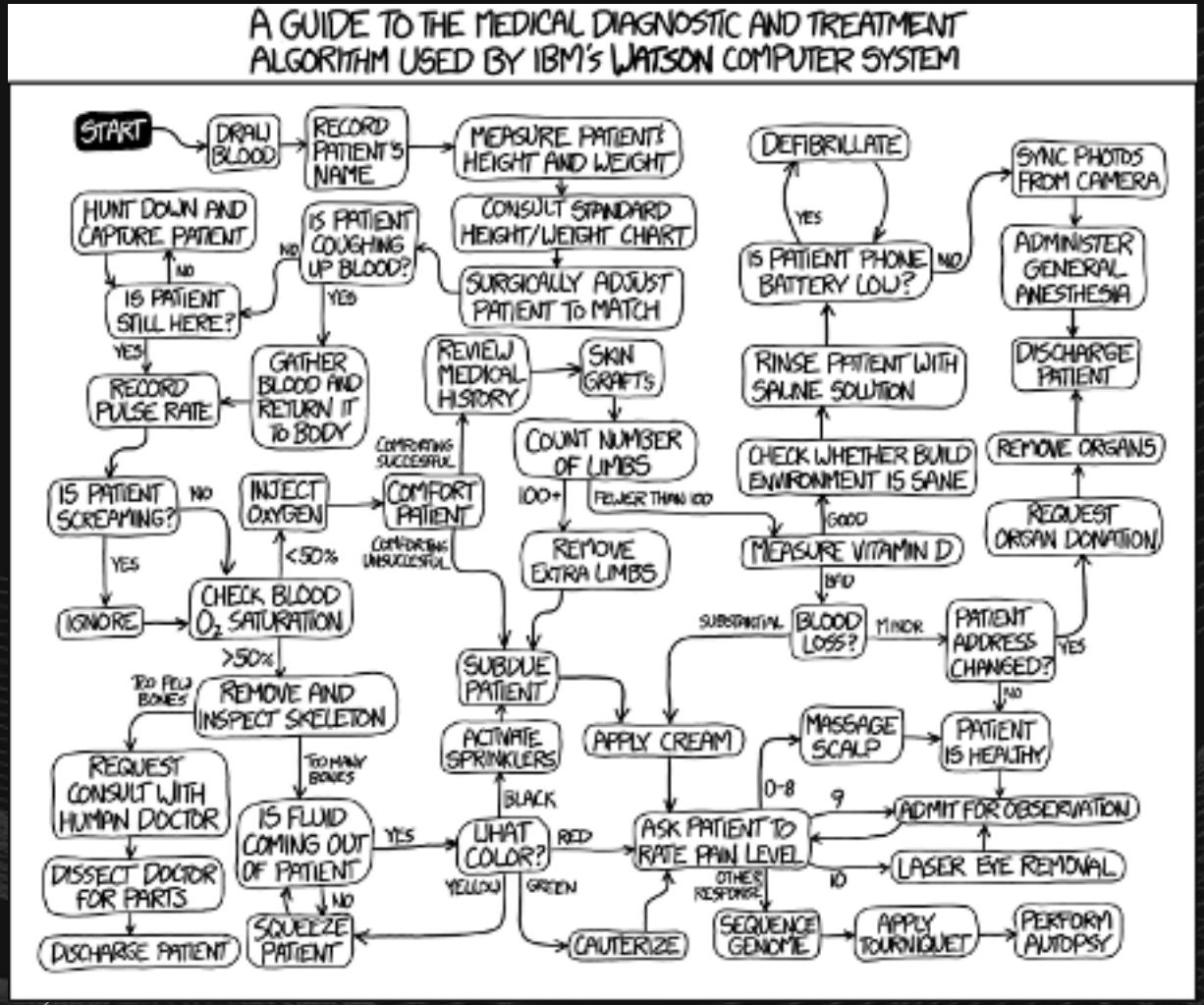
# MEETUP: APAC

## Responsible Automation, towards Interpretable and Fair AutoML

**THURSDAY, OCTOBER 8 , 2 PM - 3 PM AEST**

**In partnership with: Data Science Sydney Meetup Group**




What's wrong with this flowchart ?

https://xkcd.com/1619/

# Agenda

- Introduction
- About H2O
- Responsible Automation
- Limitations and Practical Challenges
- H2O AutoML - Fairness and Interpretability Demo
- Q & A

# Who is that talking?

## James Orton
### Data Scientist @ H2O.ai
### Australia and New Zealand

**Connect with me**

linkedin.com/in/jamesortonthedataman

james.orton@h2o.ai

## Chetan Ganjihal
### Data Scientist @ H2O.ai
### Australia and New Zealand

**Connect with me**

linkedin.com/in/chetan-ganjihal

chetan.ganjihal@h2o.ai

# About H2O

# Democratizing AI

- *Our mission to use **AI for Good** permeates into everything we do*

## TRUSTED PARTNER

AI Transformation
Bringing AI to industry by helping companies transform their businesses with H2O.ai.

## COMMUNITY

Open Source
An industry leader in providing open source, cutting edge AI & ML platforms (H2O-3).

## IMPACT/SOCIAL

AI4GOOD
Bringing AI to impact by augmenting non-profits and social ventures with technological resources and capabilities.

# AI4Good @ H2O.ai

Education

- Partner AI4All, a non profit org to improve diversity and inclusion in AI
- Mentorship and education to create talent pool for underrepresented groups

Wildlife Conservation

- Work with Wildbook to conserve wildlife
- Aid in research to detect population of threatened species
- Use of computer vision, citizen science and AI to speed population analysis and find new insights to fight extinction

# Responsible Automation

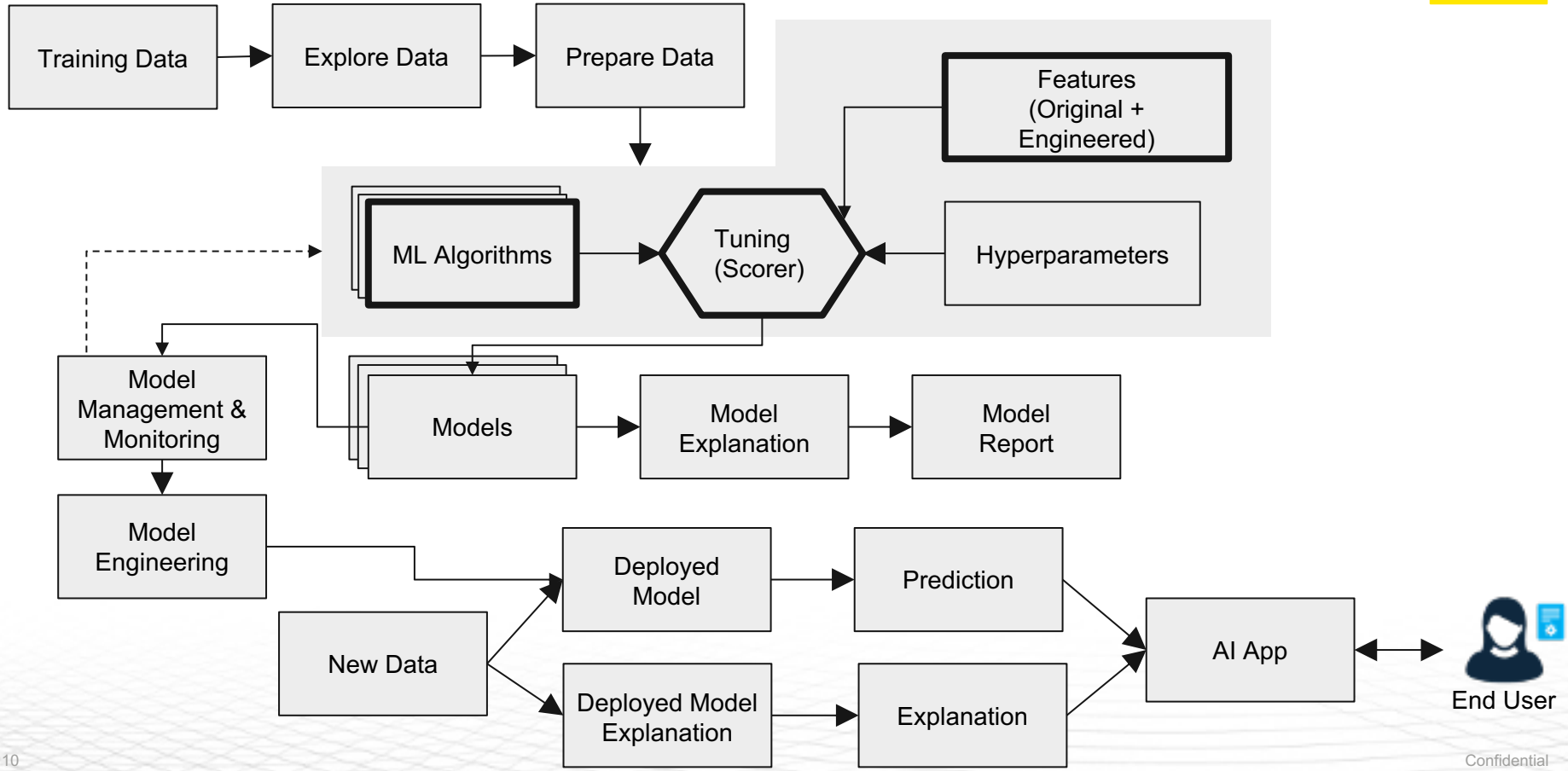
## What, Why and How ?

# Extreme Automation

Connected World

- Automation
- AI
- Algorithms

# Typical ML Workflow



# What is AutoML?

**Automated machine learning** (**AutoML**) is the process of automating the process of applying machine learning to real-world problems. AutoML covers the complete pipeline from the raw dataset to the deployable machine learning model. AutoML was proposed as an artificial intelligence-based solution to the ever-growing challenge of applying machine learning.

# Side effects of AutoML

- Limited oversight in model training
  - Potential for accidental misuse
  - Current AutoML systems have limited guardrails
- Models easily deployed in prod (bad models also)
- Troubleshooting and debugging models may be time consuming

# Why should you care?


Microsoft's AI Twitter bot goes dark after racist, sexist tweets

**Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.**

Everything that went wrong with the botched A-Levels algorithm

WELFARE

**What is robodebt?**

There are many questions about the government's controversial robodebt scheme. Let's start with what it actually is.

Racial bias in a medical algorithm favors white patients over sicker black patients

'Rogue' Algorithm Blamed for Historic Crash of the British Pound

*Apple Card Investigated After Gender Discrimination Complaints*

# Why should you care?

## Social Good

Play a positive role in creating a more just, fair and equitable society.

## Regulation

GDPR, FRCA and many many more.

## Trust

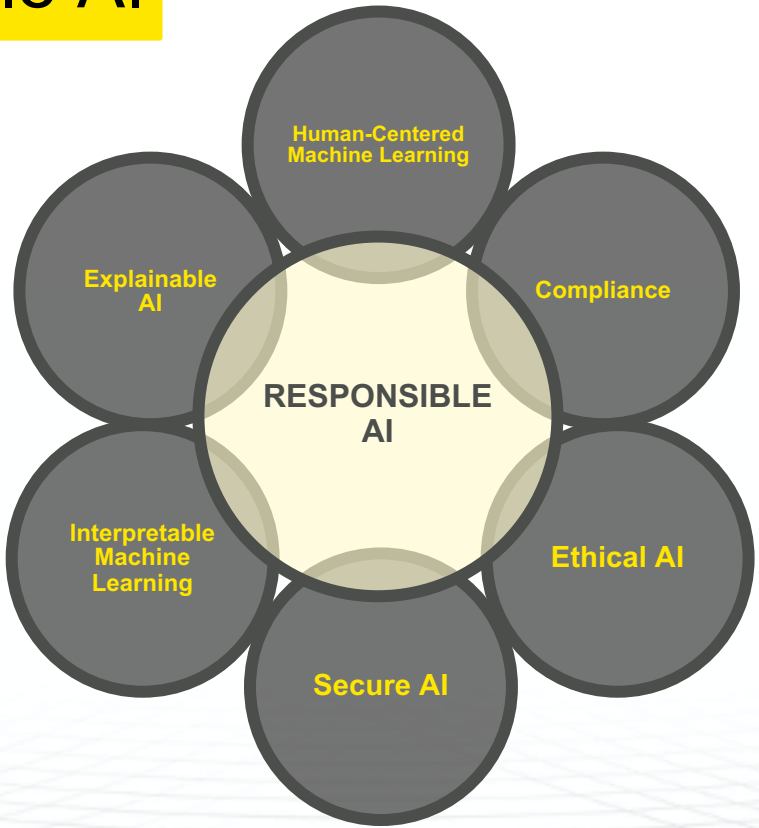
Do your stakeholders trust your ML? Does the public?

## Robustness

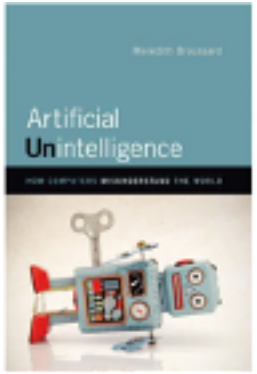
If you don’t understand your model it is more likely to fail

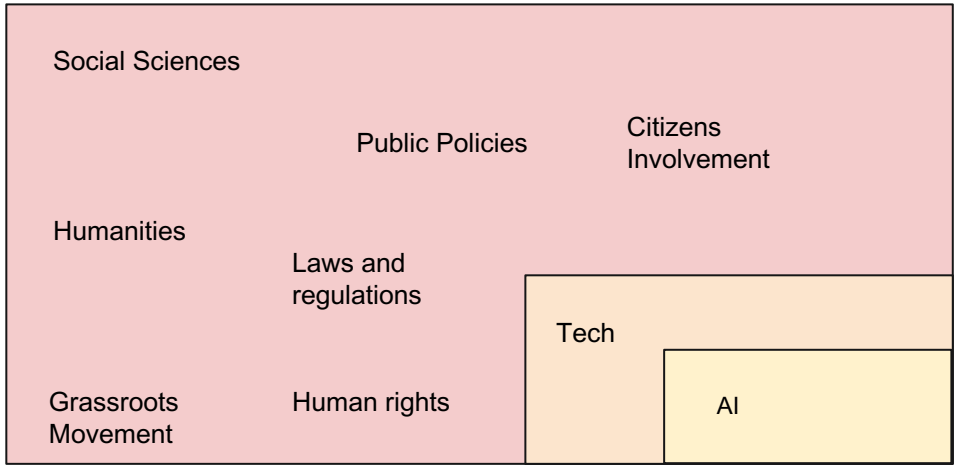
# Responsible AI

- **Explainable AI:** Focuses on the ability analyze a ML model ***after*** it has been developed
- **Interpretable Machine Learning:** Transparent model architectures and increasing how intuitive and understandable ML models can be
- **Ethical AI:** Sociological fairness in machine learning predictions (i.e., whether one category of person is being weighted unequally)
- **Secure AI:** Debugging and deploying ML models with similar counter-measures against insider and cyber threats as would be seen in traditional software
- **Human-Centered ML:** User interactions with AI and ML systems



# How do we fix misplaced faith in AI ?



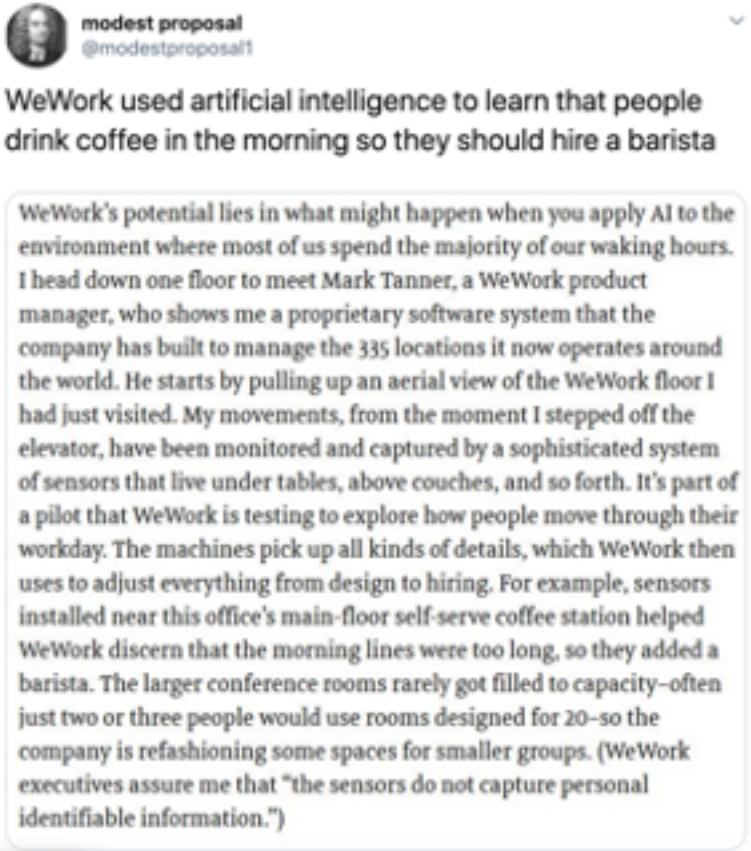




## Technochauvinism

Technochauvinism is simply the belief that technology is always the solution to every problem.

- Meredith Broussard, Artificial Unintelligence

# When the only tool we have is a hammer…



modest proposal
@modestproposal1

WeWork used artificial intelligence to learn that people drink coffee in the morning so they should hire a barista

WeWork's potential lies in what might happen when you apply AI to the environment where most of us spend the majority of our waking hours. I head down one floor to meet Mark Tanner, a WeWork product manager, who shows me a proprietary software system that the company has built to manage the 335 locations it now operates around the world. He starts by pulling up an aerial view of the WeWork floor I had just visited. My movements, from the moment I stepped off the elevator, have been monitored and captured by a sophisticated system of sensors that live under tables, above couches, and so forth. It's part of a pilot that WeWork is testing to explore how people move through their workday. The machines pick up all kinds of details, which WeWork then uses to adjust everything from design to hiring. For example, sensors installed near this office's main-floor self-serve coffee station helped WeWork discern that the morning lines were too long, so they added a barista. The larger conference rooms rarely got filled to capacity–often just two or three people would use rooms designed for 20–so the company is refashioning some spaces for smaller groups. (WeWork executives assure me that "the sensors do not capture personal identifiable information.")

Do we really need AI ?

# Don't miss the forest for the trees...

"Can I minimize differences in accuracy between subgroups" is less important than "should this be built at all"

@rctatman

Dr. Rachael Tatman

http://www.rctatman.com/talks/what-i-wont-build

Slide borrowed with permission from Erin LeDell, Chief ML Scientist & H2O AutoML lead

# More Questions to Ask

- How was the data acquired ?
- What labour was used ?
- What was it meant to be used for ?
- Have you thought about Consent, Privacy and Bias
- Will there be a shift in power ?
- Who will be benefited and who will be harmed ?

Datasheets for datasets

Model Cards

Explainable AI

Interpretability

# H2O AutoML - Fairness and Interpretability

# H2O Core: Distributed Machine Learning at Scale

Distributed, in-memory machine learning for Big Data

Designed to run on top of Hadoop, Spark and Kubernetes

Production model code ready for deployment

Ideal for data scientists looking to build ML models at scale using Python and R

# **H2O Open Source:** Fast, World-class Algos at Scale

**Familiar languages & IDEs**






**Familiar algos in H2O**

**Common**

- Quantiles
- Early Stopping

**Supervised**

- Cox Proportional Hazards (CoxPH)
- Deep Learning (Neural Networks)
- Distributed Random Forest (DRF)
- Generalized Linear Model (GLM)
- Gradient Boosting Machine (GBM)
- Naive Bayes Classifier
- Stacked Ensembles
- Support Vector Machine (SVM)
- XGBoost

**Unsupervised**

- Aggregator
- Generalized Low Rank Models (GLRM)
- Isolation Forest
- K-Means Clustering
- Principal Component Analysis (PCA)

**Checkpointing Models**

**Grid (Hyperparameter) Search**

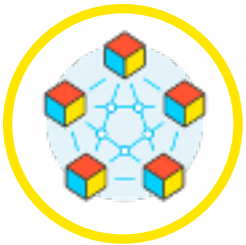
**AutoML: Automatic Machine Learning**

**Distributed in-memory compute**

***Fast* & *scalable* to TBs**


**Seamless native integration to Hadoop / Spark framework**

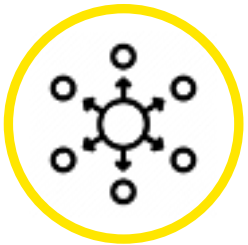
# H2O Open Source: Benefits



**Distributed in-memory machine learning for Big Data**

**High speed and accuracy model training**

**No data sampling**
**No approximations**



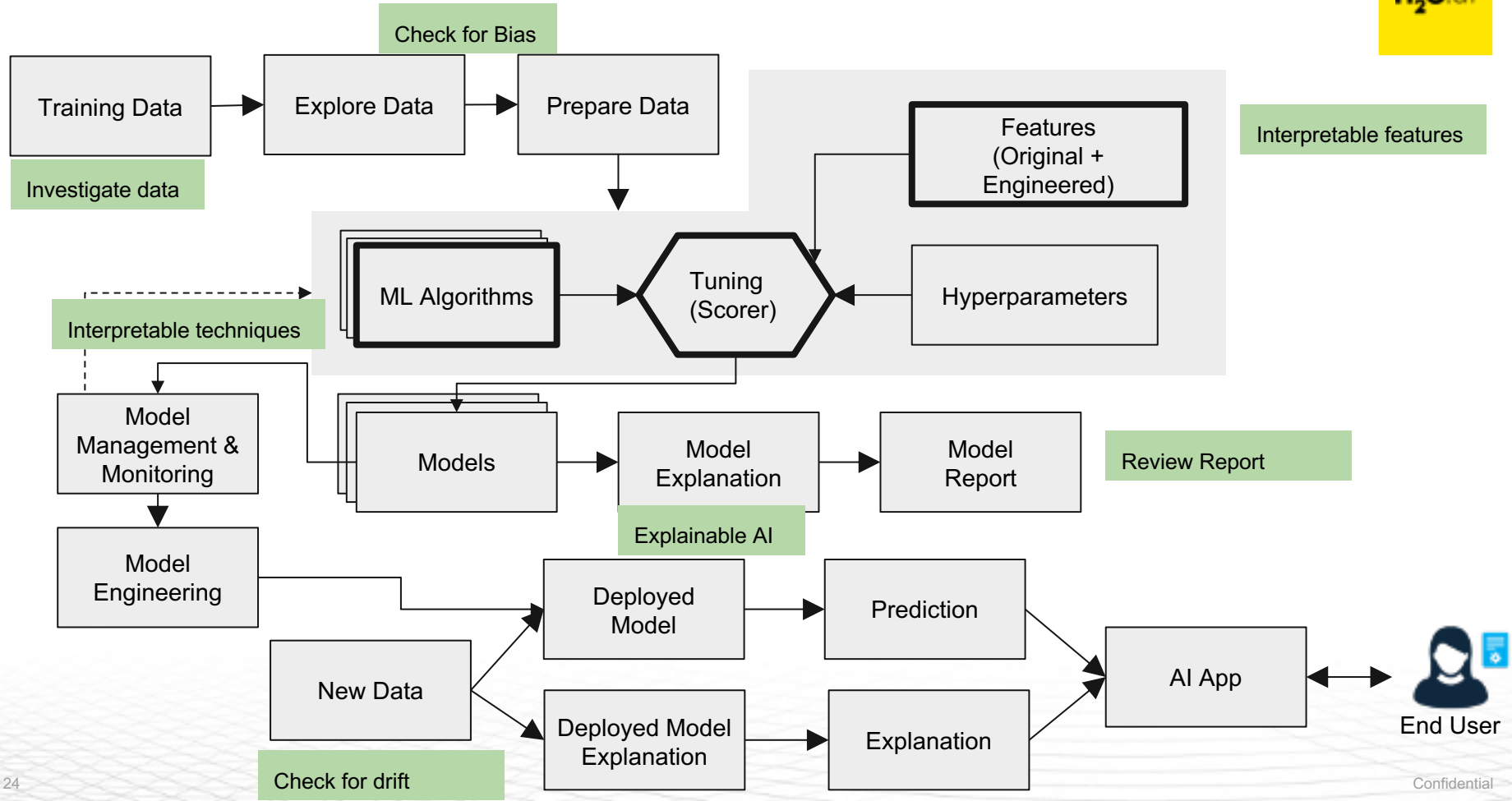
**No code rewrite going from single node to distributed cluster model training**

**Cutting edge supervised and unsupervised ML algorithms with consistent API in Python and R**

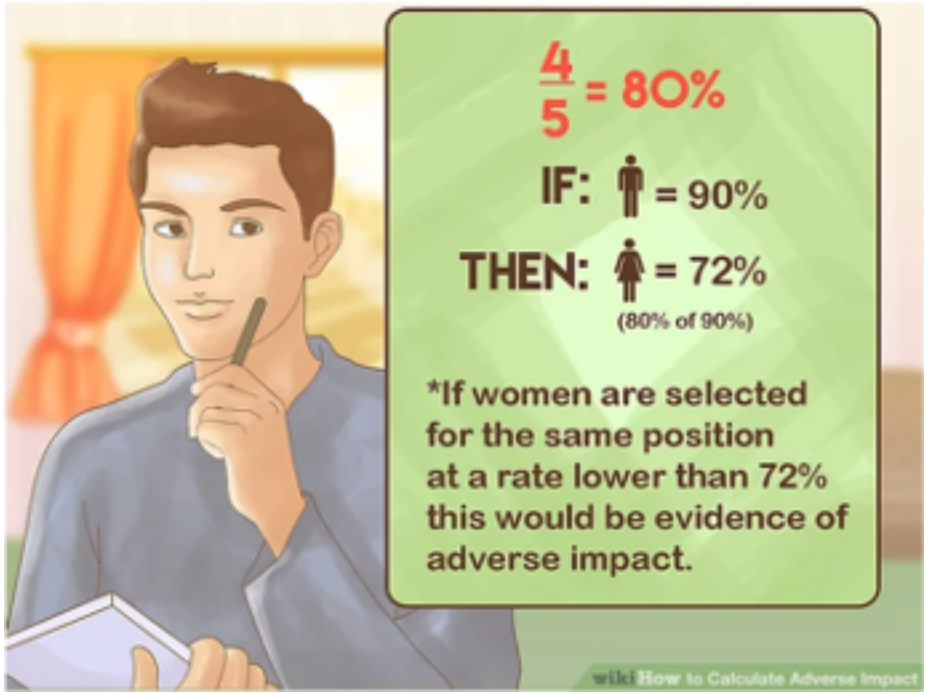
**Easy model (MOJO) deployment**


**Automated Machine Learning**

# Modified ML Workflow



# Disparate/ Adverse Impact: Demo




Assesses fairness across protected groups

https://www.wikihow.com/Calculate-Adverse-Impact

Slide borrowed with permission from Erin LeDell, Chief ML Scientist & H2O AutoML lead

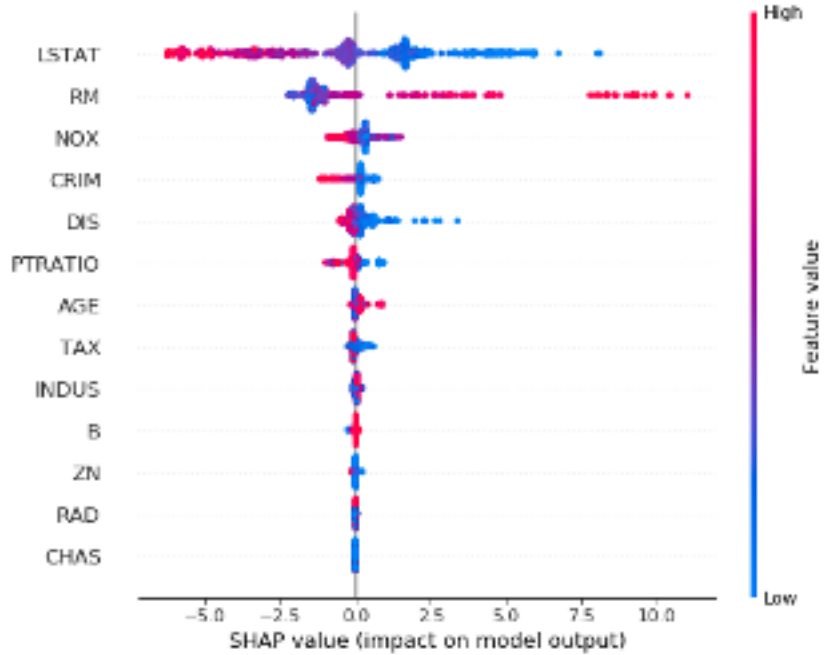
# H2O Monotonicity Constraint

Interpretable Models

```
In [55]: xgb_mono = H2OXGBoostEstimator(monotone_constraints=monotone_constraints)
         xgb_mono.train(x=feature_names, y="target", training_frame=train, validation_frame=test)

         xgboost Model Build progress: |████████████████████████████████| 100%

In [56]: xgb_mono.model_performance()

         ModelMetricsRegression: xgboost
         ** Reported on train data. **

         MSE: 0.450582394315
         RMSE: 0.671254343983
         MAE: 0.493647644764
         RMSLE: 0.217638157557
         Mean Residual Deviance: 0.450582394315
```

# H2O shap values

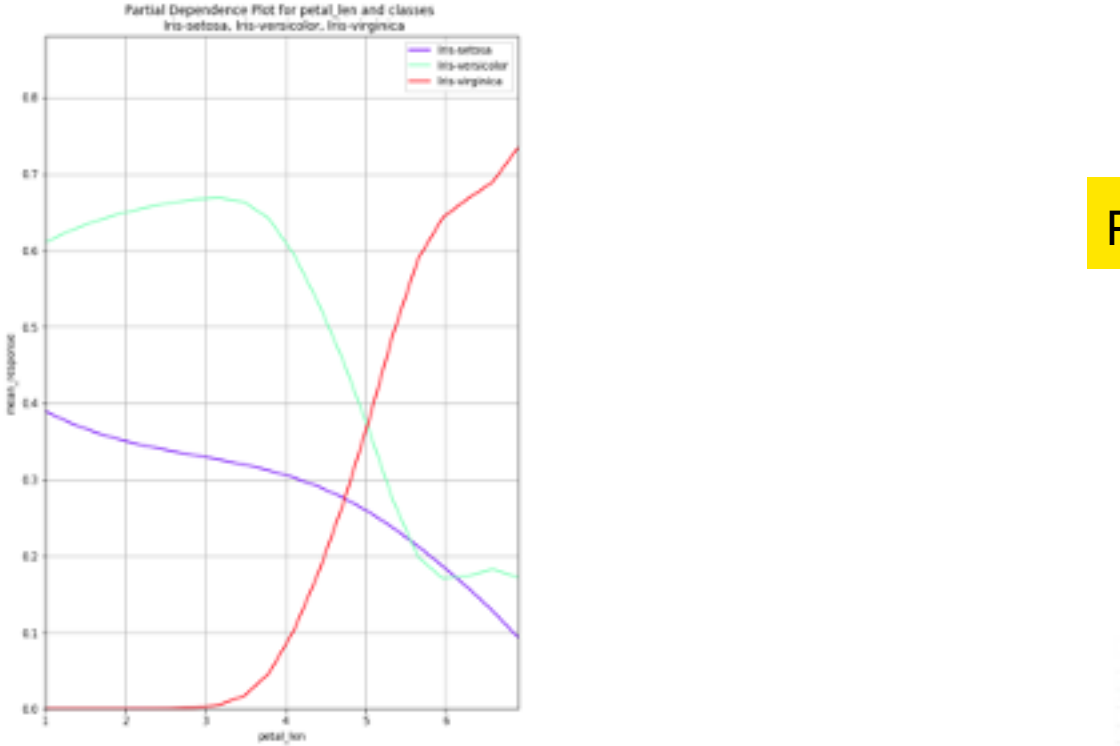
```
In [8]: # summarize the effects of all the features
        shap.summary_plot(shap_values, X)
```




Post-hoc Explanations

# H2O Partial Dependence

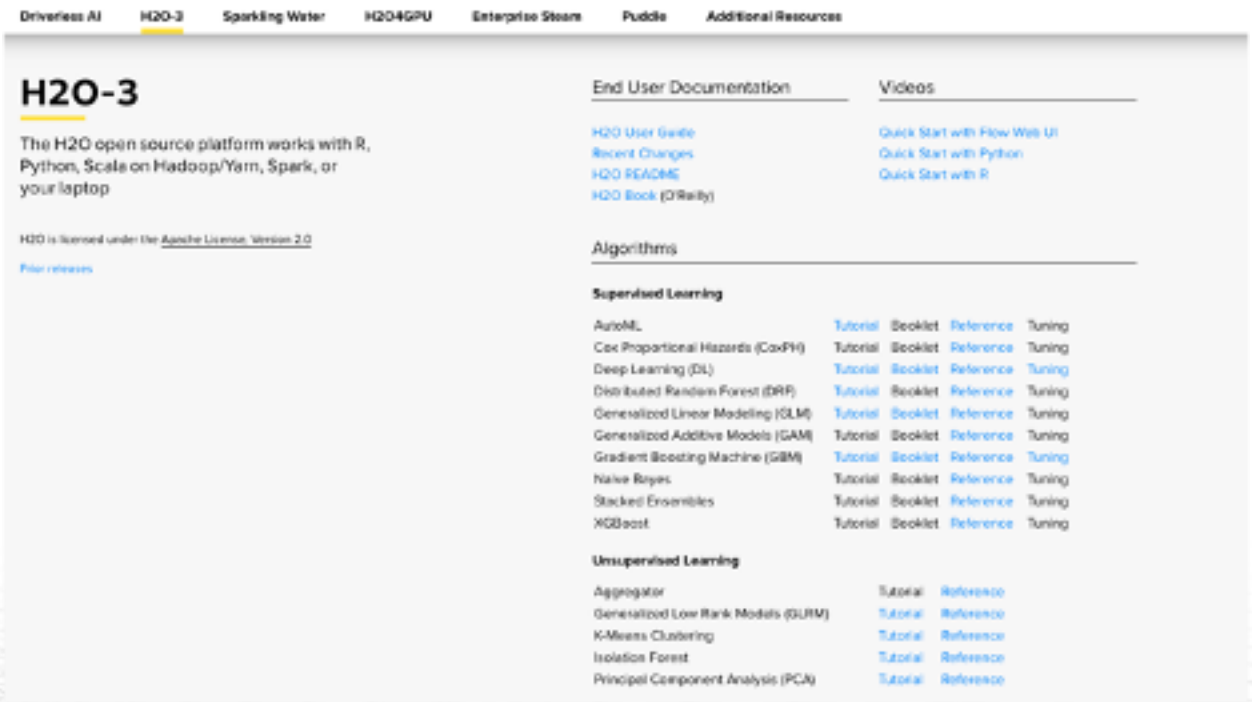
```
In [14]: # h2o multinomial PDP all classes
data = model.partial_plot(data=iris, cols=["petal_len"], plot_stddev=False, plot=True, targets=["Iris-setosa", "Iris-versicolo
r", "Iris-virginica"])
```




Post-hoc Explanations

# H2O Open Source Reference

http://docs.h2o.ai/




# Limitations and practical challenges

# Limitations and challenges

Technical challenges

- Active area of research
- Posthoc explanations are compute intensive
- Accuracy vs Fairness
- Simplified model documentation

Business challenges

- Explainability and impact analysis which business users can understand
- Lack of governance and controls
- Shortage of good commercial and opensource XAI tools (changing fast)

# Resources

# Interpretable Methods

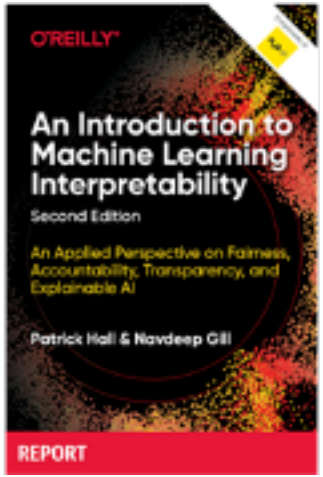
## Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

*Cynthia Rudin*
*Duke University*
*cynthia@cs.duke.edu*

**Abstract**

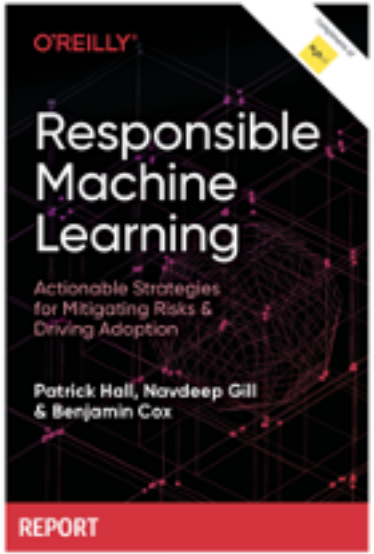
Black box machine learning models are currently being used for high stakes decision-making throughout society, causing problems throughout healthcare, criminal justice, and in other domains. People have hoped that creating methods for explaining these black box models will alleviate some of these problems, but trying to *explain* black box models, rather than creating models that are *interpretable* in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society. There is a way forward – it is to design models that are inherently interpretable. This manuscript clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare, and computer vision.

# H2O Resources




- Trust
- Post-hoc explanation
- Interpretable Models

# H2O Resources




- **People: Humans in the Loop** — Why an organization's ML culture is an important aspect of responsible ML practice
- **Processes: Taming the Wild West of Machine Learning Workflows** — Suggestions for changing or updating your processes to govern ML assets
- **Technology: Engineering ML for Human Trust and Understanding** — Tools that can help organizations build human trust and understanding into their ML systems
- **Actionable Responsible ML Guidance** — Core considerations for companies that want to drive value from ML

# Not everyone loves data

## Explainable AI: Beware of Inmates Running the Asylum

**Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences**

**Tim Miller*** and **Piers Howe†** and **Liz Sonenberg***
*School of Computing and Information Systems
†Melbourne School of Psychological Sciences
University of Melbourne, Australia
{tmiller,pdhowe,l.sonenberg}@unimelb.edu.au

# H2O Resources

## Machine Learning: Considerations for Fairly and Transparently Expanding Access to Credit

# Thank You

**We have to ask what is lost, who is harmed, and what should be forgotten with the embrace of artificial intelligence in decision making.**

Noble, Safiya Umoja. Algorithms of Oppression, 2018. NYU Press.