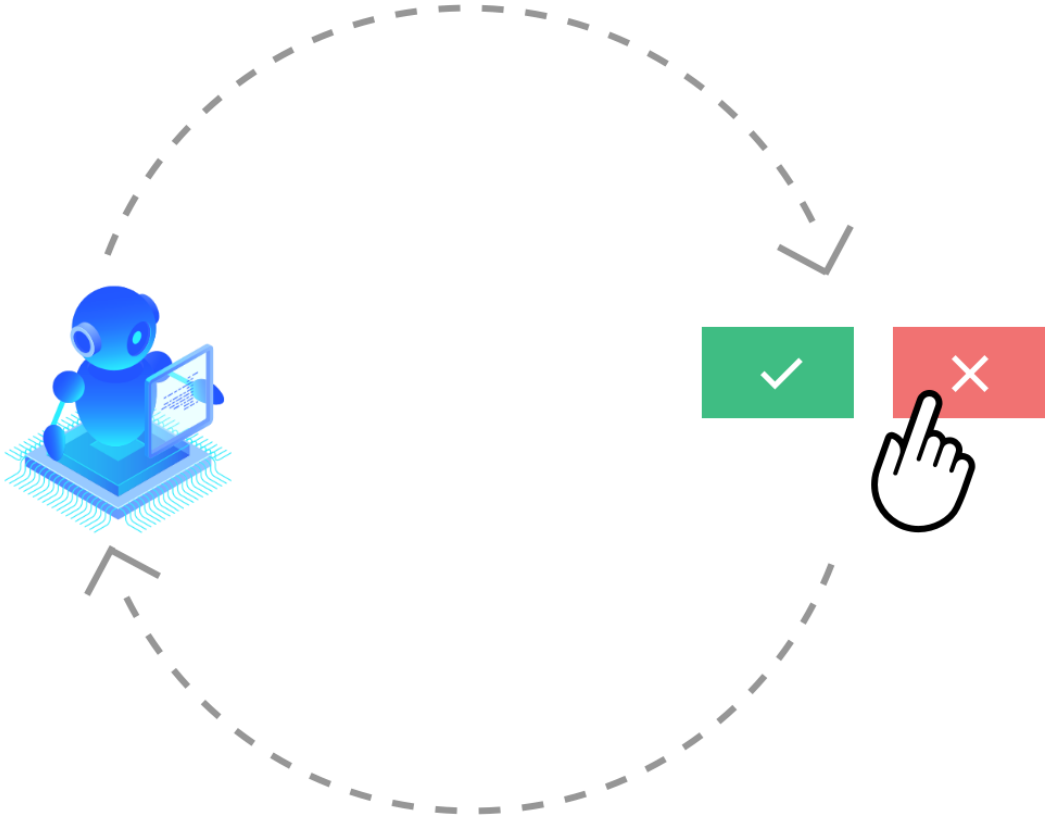
- Integrate not only pipelines of AI/ML models
- But also, keep human-in-the-loop
- Explanations + active learning

Context-Specific Explanations



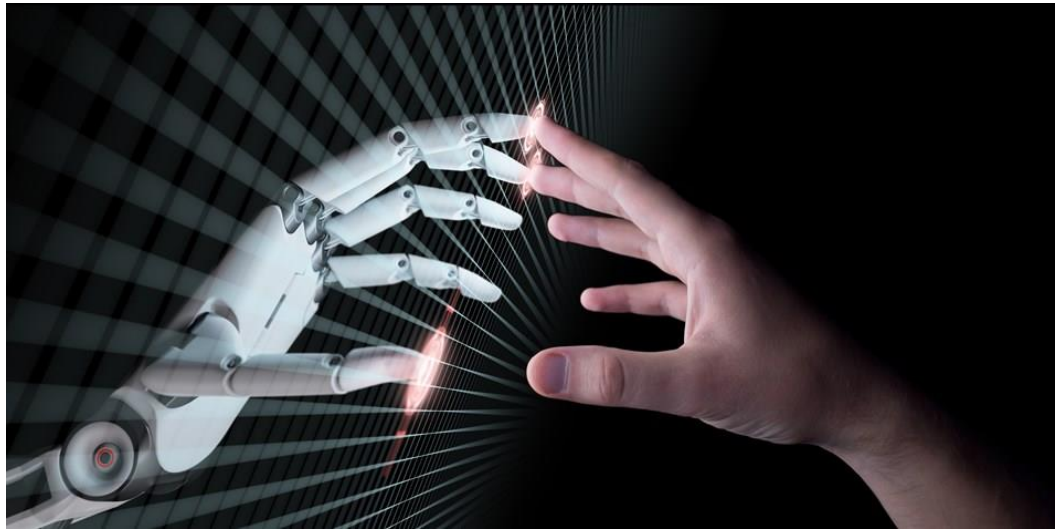
Akula et al.

Customized Explanations

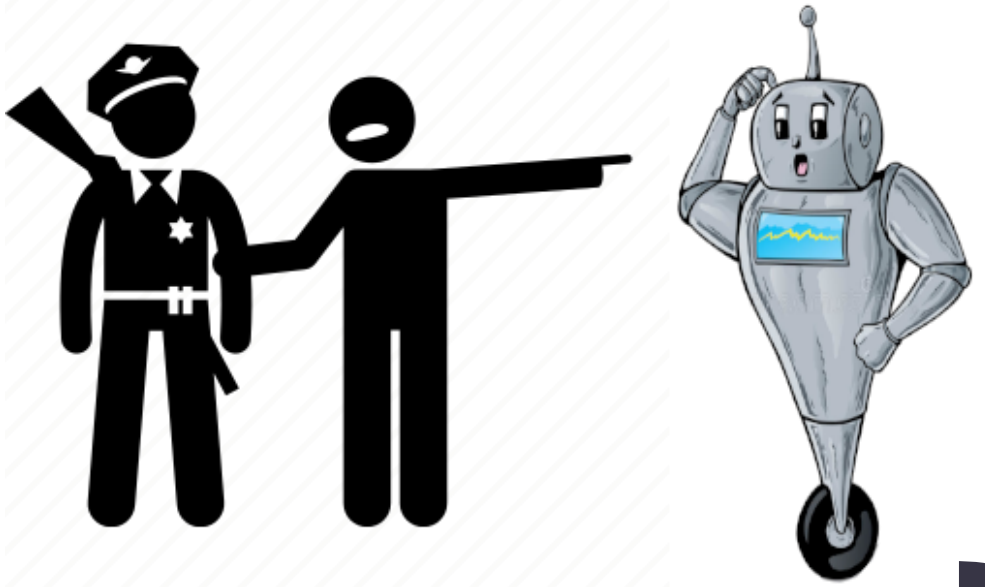


Improve Automatic Learning With A Human Touch

- AI learns automatically how to optimize an objective function based on data
- Self-supervision
- Human-out-of-the-loop?

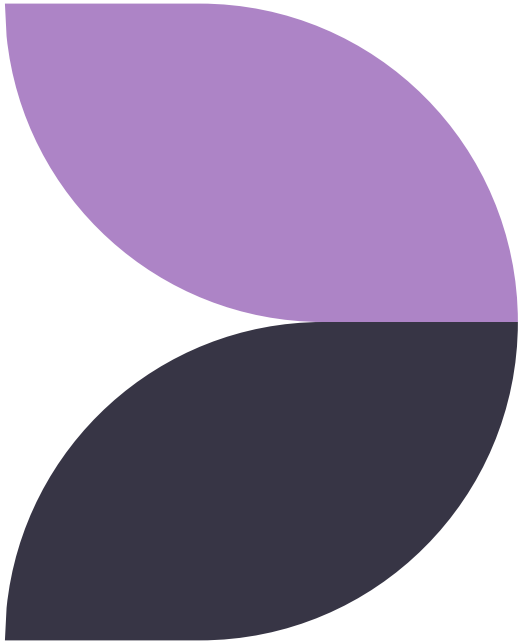


Automatic Learning And The Failure Of Intent. Who's To Blame?



Artificial + Human Intelligence

CONCLUSION



Recap

- Illustrative Example (Using PyAgrum + Microsoft Azure)
- Inference With Forward Probabilities
- The Case For Inverse Probabilities
- Black Boxes And Explainable Techniques
- Considerations For The Future

Takeaways

- Open-source rocks! *(aGrUM/pyAgrum)*
- Democratize AI beyond PhDs and into mainstream
- Azure high-performance computing (HPC) and others
- Collaborative AI: Explainable + Interactive
- Let’s build trustable AI!

Key References

Ducamp, G., Gonzales, C., & Wuillemin, P. H. (2020, February). aGrUM/pyAgrum: a toolbox to build models and algorithms for Probabilistic Graphical Models in Python. In International Conference on Probabilistic Graphical Models. PMLR.

Turi, K. N., Buchner, D. M., & Grigsby-Toussaint, D. S. (2017). Peer Reviewed: Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors. *Preventing Chronic Disease*, *14*.

Bathaee, Y. (2017). The artificial intelligence black box and the failure of intent and causation. Harv. JL & Tech., 31, 889.

Holzinger, A. (2018, August). From machine learning to explainable AI. In 2018 world symposium on digital intelligence for systems and machines (DISA) (pp. 55-66). IEEE.

Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Akula, A. R., Liu, C., Todorovic, S., Chai, J. Y., & Zhu, S. C. (2019, January). Explainable AI as Collaborative Task Solving. In CVPR Workshops (pp. 91-94).

Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic books.

Pearl, J. (2009). Causality. Cambridge university press.

Thank you

Ketemwabi Yves Shamavu, Ph.D.

yves@skyel.net

Link to recording: https://skyel.net/pages/blog/post/1648526906.html