A First Course in Machine Learning Chapter 1

Chao-En Yu

National Tsing Hua University

Supervised Learning

- Machine learning: Building a model by learning from labeled experience (including data).
- An important and general problem in machine learning is learning or inferring a *functional relationship* between a set of *attribute* or feature variables and associated *response* or target variables.
- So that based on the model, we can *predict* the response for any set of attributes.

What items a particular customer bought?

Whether or not a customer like recommendations?

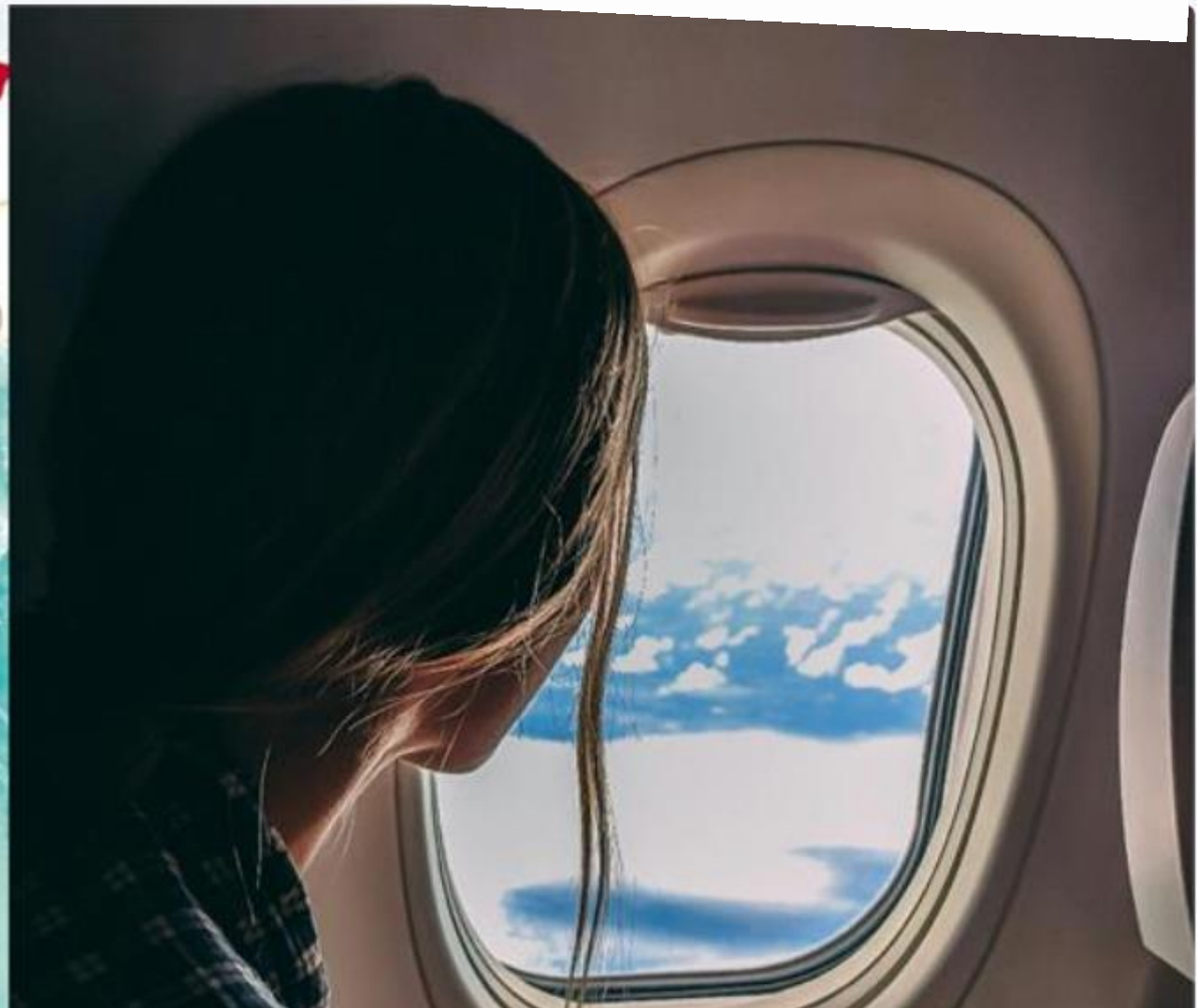
Why do We Need Predictions?

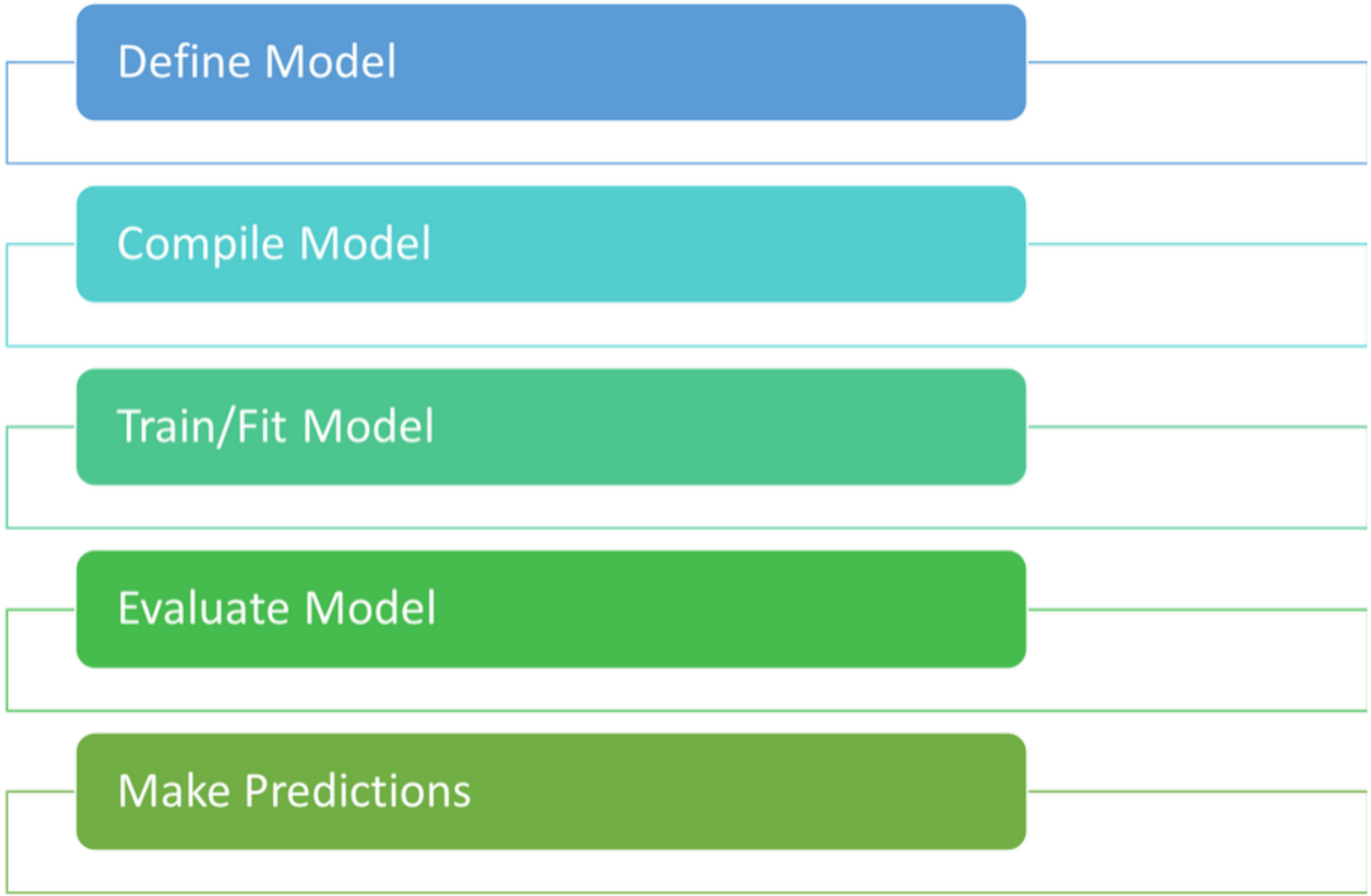
- Rationality: Using all available information to achieve goals, which includes people's predictions.
- All decisions are more or less made based on explicit or implicit predictions.
- It is not easy (but not impossible) to assess learned models without predictions in advance.
- Better and faster predictions help to make better decisions. (p.168)

竜樹諒「7／5末日預言」釀恐慌　專家估：經濟損失達5600億日圓

發佈時間：2025/07/04 18:16
更新時間：2025/07/04 18:32

K董承諾破功！日本末日謠言害星宇航空第3季虧損近4.5億元



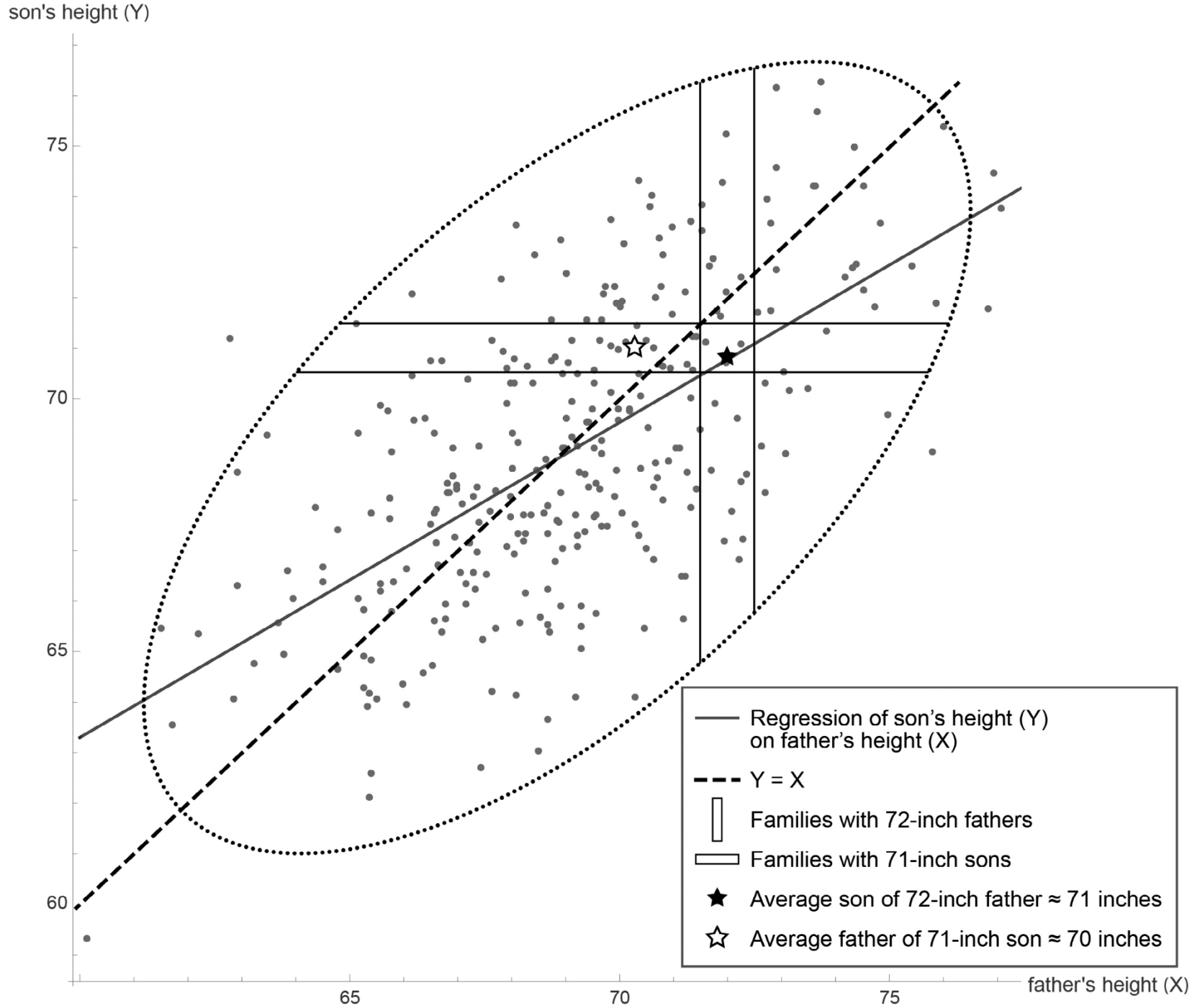

Define Model
Compile Model
Train/Fit Model
Evaluate Model
Make Predictions

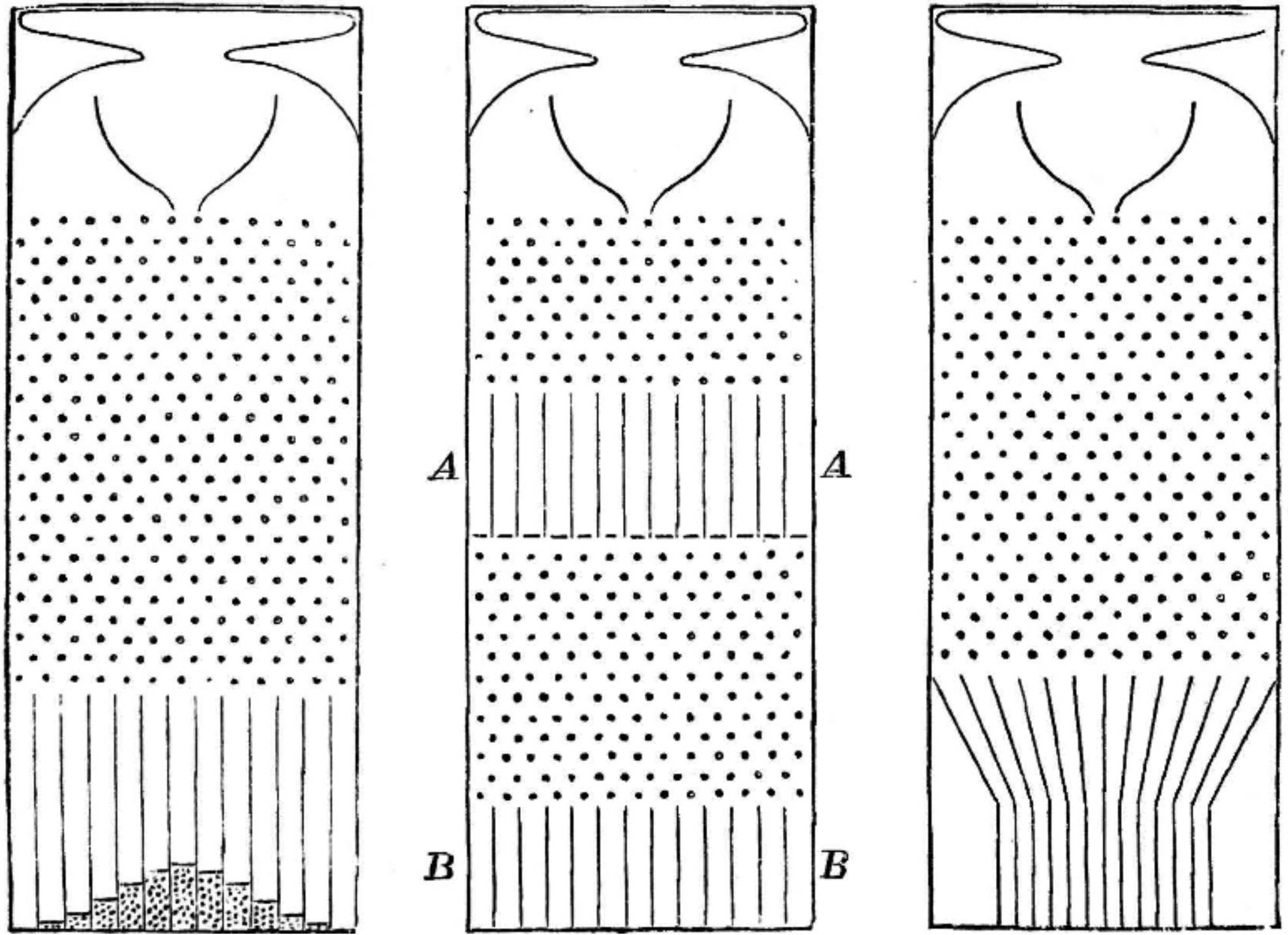
Linear Modelling

- The most straightforward of learning problems, linear modelling (learning a linear relationship between attributes and responses).

- This linear modelling is often known as *regression*, originally used in the context of genetics by Francis Galton (Friday Feb 9 1877). https://youtu.be/CdjnXh3eiy4

- The puzzle of why human heights do not spread out from one generation to the next, led him to the discovery of “regression to the mean.”


son's height (Y)
75
70
65
60
65
70
75
father's height (X)
Regression of son's height (Y) on father's height (X)
Y = X
Families with 72-inch fathers
Families with 71-inch sons
Average son of 72-inch father ≈ 71 inches
Average father of 71-inch son ≈ 70 inches


A
A
B
B

Assumptions

- The relationship between the gold medal winning time for the men's 100m at the Olympic Games held since 1896 and the year is assumed to linear:

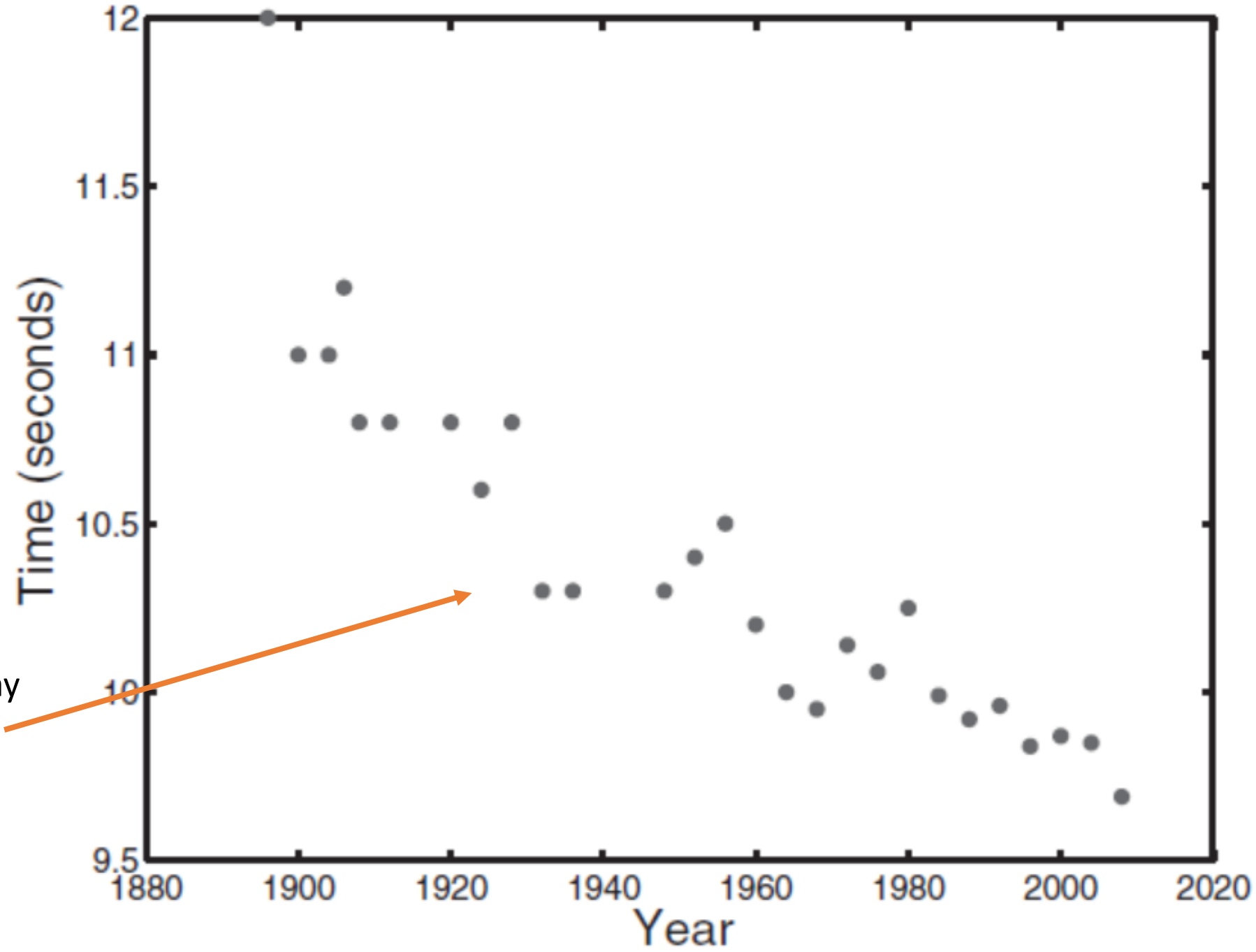
$$t = f(x; w_o, w_1) = w_o + w_1 x$$

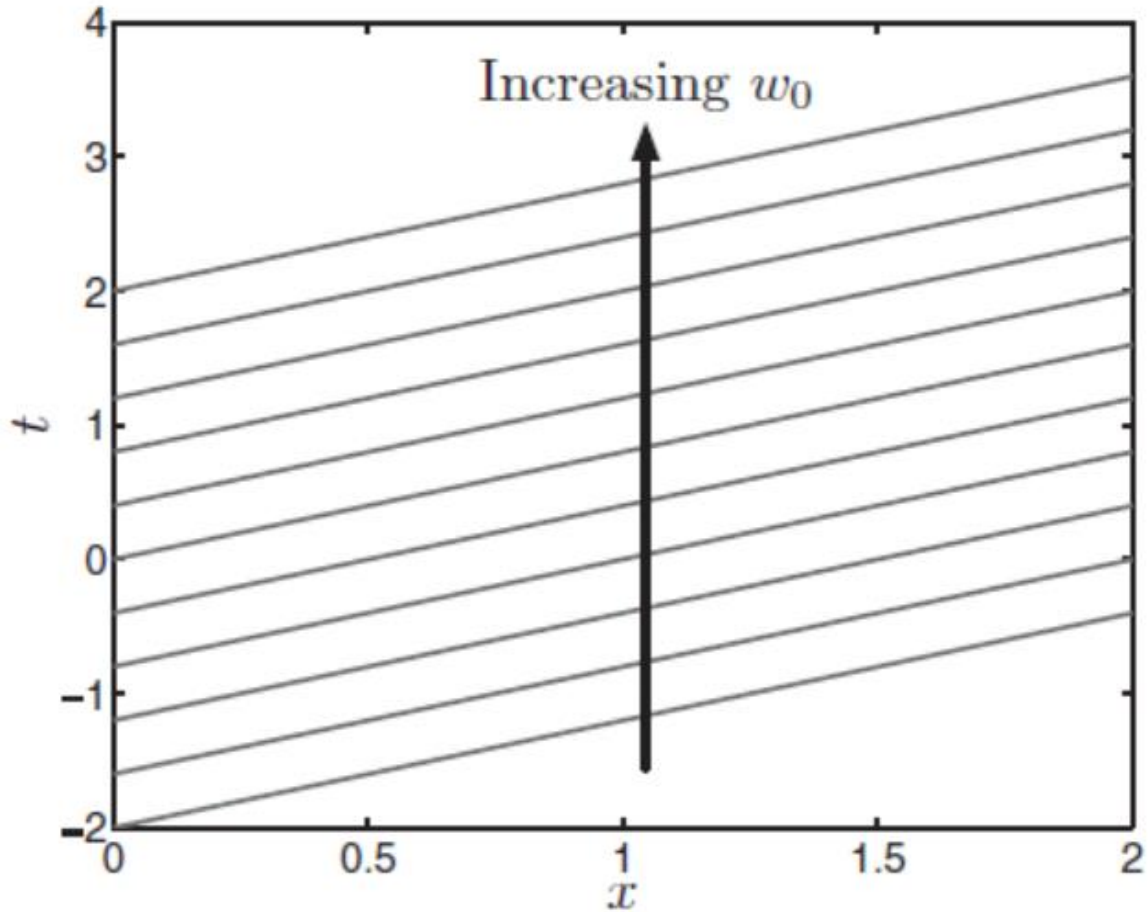
- The learning task involves using the data to choose suitable values for the two parameters w_0 and w_1.

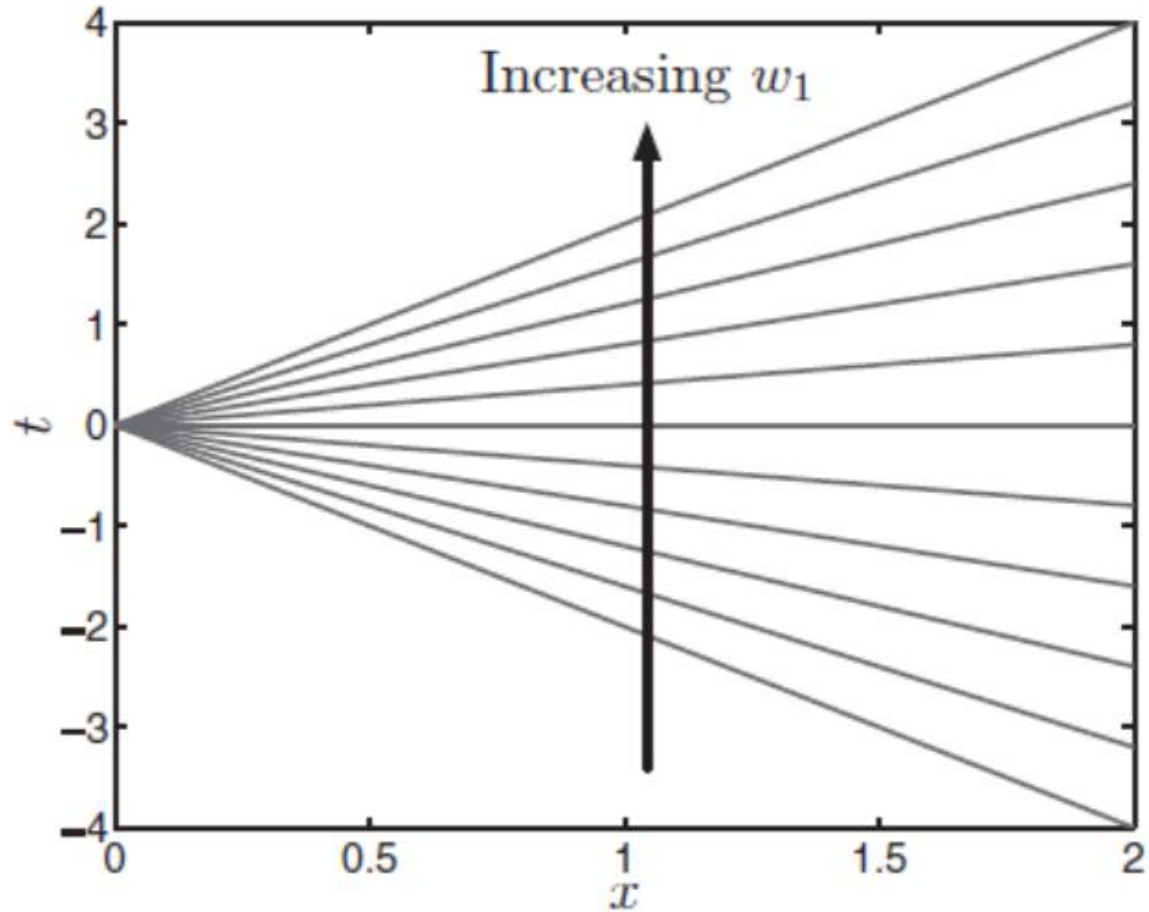
- The best solution consists of the values of w_0 and w_1 that produce a line that passes *as close as possible* to all of the data points.

Is it possible to pass through all data points?

Statistical dependence may **Not** be a causal dependence




Increasing w_0
4
3
2
1
0
−1
−2
t
0
0.5
1
1.5
2
x


Increasing w_1
4
3
2
1
0
−1
−2
−3
−4
t
0
0.5
1
1.5
2
x

Defining a Good Model

- A common way of measuring how close a model gets to one of data points is the *squared loss function*:

$$\mathcal{L}_n(t_n, f(x_n; w_o, w_1)) = (t_n - f(x_n; w_o, w_1))^2$$

- As we want a low loss for all data points, we consider finding the best values of parameters as

$$\underset{w_o, w_1}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_n(t_n, f(x_n; w_o, w_1))$$

← Mean squared error (MSE)

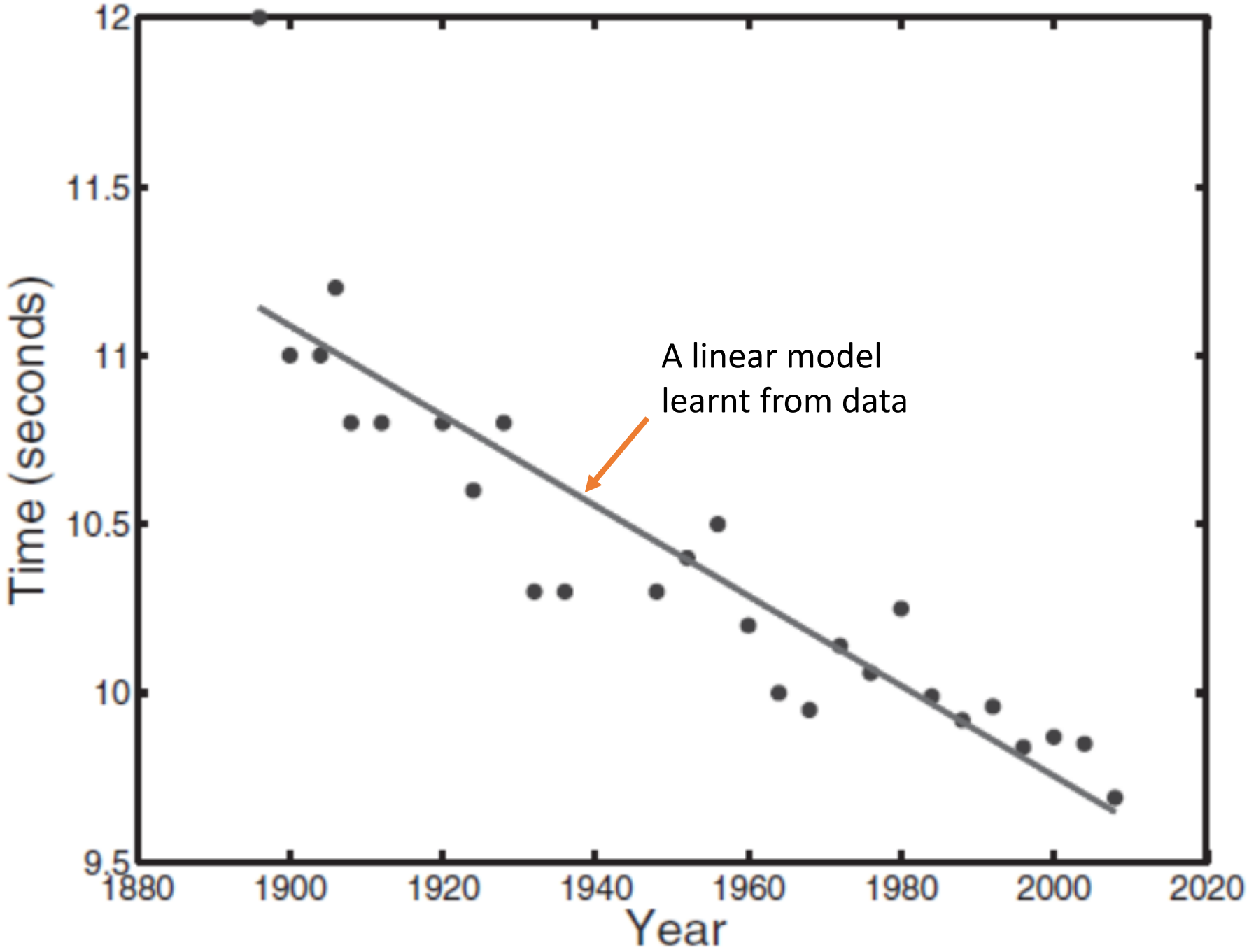
- We choose the squared loss so that an analytical solution can be derived.

Ordinary Least-Squares (OLS)

- This is called the *least-squares errors* method (introduced by Gauss (1795) and Legendre (1805)).

Olympic men's 100 m data.

n	x_n	t_n	$x_n t_n$	x_n^2
1	1896	12.00	22752.0	3.5948×10^6
2	1900	11.00	20900.0	3.6100×10^6
3	1904	11.00	20944.0	3.6252×10^6
4	1906	11.20	21347.2	3.6328×10^6
5	1908	10.80	20606.4	3.6405×10^6
6	1912	10.80	20649.6	3.6557×10^6
7	1920	10.80	20736.0	3.6864×10^6
8	1924	10.60	20394.4	3.7018×10^6


12
11.5
11
10.5
10
9.5
Time (seconds)
1880
1900
1920
1940
1960
1980
2000
2020
Year
A linear model
learnt from data

Making Predictions

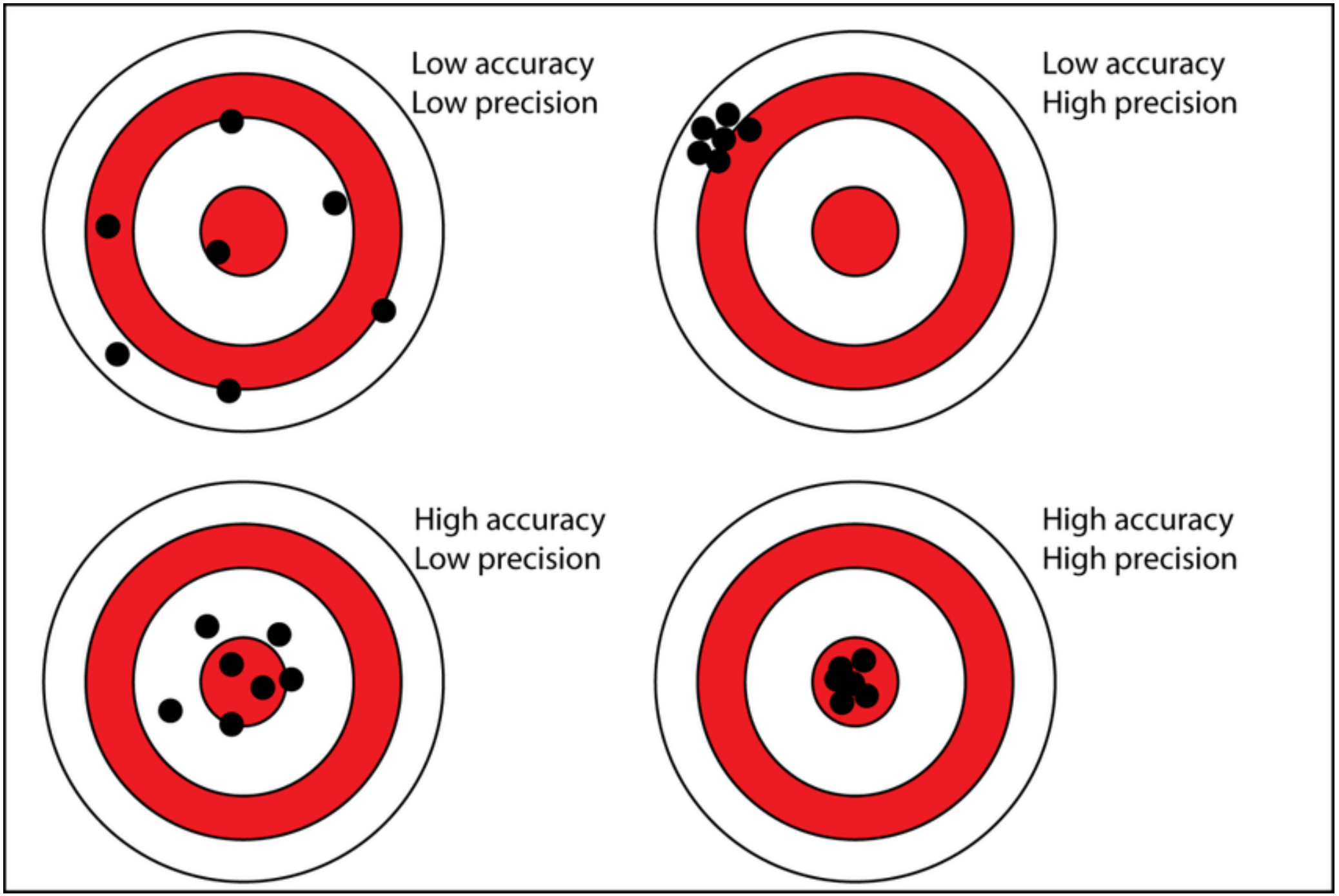
- Based on the linear regression model, we can use it to predict the winning time for a year that we have not yet observed.

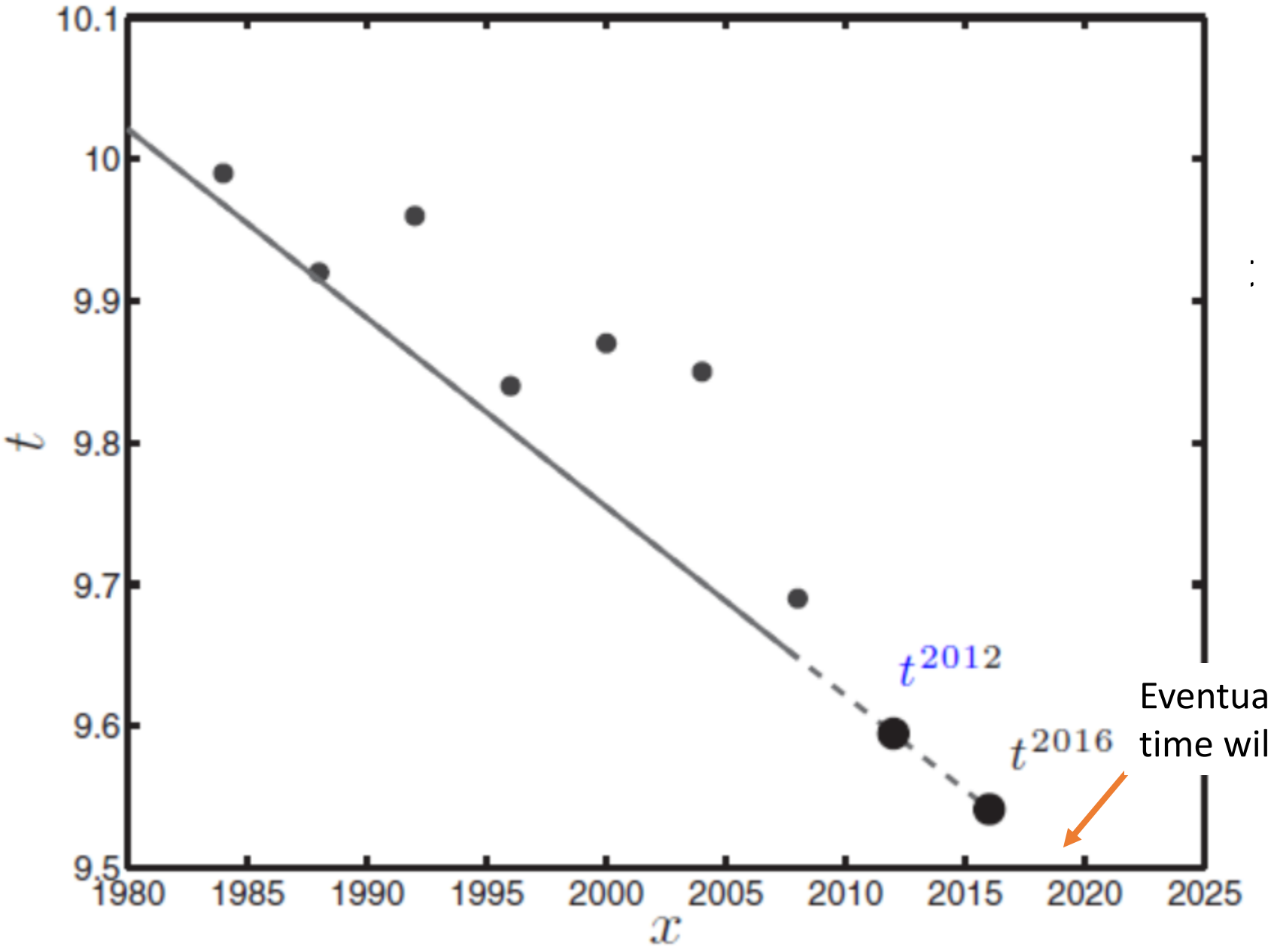
  $f(x; w_o, w_1) = 36.416 - 0.0133\ x$

  $f(2012; w_o, w_1) = 36.416 - 0.0133 * 2012 = 9.595$

  $f(2016; w_o, w_1) = 36.416 - 0.0133 * 2016 = 9.541$

- These values are incredibly *precise*. It seems unlikely to predict the outcome to a high degree of *accuracy*. It is more useful to express a range of values.


Low accuracy
Low precision
Low accuracy
High precision
High accuracy
Low precision
High accuracy
High precision



Eventually the winning time will be 0?

Non-Linear Response

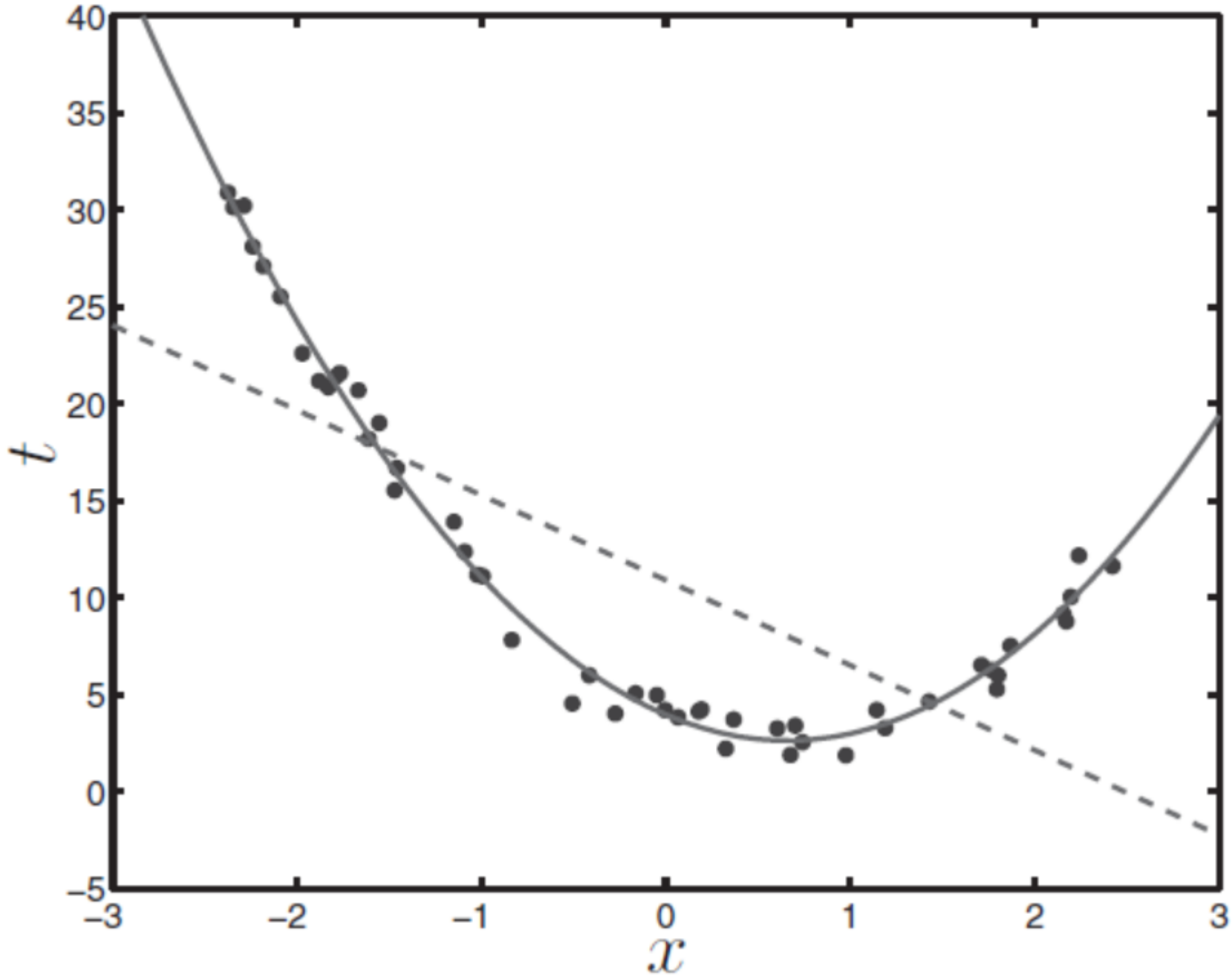
- The **linearity in the parameters** is desirable from a computational point of view. Consider augmenting our data matrix $\mathbf{X}$ with additional column x_n^2.

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{bmatrix}$$

- We are now fitting a *quadratic* function

$$f(x; \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} = w_0 + w_1 x + w_2 x^2$$

and the normal equation can be used to find $\widehat{\mathbf{w}}_{OLS}$.


40
35
30
25
20
15
10
5
0
−5
−3
−2
−1
0
1
2
3
t
x

Fitting Polynomial Functions

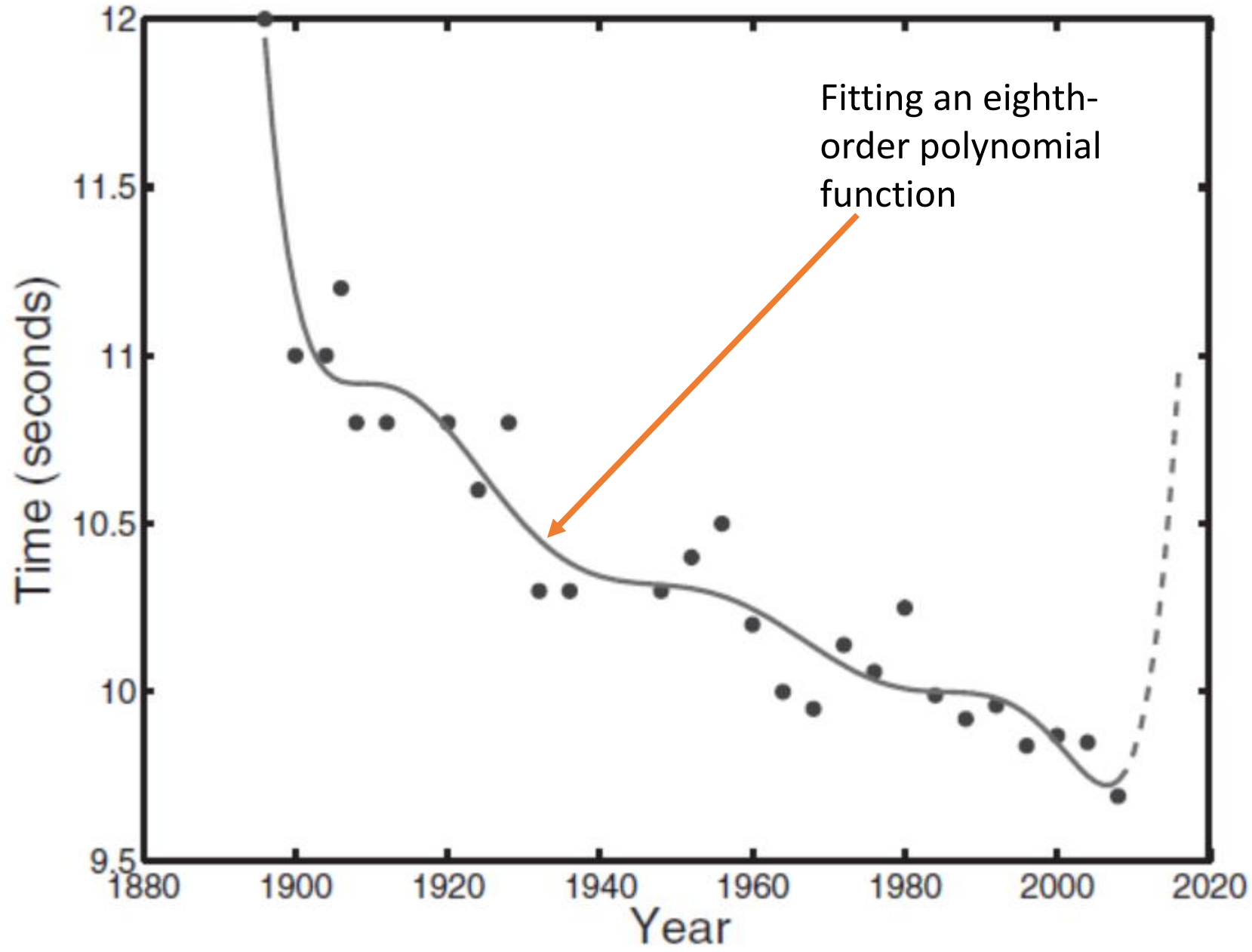
- We can add as many powers of as we like to get a polynomial function of any order $K > 0$.

$$\mathbf{X} = \begin{bmatrix} x_1^0 & \cdots & x_1^K \\ \vdots & \ddots & \vdots \\ x_N^0 & \cdots & x_N^K \end{bmatrix}$$

You usually include all available data, right?
The principle of total evidence by Rudoff Carnap (1947)

- Our function can be written in a more general form

$$f(x:\mathbf{w}) = \sum_{k=0}^{K} w_k x^k$$


Fitting an eighth-order polynomial function
Time (seconds)
Year
12
11.5
11
10.5
10
9.5
1880
1900
1920
1940
1960
1980
2000
2020

Model Complexity

- Is the eighth-order model better than the first-order model? To make the most accurate predictions, the best model should generalize beyond the training data, and perform best on the unseen future data.

- The eighth-order polynomial gets closer to the observed data than the first-order model. ($\mathcal{L}^8 = 0.459, \mathcal{L}^1 = 1.358$)

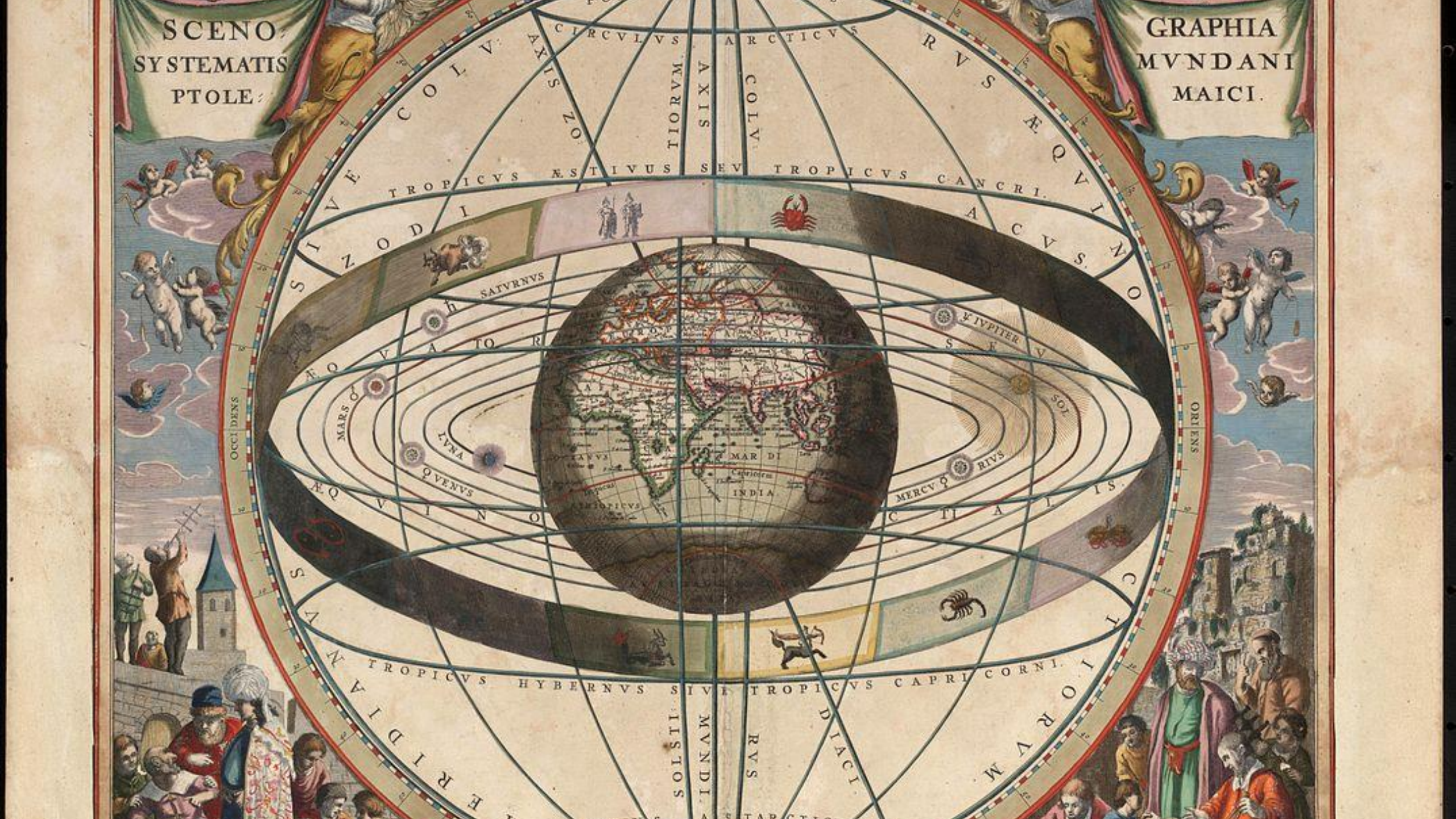
- The (dashed) predictions outside the range of the observed data do not look sensible.

Occam's Razor

- A problem-solving principle:
  If you have two competing ideas to explain the same phenomenon, you should prefer the simpler one.

- Numquam ponenda est pluralitas sine necessitate
  (Plurality should never be placed without necessity)

- Heliocentricism uses 7 assumptions. Geocentricism needs a lot.

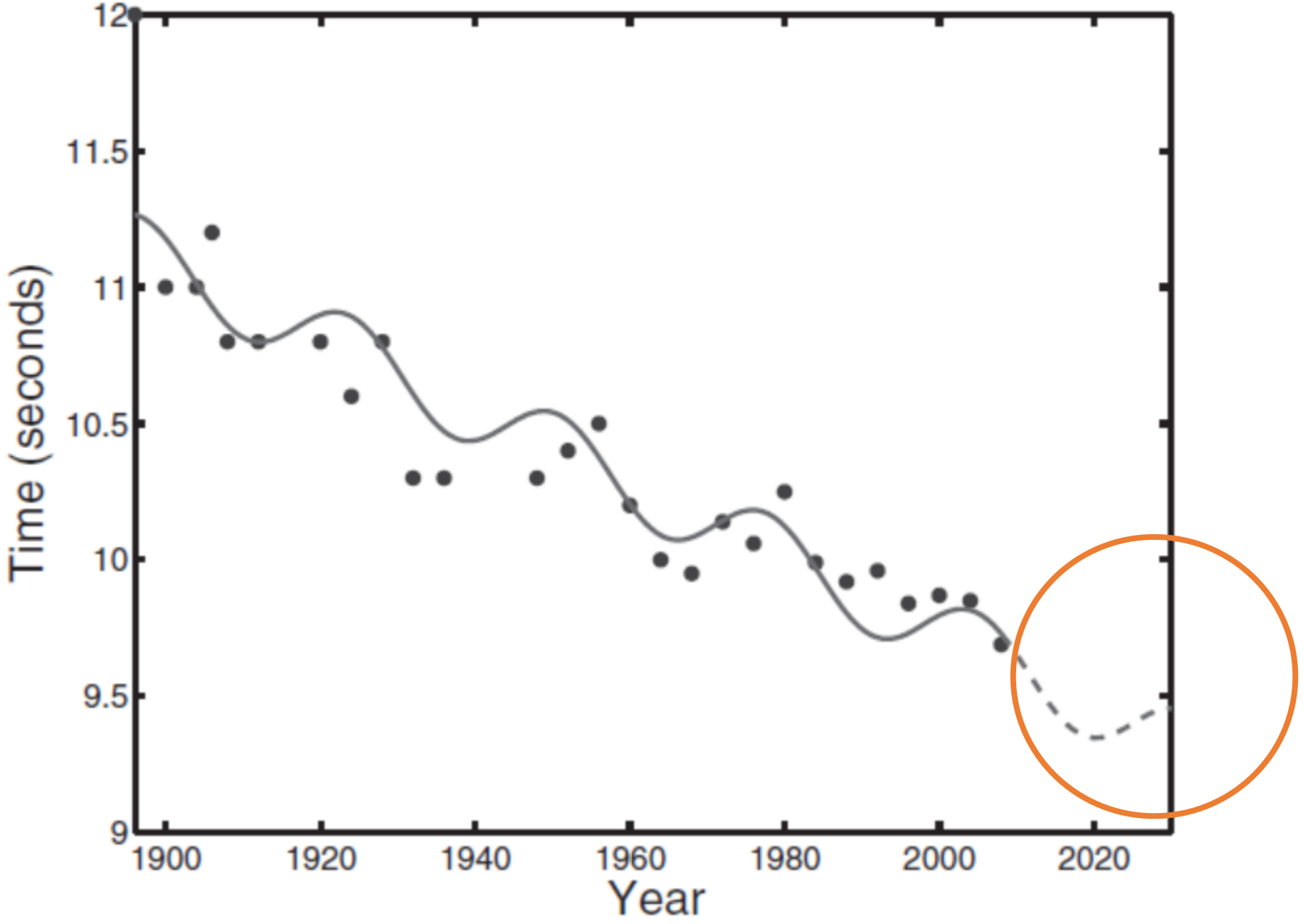

SCENO
SYSTEMATIS
PTOLE
GRAPHIA
MVNDANI
MAICI.
CIRCVLVS ARCTICVS
AXIS MVNDI
COLV
AXIS ZO
SIVE COLV
TROPICVS ÆSTIVVS SEV TROPICVS CANCRI.
ZODI
ACVS.
SATVRNVS
IVPITER
MARS
LVNA
VENVS
SOL
MERCVRIVS
MAR DI
INDIA
OCCI DENS
ORI ENS
TROPICVS HYBERNVS SIVE TROPICVS CAPRI CORNI.
SOLSTI
DIACI

Non-Polynomial Functions

- We are not restricted to polynomial functions.

$$\mathbf{X} = \begin{bmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_K(x_1) \\ h_1(x_2) & h_2(x_2) & \cdots & h_K(x_2) \\ \vdots & \vdots & \cdots & \vdots \\ h_1(x_N) & h_2(x_N) & \cdots & h_K(x_N) \end{bmatrix}$$

- For a LS fit of $f(x; \mathbf{w}) = w_0 + w_1 x + w_2 \sin\left(\frac{x-a}{b}\right)$ with $a = 2660$ and $b = 4.3$, $\mathcal{L} = 1.1037$
  (The non-linear basis function causes oscillations).


12
11.5
11
10.5
10
9.5
9
Time (seconds)
1900
1920
1940
1960
1980
2000
2020
Year

Model Selection

- We would want to choose a model that both
  (i) Accurately captures the regularities in its training data
  (ii) Generalize well to the unseen data

- Overfitting occurs (a model overfits) when the model *suffers from an overreliance on* or *pays too much attention to* the training data.

- The model does not generalize well to *new* data, and the quality of predictions deteriorates rapidly.

Optimal Model Complexity

- We would want to choose a model that both
  (i) Accurately captures the regularities in its training data
  (ii) Generalize well to the unseen data

- Overfitting occurs (a model overfits) when the model *suffers from an overreliance on* or *pays too much attention to* the training data.

- The model does not generalize well to *new* data, and the quality of predictions deteriorates rapidly.

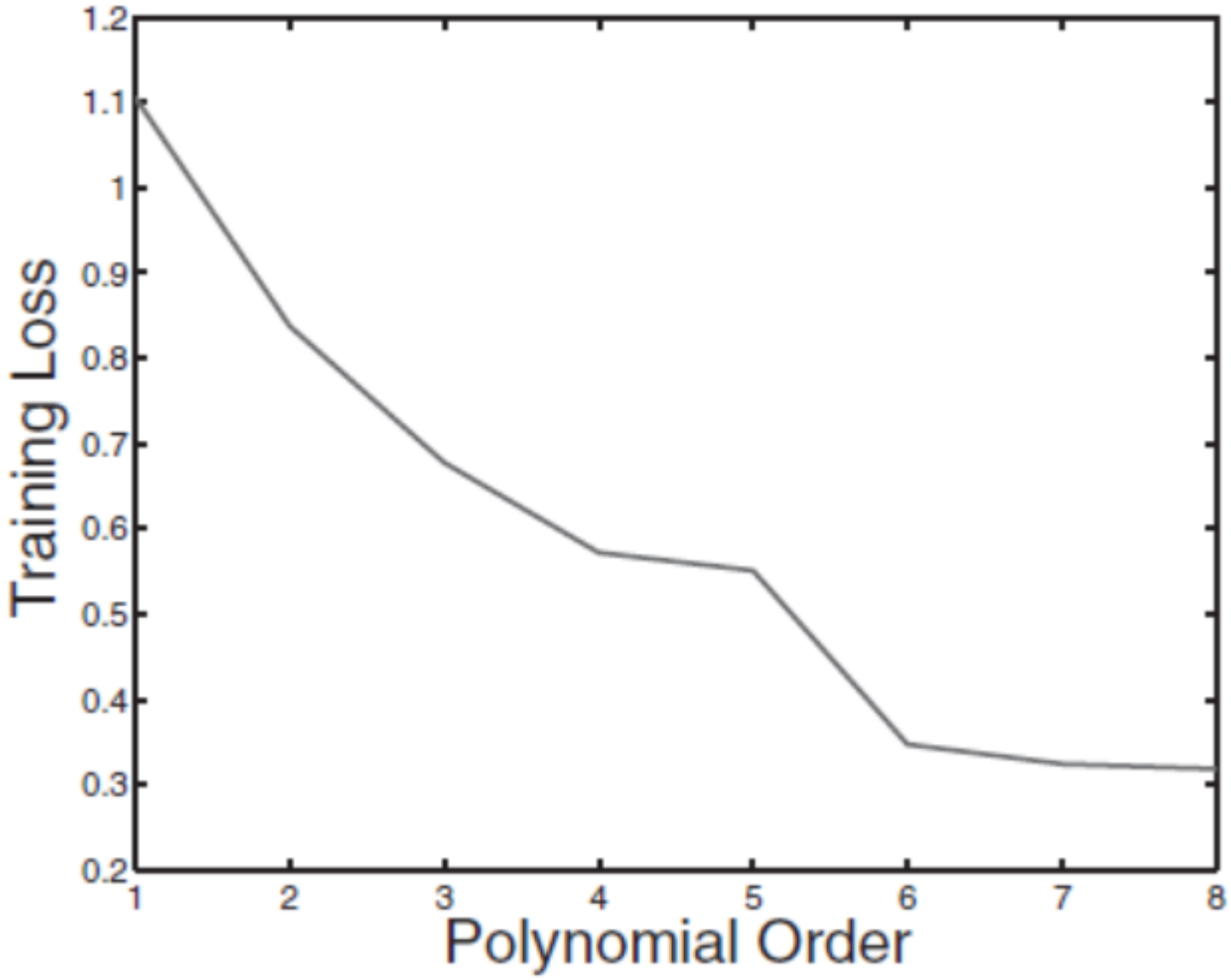
Validation

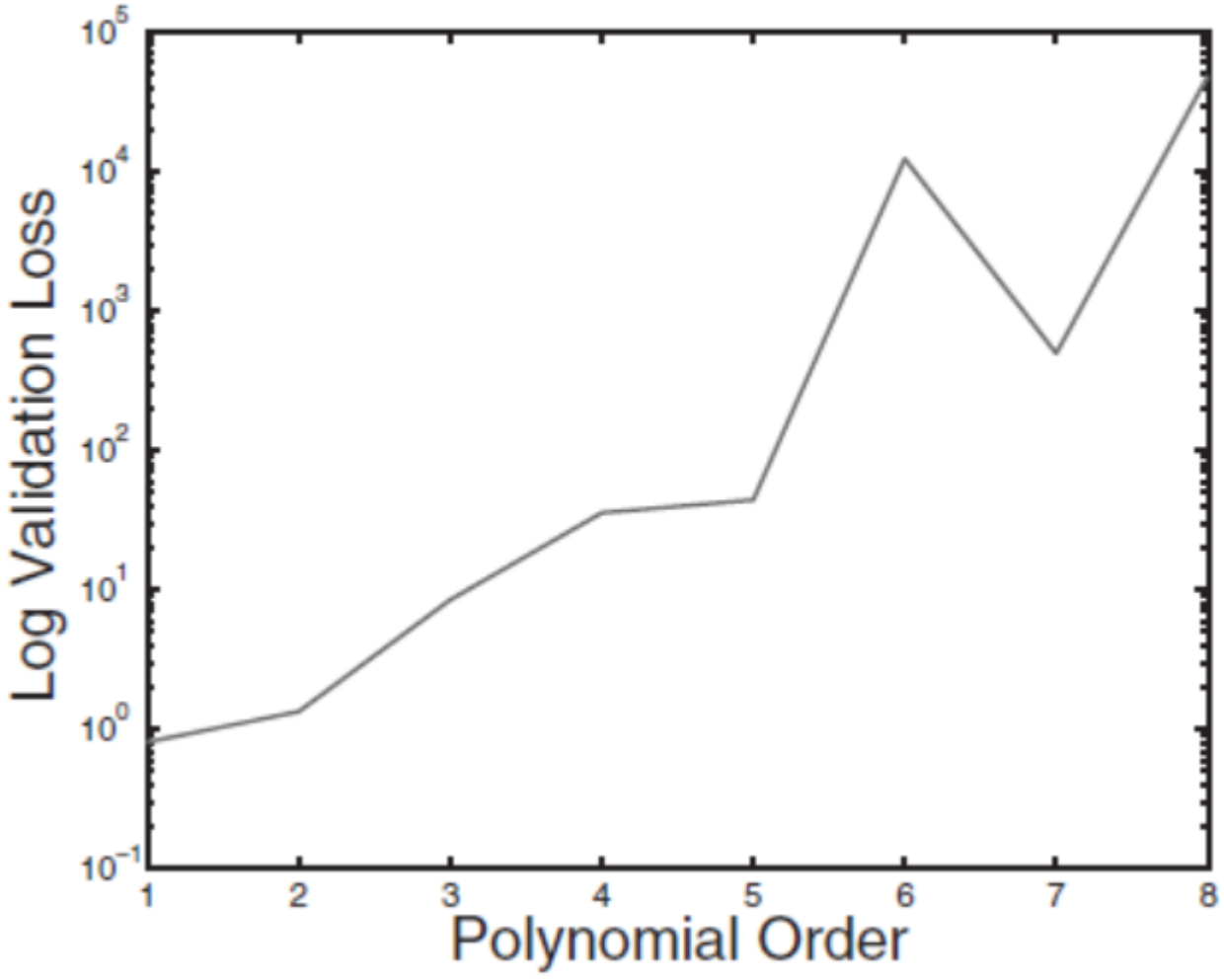
- One common way to overcome the overfitting problem is to use a second dataset (a *validation set*). Ex. Removing all Olympics since 1980 from the training set and make these the validation set.

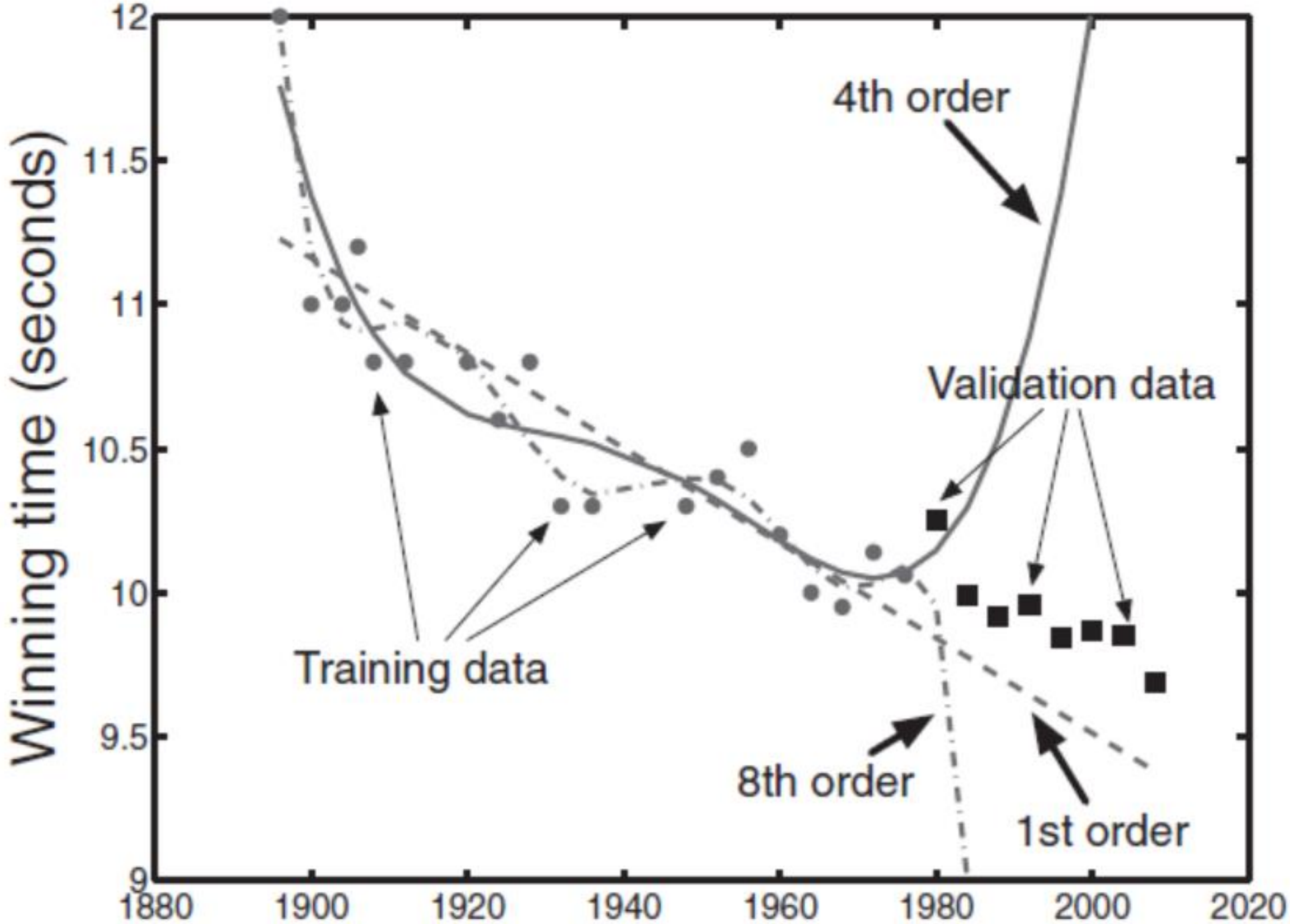
- The training loss decreases monotonically as the polynomial order (model complexity) increases.

- The validation loss suggests that a first-order polynomial has the best *generalization* ability and will produce the most reliable prediction. (np.split())

Validation

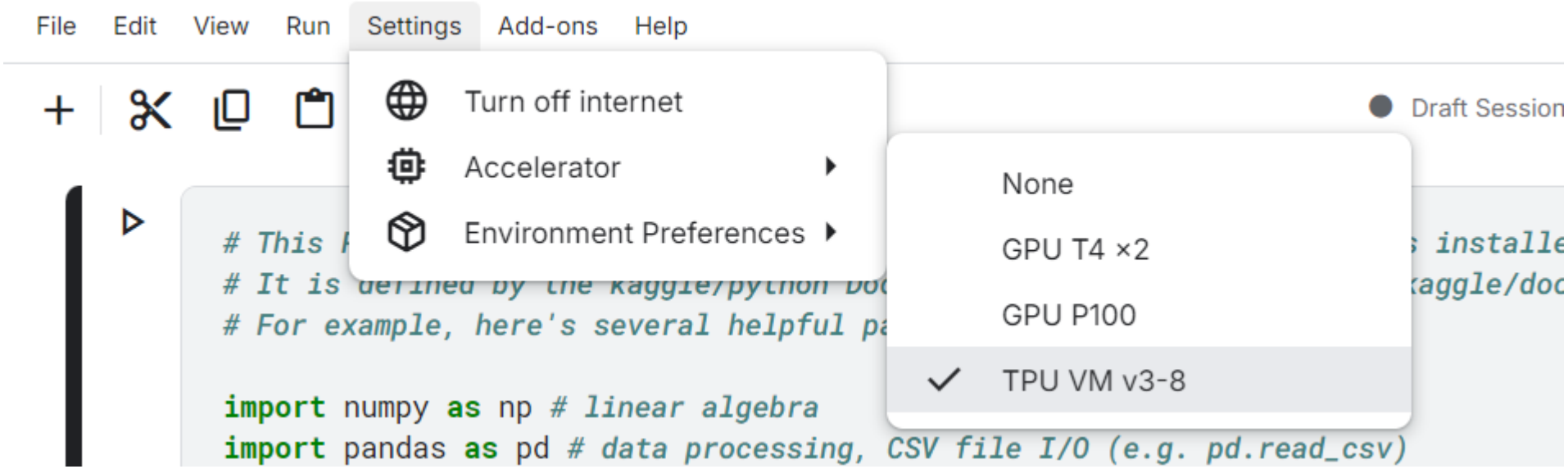

Training Loss
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
1
2
3
4
5
6
7
8
Polynomial Order


Log Validation Loss
10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
1
2
3
4
5
6
7
8
Polynomial Order


Winning time (seconds)
12
11.5
11
10.5
10
9.5
9
1880
1900
1920
1940
1960
1980
2000
2020
4th order
Validation data
Training data
8th order
1st order

Kaggle Platform

- Kaggle is a data science *competition* platform and online community for data scientists and machine learning practitioners under Google LLC.



California Housing Dataset

- The data contains information from the 1990 California census, one row for per census block group. https://www.kaggle.com/datasets/camnugent/california-housing-prices

- A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (typically has a population of 600 to 3,000 people).

Exam 2a

- What is the best *exclusion* of one explanatory variable for the California housing dataset (with the lowest log(validation loss))?

- What is log(validation loss) for the above best model?

https://forms.gle/qZc9qbvZzjWpxuMk7