yields.io

An AI Platform for

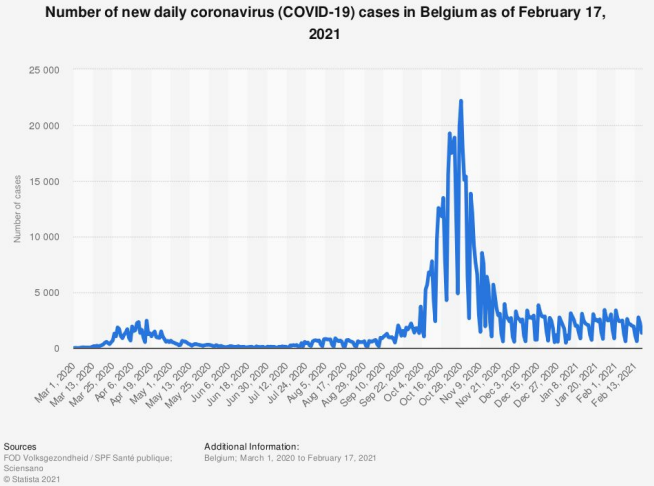
MODEL RISK MANAGEMENT


1
01
110
001
10010011001
0110011101001
0110101101001
00101001000111010010
01
0010010

Data use

Displaying data to discover obvious patterns

BI tools

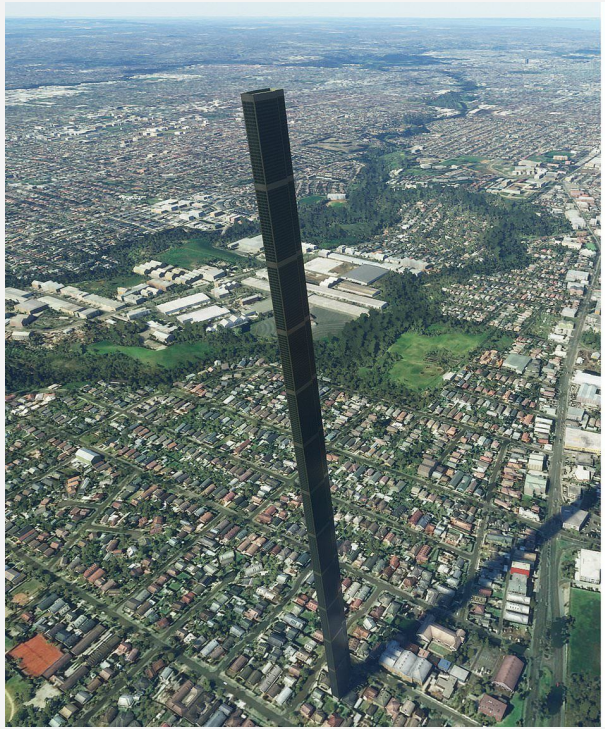


Infer from data to discover hidden patterns

Algorithms

Incomplete data

Anomalies



Many more examples

Specification gaming examples in AI - master list : Sheet1

	Submit more examples through this Google form:	*https://docs.google.com/*	*More information in this blog post:*	*https://medium.com/@d*		
Title	**Description**	**Authors**	**Original source**	**Original source link**	**Video / Image**	**Source / Credit**
Aircraft landing	Evolved algorithm for landing aircraft exploited overflow errors in the physics simulator by creating large forces that were estimated to be zero, resulting in a perfect score	Feldt, 1998	Generating diverse software versions with genetic programming: An experimental study.	http://ieeexplore.ieee.or		Lehman et al, 2018
Bicycle	Reward-shaping a bicycle agent for not falling over & making progress towards a goal point (but not punishing for moving away) leads it to learn to circle around the goal in a physically stable loop.	Randlov & Alstrom, 1998	Learning to Drive a Bicycle using Reinforcement Learning and Shaping	https://pdfs.semanticscl		Gwern Branwen
Block moving	A robotic arm trained using hindsight experience replay to slide a block to a target position on a table achieves the goal by moving the table itself.	Chopra, 2018	GitHub issue for OpenAI gym environment FetchPush-v0	https://github.com/open		Matthew Rahtz
Boat race	Reinforcement learning agent goes in a circle hitting the same targets instead of finishing the race	Amodei & Clark, 2016	Faulty reward functions in the wild	https://blog.openai.com	https://www.youtu	
Classifiers	A task is specified by using a set of goal images and training a classifier to distinguish goal from non-goal images, with the success probabilities from the classifier used as task reward. "In this task, the goal is to push the green object onto the red marker. While the classifier outputs a success probability of 1.0, the robot does not solve the task. The RL algorithm has managed to exploit the classifier by moving the robot arm in a peculiar way, since the classifier was not trained on this specific kind of negative examples."	Singh, 2019	End-to-End Deep Reinforcement Learning without Reward Engineering	https://bair.berkeley.edu	https://bair.berkel	Jan Leike
Ceiling	A genetic algorithm was instructed to try and make a creature stick to the ceiling for as long as possible. It was scored with the average height of the creature during the run. Instead of sticking to the ceiling, the creature found a bug in the physics engine to snap out of bounds.	Higueras, 2015	Genetic Algorithm Physics Exploiting	https://youtu.be/ppf3Vq	https://youtu.be/p	Jesús Higueras
CycleGAN steganography	CycleGAN algorithm for converting aerial photographs into street maps and back steganographically encoded output information in the intermediary image without it being humanly detectable.	Chu et al, 2017	CycleGAN, a Master of Steganography	https://arxiv.org/abs/17		Tech Crunch / Gwern Brar
Data order patterns	Neural nets evolved to classify edible and poisonous mushrooms took advantage of the data being presented in alternating order, and didn't actually learn any features of the input images	Ellefsen et al, 2015	Neural modularity helps organisms evolve to learn new skills without forgetting old skills	http://journals.plos.org/p		Lehman et al, 2018
Dying to Teleport	PlayFun algorithm deliberately dies in the Bubble Bobble game as a way to teleport to the respawn location	Murphy, 2013	The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel	http://www.cs.cmu.edu/		Alex Meiburg
Eurisko - authorship	Game-playing agent accrues points by falsely inserting its name as the author of high-value items	Johnson, 1984	Eurisko, The Computer With A Mind Of Its Own	http://aliciapatterson.org		Catherine Olsson / Stuart Armstrong
Eurisko - fleet	Eurisko won the Trillion Credit Squadron (TCS) competition two years in a row creating fleets that exploited loopholes in the game's rules, e.g. by spending the trillion credits on creating a very large number of stationary and defenseless ships	Lenat, 1983	Eurisko, The Computer With A Mind Of Its Own	http://aliciapatterson.org		Haym Hirsh

* https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml

SR 11-7: Guidance on Model Risk Management

BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM
WASHINGTON, D.C. 20551

DIVISION OF BANKING
SUPERVISION AND REGULATION

SR 11-7
April 4, 2011

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

SR 11-7: Guidance on Model Risk Management

BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM
WASHINGTON, D.C. 20551

DIVISION OF BANKING
SUPERVISION AND REGULATION

SR 11-7
April 4, 2011

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

MRM framework

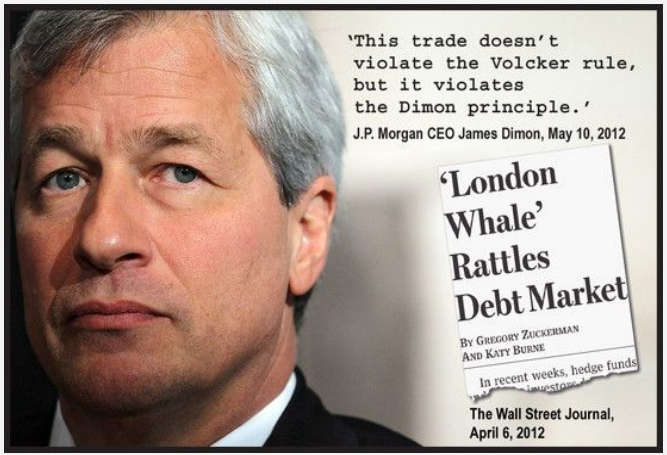
To manage this risk, an organization has to build a **model risk management framework**, which prescribes how this risk is going to be managed.

The framework provides an exhaustive description of four pillars:

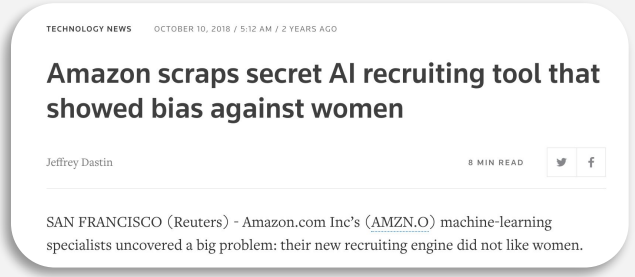
1. **Model definition**: what is a model in my organization?
2. **Model governance**: what processes do we put in place (e.g. 3 lines of defence)
3. **Model validation**: independent review (check data, perform benchmark, backtest)
4. **Model monitoring**: monitor what is running in production

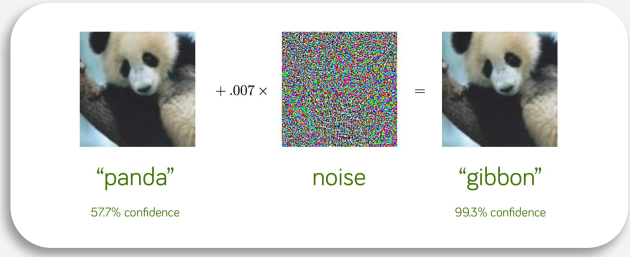
Model Risk Management Evolves

Qualitative MRM
governance

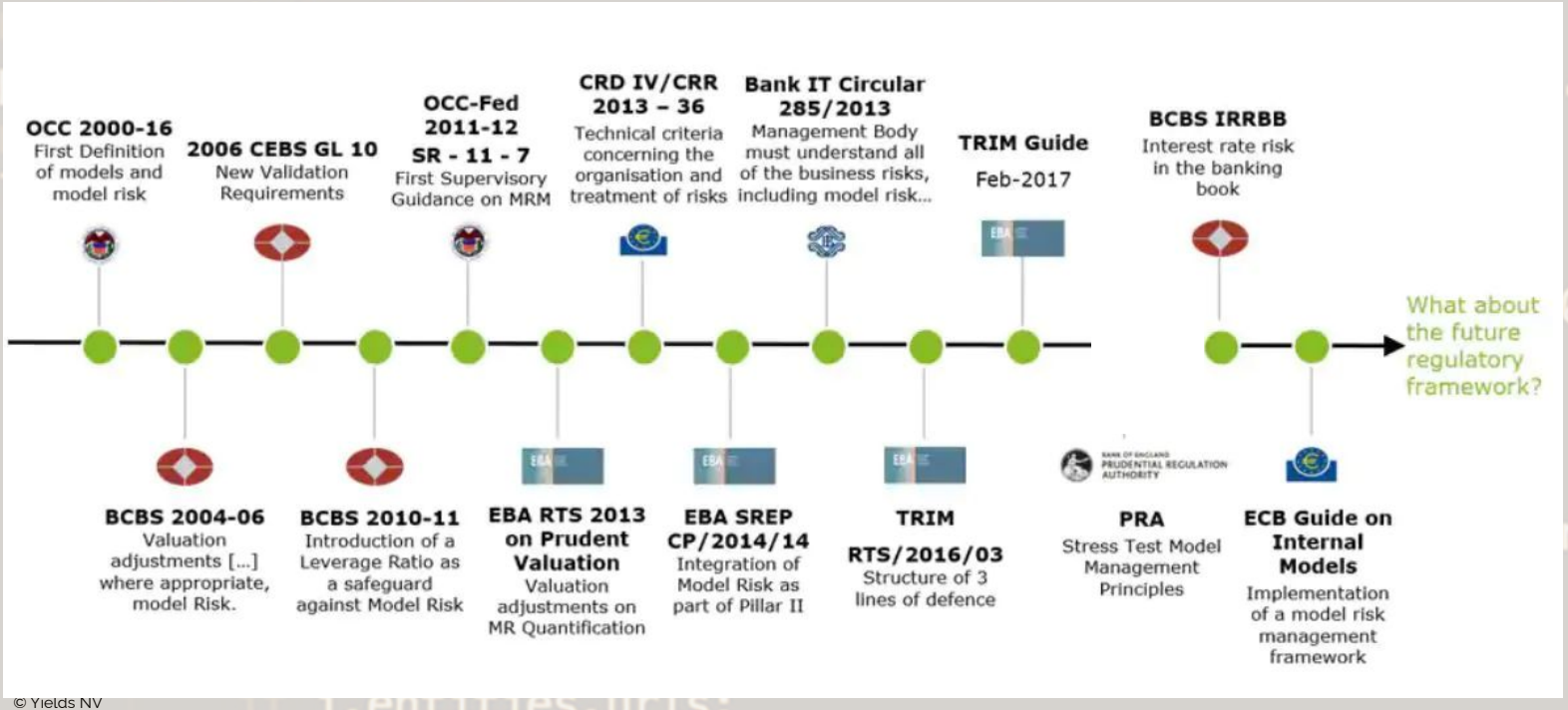


Quantitative MRM
measurements





Model Risk Management Challenges



Shortcomings of current technologies

“We’re losing track of the linkage between our Model Risk Management objects”

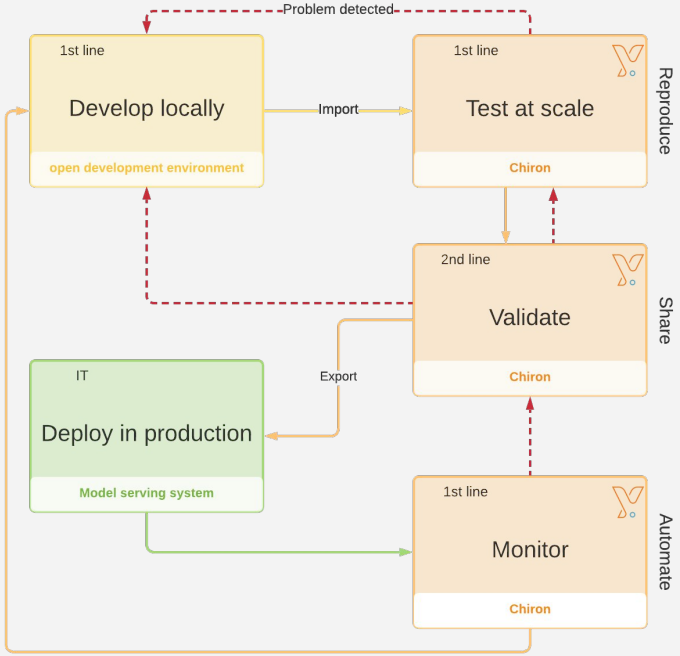
“No flexible or integrated deployment offering”

“Lack of versioning data, analytics and reports”

“We cannot monitor data and model quality evolution over time”

“No data science platform available to streamline the full Model Lifecycle”

Turn MRM into a value driver with Chiron



A Collaborative Platform

The model lifecycle is a non-sequential process with many participants. Chiron centralizes all MRM objects, and allows for organized sharing between teams. Managers have visibility over the entire process

Reproducibility

Chiron keeps track of all versions of and links between objects. This allows you to go back in time and reproduce any previous result.

Powerful Analytics

Statistical and ML techniques to assist with quantitative tasks such as data quality analysis, benchmarking and back-testing.

Immediate Access

Chiron is ready to be deployed; it is a fully Integrated yet customizable solution as a result of 540 months of cumulative development.

A European framework - 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

Qualitative and quantitative assessments

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

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:
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

Frameworks - ALTAI

fairness / bias

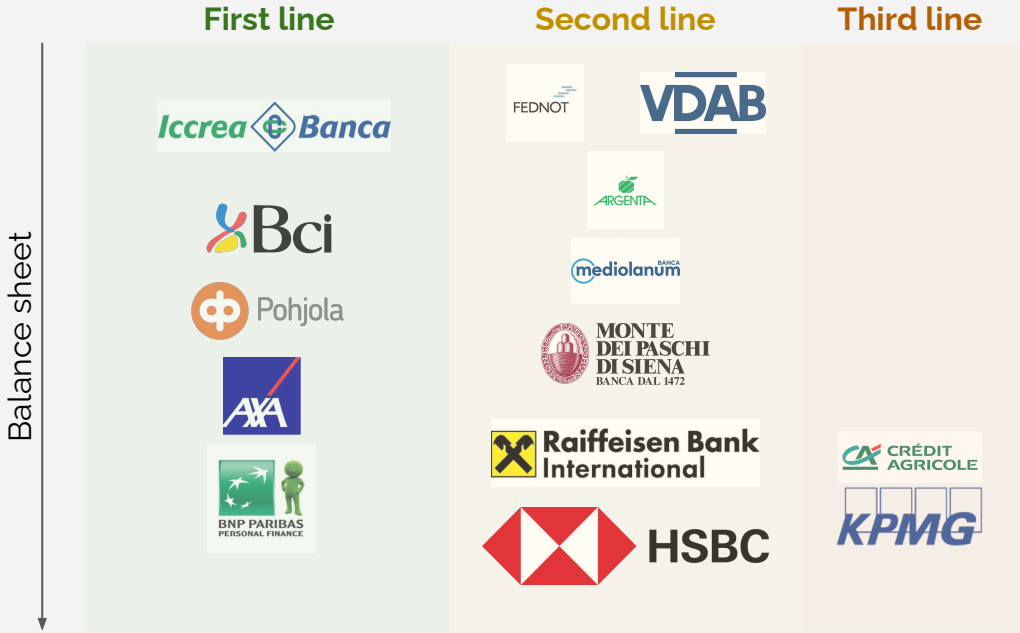
Language is a **protected** attribute.

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

Real life use cases



Risk Technology Awards 2020 Winner
Yields.io Model validation service of the year

Risk Technology Awards 2019 Winner
Yields.io Model validation service of the year

2019 FINTECH100
Leading Global Fintech Innovators

RiskTech 100 2020

RiskTech 100 2021

Yields.io

Founded in 2017

Backed by Volta,
Pamica & IMEC.iStart

Team of 20
In Brussels & London

Jos Gheerardyn
CEO/founder

Sébastien Viguié
CTO/founder

Michel Akkermans
Chairman

Peter Nowlan
Advisor

Bob Mark
Advisor

Contact

Yields NV

Brugmannlaan 63
1190 Forest
Belgium

+32 479 527 261

info@yields.io

Yields.io Ltd

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