Explainable Artificial Intelligence

Jorge Amaya
Research Expert
KU Leuven
AIDA Project

github.com/CmPA/AIDA-School

Who has already worked with ML/AI models?

Who has already published scientific results obtained using ML/AI models?

Would you agree to tie your salary to the accuracy of your ML/AI model?

A crisis in science

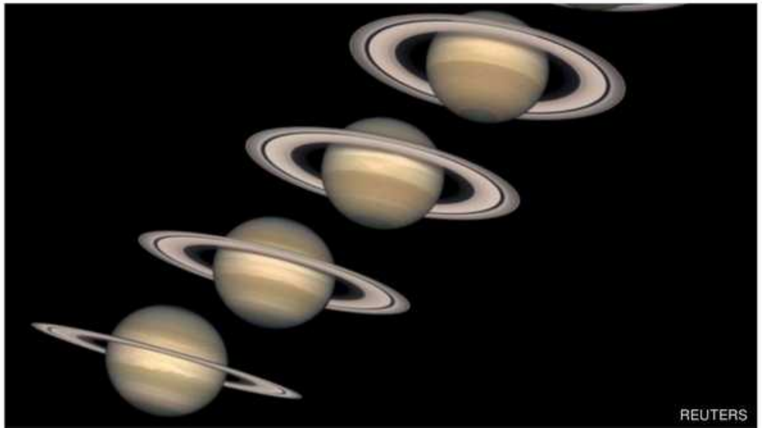
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?

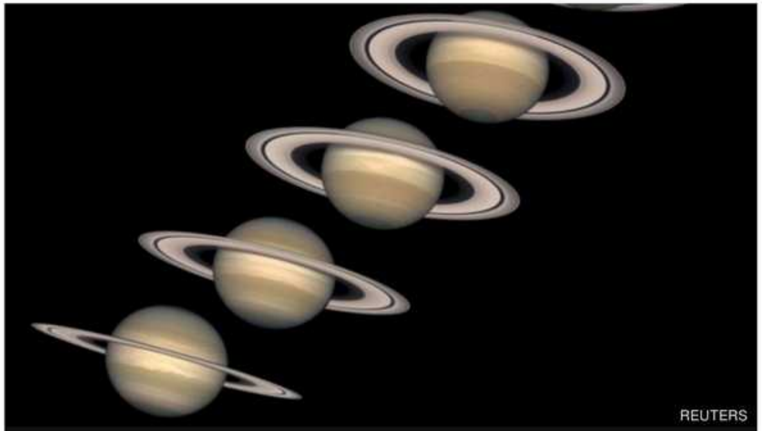
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

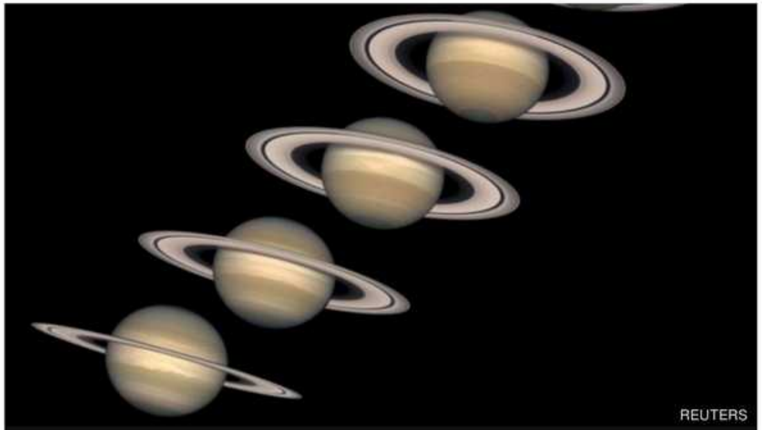
- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?

AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019 Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?
  - Exact parameters are not reported

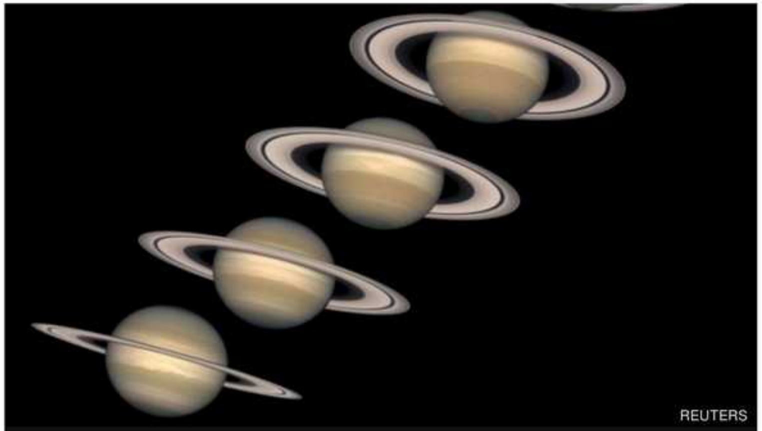
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?
  - Exact parameters are not reported
  - Data sets are not necessarily reported

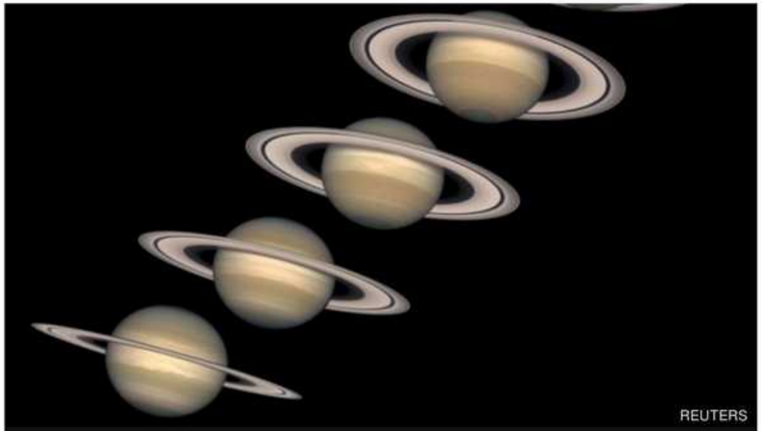
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?
  - Exact parameters are not reported
  - Data sets are not necessarily reported
  - Data pre-processing pipelines are not reported

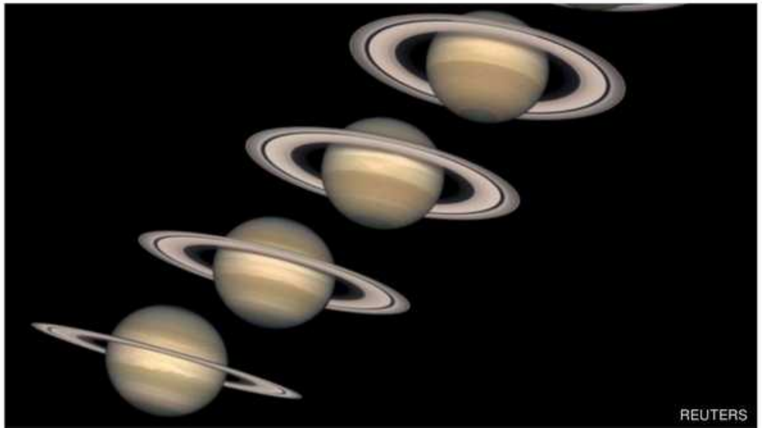
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?
  - Exact parameters are not reported
  - Data sets are not necessarily reported
  - Data pre-processing pipelines are not reported
  - Findings are not easily trusted by the community

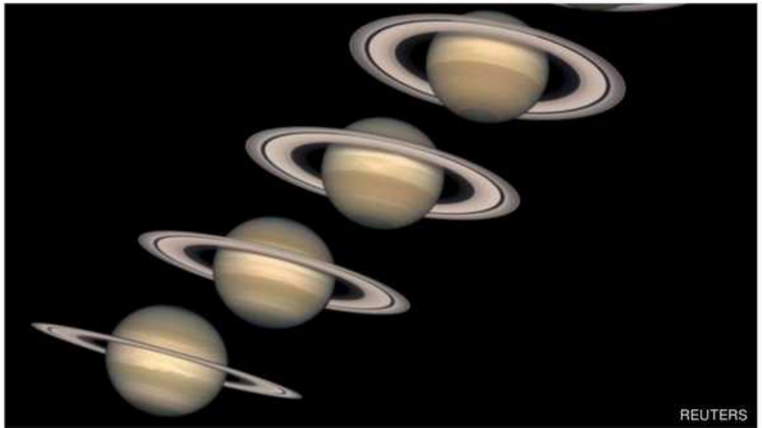
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting


REUTERS

Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?
  - Exact parameters are not reported
  - Data sets are not necessarily reported
  - Data pre-processing pipelines are not reported
  - Findings are not easily trusted by the community

- **Trust:**

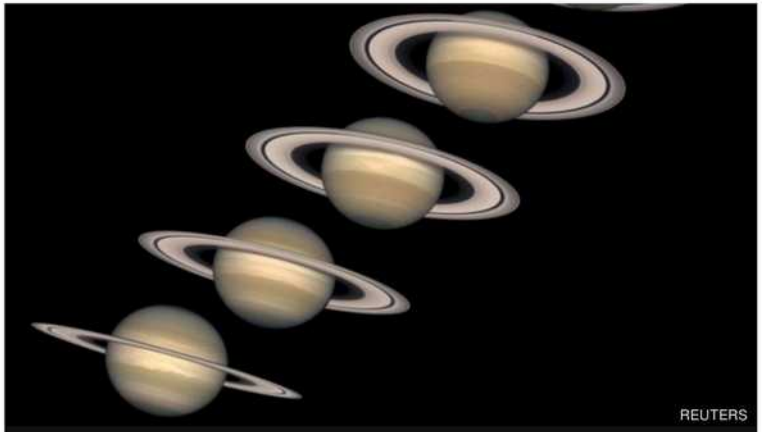
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?
  - Exact parameters are not reported
  - Data sets are not necessarily reported
  - Data pre-processing pipelines are not reported
  - Findings are not easily trusted by the community

- **Trust:**
  - Are you confident your model is accurate in all cases?

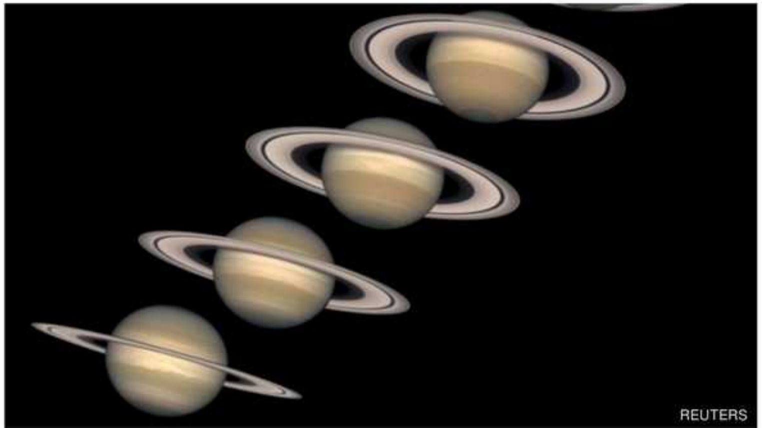
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?
  - Exact parameters are not reported
  - Data sets are not necessarily reported
  - Data pre-processing pipelines are not reported
  - Findings are not easily trusted by the community

- **Trust:**
  - Are you confident your model is accurate in all cases?
  - How does your model work internally?

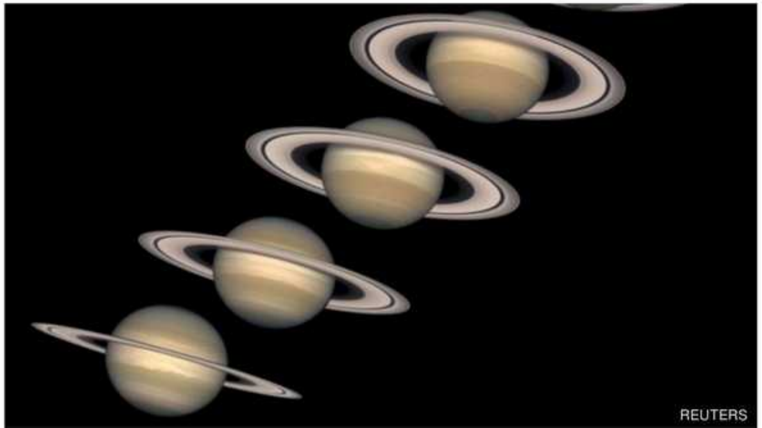
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?
  - Exact parameters are not reported
  - Data sets are not necessarily reported
  - Data pre-processing pipelines are not reported
  - Findings are not easily trusted by the community

- **Trust:**
  - Are you confident your model is accurate in all cases?
  - How does your model work internally?
  - Can you explain *why* an input gives an output?

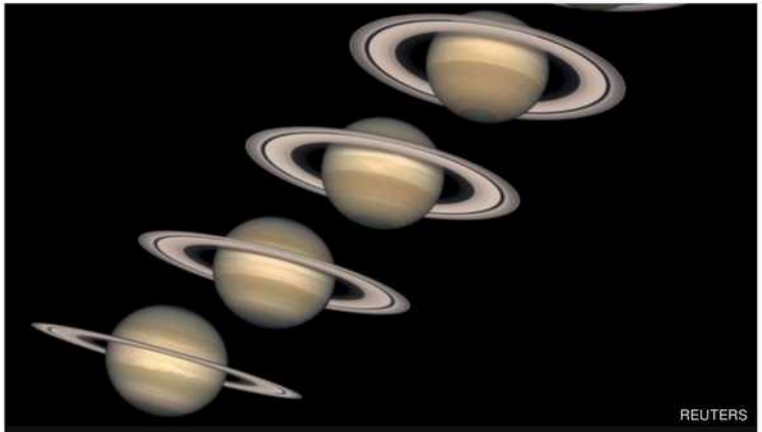
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

A crisis in science

- **Reproducibility:**
  - Patterns only in data sets and not in nature?
  - Similar studies produce non overlapping results?
  - Exact parameters are not reported
  - Data sets are not necessarily reported
  - Data pre-processing pipelines are not reported
  - Findings are not easily trusted by the community

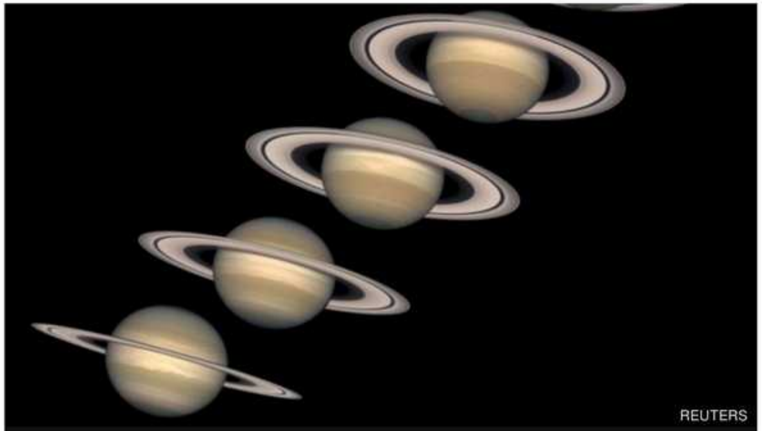
- **Trust:**
  - Are you confident your model is accurate in all cases?
  - How does your model work internally?
  - Can you explain *why* an input gives an output?
  - Can a small change in the model (hyper- parameters, data, training time) lead to different results?

AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019 Share

AAAS meeting



Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

ML data bias: horror stories

REUTERS

TECHNOLOGY NEWS OCTOBER 10, 2018 / 5:12 AM / A YEAR AGO

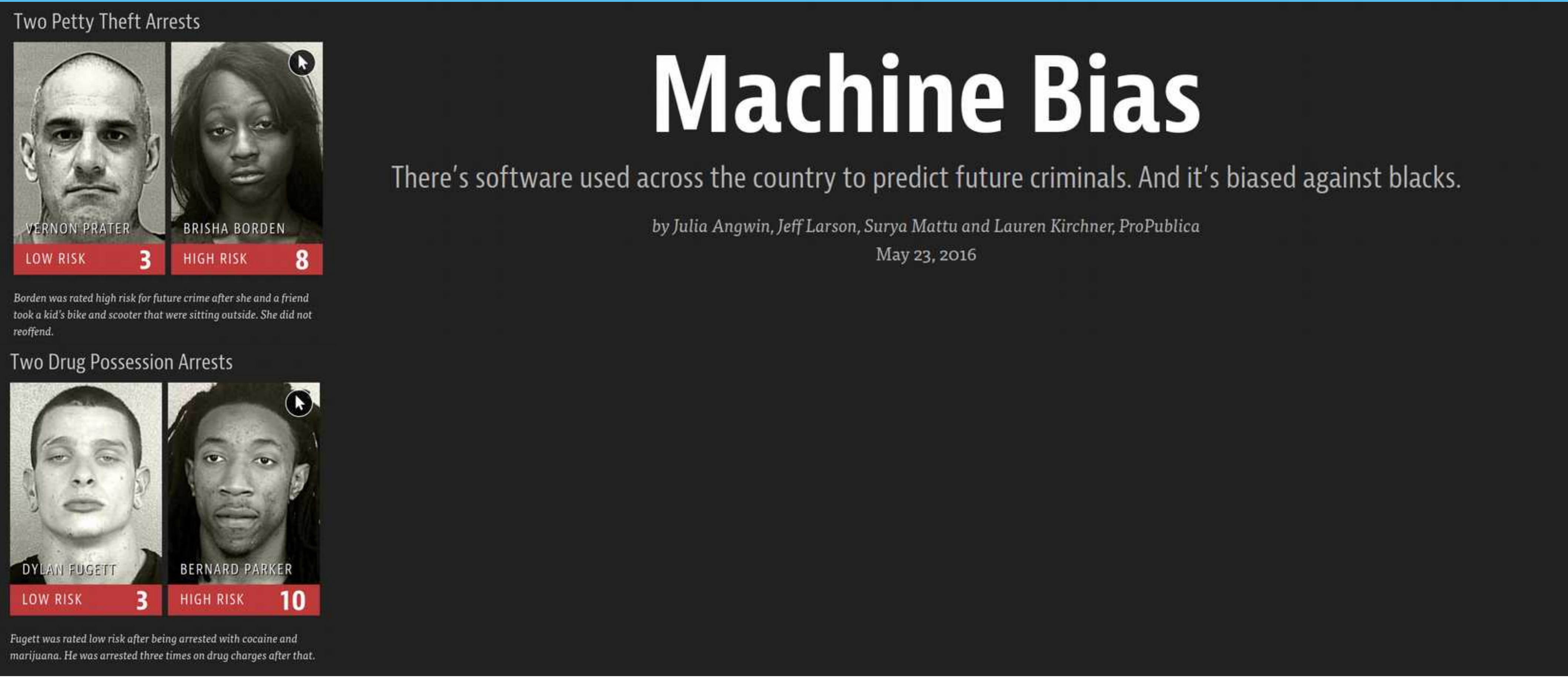
Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

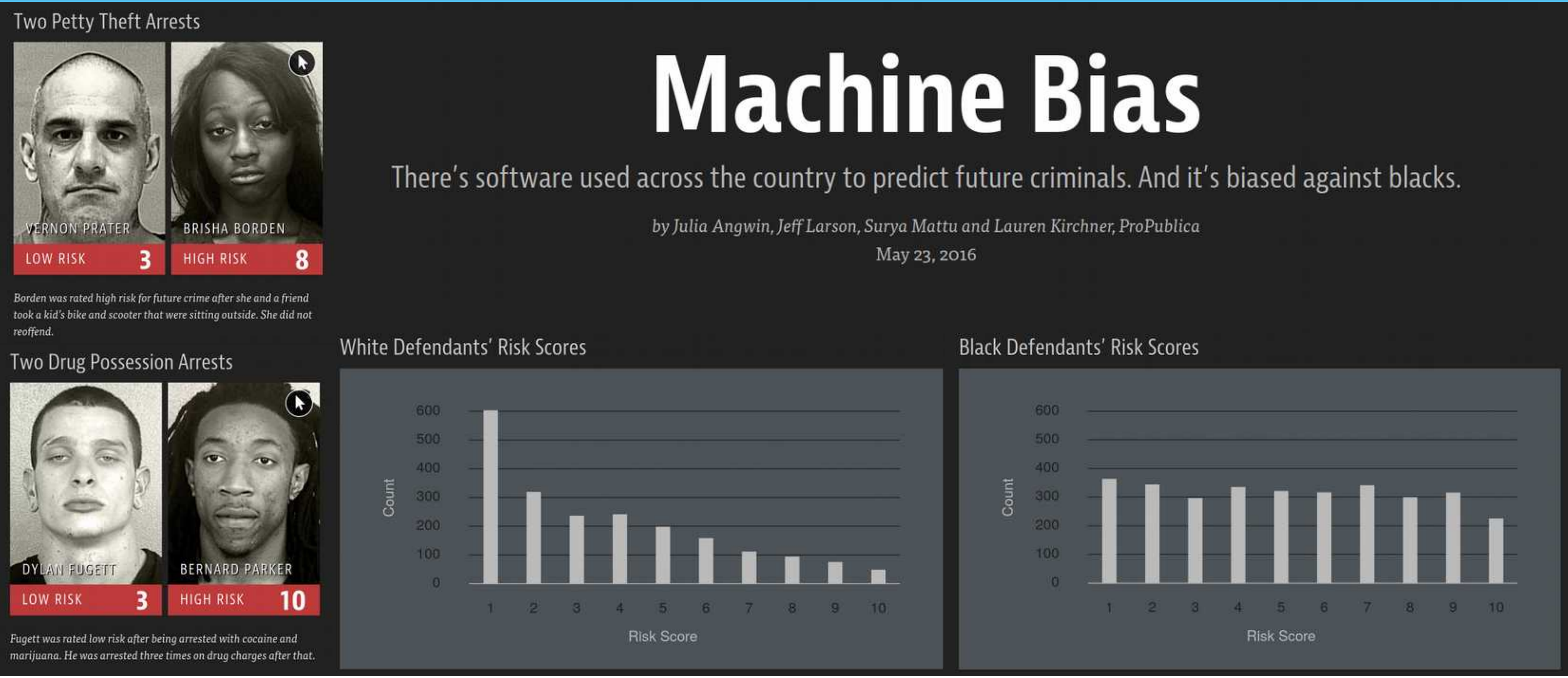
8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

ML data bias: horror stories

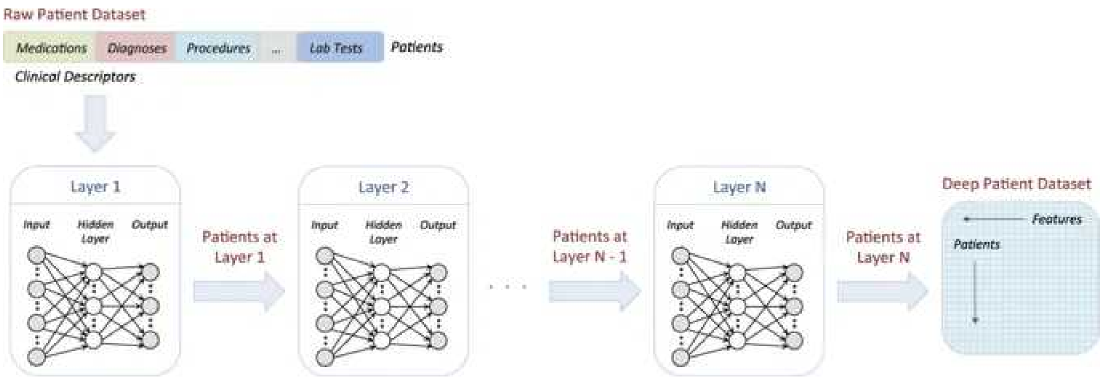


ML data bias: horror stories



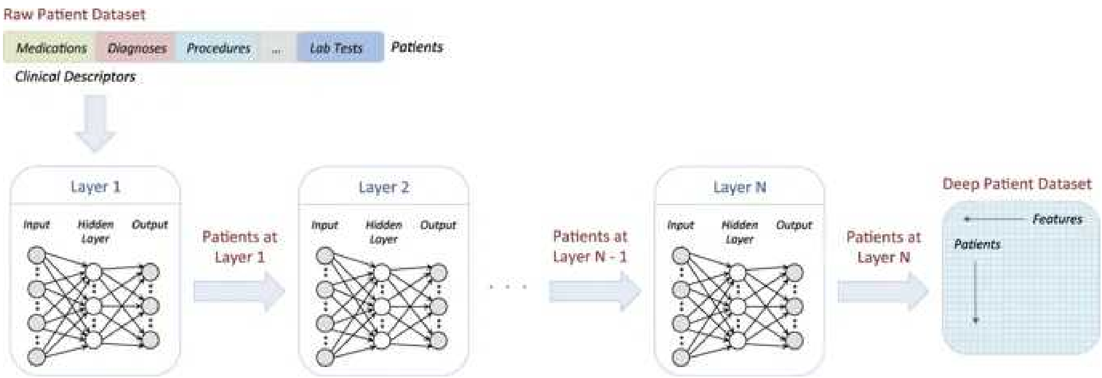
ML data bias: horror stories

Deep Patient (Mount Sinai Hospital in New York)



ML data bias: horror stories

Deep Patient (Mount Sinai Hospital in New York)

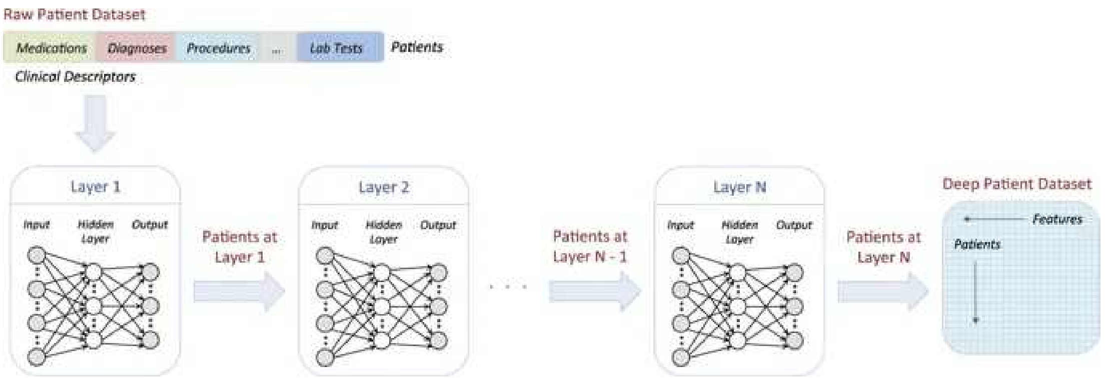


Joel Dudley, PhD

DIRECTOR, INSTITUTE FOR NEXT GENERATION HEALTHCARE
ASSOCIATE PROFESSOR | Genetics and Genomic Sciences
ASSOCIATE PROFESSOR | Population Health Science and Policy
ASSOCIATE PROFESSOR | Medicine

ML data bias: horror stories

Deep Patient (Mount Sinai Hospital in New York)



Artificial Intelligence / Machine Learning

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

by **Will Knight** Apr 11, 2017

Joel Dudley, PhD

DIRECTOR, INSTITUTE FOR NEXT GENERATION HEALTHCARE
ASSOCIATE PROFESSOR | Genetics and Genomic Sciences
ASSOCIATE PROFESSOR | Population Health Science and Policy
ASSOCIATE PROFESSOR | Medicine

"We can build these models, but we don't know how they work."

This story is part of our May/June 2017 issue

Facing the “science crisis” in ML

Facing the “science crisis” in ML

Reproducibility

How to reproduce the results of old experiments (from the literature)

Facing the “science crisis” in ML

Reproducibility

How to reproduce the results of old experiments (from the literature)

Trust

How to be confident in the results of my models and in the accuracy of my scientific discoveries

Reproducibility: the hell of ML

Reproducibility: the hell of ML

ML technique used

Reproducibility: the hell of ML

ML technique used

Model architecture

Reproducibility: the hell of ML

ML technique used

Model architecture

Hyper-parameter selection

Reproducibility: the hell of ML

ML technique used

Model architecture

Hyper-parameter selection

Data sources

Reproducibility: the hell of ML

ML technique used

Model architecture

Hyper-parameter selection

Data sources

Data pre-processing methods

Reproducibility: the hell of ML

ML technique used	Model architecture	Hyper-parameter selection
Data sources	Data pre-processing methods	Computing resources

Reproducibility: the hell of ML

- ML technique used
- Model architecture
- Hyper-parameter selection
- Data sources
- Data pre-processing methods
- Computing resources
- Software

Reproducibility: the hell of ML

ML technique used	Model architecture	Hyper-parameter selection
Data sources	Data pre-processing methods	Computing resources
Software	Evaluation metrics	

Reproducibility: the hell of ML

ML technique used	Model architecture	Hyper-parameter selection
Data sources	Data pre-processing methods	Computing resources
Software	Evaluation metrics	Budget

Reproducibility: tools

Reproducibility: tools

GitHub

Keep a repository of your code

Reproducibility: tools

GitHub

Keep a repository of your code

OSF

Keep a repository of your data

Reproducibility: tools

GitHub

Keep a repository of your code

OSF

Keep a repository of your data

jupyter

Keep a clear track of your procedures

Reproducibility: tools

GitHub — Keep a repository of your code

binder — Allow your code to be openly examined by the community

OSF — Keep a repository of your data

jupyter — Keep a clear track of your procedures

Reproducibility: tools

GitHub — Keep a repository of your code

binder — Allow your code to be openly examined by the community

OSF — Keep a repository of your data

docker — If it is impossible to share online, conteinerize

jupyter — Keep a clear track of your procedures

Reproducibility: tools

GitHub — Keep a repository of your code

binder — Allow your code to be openly examined by the community

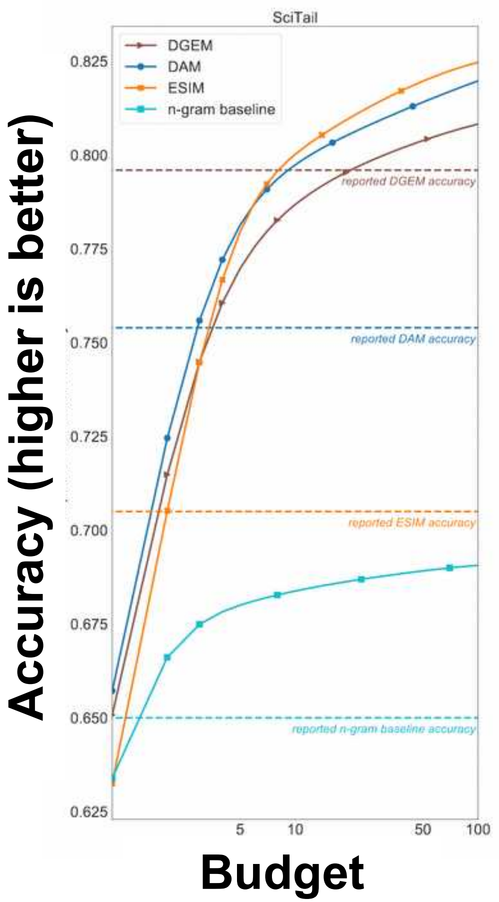
OSF — Keep a repository of your data

docker — If it is impossible to share online, conteinerize

jupyter — Keep a clear track of your procedures

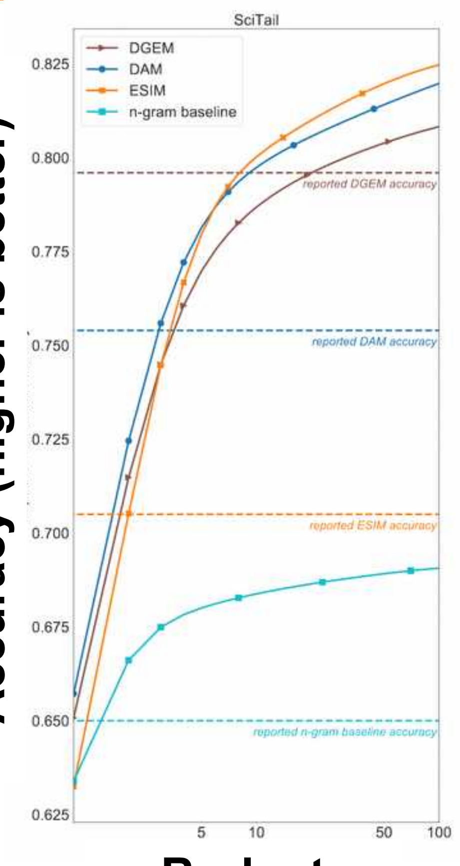
Even HPC/HPDA codes can be shared in containers

Reproducibility: budget



[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget



Accuracy (higher is better)

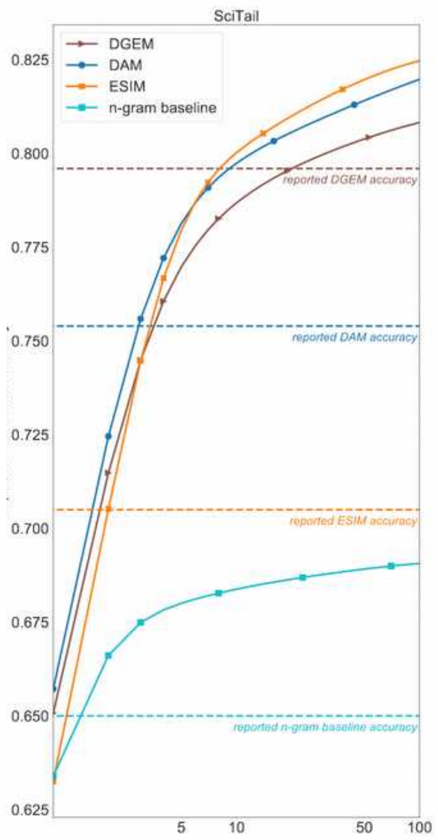
Budget

Budget:

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget

Accuracy (higher is better)



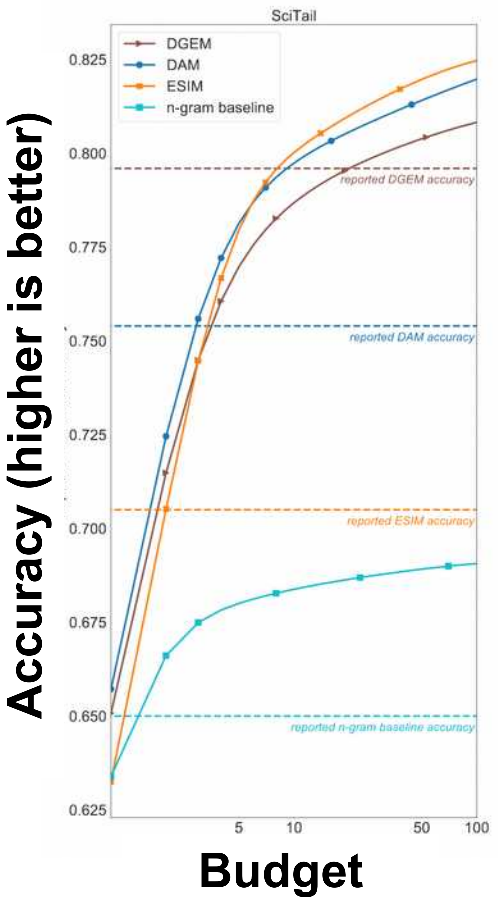
Budget

Budget:

- Training duration

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget

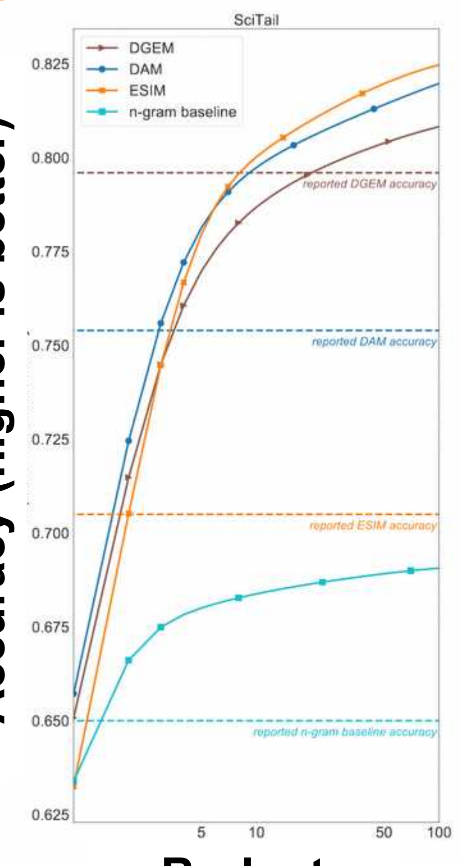


Budget:

- Training duration
- Persons month spent

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget



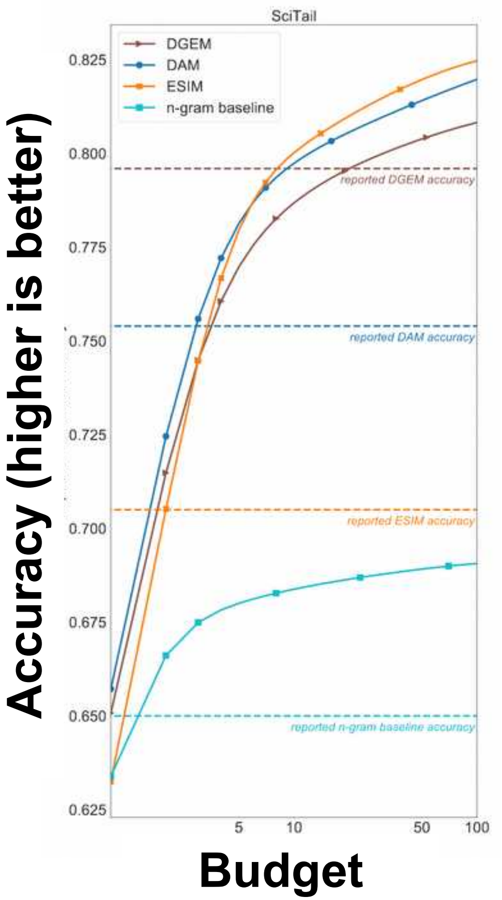
Accuracy (higher is better)

Budget

Budget:
- Training duration
- Persons month spent
- Hyper-parameter assignments

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget



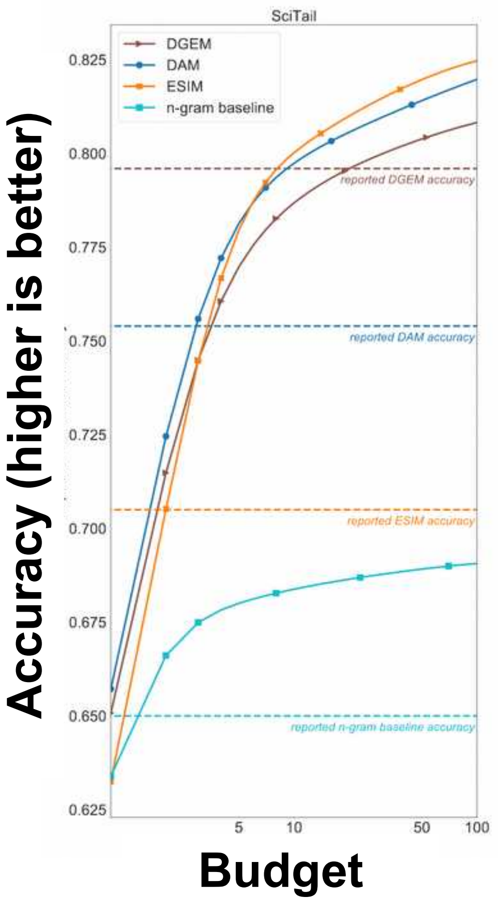
Budget:

- Training duration
- Persons month spent
- Hyper-parameter assignments

Hyper-parameter search methods:

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget



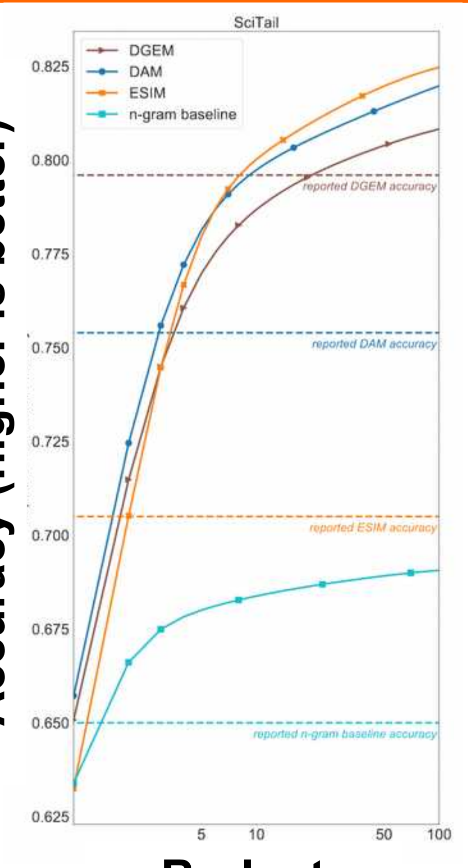
Budget:
- Training duration
- Persons month spent
- Hyper-parameter assignments

Hyper-parameter search methods:
- Manual

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget

Accuracy (higher is better)



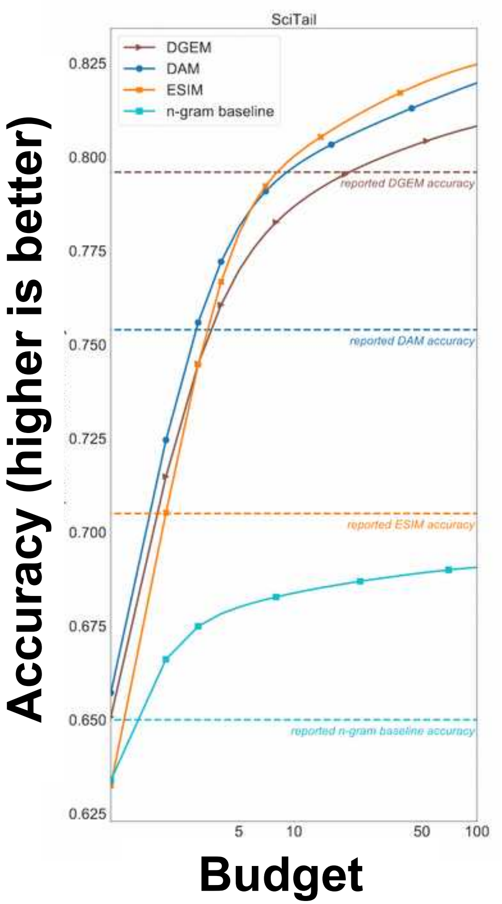
Budget

Budget:
- Training duration
- Persons month spent
- Hyper-parameter assignments

Hyper-parameter search methods:
- Manual
- Grid search

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget



Budget:

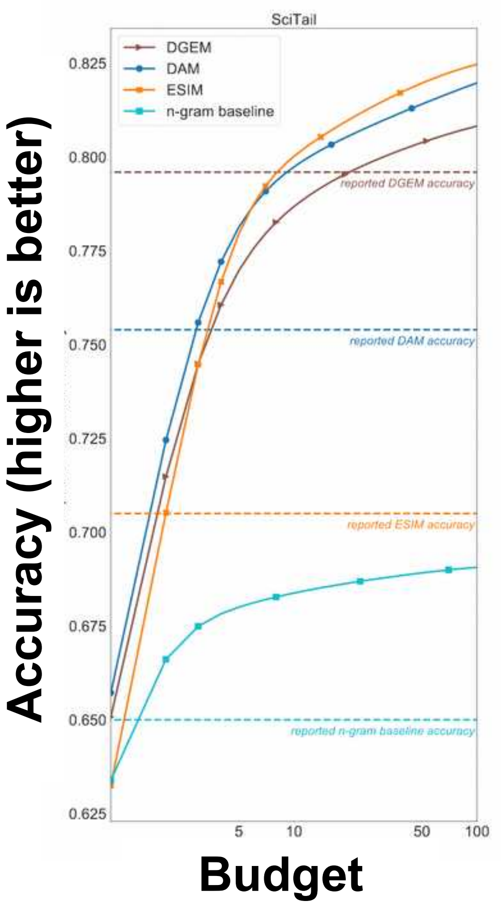
- Training duration
- Persons month spent
- Hyper-parameter assignments

Hyper-parameter search methods:

- Manual
- Grid search
- Uniform sampling

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget

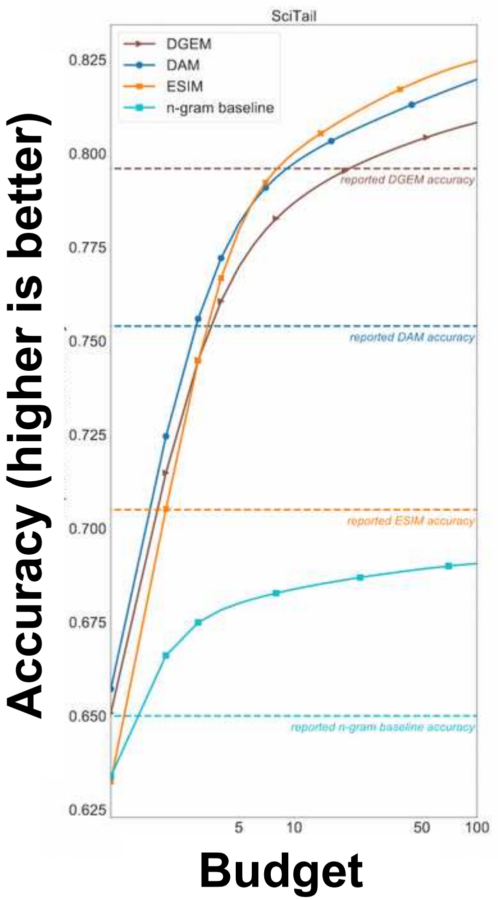


Budget:
- Training duration
- Persons month spent
- Hyper-parameter assignments

Hyper-parameter search methods:
- Manual
- Grid search
- Uniform sampling
- Bayesian selection

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: budget



Budget:

- Training duration
- Persons month spent
- Hyper-parameter assignments

Hyper-parameter search methods:

- Manual
- Grid search
- Uniform sampling
- Bayesian selection
- Evolutionary methods

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- ✔ Description of computing infrastructure

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- ✓ Description of computing infrastructure
- ✓ Average runtime for each approach

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- ✓ Description of computing infrastructure
- ✓ Average runtime for each approach
- ✓ Details of train/validation/test splits

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- ✔ Description of computing infrastructure
- ✔ Average runtime for each approach
- ✔ Details of train/validation/test splits
- ✔ Corresponding validation performance for each reported test result

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- ✓ Description of computing infrastructure
- ✓ Average runtime for each approach
- ✓ Details of train/validation/test splits
- ✓ Corresponding validation performance for each reported test result
- ✓ A link to implemented code

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- Description of computing infrastructure
- Average runtime for each approach
- Details of train/validation/test splits
- Corresponding validation performance for each reported test result
- A link to implemented code
- A link to the data used

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- ✔ Description of computing infrastructure
- ✔ Average runtime for each approach
- ✔ Details of train/validation/test splits
- ✔ Corresponding validation performance for each reported test result
- ✔ A link to implemented code
- ✔ A link to the data used

For experiments with hyperparameter search:

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- Description of computing infrastructure
- Average runtime for each approach
- Details of train/validation/test splits
- Corresponding validation performance for each reported test result
- A link to implemented code
- A link to the data used

For experiments with hyperparameter search:

- Bounds for each hyperparameter

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- Description of computing infrastructure
- Average runtime for each approach
- Details of train/validation/test splits
- Corresponding validation performance for each reported test result
- A link to implemented code
- A link to the data used

For experiments with hyperparameter search:

- Bounds for each hyperparameter
- Hyperparameter configurations for best performing models

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- Description of computing infrastructure
- Average runtime for each approach
- Details of train/validation/test splits
- Corresponding validation performance for each reported test result
- A link to implemented code
- A link to the data used

For experiments with hyperparameter search:

- Bounds for each hyperparameter
- Hyperparameter configurations for best performing models
- Number of hyperparameter search trials

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- Description of computing infrastructure
- Average runtime for each approach
- Details of train/validation/test splits
- Corresponding validation performance for each reported test result
- A link to implemented code
- A link to the data used

For experiments with hyperparameter search:

- Bounds for each hyperparameter
- Hyperparameter configurations for best performing models
- Number of hyperparameter search trials
- The method of choosing hyperparameter values

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- Description of computing infrastructure
- Average runtime for each approach
- Details of train/validation/test splits
- Corresponding validation performance for each reported test result
- A link to implemented code
- A link to the data used

For experiments with hyperparameter search:

- Bounds for each hyperparameter
- Hyperparameter configurations for best performing models
- Number of hyperparameter search trials
- The method of choosing hyperparameter values
- The criterion used to select among them

[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Reproducibility: minimal checklist

For all reported experimental results

- Description of computing infrastructure
- Average runtime for each approach
- Details of train/validation/test splits
- Corresponding validation performance for each reported test result
- A link to implemented code
- A link to the data used

For experiments with hyperparameter search:

- Bounds for each hyperparameter
- Hyperparameter configurations for best performing models
- Number of hyperparameter search trials
- The method of choosing hyperparameter values
- The criterion used to select among them
- Expected validation performance, or another measure of the mean and variance as a function of the number of hyperparameter trials.

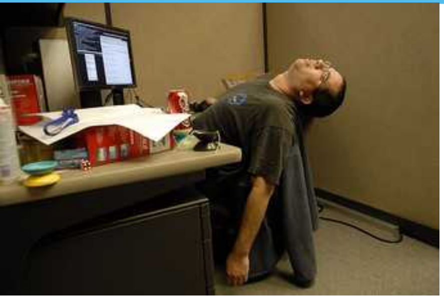
[1] Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Give enough details so you and others can reproduce your exact same results with the same budget

Reproducibility is necessary but not sufficient to explain the results

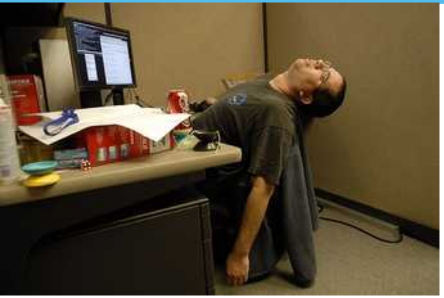
Explainable AI (XAI)

Explainable AI (XAI)



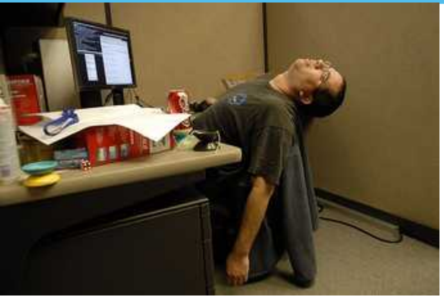
A fancy complex ML model
Clearly reproducible and traceable
Clean datasets

Explainable AI (XAI)



A fancy complex ML model
Clearly reproducible and traceable
Clean datasets

Explainable AI (XAI)



A fancy complex ML model
Clearly reproducible and traceable
Clean datasets

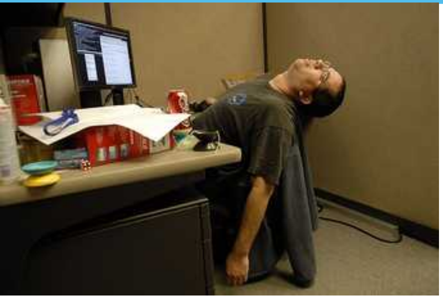
Cool new scientific discovery
Write a paper
Present in a conference

Explainable AI (XAI)

A fancy complex ML model
Clearly reproducible and traceable
Clean datasets

Cool new scientific discovery
Write a paper
Present in a conference

Explainable AI (XAI)

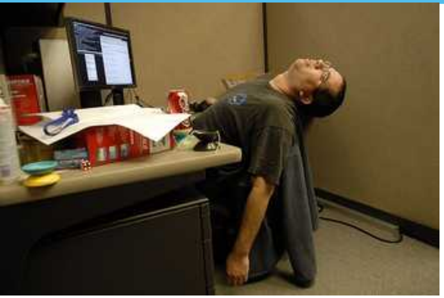


A fancy complex ML model
Clearly reproducible and traceable
Clean datasets

Cool new scientific discovery
Write a paper
Present in a conference

“How do you know what you did is correct?”
“Why should I trust your results?”

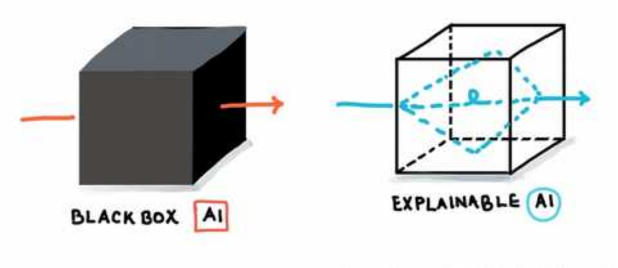
Explainable AI (XAI)



A fancy complex ML model
Clearly reproducible and traceable
Clean datasets

Cool new scientific discovery
Write a paper
Present in a conference

“How do you know what you did is correct?”
“Why should I trust your results?”



Your model might be behaving like a black box: you can not explain why it gives these results, you do not know its internal logic.

What is an explanation?

[1] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

What is an explanation?

- Explanations refer to causes [1]

[1] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

What is an explanation?

- Explanations refer to causes [1]
- An explanation is an assignment of causal responsibility

[1] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

What is an explanation?

- Explanations refer to causes [1]
- An explanation is an assignment of causal responsibility
- Answer questions:

[1] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

What is an explanation?

- Explanations refer to causes [1]
- An explanation is an assignment of causal responsibility
- Answer questions:

Question	Reasoning	Description
What?	Associative	Reason about which unobserved events could have occurred given the observed events
How?	Interventionist	Simulate a change in the situation to see if the event still happens
Why?	Counterfactual	Simulating alternative causes to see whether the event still happens

Table 3: Classes of Explanatory Question and the Reasoning Required to Answer

[1] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Explainability becomes an epistemological problem: this will not be a philosophy of science lecture!

Interpretability is the degree to which a human can understand the cause of a decision

[1] Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv Preprint arXiv:1706.07269. (2017)

Interpretability is the degree to which a human can consistently predict the model's result

[1] Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016)

Explanation in AI

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems

Some properties of an explainable model would be:

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems

Some properties of an explainable model would be:

- It is accurate

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems

Some properties of an explainable model would be:

- It is accurate
- We can trust it

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems

Some properties of an explainable model would be:

- It is accurate
- We can trust it
- We can confidently rely on its results

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems

Some properties of an explainable model would be:

- It is accurate
- We can trust it
- We can confidently rely on its results
- *It is safe*

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems

Some properties of an explainable model would be:

- It is accurate
- We can trust it
- We can confidently rely on its results
- *It is safe*
- *It is ethical*

Explanation in AI

Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems

Some properties of an explainable model would be:

- It is accurate
- We can trust it
- We can confidently rely on its results
- *It is safe*
- *It is ethical*
- *It is fair*

Disclaimer

Disclaimer

- This is a new field: the definitions, context, and tools are rapidly changing!

Disclaimer

- This is a new field: the definitions, context, and tools are rapidly changing!
- This is not an exhaustive survey, it is an introduction to the subject

Disclaimer

- This is a new field: the definitions, context, and tools are rapidly changing!
- This is not an exhaustive survey, it is an introduction to the subject
- Some keywords and definitions might change from author to author

Guidelines

Guidelines

Ethics guidelines for trustworthy AI

Guidelines

Ethics guidelines for trustworthy AI

Lawful

Ethical

Robust

Guidelines

Ethics guidelines for trustworthy AI

Lawful

Ethical

Robust

Human agency and oversight

Technical Robustness and safety

Privacy and data governance

Transparency

Diversity, non-discrimination and fairness

Societal and environmental well-being

Accountability

Explainable vs Interpretable AI

- Is it the same?

Explainable vs Interpretable AI

- Is it the same?

explanation
/ɛkspləˈneɪʃ(ə)n/

noun

a statement or account that makes something clear.
"the birth rate is central to any explanation of population trends"

Similar: clarification · simplification · description · report · version

- a reason or justification given for an action or belief.
  "Freud tried to make sex the explanation for everything"

  Similar: account · reason · justification · excuse · alibi · apologia

interpretation
/ɪntəːprɪˈteɪʃ(ə)n/

noun

the action of explaining the meaning of something.
"the interpretation of data"

Similar: explanation · elucidation · expounding · exposition · explication

- an explanation or way of explaining.
  plural noun: interpretations
  "this action is open to a number of interpretations"

  Similar: meaning · understanding · construal · connotation · reading

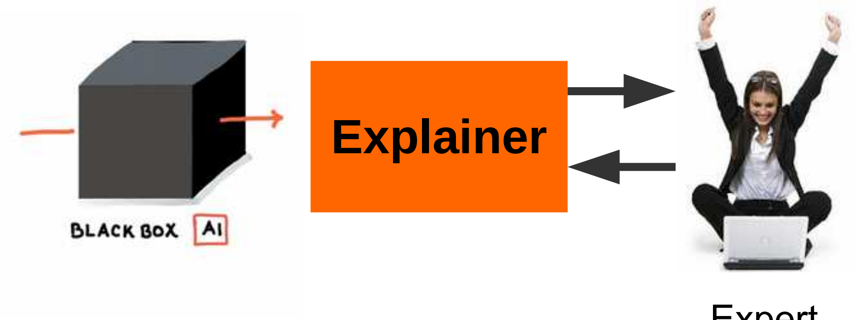
Explainable vs Interpretable AI

Explainable vs Interpretable AI

- Almost all the time used as synonyms
- **Explainable**: solutions can be understood by human experts. Black box
- **Interpretable**: solutions understood by anyone. Transparent models.

Explainable vs Interpretable AI

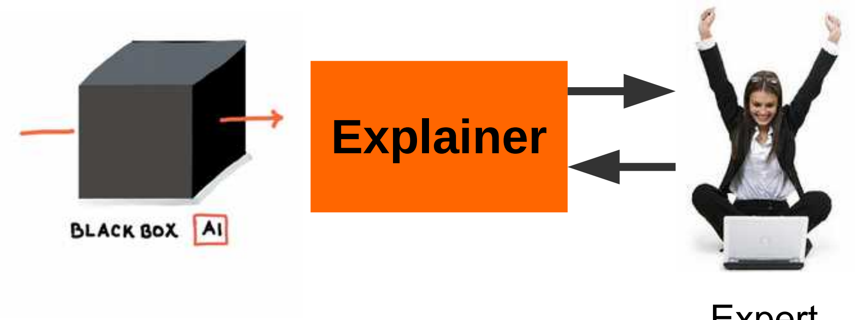
- Almost all the time used as synonyms
- **Explainable**: solutions can be understood by human experts. Black box
- **Interpretable**: solutions understood by anyone. Transparent models.



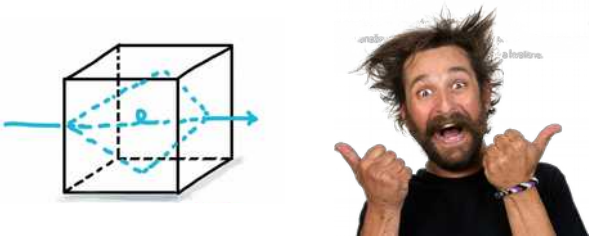
Expert

Explainable vs Interpretable AI

- Almost all the time used as synonyms
- **Explainable**: solutions can be understood by human experts. Black box
- **Interpretable**: solutions understood by anyone. Transparent models.



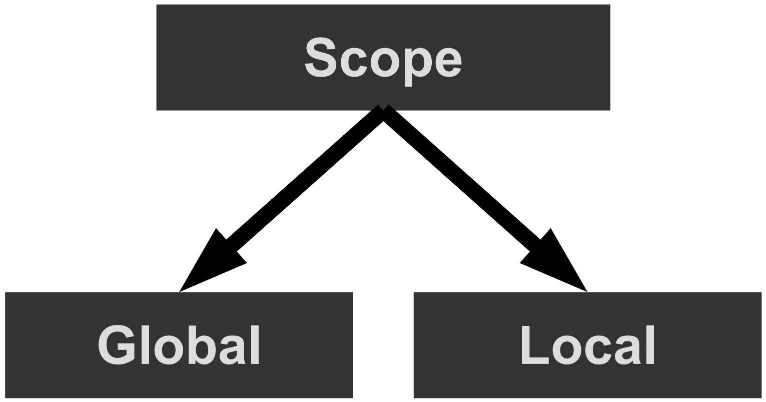
Expert



Anyone

Approaches to explain

Approaches to explain

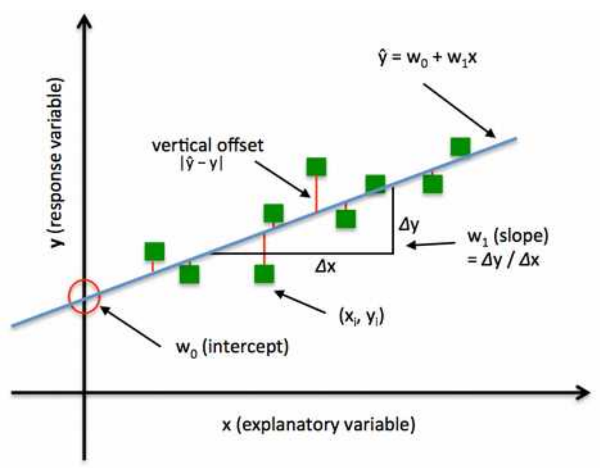


Approaches to explain

Interpretable models

Linear regression

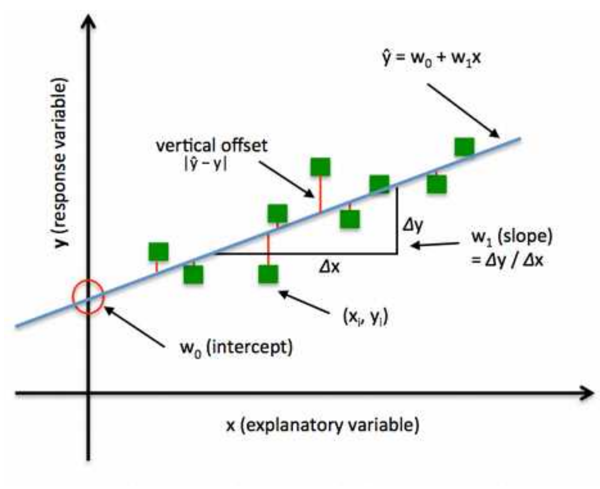
Global | Local | **Specific**



Interpretable models

Linear regression

Global | Local | **Specific**



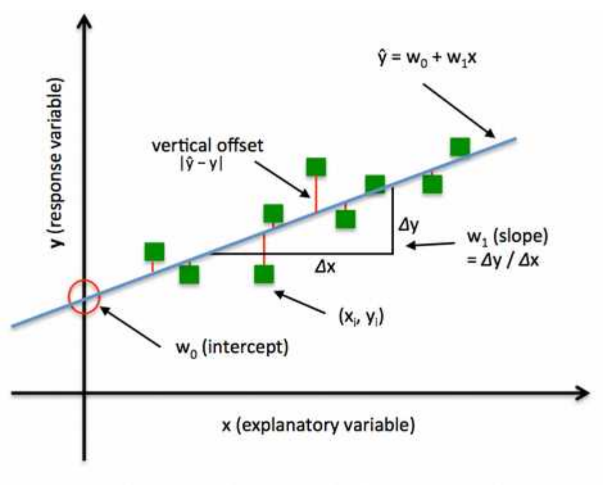
A linear model looks like this:

$$y = w_0 + w_1 * x_1 + w_2 * x_2$$

Interpretable models

Linear regression

Global | Local | **Specific**



A linear model looks like this:

$$y = w_0 + w_1 * x_1 + w_2 * x_2$$

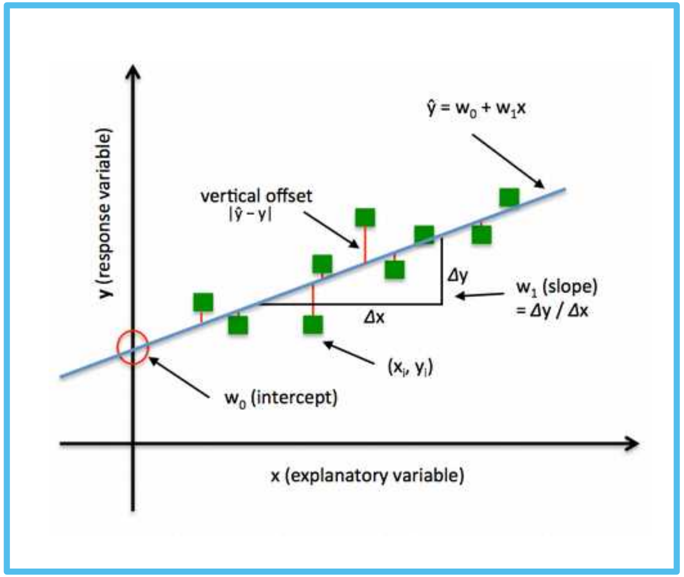
Global explanation:

w_0 , w_1, w_2 Are the level of importance of each one of the features. They give a clear connection between ***x*** and ***y***

Interpretable models

Linear regression

Global | Local | **Specific**



A linear model looks like this:

$$y = w_0 + w_1 * x_1 + w_2 * x_2$$

Global explanation:

w_0 , w_1, w_2 — Are the level of importance of each one of the features. They give a clear connection between ***x*** and ***y***

Local explanation:

$w_1 * x_1$, $w_2 * x_2$ — Give details about why the value of ***y*** was obtained

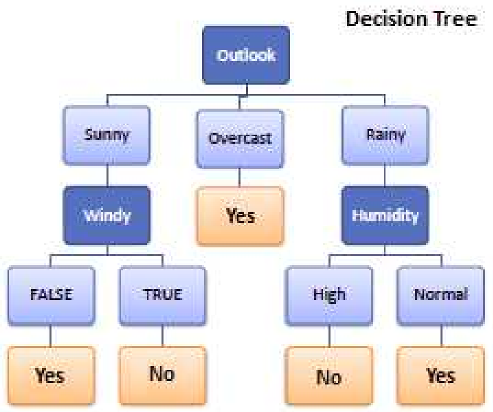
Interpretable models

Decision trees

Local

Specific

Golf today?



Interpretable models

Decision trees

Local | Specific

In this model, I can insert a value of ***x***, and I can transparently follow the logic of the model down to the leaf containing ***y***

Interpretable models

Decision trees

Local

Specific

In this model, I can insert a value of ***x***, and I can transparently follow the logic of the model down to the leaf containing ***y***

x = (Sunny, Windy)

If Sunny → If Windy → No

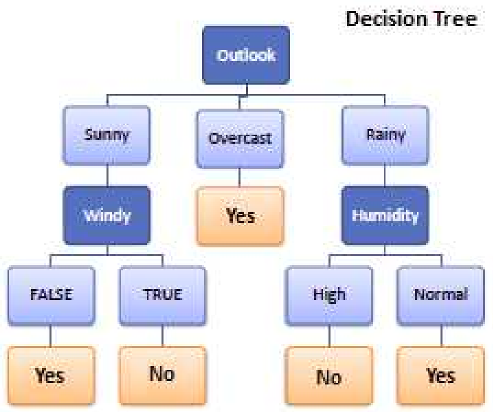
y = (No)

Interpretable models

Decision trees

Local | Specific

Golf today?



In this model, I can insert a value of ***x***, and I can transparently follow the logic of the model down to the leaf containing ***y***

x = (Sunny, Windy)
If Sunny → If Windy → No
y = (No)

Why aren't you playing golf today?
Explanation: Because it is sunny and windy.

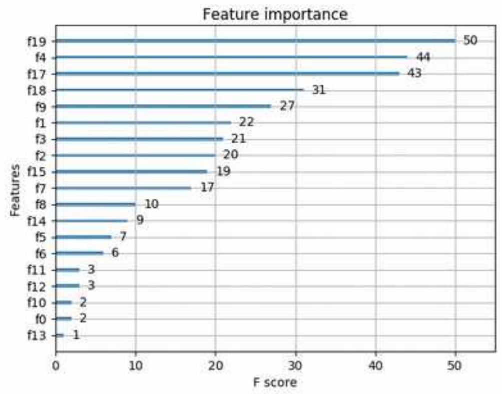
Interpretable models

Feature importance

Global

Specific

Built from decision trees

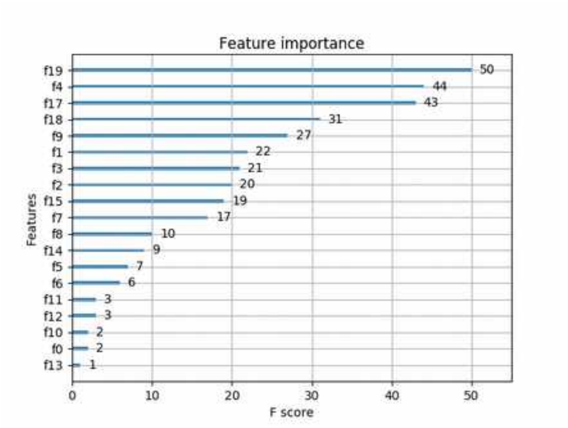


Interpretable models

Feature importance

Global | Specific

Built from decision trees



In decision tree models we can also see how all the data affects the output.

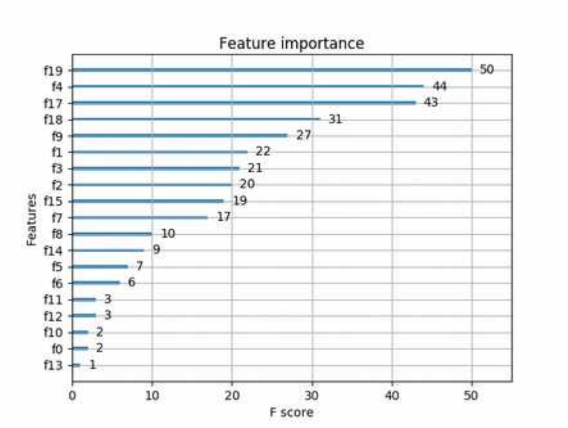
Interpretable models

Feature importance

Global

Specific

Built from decision trees



In decision tree models we can also see how all the data affects the output.

How many times a particular feature has been used to make a decision in the tree.

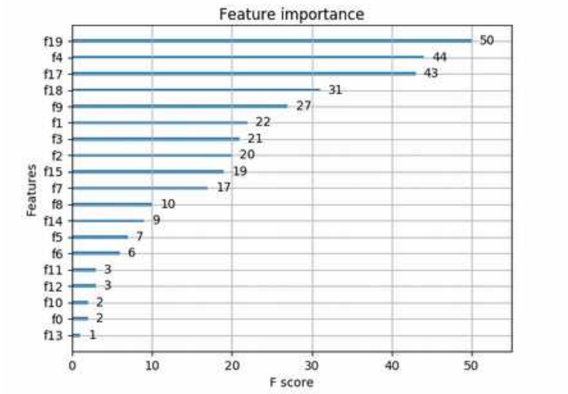
Interpretable models

Feature importance

Global

Specific

Built from decision trees



In decision tree models we can also see how all the data affects the output.

How many times a particular feature has been used to make a decision in the tree.

Feature importance in decision trees indicates how much each feature has been used to make a decision.

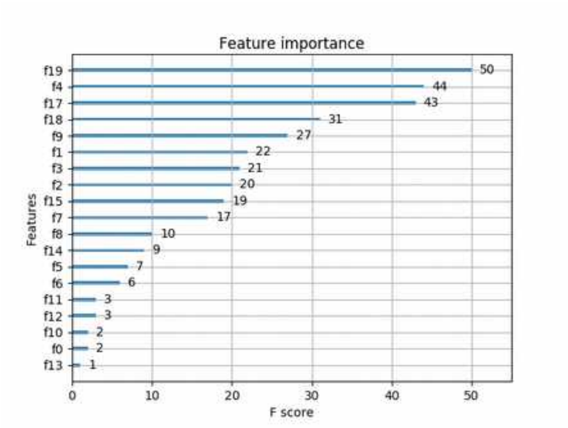
Interpretable models

Feature importance

Global

Specific

Built from decision trees



In decision tree models we can also see how all the data affects the output.

How many times a particular feature has been used to make a decision in the tree.

Feature importance in decision trees indicates how much each feature has been used to make a decision.

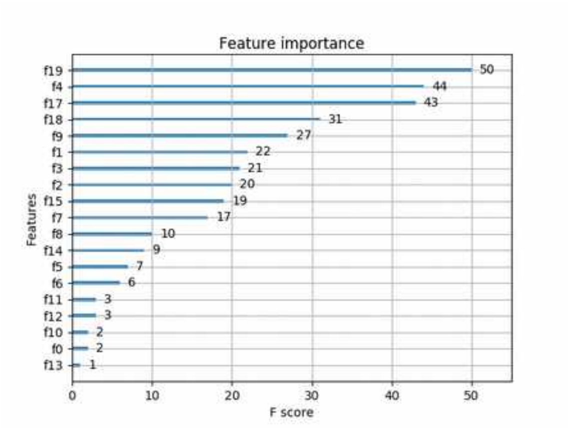
Why is the model predicting that the golf course will not have clients tomorrow?

Interpretable models

Feature importance

Global | Specific

Built from decision trees



In decision tree models we can also see how all the data affects the output.

How many times a particular feature has been used to make a decision in the tree.

Feature importance in decision trees indicates how much each feature has been used to make a decision.

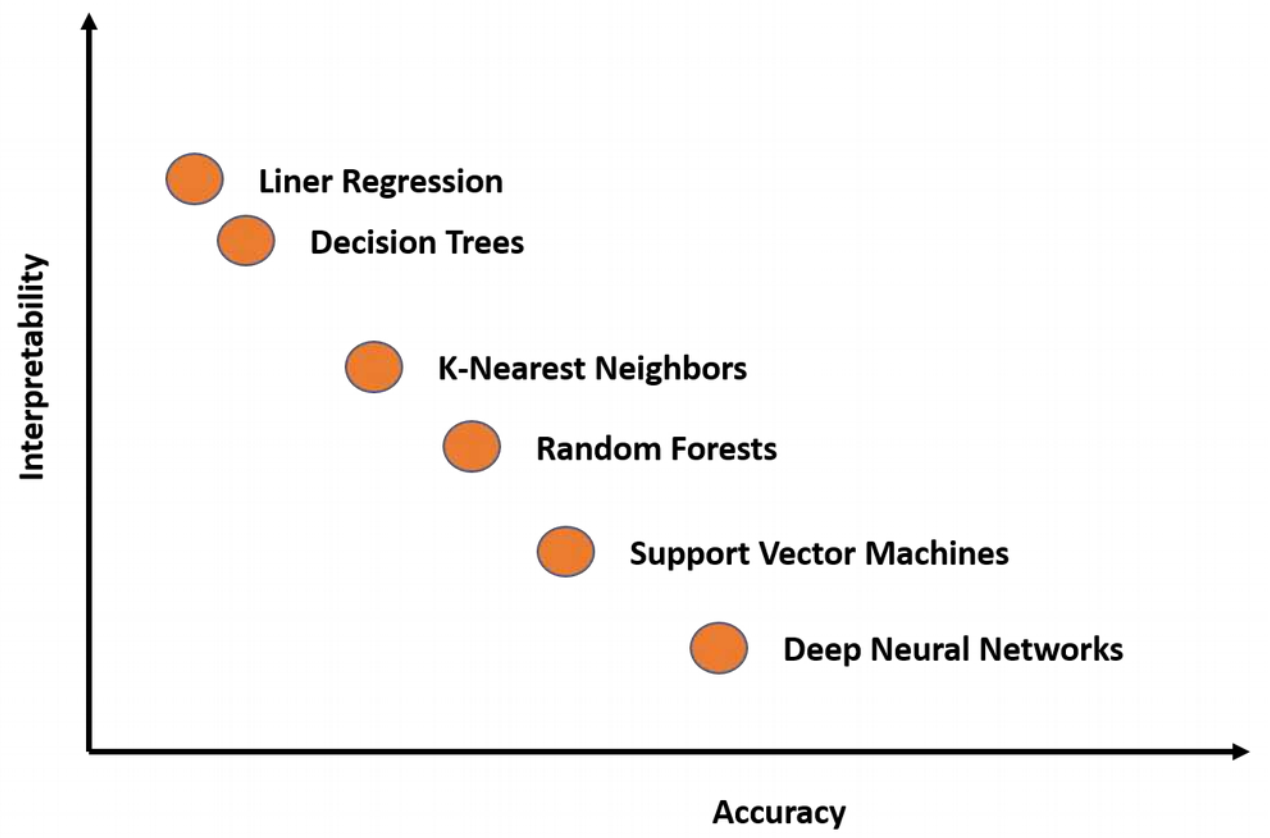
Why is the model predicting that the golf course will not have clients tomorrow?

Explanation: Sunny is the most important feature and tomorrow is not sunny

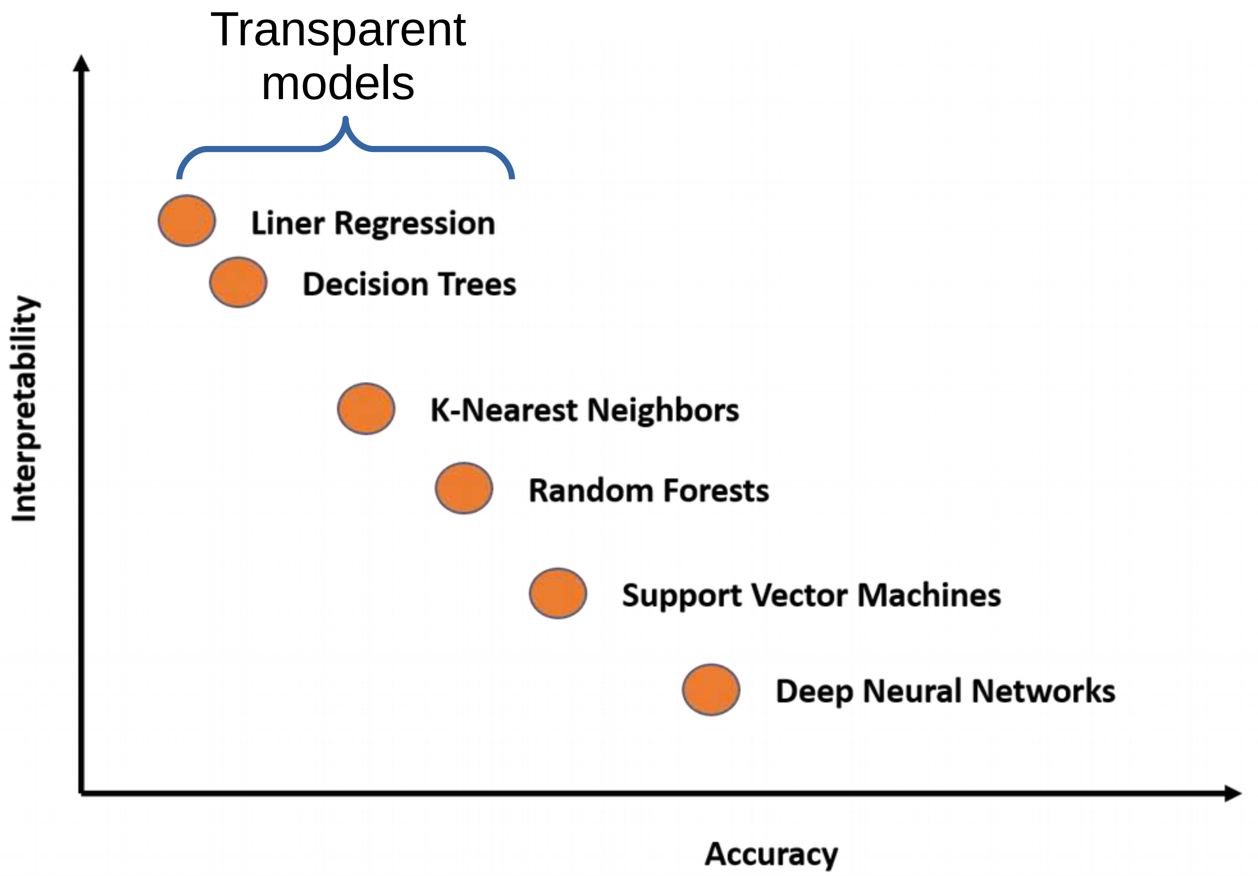
Interpretable models

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	Yes	Yes	No	regr
Logistic regression	No	Yes	No	class
Decision trees	No	Some	Yes	class,regr
RuleFit	Yes	No	Yes	class,regr
Naive Bayes	No	Yes	No	class
k-nearest neighbors	No	No	No	class,regr

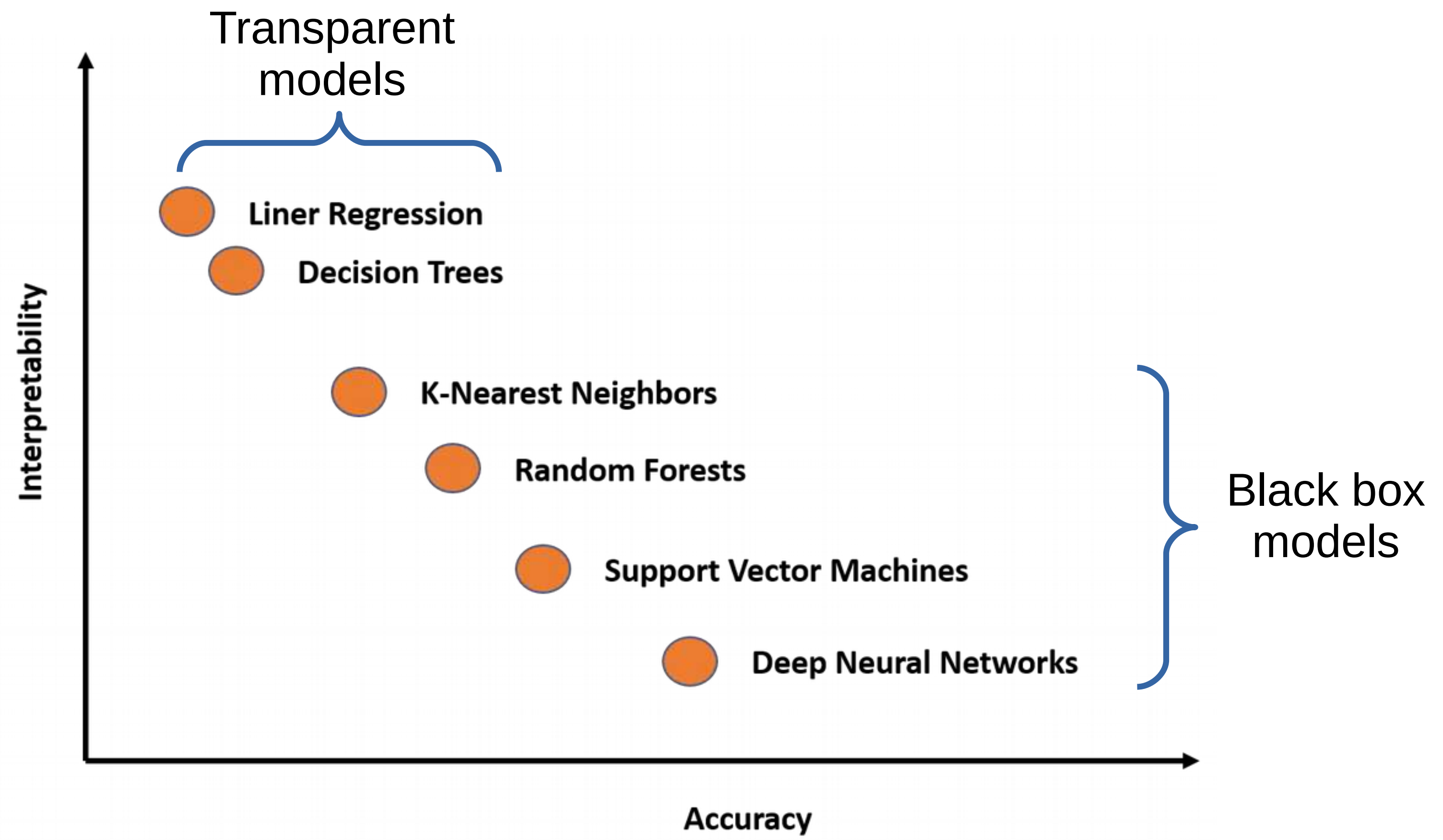
Interpretability challenge



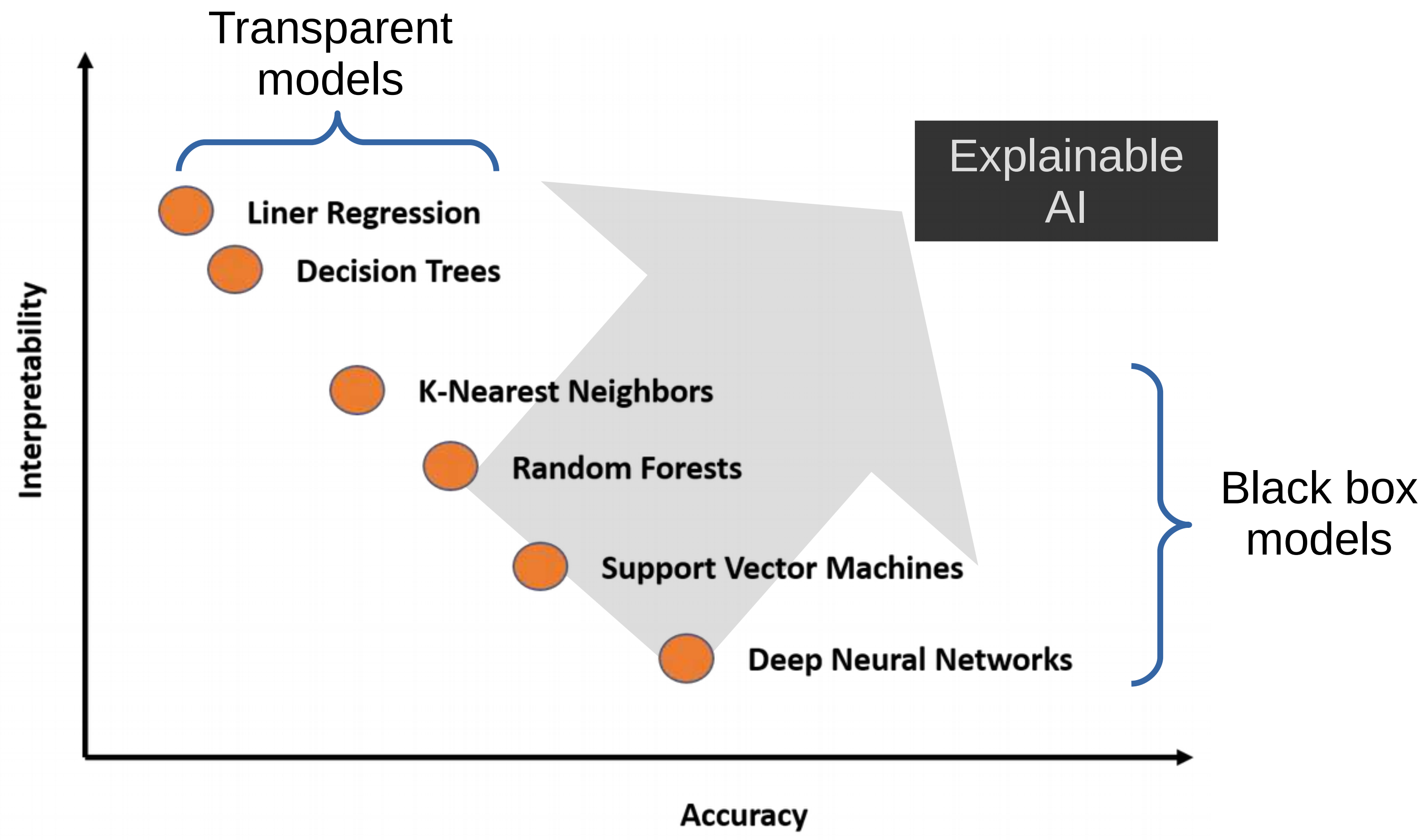
Interpretability challenge



Interpretability challenge



Interpretability challenge



Two main techniques

Two main techniques

Feature sensitivity

Test how changes in the inputs affect the outputs.

Two main techniques

Feature sensitivity

Test how changes in the inputs affect the outputs.

Surrogate models

Use transparent, explainable models, to interpret black box models

Feature sensitivity

Permutation Importance

Feature sensitivity

Permutation Importance

Idea: shuffle data in one column. How much does the output change?

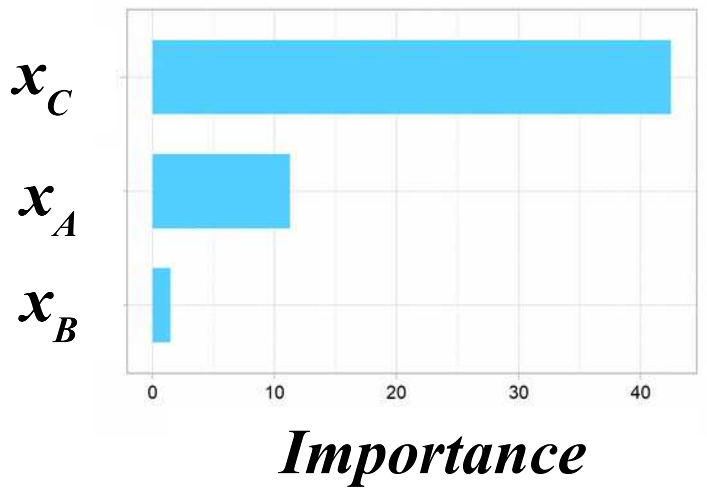
Feature sensitivity

Permutation Importance

Idea: shuffle data in one column. How much does the output change?

Large change: feature is more “important”

X_A	X_B	X_C	Y
xa1	xb1	xc1	y1
xa2	xb2	xc2	y2
xa3	xb3	xc3	y3
xa4	xb4	xc4	y4
xa5	xb5	xc5	y5
xa6	xb6	xc6	y6



Feature sensitivity

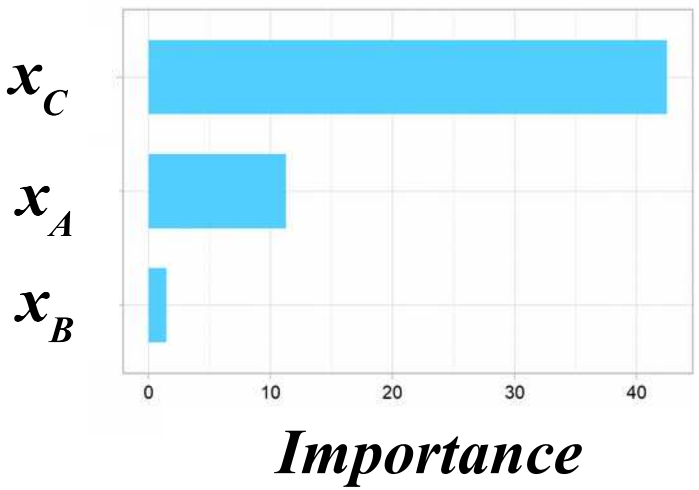
Permutation Importance

Agnostic

Idea: shuffle data in one column. How much does the output change?

Large change: feature is more “important”

X_A	X_B	X_C	Y
xa1	xb1	xc1	y1
xa2	xb2	xc2	y2
xa3	xb3	xc3	y3
xa4	xb4	xc4	y4
xa5	xb5	xc5	y5
xa6	xb6	xc6	y6



Feature sensitivity

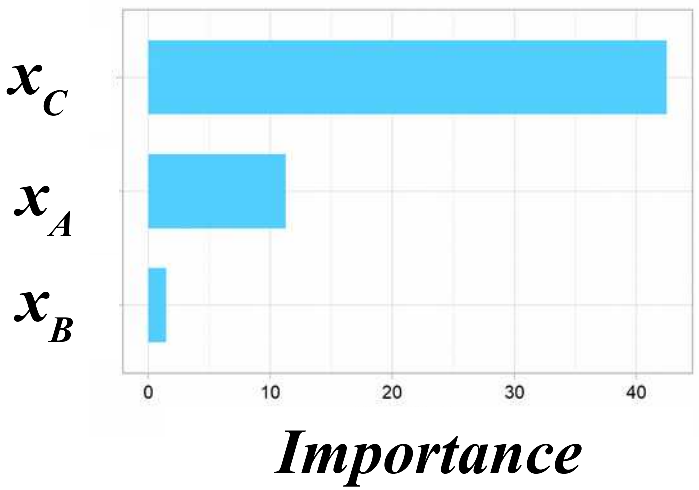
Permutation Importance

Agnostic **Global**

Idea: shuffle data in one column. How much does the output change?

Large change: feature is more "important"

X_A	X_B	X_C	Y
xa1	xb1	xc1	y1
xa2	xb2	xc2	y2
xa3	xb3	xc3	y3
xa4	xb4	xc4	y4
xa5	xb5	xc5	y5
xa6	xb6	xc6	y6



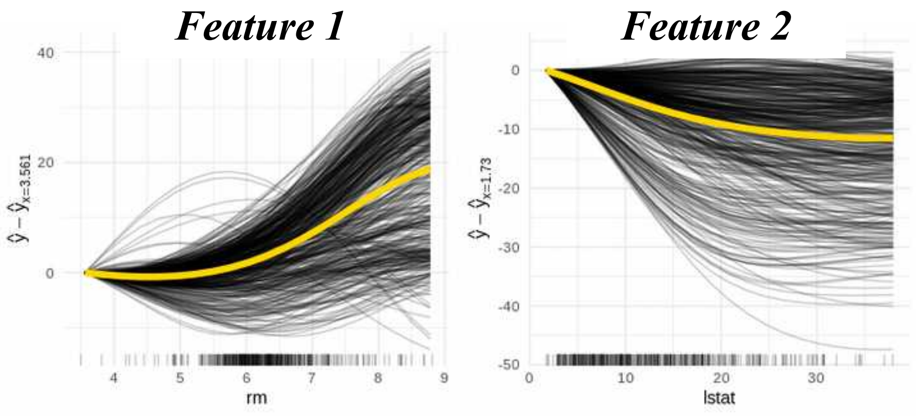
Feature sensitivity

Individual Conditional Expectation curves (ICE) and Partial Dependence Plots (PDP)

Feature sensitivity

Individual Conditional Expectation curves (ICE) and Partial Dependence Plots (PDP)

Idea: Continuously change the value of one feature. Do so for each one of the entries in the data set. This creates an error in the output, compared to what the trained model is expected to show.

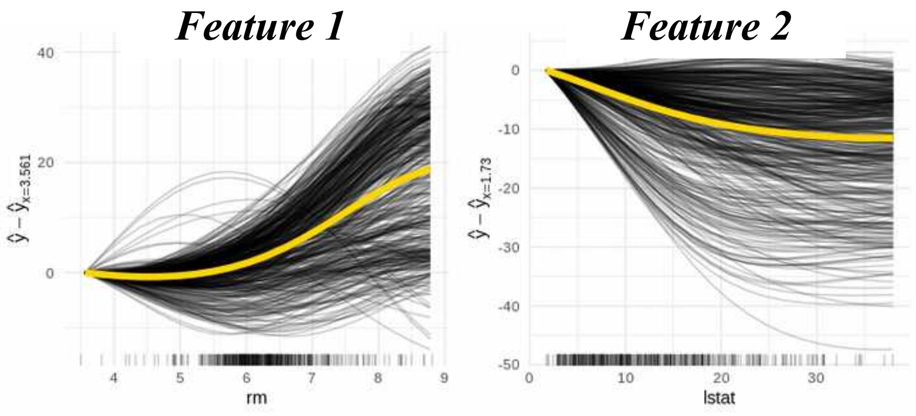


Feature sensitivity

Individual Conditional Expectation curves (ICE) and Partial Dependence Plots (PDP)

Idea: Continuously change the value of one feature. Do so for each one of the entries in the data set. This creates an error in the output, compared to what the trained model is expected to show.

This error is used to create the ICE curves (black). Take the average of the ICE curves to create the PDP (yellow)



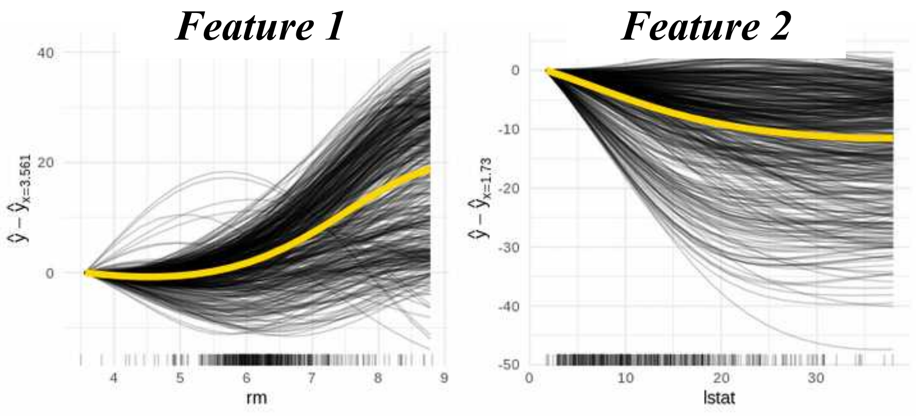
Feature sensitivity

Individual Conditional Expectation curves (ICE) and Partial Dependence Plots (PDP)

Idea: Continuously change the value of one feature. Do so for each one of the entries in the data set. This creates an error in the output, compared to what the trained model is expected to show.

This error is used to create the ICE curves (black). Take the average of the ICE curves to create the PDP (yellow)

Explanation: The output has a (linear/monotonic/exponential) relationship



Feature sensitivity

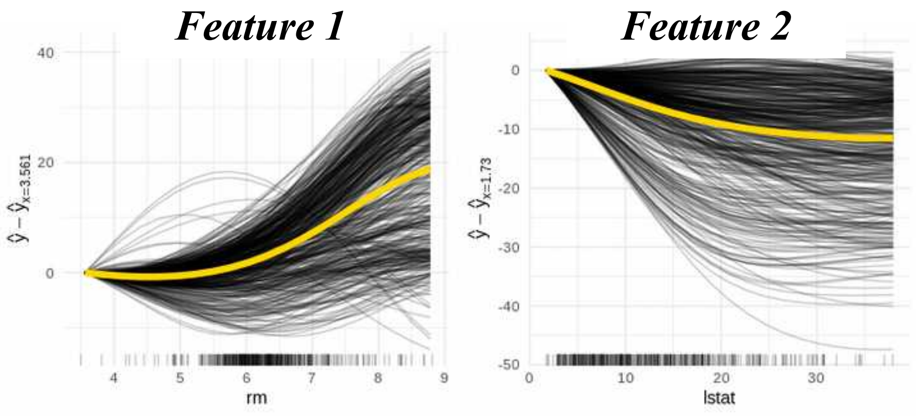
Individual Conditional Expectation curves (ICE) and Partial Dependence Plots (PDP)

Idea: Continuously change the value of one feature. Do so for each one of the entries in the data set. This creates an error in the output, compared to what the trained model is expected to show.

This error is used to create the ICE curves (black). Take the average of the ICE curves to create the PDP (yellow)

Explanation: The output has a (linear/monotonic/exponential) relationship

Agnostic



Feature sensitivity

Individual Conditional Expectation curves (ICE) and Partial Dependence Plots (PDP)

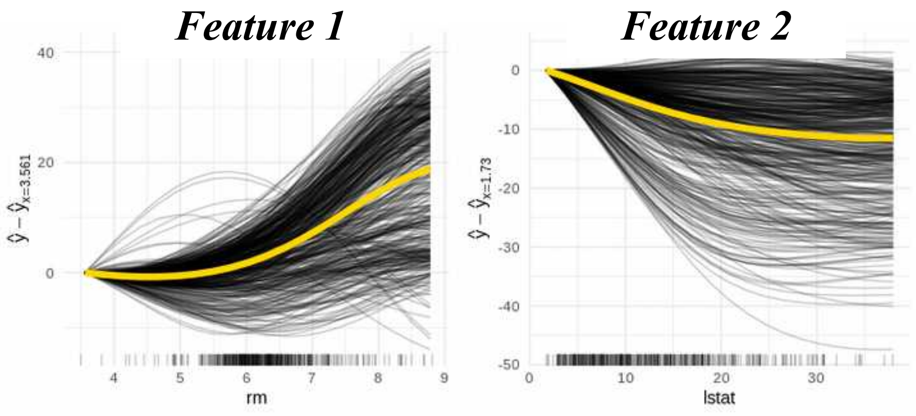
Idea: Continuously change the value of one feature. Do so for each one of the entries in the data set. This creates an error in the output, compared to what the trained model is expected to show.

This error is used to create the ICE curves (black). Take the average of the ICE curves to create the PDP (yellow)

Explanation: The output has a (linear/monotonic/exponential) relationship

Agnostic

Global



Feature sensitivity

SHapley Additive exPlanation (SHAP)

Feature sensitivity

SHapley Additive exPlanation (SHAP)

Idea: each feature is an actor that contributes to the output of the model. Shapley Values are used in game theory to fairly distribute gains and costs to features working in coalition.

Feature sensitivity

SHapley Additive exPlanation (SHAP)

Idea: each feature is an actor that contributes to the output of the model. Shapley Values are used in game theory to fairly distribute gains and costs to features working in coalition.

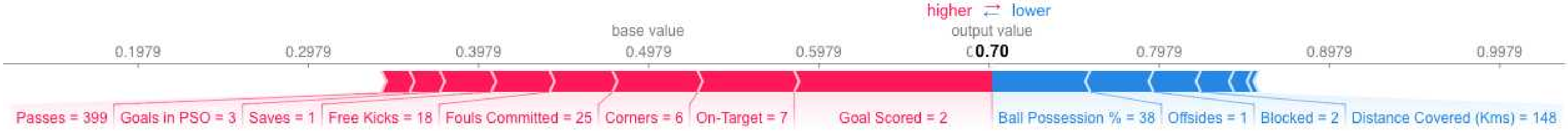
Explanation: This feature has a possitive/negative contribution pulling it away from/towards the average target value.

Feature sensitivity

SHapley Additive exPlanation (SHAP)

Idea: each feature is an actor that contributes to the output of the model. Shapley Values are used in game theory to fairly distribute gains and costs to features working in coalition.

Explanation: This feature has a possitive/negative contribution pulling it away from/towards the average target value.



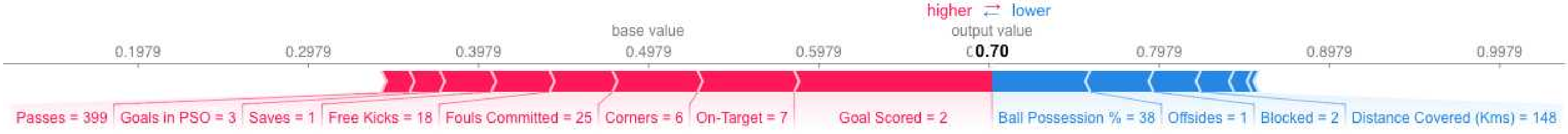
Feature sensitivity

SHapley Additive exPlanation (SHAP)

Idea: each feature is an actor that contributes to the output of the model. Shapley Values are used in game theory to fairly distribute gains and costs to features working in coalition.

Explanation: This feature has a possitive/negative contribution pulling it away from/towards the average target value.

Agnostic



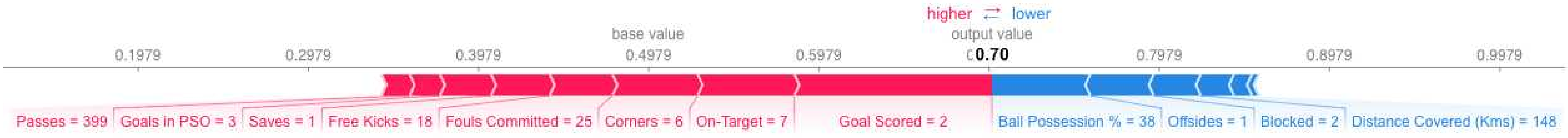
Feature sensitivity

SHapley Additive exPlanation (SHAP)

Agnostic **Local**

Idea: each feature is an actor that contributes to the output of the model. Shapley Values are used in game theory to fairly distribute gains and costs to features working in coalition.

Explanation: This feature has a possitive/negative contribution pulling it away from/towards the average target value.



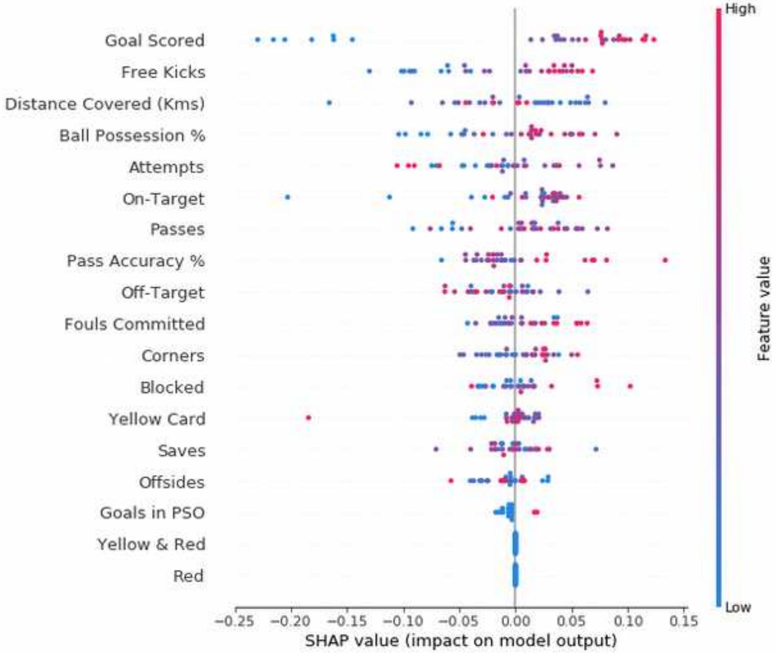
Feature sensitivity

SHapley Additive exPlanation (SHAP)

Idea: each feature is an actor that contributes to the output of the model. Shapley Values are used in game theory to fairly distribute gains and costs to features working in coalition.

Explanation: This feature has a possitive/negative contribution pulling it away from/towards the average target value.

Agnostic **Global**



Surrogate models

Agnostic

Global

Global surrogate

Surrogate models

Agnostic

Global

Global surrogate

Replace the complex model using a simpler, inherently explainable, model

Surrogate models

Agnostic

Global

Global surrogate

Replace the complex model using a simpler, inherently explainable, model

- The target data of the simpler model will be the output of the complex model

Surrogate models

Agnostic

Global

Global surrogate

Replace the complex model using a simpler, inherently explainable, model

- The target data of the simpler model will be the output of the complex model
- The error between the two models is minimized

Surrogate models

Agnostic

Global

Global surrogate

Replace the complex model using a simpler, inherently explainable, model

- The target data of the simpler model will be the output of the complex model
- The error between the two models is minimized
- The accuracy of the explainable model is presented

Surrogate models

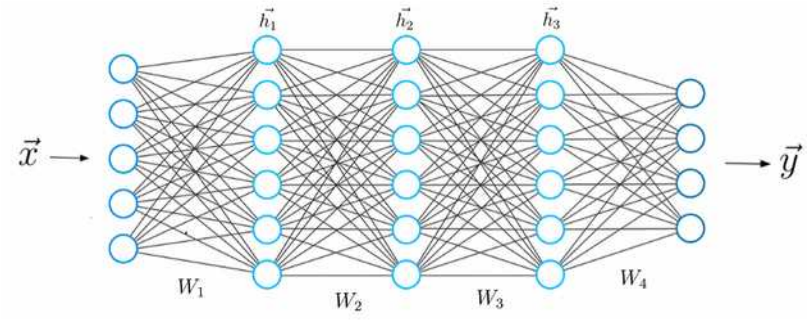
Agnostic

Global

Global surrogate

Replace the complex model using a simpler, inherently explainable, model

- The target data of the simpler model will be the output of the complex model
- The error between the two models is minimized
- The accuracy of the explainable model is presented



Surrogate models

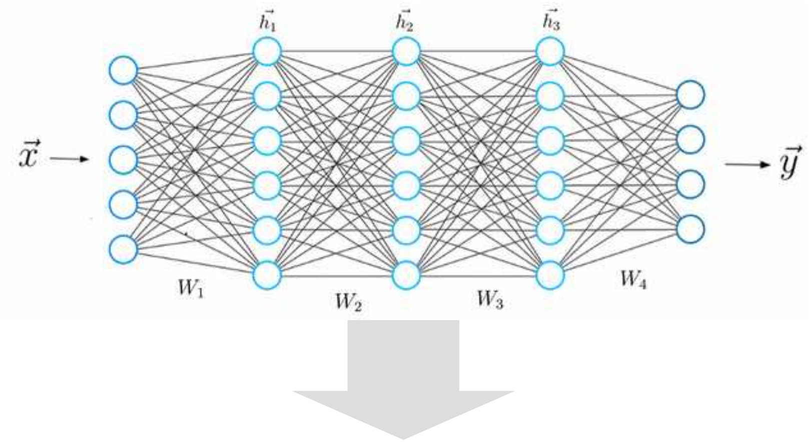
Agnostic

Global

Global surrogate

Replace the complex model using a simpler, inherently explainable, model

- The target data of the simpler model will be the output of the complex model
- The error between the two models is minimized
- The accuracy of the explainable model is presented



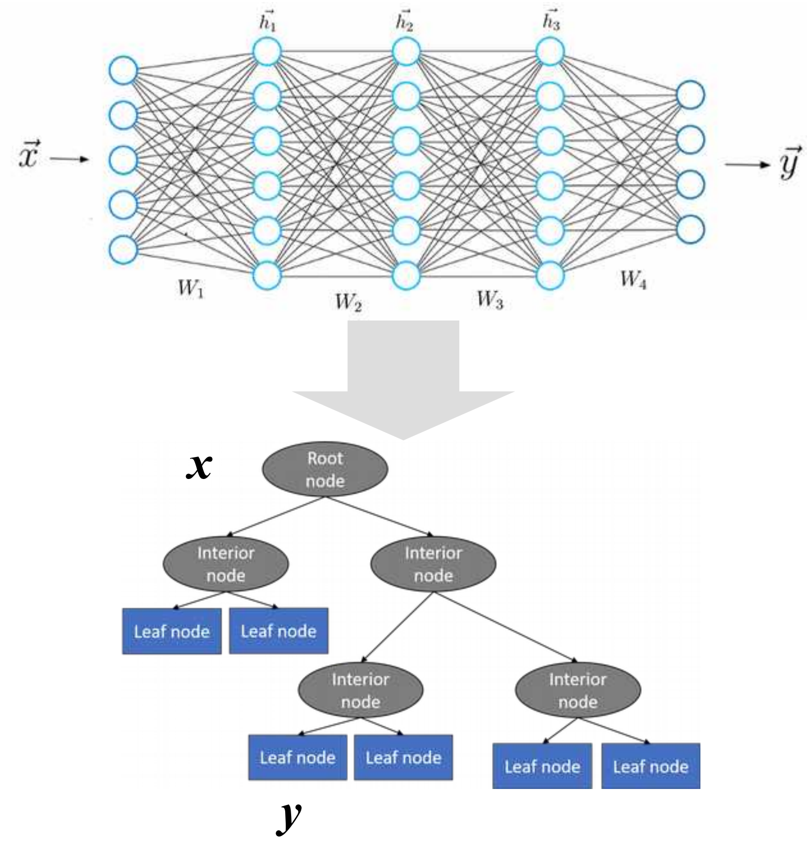
Surrogate models

Agnostic **Global**

Global surrogate

Replace the complex model using a simpler, inherently explainable, model

- The target data of the simpler model will be the output of the complex model
- The error between the two models is minimized
- The accuracy of the explainable model is presented



Surrogate models

Agnostic

Local

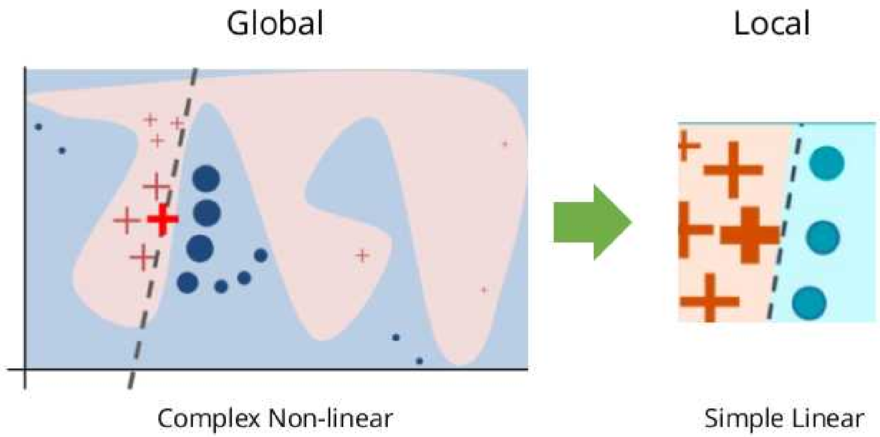
Local Interpretable Model-Agnostic (LIME)

Surrogate models

Agnostic

Local

Local Interpretable Model-Agnostic (LIME)



Surrogate models

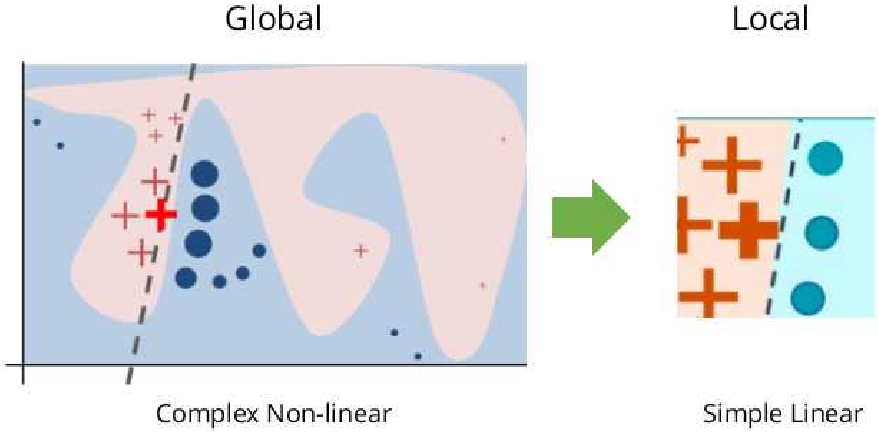
Agnostic

Local

Local Interpretable Model-Agnostic (LIME)

Replace the model in the local space around a single entry, by a linear model.:

- Create artificial data points around the entry



Surrogate models

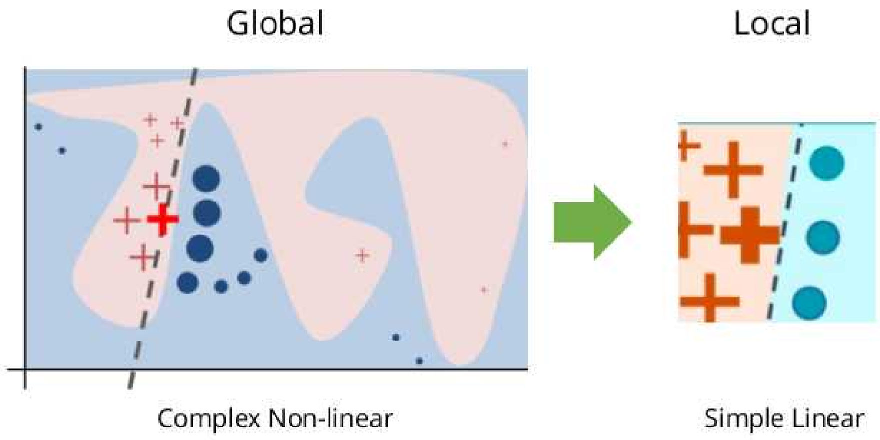
Agnostic

Local

Local Interpretable Model-Agnostic (LIME)

Replace the model in the local space around a single entry, by a linear model.:

- Create artificial data points around the entry
- Give a higher weight to points closer to the original data point



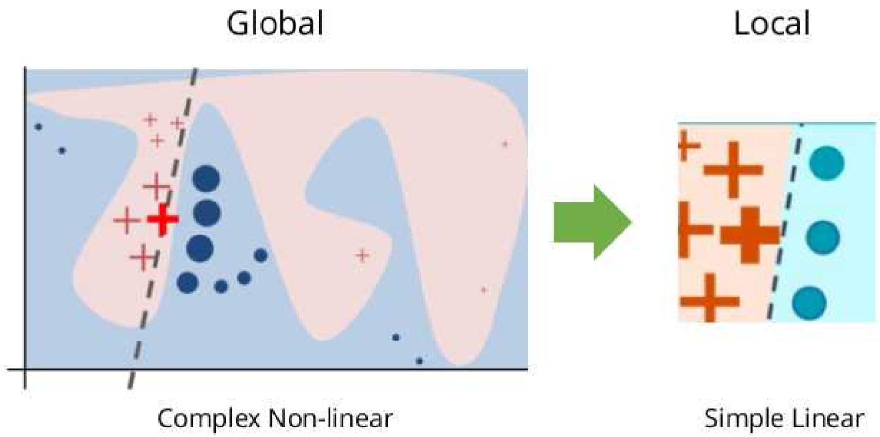
Surrogate models

Agnostic **Local**

Local Interpretable Model-Agnostic (LIME)

Replace the model in the local space around a single entry, by a linear model.:

- Create artificial data points around the entry
- Give a higher weight to points closer to the original data point
- Calculate the outputs from the complex model



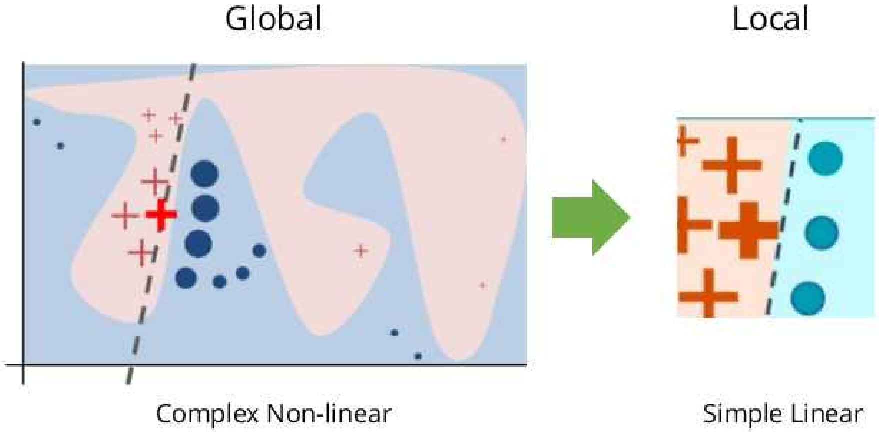
Surrogate models

Agnostic **Local**

Local Interpretable Model-Agnostic (LIME)

Replace the model in the local space around a single entry, by a linear model.:

- Create artificial data points around the entry
- Give a higher weight to points closer to the original data point
- Calculate the outputs from the complex model
- Train a linear model using these points



Surrogate models

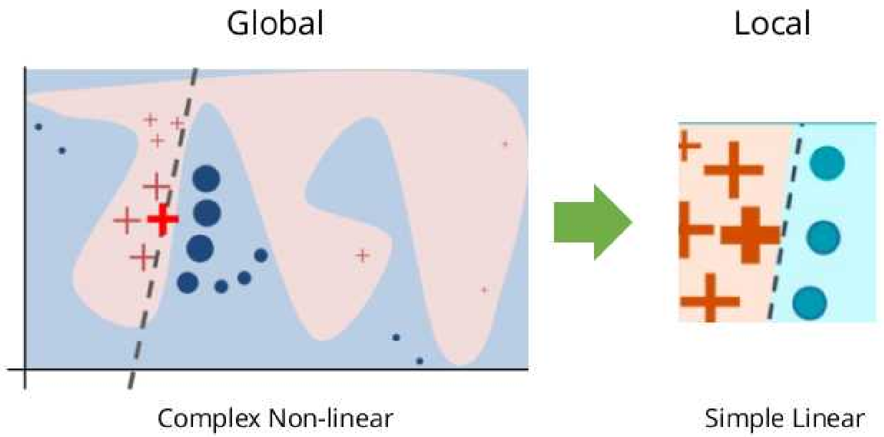
Agnostic **Local**

Local Interpretable Model-Agnostic (LIME)

Replace the model in the local space around a single entry, by a linear model.:

- Create artificial data points around the entry
- Give a higher weight to points closer to the original data point
- Calculate the outputs from the complex model
- Train a linear model using these points

- **Interpretation**: the linear model explains why the original point belongs to one class or the other



Software tools for XAI

Permutation importance with ELI5

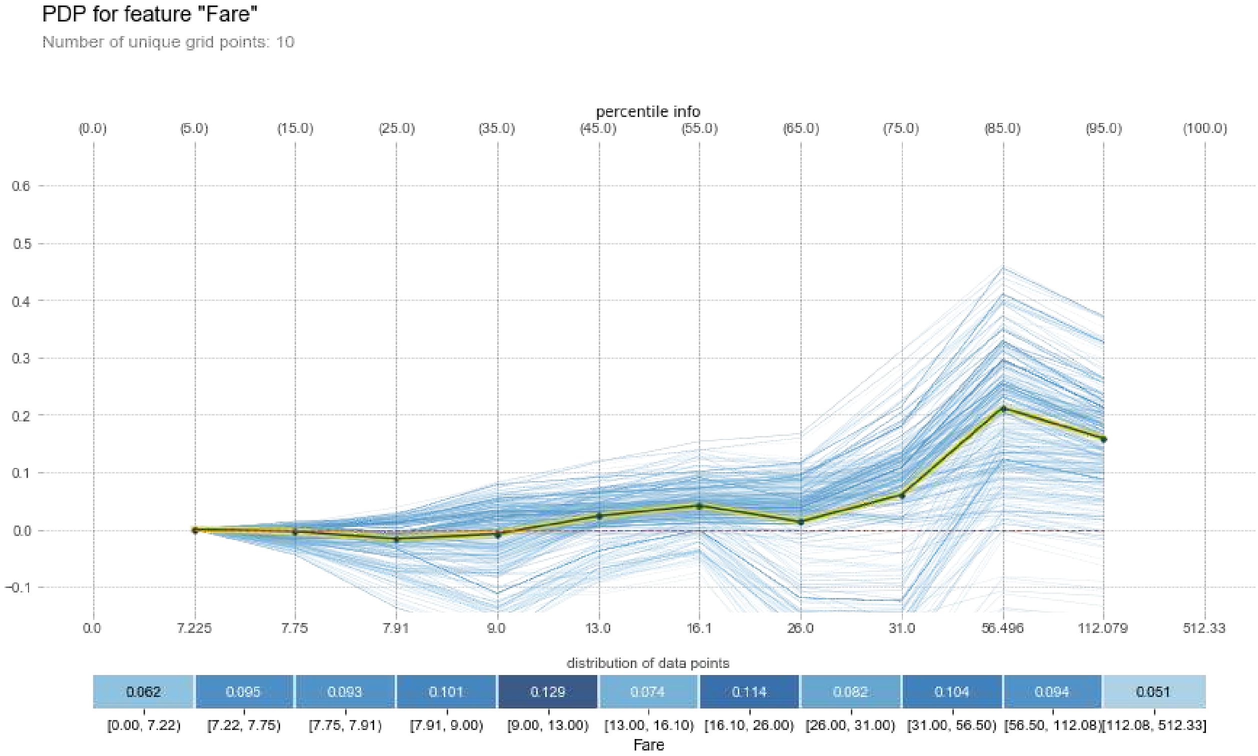
```
import eli5
from eli5.sklearn import PermutationImportance

perm = PermutationImportance(my_model, random_state=1).fit(val_X,
val_y)
eli5.show_weights(perm, feature_names = val_X.columns.tolist())
```

Weight	Feature
0.0750 ± 0.1159	Goal Scored
0.0625 ± 0.0791	Corners
0.0437 ± 0.0500	Distance Covered (Kms)
0.0375 ± 0.0729	On-Target
0.0375 ± 0.0468	Free Kicks
0.0187 ± 0.0306	Blocked
0.0125 ± 0.0750	Pass Accuracy %
0.0125 ± 0.0500	Yellow Card
0.0063 ± 0.0468	Saves
0.0063 ± 0.0250	Offsides
0.0063 ± 0.1741	Off-Target
0.0000 ± 0.1046	Passes
0 ± 0.0000	Red
0 ± 0.0000	Yellow & Red
0 ± 0.0000	Goals in PSO
-0.0312 ± 0.0884	Fouls Committed
-0.0375 ± 0.0919	Attempts
-0.0500 ± 0.0500	Ball Possession %

Software tools for XAI

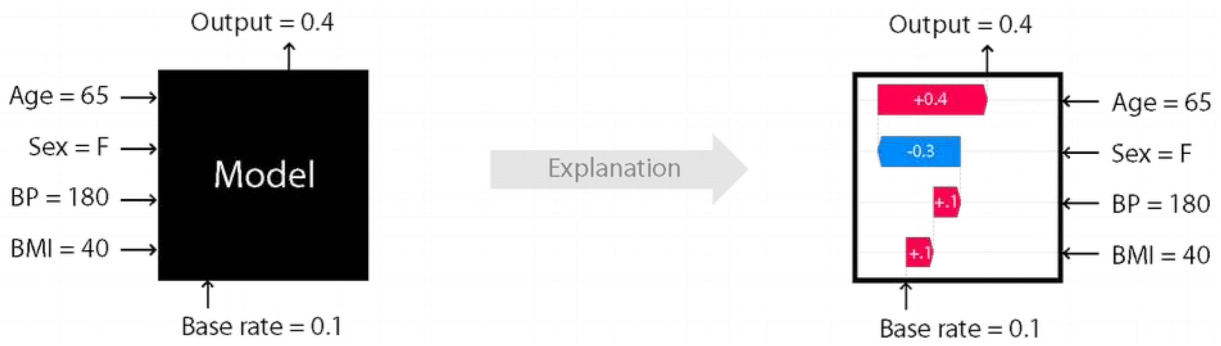
Partial Dependence Plots with PDPbox



Software tools for XAI

SHAP values with the Shap library:

SHAP



Discussion

Discussion

- Reproducibility requires a careful setup of your ML experiment and a detailed description of all of its components, including data sources

Discussion

- Reproducibility requires a careful setup of your ML experiment and a detailed description of all of its components, including data sources
- To be explainable a model needs to be trustworthy

Discussion

- Reproducibility requires a careful setup of your ML experiment and a detailed description of all of its components, including data sources
- To be explainable a model needs to be trustworthy
- Interpretations in: scope (global, local), model type (specific, agnostic)

Discussion

- Reproducibility requires a careful setup of your ML experiment and a detailed description of all of its components, including data sources
- To be explainable a model needs to be trustworthy
- Interpretations in: scope (global, local), model type (specific, agnostic)
- Two types of approach: feature sensitivity, surrogate models

Discussion

- Reproducibility requires a careful setup of your ML experiment and a detailed description of all of its components, including data sources
- To be explainable a model needs to be trustworthy
- Interpretations in: scope (global, local), model type (specific, agnostic)
- Two types of approach: feature sensitivity, surrogate models
- Multiple new and interesting techniques

Conclusion

Conclusion

- Please publish your models so others can reproduce your results.

Conclusion

- Please publish your models so others can reproduce your results.
- Include in your publications a value for the *budget* of your model.

Conclusion

- Please publish your models so others can reproduce your results.
- Include in your publications a value for the *budget* of your model.
- To convince your peers, try to include in your model evaluation any of the XAI techniques described.

Conclusion

- Please publish your models so others can reproduce your results.
- Include in your publications a value for the *budget* of your model.
- To convince your peers, try to include in your model evaluation any of the XAI techniques described.
- XAI is an additional layer to your models that will become more and more important in the next years.

Conclusion

- Please publish your models so others can reproduce your results.
- Include in your publications a value for the *budget* of your model.
- To convince your peers, try to include in your model evaluation any of the XAI techniques described.
- XAI is an additional layer to your models that will become more and more important in the next years.
- XAI is currently under constant evolution; keep an eye on the latest results and tools.

XAI: one little more layer to worry about...
...but an important one!

Enjoy XAI
Thank you!

This school has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 776262 (AIDA, www.aida-space.eu)