UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Variable selection in genomics

— methods, challenges, and possibilities

Trial talk in defense of the degree of Ph. D.
Einar Holsbø
February 8th, 2019

UNIVERSITY OF TROMSØ

Variable selection in genomics

— methods, challenges, and possibilities

Variable selection in genomics

— methods, challenges, and possibilities

Variable selection

Identifying a suitable subset of variables
as relevant for your response and the modeling thereof

Variable selection

Identifying a suitable subset of variables
as relevant for your response and the modeling thereof

(identifying what is irrelevant and can be thrown away)

Variable selection

Maybe aka. “Data mining”

Identifying a suitable subset of variables
as relevant for your response and the modeling thereof

(identifying what is irrelevant and can be thrown away)

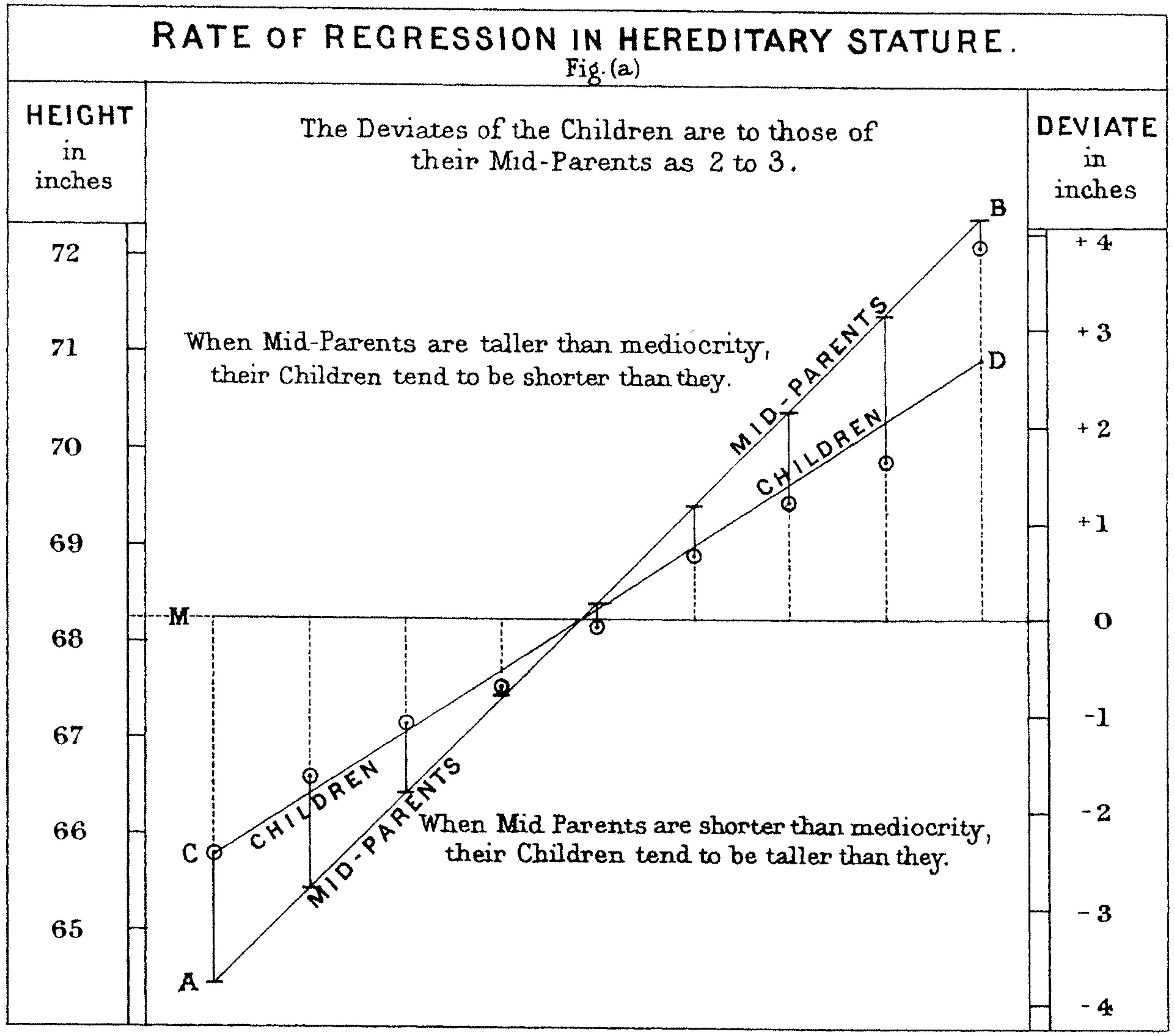
ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.

By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.



1886

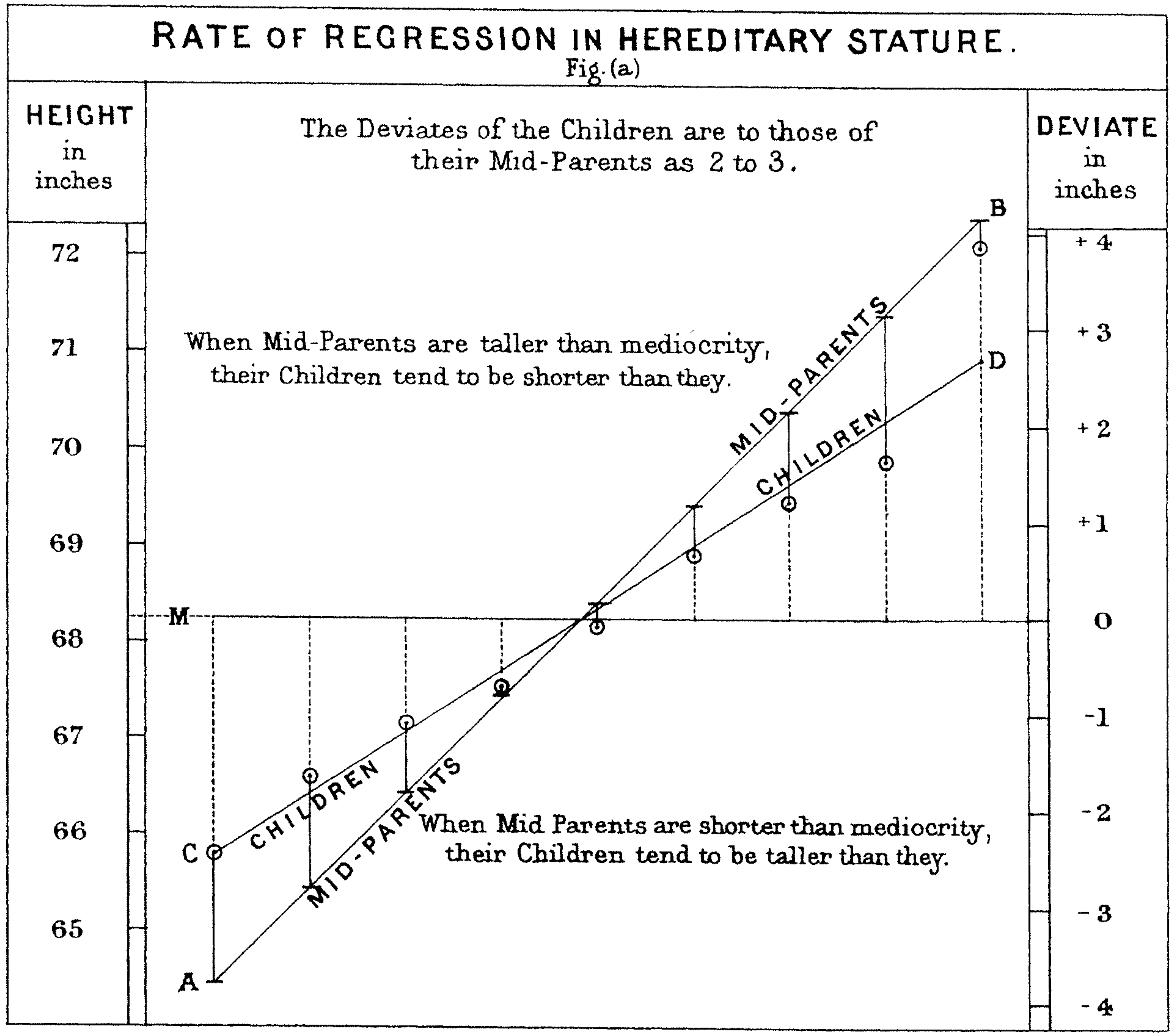
ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

[With Plates IX and X.]

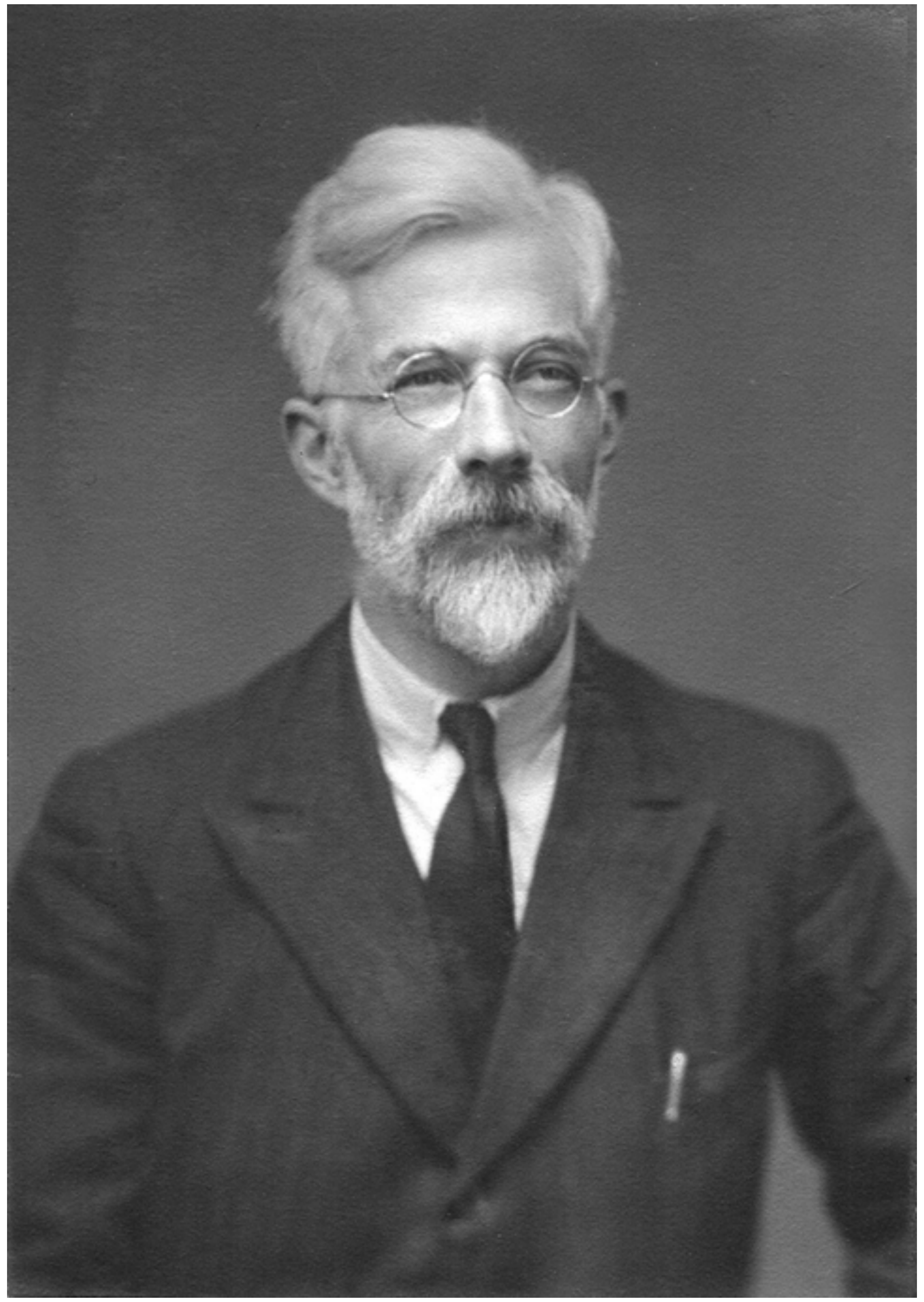
This memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.



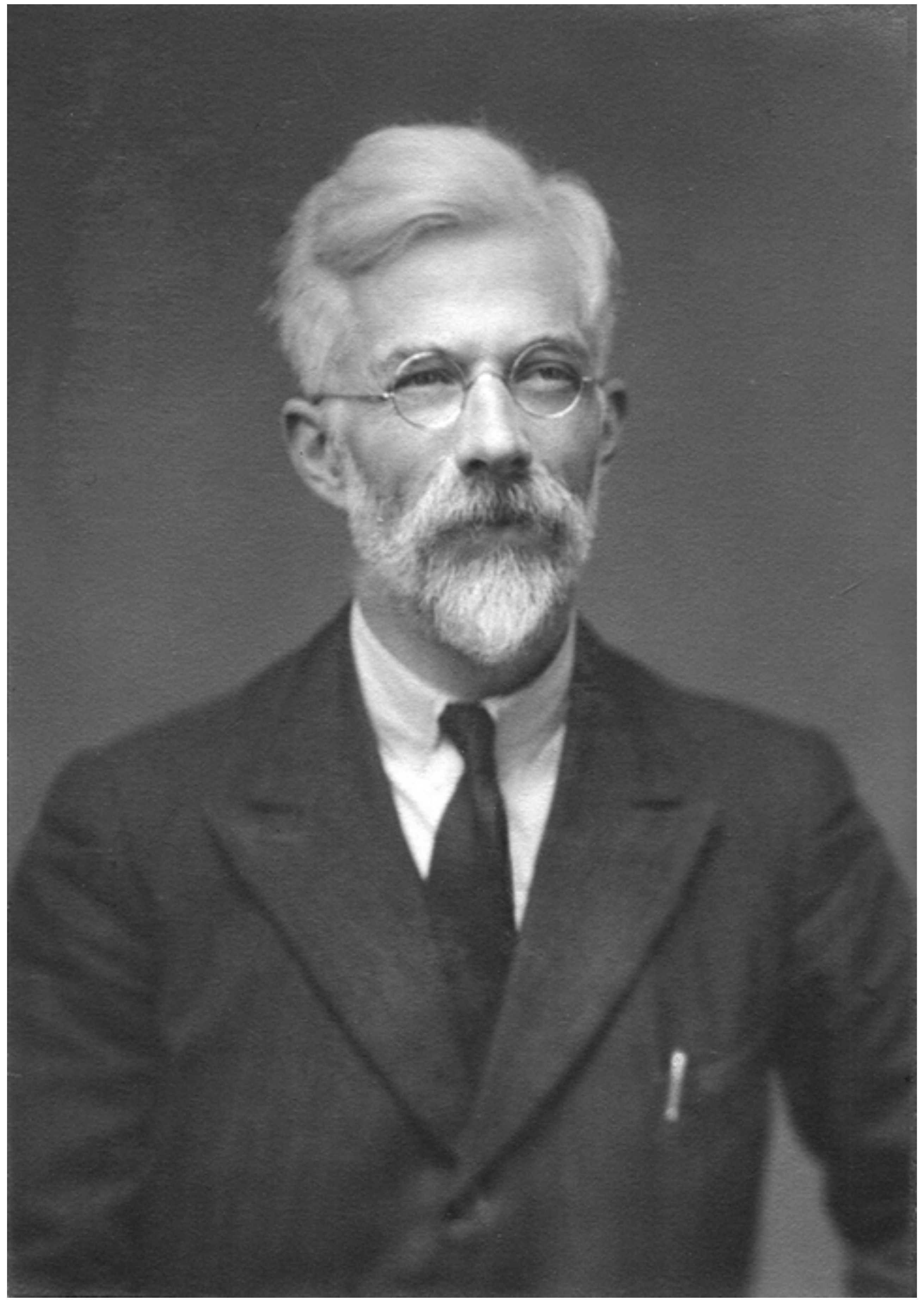
2 variables, 100s of observations

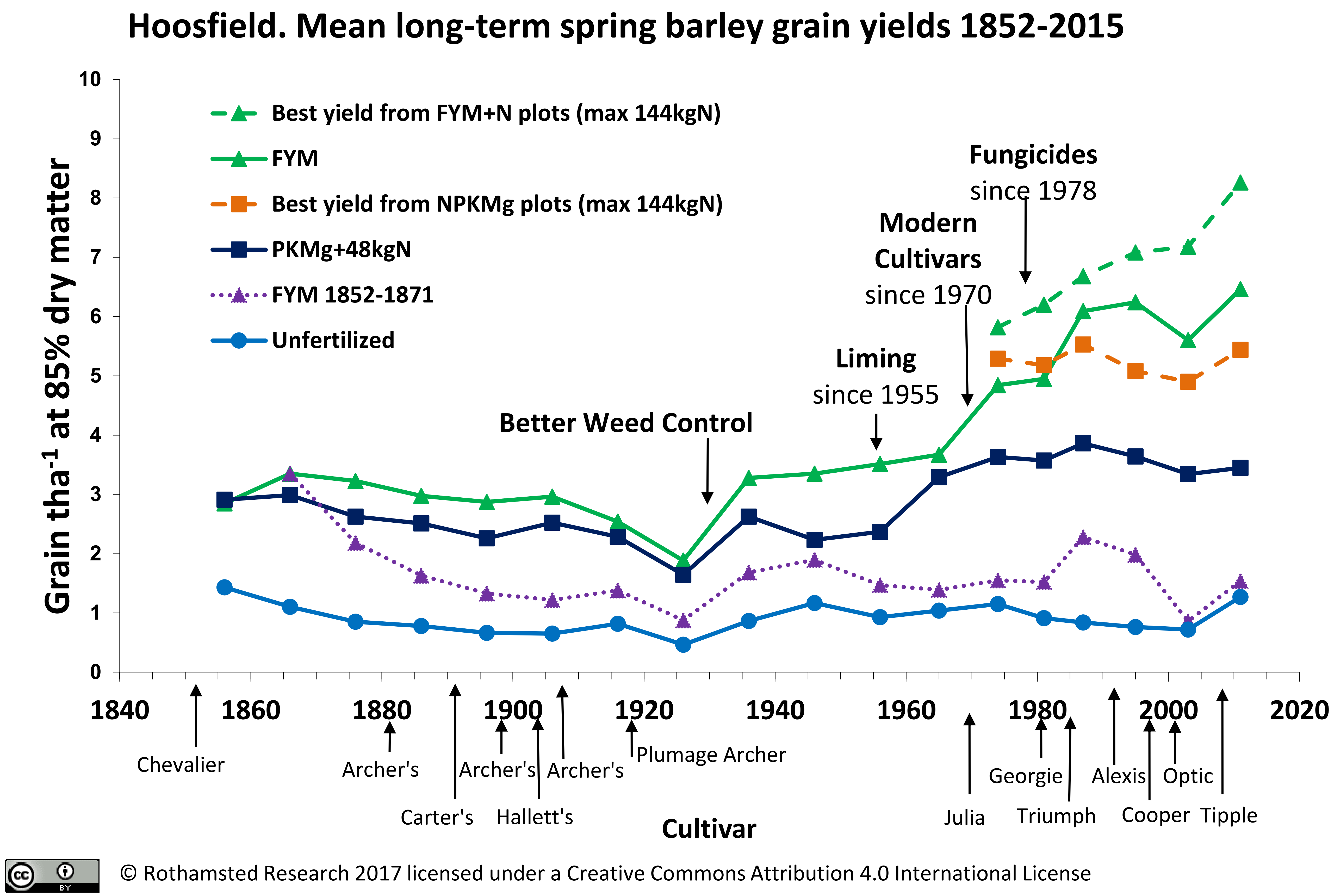
1886

Rothamsted experimental station

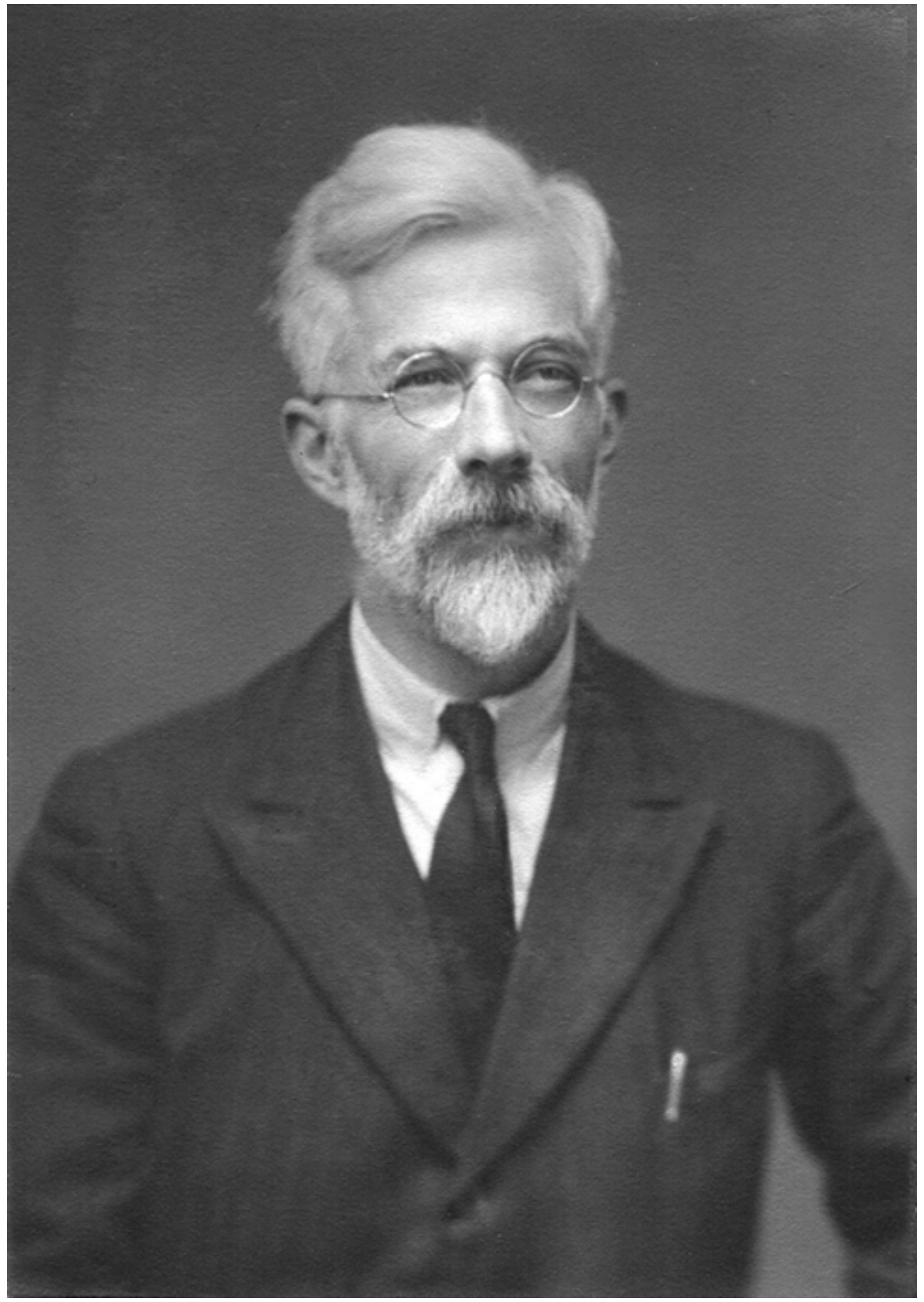


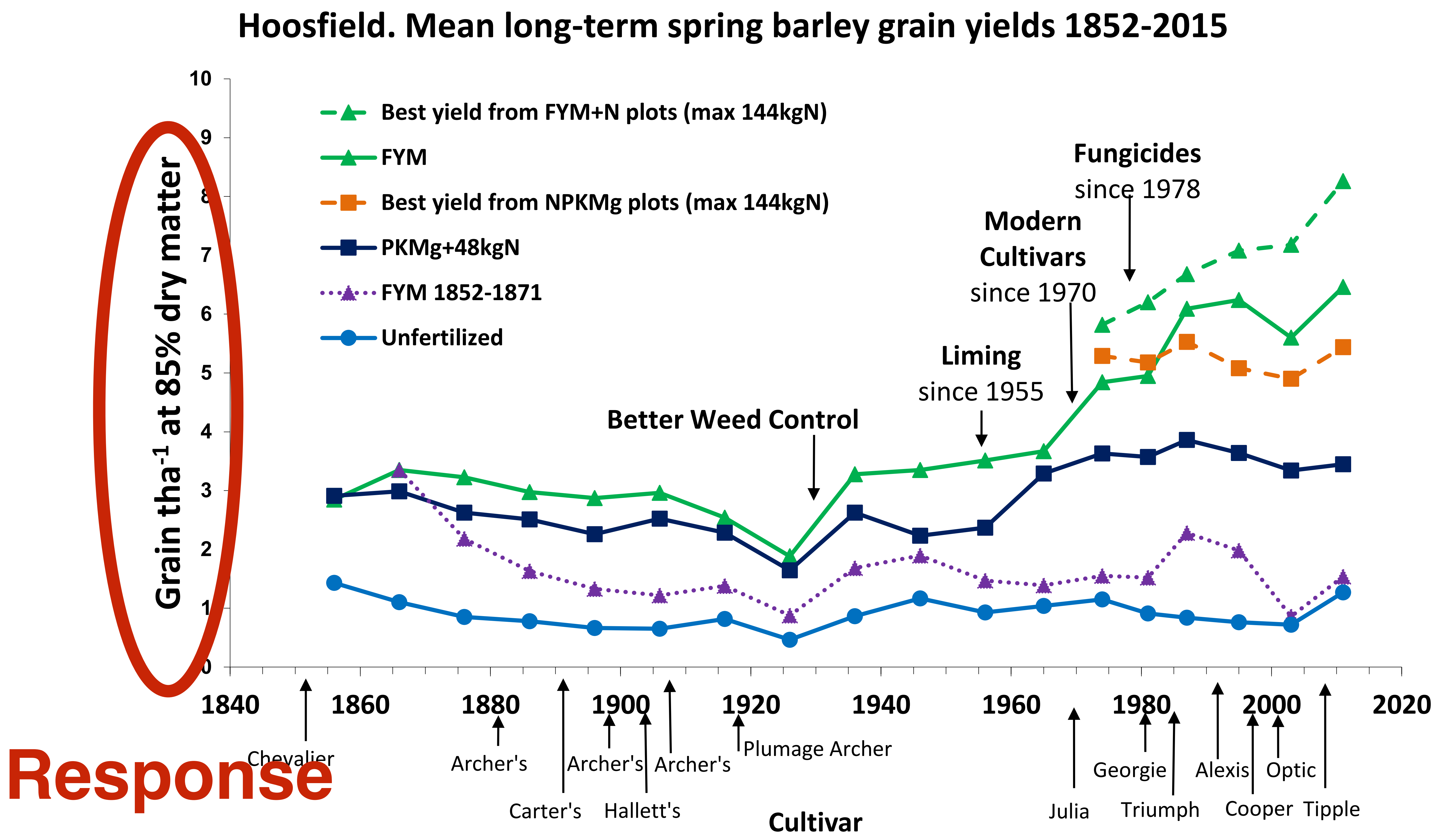
Rothamsted experimental station



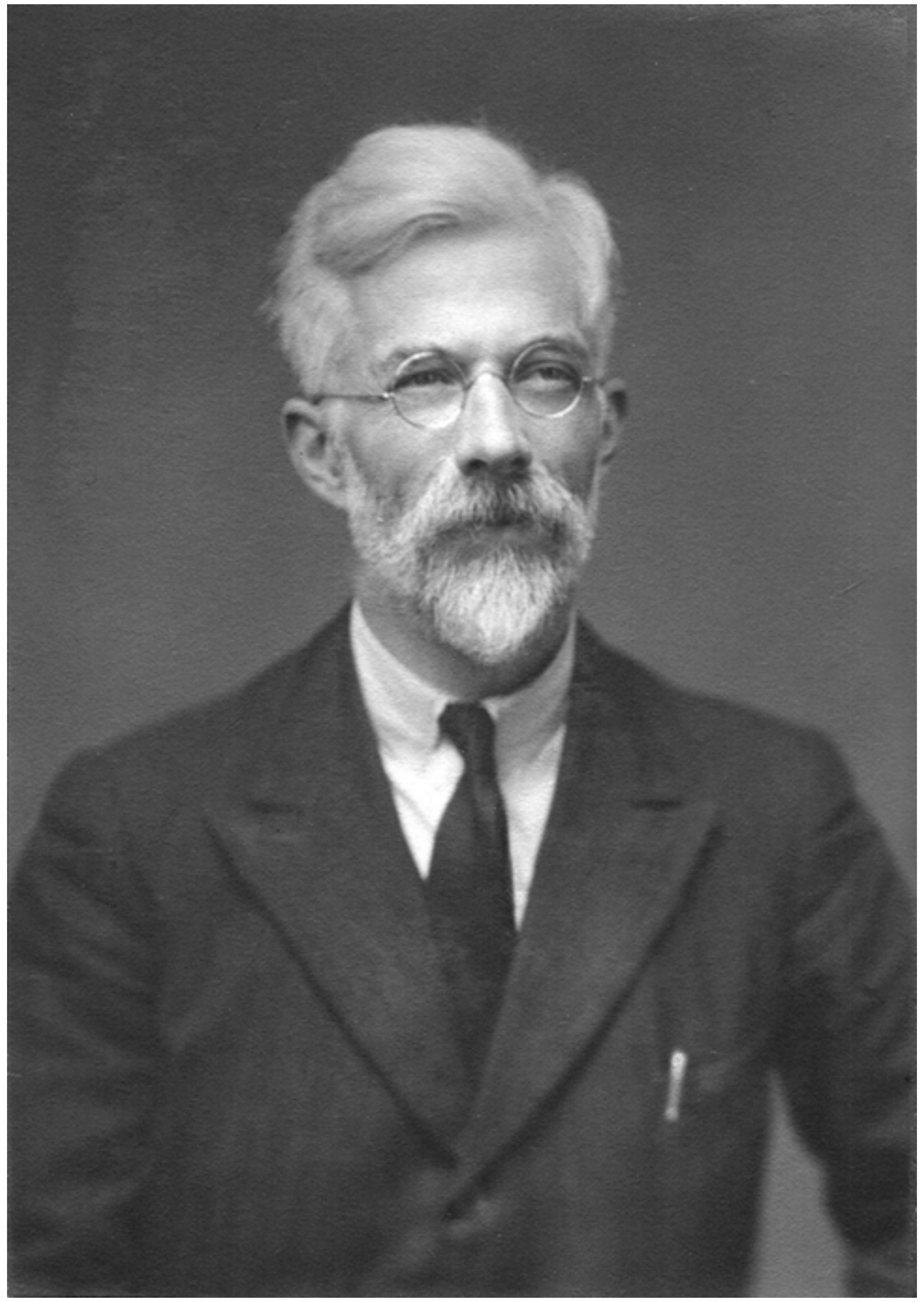


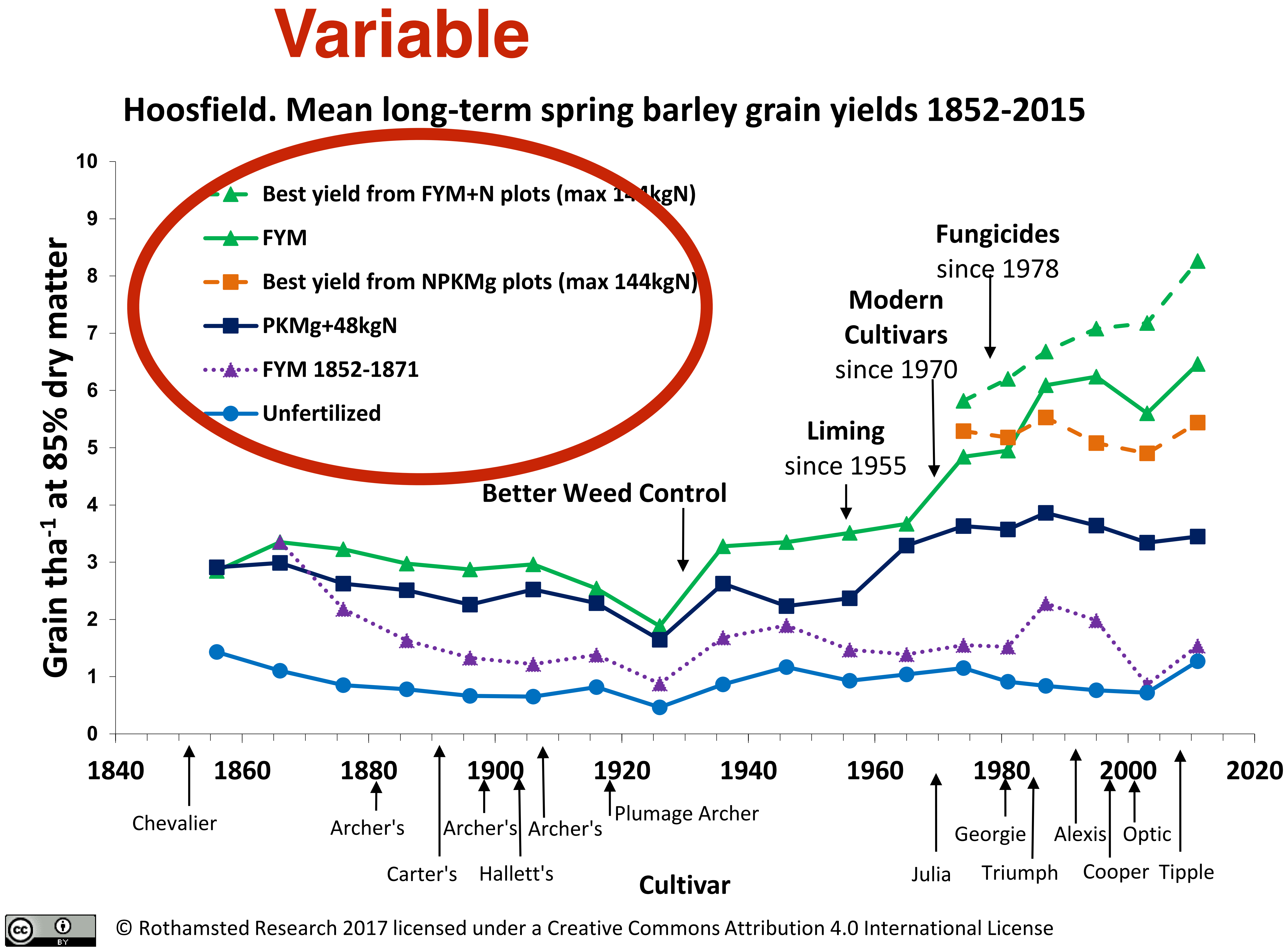
Rothamsted experimental station



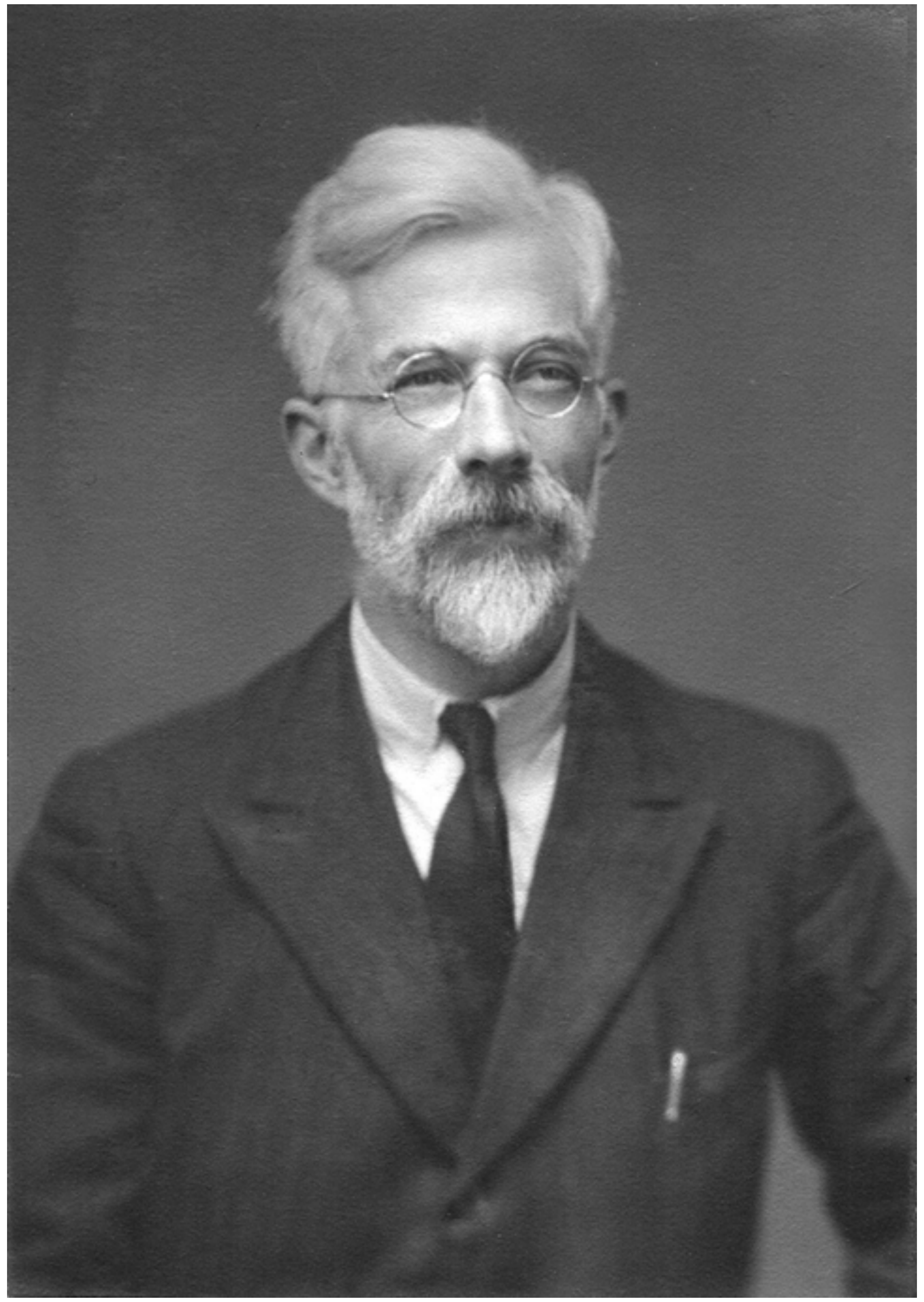


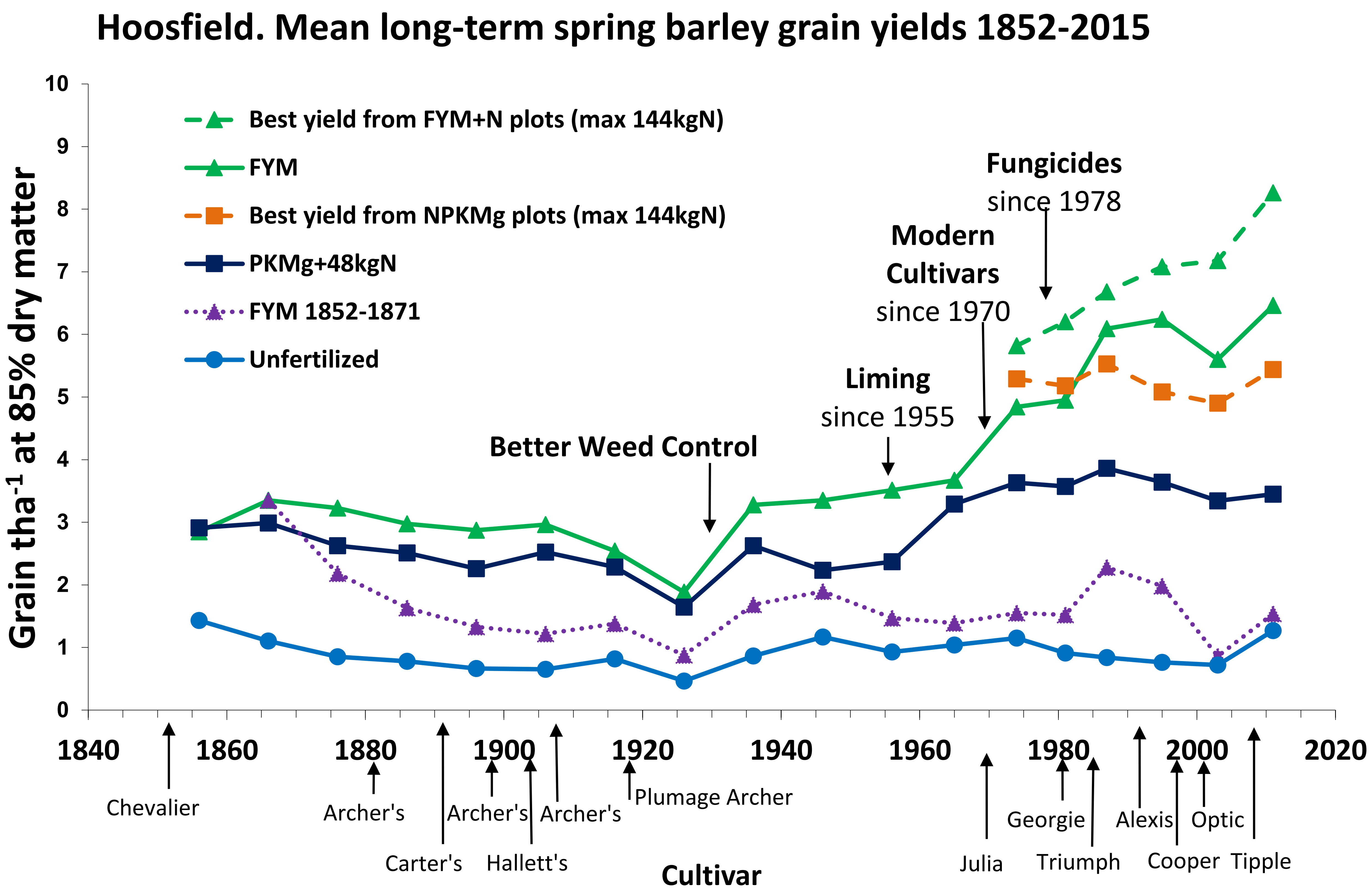
Rothamsted experimental station





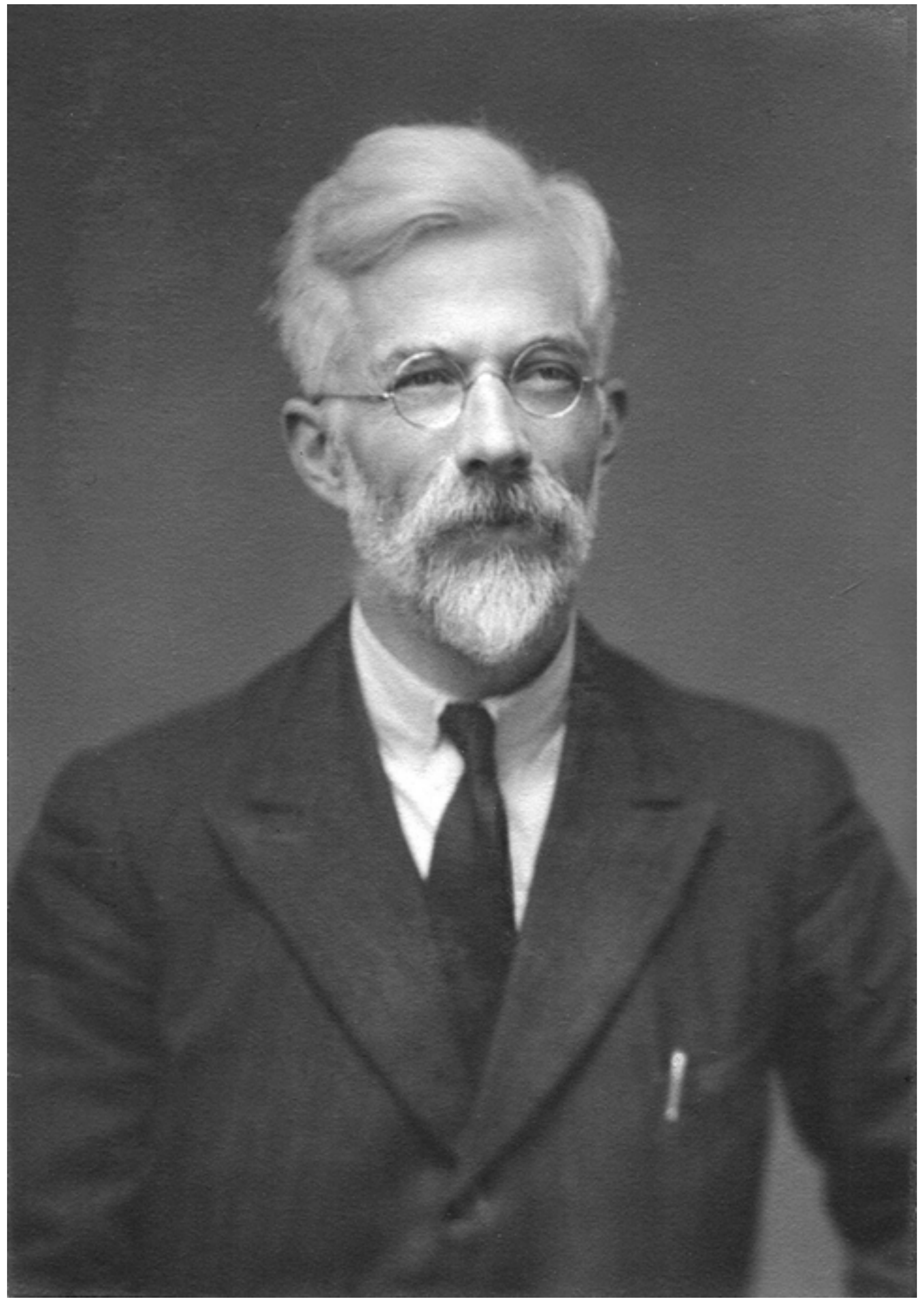
Rothamsted experimental station





6 variables, "enough" observations

Rothamsted experimental station



- Experimental design
- Small sample inference
- Comparison of multiple contrasts
- Hypothesis testing
- &c.

Why do you have irrelevant variables?

– Ronald Fisher, probably

Variable selection in genomics

— methods, challenges, and possibilities

Variable selection in genomics

— methods, challenges, and possibilities

~100 years later…

~100 years later…

Genome [n.] – the complete set of genes or genetic material present in a cell or organism.

~100 years later…

Genome [n.] – the complete set of genes or genetic material present in a cell or organism.

Genomics [n., pl.] – (treated as singular) the study of the structure, function, evolution, and mapping of genomes.

~100 years later…

-ome [suffix] – "all of them/it"
-omics [suffix] – the study of all the different things

(my interpretation)

Central dogma of molecular biology

Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258):561.

Central dogma of molecular biology

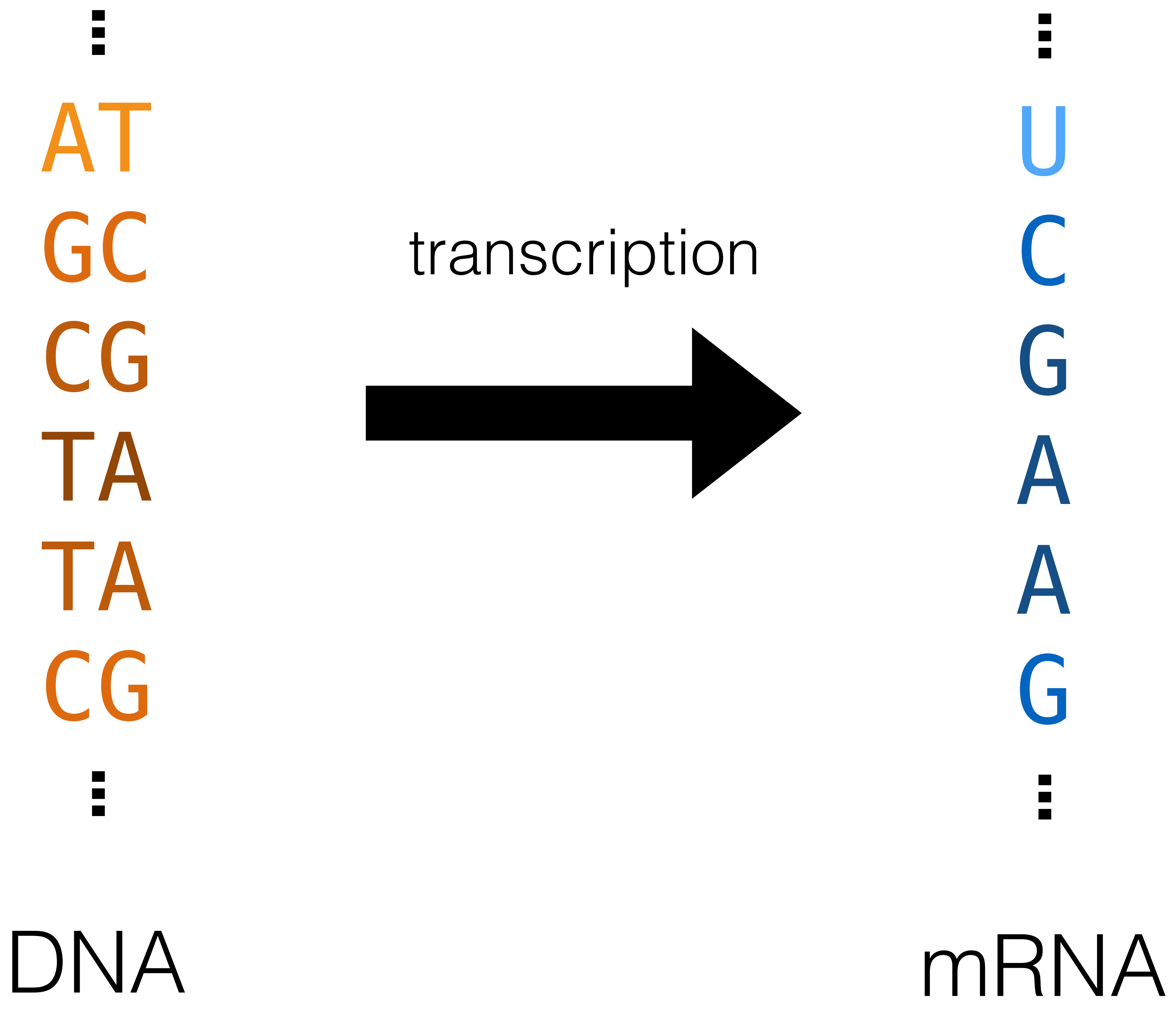
Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258):561.

⋮
AT
GC
CG
TA
TA
CG
⋮

DNA

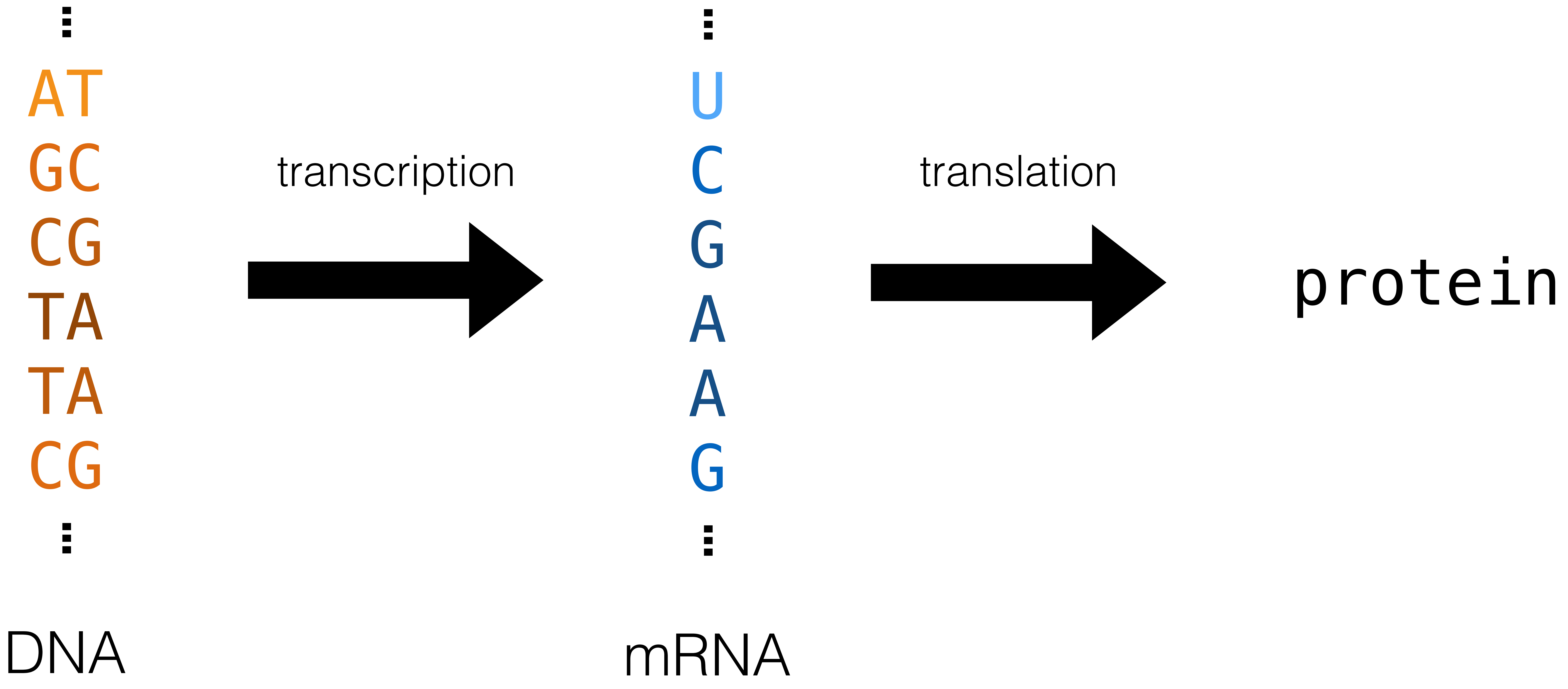
Central dogma of molecular biology

Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258):561.



Central dogma of molecular biology

Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258):561.



Central dogma of molecular biology

Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258):561.

⋮
U
C
G
A
A
G
⋮

mRNA

Central dogma of molecular biology

Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258):561.

⋮
U
C
G
A
A
G
⋮

mRNA

Transcriptomics [n.]

subfield to genomics to do with gene transcripts and the function of the genome.

So why variable selection?

So why variable selection?

- There are about 20.000 protein-coding genes

So why variable selection?

- There are about 20.000 protein-coding genes
- We are able to measure them all simultaneously

So why variable selection?

- There are about 20.000 protein-coding genes
- We are able to measure them all simultaneously
- Which ones are associated with a given process?

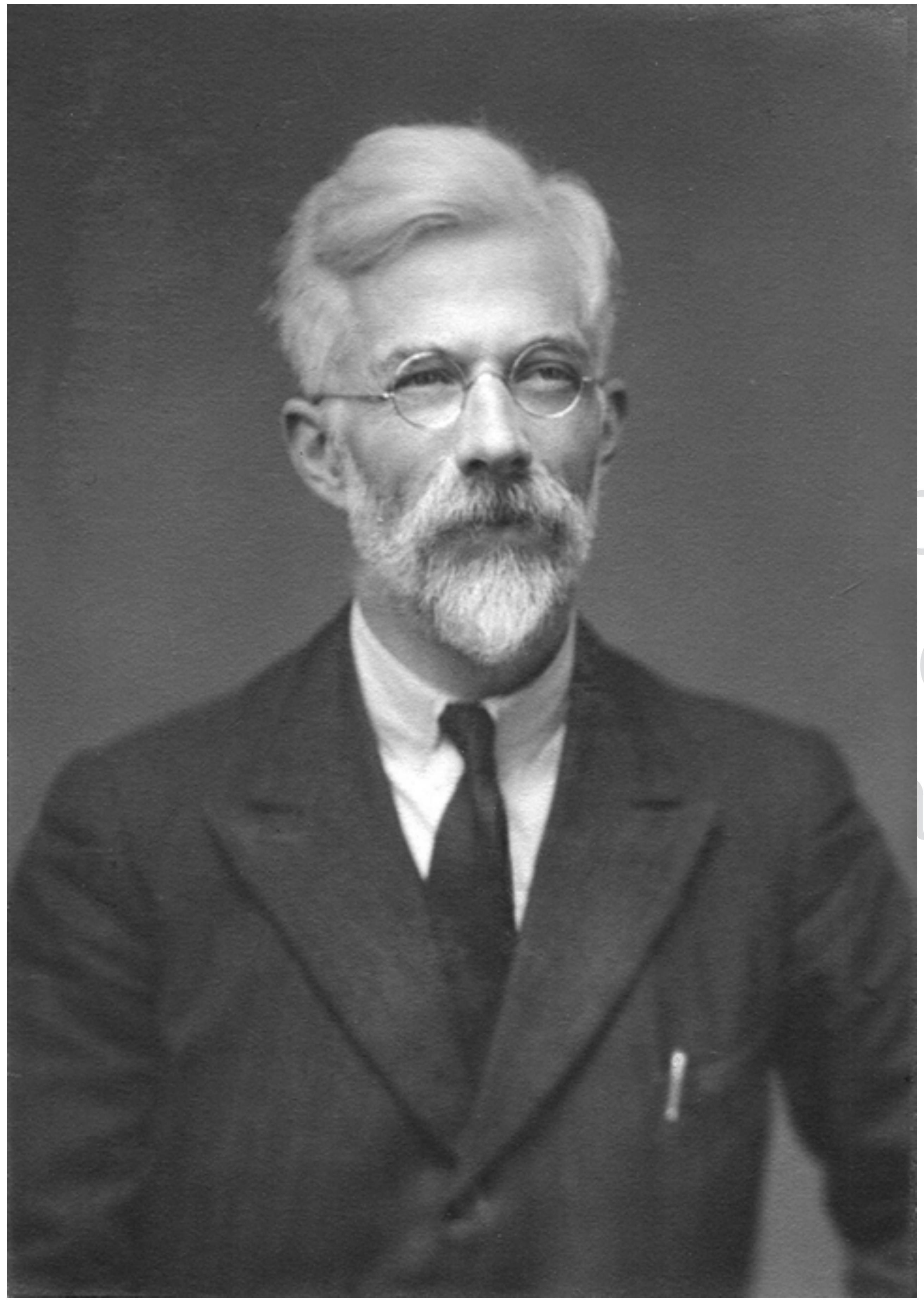
So why variable selection?

- There are about 20.000 protein-coding genes
- We are able to measure them all simultaneously
- **Which ones are associated with a given process?**

So why variable selection?

- There are about **20.000** protein-coding genes
- We are able to measure them all simultaneously
- Which ones are associated with disease?

So why variable selection?



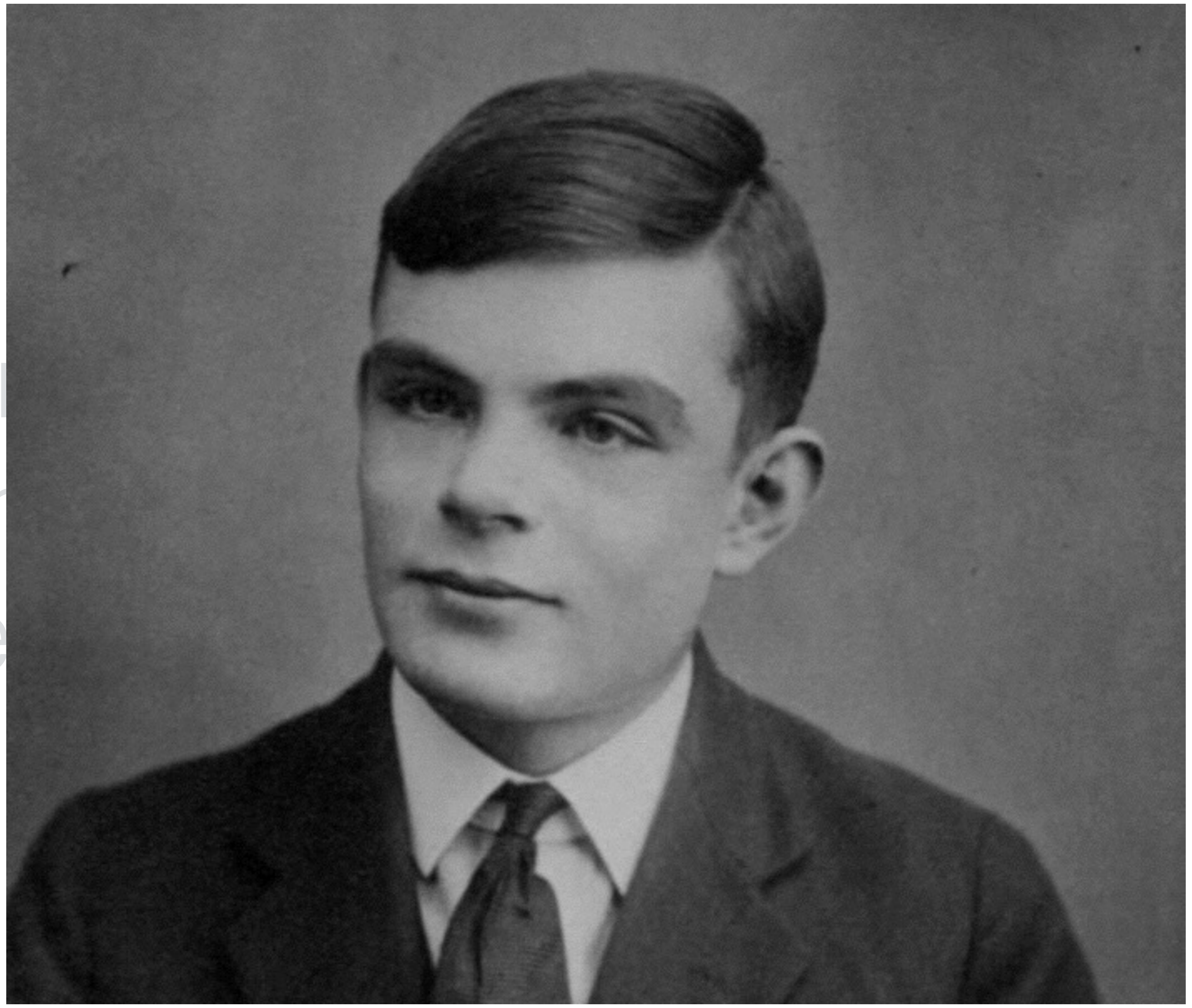
out **20.000** protein-coding genes
o measure them all simultaneously
re associated with disease?

So why variable selection?

ut **20.000** protein-coding genes
o measure them all simultaneously
re associated with disease?

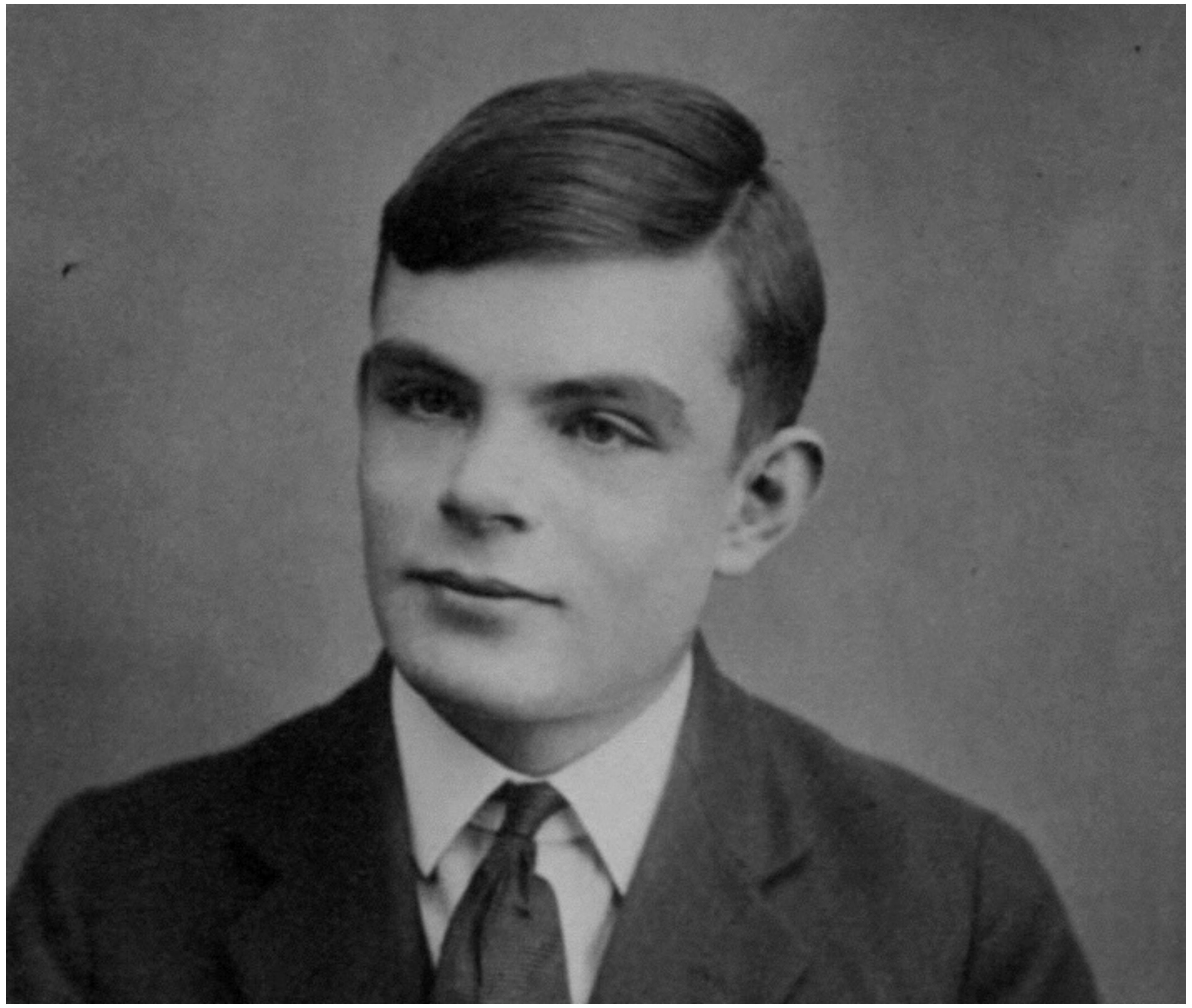
So why variable selection?

ut **20.000** p
o measure th
re associate



So why variable selection?

The computer as both problem and solution



Genomics light: 36 measurements, 20 obs


experiment_1
experiment_10
experiment_11
experiment_12
experiment_13
experiment_14
experiment_15
experiment_16
experiment_17
experiment_18
experiment_19
experiment_2
experiment_20
experiment_21
experiment_22
experiment_23
experiment_24
experiment_25
experiment_26
experiment_27
experiment_28
experiment_29
experiment_3
experiment_30
experiment_31
experiment_32
experiment_33
experiment_34
experiment_35
experiment_36
experiment_4
experiment_5
experiment_6
experiment_7
experiment_8
experiment_9
0.9
0.6
0.3
0.0
−10
−5
0
5
class
one
two

Danger of modeling noise, or “overfitting”

Want to find true signal, discard noise

Message

Message

(i) Old times: careful choice of variables; These times: measure everything

Message

(i) Old times: careful choice of variables; These times: measure everything

(ii) In genomics there are **many** variables to choose from.

Message

(i) Old times: careful choice of variables; These times: measure everything

(ii) In genomics there are **many** variables to choose from.

(iii) So little known that careful choice of variables is virtually impossible

Message

(i) Old times: careful choice of variables; These times: measure everything

(ii) In genomics there are **many** variables to choose from.

(iii) So little known that careful choice of variables is virtually impossible

(iv) Be careful

Variable selection in genomics

— methods, challenges, and possibilities

Variable selection in genomics

— methods, challenges, and possibilities

❧ Leaving genomics behind, mostly

Linear model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

Linear model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$$\text{response} = \sum \text{weights} \times \text{variables}$$

Linear model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$$\text{response} = \sum \text{weights} \times \text{variables}$$

outcome of interest

measurements

Linear model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$$\text{response} = \sum \text{weights} \times \text{variables}$$

outcome of interest (→ response)

measurements (→ variables)

find the βs/weights

A typical taxonomy

- Filters
- Wrappers
- Embedded methods

A typical taxonomy

- Filters
- Wrappers
- Embedded methods

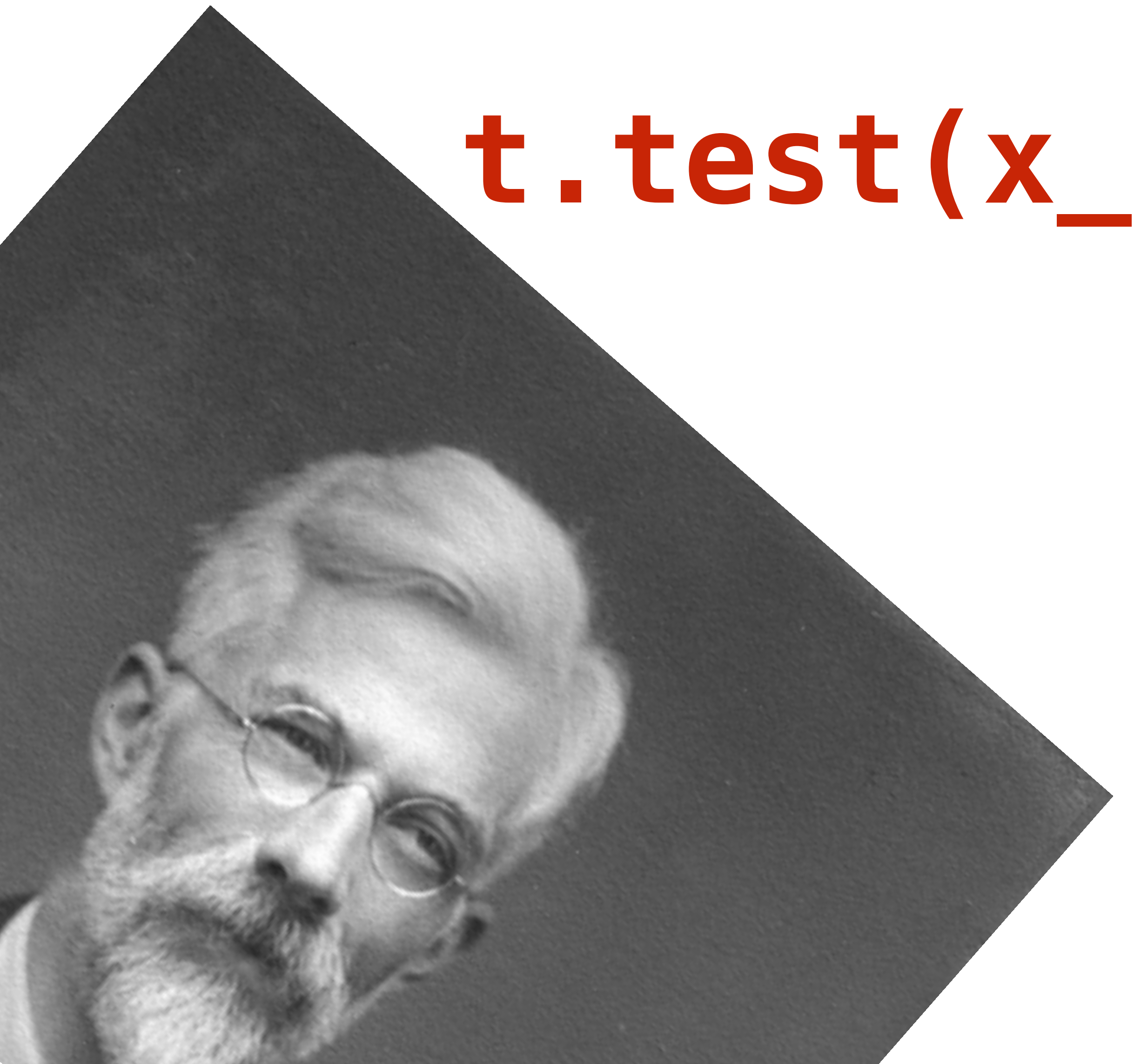
Filters

rank variables ➠ select "best" ones ➠ put in model

Filters

rank variables ➠ select “best” ones ➠ put in model

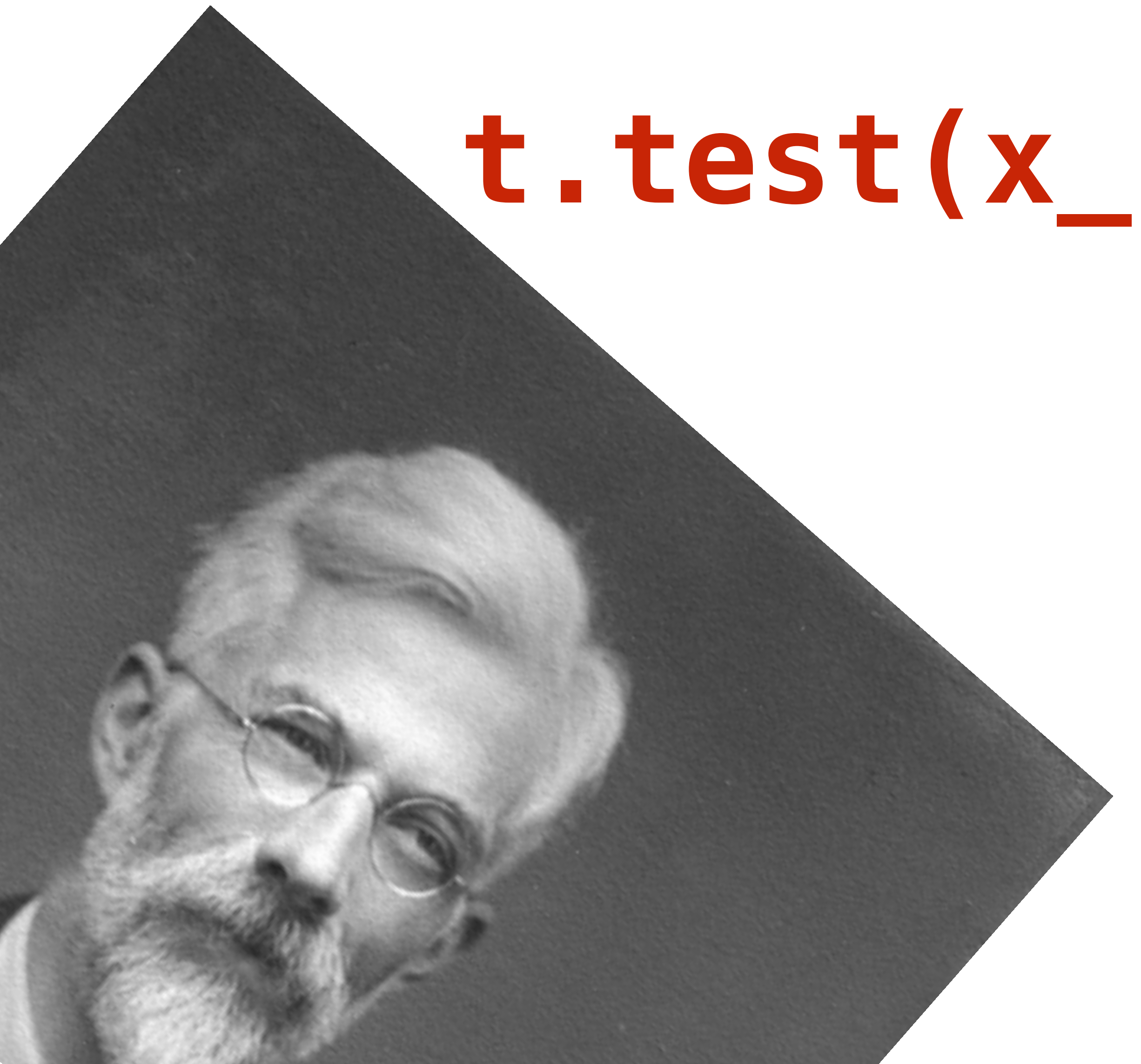
```
t.test(x_1, y) …
```



Filters

rank variables ➠ select “best” ones ➠ put in model

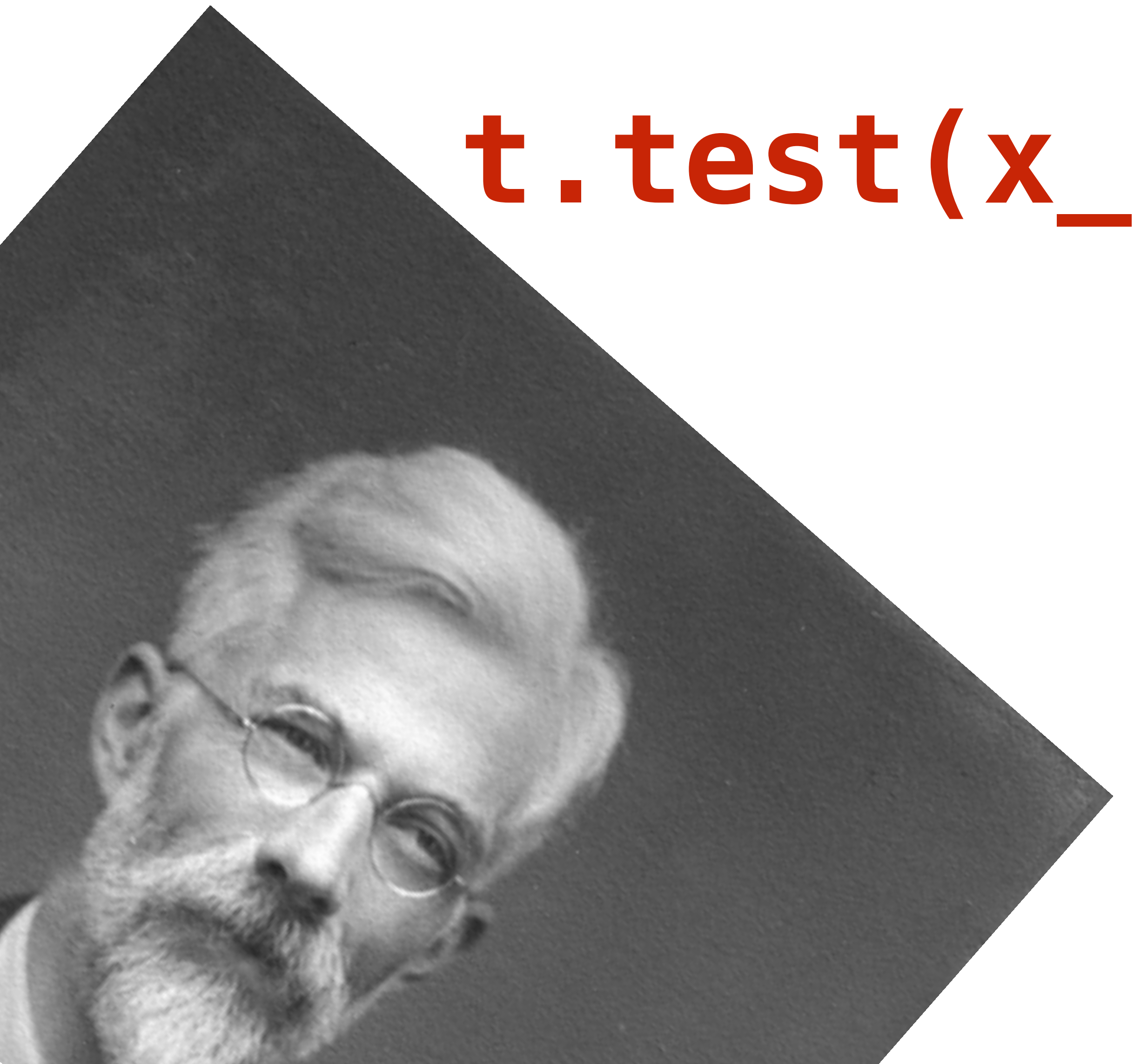
`t.test(x_1, y)` … **$p < .1$, top 10, etc.**



Filters

rank variables ➠ select "best" ones ➠ put in model

`t.test(x_1, y)` … **p < .1, top 10, etc.** **maybe linear**



Some transcriptome “filters”

- Significance analysis of microarrays (SAM) (also SAMSeq)
- Linear models for microarray/RNASeq data (LIMMA)
- K top-scoring pairs (K-tsp)

A typical taxonomy

- Filters
- Wrappers
- Embedded methods

Wrappers

evaluate fit

candidate subset → put in model

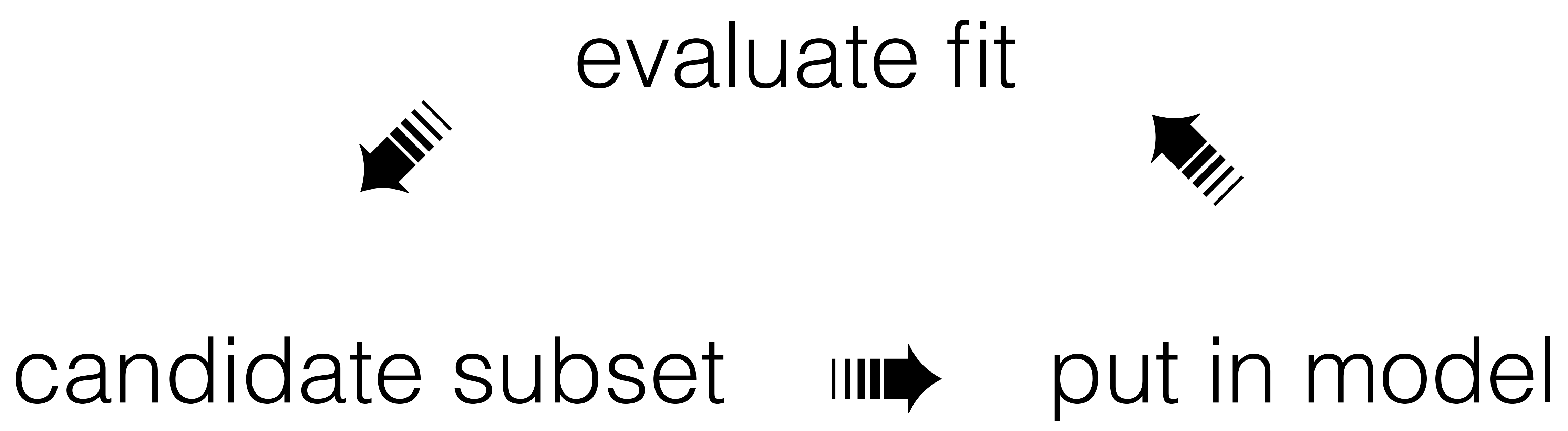
Wrappers

evaluate fit

candidate subset → put in model

stagewise,
simulated annealing, …

Wrappers



stagewise,
simulated annealing, …

maybe linear

Wrappers

metrics: R^2, empirical risk, …

evaluate fit

candidate subset ➠ put in model

stagewise,
simulated annealing, …

maybe linear

A typical taxonomy

- Filters
- Wrappers
- Embedded methods

Embedded

Combined model estimation and variable selection

Embedded

Combined model estimation and variable selection

optimize *model fit – model complexity*

N << d

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

N << d

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

N << d

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

∞ solutions

N << d

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

Standard rule-of-thumb calculations suggest
10–20 observations per parameter:

200 000 may be too few!

N << d

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

Idea: constrain the solution $\boldsymbol{\beta}$ to lie within a certain region

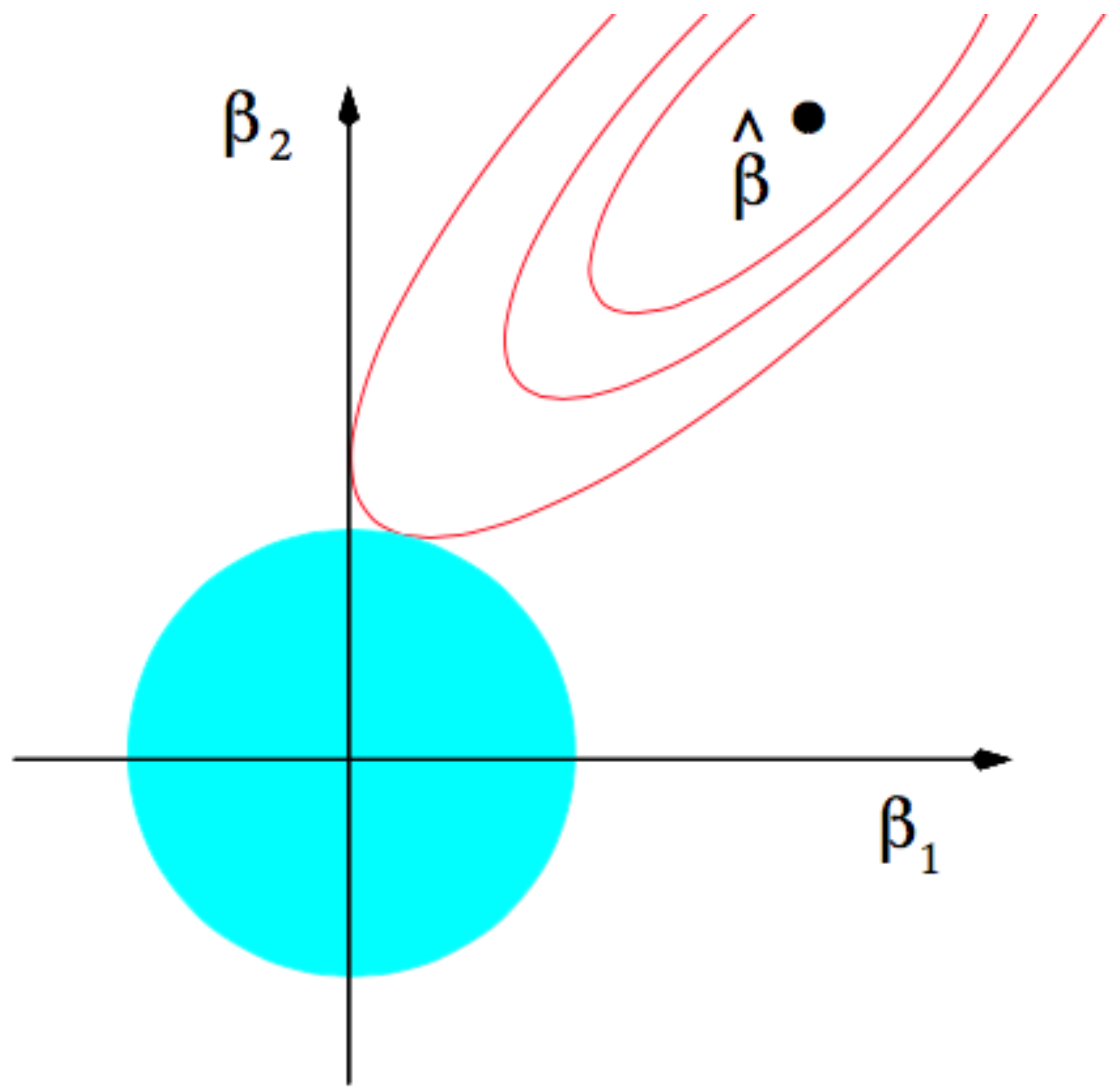
N << d

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

Idea: constrain the solution $\boldsymbol{\beta}$ to lie within a certain region

Eg find $\boldsymbol{\beta}$ s.t. $$\sum \beta_i^2 \leq t$$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

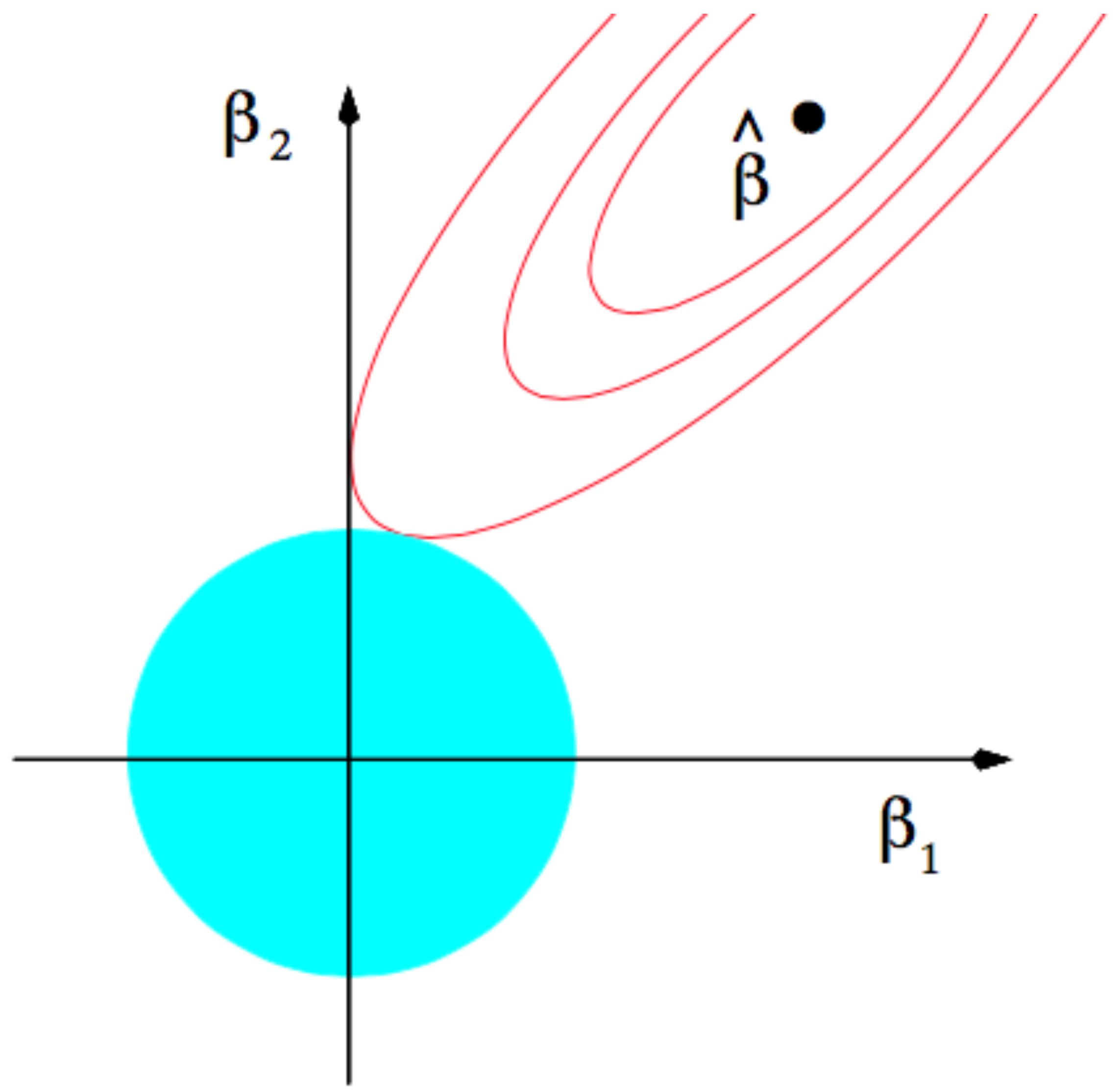


$$\sum \beta_i^2 \leq t$$

"ridge"

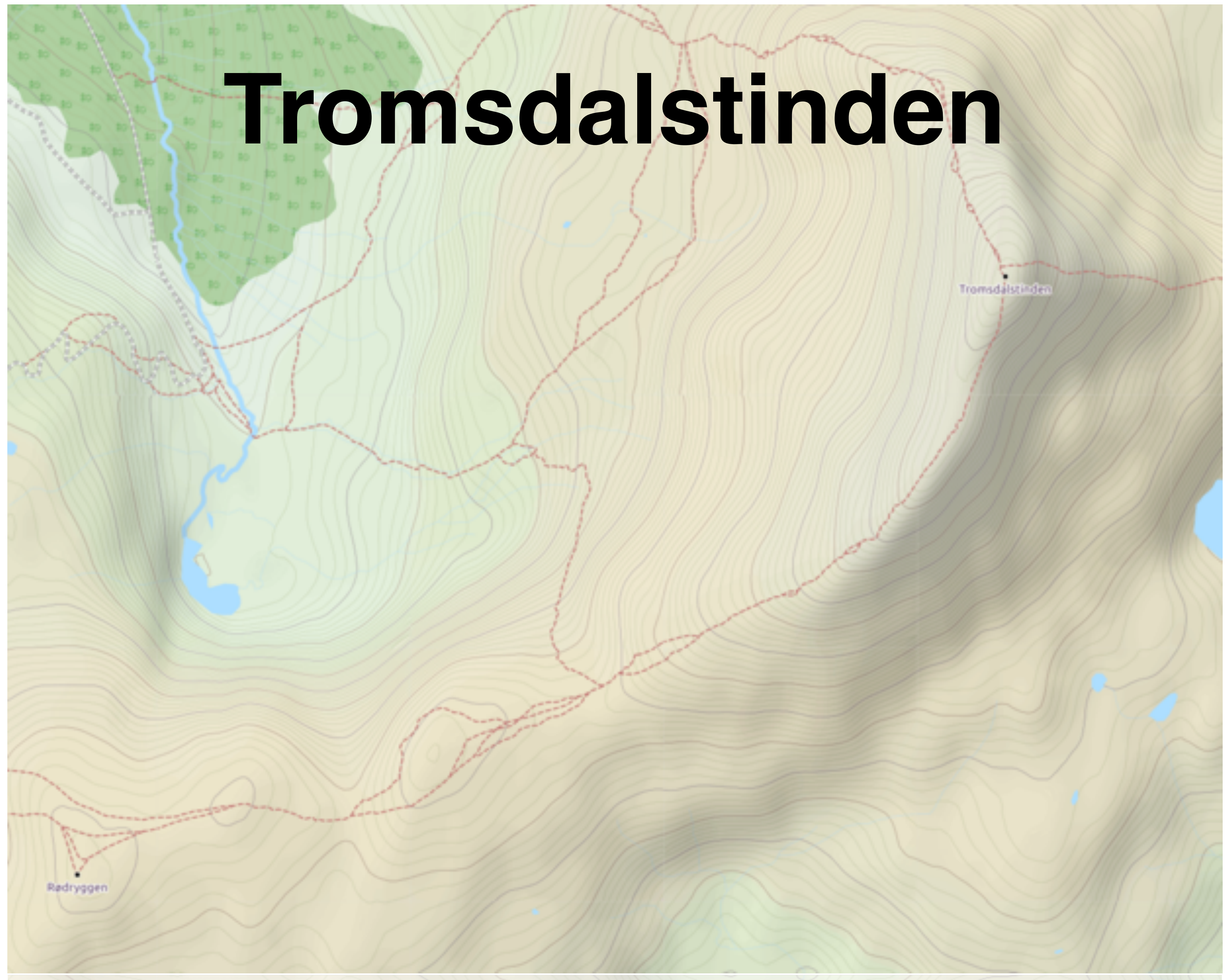
Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$



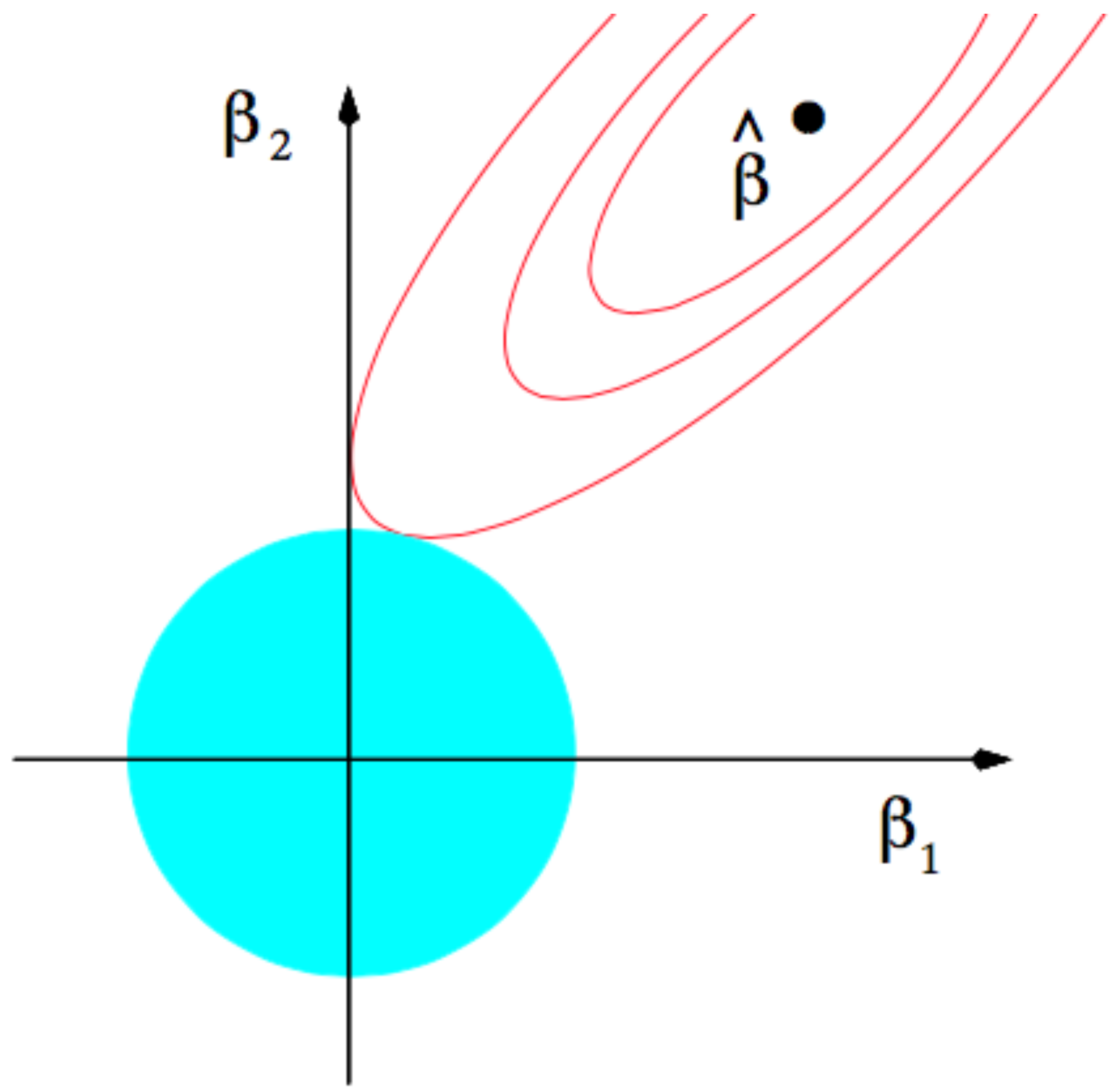
$$\sum \beta_i^2 \leq t$$

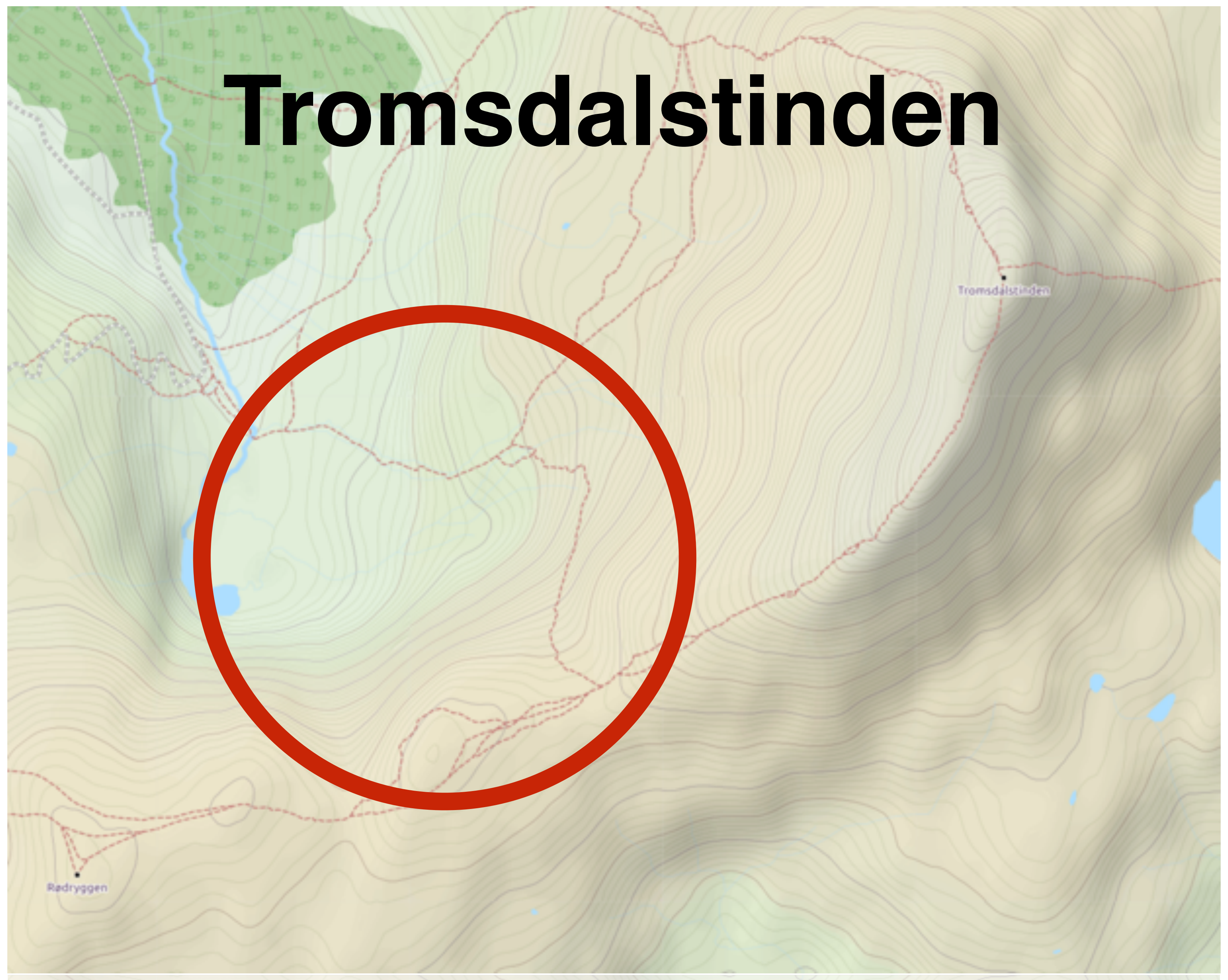
"ridge"



Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$



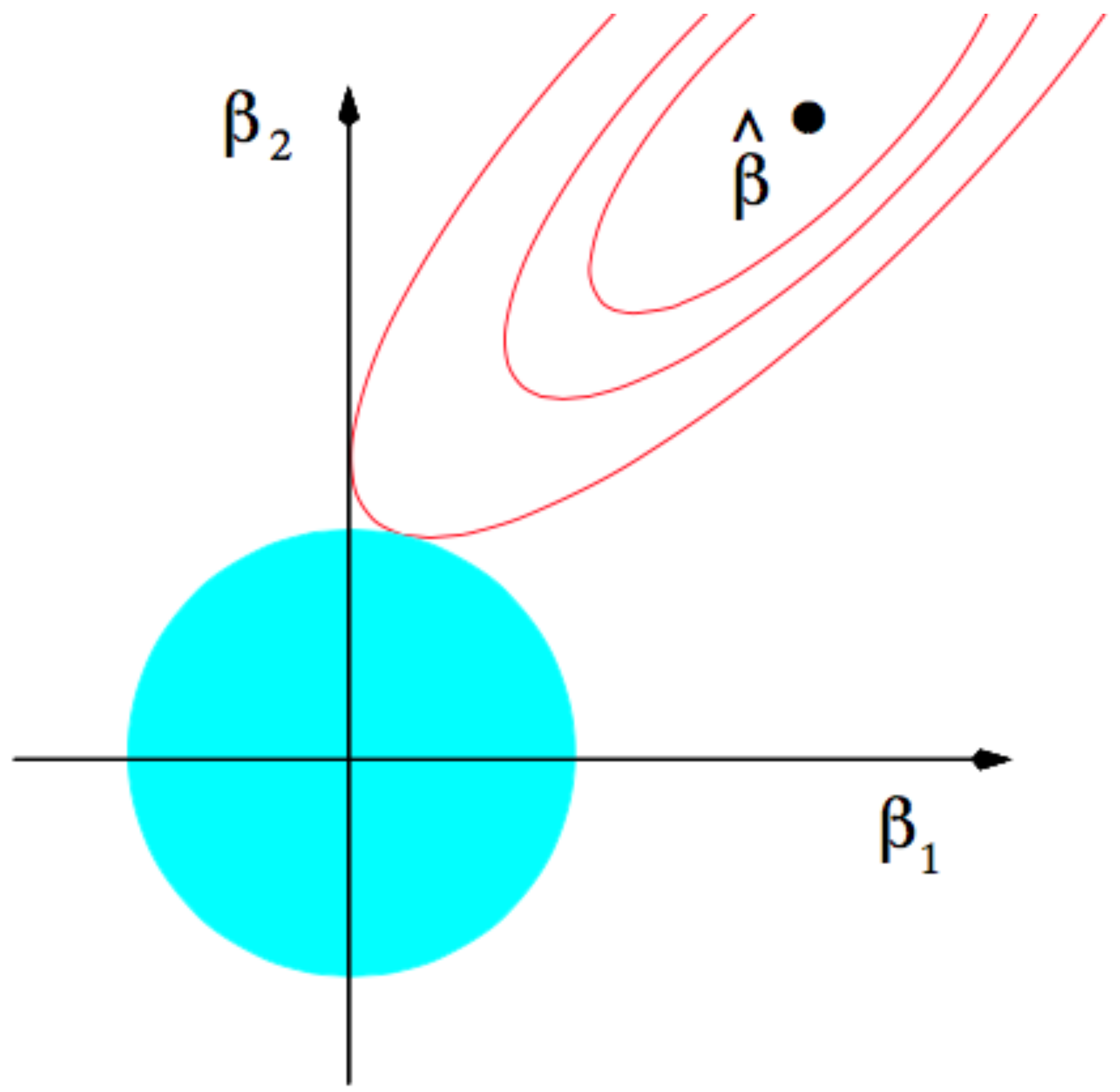


$$\sum \beta_i^2 \leq t$$

"ridge"

Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

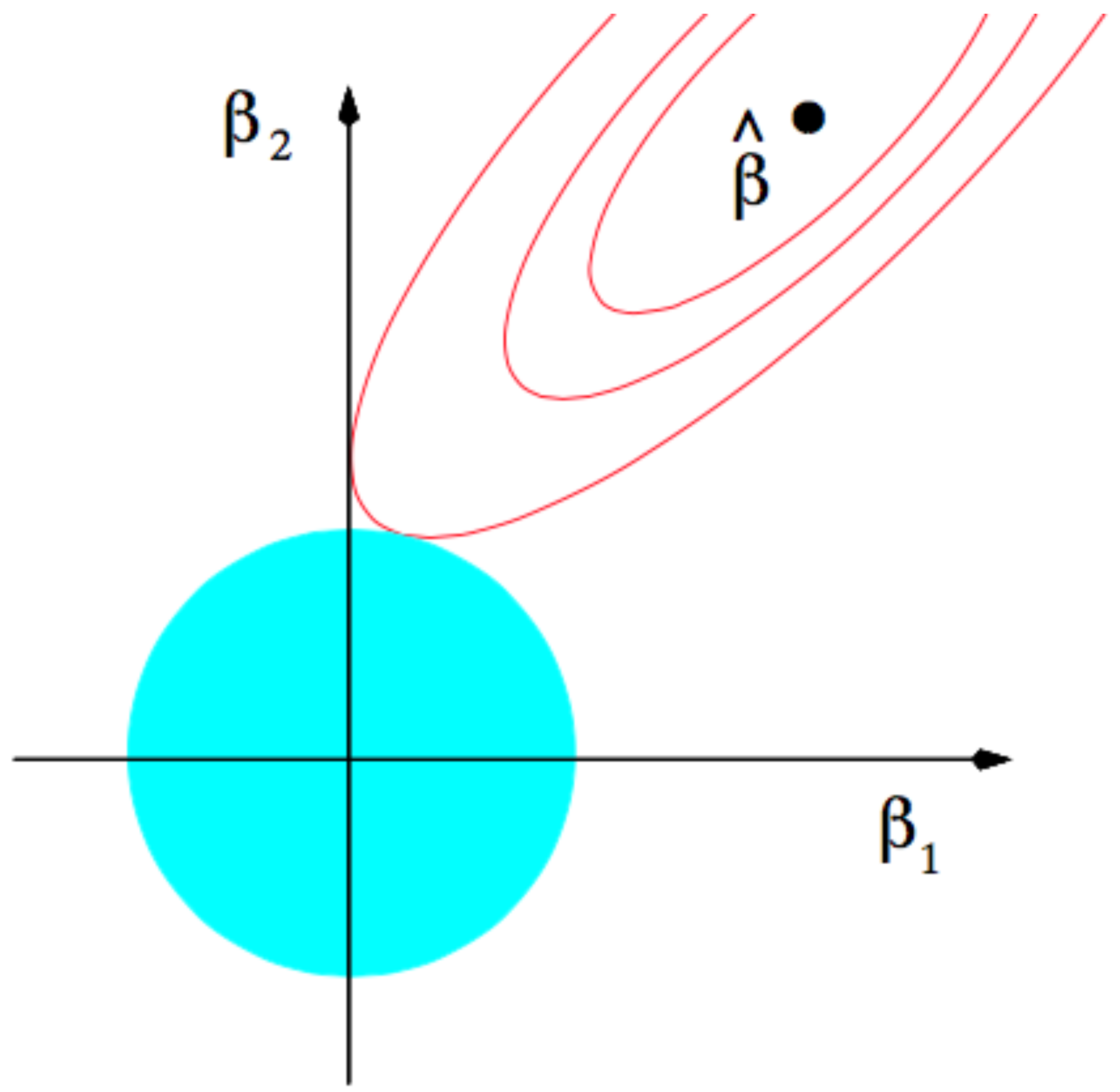


$$\sum \beta_i^2 \leq t$$

"ridge"

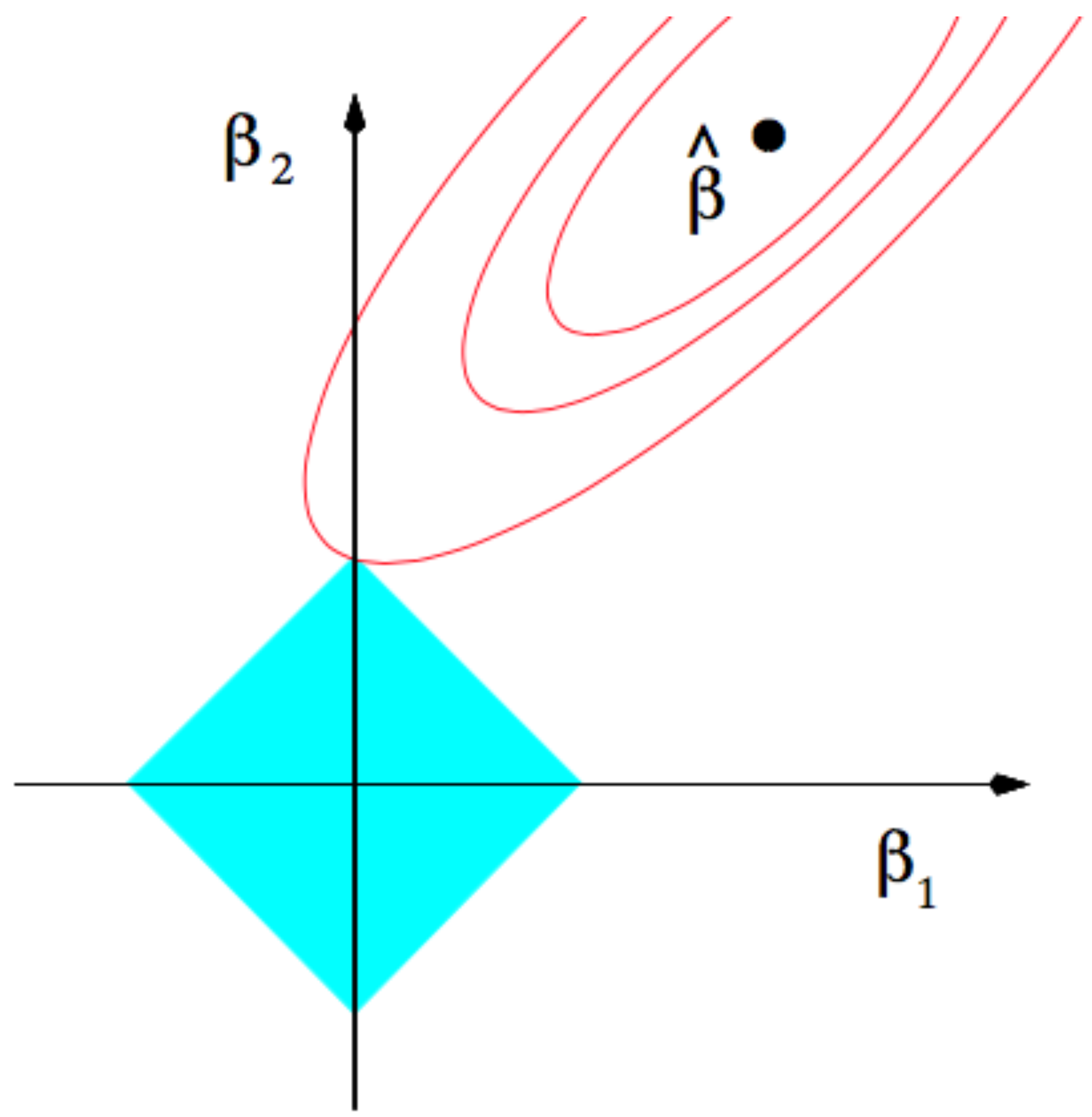
Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$



$$\sum \beta_i^2 \leq t$$

"ridge"



$$\sum |\beta_i| \leq t$$

"lasso"

Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

t usually a data-dependent decision

$$\sum \beta_i^2 \leq t$$

“ridge”

$$\sum |\beta_i| \leq t$$

“lasso”

Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

$$\sum |\beta_i| \leq t$$

"optimize *model fit – model complexity*"

$$\sum |\beta_i| \leq t$$

measure of model complexity

"optimize *model fit – model complexity*"

$$\sum |\beta_i| \leq t$$

Figure from Christophe Giraud “Introduction to High-Dimensional Statistics”

$$\sum |\beta_i| \leq t$$

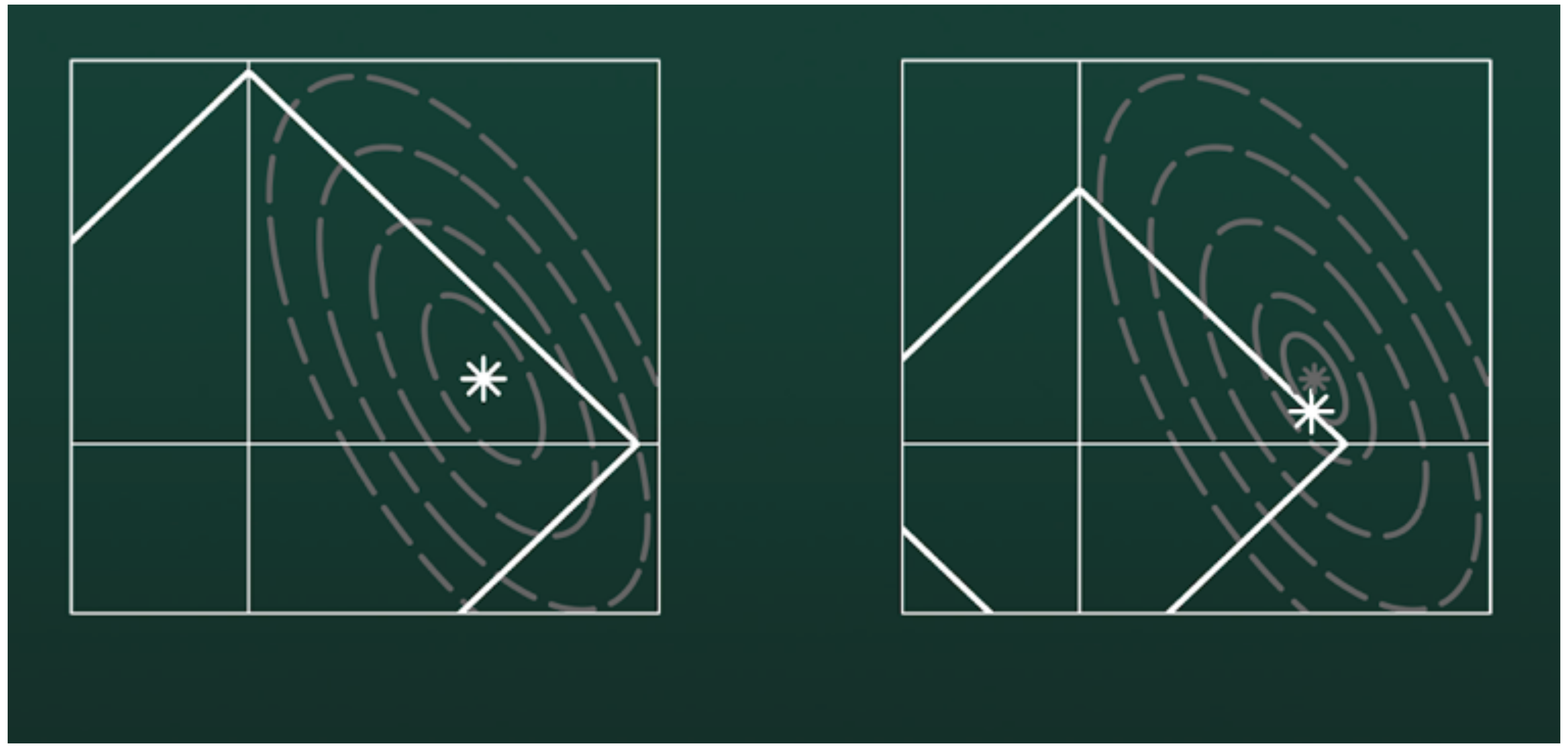


Figure from Christophe Giraud “Introduction to High-Dimensional Statistics”

$$\sum |\beta_i| \leq t$$

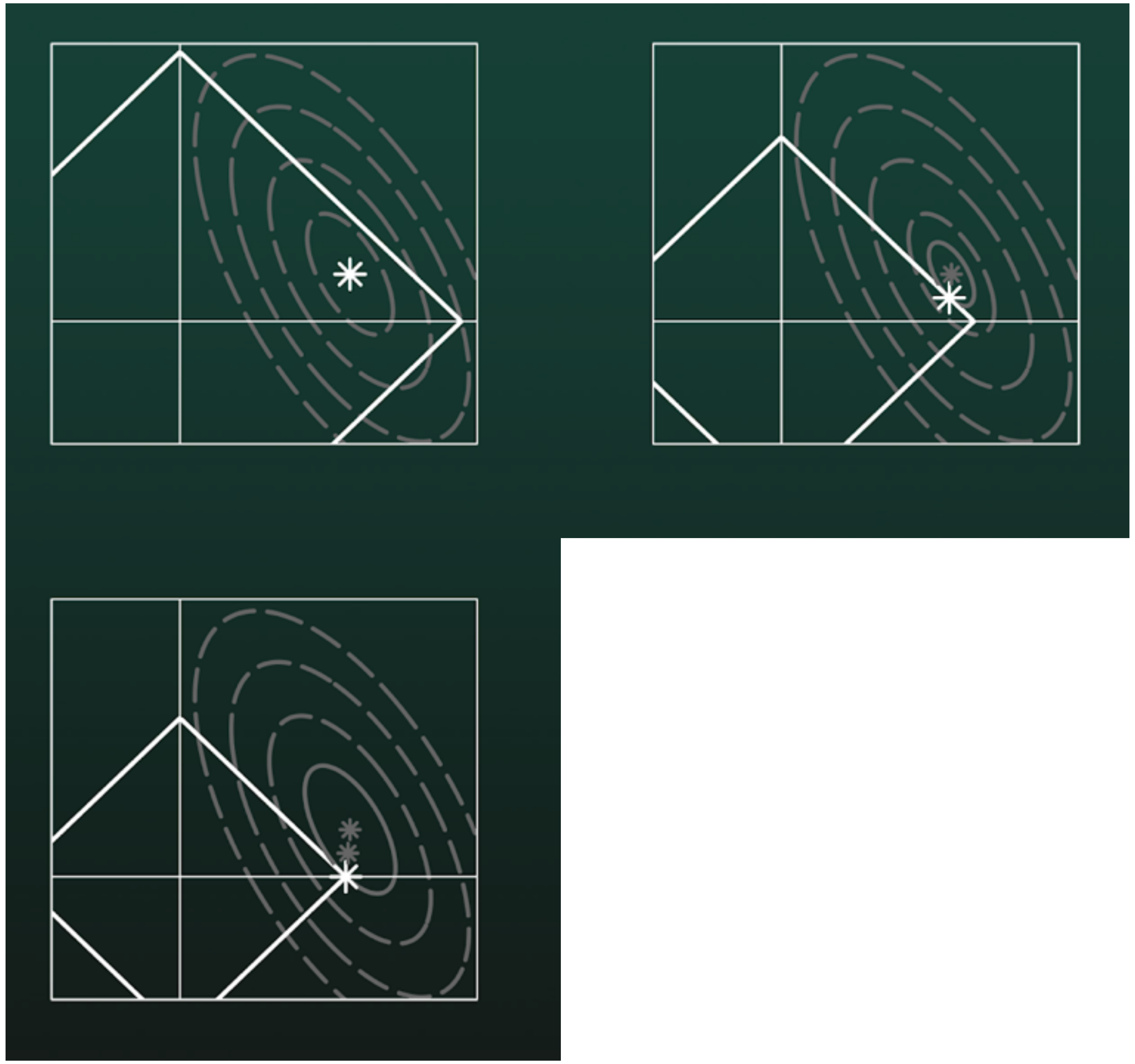


Figure from Christophe Giraud “Introduction to High-Dimensional Statistics”

$$\sum |\beta_i| \leq t$$

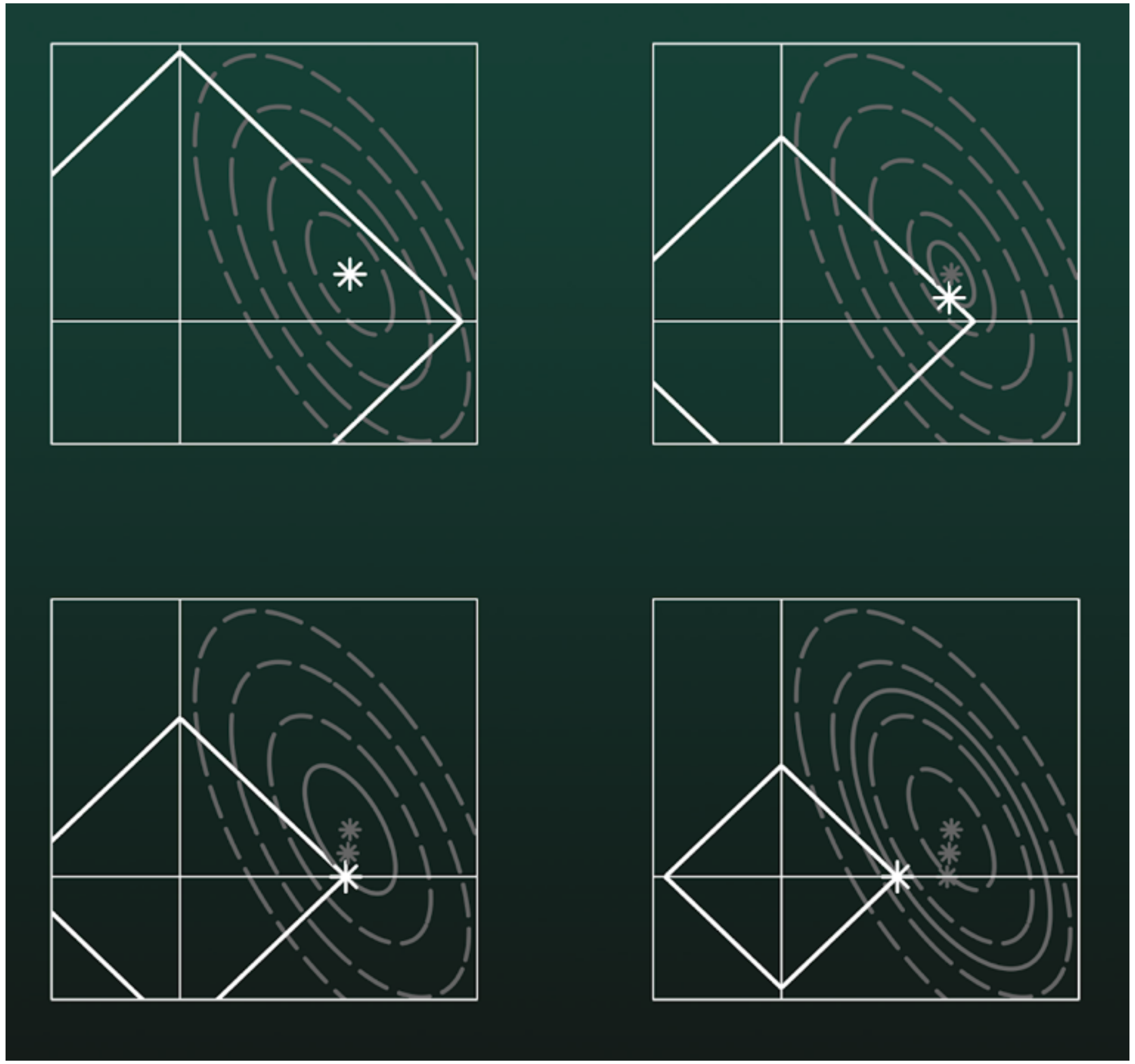


Figure from Christophe Giraud “Introduction to High-Dimensional Statistics”

End-result: a model with many coefficients = 0

$$\sum |\beta_i| \leq t$$

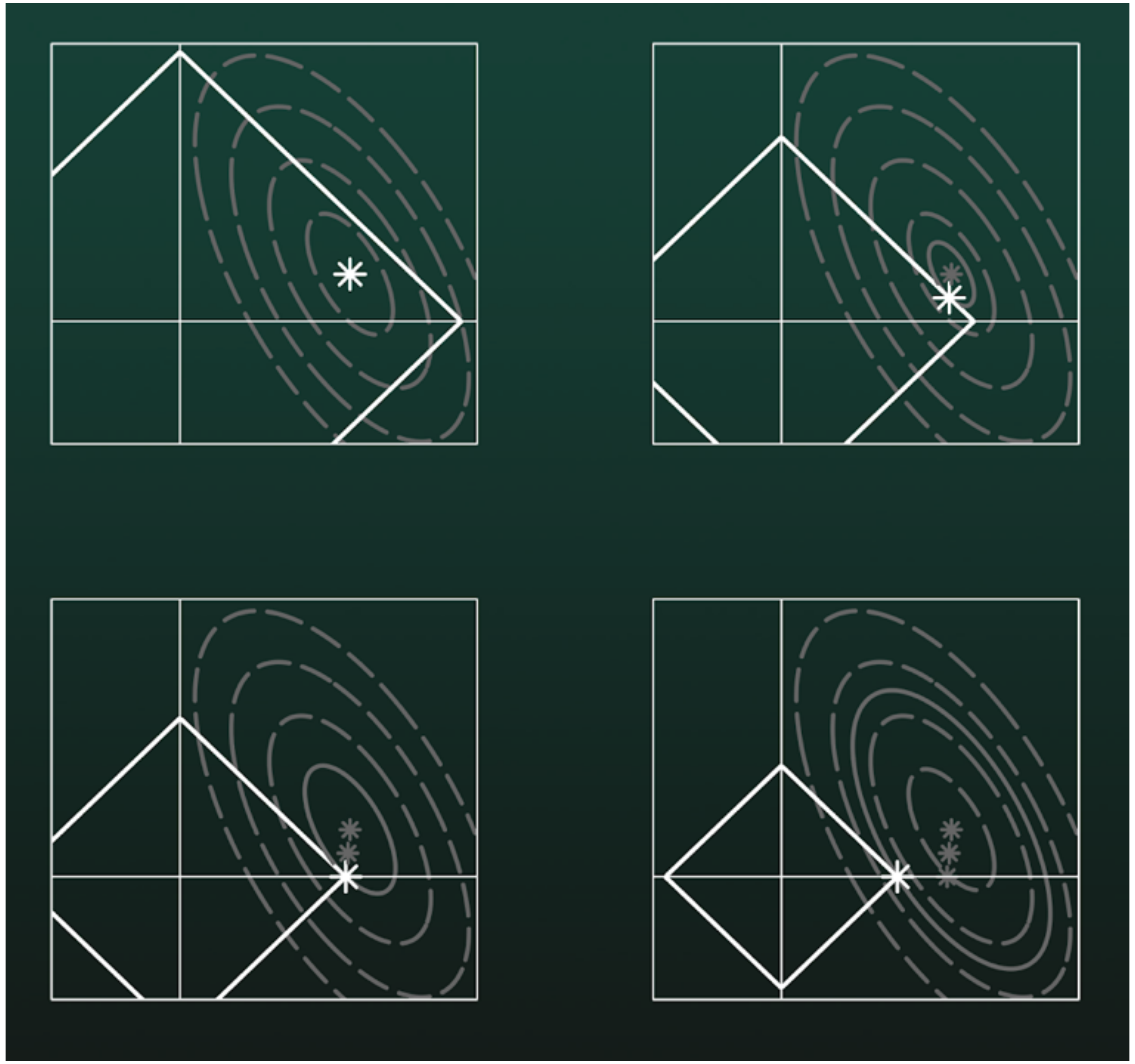


Figure from Christophe Giraud “Introduction to High-Dimensional Statistics”

Variable selection in genomics

— methods, challenges, and possibilities

Variable selection in genomics

— methods, challenges, and possibilities

Filters and wrappers considered harmful

Filters and wrappers considered harmful

Coefficients biased away from 0 ➠ "overfitting"

Filters and wrappers considered harmful

Collinearity introduces arbitrariness ➠ instability

Filters and wrappers considered harmful

Standard errors too small ➠ overconfidence

Filters and wrappers considered harmful

Use of arbitrary inclusion criteria

Filters and wrappers considered harmful

Even if we only care about predictions, the overfitting should worry us

Embedded methods

Embedded methods

Also unstable under collinearity

Embedded methods

Only real contenders of penalized likelihood variety (eg. LASSO)

Embedded methods

Difficult to sensibly use categorical variables

Embedded methods

Difficult to embed prior information (pathway info &c.)

All variable-selected models difficult to interpret

Variable selection in genomics

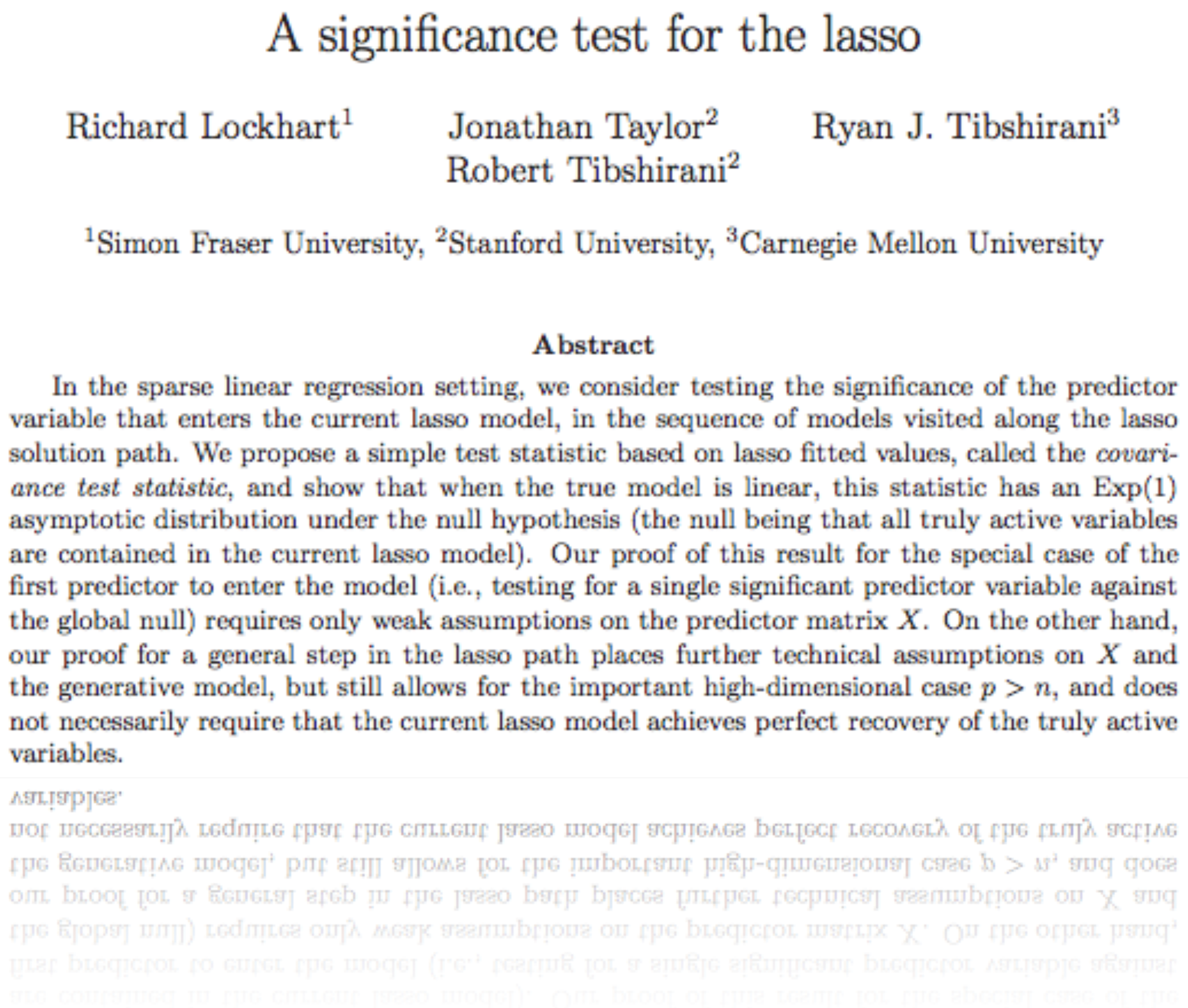
— methods, challenges, and possibilities

Variable selection in genomics

— methods, challenges, and possibilities

Post–selection inference

Post–selection inference



A significance test for the lasso

Richard Lockhart[1] Jonathan Taylor[2] Ryan J. Tibshirani[3]
Robert Tibshirani[2]

[1]Simon Fraser University, [2]Stanford University, [3]Carnegie Mellon University

Abstract

In the sparse linear regression setting, we consider testing the significance of the predictor variable that enters the current lasso model, in the sequence of models visited along the lasso solution path. We propose a simple test statistic based on lasso fitted values, called the *covariance test statistic*, and show that when the true model is linear, this statistic has an Exp(1) asymptotic distribution under the null hypothesis (the null being that all truly active variables are contained in the current lasso model). Our proof of this result for the special case of the first predictor to enter the model (i.e., testing for a single significant predictor variable against the global null) requires only weak assumptions on the predictor matrix X. On the other hand, our proof for a general step in the lasso path places further technical assumptions on X and the generative model, but still allows for the important high-dimensional case $p > n$, and does not necessarily require that the current lasso model achieves perfect recovery of the truly active variables.

Classical inference treats hypothesis as fixed; now it is often random

Post–selection inference

UiT
THE ARCTIC UNIVERSITY OF NORWAY

Faculty of Science and Technology
Department of Computer Science

Small data: practical modeling issues in human-model -omic data

—

Einar Holsbø
A Dissertation for the degree of Philosophiae Doctor — 2018

Resampling, data splitting possible, can be hard to get right

Post–selection inference

Bayesian methodology mostyl sidesteps the inferential problems.

More work to model, compute-heavy. “Subjective.”

Reducing number of variables blinded to **Y**

Reducing number of variables blinded to **Y**

- Remove low-variance variables

Reducing number of variables blinded to **Y**

- Remove low-variance variables
- Remove mostly-missing variables

Reducing number of variables blinded to **Y**

- Remove low-variance variables
- Remove mostly-missing variables
- Statistical tricks to combine collinear variables &c. (see refs)

Reducing number of variables blinded to **Y**

- Remove low-variance variables
- Remove mostly-missing variables
- Statistical tricks to combine collinear variables &c. (see refs)
- Domain knowledge

To summarize

To summarize

- **Variable selection is a modern “problem”**

To summarize

- **Variable selection is a modern “problem”**
- **Genomics is an archetypal application area**

To summarize

- **Variable selection is a modern “problem”**
- **Genomics is an archetypal application area**
- **Penalized likelihood methods probably most reliable**

To summarize

- **Variable selection is a modern “problem”**
- **Genomics is an archetypal application area**
- **Penalized likelihood methods probably most reliable**
- **Inference is tricky**

To summarize

- **Variable selection is a modern “problem”**
- **Genomics is an archetypal application area**
- **Penalized likelihood methods probably most reliable**
- **Inference is tricky**
- **Domain knowledge is both a challenge and a possibility**

Data seldom, if ever, speaks for itself. To use data effectively requires valid and revealing conceptual frameworks for understanding and interpreting patterns in data.

Nobel laureate Lars Hansen (emphasis mine).

https://qz.com/1417105/lars-hansen-a-nobel-winning-economist-has-tips-for-dealing-with-uncertainty/

UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

THANK YOU

Einar Holsbø
February 8th, 2019

UNIVERSITY OF TROMSØ

Bibliography

- Harrell: “Regression modeling strategies”
- Hastie &al.: “Elements of statistical learning”
- Hira & Gillies: “A review of feature selection and feature extraction methods applied on microarray data”
- The methods SAM, LIMMA, and k-TSP