«Honni soit qui mal y science» Petite balade dans la science, la malscience... et les statistiques

Guy Tremblay
Professeur
Département d'informatique

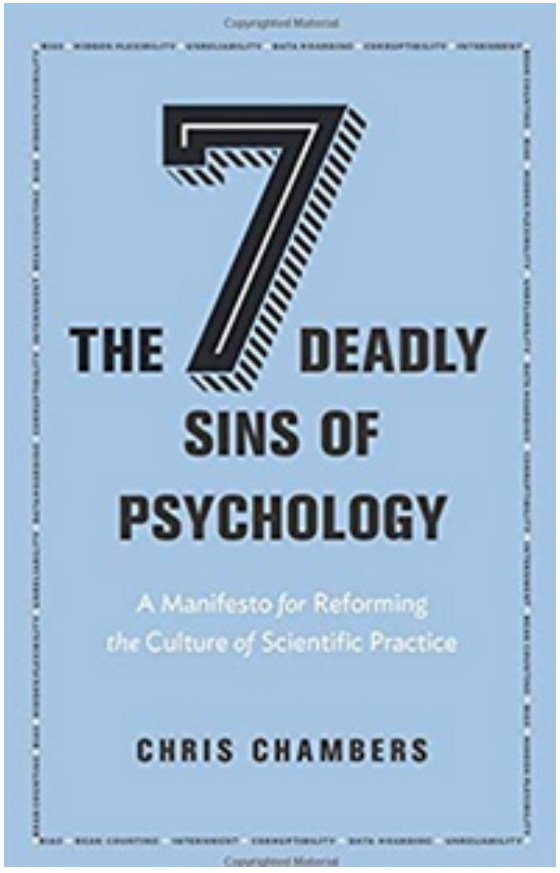
UQAM
http://www.labunix.uqam.ca/~tremblay_gu

Séminaire Latece
19 juin 2019

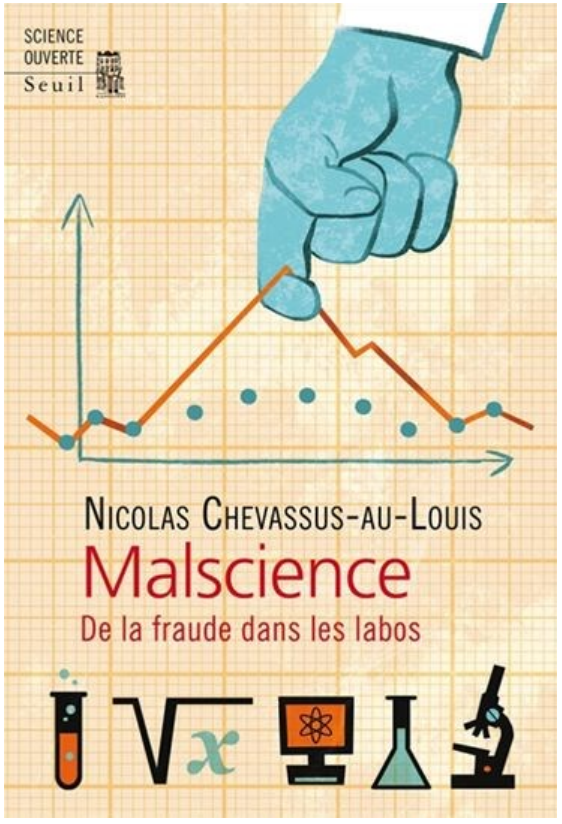
Have you ever noticed that all the instruments searching for intelligent life are pointed away from earth?

Aperçu

Trois lectures récentes. . .



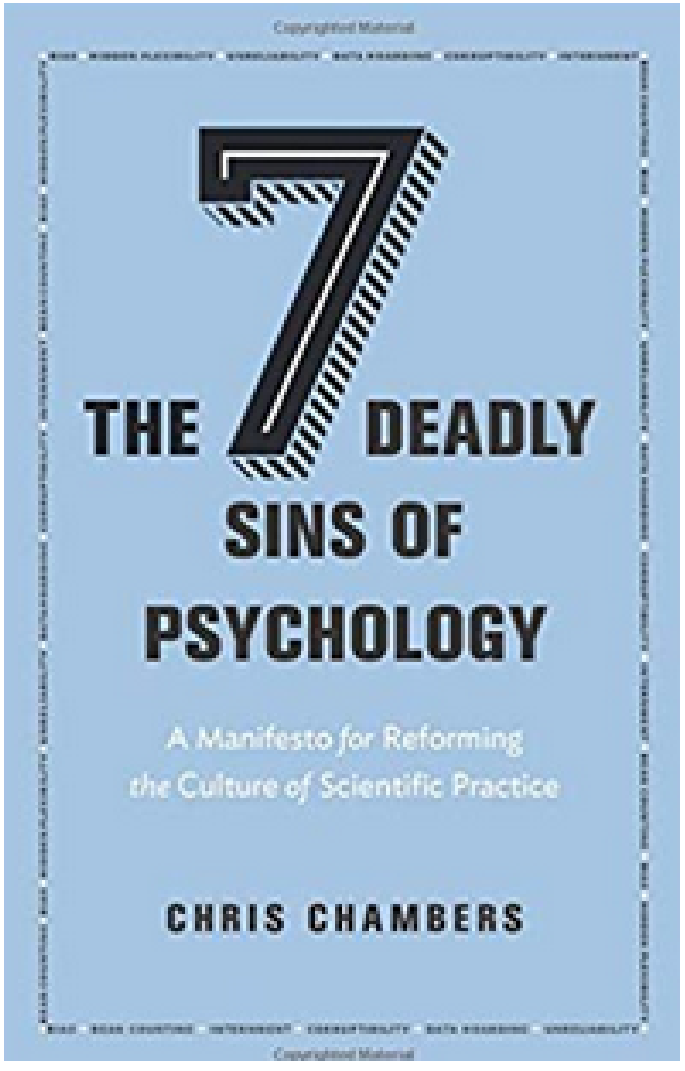
(Chambers, 2017)



(Chevaussus-au-Louis, 2016)

(NAS, 2018)

Dans ce séminaire, on va traiter de «malscience»... pas nécessairement de fraude



Quel est l'intérêt pour des chercheurs en informatique ou génie logiciel ?

- Essor, depuis 15–20 ans, du génie logiciel empirique — *Empirical Software Engineering*
  - *Empirical Software Engineering* (Journal, 1996)
  - *Evaluation and Assessment in Software Engineering* (Conférence, 1996)
  - *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (Conférence, 2007)

$\Rightarrow$ Utilisation de plus en plus fréquentes «d'expérimentations»

Quel est l'intérêt pour des chercheurs en informatique ou génie logiciel ?

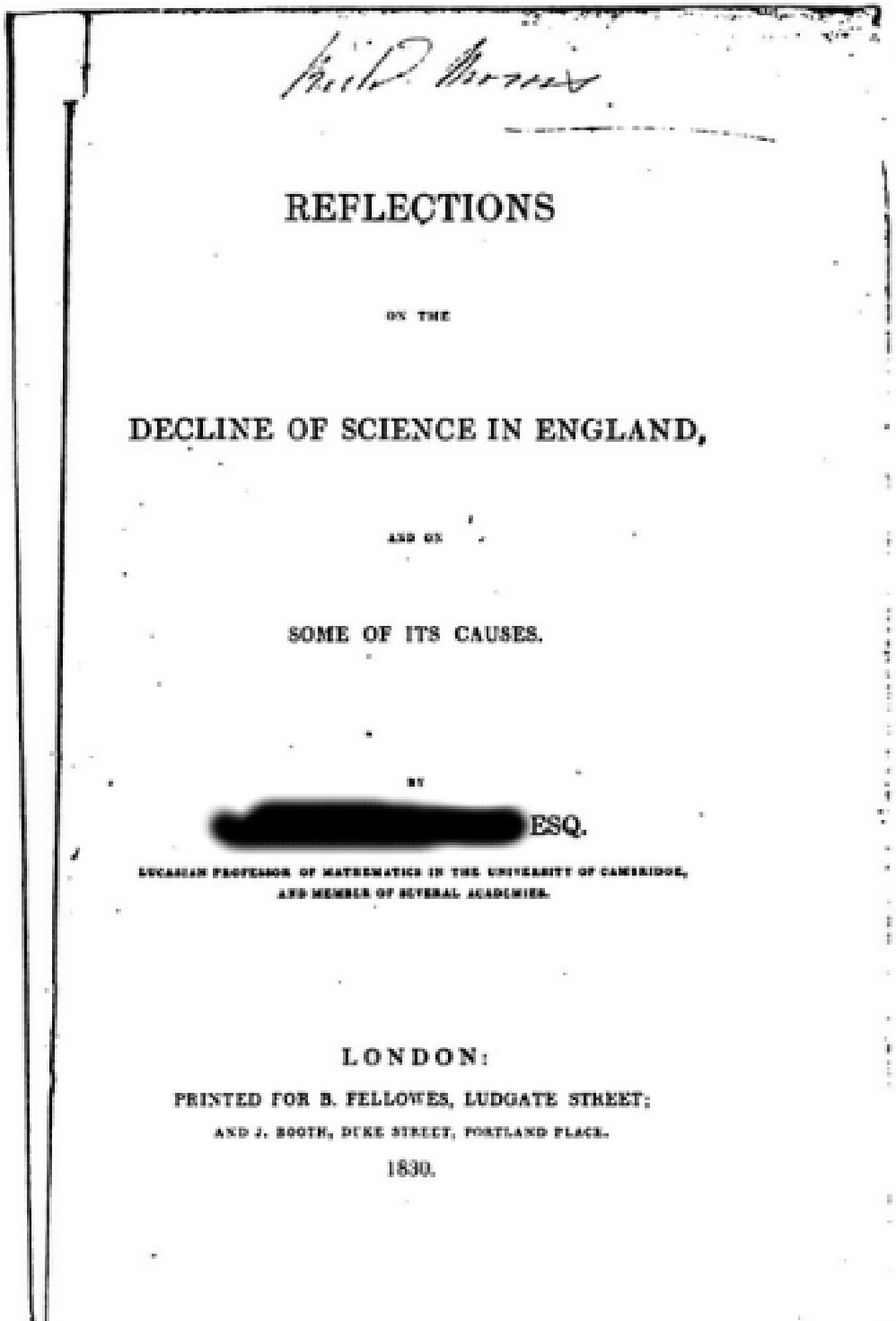
- Essor, depuis 15–20 ans, du génie logiciel empirique — *Empirical Software Engineering*
  - *Empirical Software Engineering* (Journal, 1996)
  - *Evaluation and Assessment in Software Engineering* (Conférence, 1996)
  - *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (Conférence, 2007)

$\Rightarrow$ Utilisation de plus en plus fréquentes «d'expérimentations»

- Expérimentations
  - $\Rightarrow$ Phénomènes irréguliers ou aléatoires (individus, contextes, etc.)
  - $+$ Erreurs expérimentales
  - $+$ Utilisation d'échantillons

  $\Rightarrow$ Utilisation de méthodes et d'inférences statistiques

Anecdote : il existe une analyse — très ancienne — de la «malscience» par un «informaticien» !?



REFLECTIONS

ON THE

DECLINE OF SCIENCE IN ENGLAND,

AND ON

SOME OF ITS CAUSES.

BY

[redacted] ESQ.

LUCASIAN PROFESSOR OF MATHEMATICS IN THE UNIVERSITY OF CAMBRIDGE,
AND MEMBER OF SEVERAL ACADEMIES.

LONDON:

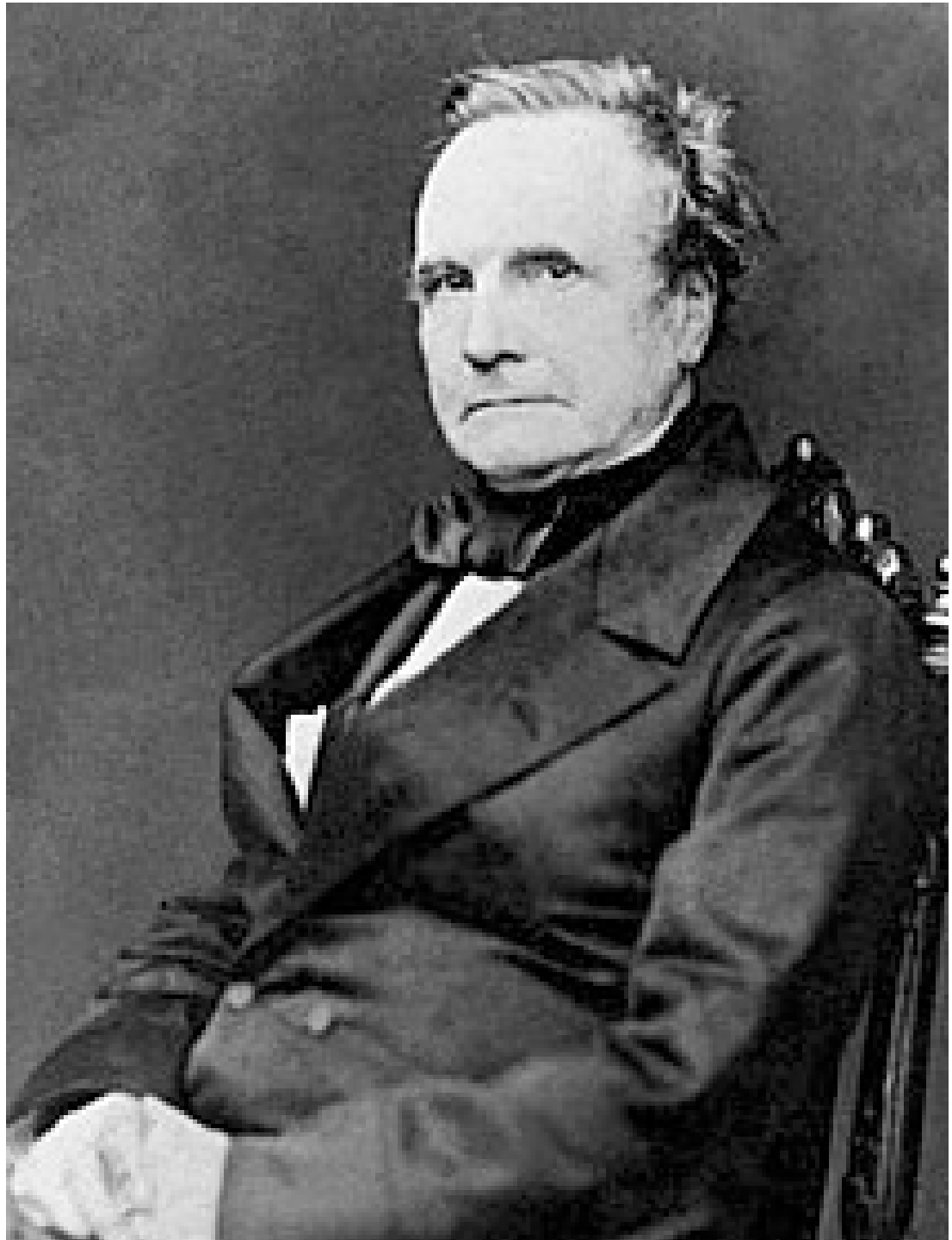
PRINTED FOR B. FELLOWES, LUDGATE STREET;

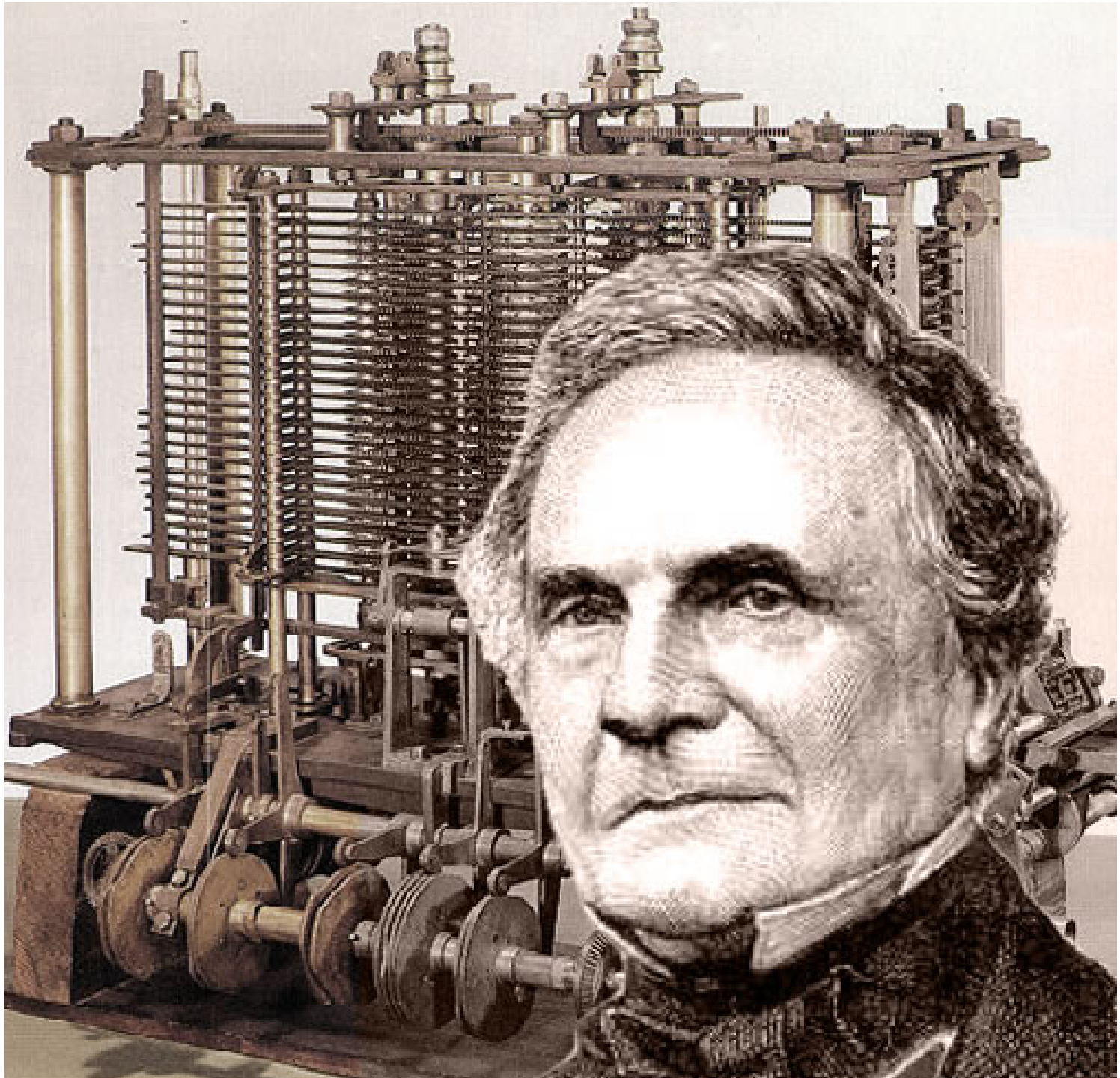
AND J. BOOTH, DUKE STREET, PORTLAND PLACE.

1830.

Quatre formes de «mauvaise science»

1. Canular
2. Falsification ou invention des données
3. Taillage des données
4. Cuisinage des données





REFLECTIONS

ON THE

DECLINE OF SCIENCE IN ENGLAND,

AND ON

SOME OF ITS CAUSES.

BY

CHARLES BABBAGE, ESQ.

LUCASIAN PROFESSOR OF MATHEMATICS IN THE UNIVERSITY OF CAMBRIDGE,
AND MEMBER OF SEVERAL ACADEMIES.

Aperçu

Aperçu

The Irreproducibility Crisis Report

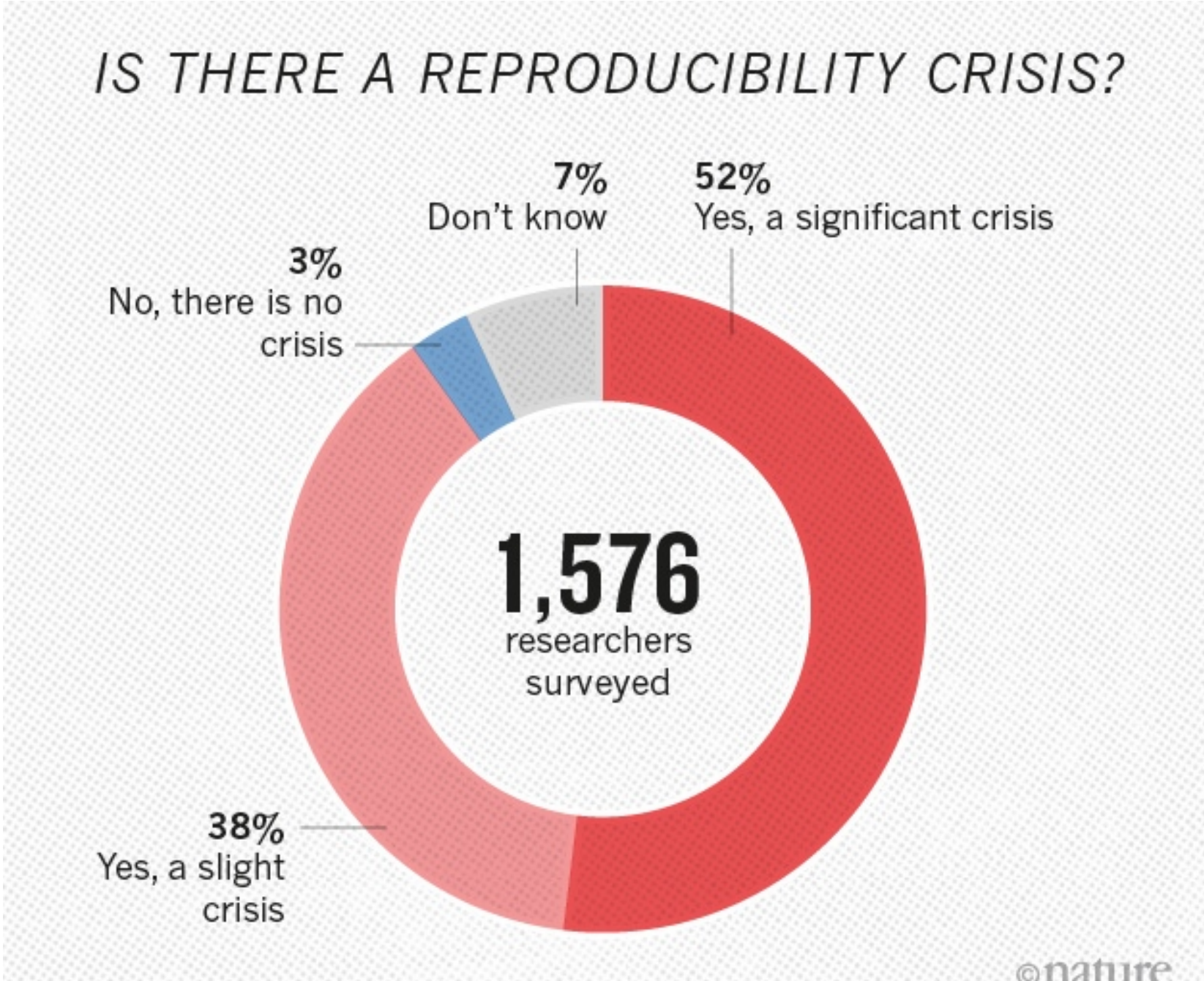
Causes, Consequences, and the Road to Reform

A reproducibility crisis afflicts a wide range of scientific and social-scientific disciplines, from epidemiology to social psychology. [. . .] Many supposedly scientific results cannot be reproduced reliably in subsequent investigations, and offer no trustworthy insight into the way the world works.

National Association of Scholars, 2018

Sondage fait par la revue *Nature* (2016)

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970



2005 : Article sur les «faux» résultats de recherche

Notamment en médecine selon l'analyse de l'auteur

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may

---

«*Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true.*
[. . .]
[This is in part because of the] ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by **formal statistical significance**, *typically for* **a p-value less than 0.05**.»

2012 : Article sur la non-reproductibilité d'études sur le cancer

NATURE | NEWS

Biotech giant publishes failures to confirm high-profile science

Amgen posts three studies at new online channel for discussing reproducibility.

Monya Baker

04 February 2016

Amgen researchers made headlines when they declared that they had been unable to reproduce the findings in 47 of 53 «landmark» [cancer and hematology] papers.

2015 : Article sur la non-reproductibilité d'études en psychologie

Science

Estimating the reproducibility of psychological science

Open Science Collaboration

Science **349** (6251), aac4716.
DOI: 10.1126/science.aac4716

«Aarts et al. describe the replication of 100 experiments reported in papers published in 2008 in three high-ranking psychology journals. [. . .] they find that about one-third to one-half of the original findings were also observed in the replication study [donc 50–60% non reproductibles].»

2016 : «*Disastrous blog post*» de B. Wansik

- Ancien professeur/chercheur — nutrition, comportement du consommateur, etc.
- Ancien directeur exécutif du *USDA's Center for Nutrition Policy and Promotion*
- Articles cités plus de 20 000 fois
- Mais. . .

2016 : «*Disastrous blog post*» de B. Wansik

- Ancien professeur/chercheur — nutrition, comportement du consommateur, etc.
- Ancien directeur exécutif du *USDA's Center for Nutrition Policy and Promotion*
- Articles cités plus de 20 000 fois
- Mais depuis 2017 : 17 articles retirés par des revues, dont 6 (le même jour) du *Journal of the American Medical Association*

"I already wrote the paper.
That's why it's so hard to
get the right data."

2016 : «*Disastrous blog post*» de B. Wansik

When [this graduate student] arrived, I gave her a data set of a [. . .] failed study which had null results [. . .]. I said, "This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set."

I had three ideas for potential Plan B, C, & D directions (since Plan A [the one-month study with null results] had failed). I told her what the analyses should be and what the tables should look like. [. . .] Six months after arriving, . . . [she] had one paper accepted, two papers with revision requests, and two others that were submitted (and were eventually accepted).

Autre symptôme : Augmentation importante du nombre de rétractations

nature International weekly journal of science

nature news home | news archive | specials | opinion | features | news blog | nature journal

comments on this story

Stories by subject

- Health and medicine
- Lab life
- Policy

Stories by keywords

Published online 5 October 2011 | *Nature* **478**, 26-28 (2011) | doi:10.1038/478026a

News Feature

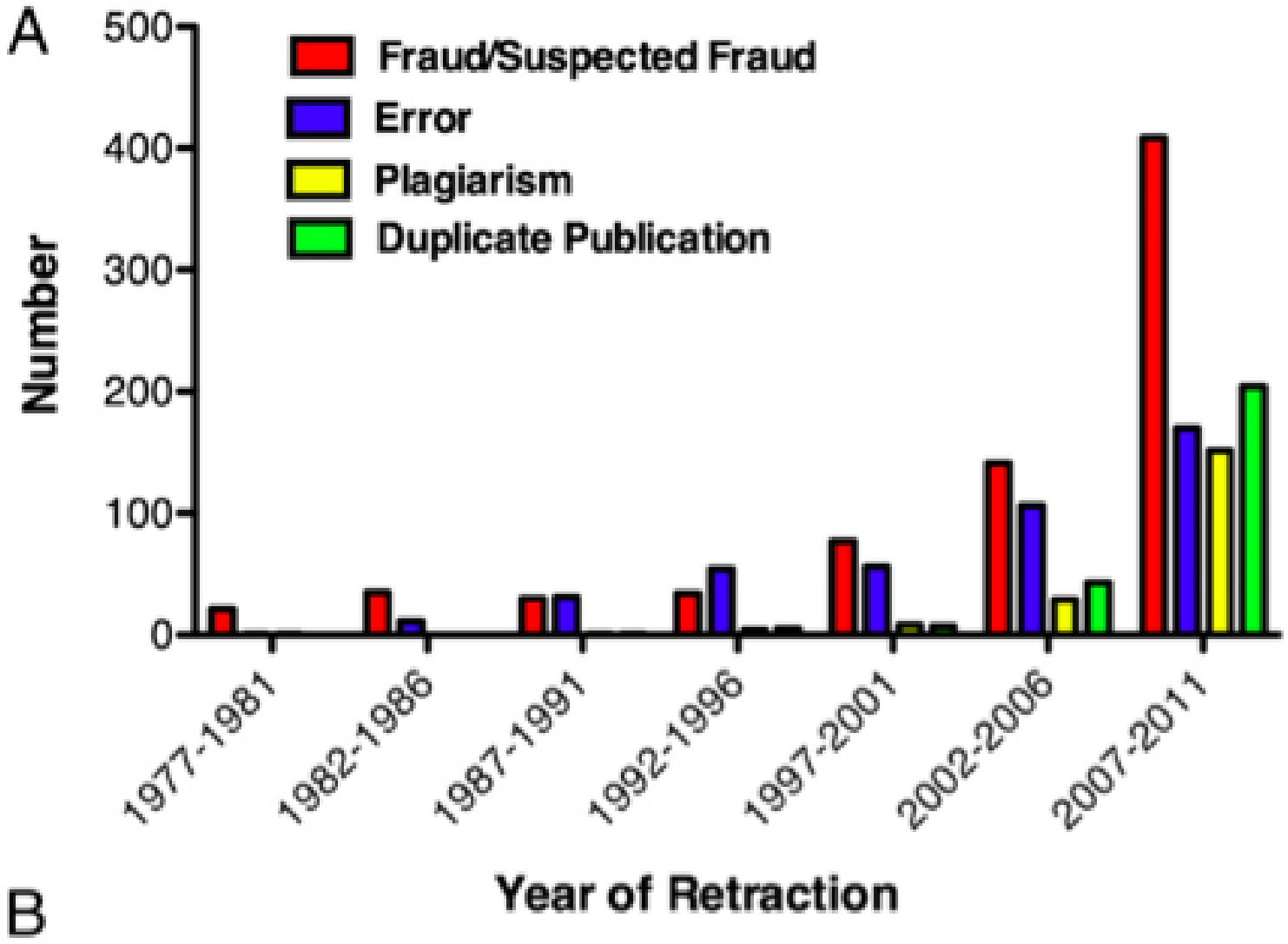
Science publishing: The trouble with retractions

A surge in withdrawn papers is highlighting weaknesses in the system for handling them.

Richard Van Noorden

- Nombre de rétractations $\approx$ multiplié par 10–12
- Les journaux prestigieux (e.g., Science, Nature, Cell) sont les plus touchés !

Autre symptôme : Augmentation importante du nombre de rétractations

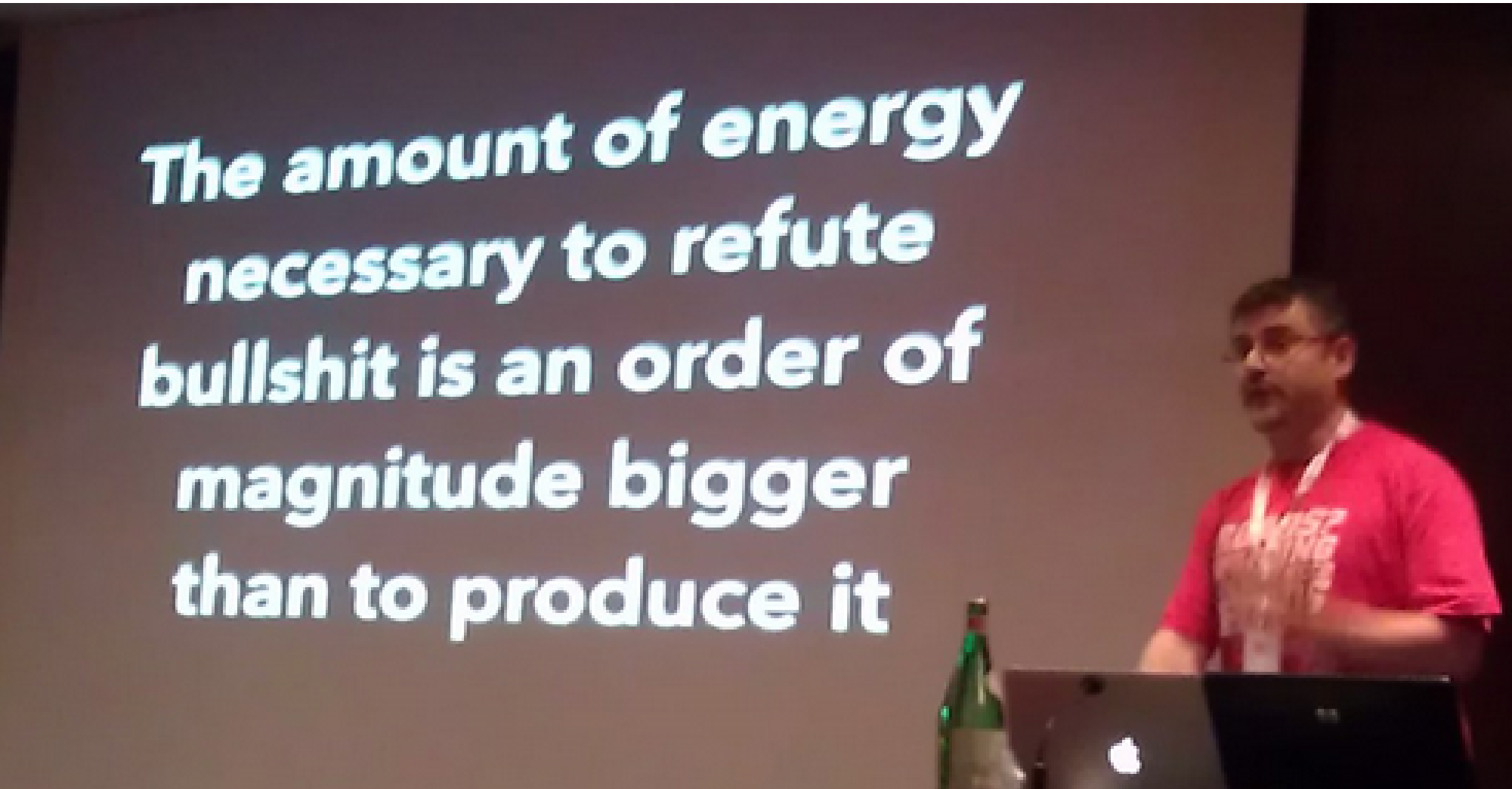


Le problème = Les rétractations ont souvent. . . peu d'effets ☹

Loi de Brandolino = ?

Le problème = Les rétractations ont souvent… peu d'effets ☹

Loi de Brandolino = *Bullshit asymetry principle*



Connaissez-vous des exemples ?

Un exemple criant : Article du Lancet (1998) sur les liens entre autisme et vaccin contre la rougeole

Vaccin MMR (*Measles, Mumps, and Rubella*) = Rougeole + Oreillons + Rubéole

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for 1 week, accompanied by their parents.

Clinical investigations

We took histories, including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S).

Un exemple criant : Article du Lancet (1998) sur les liens entre autisme et vaccin contre la rougeole

Vaccin MMR (*Measles, Mumps, and Rubella*) = Rougeole + Oreillons + Rubéole

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for 1 week, accompanied by their parents.

Clinical investigations

We took histories, including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S).

- Plus de 700 citations jusqu'en 2000

L'article a été rétracté en 2010

- Rétractation suite à une enquête (2004–10 !) par B. Deer, journaliste au *Sunday Times*
- Sur les 12 enfants :
  - 3 n'avaient aucun symptome d'autisme
  - 5 avaient développé les symptomes avant la vaccination
- Information omise dans l'article : Tests négatifs (par un assistant de Wakefield) sur la présence d'ARN de rougeole !

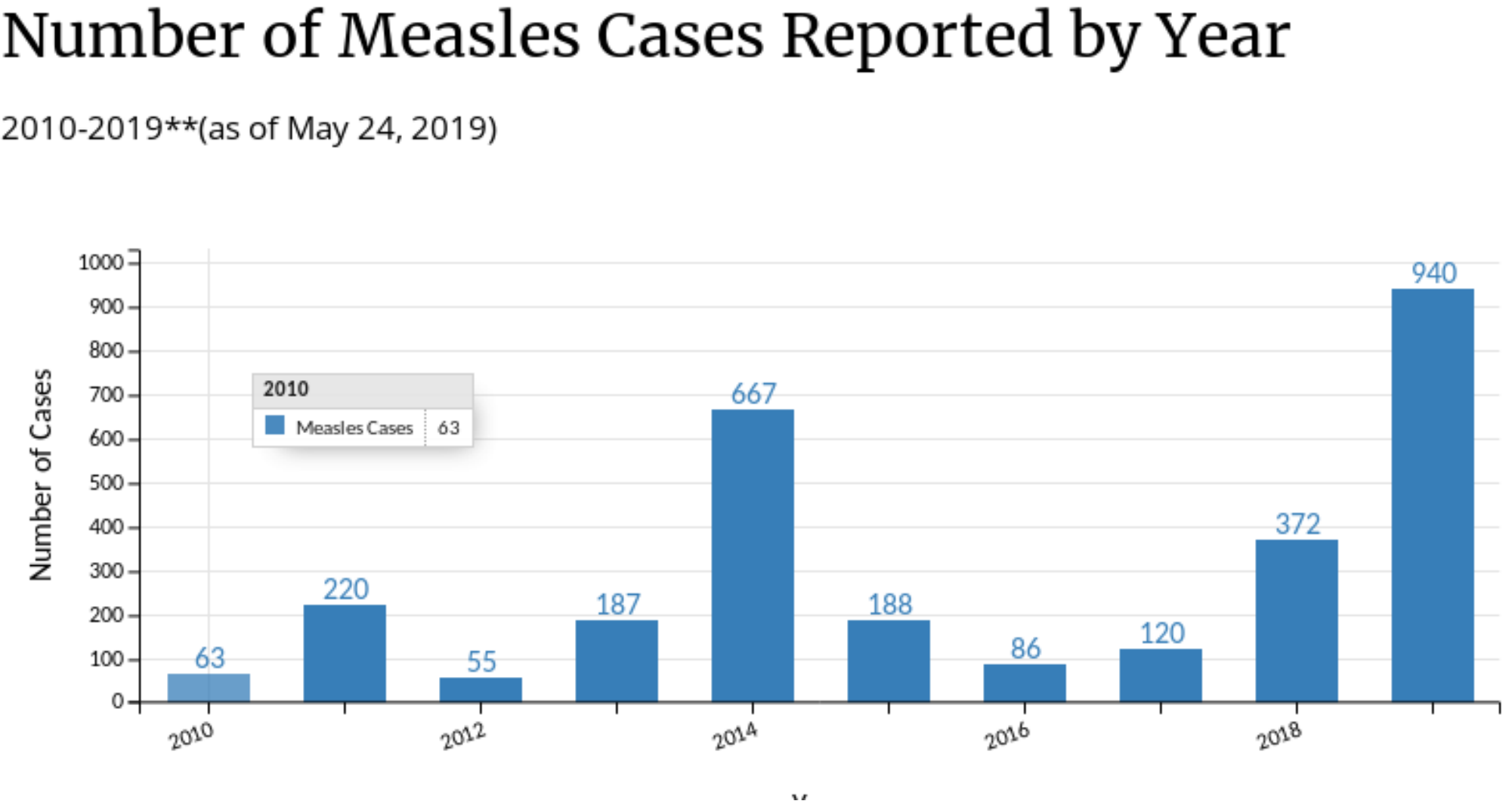
Et maintenant, en 2019. . .

A. Wakefield

- Interdit de pratique en Angleterre
- Travaille aux USA comme médecin conseil pour des associations anti-vaccins

Et maintenant, en 2019. . .

Nombre de cas aux USA — Situation semblable dans plusieurs autres pays ☹



Ajout de dernière minute : La Presse, 18 juin 2019

Publié le 18 juin 2019 à 18h40 | Mis à jour à 18h42

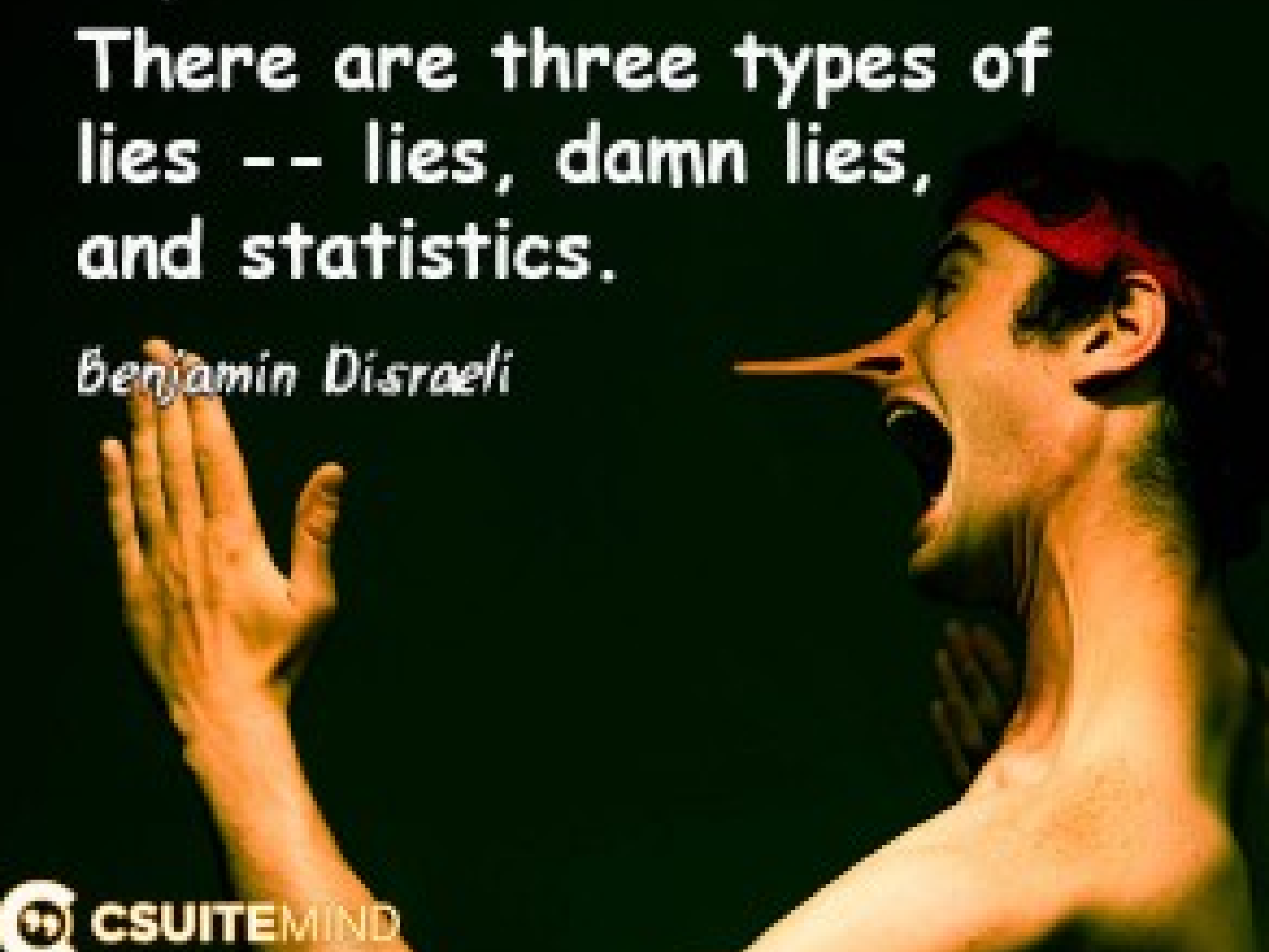
Laval: des passants possiblement contaminés à la rougeole

Le virus de la rougeole pourrait avoir été transmis pourrait avoir été transmis à des passants au Carrefour Laval.

Aperçu

There are three types of lies -- lies, damn lies, and statistics.

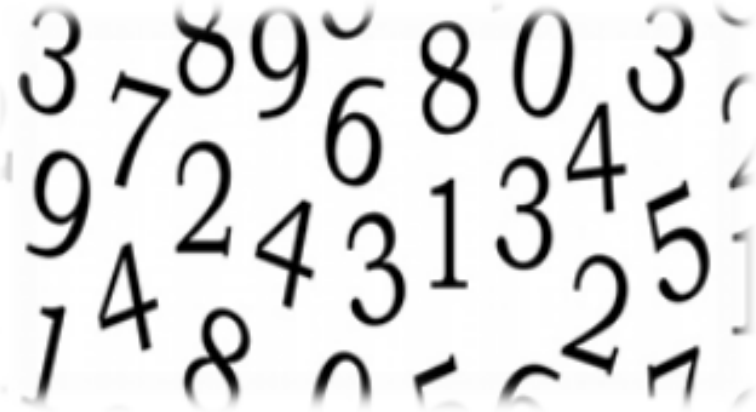
Benjamin Disraeli



Aimez-vous les statistiques ?

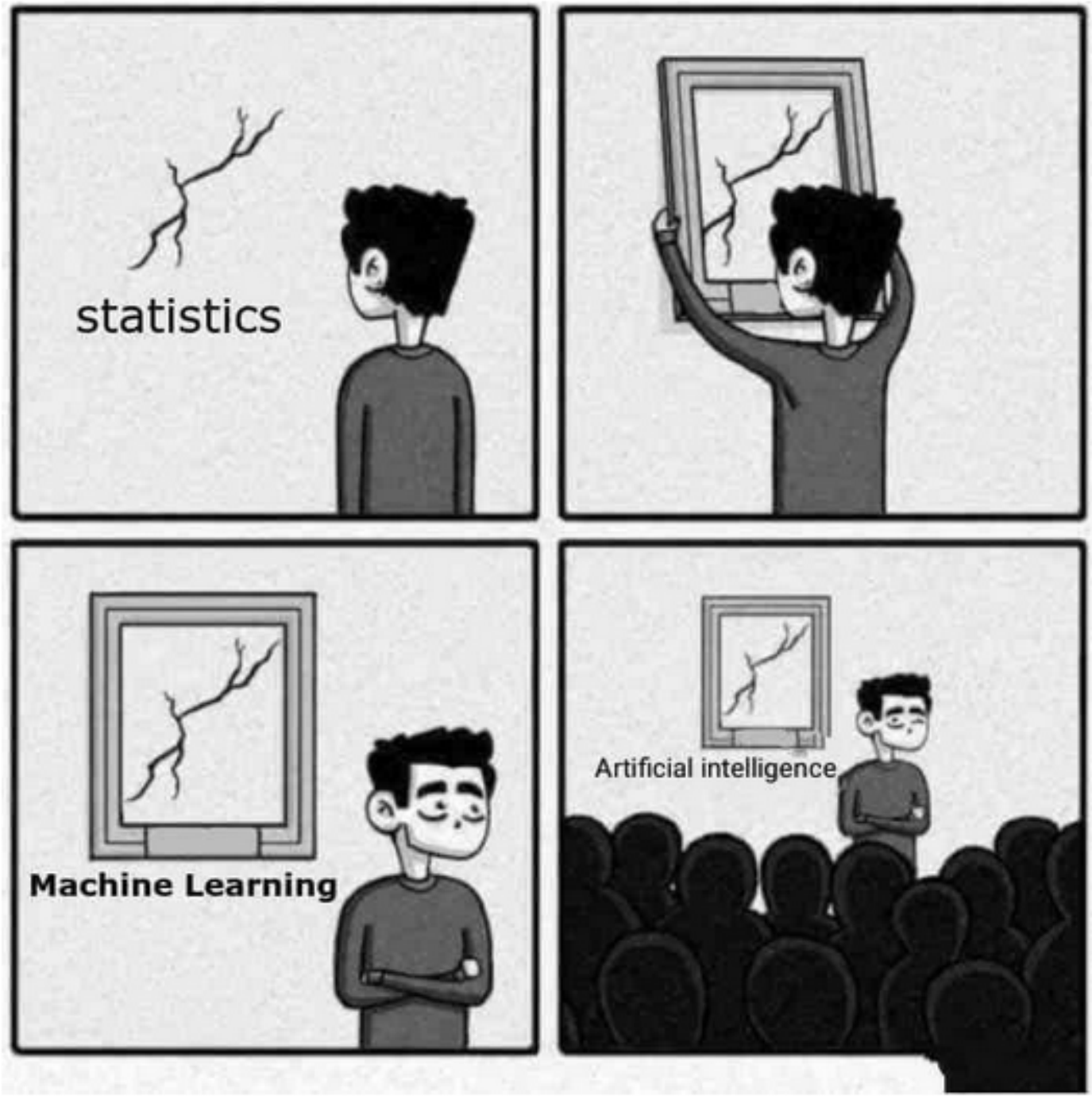
https://www.youtube.com/watch?v=ldy9RiRRZ3Y

STATISTICS EVERYWHERE!!!!



https://slideplayer.com/slide/8773876

http://towardsdatascience.com



L'utilisation — la mauvaise utilisation ? — des statistiques joue un rôle clé dans la crise en science

SIGNIFICANCE

Business Culture Politics Scie

Cargo-cult statistics and scientific crisis

Written by Philip B. Stark and Andrea Saltelli on 05 July 2018. Posted in Science

AMERICAN Scientist

The Statistical Crisis in Science

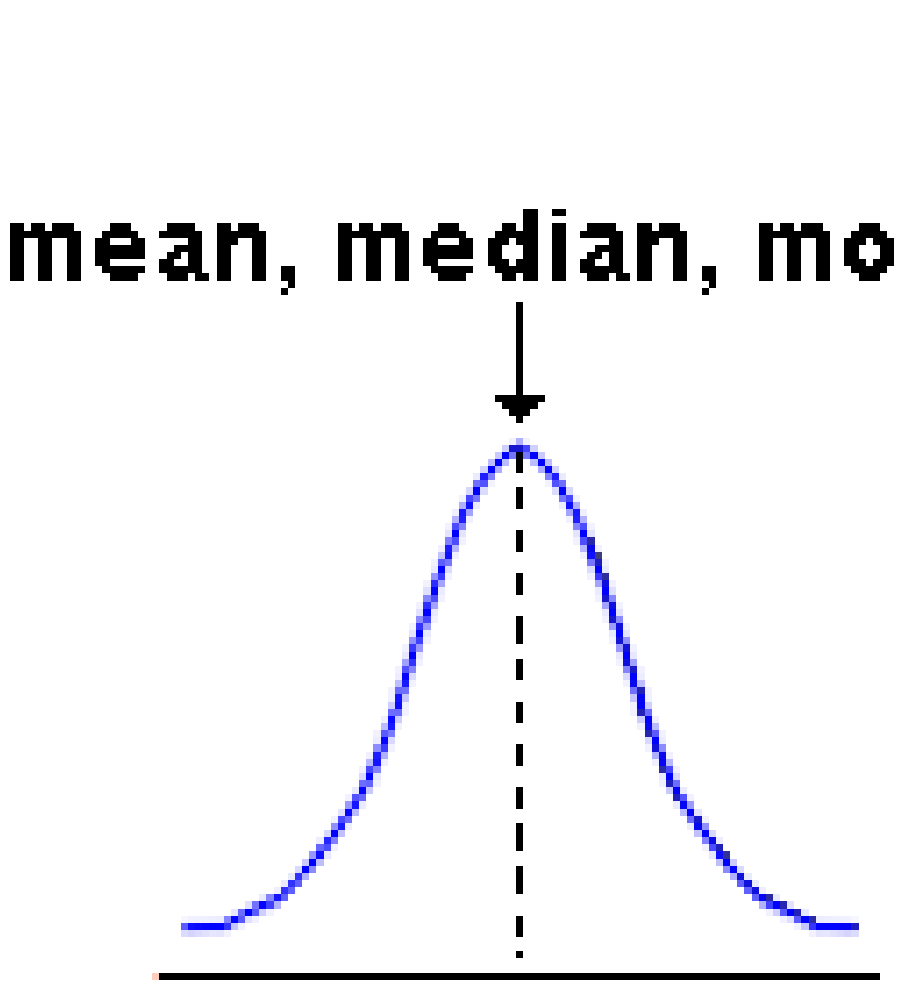
BY ANDREW GELMAN, ERIC LOKEN

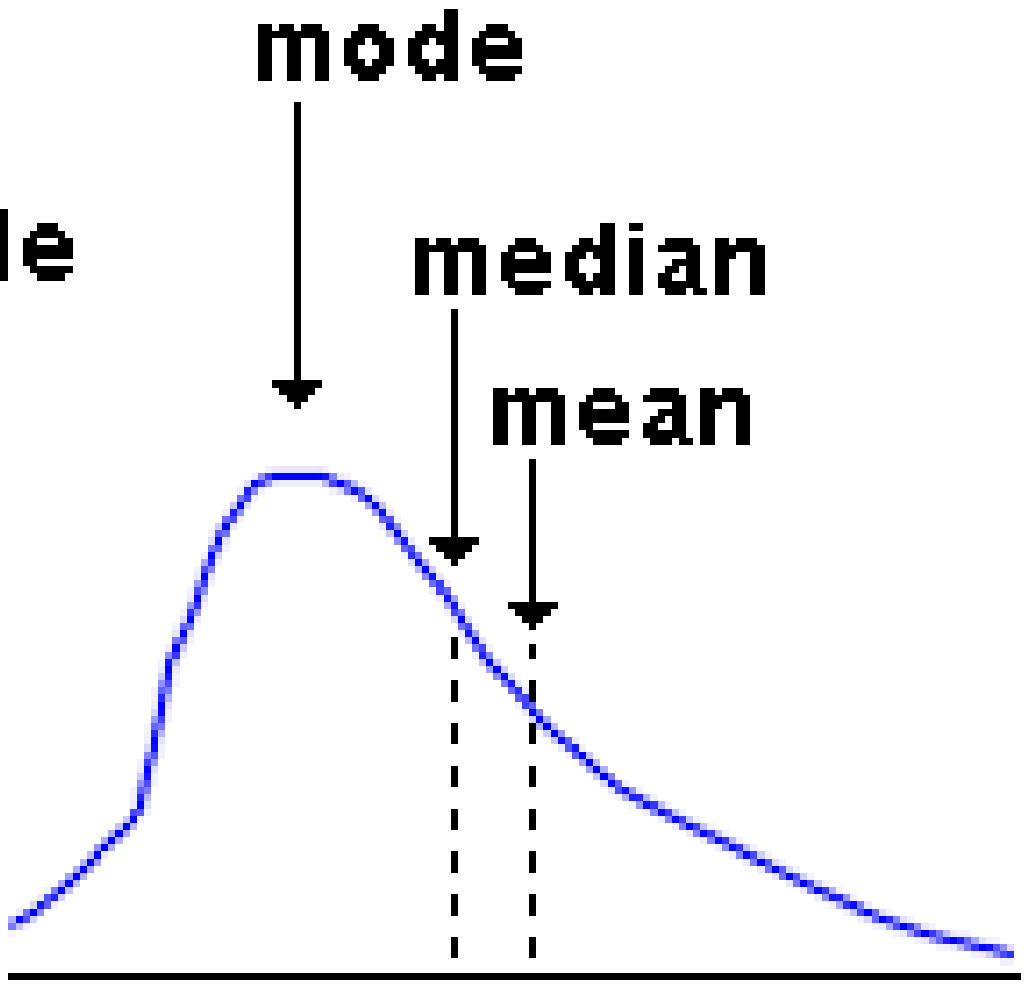
Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Mesures de tendance centrale

Mesure de tendance centrale = Valeur autour de laquelle se concentrent les données

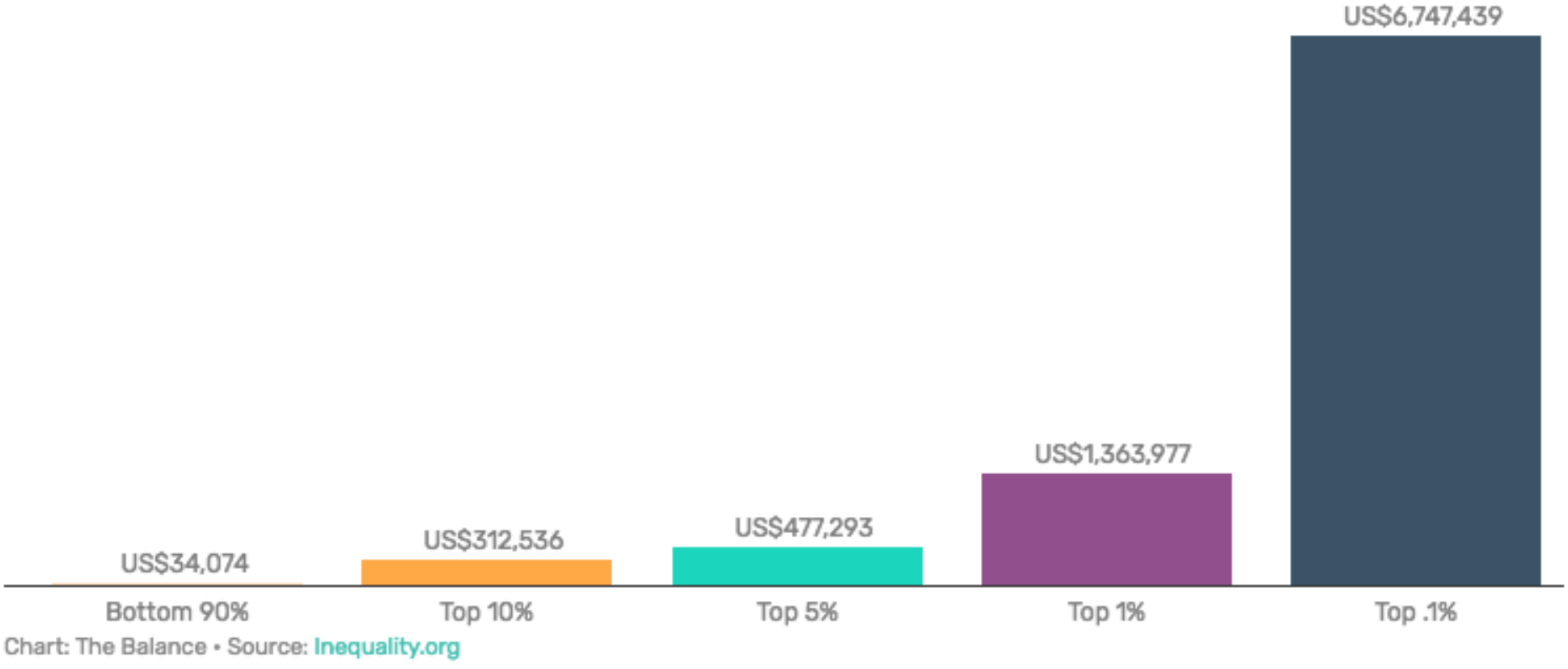
https://vula.uct.ac.za





Revenus des ménages aux USA

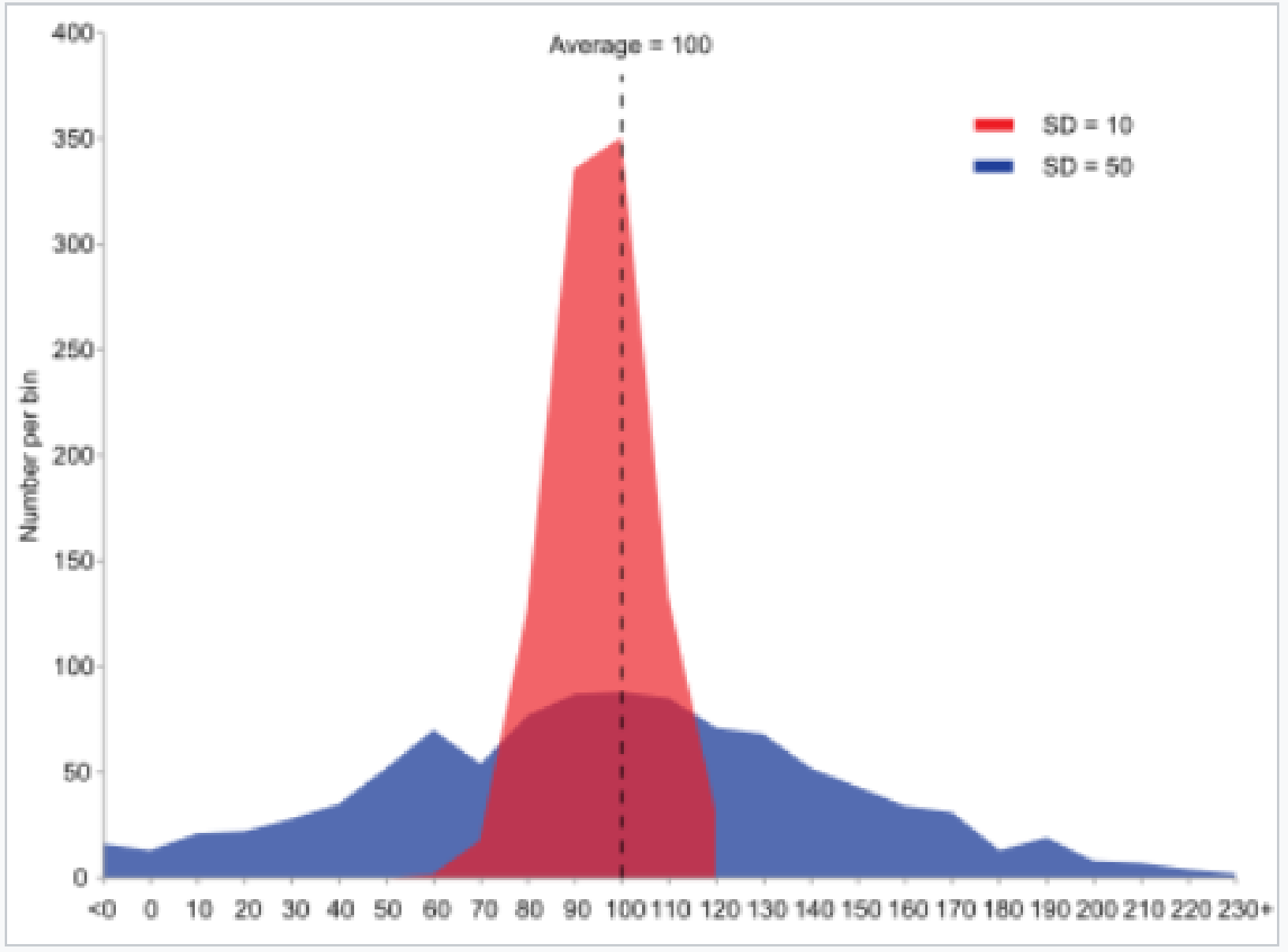
Moyenne $\approx 0.9 \times 34\,074\$ + 0.1 \times 312\,536\$ = 61\,920\$$



Mesures de dispersion

Mesure de dispersion = Décrit la variabilité des différentes valeurs

https://en.wikipedia.org/wiki/Statistical_dispersion



Mesure de dispersion = Décrit la variabilité des différentes valeurs

Écart-type

Soit $xs = \{x_0, x_1, \ldots, x_{n-1}\}$ et $m = \text{Moyenne}(xs)$

$$\text{Écart-type}(xs) = \sqrt{\frac{\sum_{i=0}^{n-1}(x_i - m)^2}{n-1}}$$

Mesure d’association

Mesure d'assocation la plus commune = Coefficient de corrélation (linéaire)

Coefficient de corrélation entre deux mesures

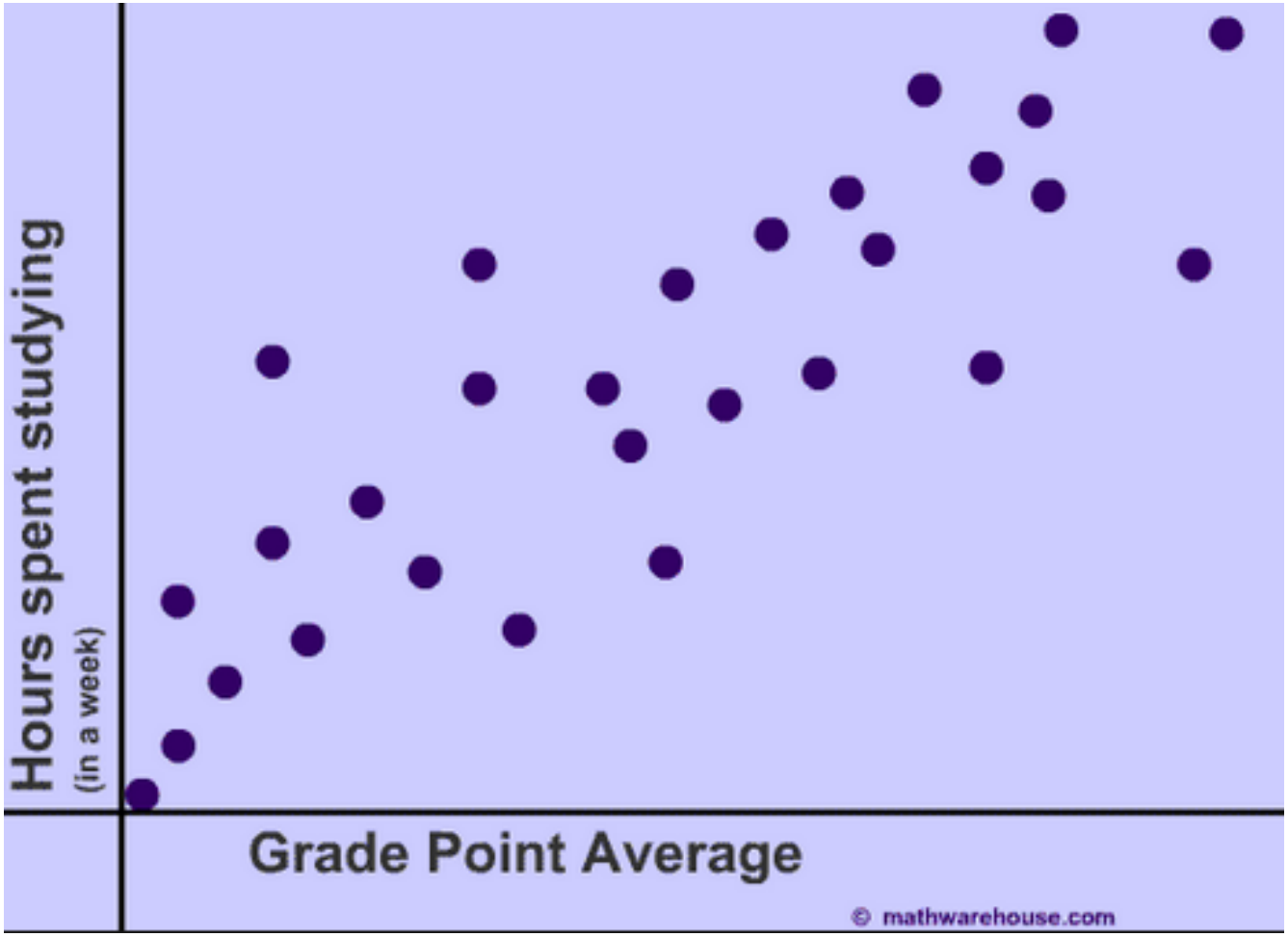
«*standardized way of describing the amount by which [two measures] covary*»

«*Statistical Methods and Measurement*», J. Rosenberg [SSS08]

Exemples de corrélation — positive

Nombre d'heures d'étude vs. résultat académique

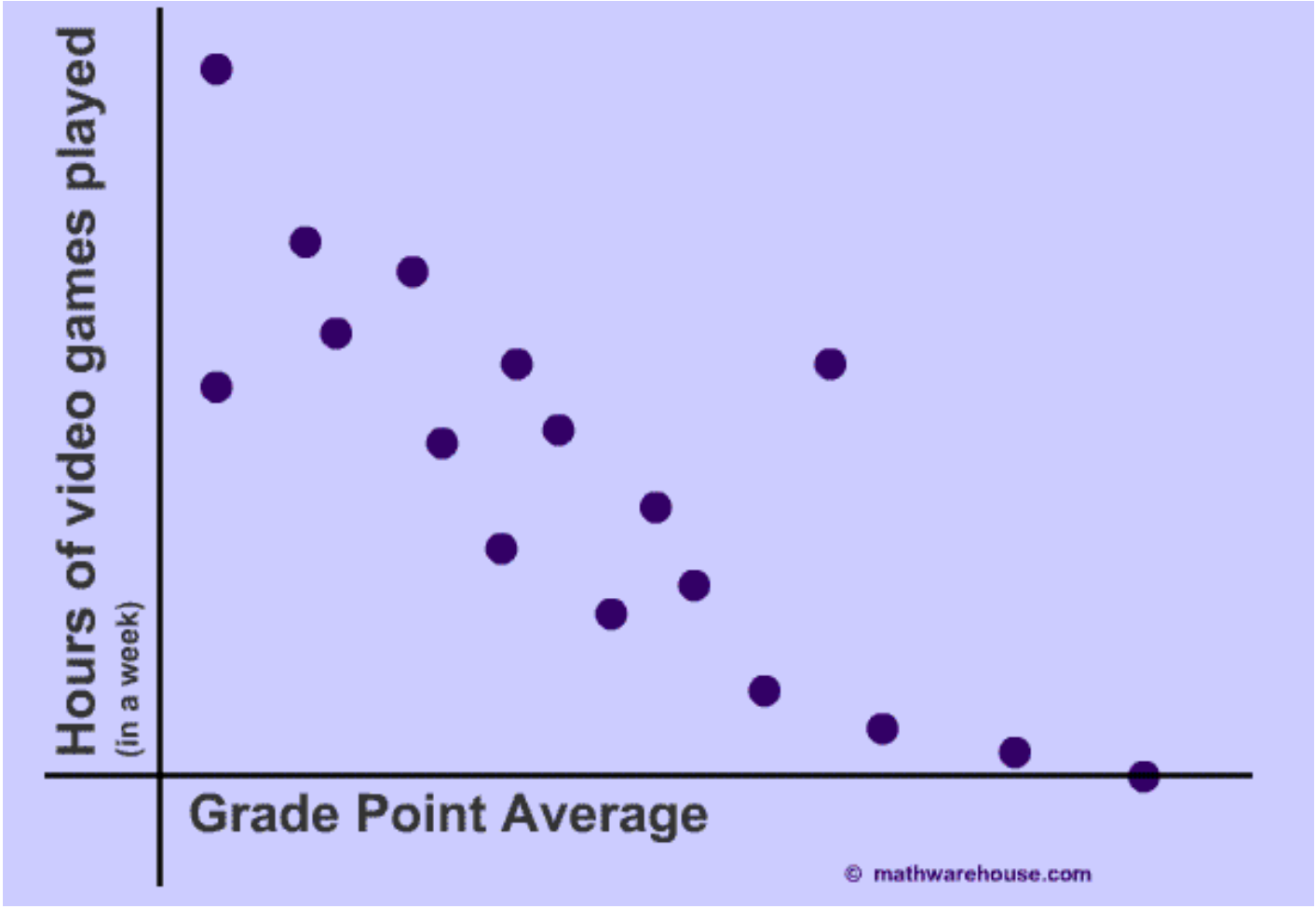
https://www.mathwarehouse.com/statistics/correlation-coefficient/how-to-calculate-correlation-coefficient.php



Exemples de corrélation — négative

Nombre d'heures de jeux vidéos vs. résultat académique

https://www.mathwarehouse.com/statistics/correlation-coefficient/how-to-calculate-correlation-coefficient.php



La corrélation de Pearson

Corrélation de Pearson (entre deux séries de données !)

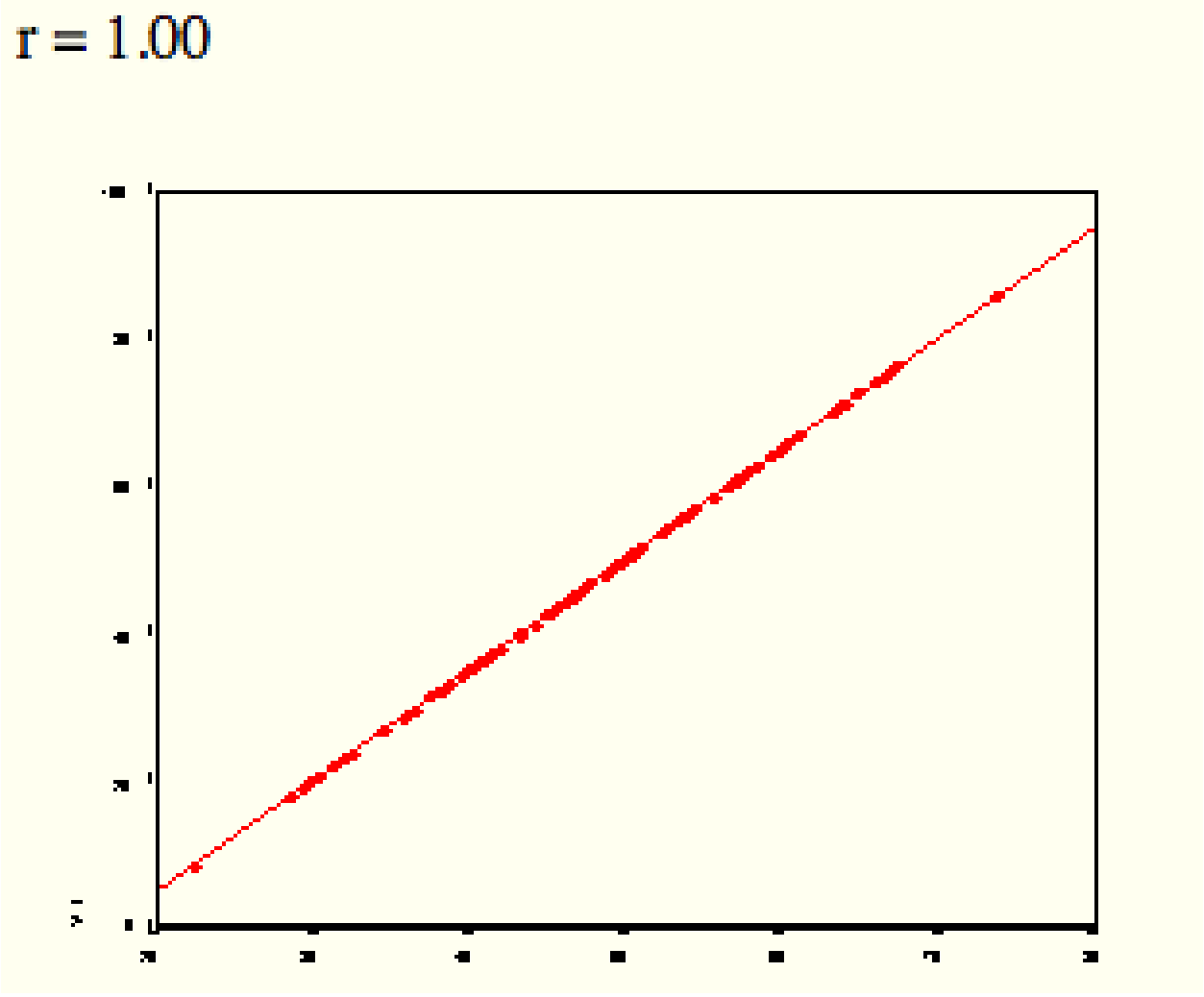
Soit $xs = [x_0, x_1, \ldots, x_{n-1}]$
Soit $ys = [y_0, y_1, \ldots, y_{n-1}]$

corrélation(xs, ys) = «degré de relation linéaire qui existe entre deux (séries de) mesures»

$$\text{corrélation}(xs, ys) = \frac{\sum_{i=0}^{n-1} \frac{(x_i - m_x)}{et_x} \frac{(y_i - m_y)}{et_y}}{n-1}$$

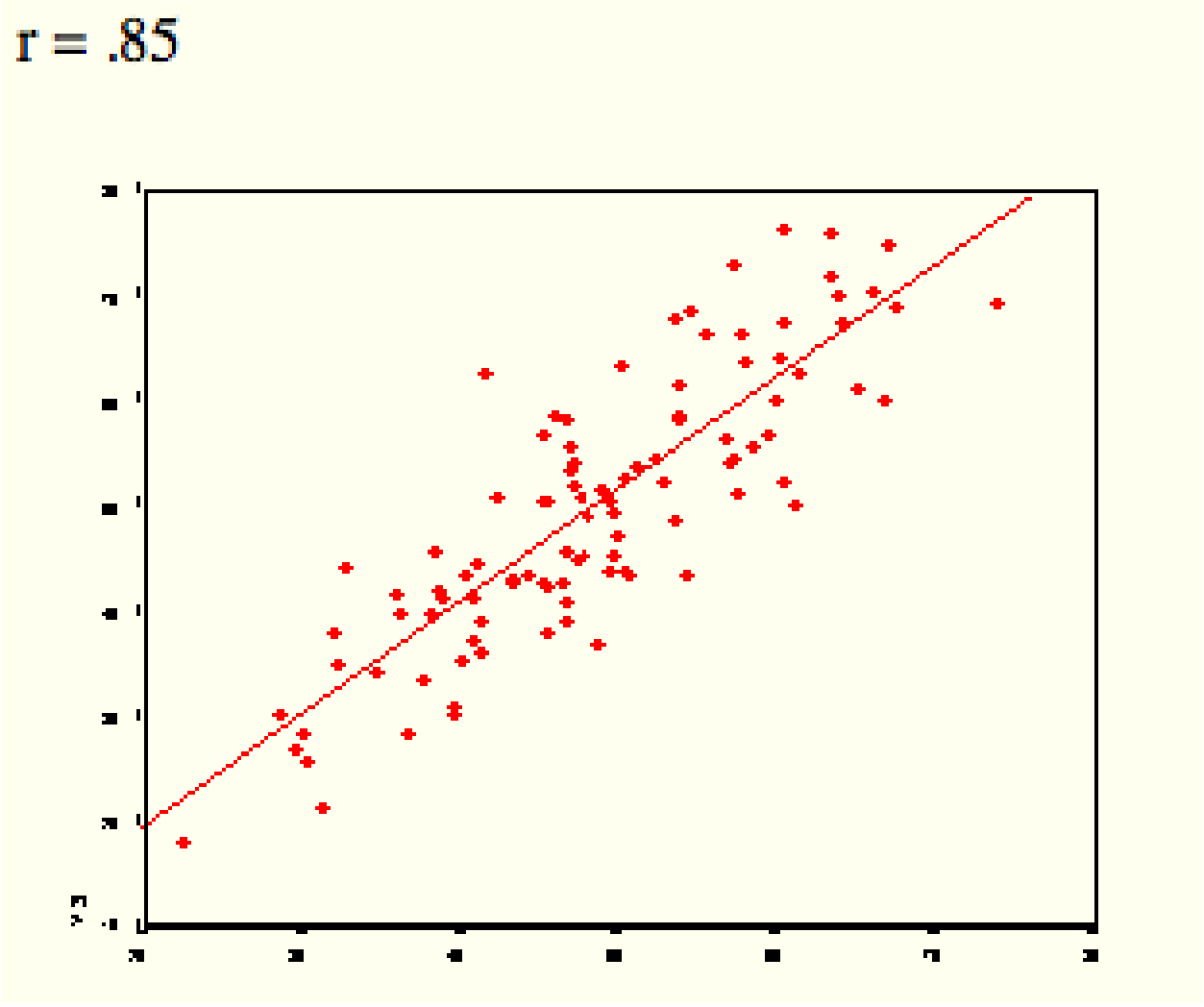
Le coefficient de corrélation varie de -1.0 à $+1.0$

Source: `http://faculty.cbu.ca/~erudiuk/IntroBook/sbk17.htm`



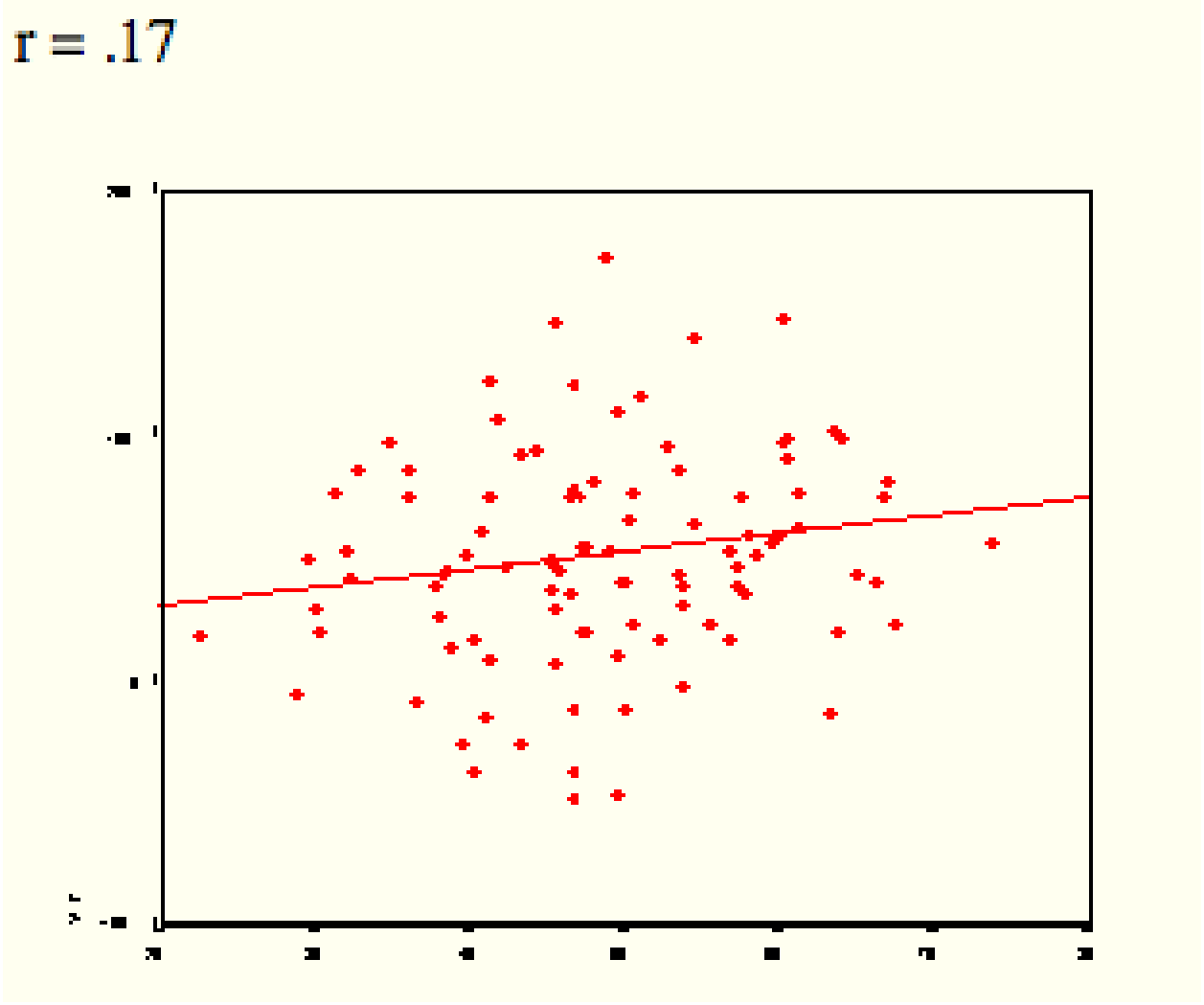
Le coefficient de corrélation varie de -1.0 à $+1.0$

Source: `http://faculty.cbu.ca/~erudiuk/IntroBook/sbk17.htm`

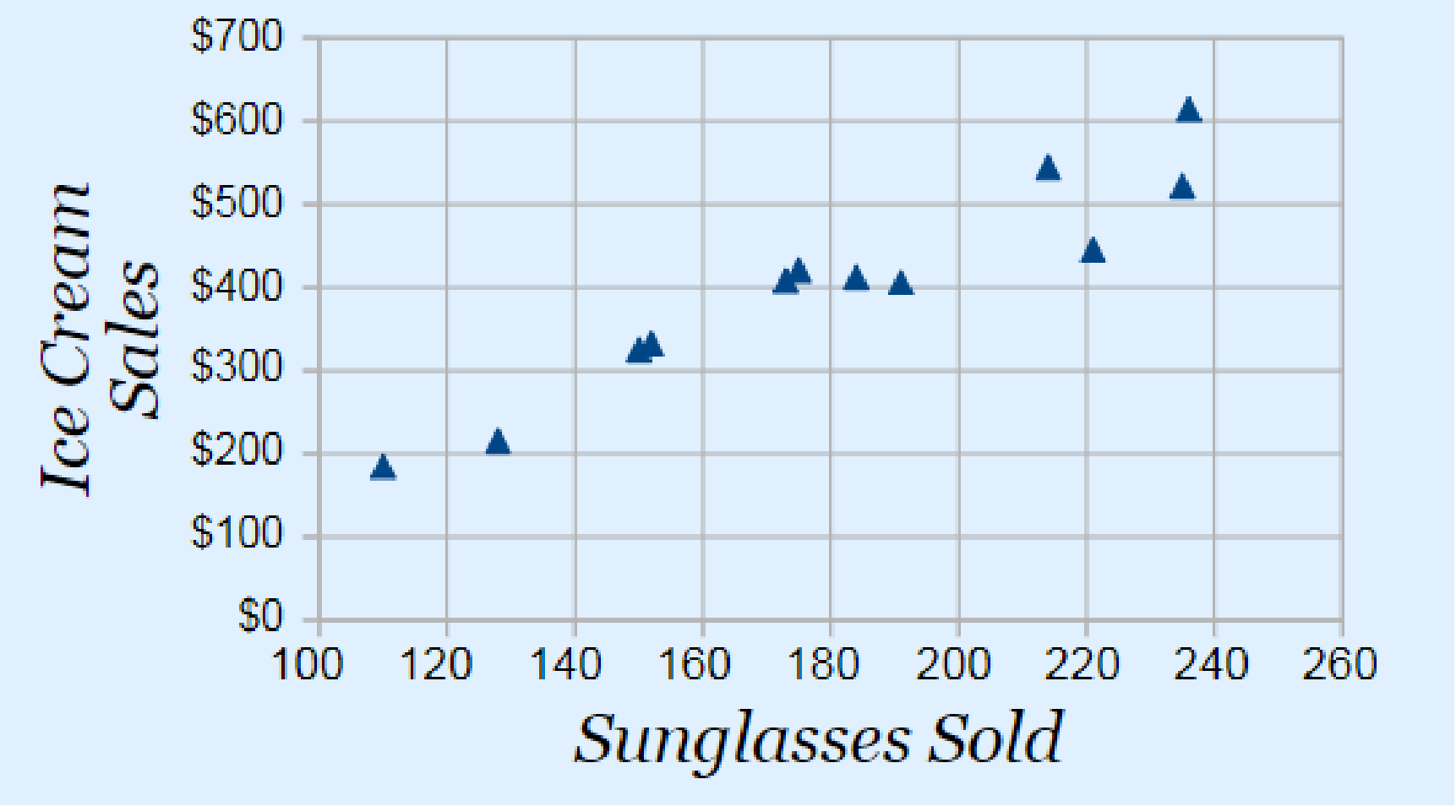


Le coefficient de corrélation varie de -1.0 à $+1.0$

Source: `http://faculty.cbu.ca/~erudiuk/IntroBook/sbk17.htm`

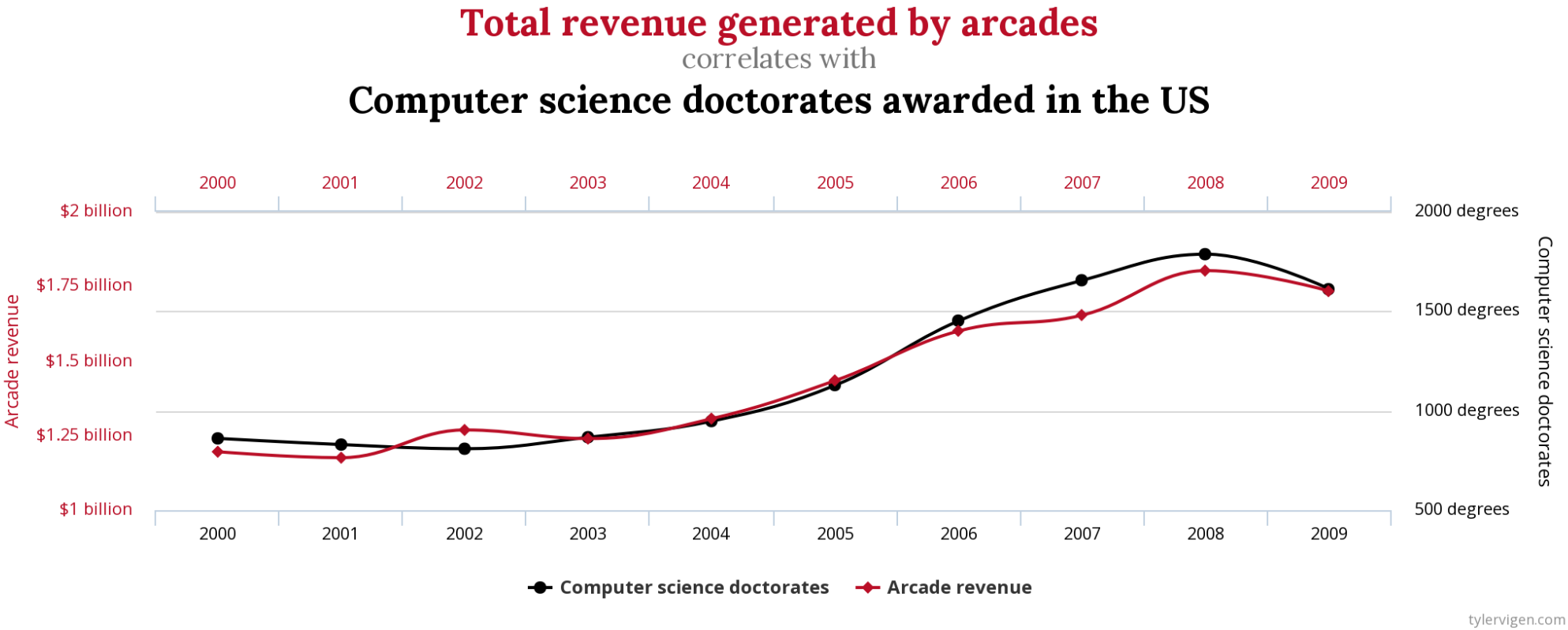


La présence d'une corrélation n'implique pas un lien de causalité !



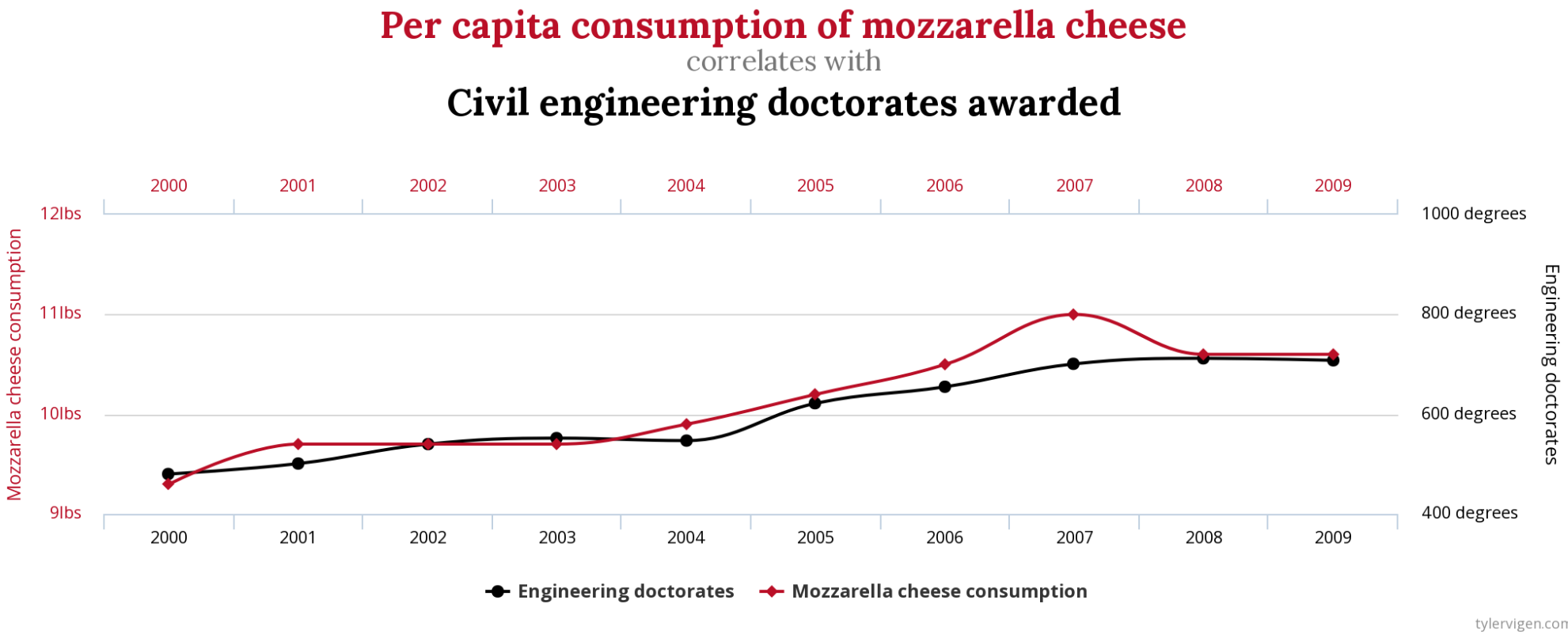
En cherchant, on trouve de nombreuses corrélations !

http://www.tylervigen.com/spurious-correlations



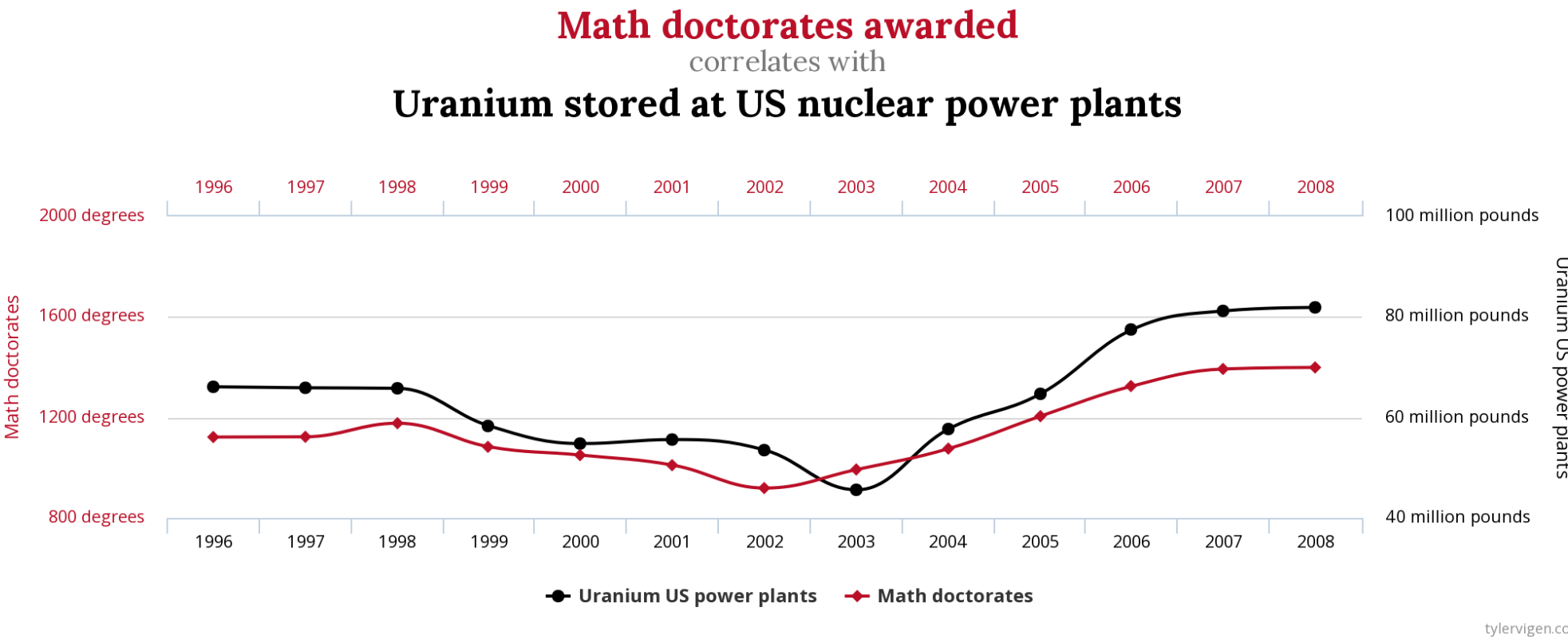
En cherchant, on trouve de nombreuses corrélations !

http://www.tylervigen.com/spurious-correlations



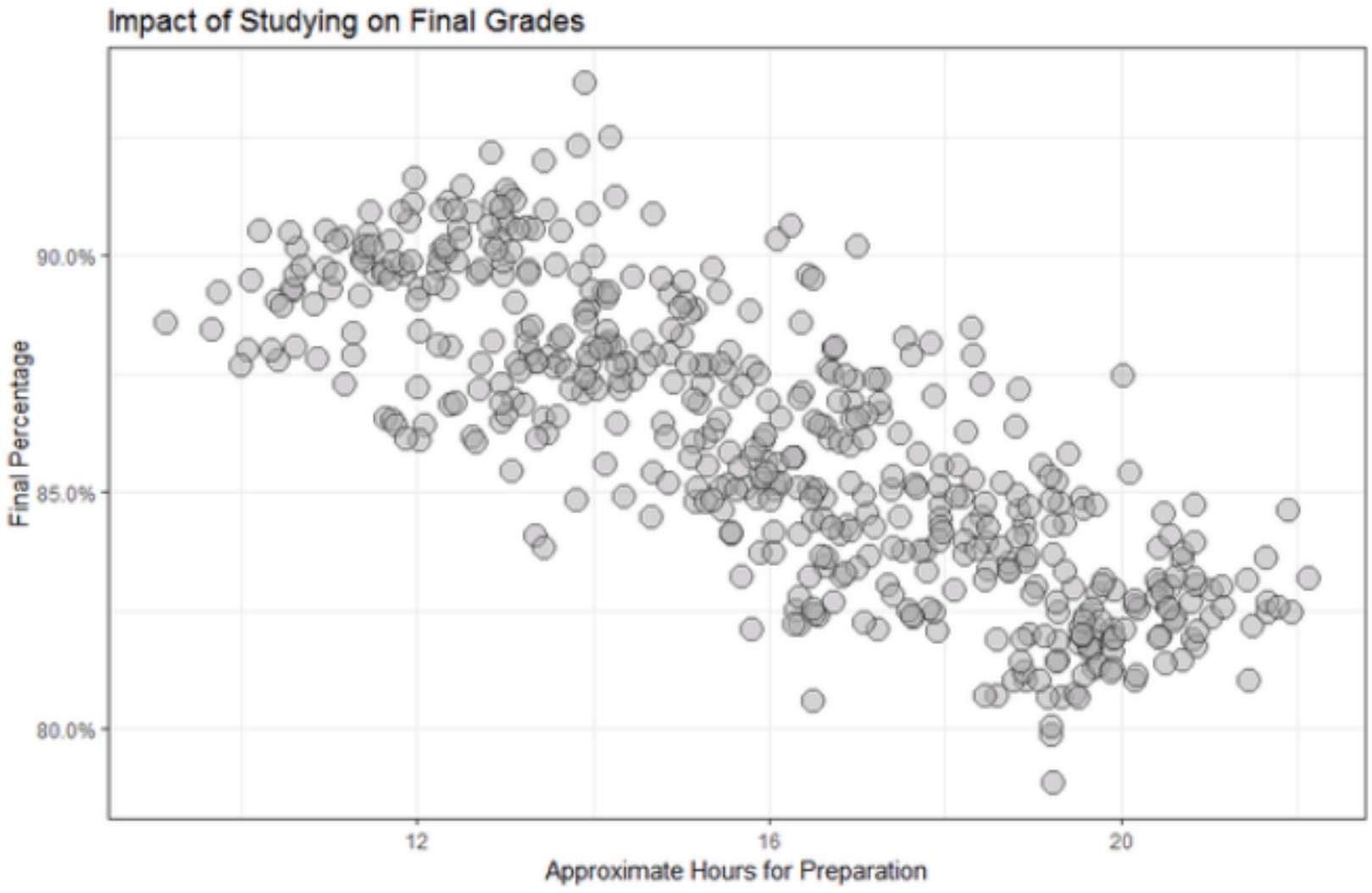
En cherchant, on trouve de nombreuses corrélations !

http://www.tylervigen.com/spurious-correlations



Coefficient de corrélation et paradoxe de Simpson ⋆

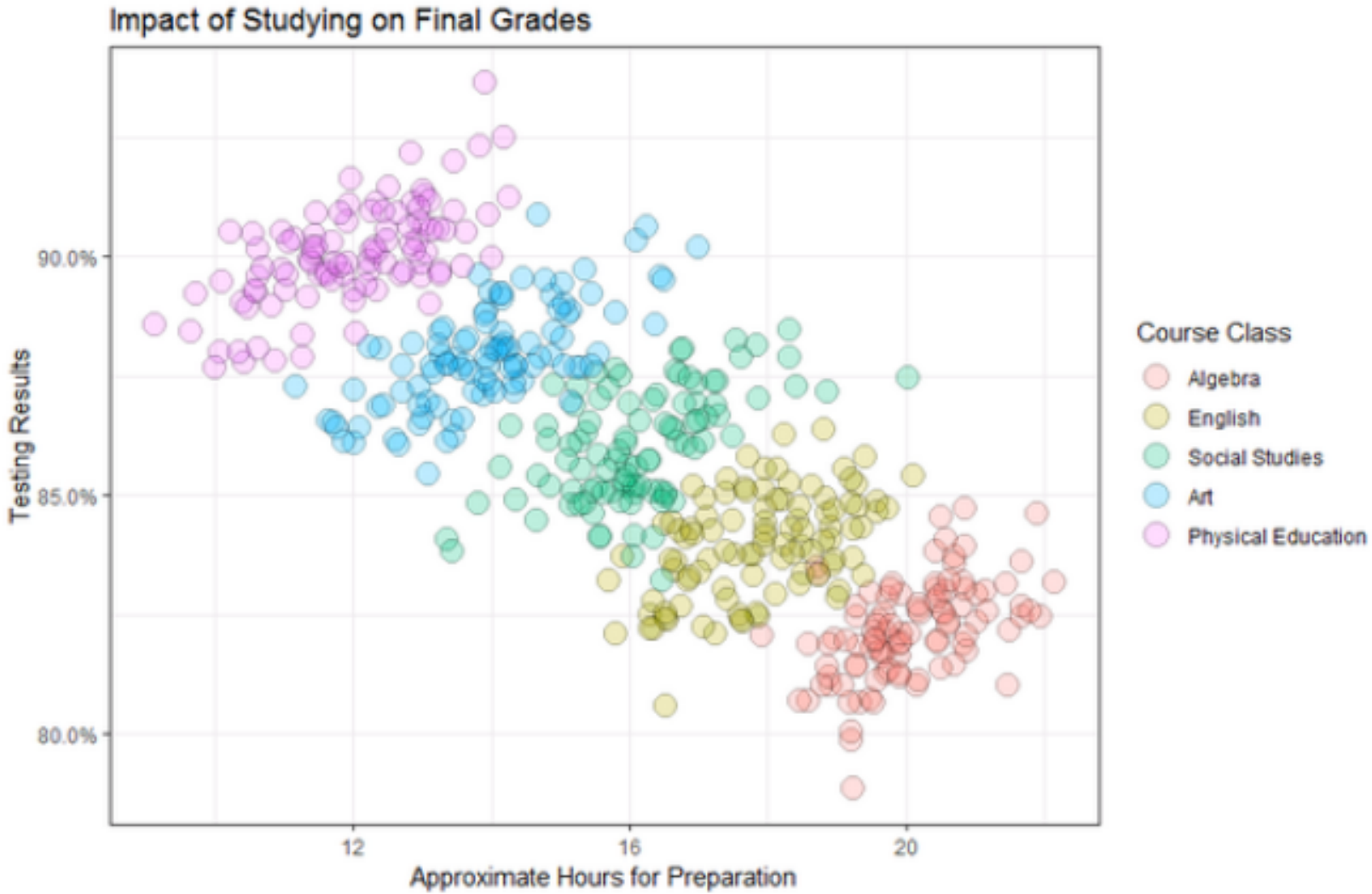
Source: `https://www.quora.com/What-is-Simpsons-paradox`



Coefficient de corrélation et paradoxe de Simpson ⋆

Corrélation négative pour l'ensemble, corrélations positives pour les sous-ensembles

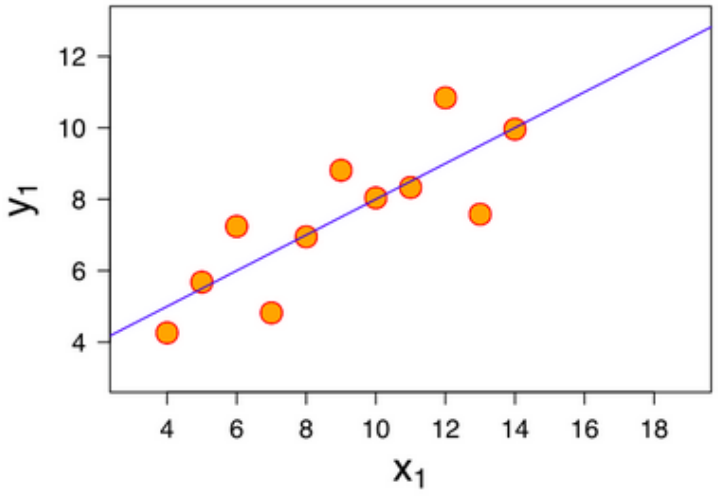
Source: `https://www.quora.com/What-is-Simpsons-paradox`

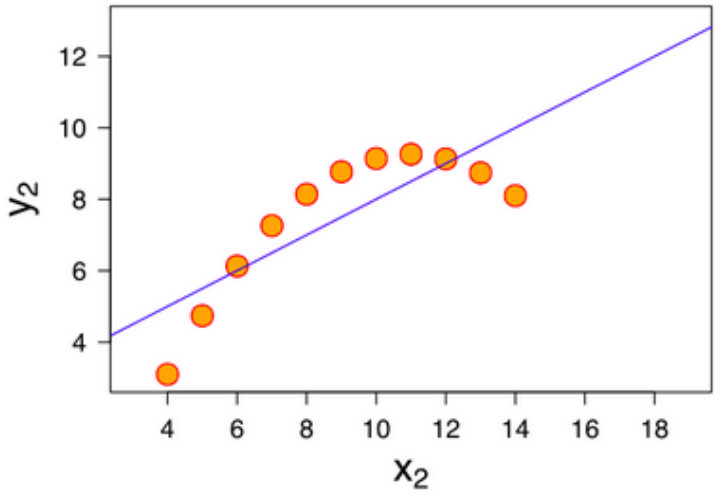


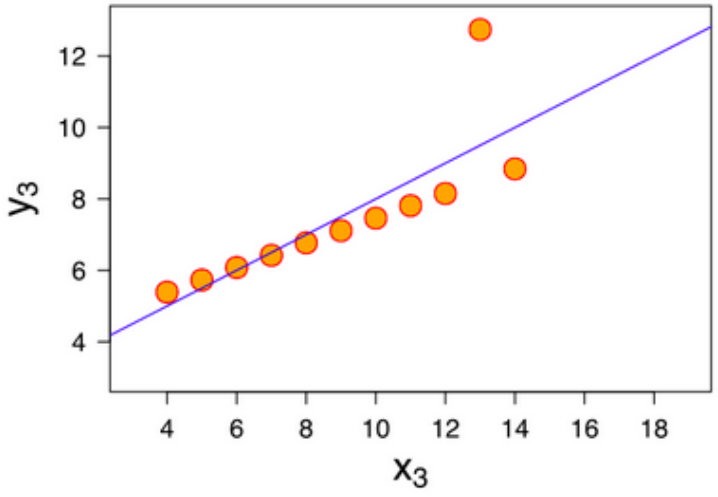
Distributions de données

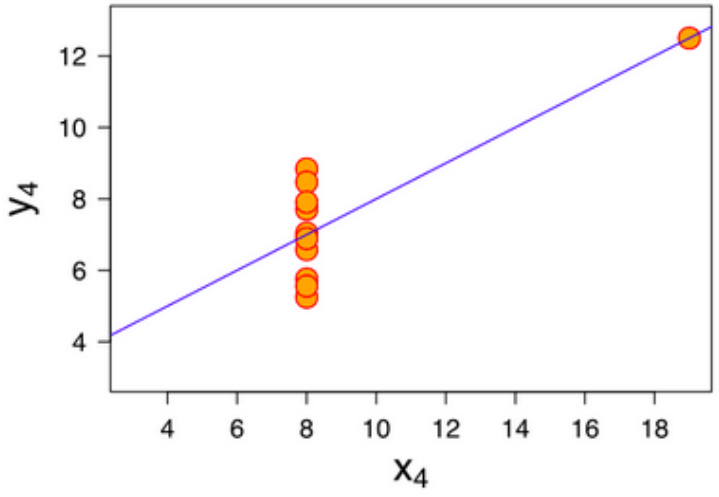
Les indicateurs sont utiles. . . mais parfois trompeurs

Qu'ont de particulier ces quatre groupes de données (*Anscomb Quartet*) ?



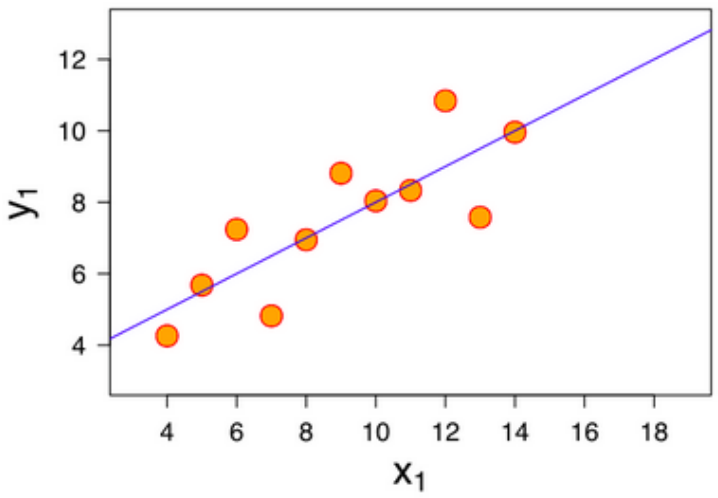


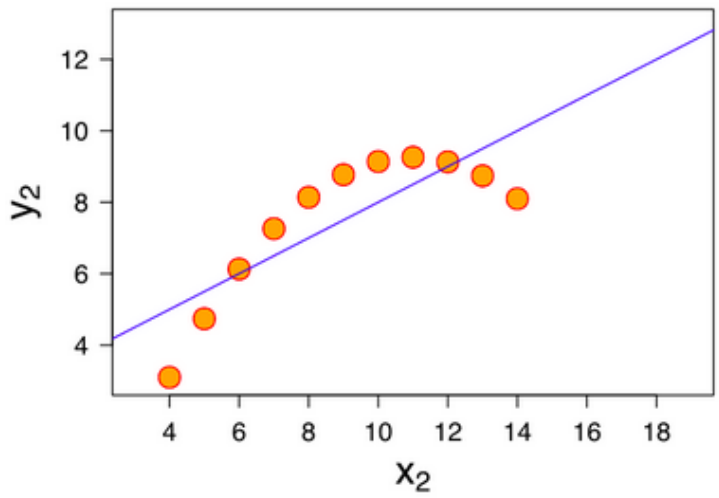


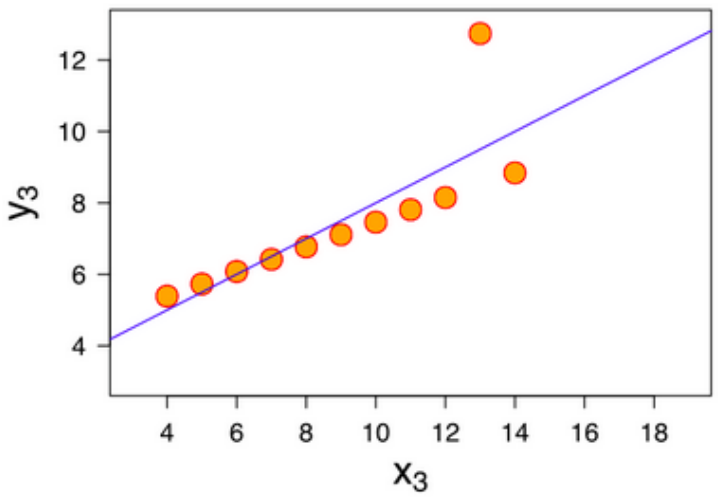


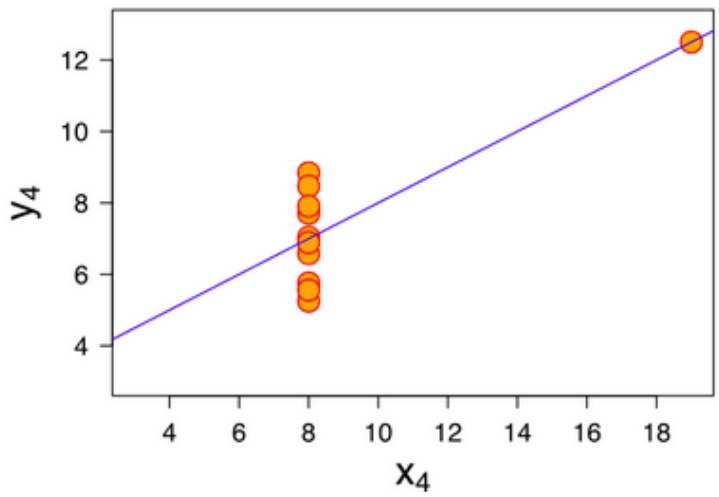
Les indicateurs sont utiles. . . mais parfois trompeurs

Qu'ont de particulier ces quatre groupes de données (*Anscomb Quartet*) ?









Mêmes moyennes, écart-types et corrélations (+0.816) !

Les indicateurs sont utiles. . . mais parfois trompeurs

Douze (12) groupes avec mêmes moyennes, écart-types et corrélations (+0.32)

«*Stat Stats, Different Graphs : Generating Datasets with Varied Appearances and Identical Statistics through Simulated Annealing*», Metjka et Fitzmaurice, 2017

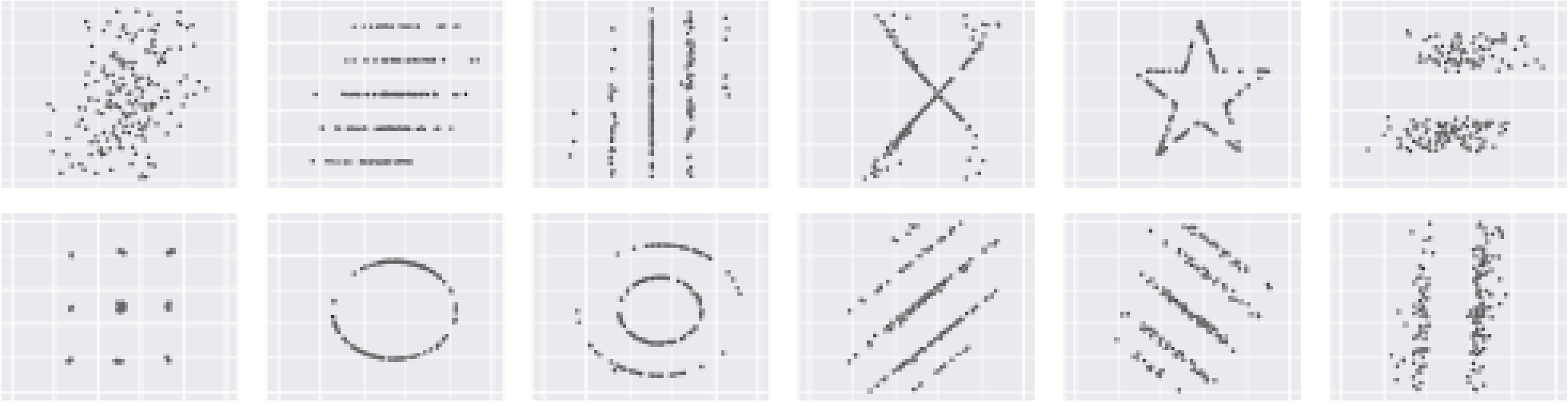


Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. ($\bar{x}$ =54.02, $\bar{y}$ = 48.09, sd_x = 14.52, sd_y = 24.79, *Pearson's r* = +0.32)

Les indicateurs sont utiles. . . mais parfois trompeurs

Douze (12) groupes avec mêmes moyennes, écart-types et corrélations (+0.32)

«*Stat Stats, Different Graphs : Generating Datasets with Varied Appearances and Identical Statistics through Simulated Annealing*», Metjka et Fitzmaurice, 2017

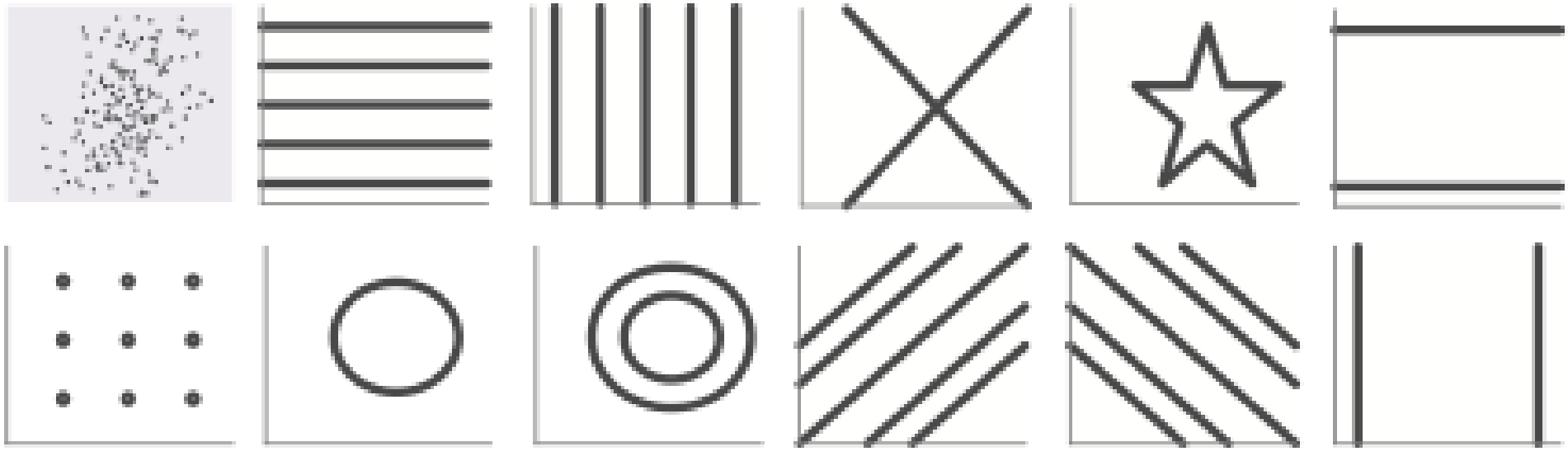
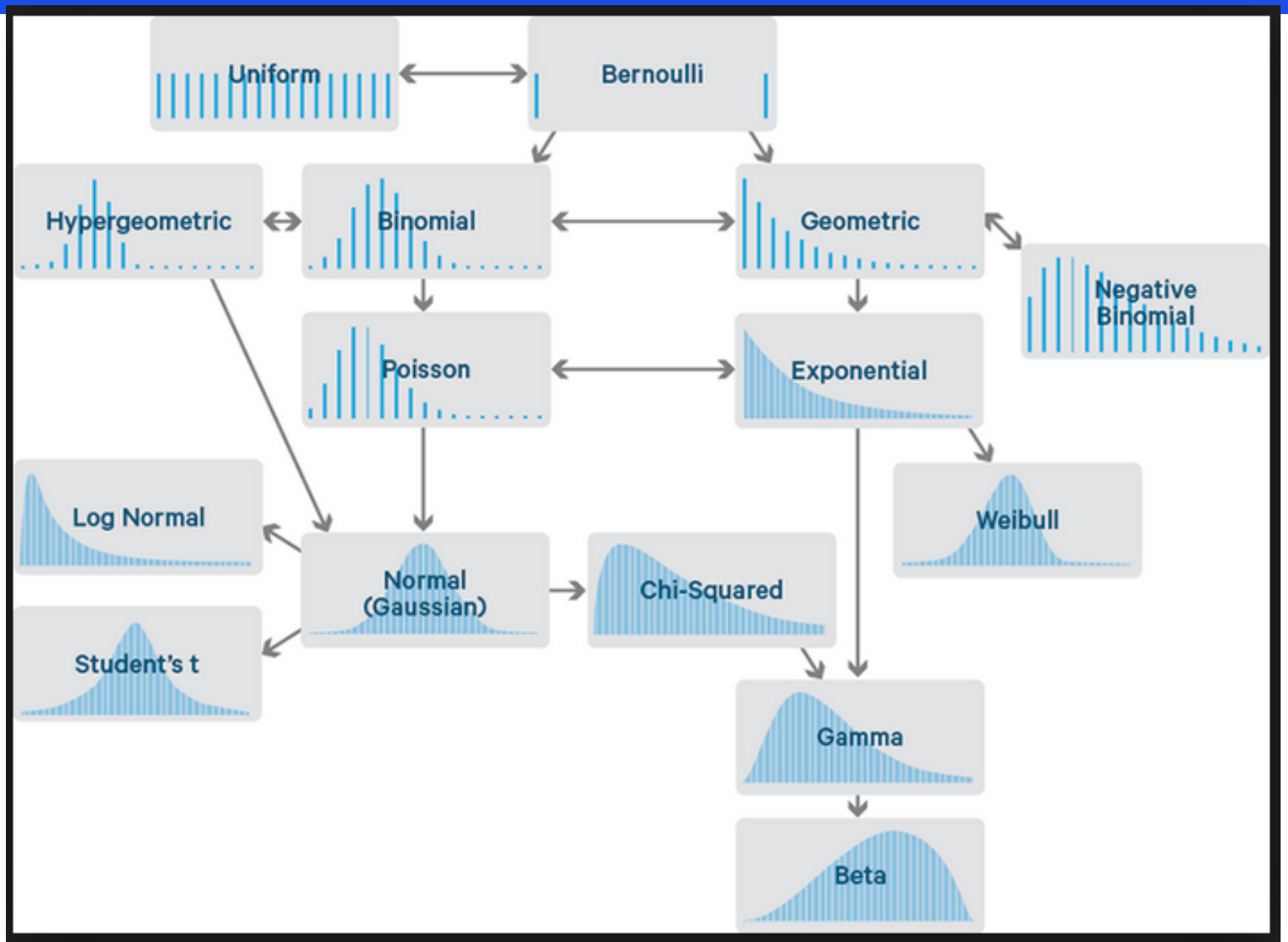
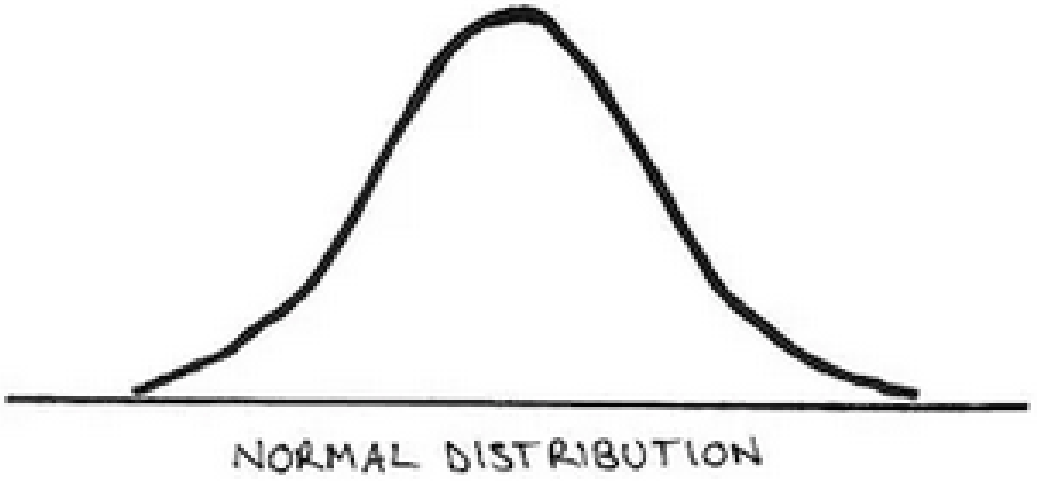


Figure 3. The initial data set (top-left), and line segment collections used for directing the output towards specific shapes. The results are seen in Figure 1.

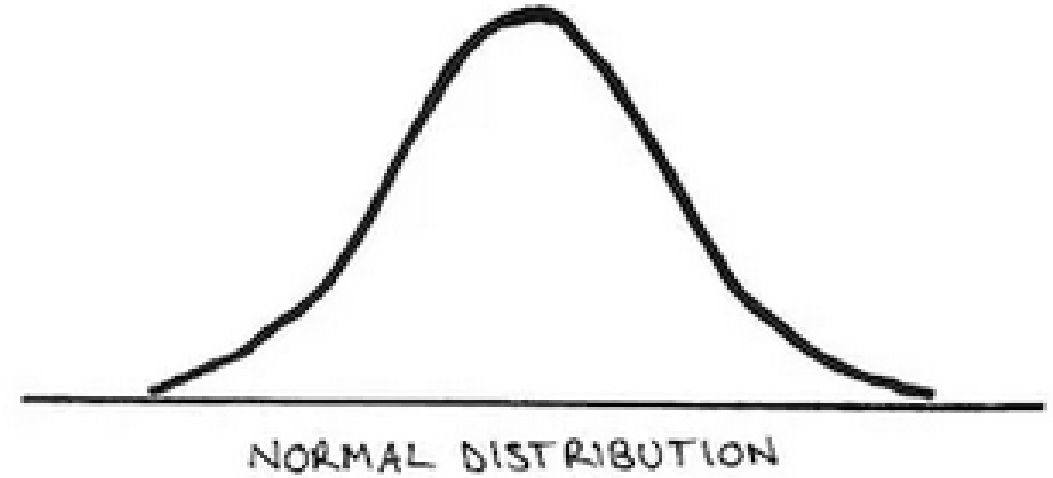
Il existe de nombreuses distributions de données

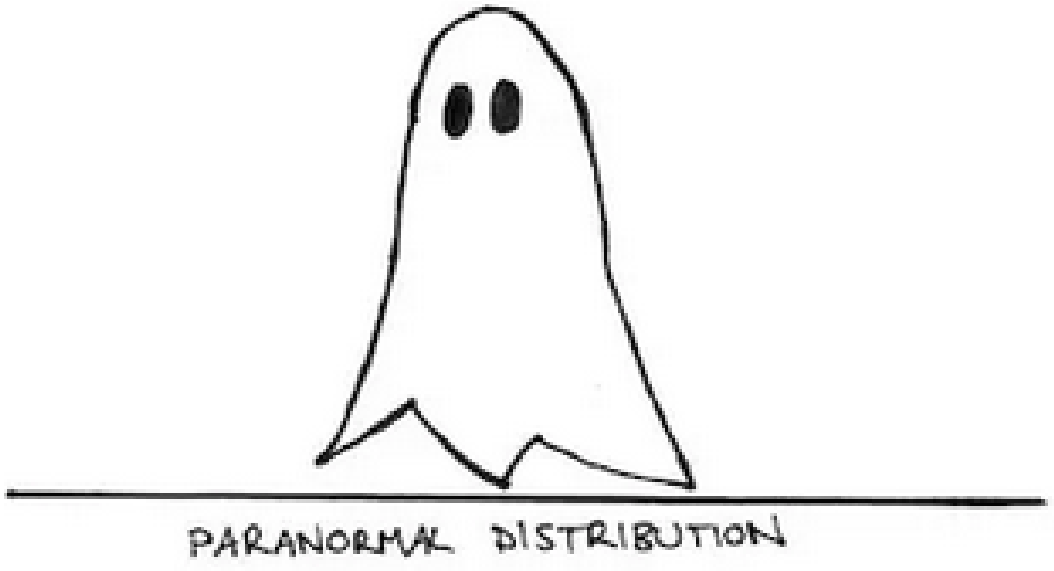


Une distribution souvent rencontrée = Distribution normale



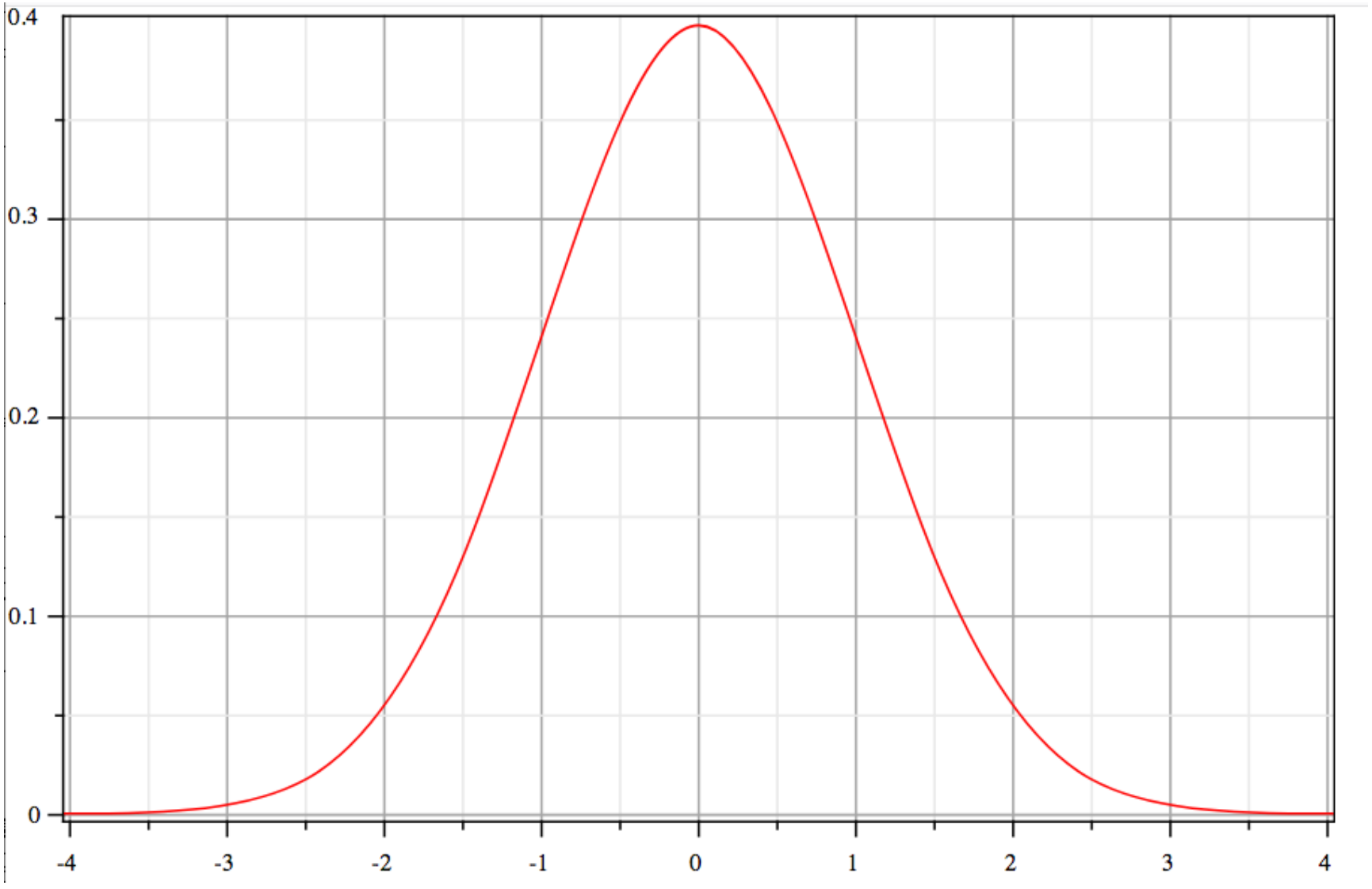
Une distribution souvent rencontrée = Distribution normale



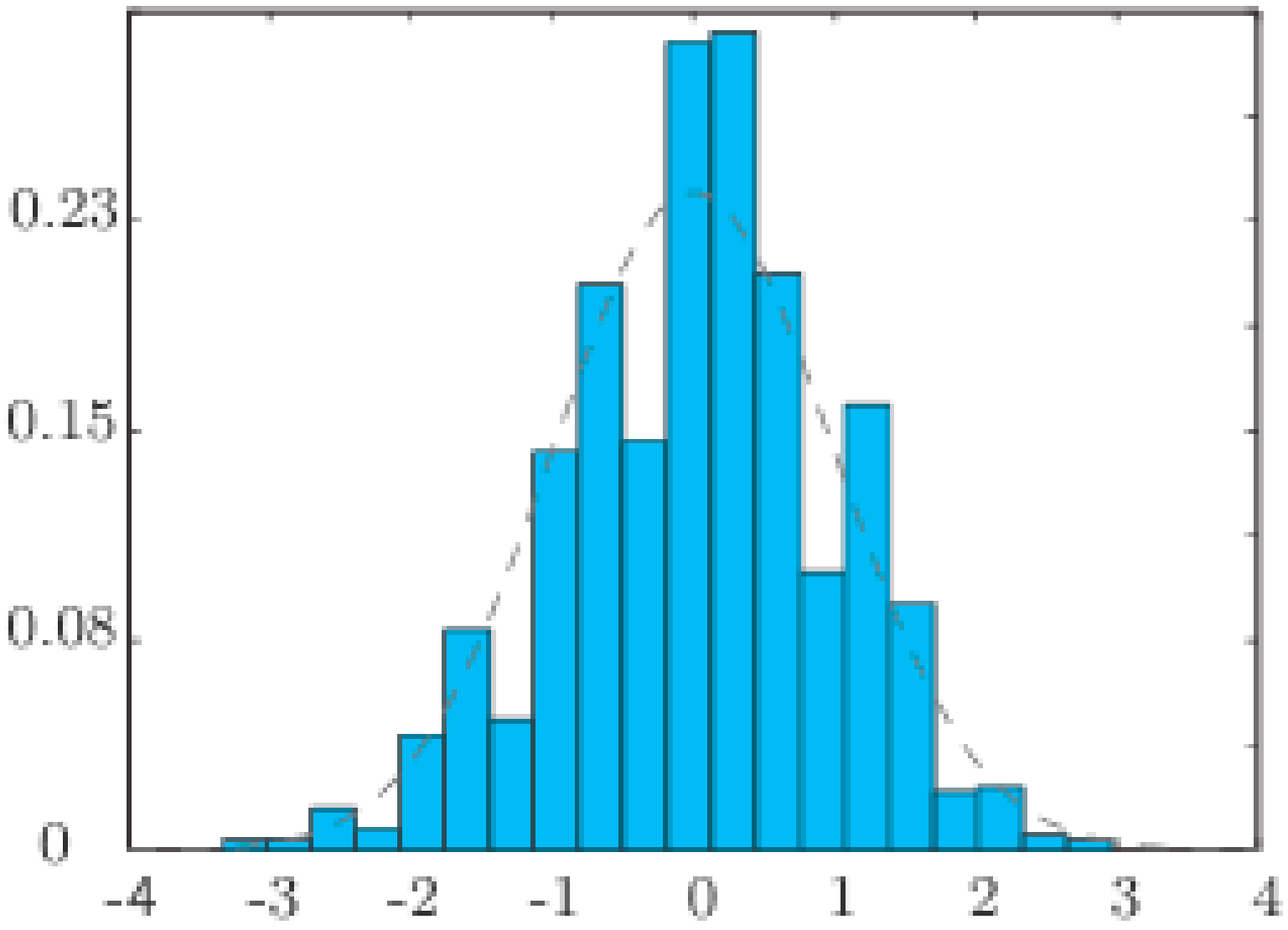


Distribution normale (continue) : $\mathcal{N}(0, 1)$

https://upload.wikimedia.org/wikipedia

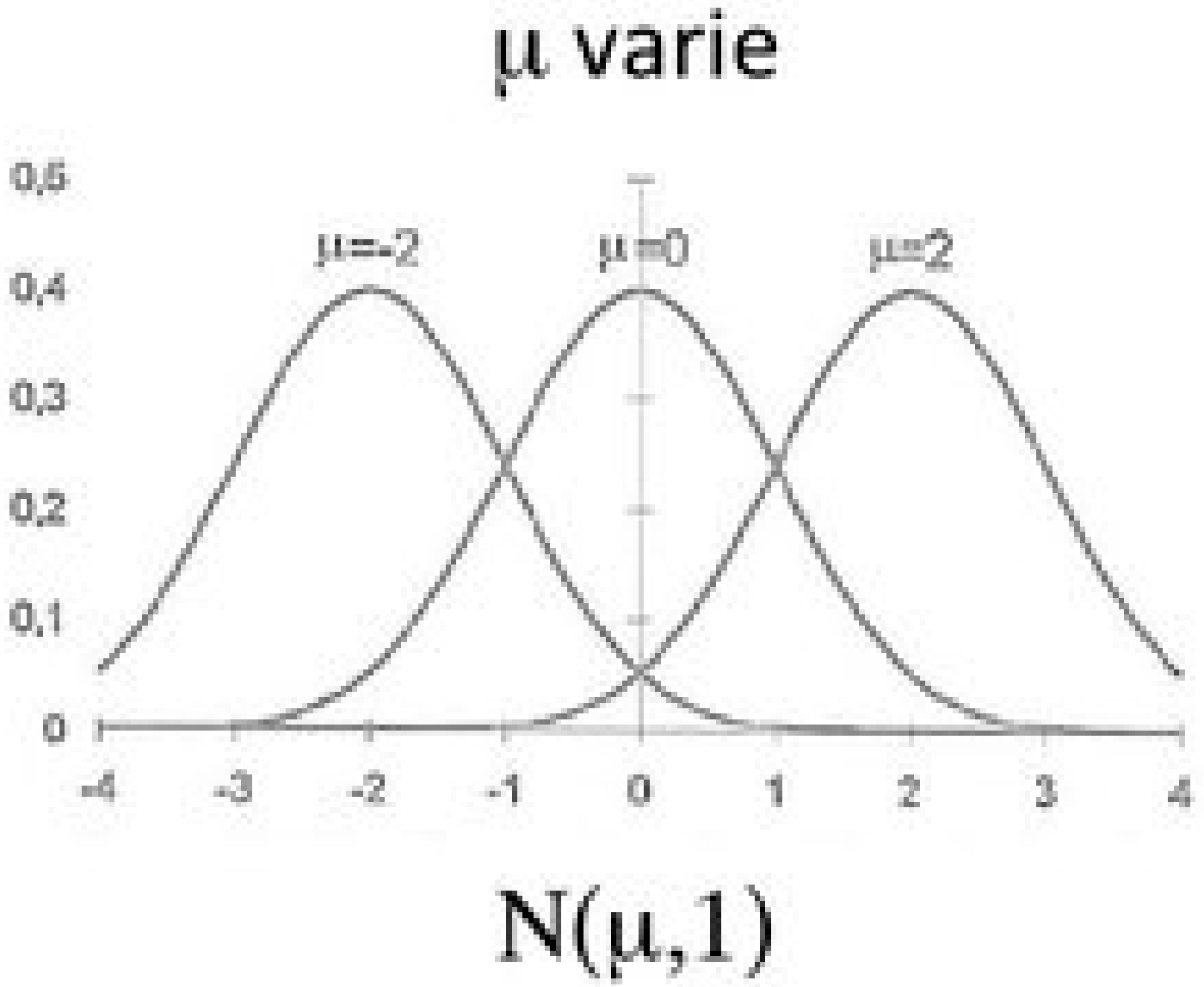


Distribution normale (discrète)



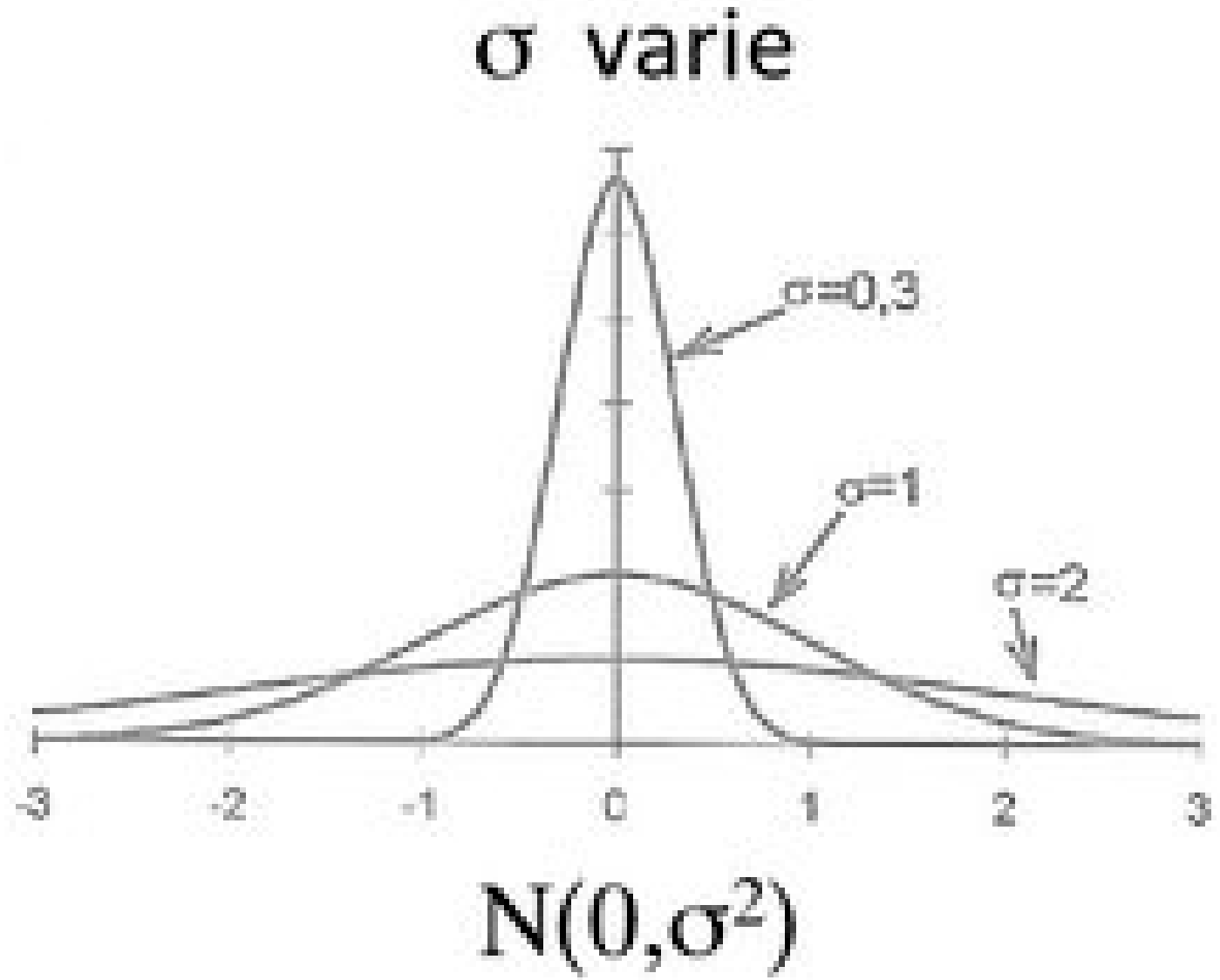
Distribution normale : Variation de μ

https://upload.wikimedia.org/wikipedia



Distribution normale : Variation de σ

https://upload.wikimedia.org/wikipedia



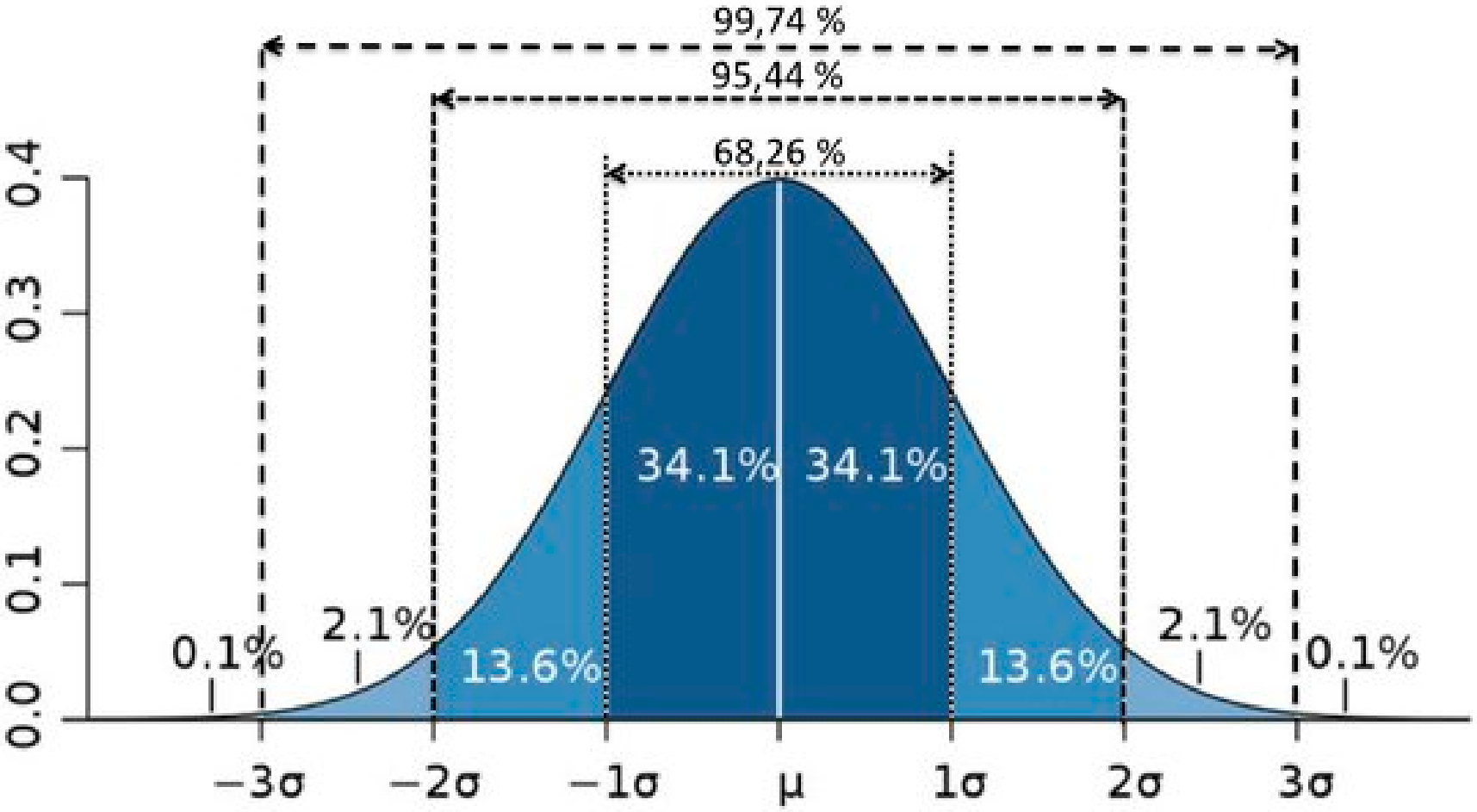
Distribution normale : $\mathcal{N}(\mu, \sigma^2)$

http://www.ilovestatistics.be/probabilite/loi-normale.html

Quelle information nous donne σ dans une distribution normale ?

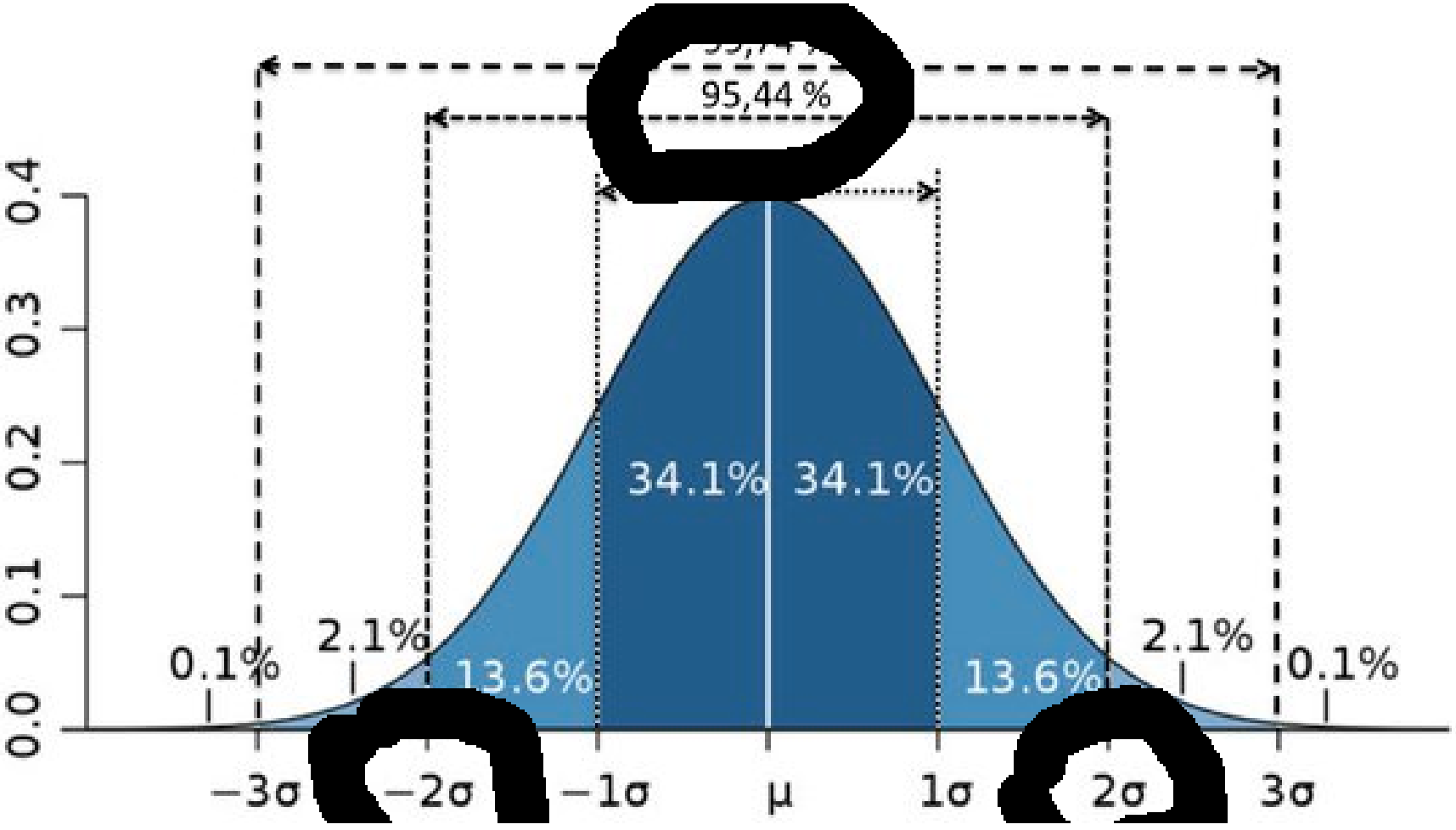
Distribution normale : $\mathcal{N}(\mu, \sigma^2)$

http://www.ilovestatistics.be/probabilite/loi-normale.html



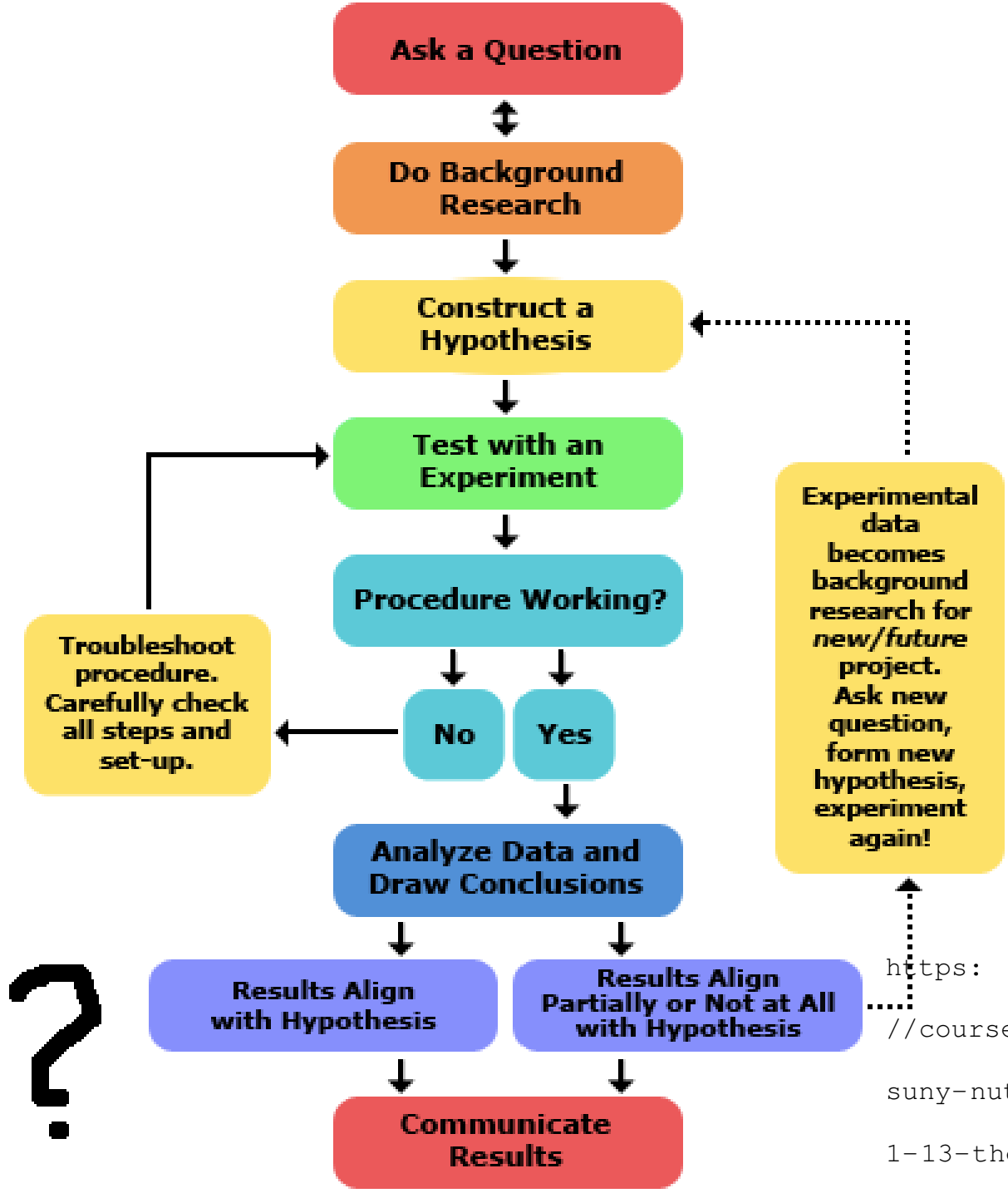
Distribution normale : $\mathcal{N}(\mu, \sigma^2)$

http://www.ilovestatistics.be/probabilite/loi-normale.html



Aperçu

La méthode scientifique



https://courses.lumenlearning.com/suny-nutrition/chapter/1-13-the-scientific-method/

Pourquoi les statistiques sont souvent utilisées ?

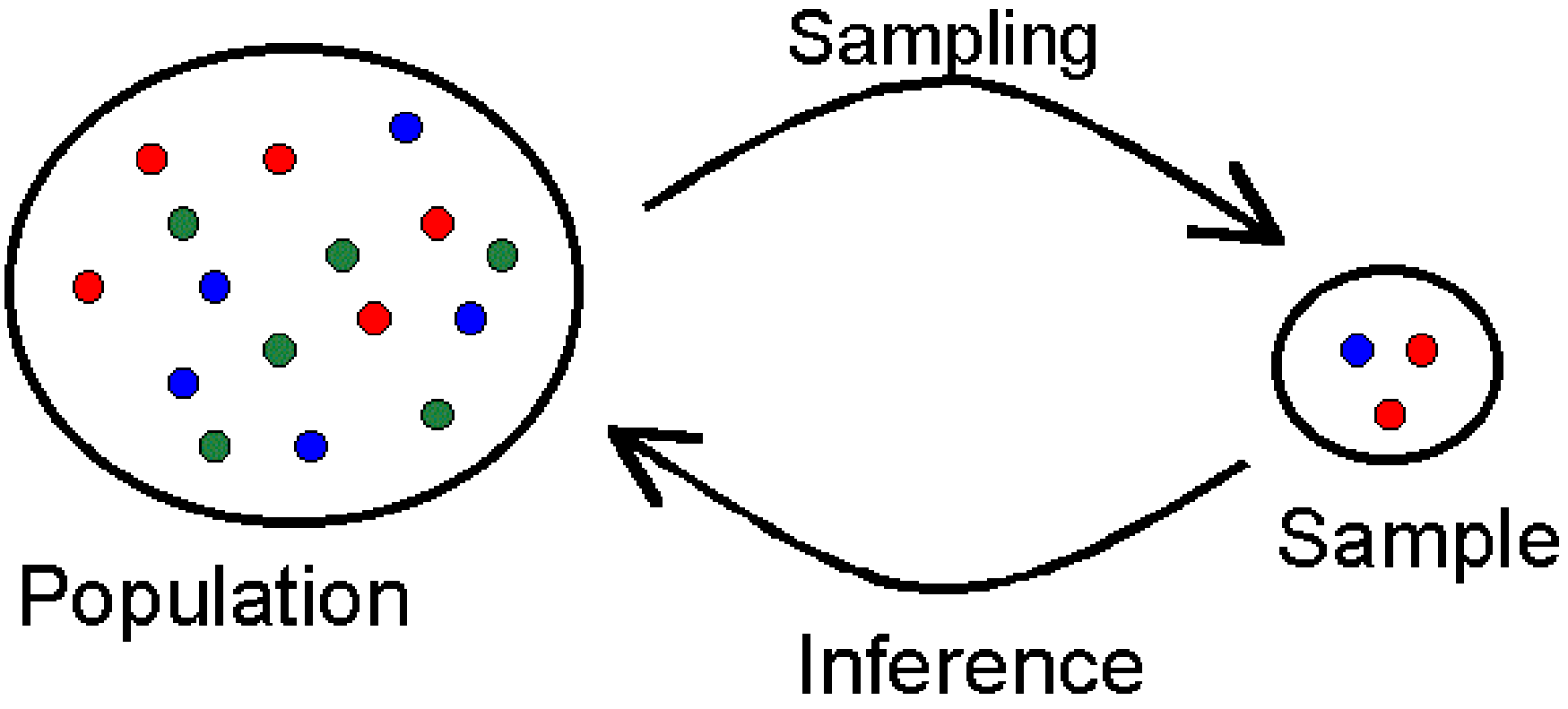
"Data don't make any sense,
we will have to resort to statistics."

Pourquoi les statistiques sont souvent utilisées ?

- Phénomènes irréguliers, aléatoires, ...
- Imprécision des mesures expérimentales
- Raisonnement sur des échantillons
- ...

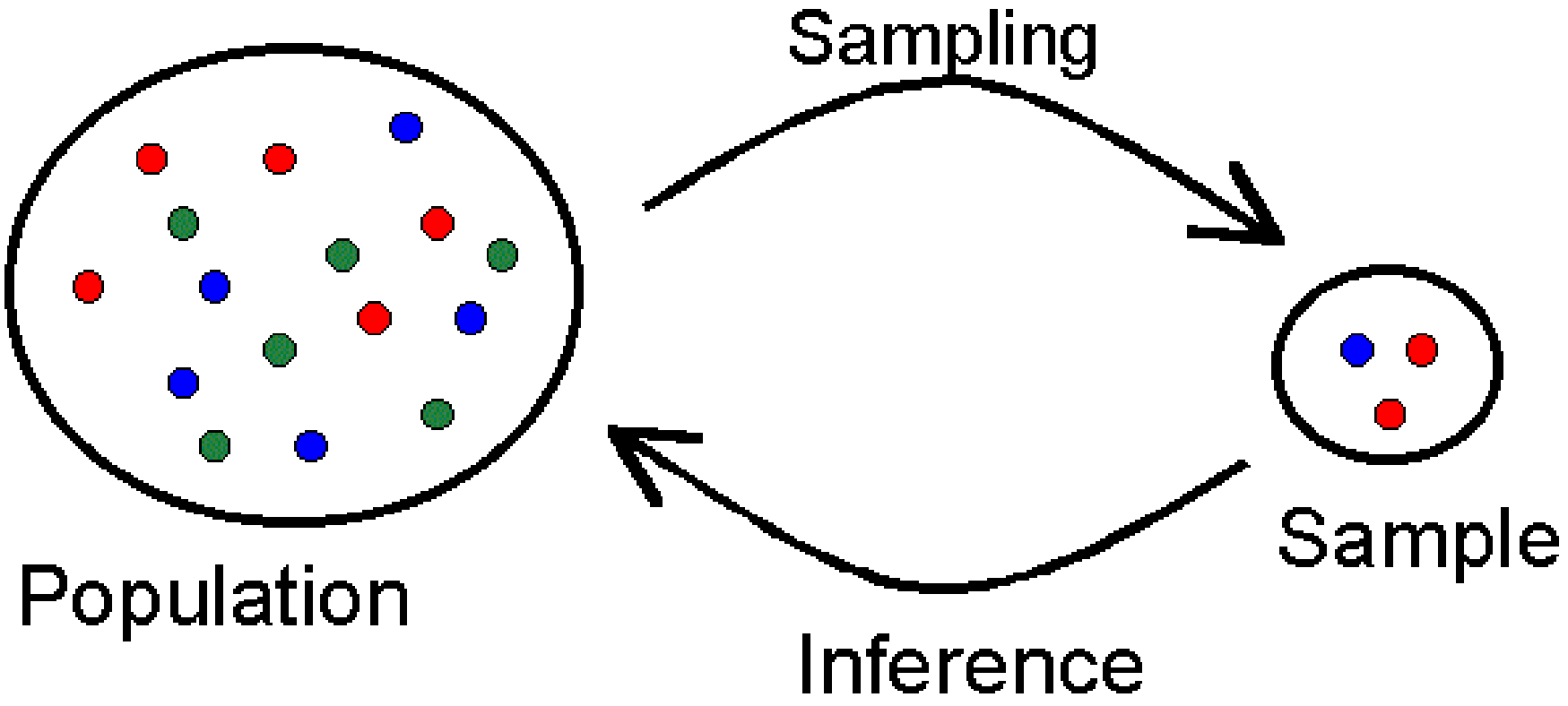
Pourquoi les statistiques sont souvent utilisées ?

http://palin.co.in/difference-between-population-and-sampling-with-example



Pourquoi les statistiques sont souvent utilisées ?

http://palin.co.in/difference-between-population-and-sampling-with-example



Objectif de l'inférence statistique

Pouvoir affirmer, avec une certaine «confiance», qu'un phénomène (effet) **n'est pas entièrement dû au hasard**

Un exemple (inventé) en lien avec le génie logiciel et l'enseignement

Description du contexte (fictif !)

Cours INF3456 utilisant le langage *L*

- Offert au bac depuis 9 trimestres
- $\approx$ 30–40 étudiant.e.s par trimestre
- Langage utilisé = *L*
- Aucun IDE disponible pour *L*, mais. . .

Description du contexte (fictif !)

Cours INF3456 utilisant le langage *L*

- Offert au bac depuis 9 trimestres
- $\approx$ 30–40 étudiant.e.s par trimestre
- Langage utilisé = *L*
- Aucun IDE disponible pour *L*, mais. . .

Nouvel IDE pour le langage *L*

- Un professeur a développé un nouvel IDE pour *L*
- Il aimerait savoir si l'utilisation de l'IDE améliore l'apprentissage de *L* par les étudiant.e.s

Description de l'expérience

Données déjà connues ≈ Population

Données connues

- On connait les résultats des neuf (9) trimestres antérieurs (300 étudiant.e.s) :

⇒ moyenne = 69.78 % (écart-type = 9.72)

```
[40- 45): **
[45- 50): *****
[50- 55): ***********
[55- 60): ******************************
[60- 65): ****************************************
[65- 70): ************************************************************
[70- 75): ************************************************************
[75- 80): ****************************************
[80- 85): ***************************
[85- 90): ************
[90- 95): *
[95-100): ***
```

Description de l'expérience

Résultats obtenus à l'hiver 2019 = Échantillon

Résultats obtenus suite à l'utilisation du nouvel IDE

- Nombre d'étudiant.e.s = 30
- Moyenne = 73.22 % (écart-type = 14.11)

```
[35- 40): *
[40- 45):
[45- 50): *
[50- 55):
[55- 60): **
[60- 65): **
[65- 70): ******
[70- 75): *******
[75- 80): **
[80- 85): ****
[85- 90): *
[90- 95): **
[95-100): **
```

Que peut-on conclure quant à l'utilisation de l'IDE ?

Résultats sans IDE (300 étudiant.e.s)

- Moyenne = 69.78 %
- Écart-type = 9.72

Résultats avec IDE (30 étudiant.e.s)

- Moyenne = 73.22 %
- Écart-type = 14.11

Que peut-on conclure quant à l'utilisation de l'IDE ?

Résultats sans IDE (300 étudiant.e.s)

- Moyenne = 69.78 %
- Écart-type = 9.72

Résultats avec IDE (30 étudiant.e.s)

- Moyenne = 73.22 %
- Écart-type = 14.11

1. Aide les étudiant.e.s ?
   (la moyenne **semble** avoir augmenté)

Que peut-on conclure quant à l'utilisation de l'IDE ?

Résultats sans IDE (300 étudiant.e.s)

- Moyenne = 69.78 %
- Écart-type = 9.72

Résultats avec IDE (30 étudiant.e.s)

- Moyenne = 73.22 %
- Écart-type = 14.11

1. Aide les étudiant.e.s ?
   (la moyenne **semble** avoir augmenté)
2. Aide certaines personnes, mais désavantage d'autres ?
   (l'écart-type semble avoir augmenté)

Que peut-on conclure quant à l'utilisation de l'IDE ?

Résultats sans IDE (300 étudiant.e.s)

- Moyenne = 69.78 %
- Écart-type = 9.72

Résultats avec IDE (30 étudiant.e.s)

- Moyenne = 73.22 %
- Écart-type = 14.11

1. Aide les étudiant.e.s ?
   (la moyenne **semble** avoir augmenté)
2. Aide certaines personnes, mais désavantage d'autres ?
   (l'écart-type semble avoir augmenté)
3. Aucun effet ?
   (différences dûes uniquement «au hasard» (échantillonnage))

Approche NHST pour l'inférence statistique (sur la moyenne)

Null Hypothesis Significance Testing

On formule l'hypothèse qu'on veut vérifier

- H : L'utilisation de l'IDE permet d'améliorer la moyenne

Approche NHST pour l'inférence statistique (sur la moyenne)

Null Hypothesis Significance Testing

On formule l'hypothèse qu'on veut vérifier

- H : L'utilisation de l'IDE permet d'améliorer la moyenne

On formule une hypothèse nulle (aucun effet = **juste le hasard !**)

- H_0 : L'utilisation de l'IDE... n'a aucun effet sur la moyenne

Approche NHST pour l'inférence statistique

Reductio ad unlikely

On procède à une sorte de raisonnement «par l'absurde» (*reductio ad absurdum*) mais avec des statistiques

- Supposons que l'hypothèse nulle (juste le hasard) soit vraie.

Approche NHST pour l'inférence statistique

Reductio ad unlikely

On procède à une sorte de raisonnement «par l'absurde» (*reductio ad absurdum*) mais avec des statistiques

- Supposons que l'hypothèse nulle (juste le hasard) soit vraie.
- Est-ce «surprenant» d'avoir obtenu les résultats observés ?

Approche NHST pour l'inférence statistique

Reductio ad unlikely

On procède à une sorte de raisonnement «par l'absurde» (*reductio ad absurdum*) mais avec des statistiques

- Supposons que l'hypothèse nulle (juste le hasard) soit vraie.
- Est-ce «surprenant» d'avoir obtenu les résultats observés ?
  - Si le résultat **n'est pas surprenant**,
    alors on ne rejette pas l'hypothèse nulle :
    l'intervention ne semble pas avoir d'intérêt ☹

Approche NHST pour l'inférence statistique

Reductio ad unlikely

On procède à une sorte de raisonnement «par l'absurde» (*reductio ad absurdum*) mais avec des statistiques

- Supposons que l'hypothèse nulle (juste le hasard) soit vraie.
- Est-ce «surprenant» d'avoir obtenu les résultats observés ?
  - Si le résultat n'est pas surprenant,
    alors on ne rejette pas l'hypothèse nulle :
    l'intervention ne semble pas avoir d'intérêt ☹

    Le hasard permet d'expliquer le résultat !

Approche NHST pour l'inférence statistique

Reductio ad unlikely

On procède à une sorte de raisonnement «par l'absurde» (*reductio ad absurdum*) mais avec des statistiques

- Supposons que l'hypothèse nulle (juste le hasard) soit vraie.
- Est-ce «surprenant» d'avoir obtenu les résultats observés ?
  - Si le résultat n'est pas surprenant,
    alors on ne rejette pas l'hypothèse nulle :
    l'intervention ne semble pas avoir d'intérêt ☹
  - Si le résultat est très (!) **«surprenant !»**,
    alors on rejette l'hypothèse nulle :
    l'intervention semble avoir un intérêt ☺

Approche NHST appliquée à notre exemple (IDE pour L)

Caractéristiques de la population avec H_0

Soit une population avec :

- Moyenne = 69.78%
- Écart-type = 9.72

Propriété statistique (distribution d'échantillonnage de la moyenne)

Si on prend divers échantillons de taille N, alors les moyennes vont suivre une loi

$$\mathcal{N}(69.78\%, \frac{9.72^2}{N}))$$

Approche NHST appliquée à notre exemple (IDE pour L)

Caractéristiques de la population avec H_0

Soit une population avec :

- Moyenne = 69.78%
- Écart-type = 9.72

Propriété statistique (distribution d'échantillonnage de la moyenne)

Si on prend divers échantillons de taille N, alors les moyennes vont suivre une loi

$$\mathcal{N}(69.78\%, \frac{9.72^2}{N}))$$

Donc, pour $N = 30$ (notre échantillon) :

- Moyenne = 69.78%
- Écart-type = $\frac{9.72}{\sqrt{30}} = 1.774621$

Un échantillon dont la moyenne est de 73.22 — donc avec un écart de 3.44 (73.22 − 69.78) — est-il surprenant ?

Rappel

$X \sim \mathcal{N}(69.78, 1.774621^2)$

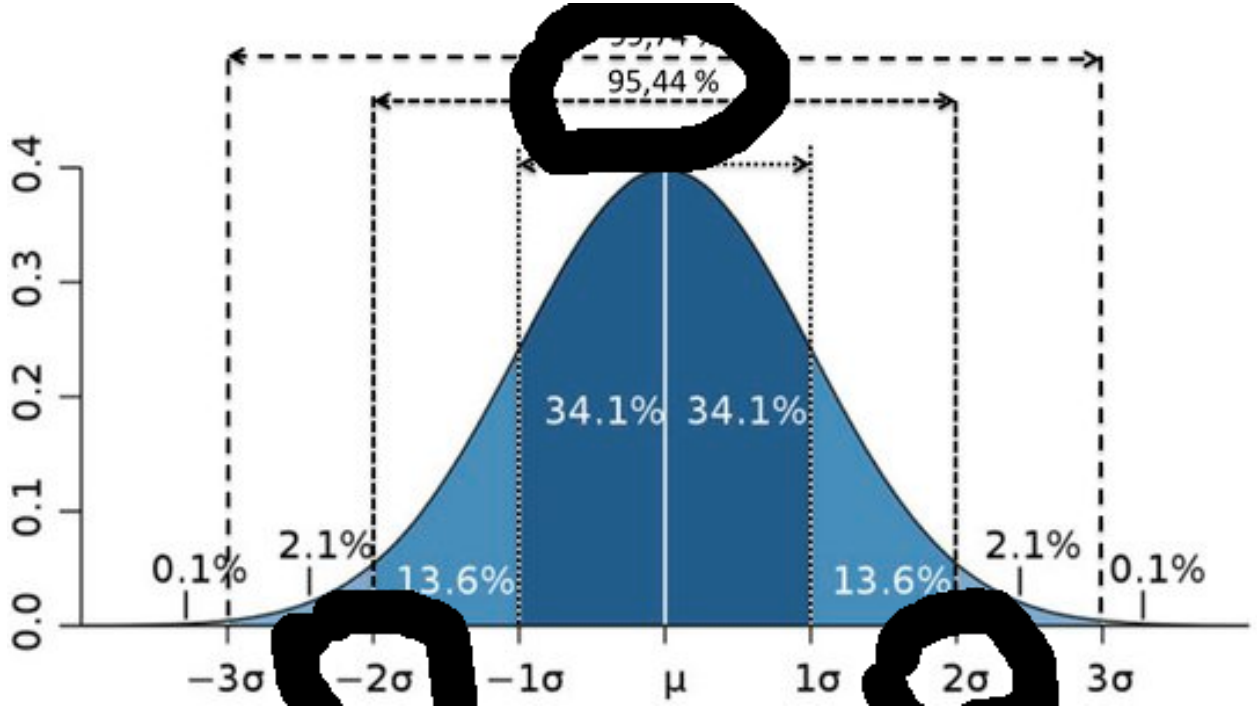
- Écart de 3.44 $\Rightarrow$ 1.9384 écart-types

Un échantillon dont la moyenne est de 73.22 — donc avec un écart de 3.44 (73.22 − 69.78) — est-il surprenant ?

Rappel

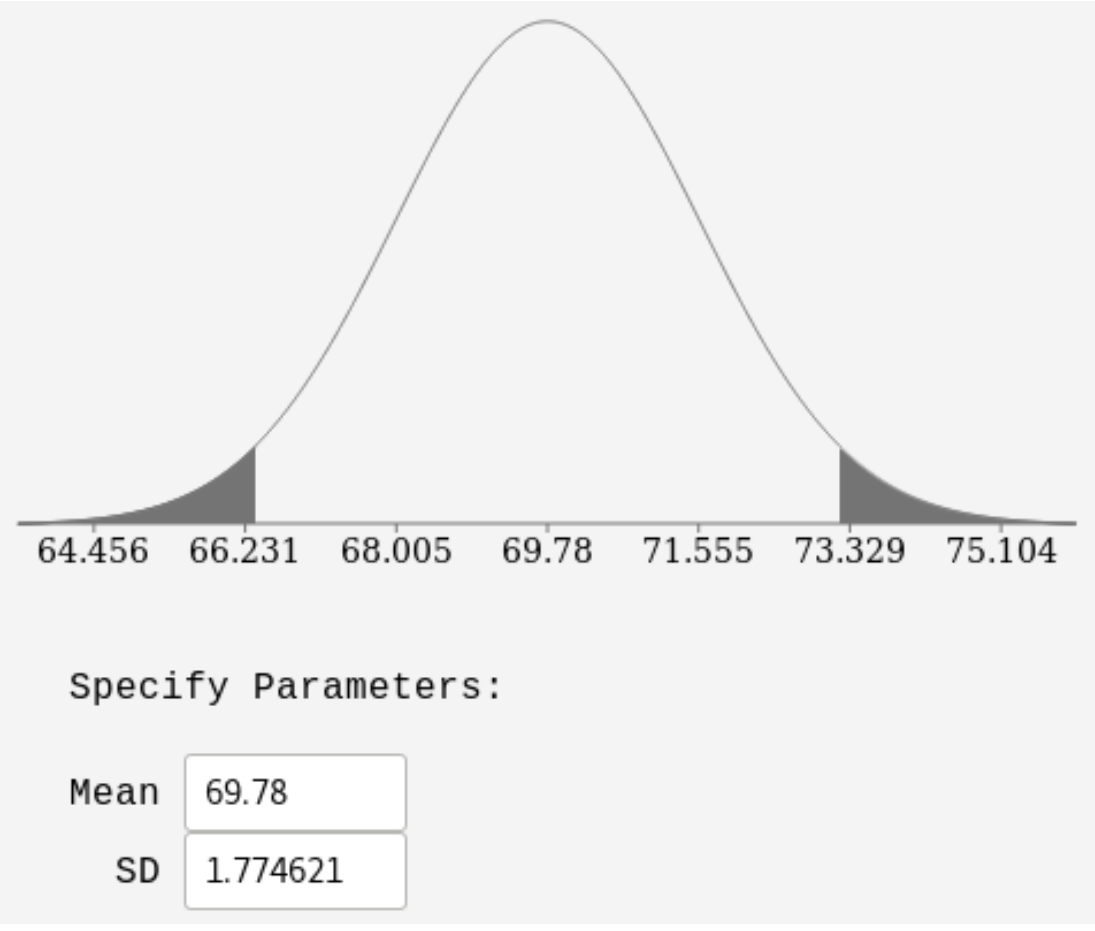
$X \sim \mathcal{N}(69.78, 1.774621^2)$

- Écart de 3.44 $\Rightarrow$ 1.9384 écart-types



Plus précisément : Un échantillon dont la moyenne a un écart supérieur à 3.44 $(73.22 - 69.78)$ est-il surprenant ?

Plus précisément : Un échantillon dont la moyenne a un écart supérieur à 3.44 (73.22 – 69.78) est-il surprenant ?



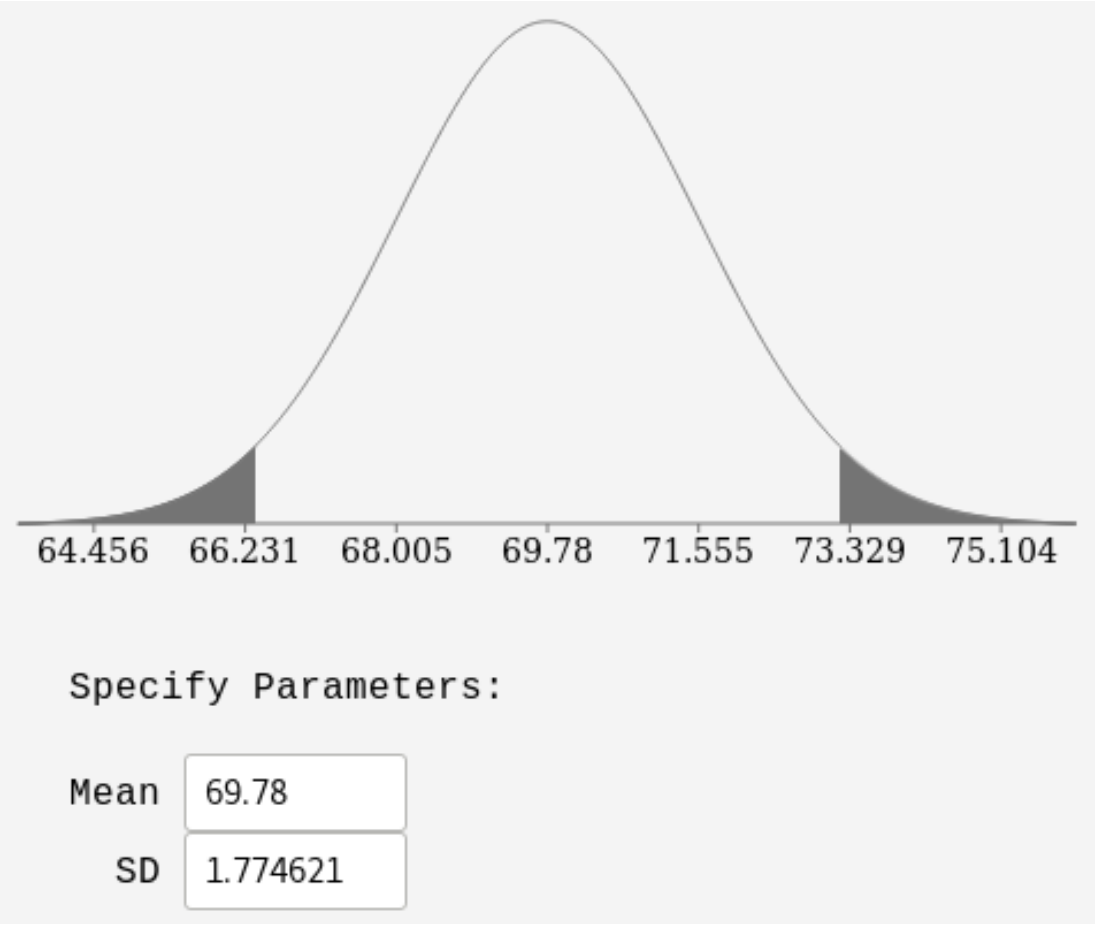
Results:

Area (probability) = 0.0526

Recalculate

Plus précisément : Un échantillon dont la moyenne a un écart supérieur à 3.44 (73.22 – 69.78) est-il surprenant ?

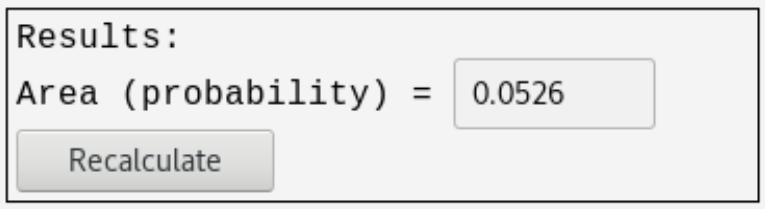
Écart > 3.44 $\Rightarrow$ 1.9384 écart-type ou + $\Rightarrow$ **p-value =** 0.0526 $>$ 0.05 ☹



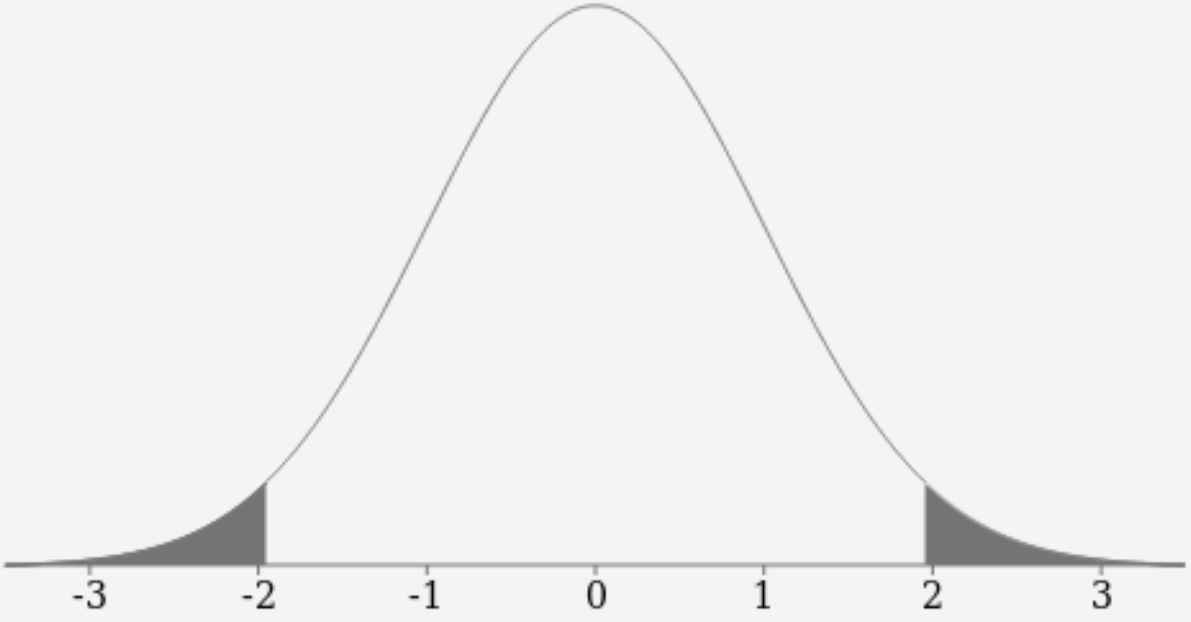
Si seul le hasard joue :

- On obtiendrait un tel écart dans 5.26% des cas

$\Rightarrow$ Pas surprenant !



Quand peut-on conclure qu'un résultat est surprenant ? Réponse «standard» = $p < 0.05$!



Cas $\mathcal{N}(0, 1)$

Specify Parameters:

Mean 0

SD 1

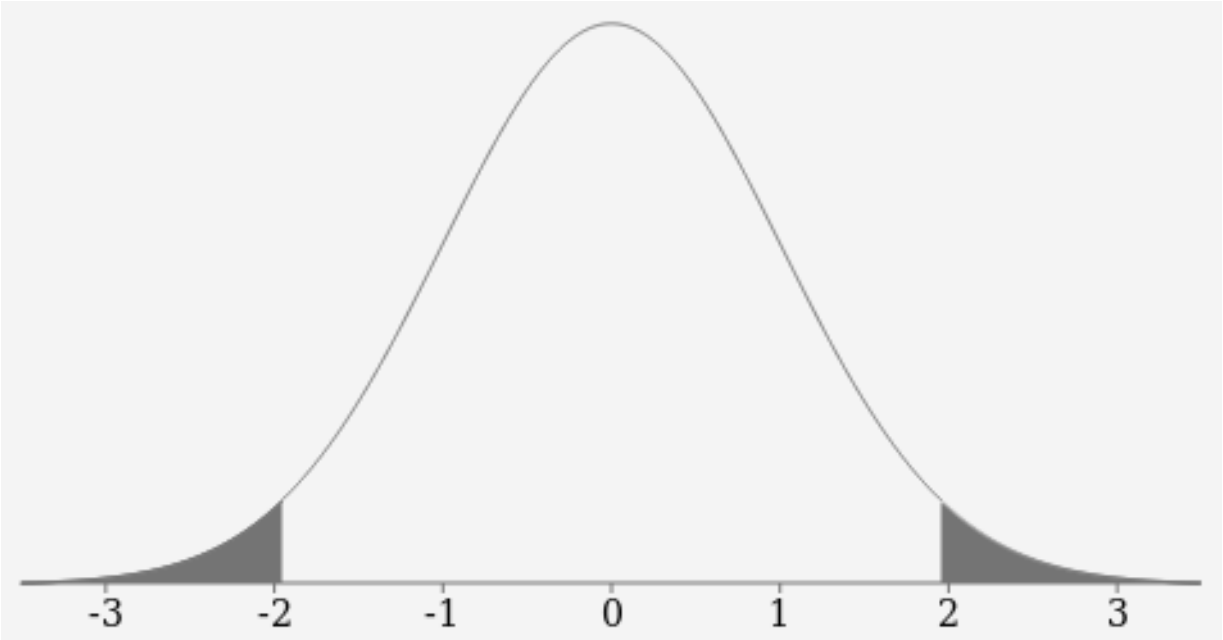
Outside -1.96 and 1.96

Results:

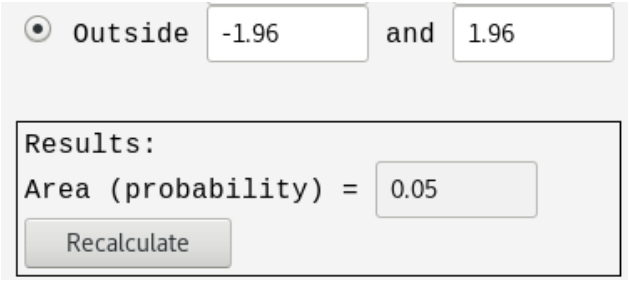
Area (probability) = 0.05

Recalculate

Quand peut-on conclure qu'un résultat est surprenant ? Réponse «standard» = $p < 0.05$!



Cas $\mathcal{N}(0, 1)$



Pour $X \sim \mathcal{N}(\mu, \sigma^2)$: Si seul le hasard est en jeu, alors $X \in [\mu - 1.96\sigma, \mu + 1.96\sigma]$ 19 fois sur 20

L’expression «19 fois sur 20», ça vous dit quelque chose ?

L'expression «19 fois sur 20», ça vous dit quelque chose ?

Résultat d'un sondage présenté sur le site Web de La Presse

Publié le 24 mai 2019 à 06h26 | Mis à jour à 06h26

Ontario : Doug Ford et son parti en chute libre

Les intentions de vote du Parti progressiste-conservateur de l'Ontario dégringolent et le taux d'insatisfaction envers le premier ministre Doug Ford n'a jamais été aussi élevé selon un sondage Recherche Mainstreet réalisé mardi et mercredi derniers.

[. . .]

Le sondage Mainstreet a été réalisé auprès de 996 personnes en Ontario. Sa marge d'erreur est de plus ou moins 3,1 %, **19 fois sur 20**.

Pourquoi utilise-t-on $p < 0.05$?

Suggestion de R.A. Fisher (1890–1962)

- Une suggestion… qui est devenue une sorte de convention — de «dogme !» — dans plusieurs domaines :
  - Sciences bio-médicales
  - Psychologie
  - Sciences sociales
  - Sondages

Pourquoi utilise-t-on $p < 0.05$?

Suggestion de R.A. Fisher (1890–1962)

- Une suggestion… qui est devenue une sorte de convention — de «dogme !» — dans plusieurs domaines :
  - Sciences bio-médicales
  - Psychologie
  - Sciences sociales
  - Sondages

«*Statistical errors*», R. Nuzzo, Nature, 2014

The irony is that when UK statistician Ronald Fisher introduced the P-value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the old-fashioned sense : **worthy of a second look**.

Des domaines utilisent des valeurs inférieures à 0.05 !

Physique des particules

High-energy physics requires even lower p-values to announce evidence or discoveries. The threshold for "evidence of a particle," corresponds to p=0.003, and the standard for "discovery" is p=0.0000003.

Et que se passe-t-il si, dans nos résultats de l'hiver 2019 pour INF3456, on change une (1) donnée ?

Après avoir fait l'analyse ci-haut, on décide de réviser la correction et une note (la plus faible) est augmentée «un peu» :

$$33.9 \rightarrow 35.9$$

Et que se passe-t-il si, dans nos résultats de l'hiver 2019 pour INF3456, on change une (1) donnée ?

Après avoir fait l'analyse ci-haut, on décide de réviser la correction et une note (la plus faible) est augmentée «un peu» :

$33.9 \rightarrow 35.9$

$\Rightarrow$ moyenne échantillon : $73.22 \rightarrow 73.32$

$\Rightarrow$ 1.9948 écart-types (de 69.78)

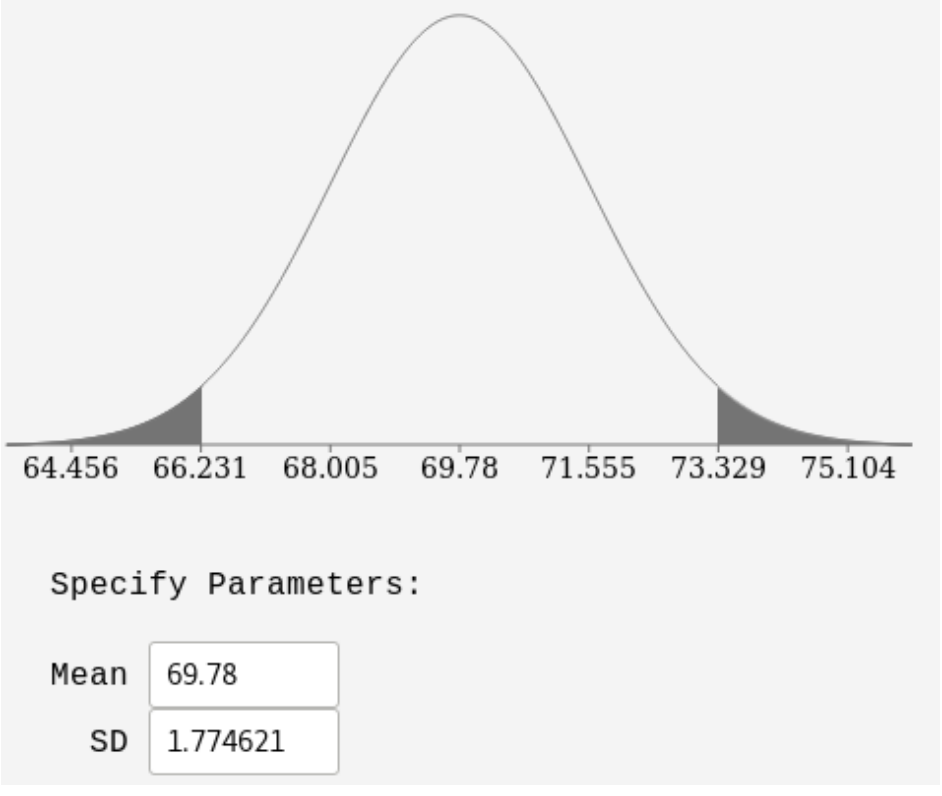
Et que se passe-t-il si, dans nos résultats de l'hiver 2019 pour INF3456, on change une (1) donnée ?

Après avoir fait l'analyse ci-haut, on décide de réviser la correction et une note (la plus faible) est augmentée «un peu» :

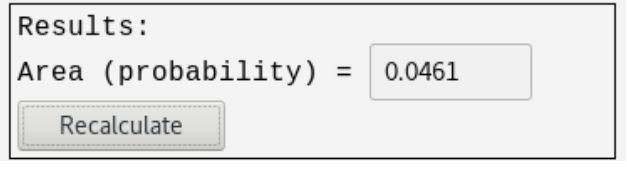
$33.9 \rightarrow 35.9$

$\Rightarrow$ moyenne échantillon : $73.22 \rightarrow 73.32$

$\Rightarrow$ 1.9948 écart-types (de 69.78)



On a alors $p < 0.05$ et on peut dire que notre résultat est «**statistiquement significatif**»



Aperçu

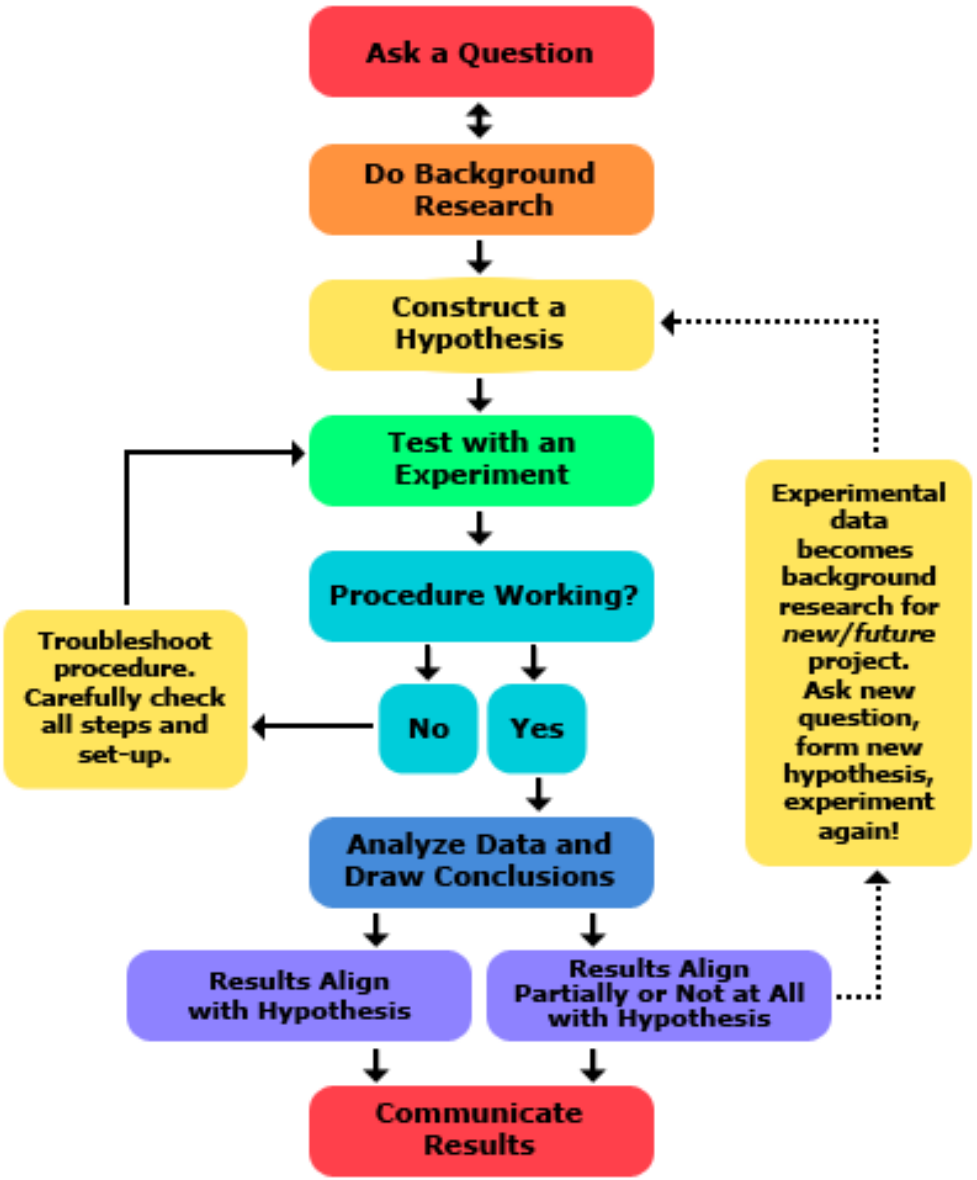
La crise n'est pas nécessairement dûe à des «fraudes»

Outright fraud is almost certainly just a small part of that problem, but high-profile examples have exposed a greyer area of bad or lazy scientific practice that many had preferred to brush under the carpet.

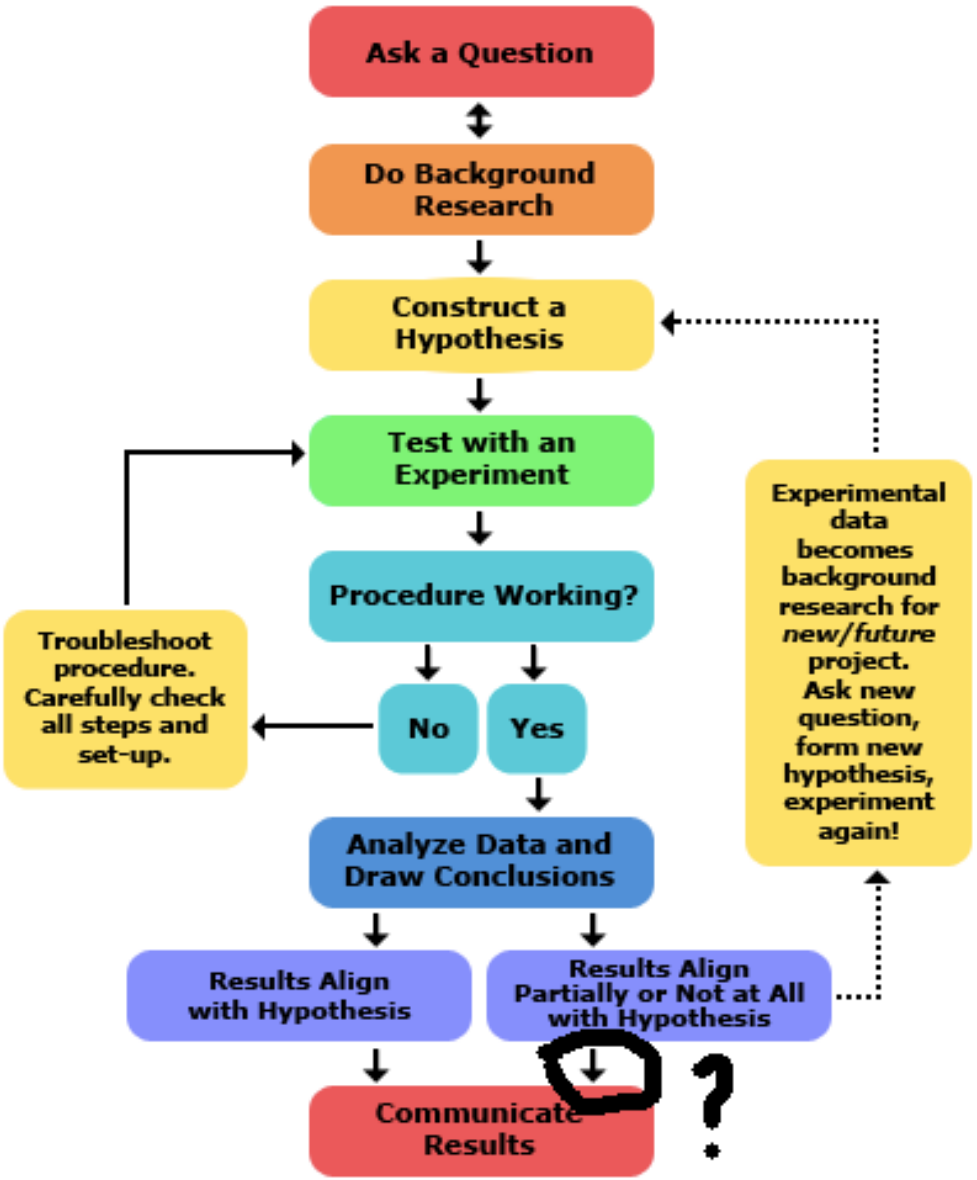
«False positives : fraud and misconduct are threatening scientific research», A. Jha, The Guardian, 2012

5.1 Valorisation des résultats «positifs» et de la «nouveauté»

Peut-on publier n'importe quels résultats ?

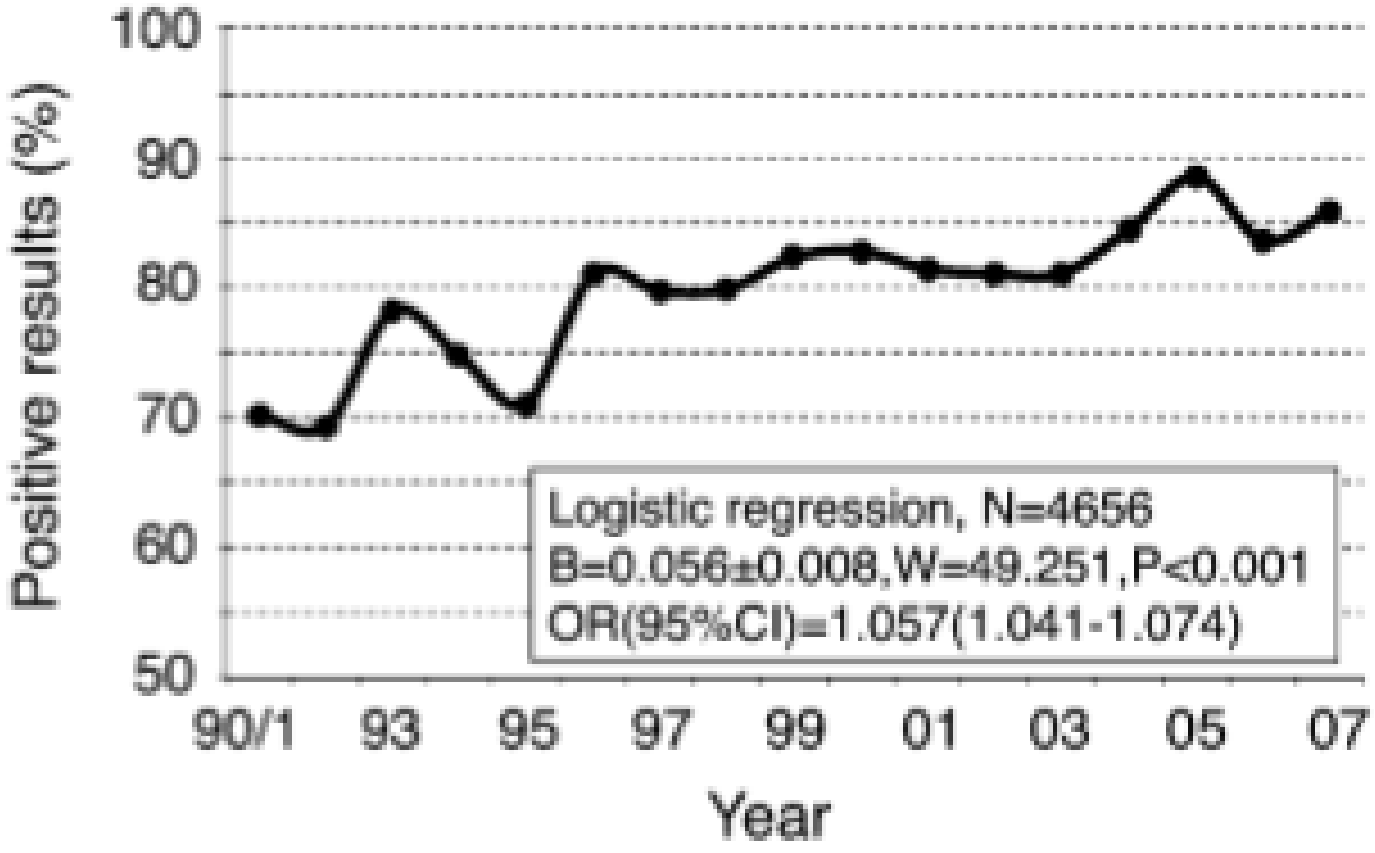


Peut-on publier n'importe quels résultats ?



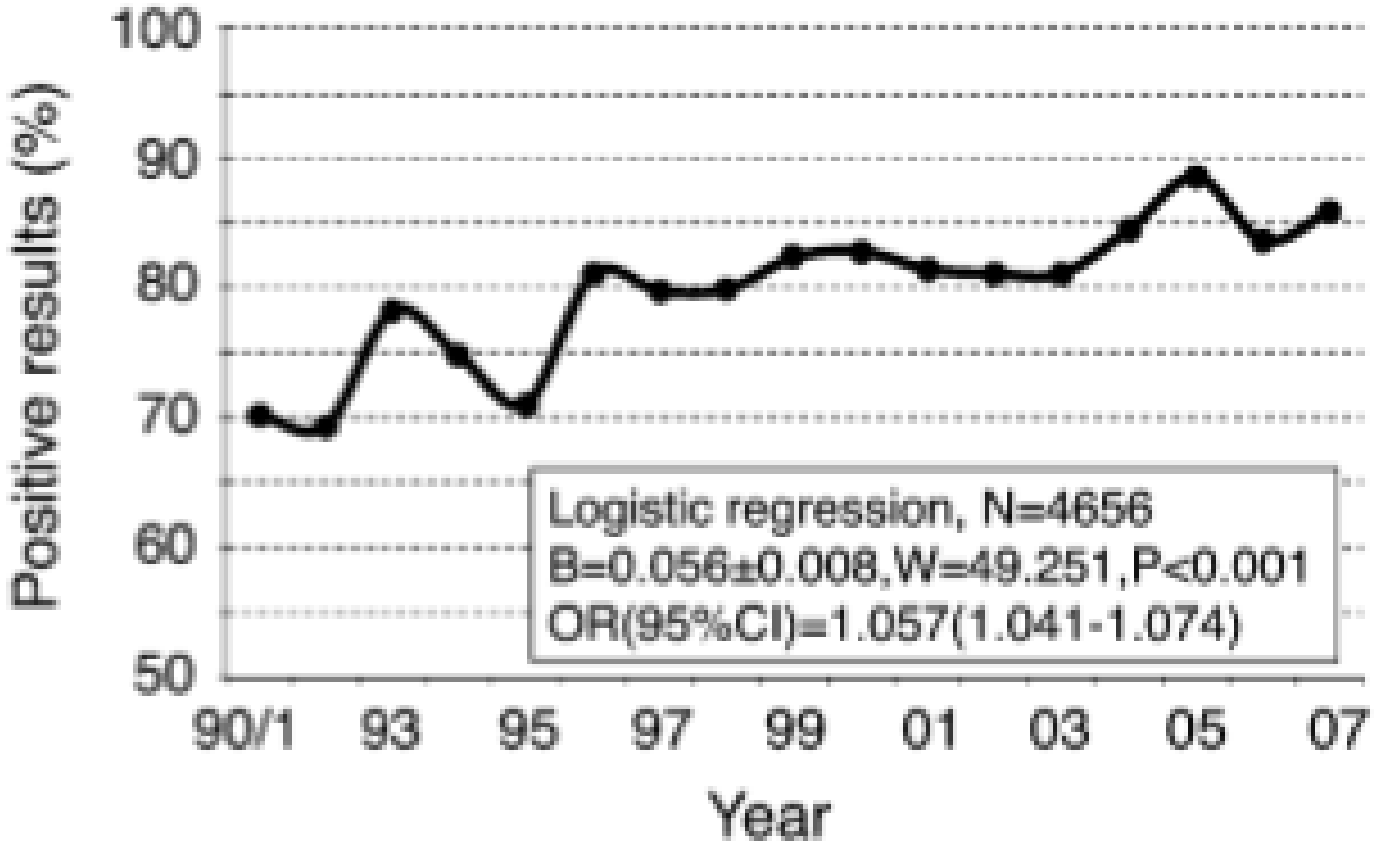
Évolution du pourcentage d'articles publiés présentant des résultats positifs

- Fanelli (2010) : 2000 articles dans divers domaines (bio, psycho, physique, chimie, etc.) — *space science* : 70%, . . . , psycho : 91%.



Évolution du pourcentage d'articles publiés présentant des résultats positifs

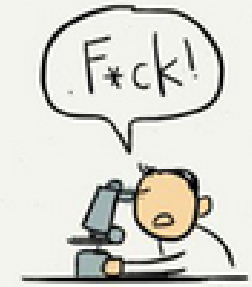
- Fanelli (2010) : 2000 articles dans divers domaines (bio, psycho, physique, chimie, etc.) — *space science* : 70%, . . . , psycho : 91%.
- Autre étude : biologie moléculaire et médecine clinique : 100%

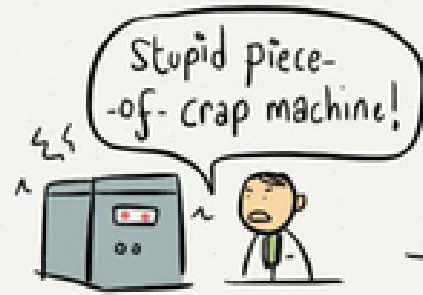


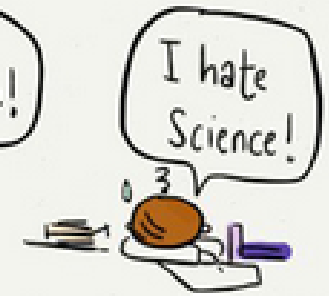
Some people think scientists exclaim

When doing experiments.

But they're way more likely to say...







twistedoodles.com

Les articles scientifiques racontent une histoire, pas la réalité

Pour le béotien qui l'aborde, la littérature scientifique étonne en effet par son étonnante efficacité. Exceptionnels sont les articles qui décrivent un échec, une fausse piste, une impasse. Tout se passe comme si les chercheurs n'avaient toujours que de bonnes idées. Supposés interroger la nature, leurs expériences ont presque toujours le bon goût de confirmer l'hypothèse qui avait conduit à leur élaboration.

«Malscience — De la fraude dans les labos»,
N. Chevassus-au-Louis (2016)

«Le côté sombre de la science», S. Larivée, Revue de psychoéducation, 2017

Tableau 1. Revues qui publient uniquement des résultats négatifs

Nom de la revue	Depuis	Statut actuel
The All Results Journal: Biol	2010	Actif
The All Results Journal: Chem	2010	Actif
The All Results Journal: Phys	2011	Actif
The All Results Journal: Nano	2015	Actif
Cortex	2013	Actif
Journal of Pharmaceutical Negative Results	2010	Actif
Journal of Negative Results – Ecology et Evolutionary Biology	2004	Interrompu
Journal of Negative Results in BioMedicine	2002	Actif
Journal of Negative Results in Speech and Audio Sciences	2004	?
New Negatives in Plant Science	2014	Actif
Plos One	2014	?
Journal of Negative Observation in Genetic Oncology	1997	Interrompu
Negat	?	?
Negations	?	Actif
Negative Capability	?	Interrompu
Contingent Negative Variation	?	?
Yixue Zhengming	?	Actif
Negative Pessure Wound Therapy	?	Actif
Journal of Negative and No Positive Results	?	Actif
Making Digital Negatives With an Ink-Jet Printer	?	Actif
Journal of Articles in Support of the Null Hypothesis	2002	Actif
Journal of Errology	?	Interrompu
Journal of Interesting Negative Results	2008	Interrompu
Nature Negative Results section	2010	Actif
The Journal of Spurious Correlations	2005	Interrompu
The Null Journal	2009	Interrompu
University of Colorado Database of Negative Results	2011	Interrompu
The International Journal of Negative & Null Results	?	Interrompu
Negative Results	2016	Actif

Il est très difficile de publier des résultats négatifs : Un exemple «intéressant»

Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect.

By Bem, Daryl J.

Journal of Personality and Social Psychology, Vol 100(3), Mar 2011, 407-425

Abstract

The term psi denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and premonition (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and *premonition* are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by “time-reversing” well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed. (PsycINFO Database Record (c) 2016 APA, all rights reserved)

Il est très difficile de publier des résultats négatifs : Un exemple «intéressant»

Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect.

By Bem, Daryl J.

Journal of Personality and Social Psychology, Vol 100(3), Mar 2011, 407-425

Abstract

The term psi denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and premonition (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and *premonition* are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed. (PsycINFO Database Record (c) 2016 APA, all rights reserved)

- Une équipe a tenté (3 fois !) de reproduire son expérience. . .

mais sans succès ☹

Il est très difficile de publier des résultats négatifs : Un exemple «intéressant»

Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect.

By Bem, Daryl J.
Journal of Personality and Social Psychology, Vol 100(3), Mar 2011, 407-425

Abstract

The term psi denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and premonition (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and *premonition* are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed. (PsycINFO Database Record (c) 2016 APA, all rights reserved)

- Une équipe a tenté (3 fois !) de reproduire son expérience. . .

mais sans succès ☹

- Dixit le *Journal of Pers. and Soc. Psy.* : «*[we do] not publish replication studies, whether successful or unsuccessful*» !

La réplication d'expériences est cruciale pour «confirmer» qu'un résultat est significatif

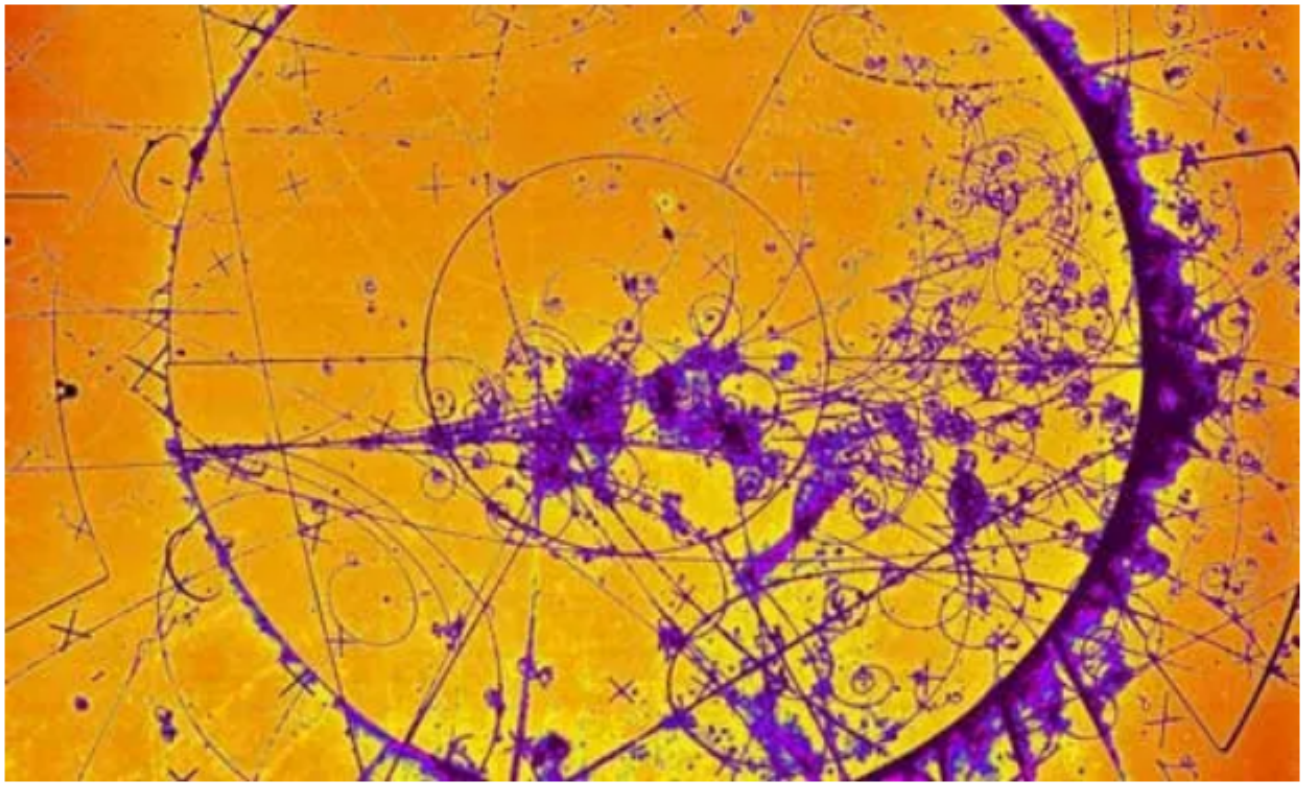
Et la réplication d'expériences est aussi cruciale pour «**infirmer**» un résultat

Détection de neutrinos plus rapides que la lumière ?

2011

Neutrinos still faster than light in latest version of experiment

Finding that contradicts Einstein's theory of special relativity is repeated with fine-tuned procedures and equipment



▲ Scientists from Cern have repeated their finding of neutrinos travelling faster than the speed of light. Photograph: Cern/Science Photo Library

Détection de neutrinos plus rapides que la lumière ? Non !

2012 : Erreur causée par a loose fiber-optic cable !

Flaws found in faster-than-light neutrino measurement

Two possible sources of error uncovered.

Eugenie Samuel Reich

22 February 2012

The OPERA collaboration, which made headlines in September with the revolutionary claim that it had clocked neutrinos travelling faster than the speed of light, has identified two possible sources of error in its experiment. If true, its initial result would have violated Einstein's special theory of relativity, a cornerstone of modern physics.

OPERA had collected data suggesting that neutrinos generated at CERN near Geneva in Switzerland and sent 730 kilometres to its detector

La valorisation des résultats positifs peut conduire à des pratiques «douteuses»

Pers Soc Psychol Rev. 1998;2(3):196-217.

HARKing: hypothesizing after the results are known.

Kerr NL[1].

⊕ **Author information**

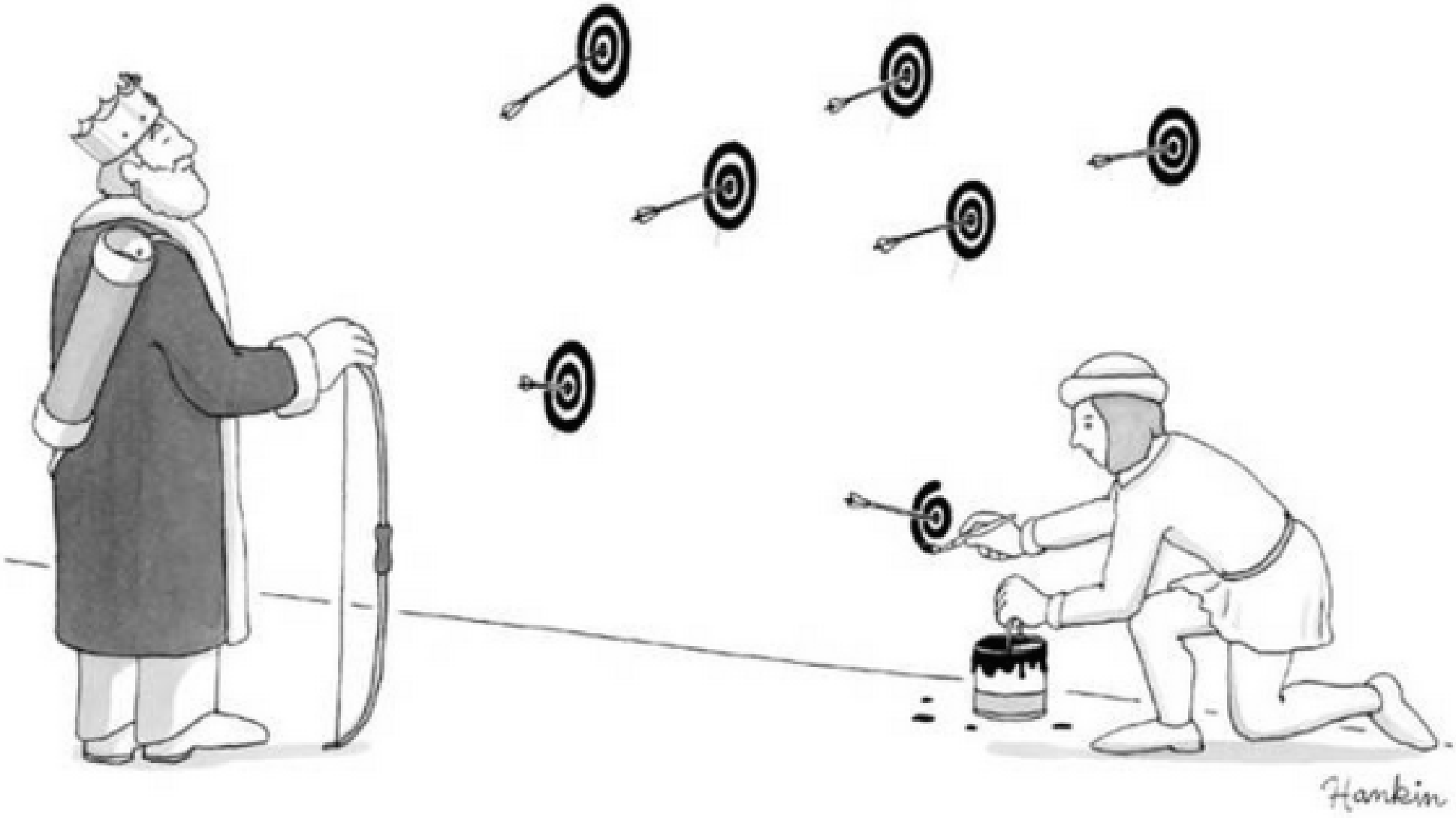
Abstract

This article considers a practice in scientific communication termed HARKing (Hypothesizing After the Results are Known). HARKin defined as presenting a post hoc hypothesis (i.e., one based on or informed by one's results) in one's research report as i f it were, a priori hypotheses. Several forms of HARKing are identified and survey data are presented that suggests that at least some forms HARKing are widely practiced and widely seen as inappropriate. I identify several reasons why scientists might HARK. Then I discu reasons why scientists ought not to HARK. It is conceded that the question of whether HARKing ' s costs exceed its benefits is a co that ought to be addressed through research, open discussion, and debate. To help stimulate such discussion (and for those such a who suspect that HARKing's costs do exceed its benefits), I conclude the article with some suggestions for deterring HARKing.

HARKing

«[P]resenting a post hoc hypothesis in the introduction of a research report as if it were an a priori hypothesis.»

Note : *Hark ! = Listen ! (Oxford Dictionary)*


Hankin

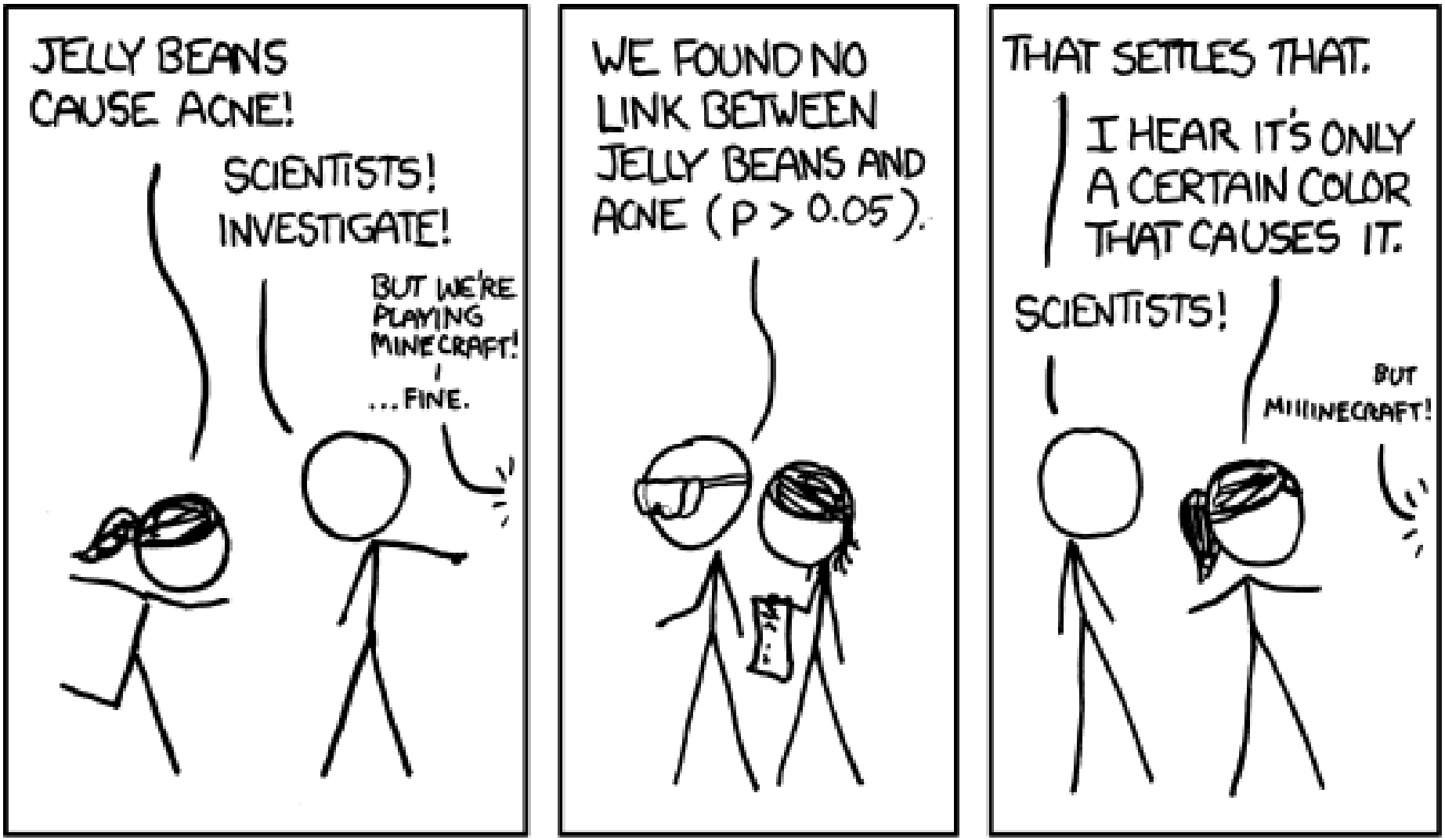
Self-Admission Rates of HARKing in Self-Report Surveys

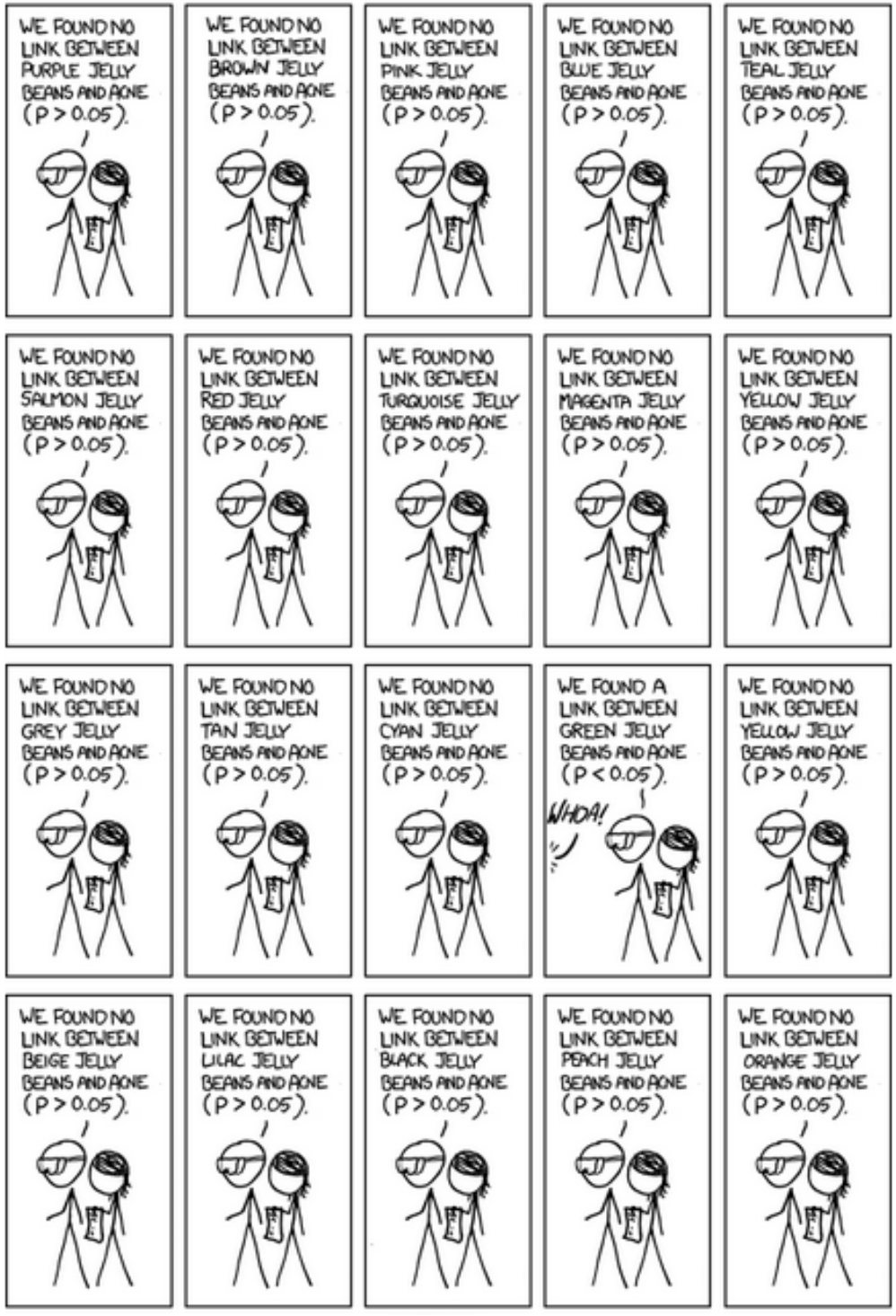
Survey	Population	Survey Item	*N*	Self-Admission Rate
John, Loewenstein, and Prelec (2012)	USA psychologists	"In a paper, reporting an unexpected finding as having been predicted from the start."	2,155	27.0%
Agnoli, Wicherts, Veldkamp, Albiero, and Cubelli (2017)	Italian psychologists	"In a paper, reporting an unexpected finding as having been predicted from the start."	277	37.4%
Bosco, Aguinis, Field, Pierce, and Dalton (2016, Study 1)	Researchers who published in *Personnel Psychology* and the *Journal of Applied Psychology* during 2005 to 2010	"whether any changes in hypotheses had occurred between the completion of data collection and subsequent publication."	53	38%
Fiedler and Schwarz (2016)	German psychologists	"Reporting an unexpected finding as having been predicted from the start."	1,138	47%
Banks et al. (2016, Studies 1 & 2)	Management researchers	"selectively reported hypotheses on the basis of statistical significance…and presented a post hoc hypothesis as if it were developed a priori."	749	50%
Motyl et al. (2017, Study 1)	Personality and social psychologists from Australian, European, and the USA	"Report that unexpected findings were expected."	1,166	58%
			Mean	43%

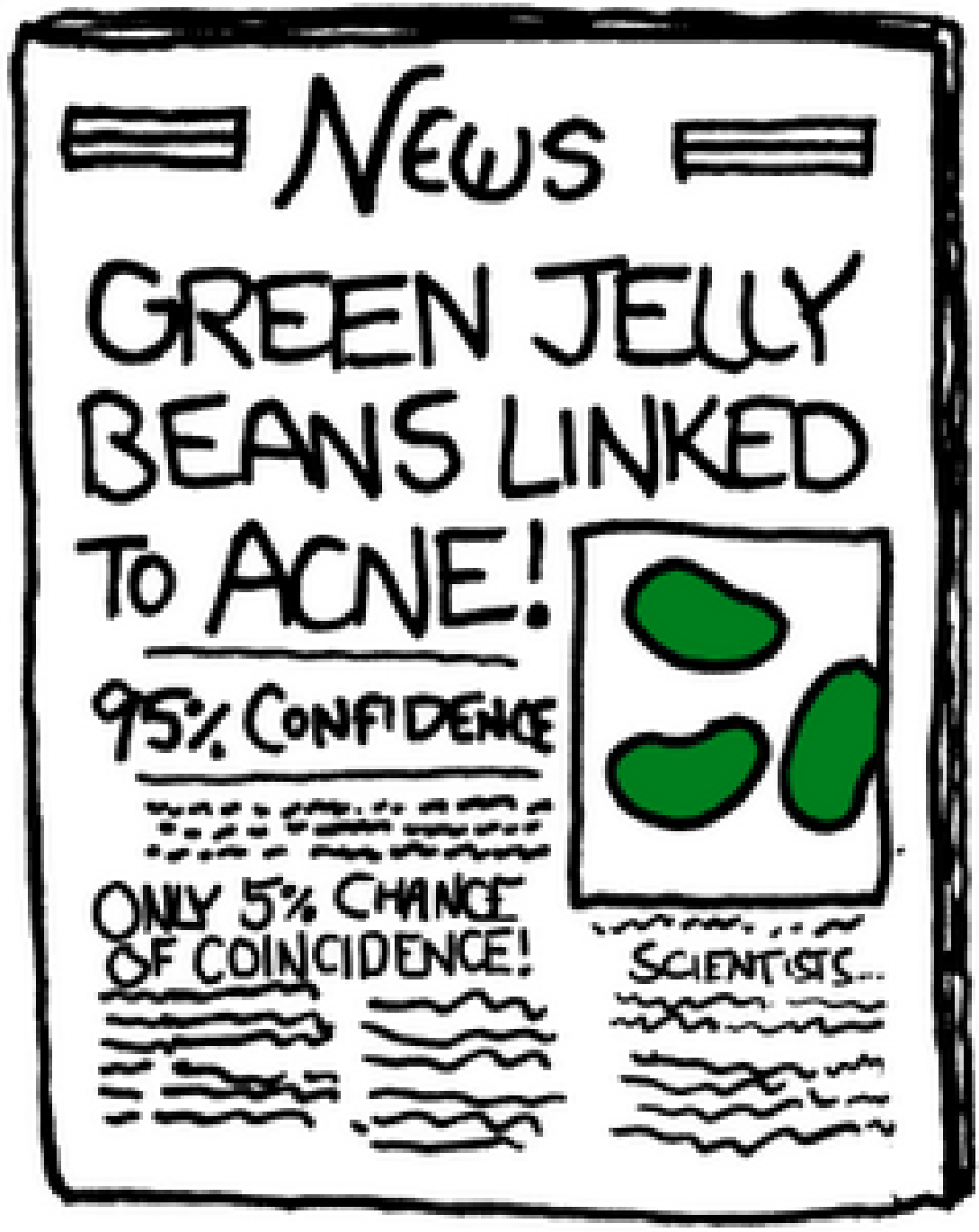
Note. Self-admission rates are for undertaking the stated behavior "at least once." Self-admission rates are likely to be underestimates because researchers tend to underreport practices that they perceive to be undesirable (Agnoli et al., 2017).

«Il est probable que des choses improbables se produiront.»

Aristote


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

5.2 Flexibilité des protocoles et des analyses

Les chercheurs, dans leurs expériences et analyses, jouissent d'une certaine (grande ?) «latitude»

- Exclusion de certaines valeurs/participants (*outliers*) . . .
  ou pas ?
- Fin de la collecte des données. . .
  ou poursuite ?
- Utilisation d'une analyse statistique. . .
  ou une autre ?

IF YOU TORTURE THE DATA LONG ENOUGH, IT *WILL* CONFESS.

Une technique de torture = *p-hacking*

Une technique de torture = *p-hacking*

P-hacking

[p-hacking] occurs when researchers collect or select data or statistical analyses until **nonsignificant** *results become* **significant**.

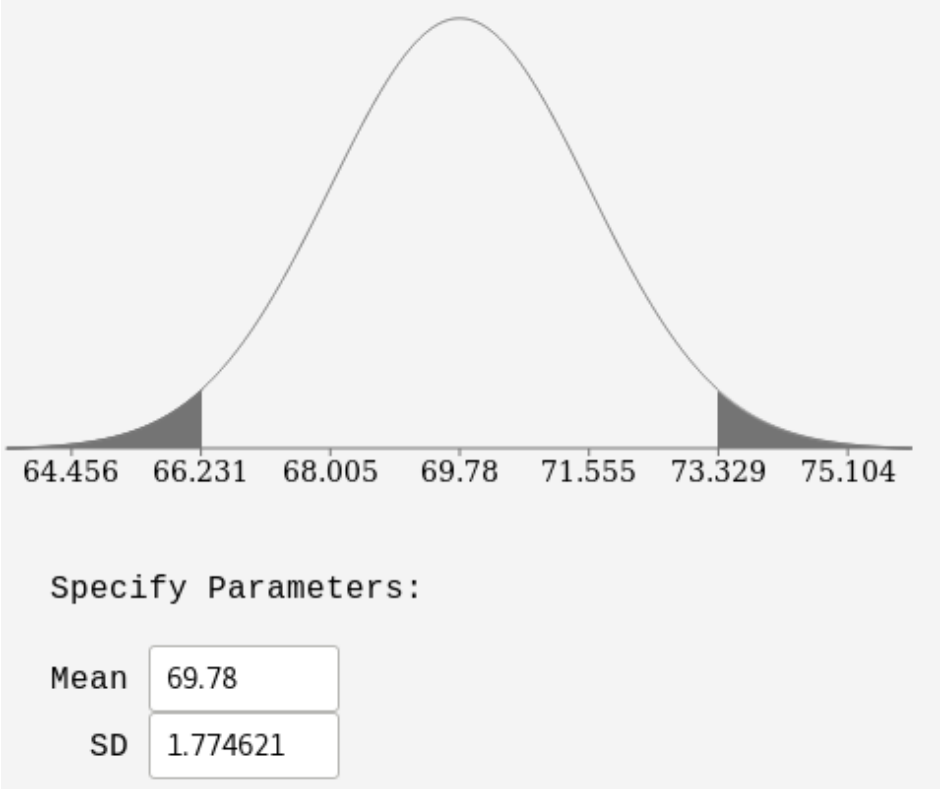
«The Extent and Consequences of P-Hacking in Science», *Head et al. (2015)*

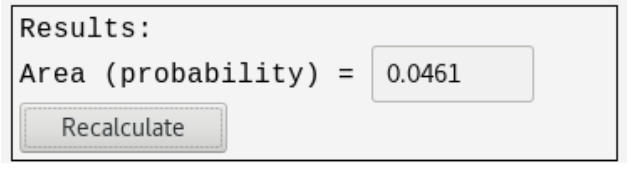
Rappelez-vous l'expérience sur l'IDE pour L

Correction révisée + une (1) note modifiée :

$33.9 \to 35.9 \Rightarrow$ Moyenne : $73.22 \to 73.32$

- Avant : $p = 0.0526 > 0.05$ ☹
- **Après :** $p = 0.0461 < 0.05$ ☺

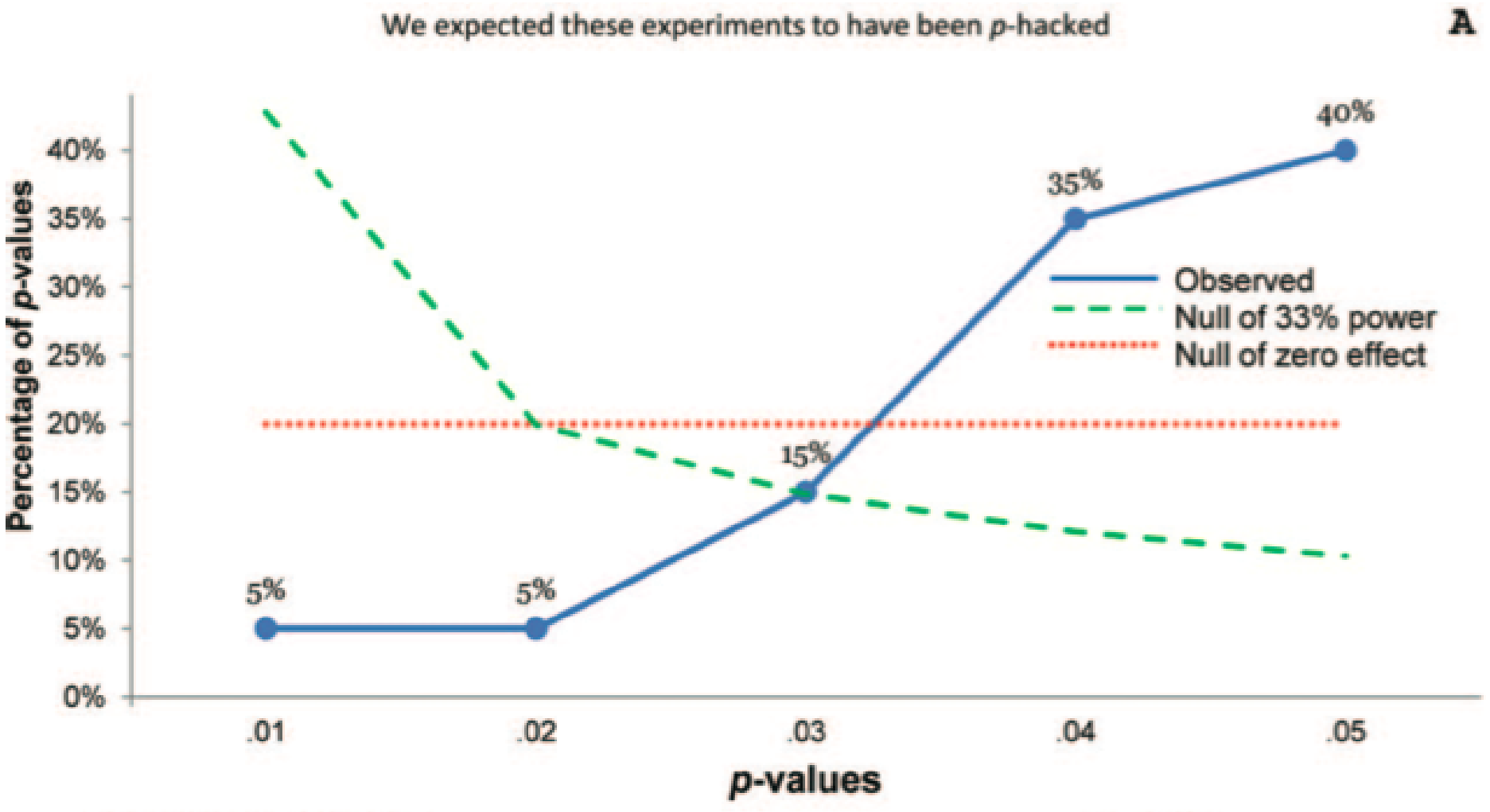




Est-ce que ce genre de bidouillage est fréquent ?

Est-ce que ce genre de bidouillage est fréquent ?

Oui !



Des analyses différentes sur les mêmes données, peuvent conduire à des conclusions différentes !

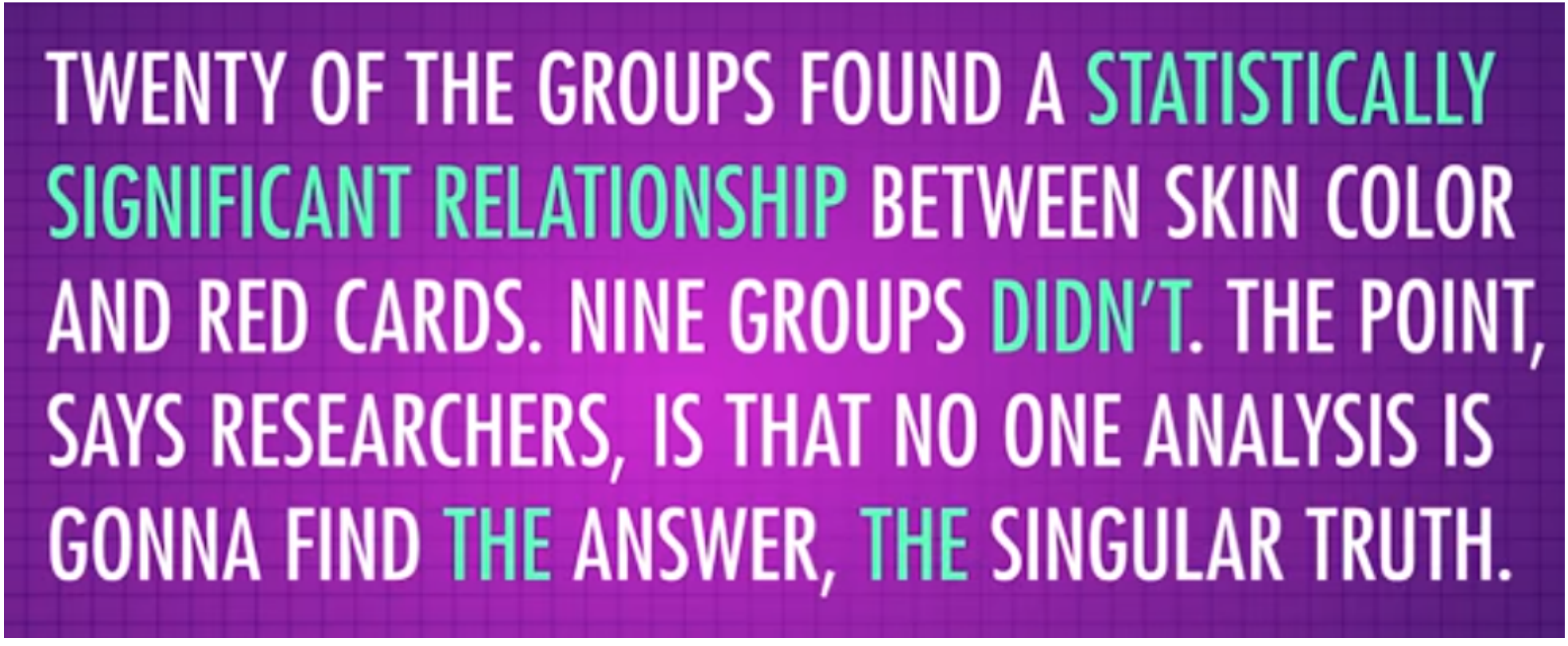
https://www.youtube.com/watch?v=vBzEGSm23y8

Question : Les arbitres donnent-ils plus souvent des pénalités aux joueurs à peau foncée qu'à ceux à peau claire ?

Des analyses différentes sur les mêmes données, peuvent conduire à des conclusions différentes !

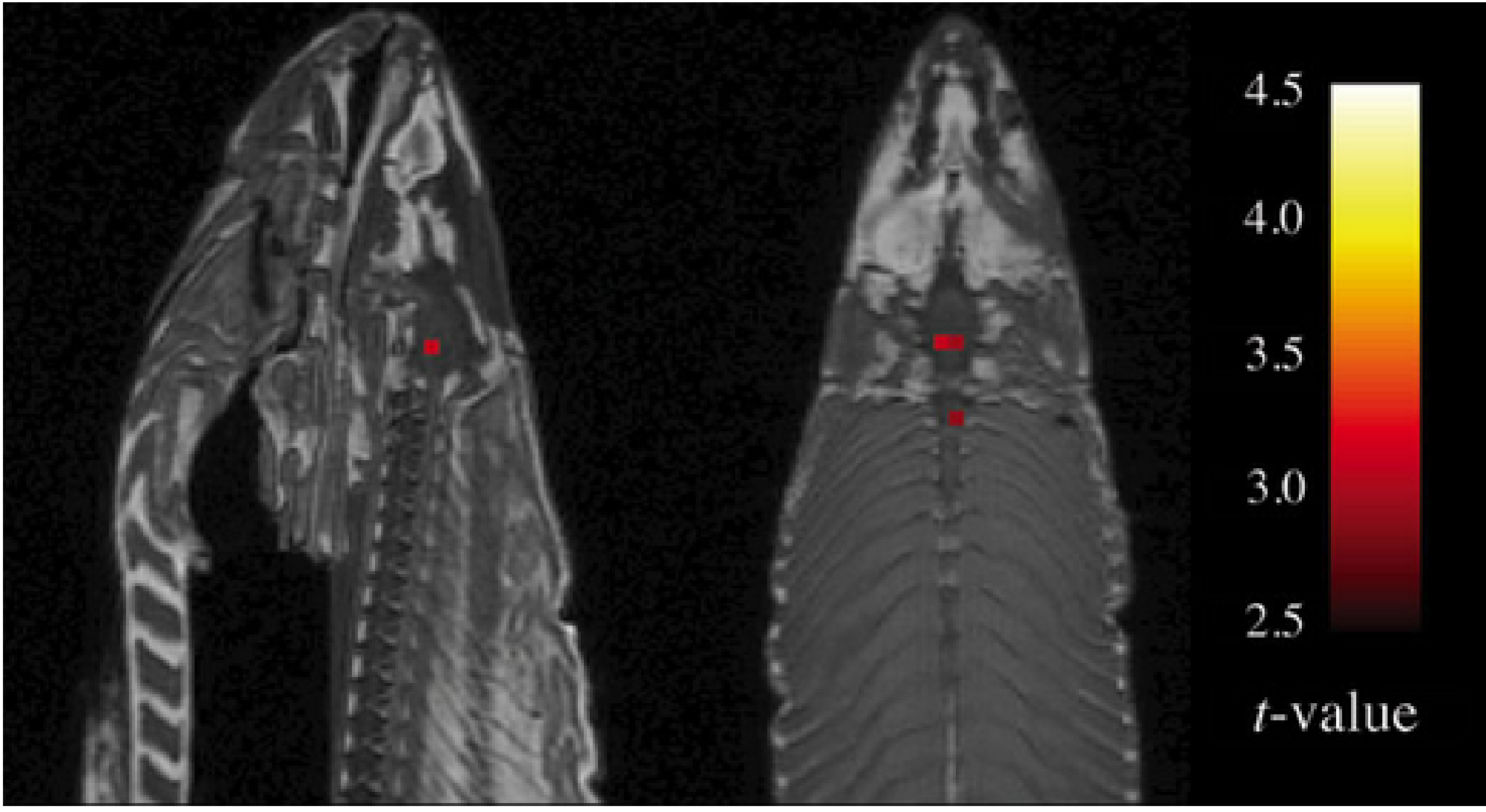
https://www.youtube.com/watch?v=vBzEGSm23y8

Question : Les arbitres donnent-ils plus souvent des pénalités aux joueurs à peau foncée qu'à ceux à peau claire ?



Autre exemple de *result fishing* : Un saumon qui réagit à des photos d'humains exprimant diverses émotions

Expériences utilisant le *Functional Magnetic Resonance Imaging* (fMRI)



Autre exemple de *result fishing* : Un saumon qui réagit à des photos d'humains exprimant diverses émotions

Expériences utilisant le *Functional Magnetic Resonance Imaging* (fMRI)

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Dangers du *data mining* ?

Data mining explicitly capitalizes on one of the key principles of both cherry-picking and question trolling—i.e., that if a researcher looks at enough sample results, he or she is bound to eventually find something that looks interesting. [. . .]

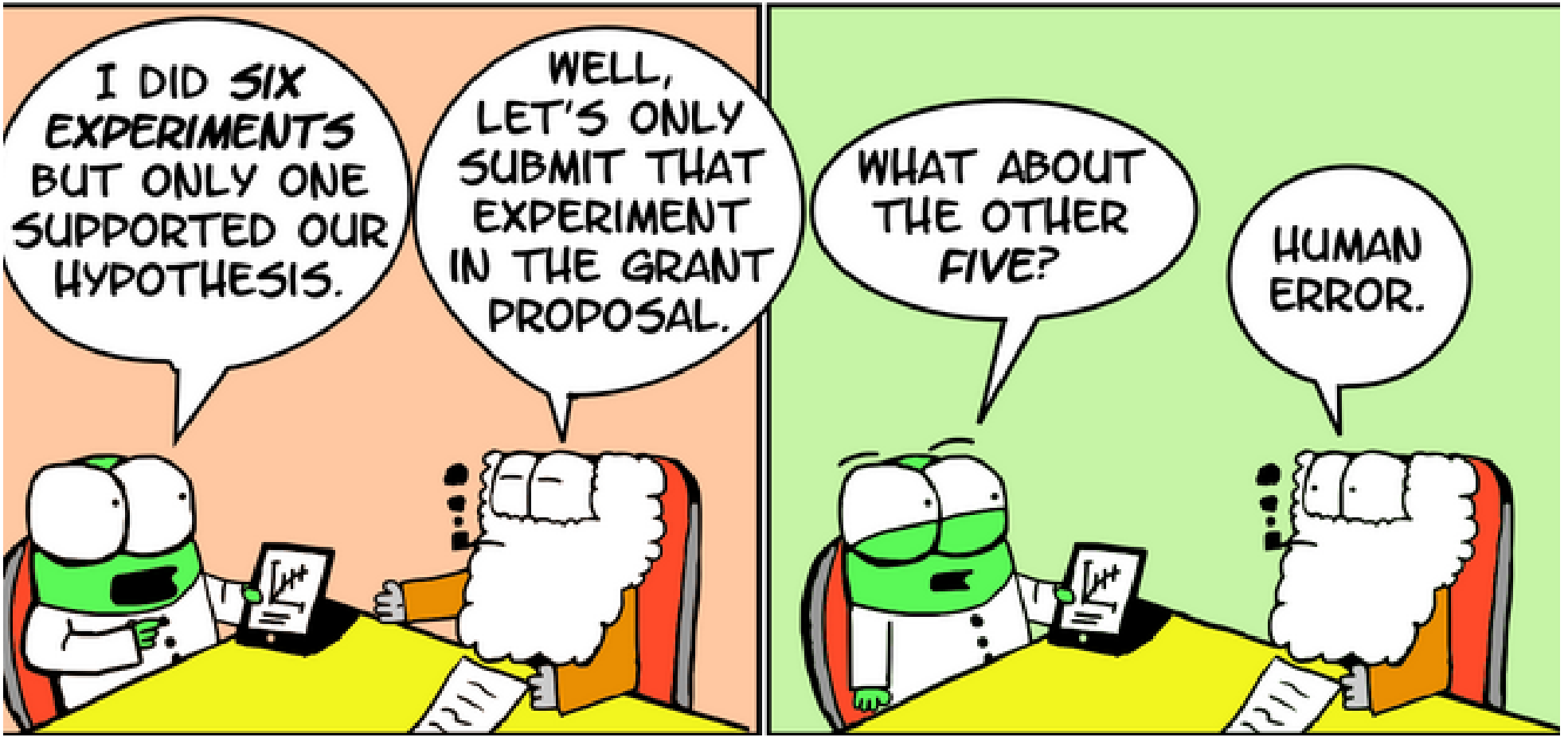
«HARKing : How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results ?», *K.R. Murphy & H. Aguinis, Journal of Business and Psychology, 2017.*

5.3 Encore d’autres facteurs

Biais de confirmation

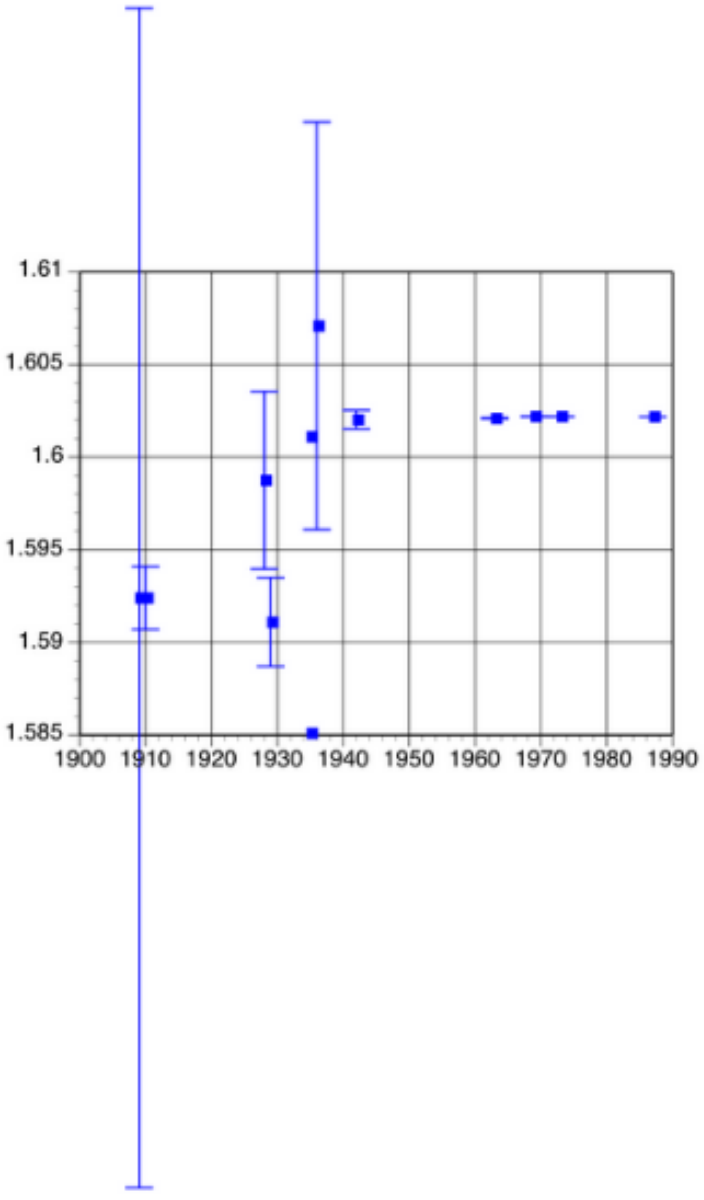
CONFIRMATION BIAS

FAVOURING EVIDENCE THAT SUPPORTS YOUR PRE-EXISTING BELIEFS WHILE IGNORING EVIDENCE THAT DOESN'T.



Détermination de la charge élémentaire de l'électron et rôle des résultats «négatifs» (non-réplication)

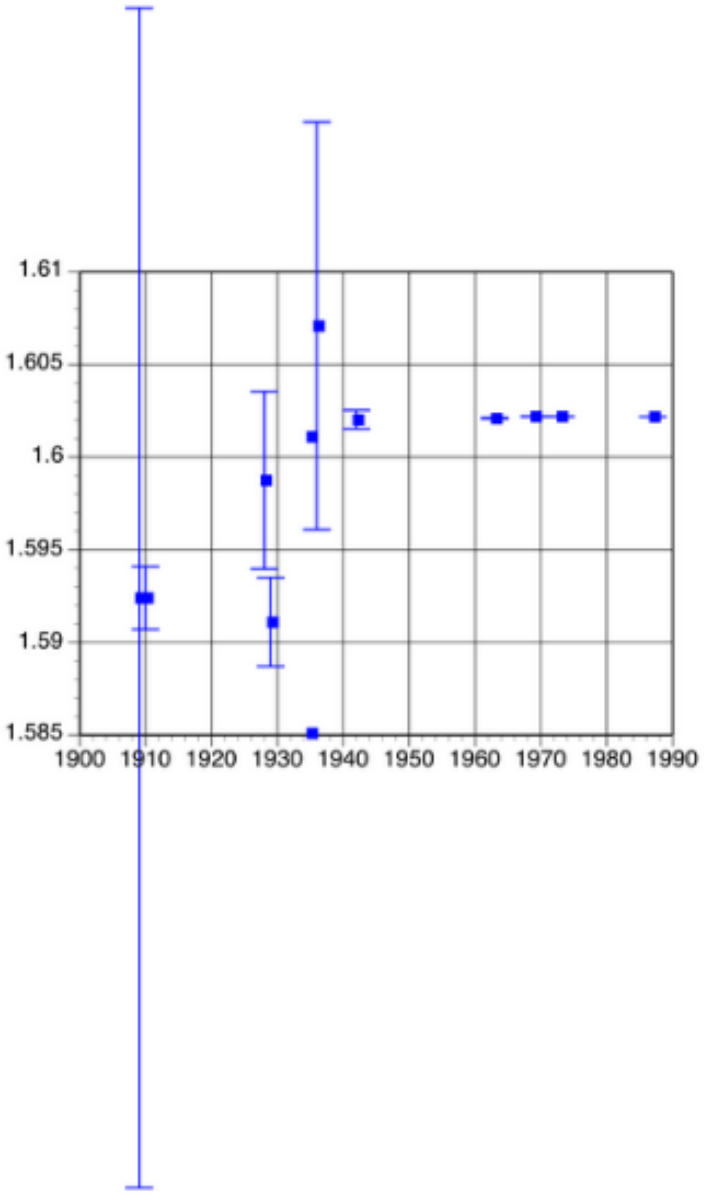
Travaux initiaux faits par R.A. Milikan ⇒ Prix Nobel de physique (1923)



Sauf que. . .

Détermination de la charge élémentaire de l'électron et rôle des résultats «négatifs» (non-réplication)

Travaux initiaux faits par R.A. Milikan ⇒ Prix Nobel de physique (1923)



Millikan (notebooks)
Millikan (published)
Erik Backlin, Nature 1929
[Birge, 1929]
Backlin and Flemberg, Nature 1936
Backlin and Flemberg, cited in HR Robinson RPP 1937
Gunnar Kellström PR 1936
[Birge, 1942]
[Dummond and Cohen, 1963]
[Taylor et al, 1969]
[Cohen and Taylor, 1973]
[Cohen and Taylor, 1987]

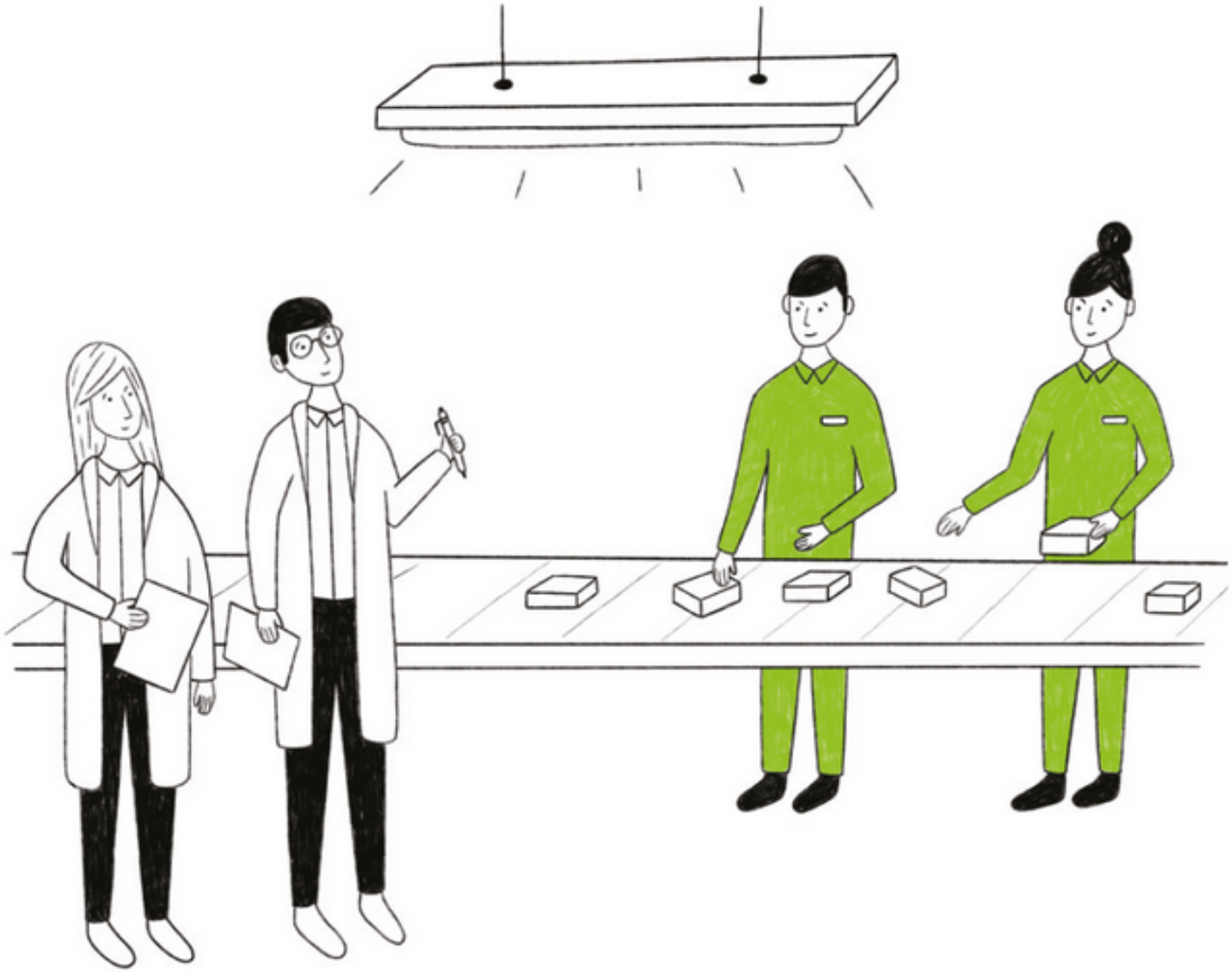
Sauf que. . .

«*Finding out that something does not work isn't going to win you a Nobel prize*»

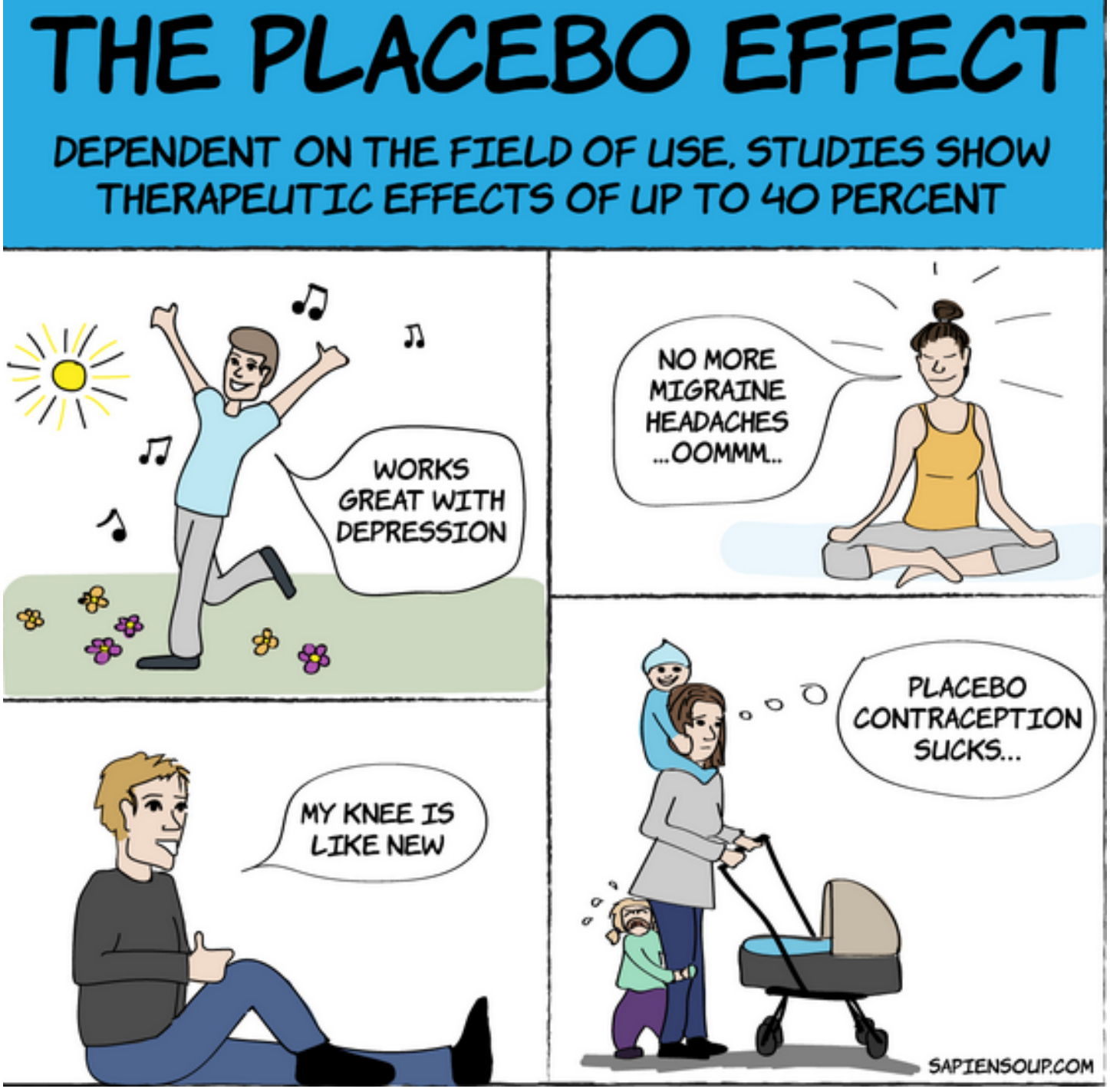
Expérimentation avec des humains et effet Hawthorne

Effet Hawthorn = Effet de l'observateur

https://www.geckoboard.com/learn/data-literacy/statistical-fallacies/hawthorne-effect/



Expérimentation avec des humains et effet placebo



https://sapiensoup.com/placebo-homeopathy


DRUGS
PLACEBOS
EX-STRENGTH PLACEBOS
DAVE CARPENTER...
CARTOONSTOCK .com
Search ID: dcr0323

Aperçu

Conclusion : Quelques pistes de solution ?

- Encourager la réplication

Conclusion : Quelques pistes de solution ?

- Encourager la réplication
- Outils pour détecter certains résultats douteux :
  - GRIM/GRIMMER (Wansik !)
  - SPRITE

Conclusion : Quelques pistes de solution ?

- Encourager la réplication
- Outils pour détecter certains résultats douteux :
  - GRIM/GRIMMER (Wansik !)
  - SPRITE
- Données ouvertes

Conclusion : Quelques pistes de solution ?

- Encourager la réplication
- Outils pour détecter certains résultats douteux :
  - GRIM/GRIMMER (Wansik !)
  - SPRITE
- Données ouvertes
- Utiliser $p < 0.01$ ou $p < 0.005$

Conclusion : Quelques pistes de solution ?

- Encourager la réplication
- Outils pour détecter certains résultats douteux :
  - GRIM/GRIMMER (Wansik !)
  - SPRITE
- Données ouvertes
- Utiliser $p < 0.01$ ou $p < 0.005$
- Laisser tomber NHST — statistiques Bayésiennes ?

Conclusion : Quelques pistes de solution ?

- Encourager la réplication
- Outils pour détecter certains résultats douteux :
  - GRIM/GRIMMER (Wansik !)
  - SPRITE
- Données ouvertes
- Utiliser $p < 0.01$ ou $p < 0.005$
- Laisser tomber NHST — statistiques Bayésiennes ?
- *Registered reports*

Registered Reports

Peer review before results are known to align scientific values and practices

Source: *Center for Open Science* : `https://osf.io/8mpji/wiki/home/`

Pour en savoir plus. . .

N. Chevassus-au Louis.

Malscience — De la fraude dans les labos.
Éditions du Seuil, 2016.

C. Chambers.

The seven deadly sins of psychology : A manifesto for reforming the culture of scientific practice.
Princeton University Press, 2017.

R.R. Haccoun and D. Cousineau.

Statistiques—Concepts et applications (Deuxième édition revue et augmentée).
Les Presses de l'Université de Montréal, 2010.

J.P.A. Ioannidis.

Why most published research findings are false.
PLoS Medicine, 2(8) :e124, 2005.

J.P.A. Ioannidis.

What have we (not) learnt from millions of scientific paper with *p* values ?
The American Statistician, 73(S1) :20–25, 2019.

Pour en savoir plus. . .

D. Randall and C. Welser.

The irreproducibility crisis of modern science—Causes, consequences, and the road to reform.
Technical report, National Association of Scholars, 2018.

F. Shull, J. Singer, and D.I.K. Sjoberg, editors.

Guide to Advanced Empirical Software Engineering.
Springer, 2008.

R.L. Wasserstein and N.A. Lazar.

The ASA's statement on *p*-values : Context, process, and purpose.
The American Statistician, 70(2) :129–133, 2016.

A. Zeller, T. Zimmermann, and C. Bird.

Failure is a four-letter word : A parody in empirical research.
In *Proc. of the 7th Int. Conf. on Predictive Models in Software Engineering*. ACM, 2011.

Remarques ?

Questions ?

Annexe. . . en forme de devinettes !

Connaissez-vous ces scientifiques ?

Ont-ils «bonne réputation» ?

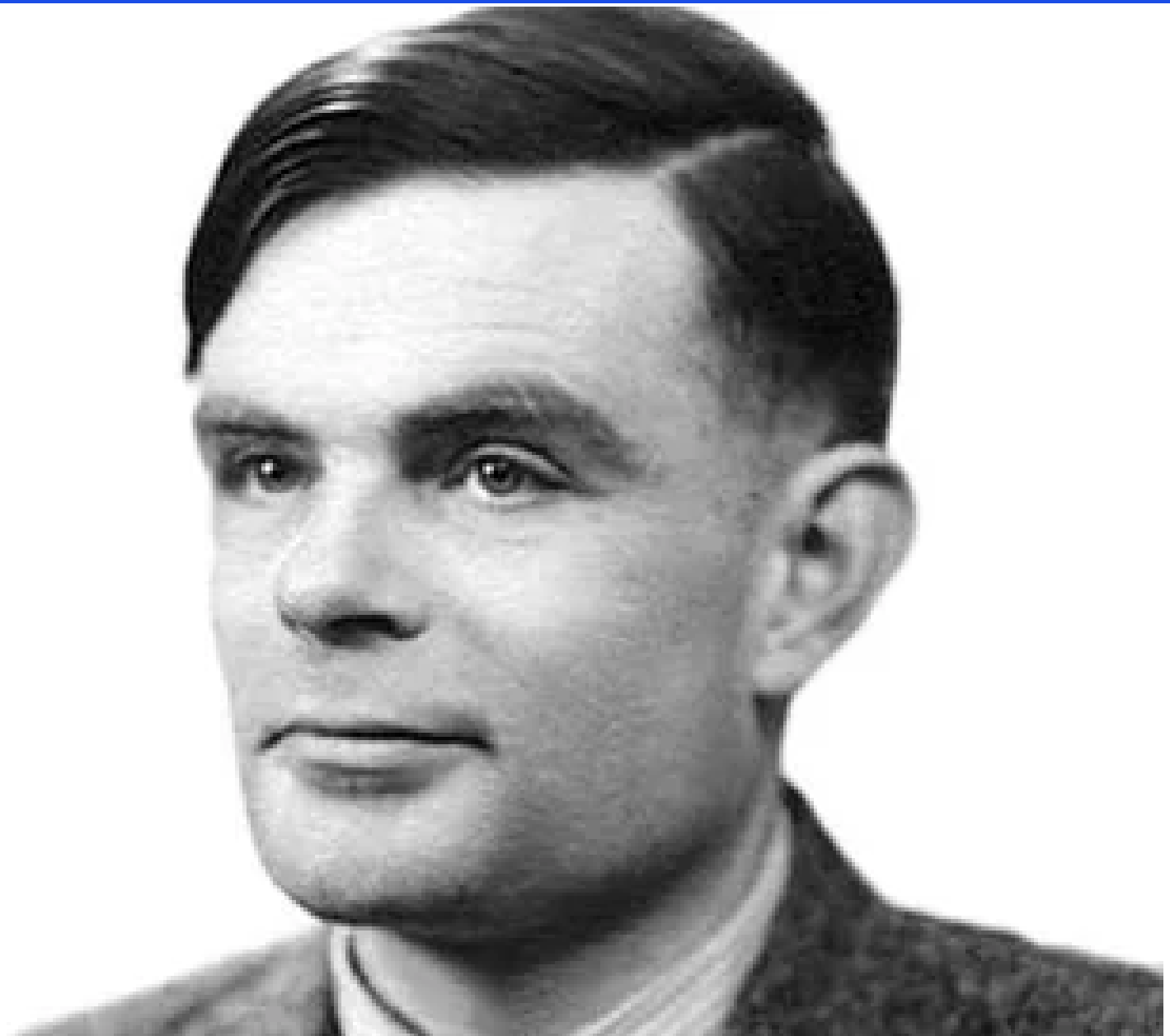
(1822–1895)

Louis Pasteur
(1822–1895)

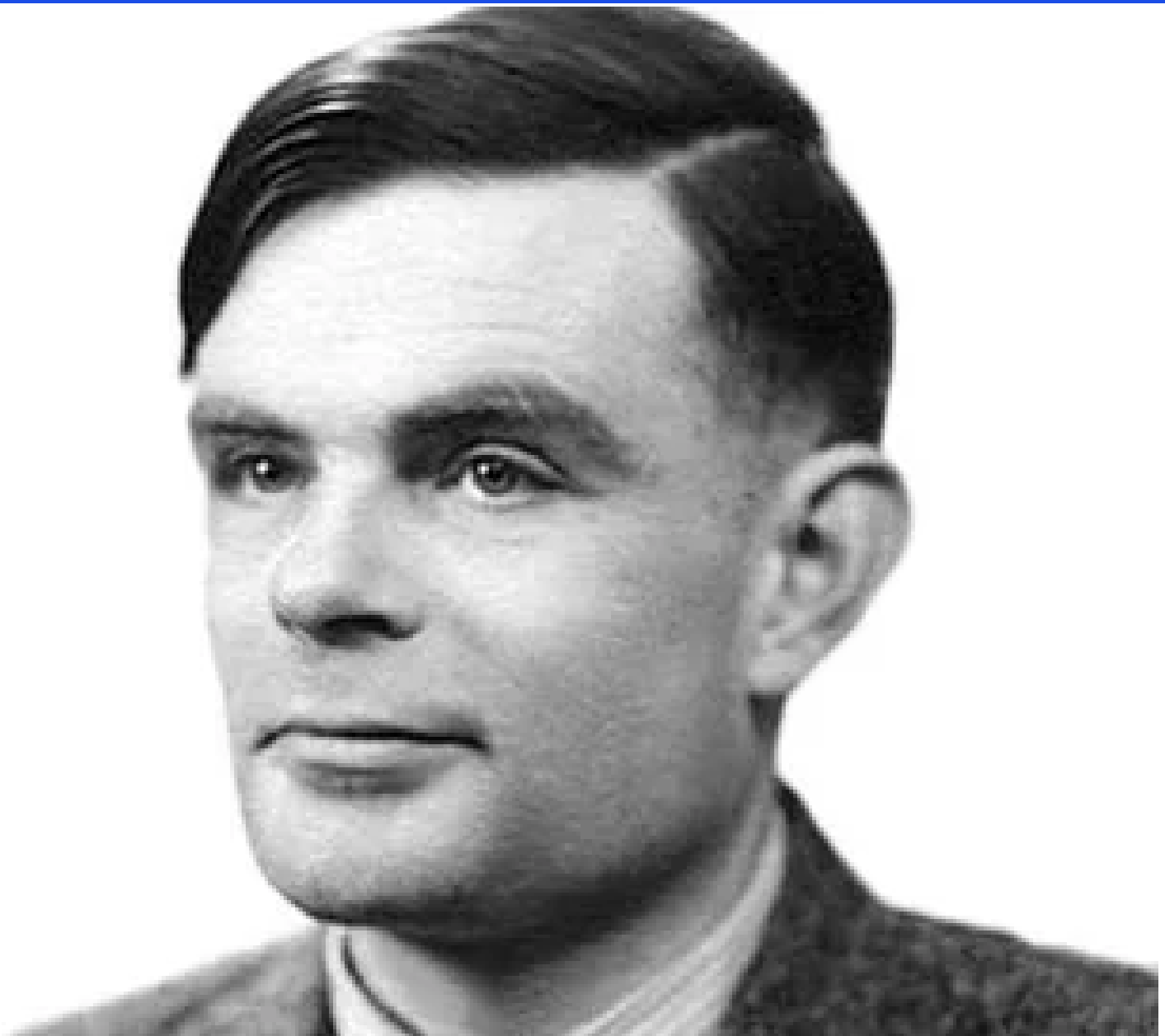
(1822-1884)

Gregor Mendel
(1822-1884)

(1912–1954)



Alan Turing
(1912–1954)



(1879–1955 & 1875–1948)

Albert Einstein. . . et sa femme Mileva Marić/Einstein (1879–1955 & 1875–1948)

(1883–1971)

Cyril Burt
(1883–1971)

(1906–1978)

Kurt Godel
(1906–1978)

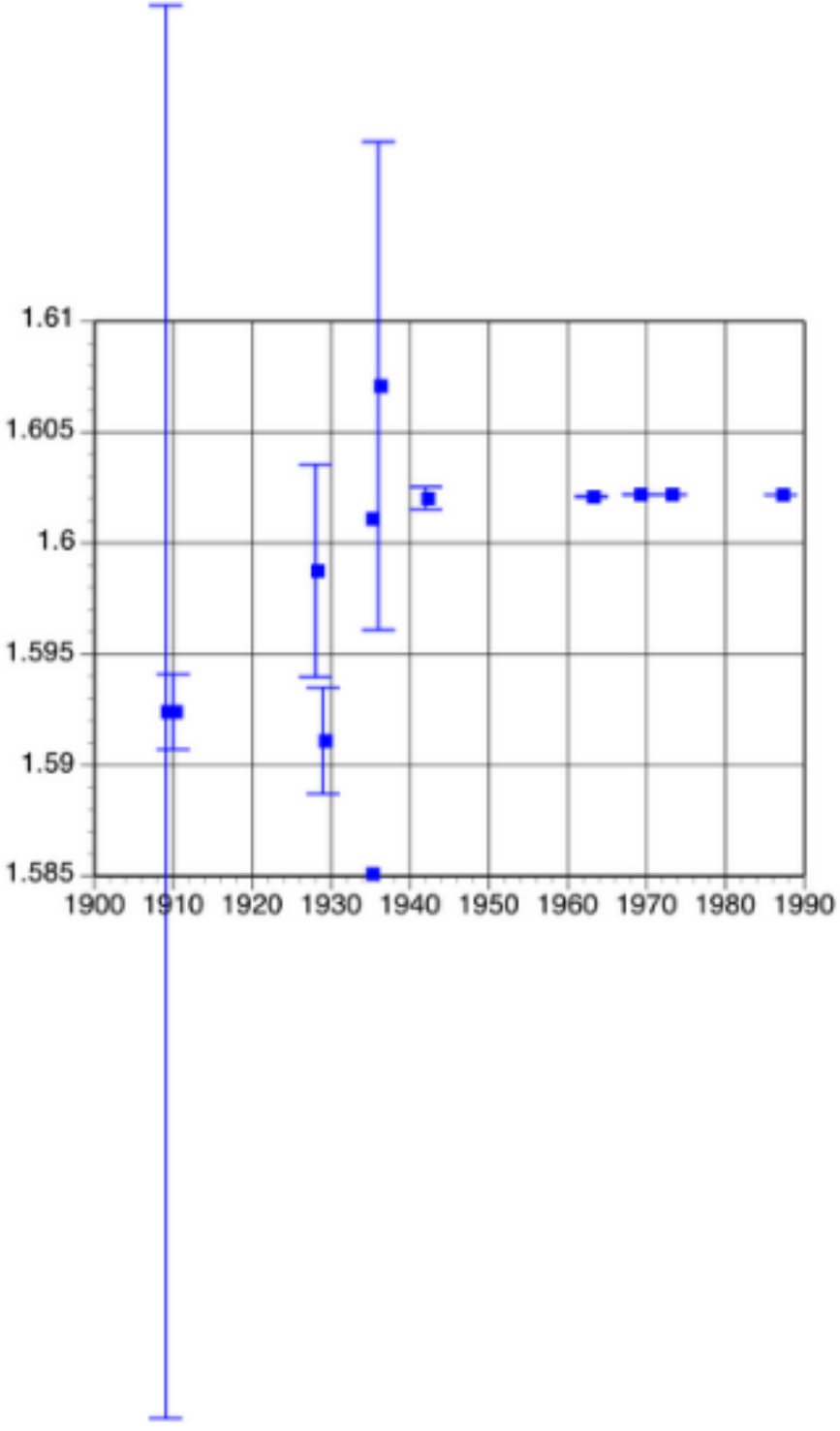
(1957–)


children
have
AUTISM?
Treating
Autism
www.TreatingAutism.com

Andrew Wakefield
(1957–)

(1868–1953)

Robert Andrews Milikan
(1868–1953)



Millikan (notebooks)
Millikan (published)
Erik Backlin, Nature 1929
[Birge, 1929]
Backlin and Flemberg, Nature 1936
Backlin and Flemberg, cited in HR Robinson RPP 1937
Gunnar Kellström PR 1936
[Birge, 1942]
[Dummond and Cohen, 1963]
[Taylor et al, 1969]
[Cohen and Taylor, 1973]
[Cohen and Taylor, 1987]

(1952–)

Didier Raoult (1952–)

- Plus de 2 200 articles
- Période 2007-13 : 636 articles (1 article à tous les 4 jours !)
  «*In April 2017, on Google Scholar citations, he cumulated over 104,000 citations and an h index of 148.*» (Wikipedia)

- Son labo emploie plus de 200 personnes.
- Avoue «ne pas relire les manuscrits comportant sa signature avant soumission» !
- Banni de publication dans une douzaine de revues (*circa.* 2006)