Frameworks for model risk management of AI

Jos Gheerardyn, Yields.io, November 2020

www.yields.io

Agenda

- Model risk components
  - Overview of market practice
  - Technological evolutions

- Adapting for AI
  - Typical ML model dependencies
  - Frameworks
    - designing AI-safety
    - assessment list for trustworthy AI
    - quantitative tests
  - Design considerations
  - Limitations

© Yields NV

MRM crash course

Definition of a model

From the definition of the **FED***:

“The term model refers to a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates.”

This is a very broad definition.

Examples

- A valuation model
- A fraud detection algorithm
- A chatbot
- A data extraction algorithm
- ...

* https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

© Yields NV

Model risk

From the definition of the **FED***:

“The use of models invariably presents model risk, which is ***the potential*** *for adverse consequences from decisions based on* ***incorrect or misused model outputs and reports****. Model risk can lead to financial loss, poor business and strategic decision-making, or damage to a banking organization’s reputation. Model risk occurs primarily for two reasons:*

1. *a model may have* ***fundamental errors*** *and produce inaccurate outputs when viewed against its design objective and intended business uses;*
2. *a model may be* ***used incorrectly or inappropriately*** *or there may be a misunderstanding about its limitations and assumptions.”*

* https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

© Yields NV

MRM framework

To manage this risk, an organization has to build a **model risk management framework**, which prescribes how this risk is going to be managed.

The framework provides an exhaustive description of four pillars:

1. Model definition
2. Model governance
3. Model validation
4. Model monitoring

© Yields NV

Model definition

Model inventory: A database that contains key features of all models present in the organization. Key attributes per model include

- Model type
- Model owner
- Development status
- Model use
- Data sources
- Pointers to documentation and source code
- Model risk tier (1-5)

Model documentation: Written by the model developer. Contains

- Model goals, assumptions and limitations
- Description of the underlying mathematics and the algorithms used
- Description of the model selection process
- Description of data cleaning + feature generation and selection process
- Overview of performed tests

© Yields NV

Model governance

Defining the stakeholders, responsibilities and business processes.

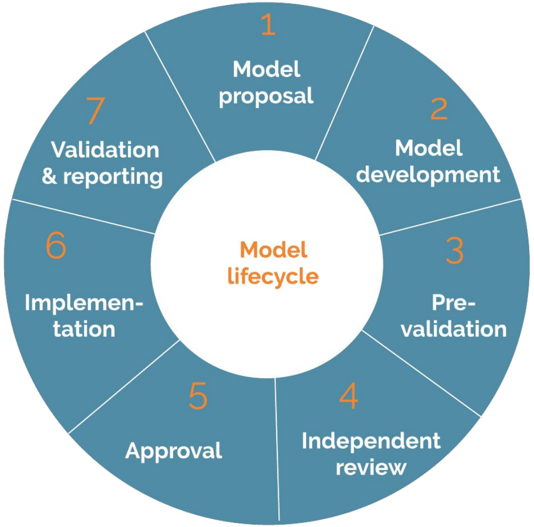
Main **stakeholders**:

1st line of defence
- model owner: responsible that the model is properly developer, maintained and used
- model developer
- model user

2nd LOD
- model validator

3rd LOD
- audit

Main **process**: the model lifecycle



Model lifecycle

1 Model proposal

2 Model development

3 Pre-validation

4 Independent review

5 Approval

6 Implementation

7 Validation & reporting

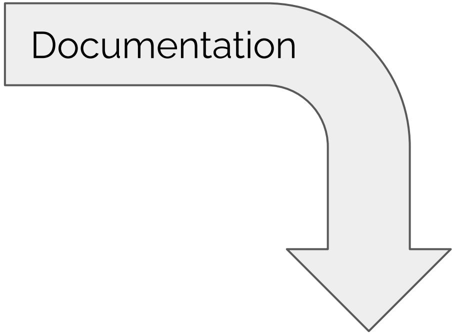
© Yields NV

Model validation

Independent validation:

- Verify that the documentation is complete
- Describe the model dependencies
- Describe the framework and assumptions
- Verify the model design and performance testing
  - Model selection
  - backtesting
  - benchmarking
  - sensitivity testing
  - model uncertainty
- Determine the limitations
- List the challenges to the first line

Note: breadth of validation depends on model risk tier.



Documentation

When?

- model has changed
- use case has changed
- too much time has past since previous validation

© Yields NV

Model monitoring

Frequent/continuous analysis of the model to detect issues quickly

- Monitor data quality
- Monitor model performance

Best practice suggests to determine thresholds (during validation) which would trigger alarms. E.g. when performance KPI drops below a given value.

There should be a process for dealing with model issues that have been discovered during model monitoring.

© Yields NV

Adapting for ML/AI

Model risk evolves

WEEKEND JOURNAL

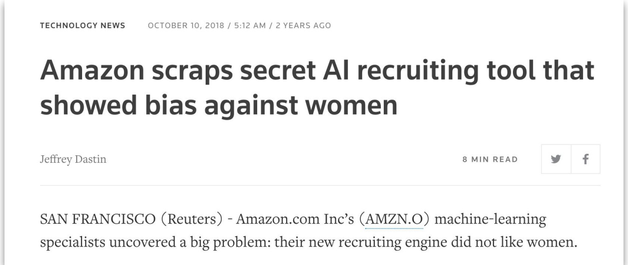
At Philippe Starck's Parisian Table

THE WALL STREET JOURNAL.

'London Whale' Warned About Risk of Trades

U.S. Sues to Block Merger Of AB InBev and Modelo

European CEOs Warn Rising Euro Is Threat

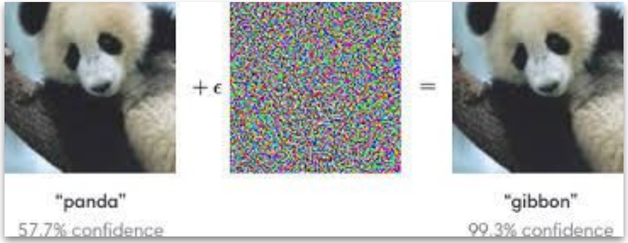


TECHNOLOGY NEWS OCTOBER 10, 2018 / 5:12 AM / 2 YEARS AGO

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.



1. Models evolve faster
2. Larger datasets & heavier computations
3. MRM emphasis shifts to the **first line**

© Yields NV

Lael Brainard - FED - Nov 2018

Our existing regulatory and supervisory guardrails are a good place to start as we assess the appropriate approach for AI processes.

The National Science and Technology Council, in an extensive study addressing regulatory activity generally, concludes that if an AI-related risk "falls within the bounds of an existing regulatory regime, . . . **the policy discussion should start by considering whether the existing regulations already adequately address the risk**, or whether they need to be adapted to the addition of AI."

A recent report by the U.S. Department of the Treasury reaches a similar conclusion with regard to financial services.

© Yields NV

See https://www.federalreserve.gov/newsevents/speech/brainard20181113a.htm

Structure of a MRM

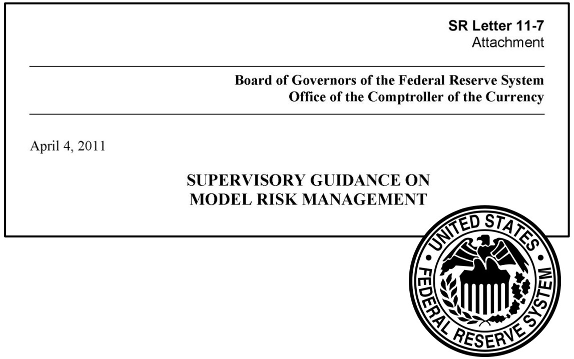
Some highlights:

1. Model dependencies

2. Framework and assumptions

3. Model design and performance testing
   Model selection, backtesting, benchmarking, sensitivity testing, model uncertainty

4. Limitations



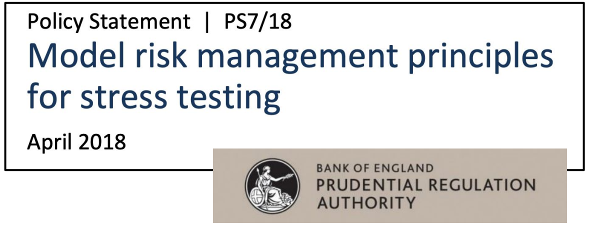
SR Letter 11-7
Attachment

Board of Governors of the Federal Reserve System
Office of the Comptroller of the Currency

April 4, 2011

SUPERVISORY GUIDANCE ON
MODEL RISK MANAGEMENT

UNITED STATES
FEDERAL RESERVE SYSTEM



Policy Statement | PS7/18

Model risk management principles for stress testing

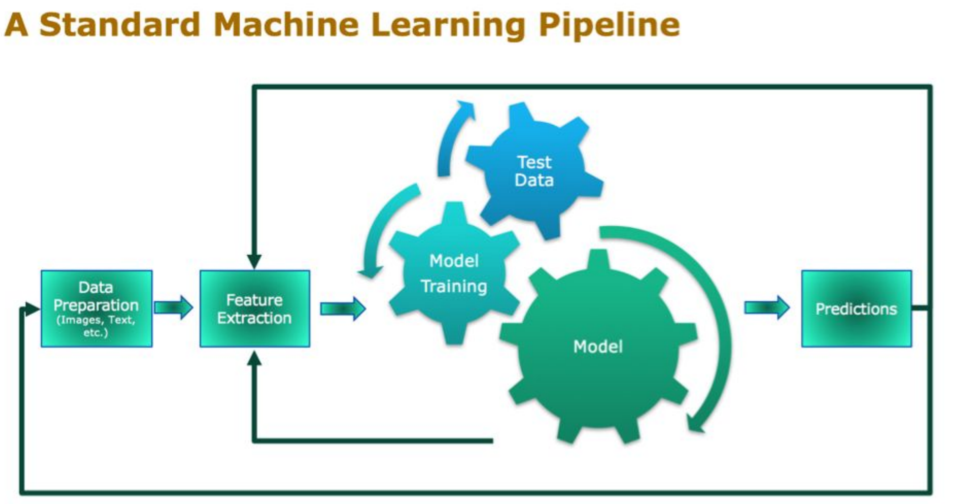
April 2018

BANK OF ENGLAND
PRUDENTIAL REGULATION AUTHORITY

© Yields NV

Model dependencies

Model dependencies


A Standard Machine Learning Pipeline
Data Preparation (Images, Text, etc.)
Feature Extraction
Test Data
Model Training
Model
Predictions

A typical pipeline consists of many non-trivial models

- Data cleaning
- Feature engineering
- Training (optimization algorithm)
- Actual model

© Yields NV

AI model risk frameworks

Frameworks - AI safety

Five principles, originally formulated in a paper by Stanford U, UC Berkeley, Google Brain and Open AI[*]

1. Avoid negative side effects

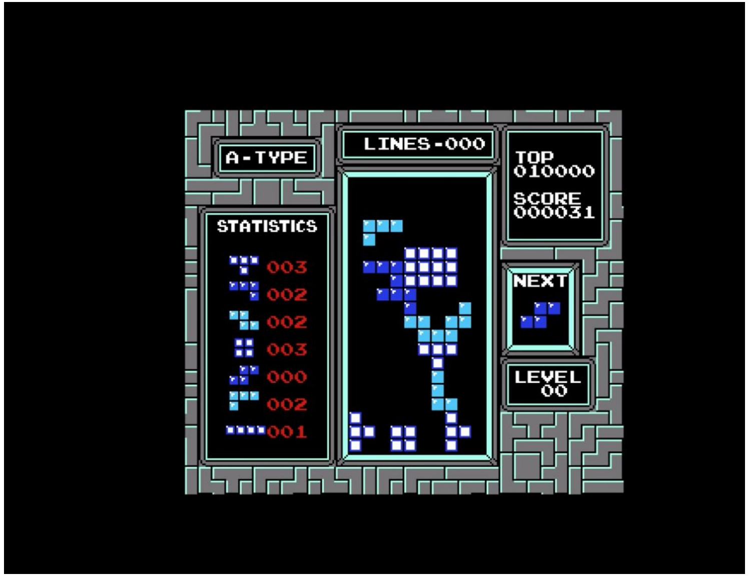
E.g. the AI tries to trigger defaults when loan margin turns negative.

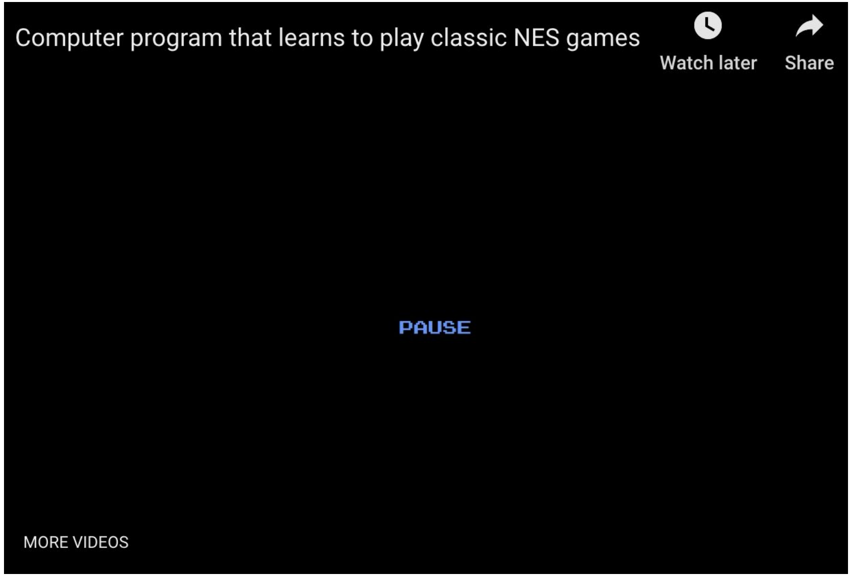
© Yields NV

[*] See https://arxiv.org/abs/1606.06565

Frameworks - AI safety

2. Reward hacking




Computer program that learns to play classic NES games

© Yields NV

Frameworks - AI safety

3. Scalable oversight

If atypical collateral (such as artwork) is encountered, the AI has to reach out to experts if there is too much uncertainty.

© Yields NV

Frameworks - AI safety

4. Safe exploration

E.g. AI tries to re-introduce Ninja loans to learn about new possible client segments

© Yields NV

Frameworks - AI safety

5. Robustness against distributional shifts

This can be very subtle: “Adversarial examples”*
An example of data hacking

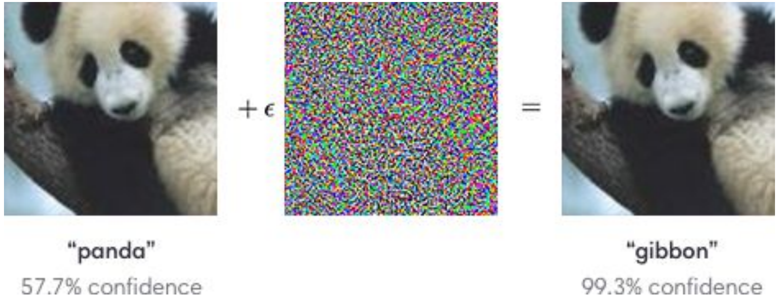
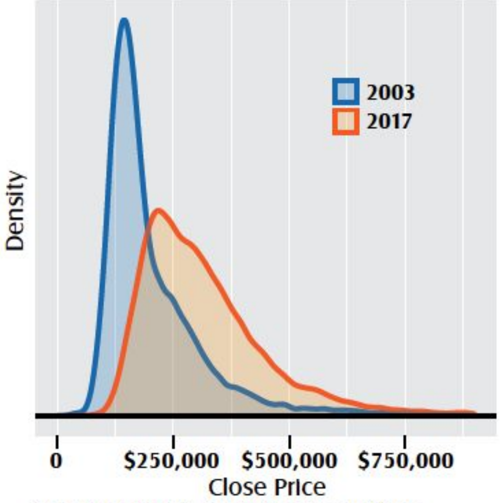


Figure 1. Texas Single-Family New-Home Sales Distribution (by Year)



Note: The probability density functions specify the probability that home sales occur within a particular range of values. The probability is measured by the area under the curve within the range (e.g., within $190,000 and $250,000).

Source: Real Estate Center at Texas A&M University

© Yields NV

* See https://arxiv.org/abs/1312.6199

Frameworks - ALTAI*

Example: Trustworthy AI

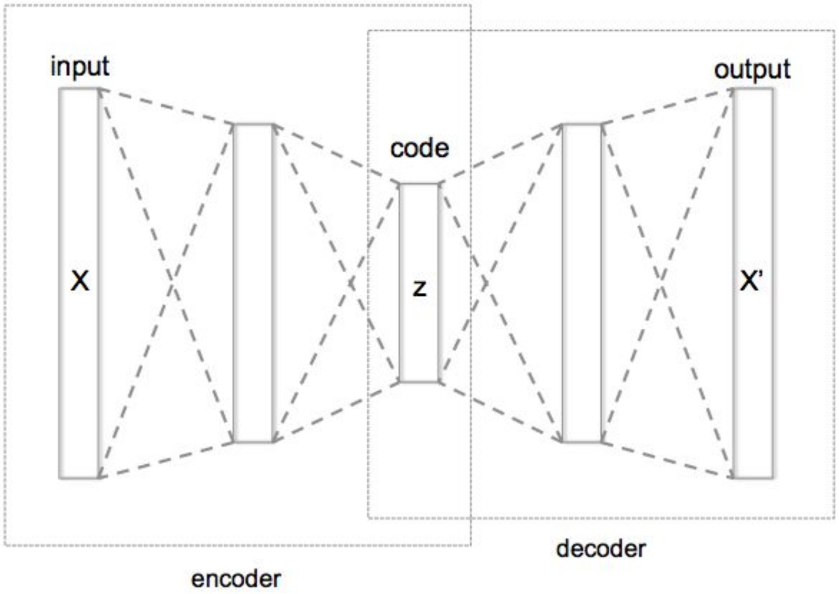
1. human agency and oversight
2. technical robustness and safety
3. privacy and data governance
4. transparency
5. diversity, non-discrimination and fairness
6. environmental and societal well-being
7. accountability

Measure
if your organisation's AI is
trustworthy
AI
ALTAI – Assessment List for Trustworthy Artifical Intelligence

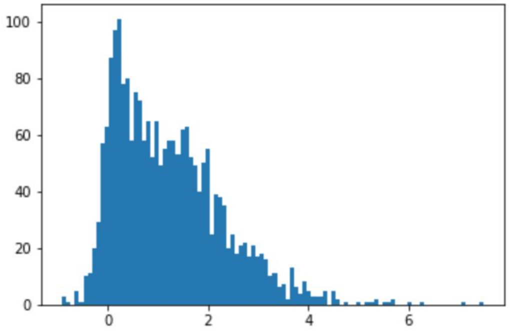
Qualitative and quantitative assessments

* See https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Autoencoder



- Minimize reconstruction error |x - x`|
- Linear autoencoder = PCA*
- Fast
- Non-linear activation



* See e.g. https://www.cs.toronto.edu/~urtasun/courses/CSC411/14_pca.pdf

Frameworks - ALTAI

technical robustness and safety

Data poisoning: check robustness of model performance relative to data quality

1. train auto-encoder
2. assign novelty score to each datapoint
3. measure model performance as a function of the novelty score

Robustness against adversarial examples*

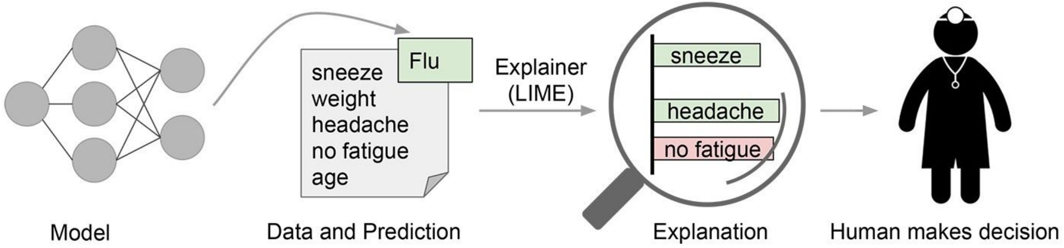
1. Measure performance loss against adversarial directions: $x_2 = x_1 + \epsilon sign(\nabla_x J)$
2. Compare loss with (low-dimensional) benchmark model

privacy

Measure disclosure risk (fraction of uniquely identifiable samples, given the value of a set of attributes)

* see https://arxiv.org/pdf/1412.6572.pdf

© Yields NV



Frameworks - ALTAI

$$\xi(x) = \underset{g\in G}{\operatorname{argmin}} \;\; \mathcal{L}(f,g,\pi_x) + \Omega(g)$$

$$\mathcal{L}(f,g,\pi_x) = \sum_{z,z'\in\mathcal{Z}} \pi_x(z)\left(f(z)-g(z')\right)^2$$

Transparency

local explainability

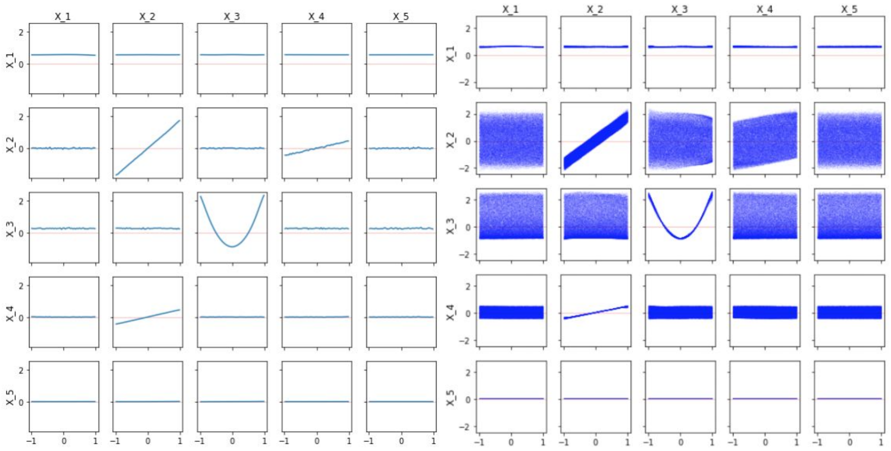
- e.g. LIME* = Local interpretable model-agnostic explanations

global explainability

- e.g. LE plots**

$$f_j^{LE}(x_j) = E_{\mathbf{X}_{-j}|X_j}\{f_j^1(X_j,\mathbf{X}_{-j})|X_j = x_j\},$$

$$f_{k,j}^{LE}(x_j) = E_{\mathbf{X}_{-j}|X_j}\{f_k^1(X_k,\mathbf{X}_{-k})|X_j = x_j\}$$



* See https://arxiv.org/pdf/1602.04938.pdf

** See https://arxiv.org/pdf/1808.07216.pdf

© Yields NV

Frameworks - ALTAI

fairness / bias

Language is a **protected** attribute.

How to create an unbiased credit model?

1. **Unawareness**
   But redundant encodings

2. **Demographic parity**
   But different PD

3. **Equalized odds**

See Hardt, Price & Srebro, Equality of Opportunity in Supervised Learning, https://arxiv.org/pdf/1610.02413.pdf

© Yields NV

Strictly confidential

Frameworks - ALTAI

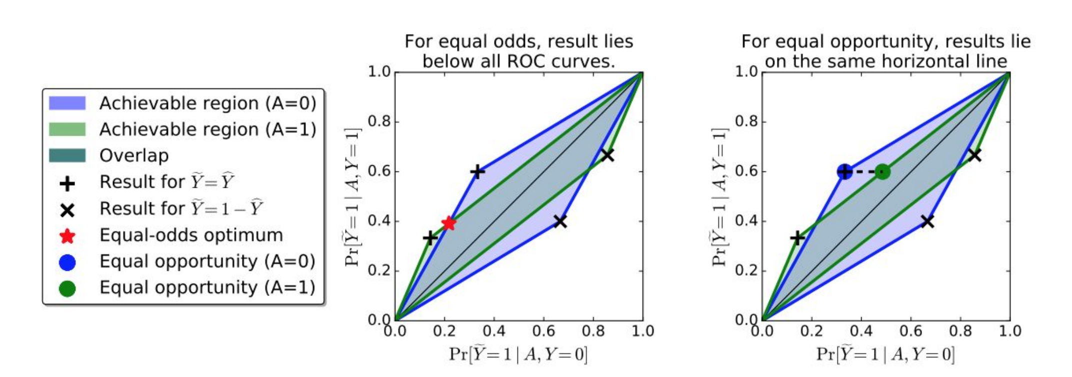
fairness / bias

1. Determine what are the protected attributes

2. Train the model on the full dataset

3. Measure the bias statistically

4. Correct the model output if needed



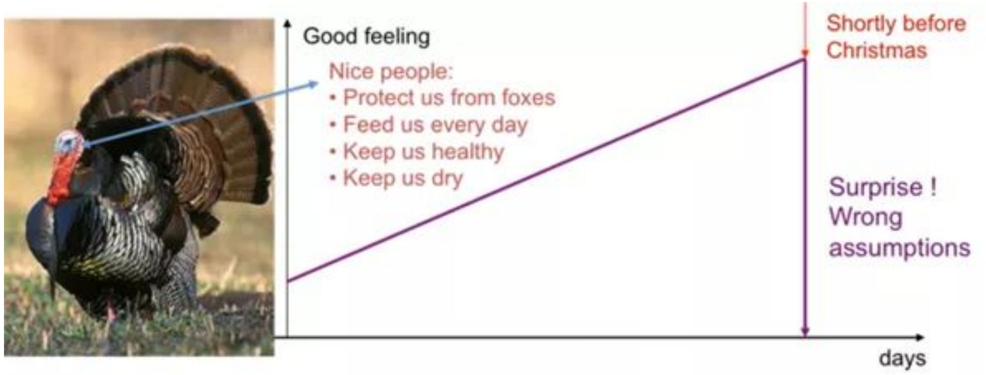
© Yields NV

Design and assumptions

Model assumptions

Use case I: Inference

The past is representative of the future
Test for distributional shifts

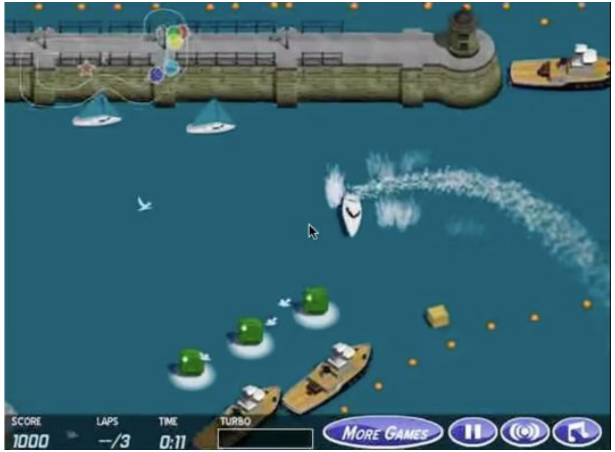


© Yields NV

Model assumptions

Use case II: Reinforcement learning

Are the rules correct?
Check consistency with underlying assumptions



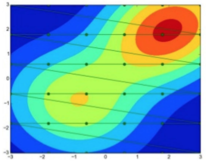
© Yields NV

Model selection

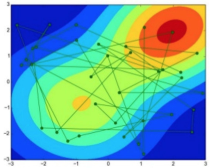
Hyperparameter “de-tuning”

ML in general and (deep) neural network algorithms have many degrees of freedom

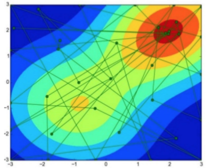
- Number of layers, number of nodes and connections
- Activation functions
- Learning rates
- Etc.



Grid



Random



SMAC*

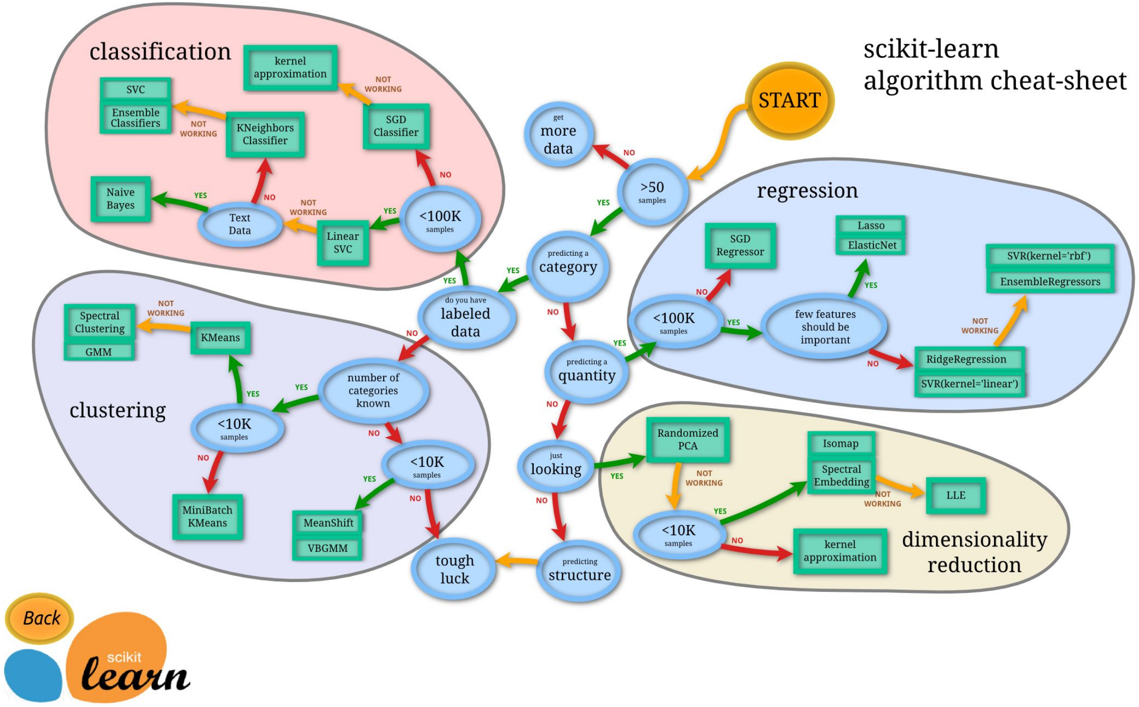
Genetic programming**

© Yields NV

* See https://www.cs.ubc.ca/~hutter/papers/10-TR-SMAC.pdf

** See https://github.com/EpistasisLab/tpot

Limitations



Conclusion

Introducing AI can have considerable **benefits**

But also introduces **risk**

To **decide** when to use AI

- Measure the added value
- And base your decision on the risk appetite
- Defined via an updated model risk management framework

© Yields NV

Thank you!

Yields NV

Parvis Sainte-Gudule 5,
1000 Brussels
Belgium

+32 479 527 261

info@yields.io

Yields.io Ltd

8 Northumberland Ave,
Westminster, London WC2N
5BY, UK

+44 7584 068899

info@yields.io

© Yields NV