Extracting Governing Equations from Data

Rui Carvalho [1,2,3]

November 20, 2019

[1]Department of Engineering, Durham University, Lower Mountjoy, South Road, Durham, DH1 3LE, UK

[2]Durham Energy Institute, Durham University, South Road, Durham, DH1 3LE, UK

[3]Institute for Data Science, Durham University, South Road, Durham, DH1 3LE, UK

Historical context

Gauss: Statistics, Optimisation and Astronomy

Gauss, Statistics, and Gaussian Elimination

G. W. STEWART*

Gaussian elimination is the algorithm of choice for the solution of dense linear systems of equations. However, Gauss himself originally introduced his elimination procedure as a way of determining the precision of least squares estimates and only later described the computational algorithm. This article tells the story of Gauss, his algorithm, and its relation to his probabilistic development of least squares.

Key Words: Gauss; Gaussian elimination; Theory of least squares.

1. INTRODUCTION

Everyone knows that Gauss invented Gaussian elimination, and for practical purposes everyone is right. What is less well known is that Gauss introduced the procedure as a mathematical tool to get at the precision of least squares estimates. In fact, the computational component in his original description is so little visible that it takes some doing to see an algorithm in it.

Gaussian elimination, therefore, was not conceived as a general numerical algorithm with applications in statistics and least squares. Rather it was a procedure that sprang from the interface of statistics and computation. Because the full story is known only to the few who have consulted the original sources, I hope my readers will be interested to see how Gauss did things. But there is more than the satisfaction of idle curiosity here. Gauss and Laplace were the premier statisticians of their day, and Gauss alone was the premier numerical analyst. Today we still have something to learn from observing Gauss's practices.

2. CHRONICLES

The principle of least squares arose from the problem of combining sets of overdetermined equations to form a square system that could be solved for the unknowns. The problem went under the name of the combination of observations, and has been well surveyed by Stigler (1986) in his *History of Statistics*. By way of background, I will relate in chronological order the major events in the story of least squares, from Gauss's first discovery to his final treatment in the 1820s.

*Professor, Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742

Journal of Computational and Graphical Statistics, Volume 4, Number 1, Pages 1–11

Carl Friedrich Gauss (1777-1855)

Gauss: Statistics, Optimisation and Astronomy

Although we tend to regard Gauss chiefly as a mathematician, it was as an astronomer that he first made his mark. On New Year's Day of 1801, the astronomer Piazzi discovered the asteroid Ceres. The new planet became unobservable after only nine degrees of an arc had been recorded, and astronomers were faced with problem of determining where to look for it next. Gauss undertook the calculation, using new techniques in physical astronomy and presumably his principle of least squares. At the end of 1801 he predicted where in the heavens the asteroid would be found, and his reputation was made.

Source: Margaret Wright, Courant Institute

Mathematical Optimisation

(Mathematical) Optimisation problem

$$\begin{aligned} &\text{minimize } f_0(x) && && (1)\\ &\text{subject to } f_i(x) \leq b_i, && i = 1, \ldots, m && (2) \end{aligned}$$

- $x = (x_1, \ldots, x_n)$: optimisation variables.
- $f_o : \mathbf{R}^n \to \mathbf{R}$: objective function.
- $f_i : \mathbf{R}^n \to \mathbf{R}, i = 1, \ldots, m$: constraint functions.

optimal solution x^* has smallest value of f_0 among all vectors that satisfy the constraints.

Solving optimisation problems

General optimisation problem

- Very difficult to solve.
- Methods involve some compromise, *e.g.*, very long computation time or not always finding a solution.

Exceptions

- Least-squares problems.
- Linear programming problems.
- Convex optimisation problems.

Mathematical Optimisation

> The great watershed in optimisation isn't between linearity and nonlinearity, but convexity and noncovexity.

Rockafellar, 1993

Broadly speaking, optimisation problems involving convex functions tend to be *nice*:

- Any minimiser is the unique global minimiser;
- Convex optimisation problems can often be solved rapidly;
- Theoretical guarantees of convergence.


The Economist
Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar
OCTOBER 19TH–25TH 2013
Economist.com
HOW SCIENCE GOES WRONG.
99
Einsteinium


The Economist
China at 70—pomp and protests
Big Tech and the state gird for battle
Europe's anti-populist backlash
What would Trump's gators cost?
Masters of the universe
How machines are taking over Wall Street

Quant funds can be divided into two groups: those like Stockfish, which use machines to mimic human strategies; and those like AlphaZero, which create strategies themselves. For 30 years quantitative investing started with a hypothesis, says a quant investor. Investors would test it against historical data and make a judgment as to whether it would continue to be useful. Now the order has been reversed. "We start with the data and look for a hypothesis," he says.

Figure 1: Source: *March of the machines*, The Economist 05/10/2019

> *Taking the original Fisherian point of view, significance testing is an effort to address the selection of an interesting finding regarding a single parameter from the background noise. Modern science faces the problem of selection of promising findings from the noisy estimates of many.*

Y. Benjamini and Y. Hechtlinger, *Biostatistics* (2014) **15**, 13-16

Fitting

Norms and l_p norms

Examples of norms on the vector space $\mathbf{R}^n$ are the so-called l_p norms, defined as

$$\|x\|_p := \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p}, \quad 1 \leq p < \infty. \tag{3}$$

- For $p = 2$ we obtain the standard Euclidean length

$$\|x\|_2 := \sqrt{\sum_{k=1}^{n} x_k^2}, \tag{4}$$

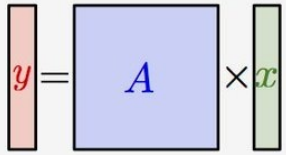
- For $p = 1$ we obtain the sum-of-absolute-values length

$$\|x\|_1 := \sum_{k=1}^{n} |x_k|. \tag{5}$$

Linear systems in a nutshell

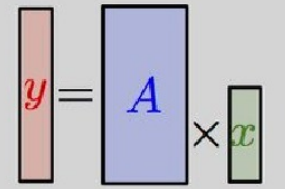
Solving $y \underset{\approx}{=} A\,x \in \mathbb{R}^m \qquad A \in \mathbb{R}^{m \times n}$

Determined $(m = n)$: $x = A^{-1}y$



Over-determined $(m > n)$: $\min_x \|Ax - y\|^2$

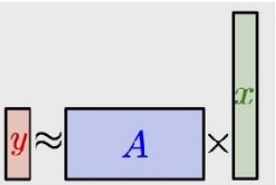
$x = (A^\top A)^{-1} A^\top y \overset{\text{def.}}{=} A^+ y$



least squares

Under-determined $(m < n)$: $\min_x \{\|x\| \;;\; Ax = y\}$

$x = A^\top (AA^\top)^{-1} y \overset{\text{def.}}{=} A^+ y$

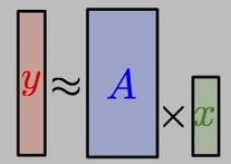


least norm problem

A ill-posed and/or noise: $\min_x \|Ax - y\|^2 + \lambda \|x\|^2$

$x = (A^\top A + \lambda \mathrm{Id}_n)^{-1} A^\top y \overset{\lambda \to 0}{\longrightarrow} A^+ y$

$= A^\top (AA^\top + \lambda \mathrm{Id}_m)^{-1} y$ (Woodbury identity)



modified from Gabriel Peyré

Least squares problem

Least squares problem: choose x to minimise $f(x) = \|Ax - b\|_2^2$ where $A \in \mathbf{R}^{m \times n}$ with $m \geq n$, and $b \in \mathbf{R}^m$ are problem data.

- $m \times n$ matrix A is tall, so $Ax = b$ is over-determined.
- For most choices of b, there is no x that satisfies $Ax = b$.
- *Residual*: $r = Ax - b$.
- Idea: make residual as small as possible, if not 0.
- Assume that the columns of A are independent (the Gram matrix $A^T A$ is invertible), the least-squares approximation problem has the unique solution:

$$x = (A^T A)^{-1} A^T b. \tag{6}$$

- Compare with the solution of the square invertible system $Ax = b$:

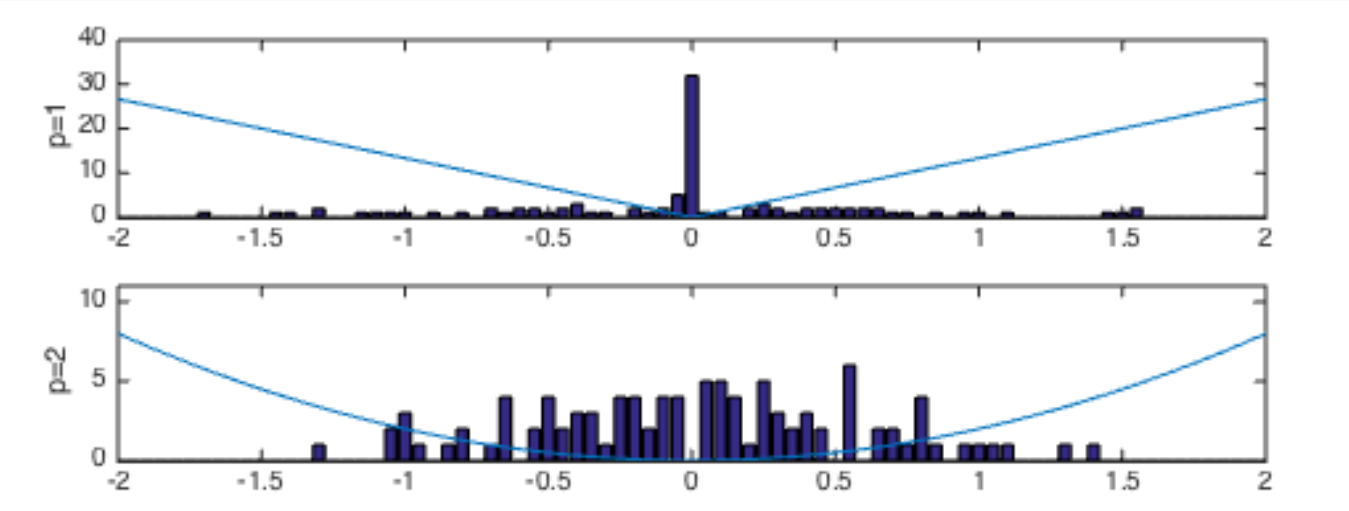
$$x = A^{-1} b \tag{7}$$

Example: Penalty function approximation

$$\begin{array}{ll} \text{minimise} & \phi(r_1) + \cdots + \phi(r_m) \\ \text{subject to} & r = Ax - b \end{array} \tag{8}$$

Histogram of residuals for the l_1 and l_2 penalty functions (m $= 100$, n $=$ 30):

$$\phi(u) = |u|, \phi(u) = u^2 \tag{9}$$



Regularised approximation

how well the data agrees with the model $Ax=b$

$$\text{minimise} \quad (\|Ax-b\|, \|x\|) \tag{10}$$

how large are your model parameters

- $A \in \mathbf{R}^{m\times n}$ is a matrix of n predictors;
- $x \in \mathbf{R}^n$ are the parameters;
- $b \in \mathbf{R}^m$ is a vector of responses.

Idea:

- We want a good fit of $Ax = b$, but we want to do it efficiently, *i.e.*, with small $\|x\|$, so we add to the objective a term that penalises large x.
- Regularisation avoids large x.

Regularised approximation

$$\text{minimise} \quad \|Ax - b\| + \lambda \|x\| \tag{11}$$

- Standardise A, so that each column has zero mean and unit variance.
- Solution for $\lambda > 0$ traces out optimal trade-off curve (sweep λ from 0 to ∞).
- Convex problem, so we know how to solve it.

Ridge regression

$$\text{minimise} \quad \|Ax - b\|_2^2 + \lambda \|x\|_2^2 \tag{12}$$

Squared objective makes problem smooth (second-derivative exists) and we have an analytical solution.

Can be solved as a least squares problem with the analytical solution:

$$x = (A^T A + \lambda I)^{-1} A^T b. \tag{13}$$

The Lasso

Lasso (*least absolute shrinkage and selection operator*)

$$\text{minimise} \quad \|Ax - b\|_2^2 + \lambda \|x\|_1 \tag{14}$$

- Statistical procedure that solves the ordinary least squares problem penalised with an l_1 norm (it promotes sparsity).
- If $\lambda = 0$ you get the least squares solution.
- if $\lambda = \infty$ you get $x = 0$.
- The Lasso tries to fit a model by selecting variables:
  - Start at $\lambda = \infty$, where you find no variables;
  - As you decrease λ the Lasso will include more and more variables, one at a time.
- Convex problem, so we know how to solve it efficiently.

The Lasso

$$\text{minimise} \quad \|Ax - b\|_2^2 + \lambda \|x\|_1 \tag{15}$$

Can also be written as:

$$\underbrace{\sum_{i=1}^{n} \left(y_i - \overbrace{\beta_0}^{\text{intercept}} - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{Residual Sum of Squares}} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{16}$$

EV charging

ELAAD dataset

- ElaadNL: Dutch research organisation involved in the development and deployment of EV charging technologies.
- $1,747$ georeferenced charging stations.
- $54,000$ users, each identified by a unique *id*.
- $1,060,763$ charging events.
- Data collected between January 2012 and March 2016.

ELAAD: spatial data

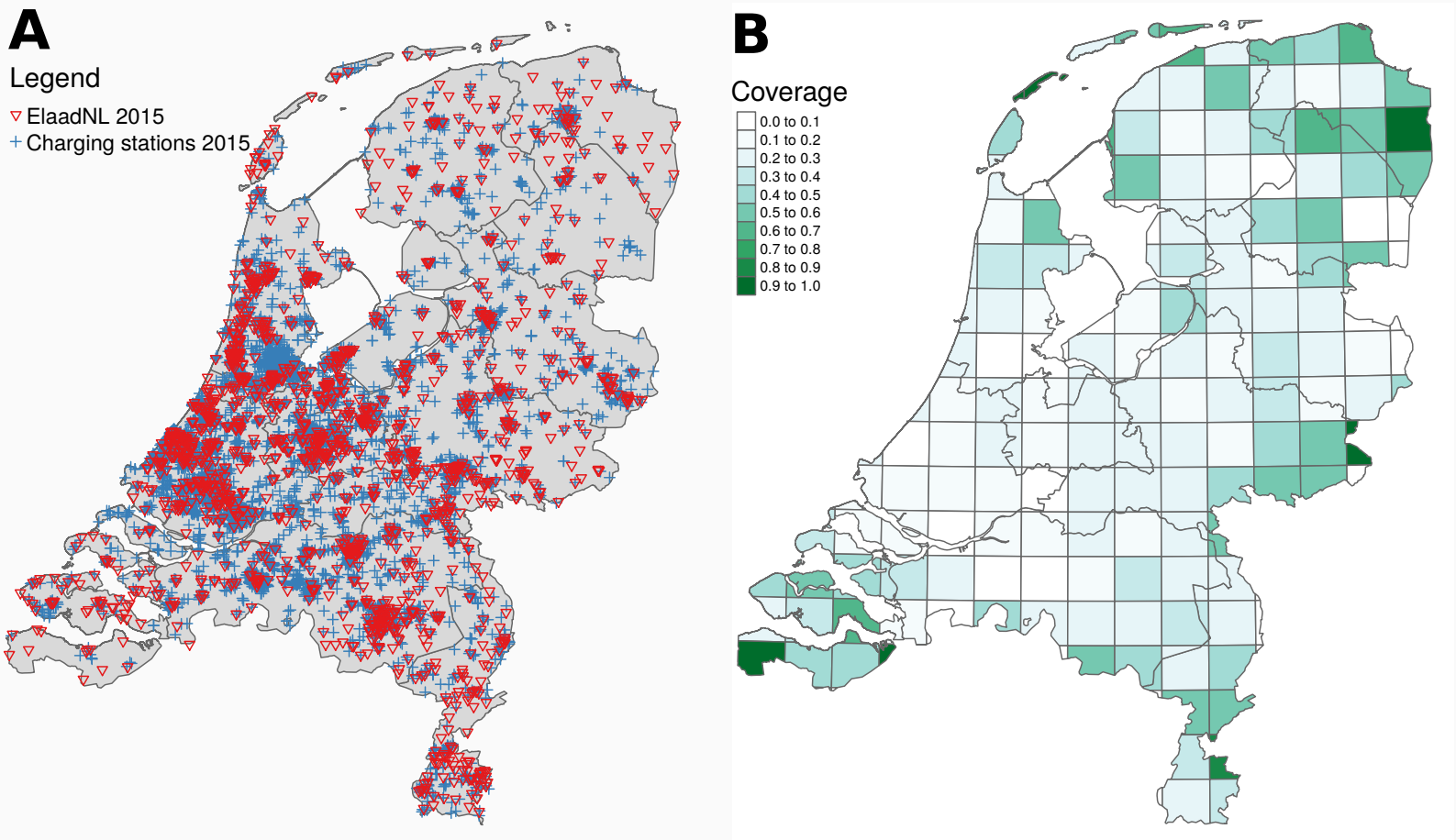


Figure 2: Public charging stations in the ElaadNL data set in 2015 (triangles) shown together with the Charging stations 2015 dataset (crosses). In the year 2015, 17 786 publicly available connectors for slow charging were operational in the NL. We identified 8 400 unique positions of charging stations, i.e. considering the distribution of connectors at charging stations observed in the ElaadL dataset, this data covers 78.3% of all stations. In panel (B), we estimated the spatial representativeness of the ElaadNL data sets by calculating the ratio between the number of station in ELaadNL and in Charging stations 2015 located in squared cells of a regular grid. In the largest cities, Amsterdam and Rotterdam, the data contains a small percentage of all charging stations.

Predictors: GIS data

- **Vector data**
  - **Polygon data**
    - population cores,
    - neighbourhoods data,
    - energy atlas,
    - liveability,
    - land use and land cover (urban atlas, CBS land cover).
  - **Polyline data**
    - traffic flow data.
  - **Point data**
    - OSM amenities,
    - OpenChargeMap.
- **Raster data**
  - LandScan - ambient population.

Vector polygon data: Population cores (2168 cores)

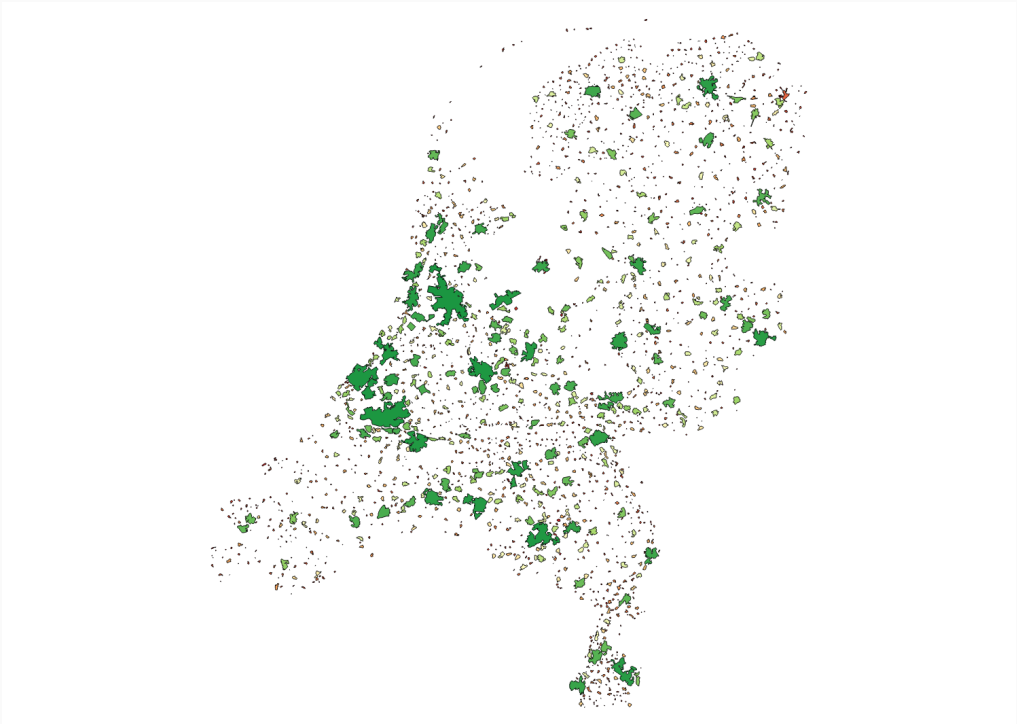


Figure 3: Population cores are continuous spatial units with at least 25 homes or 50 inhabitants (102 predictors). Source: Statistics Netherlands https://opendata.cbs.nl/

Polygon data: Population cores (2168 cores)

- number of persons in private household,
- number of persons in private households, 0 to 15 years,
- number of persons in private households, 15 to 25 years,
- number of persons in private households 25 to 45 years,
- number of persons in private households, 45 to 65 years,
- number of persons in private households 65 years or older,
- number of persons in one-person households,
- number of people in multi-person household with children,
- number of people in multi-person household without children,
- nercentage of working population, 15-24 years old,
- number of households of two persons,
- number of households of three persons,
- the number of residential units,
- ...

Lasso

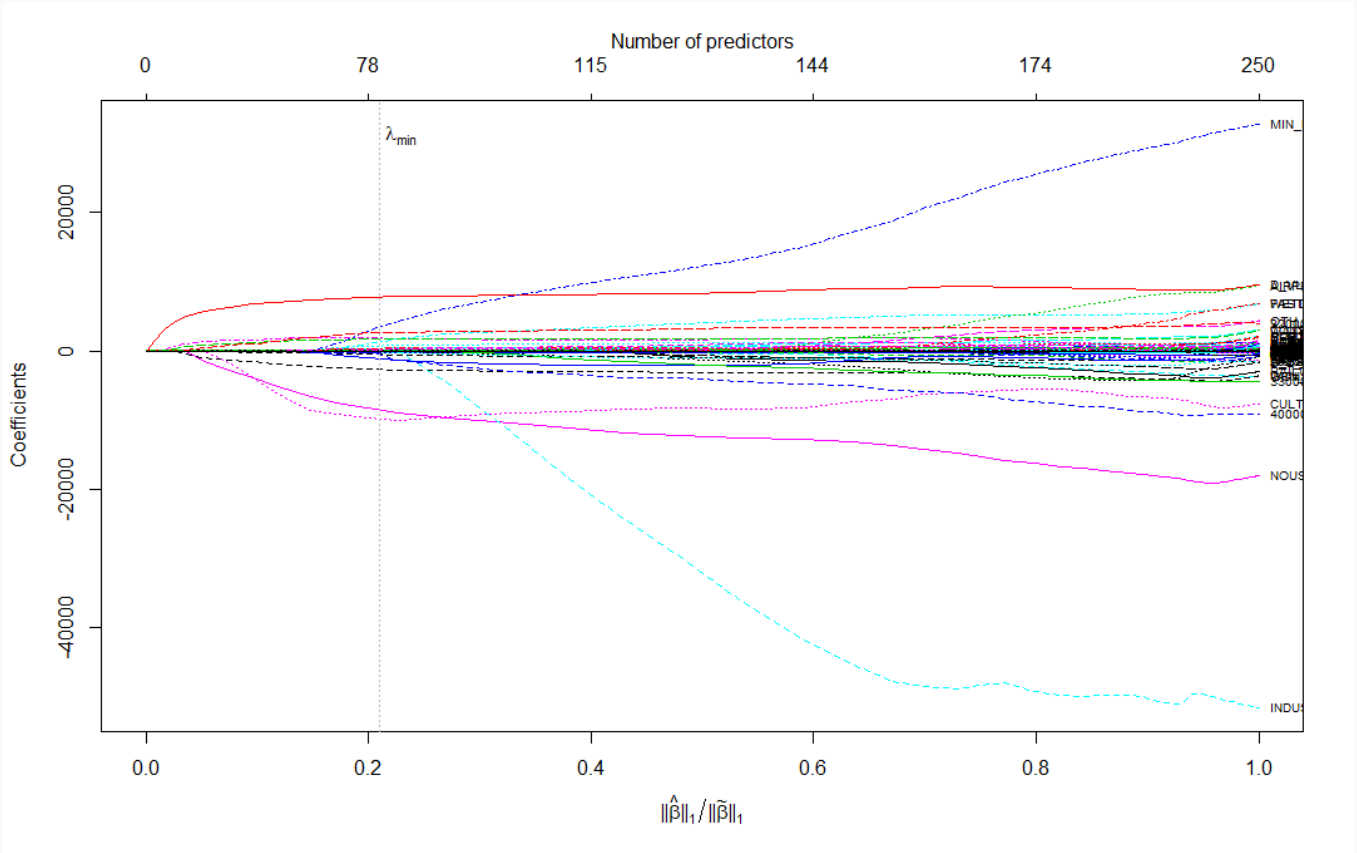
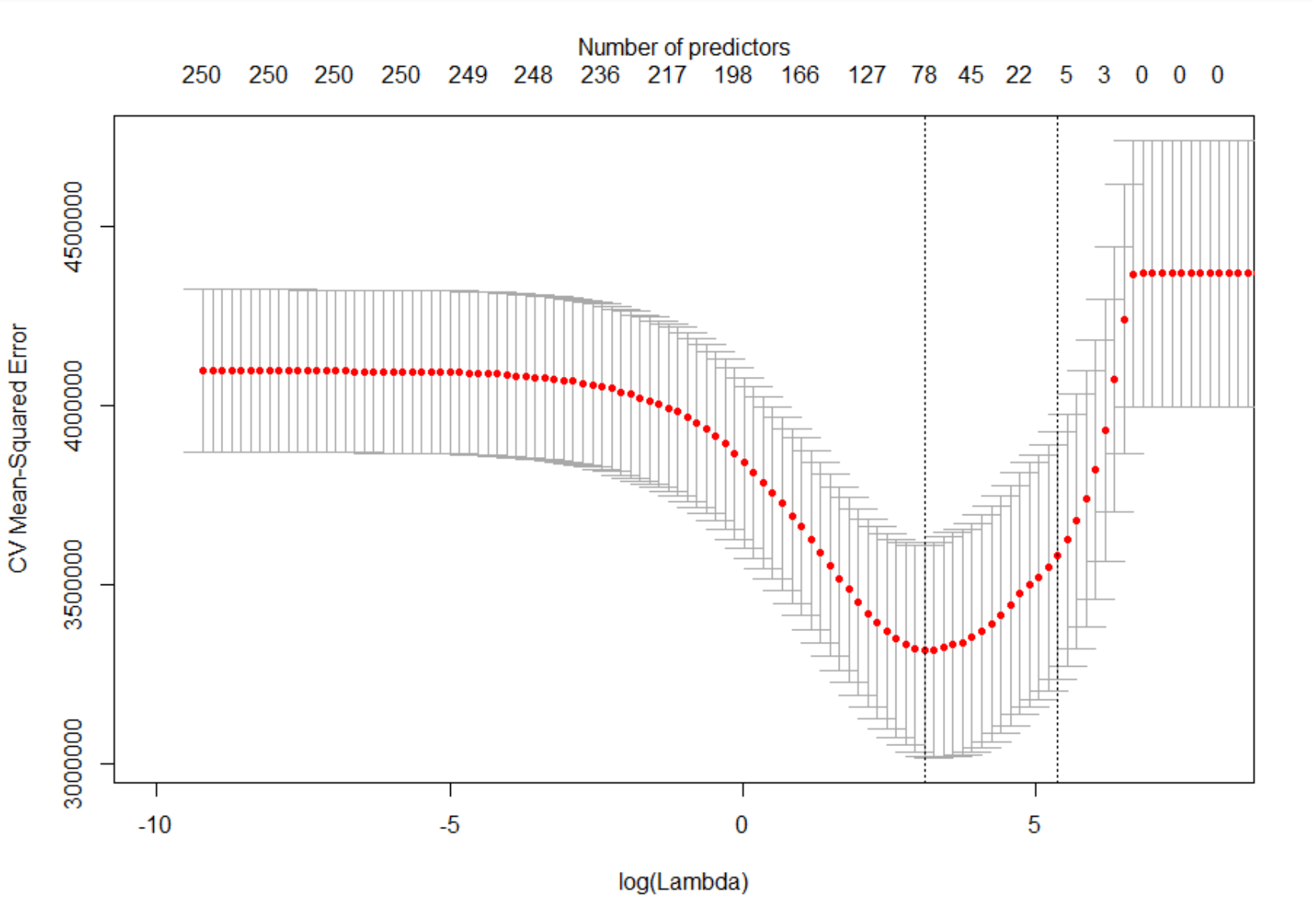


Figure 4: Lasso

$$\underbrace{\sum_{i=1}^{n}\left(y_i - \overbrace{\beta_0}^{\text{intercept}} - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{Residual Sum of Squares}} + \lambda \sum_{j=1}^{p}|\beta_j| \tag{17}$$

Lasso


Number of predictors
250 250 250 250 249 248 236 217 198 166 127 78 45 22 5 3 0 0 0
CV Mean-Squared Error
3000000
3500000
4000000
4500000
-10
-5
0
5
log(Lambda)

Lasso fit

- Lasso does variable selection on 240 predictors.
- At the optimal λ, we reduce the number of predictors to 79 (about 1/3 of the original predictors).
- $R^2 = 0.362$ at optimum value ($R^2(adjusted) = 0.316$).
- The λ_{min} is the one which minimises the cross-validation error. The $\lambda_{1\sigma}$ is the λ value within 1 standard error of λ_{min}.

Coefficient	Meaning
-27.8627	The percentage of working population working in Mining, Manufacturing and Construction.
20.3313	The percentage of working population employed in commercial services
18.169	The percentage working population engaged in agriculture, forestry and fisheries, industry, commercial and non-commercial services
-0.1884	The percentage of the number of multi-person households without children
4.1384	Number of business Services
14.5717	Property unknown (no link between the addresses of the Key Registers Addresses and the housing register Cadaster).
57.8468	Average income per inhabitant
-0.9093	Average distance of all residents in an area to the nearest shops for groceries.

Data-driven discovery of dynamical systems

Data-driven discovery of dynamical systems

Goal of computationally-oriented scientists:

> Inferring a (typically nonlinear) model from observations that both correctly identifies the underlying dynamics *and* generalises qualitatively and quantitatively to unmeasured parts of the phase, parameter, or application space.

- ODE or PDE system described by

$$u_t = N(u, x, t; \overbrace{\mu}^{\text{parameter}}) \tag{18}$$

- Our objective is to discover $N(.)$ given only time-series measurements of the system.
- A key assumption (prior) is that the true $N(.)$ is comprised of only a few terms, making the model sparse in the space of all possible combinations of functions.
- For example, Burgers' equation

$$N = -uu_x + \mu u_{xx} \tag{19}$$

and the harmonic oscillator

$$N = -i\mu x^2 u - i\hbar u_{xx}/2 \tag{20}$$

each have only two terms.

Identifying Dynamical Systems

PNAS

Discovering governing equations from data by sparse identification of nonlinear dynamical systems

Steven L. Brunton[a,1], Joshua L. Proctor[b], and J. Nathan Kutz[c]

[a]Department of Mechanical Engineering, University of Washington, Seattle, WA 98195; [b]Institute for Disease Modeling, Bellevue, WA 98005; and [c]Department of Applied Mathematics, University of Washington, Seattle, WA 98195

Edited by William Bialek, Princeton University, Princeton, NJ, and approved March 1, 2016 (received for review August 31, 2015)

Extracting governing equations from data is a central challenge in many diverse areas of science and engineering. Data are abundant whereas models often remain elusive, as in climate science, neuroscience, ecology, finance, and epidemiology, to name only a few examples. In this work, we combine sparsity-promoting techniques and machine learning with nonlinear dynamical systems to discover governing equations from noisy measurement data. The only assumption about the structure of the model is that there are only a few important terms that govern the dynamics, so that the equations are sparse in the space of possible functions; this assumption holds for many physical systems in an appropriate basis. In particular, we use sparse regression to determine the fewest terms in the dynamic governing equations required to accurately represent the data. This results in parsimonious models that balance accuracy with model complexity to avoid overfitting. We demonstrate the algorithm on a wide range of problems, from simple canonical systems, including linear and nonlinear oscillators and the chaotic Lorenz system, to the fluid vortex shedding behind an obstacle. The fluid example illustrates the ability of this method to discover the underlying dynamics of a system that took experts in the community nearly 30 years to resolve. We also show that this method generalizes to parameterized systems and systems that are time-varying or have external forcing.

dynamical systems | machine learning | sparse regression | system identification | optimization

dynamical systems from data. However, symbolic regression is expensive, does not scale well to large systems of interest, and may be prone to overfitting unless care is taken to explicitly balance model complexity with predictive power. In ref. 4, the Pareto front is used to find parsimonious models. There are other techniques that address various aspects of the dynamical system discovery problem. These include methods to discover governing equations from time-series data (6), equation-free modeling (7), empirical dynamic modeling (8, 9), modeling emergent behavior (10), and automated inference of dynamics (11–13); ref. 12 provides an excellent review.

Sparse Identification of Nonlinear Dynamics (SINDy)

In this work, we reenvision the dynamical system discovery problem from the perspective of sparse regression (14–16) and compressed sensing (17–22). In particular, we leverage the fact that most physical systems have only a few relevant terms that define the dynamics, making the governing equations sparse in a high-dimensional nonlinear function space. The combination of sparsity methods in dynamical systems is quite recent (23–30). Here, we consider dynamical systems (31) of the form

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t)). \qquad [1]$$

The vector $\mathbf{x}(t) \in \mathbb{R}^n$ denotes the state of a system at time t, and the function $\mathbf{f}(\mathbf{x}(t))$ represents the dynamic constraints that de-

Data-driven discovery of dynamical systems: method

Method:

- Construct a library $\Theta(U)$ of candidate linear, nonlinear, and partial derivative terms for the right-hand side.
- Each column of $\Theta(U)$ contains the values of a candidate term evaluated using the collected data.
- In this library, one can write the dynamics as

$$U_t = \Theta(U)\xi \tag{21}$$

  where
  - U_t is a vector of time derivatives of the measurement data.
  - ξ is a sparse vector, with each nonzero entry corresponding to a functional term to be included in the dynamics.
- Finding the sparsest vector ξ consistent with the measurement data is now feasible with advanced methods in sparse regression, which makes it possible to find the most parsimonious model while circumventing a combinatorial search.

Identifying the Lorenz Equations

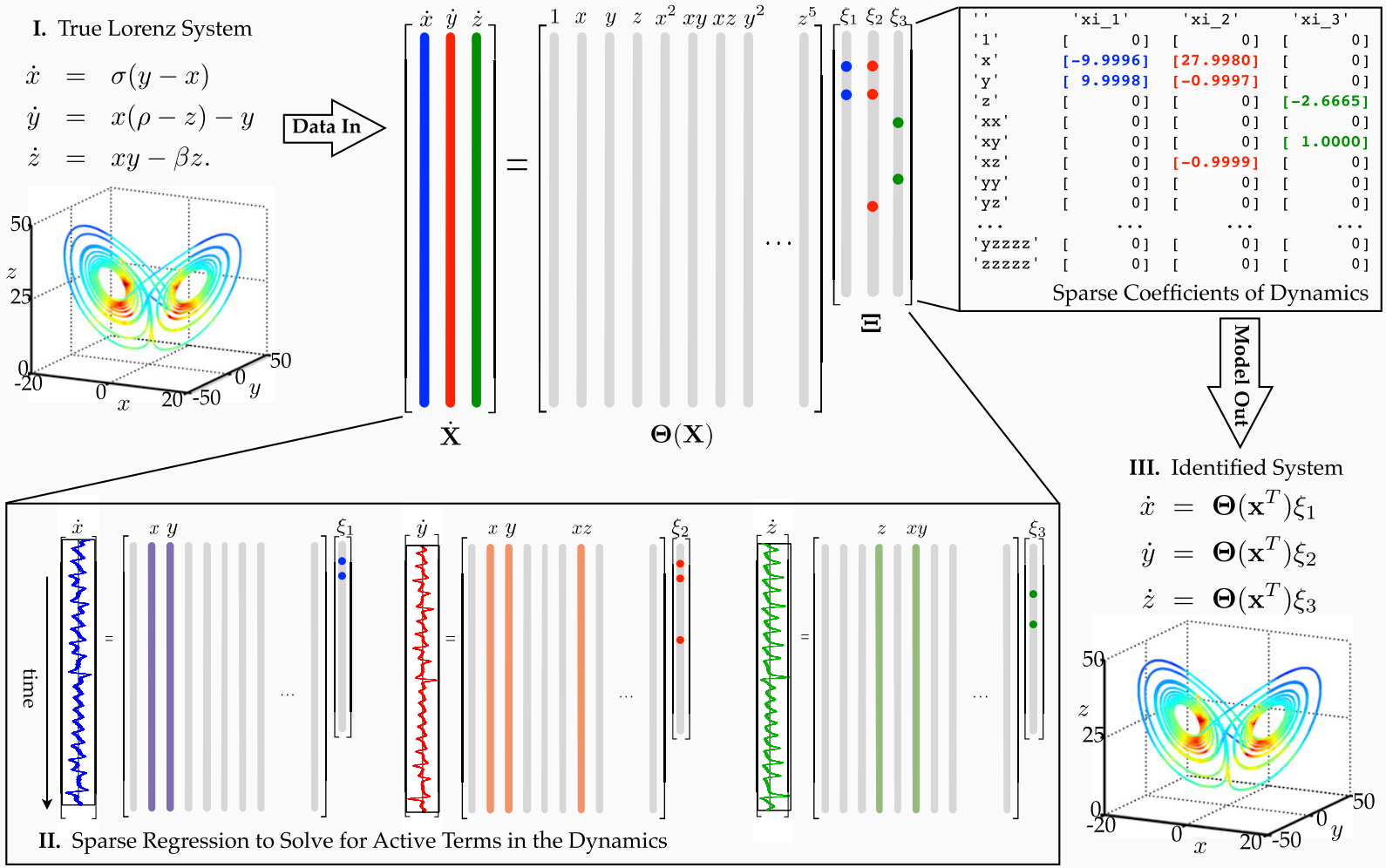


Fig. 1. Schematic of the SINDy algorithm, demonstrated on the Lorenz equations. Data are collected from the system, including a time history of the states $\mathbf{X}$ and derivatives $\dot{\mathbf{X}}$; the assumption of having $\dot{\mathbf{X}}$ is relaxed later. Next, a library of nonlinear functions of the states, $\Theta(\mathbf{X})$, is constructed. This nonlinear feature library is used to find the fewest terms needed to satisfy $\dot{\mathbf{X}} = \Theta(\mathbf{X})\Xi$. The few entries in the vectors of Ξ, solved for by sparse regression, denote the relevant terms in the right-hand side of the dynamics. Parameter values are $\sigma = 10, \beta = 8/3, \rho = 28$, $(x_0, y_0, z_0)^T = (-8, 7, 27)^T$. The trajectory on the Lorenz attractor is colored by the adaptive time step required, with red indicating a smaller time step.

Knockoffs

Most published research findings are probably false [Ioannidis]

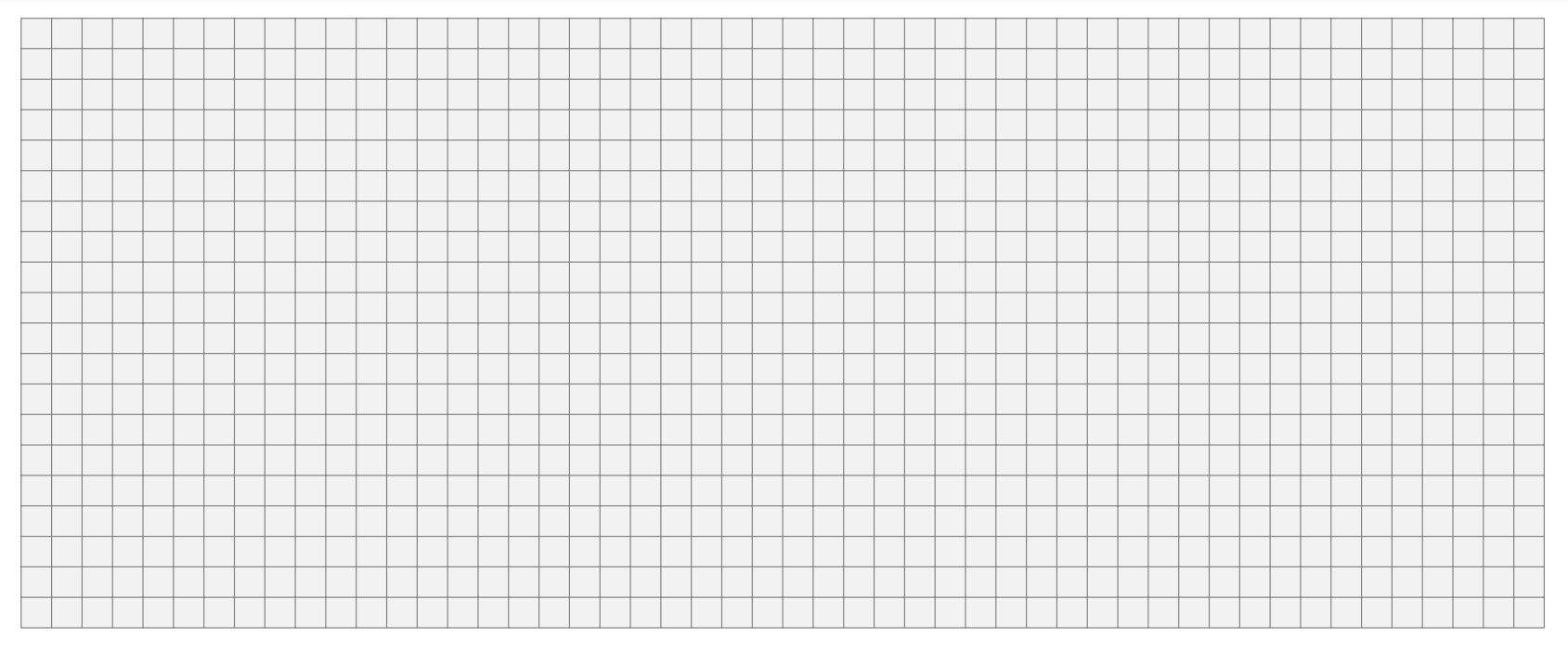


Figure 6: 1000 hypotheses to test.

credit: Emmanuel Candes (Stanford), The Economist (19/10/2013)

Most published research findings are probably false [Ioannidis]

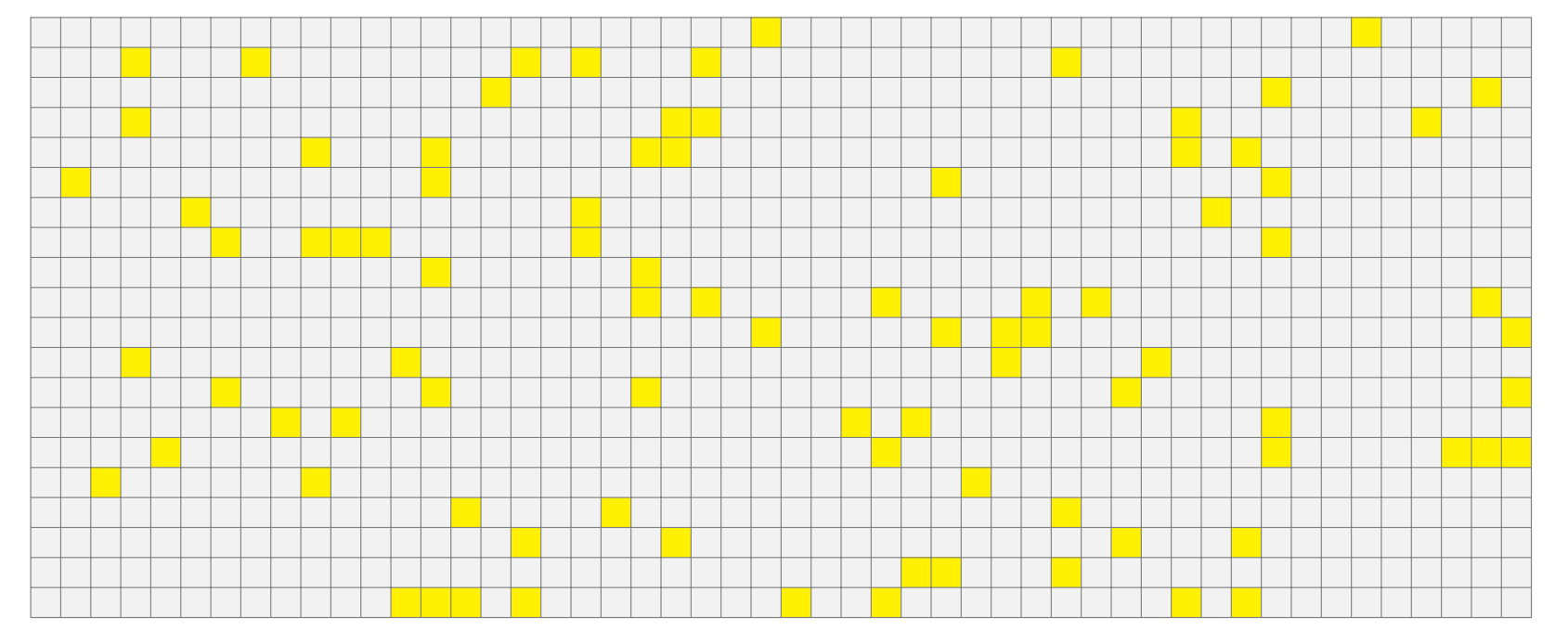


Figure 7: 1000 hypotheses, 100 potential discoveries.

Most published research findings are probably false [Ioannidis]

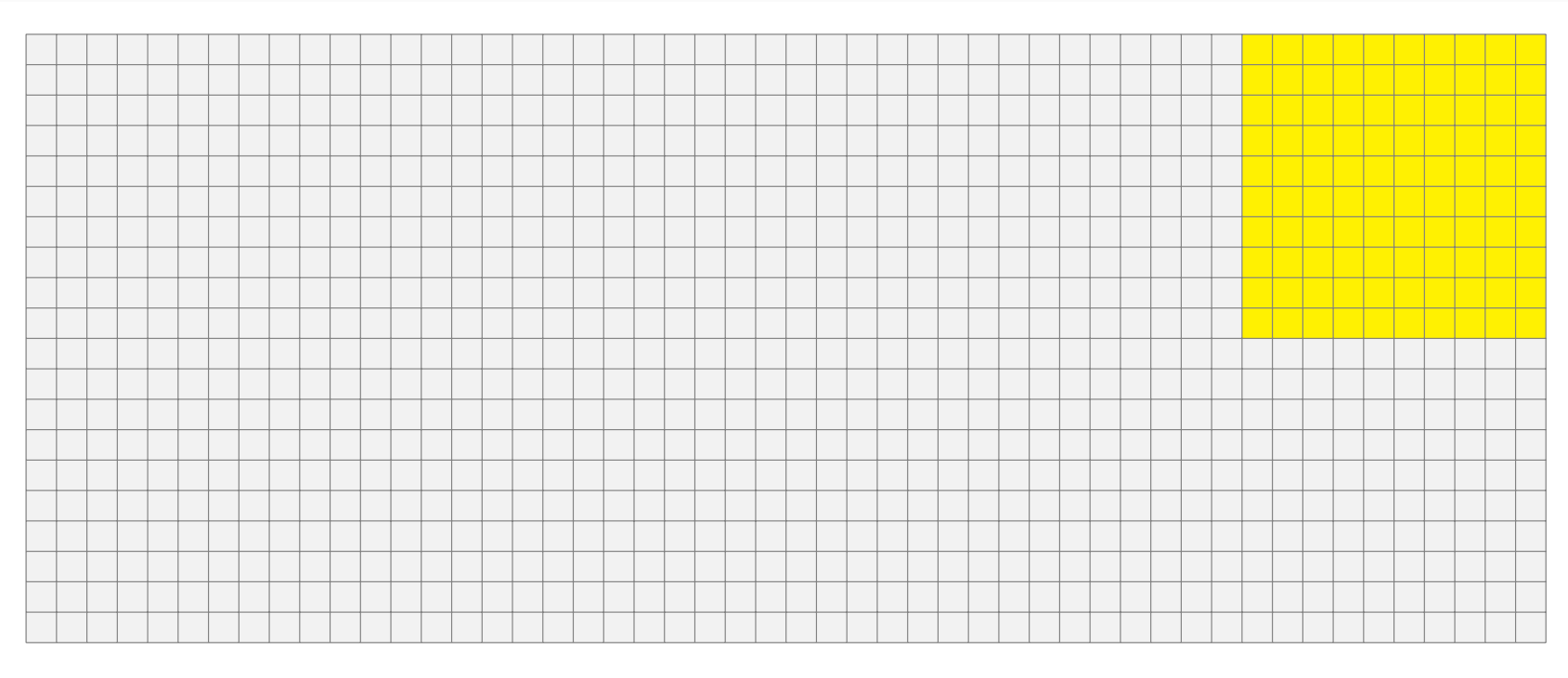
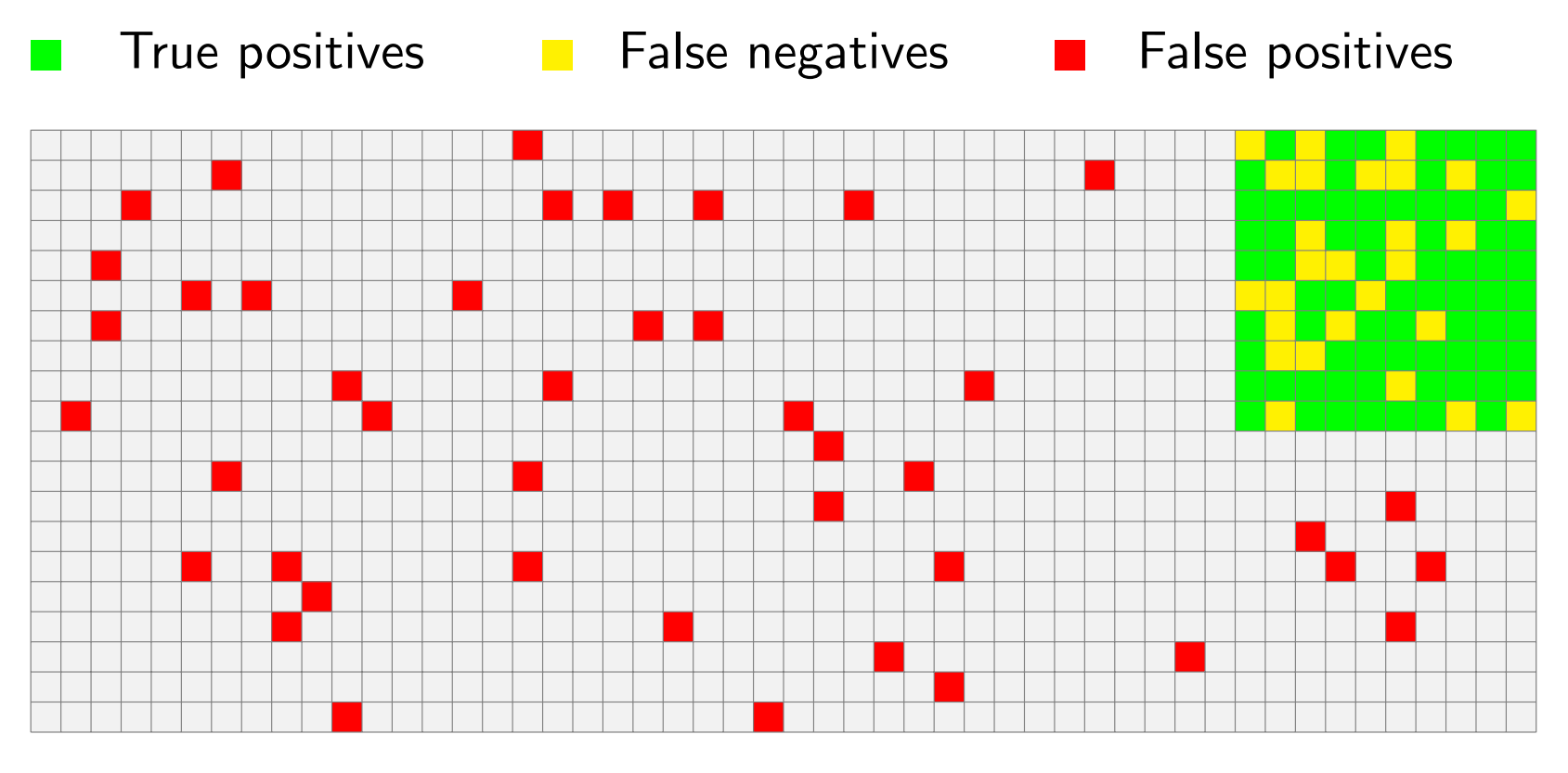


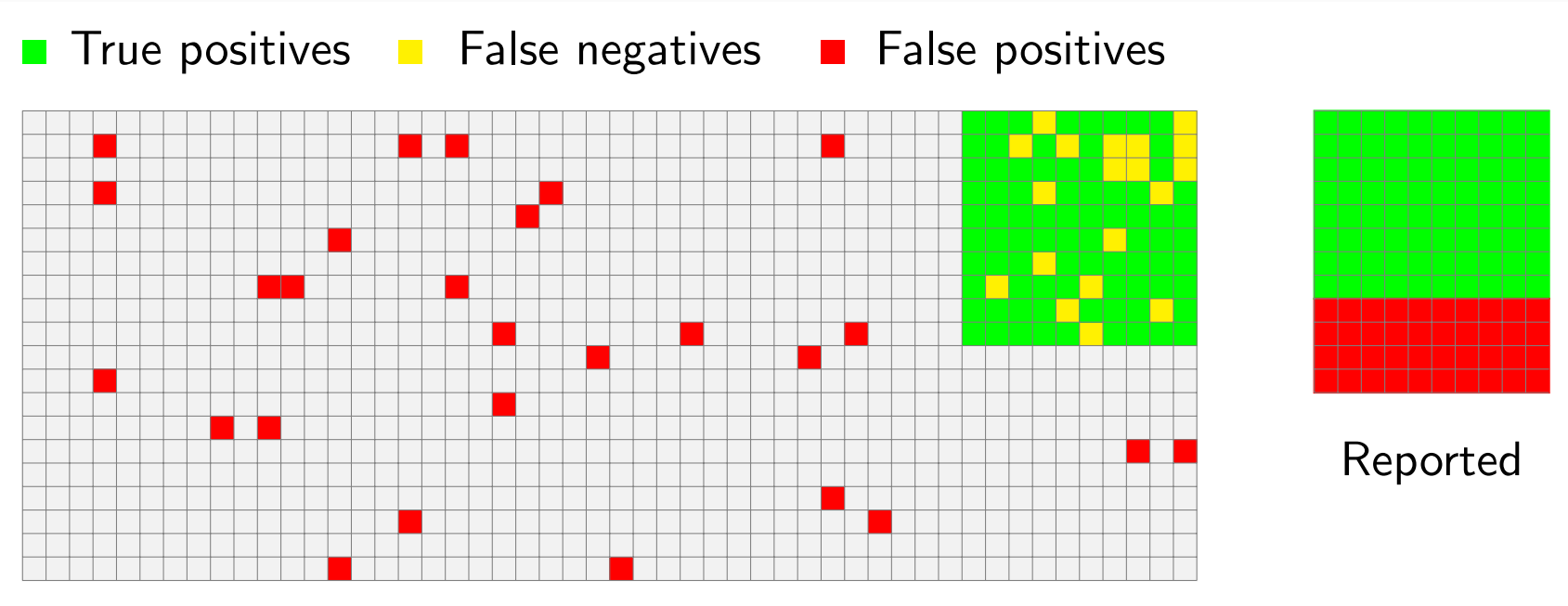
Figure 8: Out of these 1000 hypotheses, 100 hypotheses are potential discoveries (in yellow), but 900 are null (the white squares).

Most published research findings are probably false [Ioannidis]



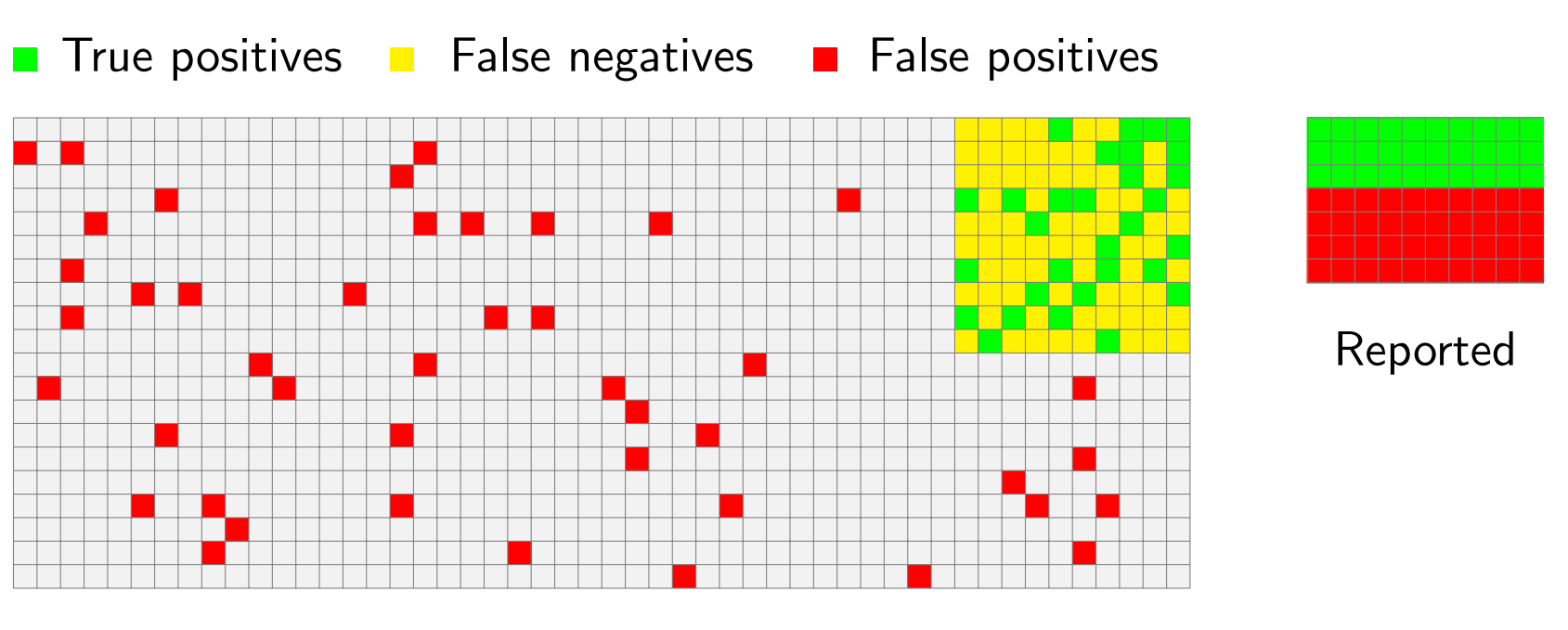
- Power≈80% → true positives≈ 80
  (I have 80% chance to declare potential discoveries [green and yellow squares] as positive [green squares])
- False positives (5% level)≈ 45
  (I detect 5% of the 900 that are completely irrelevant [red squares])

Most published research findings are probably false [Ioannidis]



- When reporting, I don't know which are true positives or false positives, so I just report the positives.
- Observe that a large fraction of reported discoveries is false.
- In this example, over 1 in 3 hypotheses are null and thus cannot be replicated.

Most published research findings are probably false [Ioannidis]



- Now suppose that we drop the power from 80% to 30%: Power$\approx$ 30%.
- I still have on average 45 nulls [red squares].
- But the number of true discoveries dropped: I now have 30 instead of 80.
- False discover proportion:

$$FDP = \frac{45}{30 + 45} = 60\%$$

- Most of what I'm reporting is false!

Small, low-powered studies are endemic in neuroscience

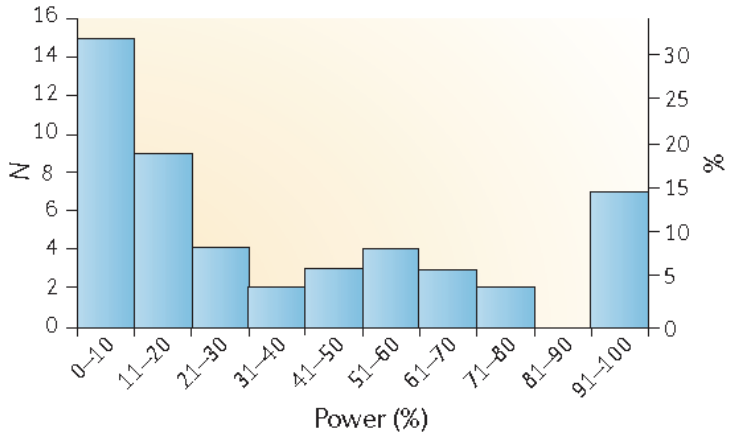


Figure 3 | **Median power of studies included in neuroscience meta-analyses.** The figure shows a histogram of median study power calculated for each of the $n = 49$ meta-analyses included in our analysis, with the number of meta-analyses (N) on the left axis and percent of meta-analyses (%) on the right axis. There is a clear bimodal distribution; $n = 15$ (31%) of the meta-analyses comprised studies with median power of less than 11%, whereas $n = 7$ (14%) comprised studies with high average power in excess of 90%. Despite this bimodality, most meta-analyses comprised studies with low statistical power: $n = 28$ (57%) had median study power of less than 31%. The meta-analyses ($n = 7$) that comprised studies with high average power in excess of 90% had their broadly neurological subject matter in common.

ANALYSIS

Power failure: why small sample size undermines the reliability of neuroscience

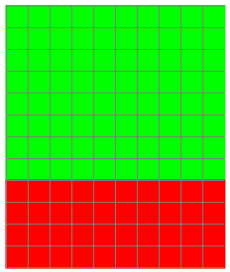
Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

Figure 9: Button et al., *Nature Neuroscience*, vol 14, 365 (2013)

False discovery rate (FDR) [Benjamini-Hochberg]

Selection problem: How do we find true associations out of a sea of possibilities?



Reported

- $H_1, \ldots, H_n$ hypotheses to be tested

$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{ false discoveries}}{\#\text{ discoveries}}\right] = \mathbb{E}\left[\frac{\#\text{ red squares}}{\#\text{ green}+\#\text{ red squares}}\right].$$

$$\frac{0}{0} = 0.$$

- FDR is the *fraction of irreproducibility*.
- Benjamini and Hochberg ('95) proposed a simple algorithm to control the FDR, *i.e.*, to control the reliability of the model.

The Knockoff filter

Why does the Lasso make errors?

- Feature correlated with noise.
- Feature correlated with a signal not included in the model.

Problem: How do we control the FDR of selected features $\left\{ i : \hat{\beta}_i \neq 0 \right\}$?

Knockoffs:

- For each feature X_j, create a fake variable $\tilde{X}_j$ (*knockoff*).
- X_j and $\tilde{X}_j$ are equally likely to be selected (when not in the model):
  - The covariance between knockoff features is the same as the covariance between the two original features.

$$\tilde{X}_j' \tilde{X}_k = X_j' X_k \qquad \text{for all } j, k \tag{22}$$

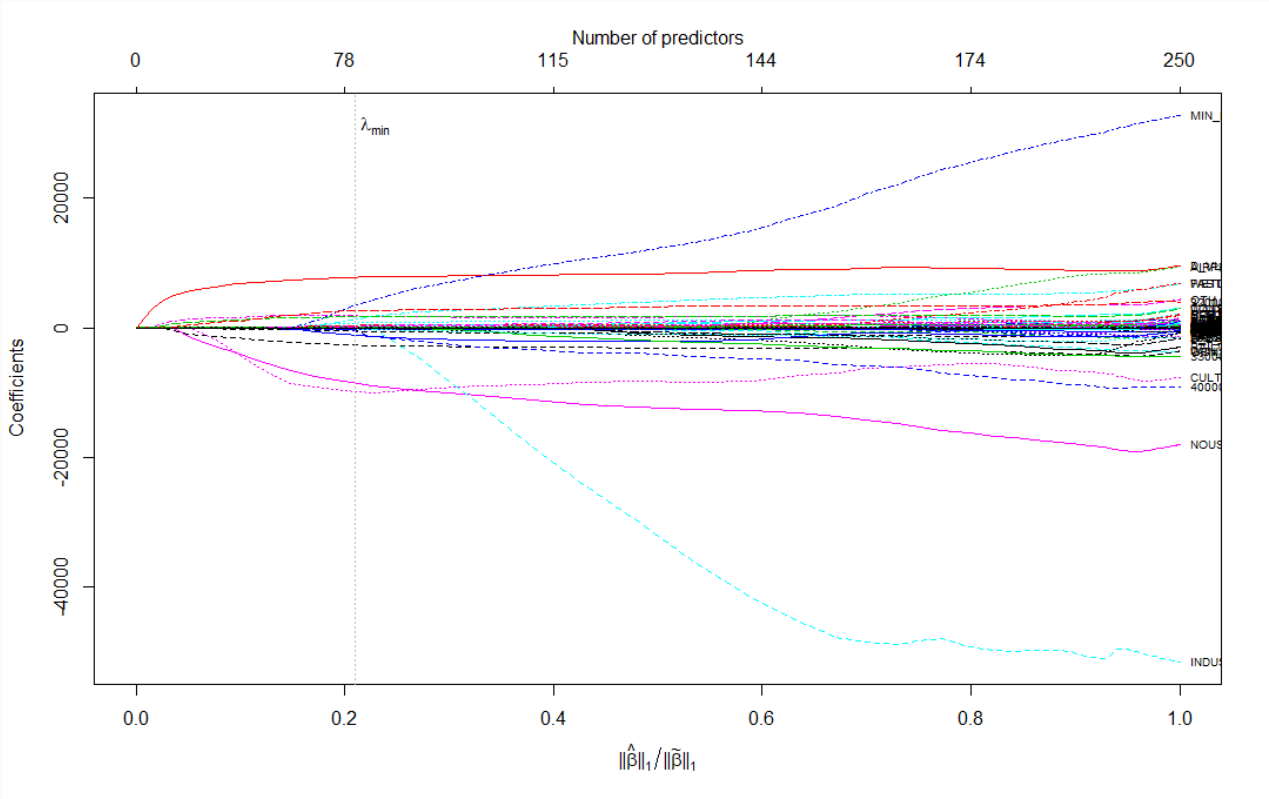
  - Knockoffs have the same covariance with a true feature that the two original true features have with each other.

$$\tilde{X}_j' X_k = X_j' X_k \qquad \text{for all } j \neq k \tag{23}$$

The Knockoff filter

Lasso:

$$\underbrace{\sum_{i=1}^{n}\left(y_i - \overbrace{\beta_0}^{\text{intercept}} - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{Residual Sum of Squares}} + \lambda \sum_{j=1}^{p}|\beta_j| \tag{24}$$



The Knockoff filter

- Lasso selects say, 52 original features and 26 knockoff features $\Leftrightarrow$ probably $\approx$ 26 false positives among the 52 original features.
- Continue along the Lasso path until the ratio between the knockoffs and the original features is below the target FDR –then stop.
- Report only the original features you have found.
- With this method, we can guarantee replicability.

Summary

- Scientists have been fitting data since the 19th century;
- Variable selection methods, such as the Lasso, are a fresh take on fitting problems;
- The knockoffs framework allows us to fit reliably;
- Need for automation in science and engineering.


Carl Friedrich Gauss (1777-1855)