



Introduction à l'Apprentissage Statistique

M2 Actuariat – ISFA – 2021/2022

Pierrick Piette
Actuaire à Seyna
pierrick.piette@gmail.com

Seyna.

In God we trust, all others bring data
- William Edwards Deming



Un peu d'histoire de la statistique

Discipline ancienne

- Les premiers écrits sont des recensements agricoles
- Prémices des statistiques descriptives

Etymologie

- *a priori* de l'allemand *Staatskunde*
- Littéralement « étude de l'Etat », qui perdure encore de nos jours

La statistique inférentielle arrive plus tard

- Al-Kindi, Déchiffrer les messages cryptés (~850)
- John Graunt et William Petty, Tables de mortalité à Londres (1662) → *Cimetières londiniens*.
- Développement des applications avec l'informatique depuis 1950 → *Machine de Turing*.

Etude des risques

Analyse des engagements d'un assureur

- Comprendre l'impact des caractéristiques **X** sur le risque **Y**

Base de données classique d'un assureur

- les **caractéristiques** de l'assuré *Sexe interdiction depuis 2012*
- les **options** du contrat
- les conditions de **marché**

↪ Ensuite, variable d'intérêt est la prime pure.

Les informations **X** jouent un rôle crucial dans la prévision de sinistralité **Y**

- Tarification
- Provisionnement
- Marketing Subsidence , sol argileux se remplit d'eau et ensuite il va se fracturer => fracture dans les maisons...

Généralités

Pourquoi modéliser ?

- A partir d'une série d'observations, phénomène trop complexe pour une description analytique

Objectifs en statistique

- **Explorer** : décrire les variables, leurs liaisons, etc.
- **Expliquer** : tester l'influence d'une variable dans un modèle supposé connu
- **Prévoir** : trouver le meilleur modèle dans un ensemble de prédicteur

Econométrie historique

- Modèles paramétriques avec variables explicatives + bruit
- Inférer les paramètres depuis les observations en contrôlant les propriétés de la partie aléatoire
- Méthode de statistique inférentielle, en délaissant un peu la partie prévision

Historique de l'apprentissage statistique

Modèles linéaires

- Méthode des moindres carrés : Legendre (1805) et Gauss (1821)
- Régression : Galton (1886)
- Logistic Regression : Berkson (1944)

Briques du Deep Learning

- Réseaux de neurones : McCulloch & Pitts (1943)
- Perceptron : Rosenblatt (1957)
- Convolutional Neural Network : Fukushima (1980)

Historique de l'apprentissage statistique

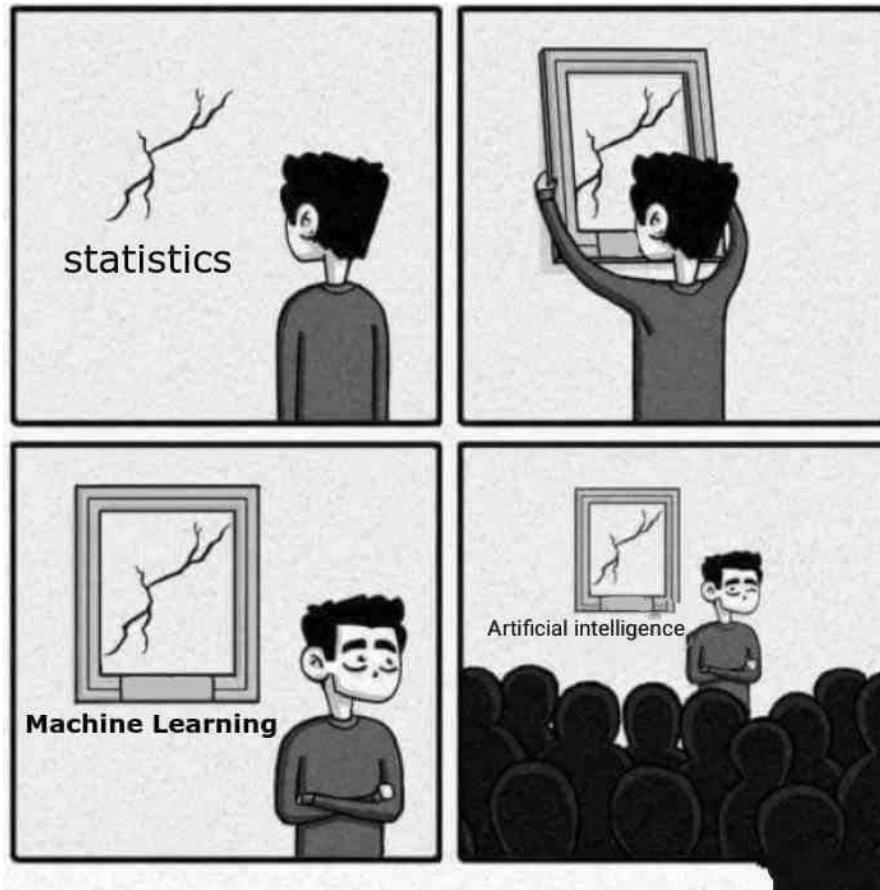
Pénalisation

- Ridge : Hoerl & Kennard (1970)
- Lasso : Tibshirani (1996)
- Elastic-Net : Zou & Hastie (2005)

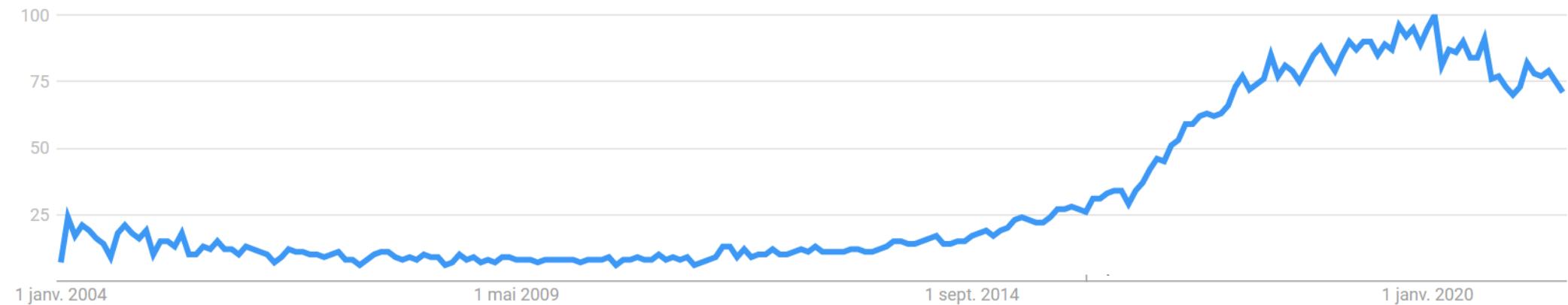
Arbres et SVM

- CART : Breiman et al. (1984)
- Boosting : Schapire (1990)
- Support Vector Machine : Boser et al. (1992) et Cortes & Vapnik (1995)
- Bagging : Breiman (1996)
- Random Forest : Breiman (2001)

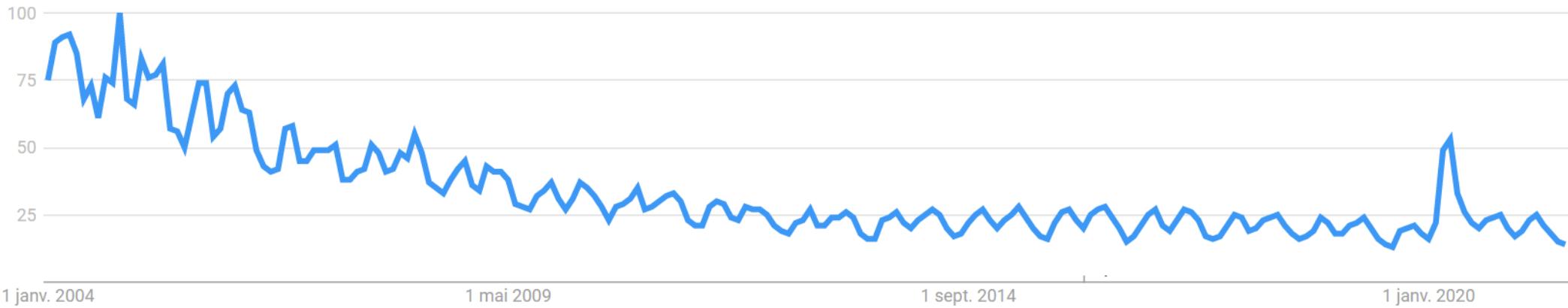
De la statistique à l'intelligence artificielle ?



Google Trend – Machine Learning



Google Trend – Statistics



Data Science

Sciences des données

- Les données sont aux centre de l'étude
- On ne parle pas de *Modelling Science*
- Le modèle arrive dans un second temps

« Garbage In, Garbage Out »

- attribuée à George Fuechsel (IBM)
- Qu'importe l'algorithme, si les données initiales sont de mauvaise qualité

Data Engineering est indispensable

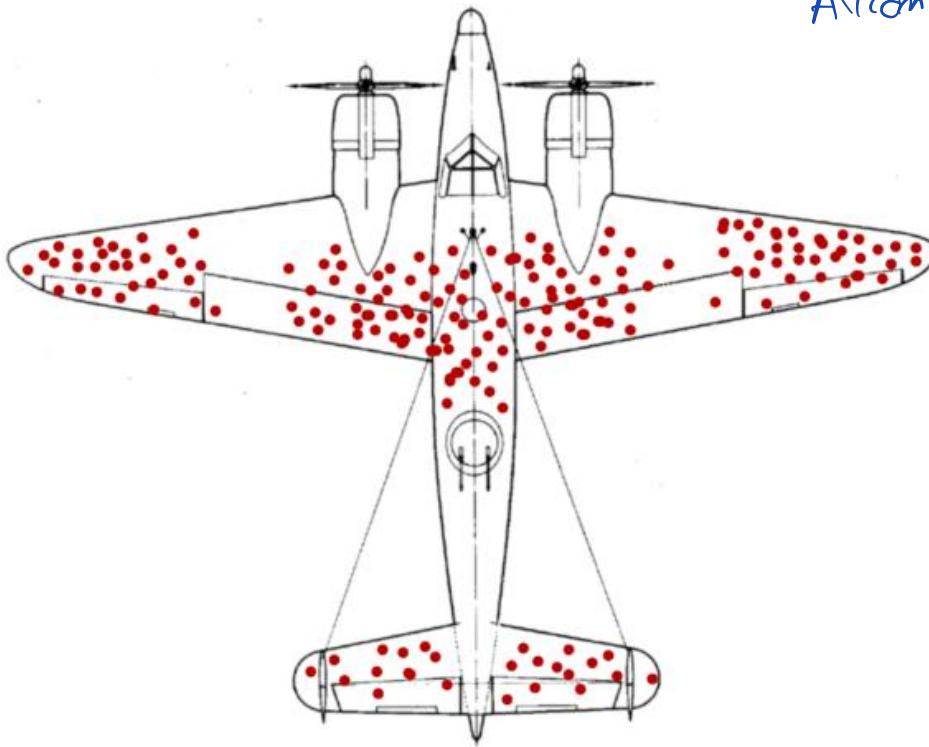
- De plus en plus séparation des métiers
- Mais besoin de comprendre tout le process



Biais de sélection

Différence entre la population observée et la population totale

Avec, parties à renforcer

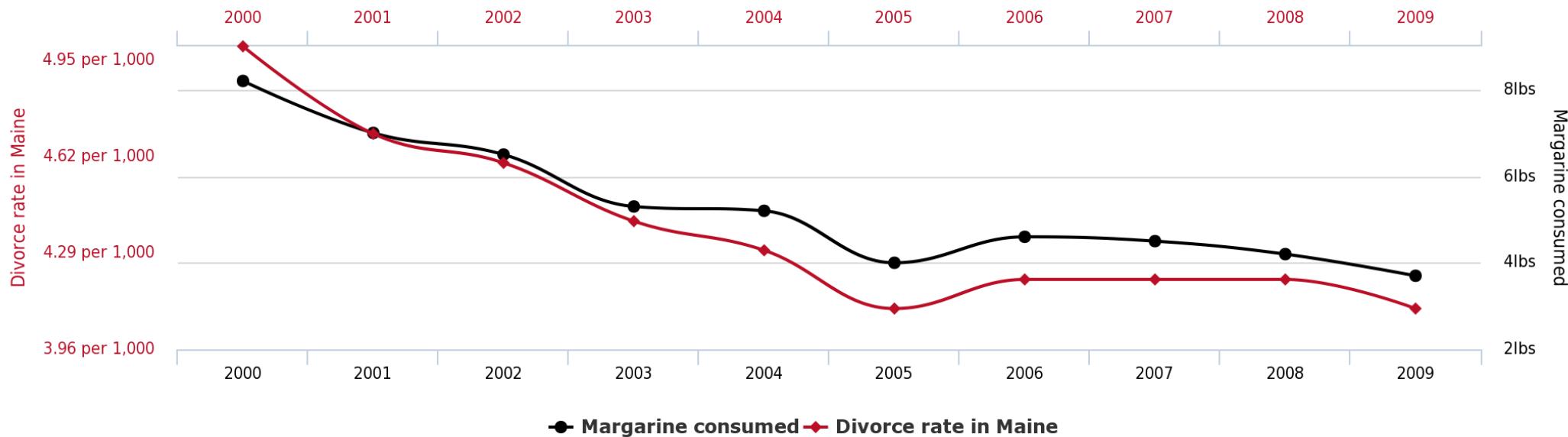


Corrélation / Causalité

Divorce rate in Maine

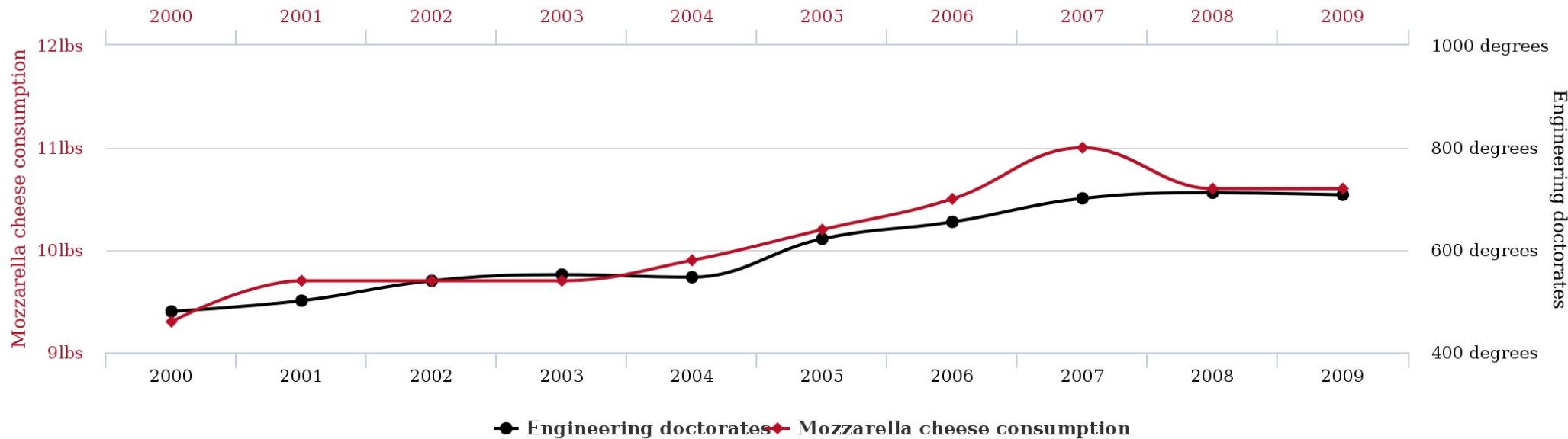
correlates with

Per capita consumption of margarine

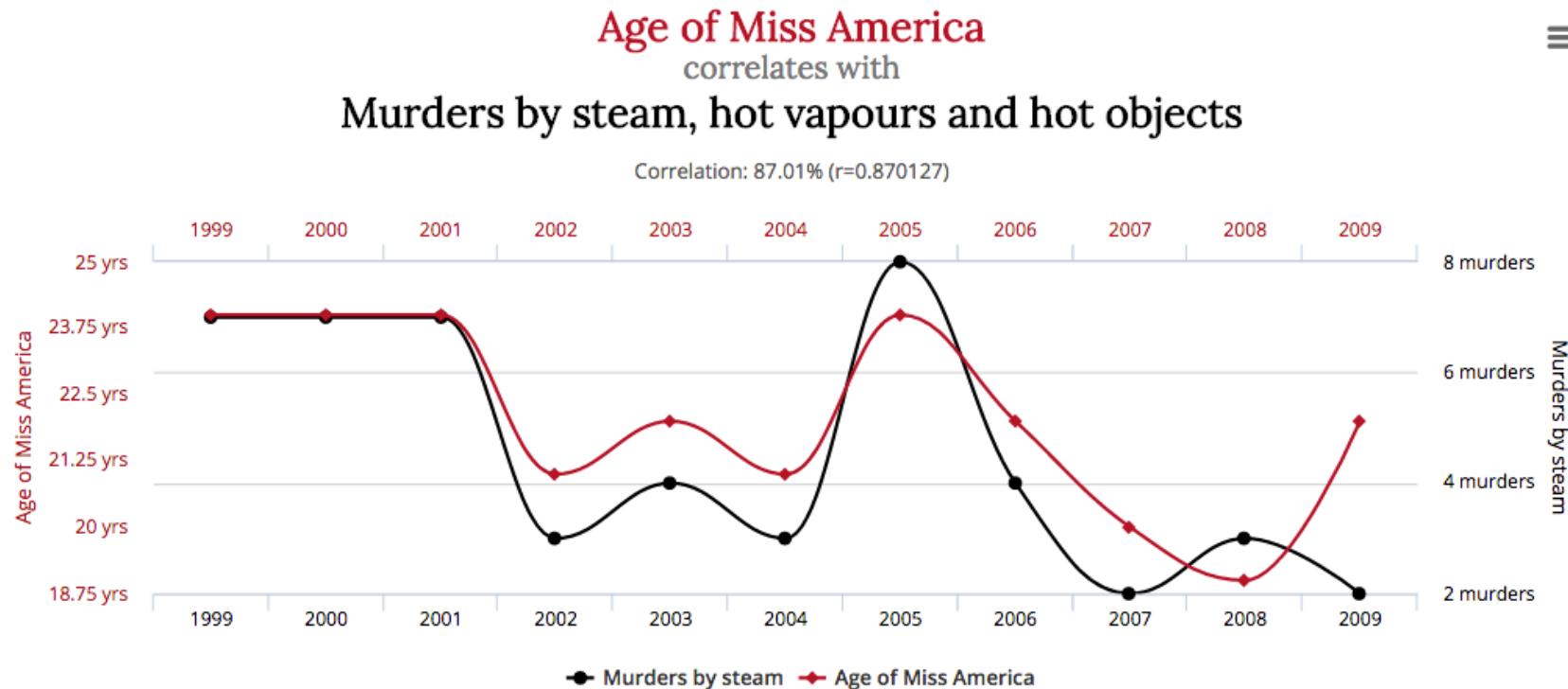


Corrélation / Causalité

Per capita consumption of mozzarella cheese
 correlates with
Civil engineering doctorates awarded



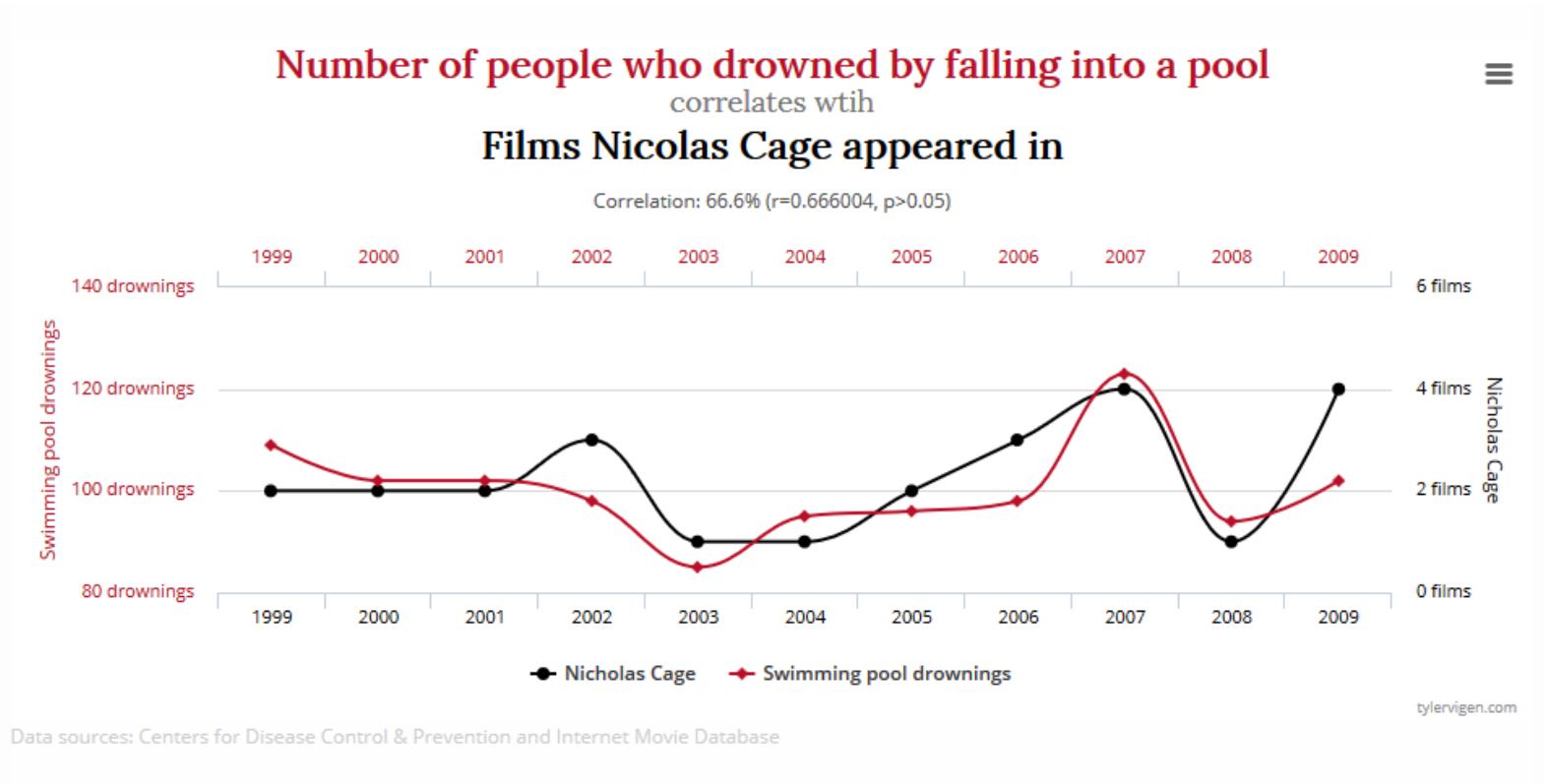
Corrélation / Causalité



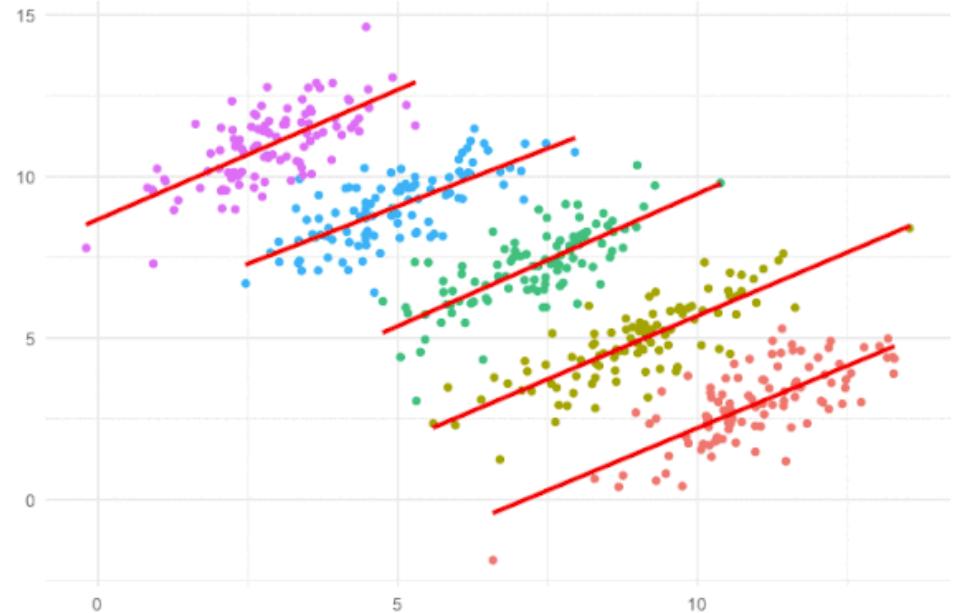
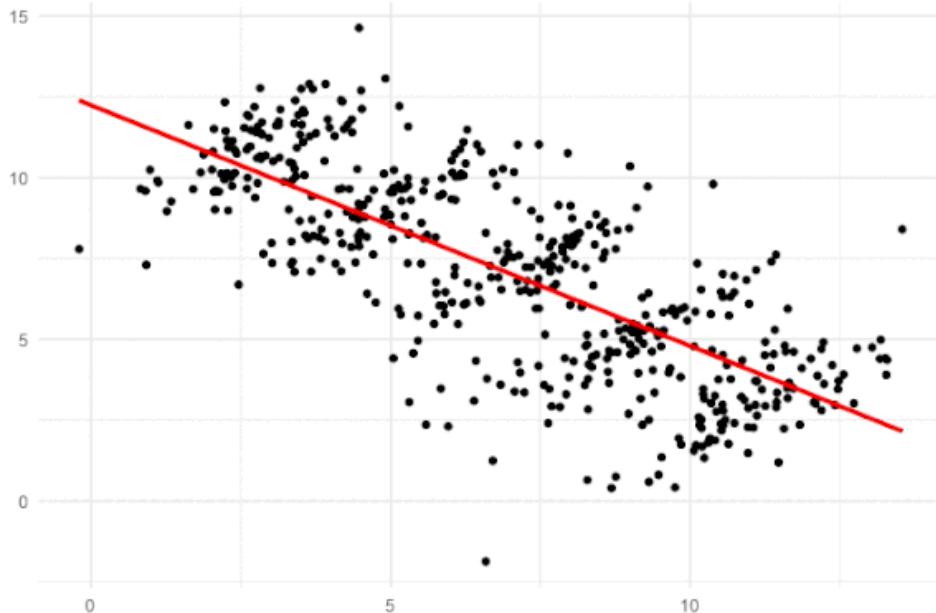
Data sources: Wikipedia and Centers for Disease Control & Prevention

tylervigen.com

Corrélation / Causalité



Paradoxe de Simpson



Biais de modèle

Les biais des données se retrouvent inévitablement dans les modèles

- L'algorithme apprend depuis les données d'entraînement
- Si la base est faussée, les prédictions ne peuvent qu'être erronées

Problématiques pratiques

- Perte d'observations dans le data flow (*missing values*, etc.)
- Oubli ou non-observation d'une variable
- Ne pas prendre du recul et oublier le bon sens

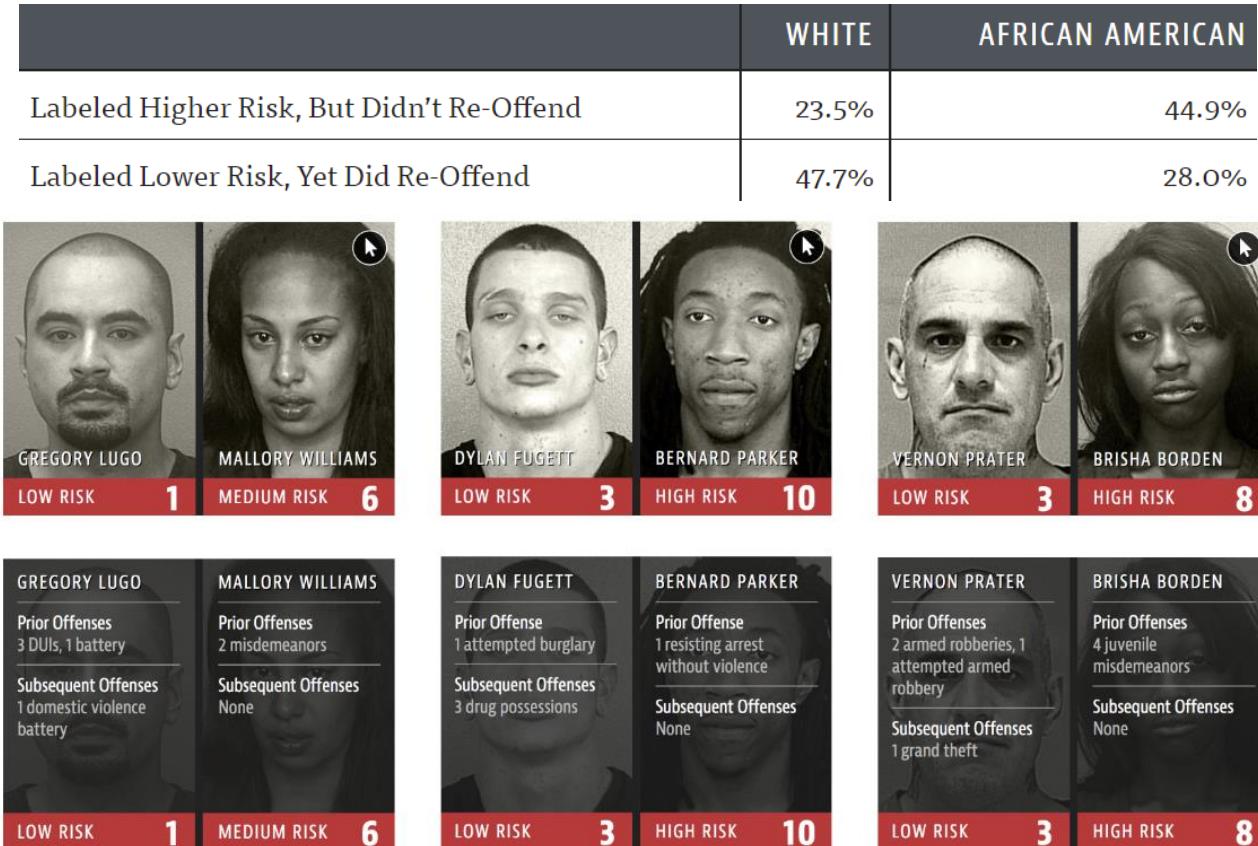
L'IA est-elle éthique ?

- L'intelligence artificielle est neutre par rapport aux données
- Mais les données proviennent de la société avec ses stéréotypes
- Prédictions sexistes ou racistes



Biais de modèle

« Aide » à la décision de justice aux U.S. pour prédire la récidive.



Un quatrième mensonge

« There are three kinds of lies

- lies,
- damned lies
- and statistics »

There are **four** kinds of lies

- lies,
- damned lies,
- statistics
- and **models**

R ou Python ?

R

- Logiciel pour les statisticiens par les statisticiens
- Fortement utilisé en recherche actuarielle
- Calculs pas forcément très optimisés
- *Statistical Learning*

Python

- Langage de programmation
- Puissance de calcul
- *Machine Learning*

Excel

- **Non !**

Analysis Tool	Similar Superhero	Super Powers in Common
		<ul style="list-style-type: none"> • Detective Work • Intelligence • Cunning • Usage of Tools • More Brain than Muscles
		<ul style="list-style-type: none"> • Muscle Power • Super Strength • Elegance • Wide Range • More Muscles than Brain

Scope du cours

Introduction à l'apprentissage statistique

- Donner les bases, les fondements mathématiques de la théorie
- Comprendre les principales problématiques
- Avoir une « boîte à outils » des principaux modèles : cours ET travaux dirigés

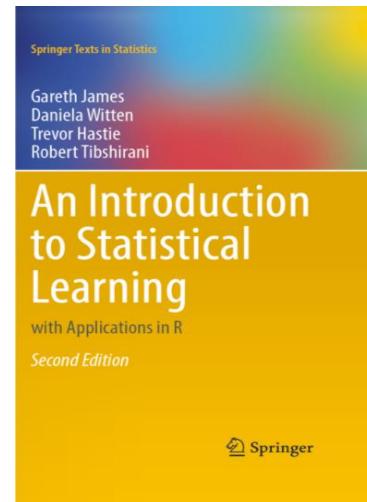
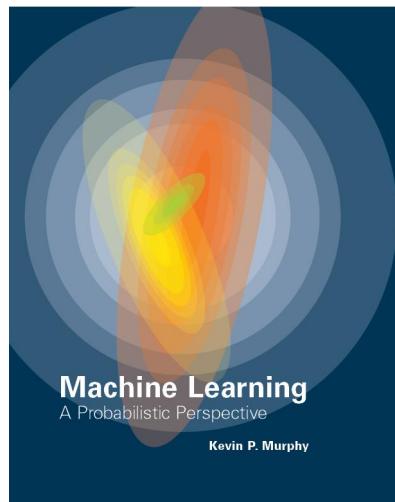
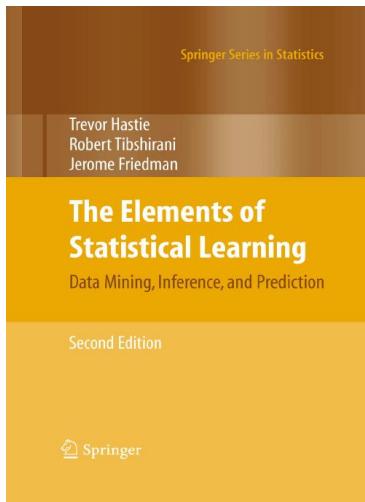
Tout ne sera pas abordé en détails

- Spécificité des séries temporelles
- Spécificité des données censurées
- Deep Learning avancé
- Unbalanced data

Bibliographie

Livres

- *The Elements of Statistical Learning*, 2009, T. Hastie, R. Tibshirani, J. Friedman
- *Machine Learning A Probabilistic Perspective*, 2012, K. P. Murphy
- *An Introduction to Statistical Learning with Applications in R*, 2013 , G. James, D. Witten, T. Hastie, R. Tibshirani



Bibliographie

Notes de cours

- *Machine Learning for Intelligent Systems*, K. Weinberger, Cornell
- *Data Analytics for Non-Life Insurance Pricing*, M. Wüthrich, ETH Zurich
- *Machine Learning*, A. Ng, Coursea / Standford

Divers

- Newsletter Veille Data, J. Del Hoyo
- Kaggle Kernels
- Stack Overflow
- * *Daily Dose of Data Science.*

Plan du cours

1. Motivations actuarielles
2. Modèles linéaires
3. Théorie de l'apprentissage statistique
4. Arbres de décision
5. Bagging
6. Boosting
7. Interprétabilité
8. Modèles à noyaux
9. Réseaux de neurones et Deep Learning