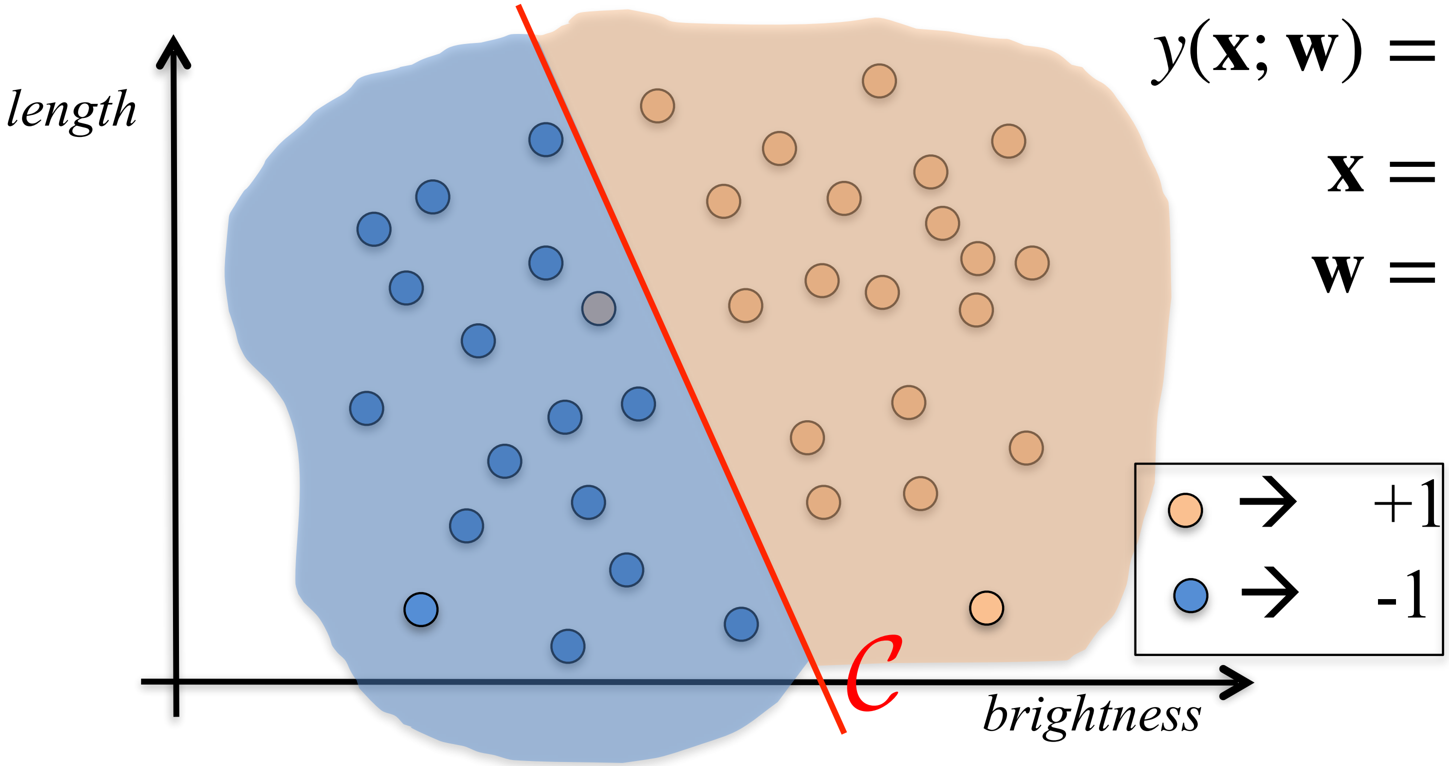
Linear Regression

Pascal Fua
IC-CVLab

Reminder: Linear 2D Model

Some algorithm →

$$\begin{pmatrix} brightness \\ length \end{pmatrix}$$



$$y(\mathbf{x};\mathbf{w}) = \mathrm{sign}(\mathrm{w_x b} + \mathrm{w_y l} + \mathrm{w_0})$$

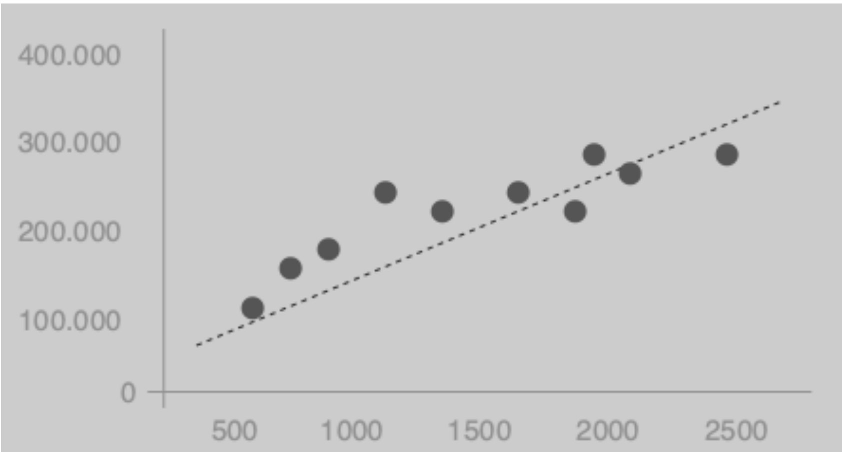
$$\mathbf{x} = [b, l]$$

$$\mathbf{w} = [w_x, w_y, w_0]$$

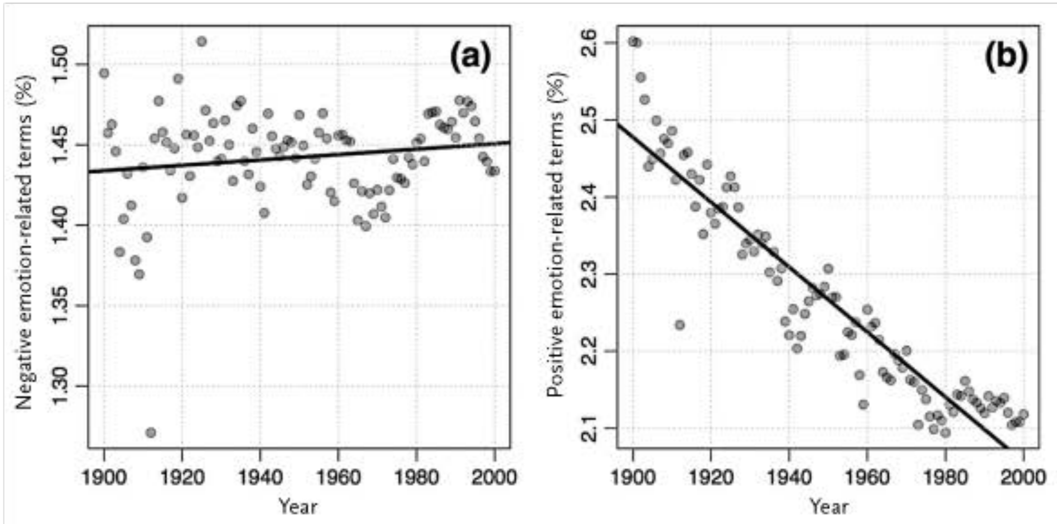
How do we find **w**?

Linear Regression

- In some cases, the value we seek to predict is not a category label.
- Instead, it is a continuous value that follows a pattern:



Price of a house as a function of its size



Proportion of negative and positive emotions in anglophone fiction.

Moretti & Sobchuk, 2019

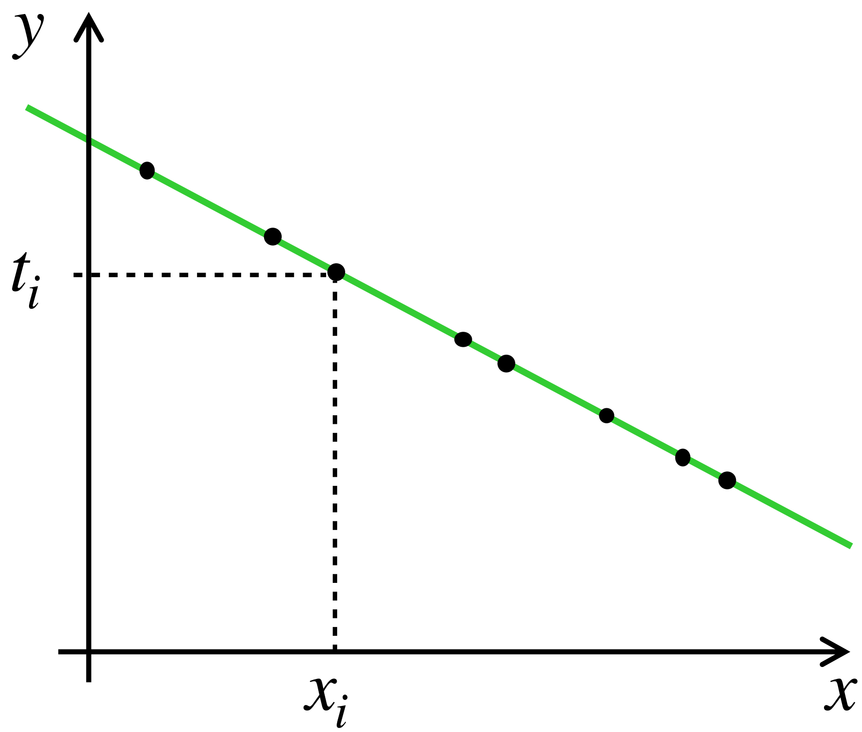
Linear Regression in 1D

We aim to solve a regression task with:

- a single input dimension;
- a single output dimension;
- a continuous output.

Line Fitting without Noise

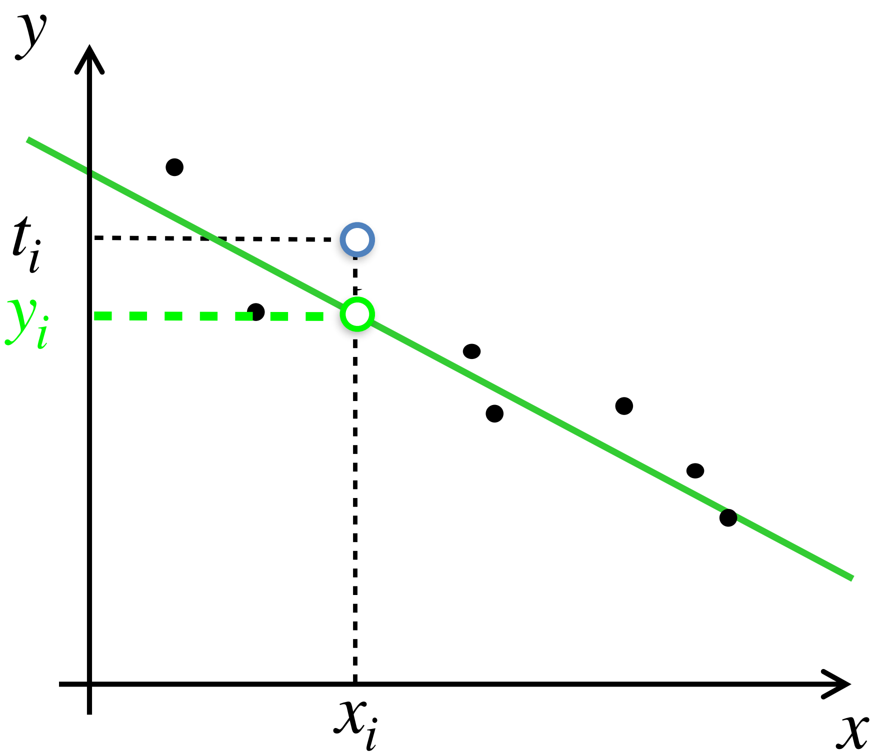
- Given N pairs $\{(x_i, t_i)\}$, find the line that passes through these observations:



- This ideal case never occurs in practice.

Regression with Noise in 1D

- Given N pairs $\{(x_i, t_i)\}$ of noisy measurements, find the line that best fits these observations:



$$y_i = w_0 + w_1 x_i$$
$$= \tilde{\mathbf{w}} \cdot \tilde{\mathbf{x}}_i \text{ with } \tilde{\mathbf{x}}_i = [1, x_i]^t$$
$$d_i = |y_i - t_i|$$

$$\Rightarrow \text{Minimize } \sum_i d_i^2 \text{ w.r.t } \tilde{\mathbf{w}}$$

Multiple Input Dimensions

In general, an input observation $\mathbf{x}_i$ is not represented by a single value

- A grayscale image can be represented by a $W \cdot H$ dimensional vector

$$\mathbf{x}_i = \text{vectorize}(\ \ )\ \in \mathbb{R}^{28\cdot 28} = \mathbb{R}^{784}$$

- Individual samples can have many attributes

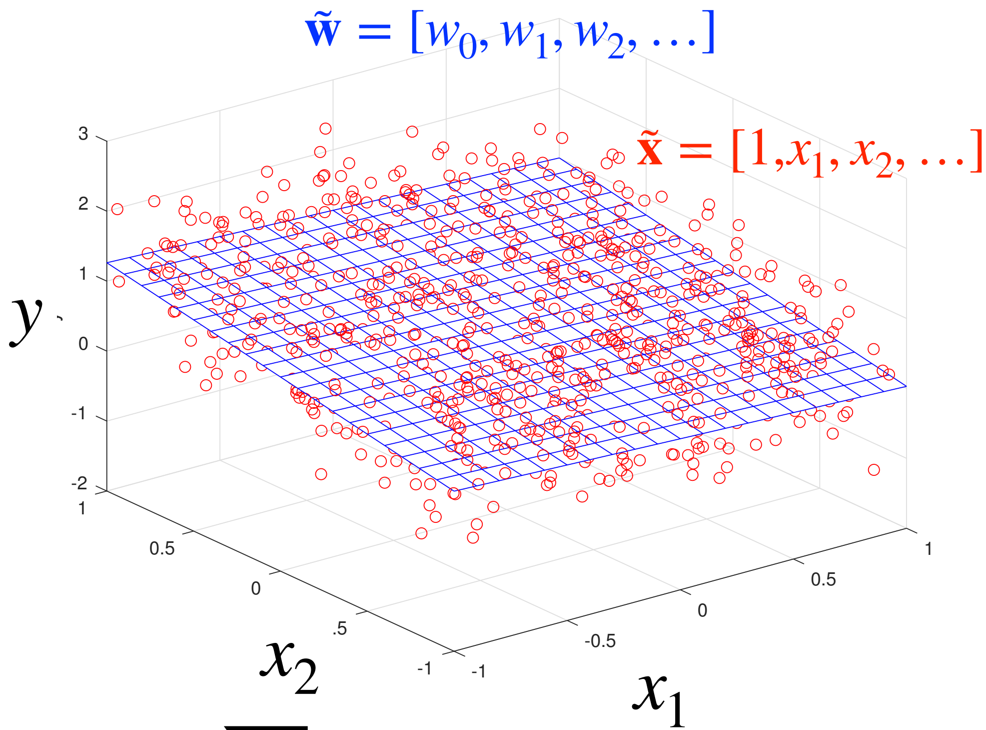
Birth weight prediction

	Age at delivery	Weight prior to pregnancy (pounds)	Smoker	Doctor visits during 1st trimester	Race	Birth Weight (grams)
Patient 1	29	140	Yes	2	Caucasian	2977
Patient 2	32	132	No	4	Caucasian	3080
Patient 3	36	175	No	0	African-Am	3600
*	*	*	*	*	*	*
*	*	*	*	*	*	*
Patient 189	30	95	Yes	2	Asian	3147

EPFL

Regression in Dimension D>1

- Given N pairs $\{(\mathbf{x}_i \in \mathbb{R}^D, t_i)\}$ of noisy measurements, find the hyperplane defined by $\tilde{\mathbf{w}} \in \mathbb{R}^{D+1}$ that best fits these observations.



$\Rightarrow$ Minimize $\sum_i d_i^2$ w.r.t where

$$y_i = \tilde{\mathbf{w}} \cdot \tilde{\mathbf{x}}_i \text{ with } \tilde{\mathbf{x}}_i = [1 \,|\, \mathbf{x}_i]^t \,,$$

$$d_i = |\, y_i - t_i \,| \,.$$

Closed-Form Solution

$$\text{Minimize} \;\; E(\tilde{\mathbf{w}}) = \frac{1}{2}\sum_i d_i^2 = \frac{1}{2}\sum_i (\tilde{\mathbf{w}} \cdot \tilde{\mathbf{x}}_i - t_i)^2$$

At the minimum

$$0 = \nabla E \; ,$$

$$= \sum_i (\tilde{\mathbf{x}}_i^T(\mathbf{w}^* \cdot \tilde{\mathbf{x}}_i) - t_i) \; ,$$

$$\Rightarrow (\sum_i \tilde{\mathbf{x}}_i \cdot \tilde{\mathbf{x}}_i^T)\mathbf{w}^* = \sum_i t_i \tilde{\mathbf{x}}_i \; .$$

$(D+1) \times (D+1)$ matrix — $(D+1) \times 1$ vectors

$\Rightarrow$ $\mathbf{w}^*$is the solution of a linear system.

Linear Regression: Validation

- Once we have the optimal parameters $\mathbf{w}^*$, we can predict y_t for any new test input $\mathbf{x}_t$ as

$$y_t = \mathbf{w}^* \cdot \left[1 \,|\, \mathbf{x}_t\right] \ .$$

- For evaluation purposes, one can compare the predictions to the true values of validation data.
- The evaluation metric may directly be the loss function, but may also differ from it.

Evaluation Metrics

- Mean Square Error (MSE)

$$MSE = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - t_i)^2$$

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - t_i)^2}$$

- Mean Absolute Error (MAE)

$$MAPE = \frac{1}{N_t} \sum_{i=1}^{N_t} \left| y_i - t_i \right|$$

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{N_t} \sum_{i=1}^{N_t} \left| \frac{y_i - t_i}{t_i} \right|$$

Example: UCI Wine Quality Dataset

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

- Predict the quality of wine based on several attributes.
- Final RMSE:
  - 0.65 for training data
  - 0.63 for test data
- The RMSE being relatively similar on training and test data is a good sign.

https://medium.com/datadriveninvestor/regression-from-scratch-wine-quality-prediction-d61195cb91c8

UCI Wine Quality Dataset

	Coeffecient
fixed acidity	0.017737
volatile acidity	-0.992560
citric acid	-0.139629
chlorides	-1.590943
free sulfur dioxide	0.005597
total sulfur dioxide	-0.003520
density	0.768590
pH	-0.437414
sulphates	0.812888
alcohol	0.301484

- One can then look at the coefficient values—the individual w_i—to assess the influence of each attribute.
- But a coefficient might be very small simply to compensate for the fact that the range of the feature is very large.
- This can be corrected by normalizing the features.

I already said that when I talked about K-Means!

Example: Age from Text

- How old is the person who wrote this?

“I can’t sleep, but this time I have school tomorrow, so I have to try I guess. My parents got all pissed at me today because I forgot how to do the homework [...]. Really mad, I ended it pissing off my mom and [...] NOTHING! Damn, when I’m at my cousin’s I have no urge to use the computer like I do here, […]"

- And this?

“[...] I was a little bit fearful of having surgery on both sides at once (reduction and lift on the right, tissue expander on the left) [...] On the good side, my son and family live near the plastic surgeon’s office and the hospital, [...], at least from my son and my granddaughter […]"

Example: Age Prediction from Text

- $\mathbf{x}_i$: number of times each word in the directory appears
- y_i : Age of the writer
- Linear regression assigns a coefficient to each word

like	-1.295
gender-male	-0.539
LIWC-School	-0.442
just	-0.354
LIWC-Anger	-0.303
LIWC-Cause	-0.290
mom	-0.290
so	-0.271
definitely	-0.263
LIWC-Negemo	-0.256

Seems younger

years	0.601
POS - dt	0.485
LIWC - Incl	0.483
POS - prp vbp	0.337
granddaughter	0.332
grandchildren	0.293
had	0.277
daughter	0.272
grandson	0.245
ah	0.243

Seems older

LIWC = Linguistic Inquiry and Word Count

N'Guyen et al. 2011

Example: Age Prediction from Text

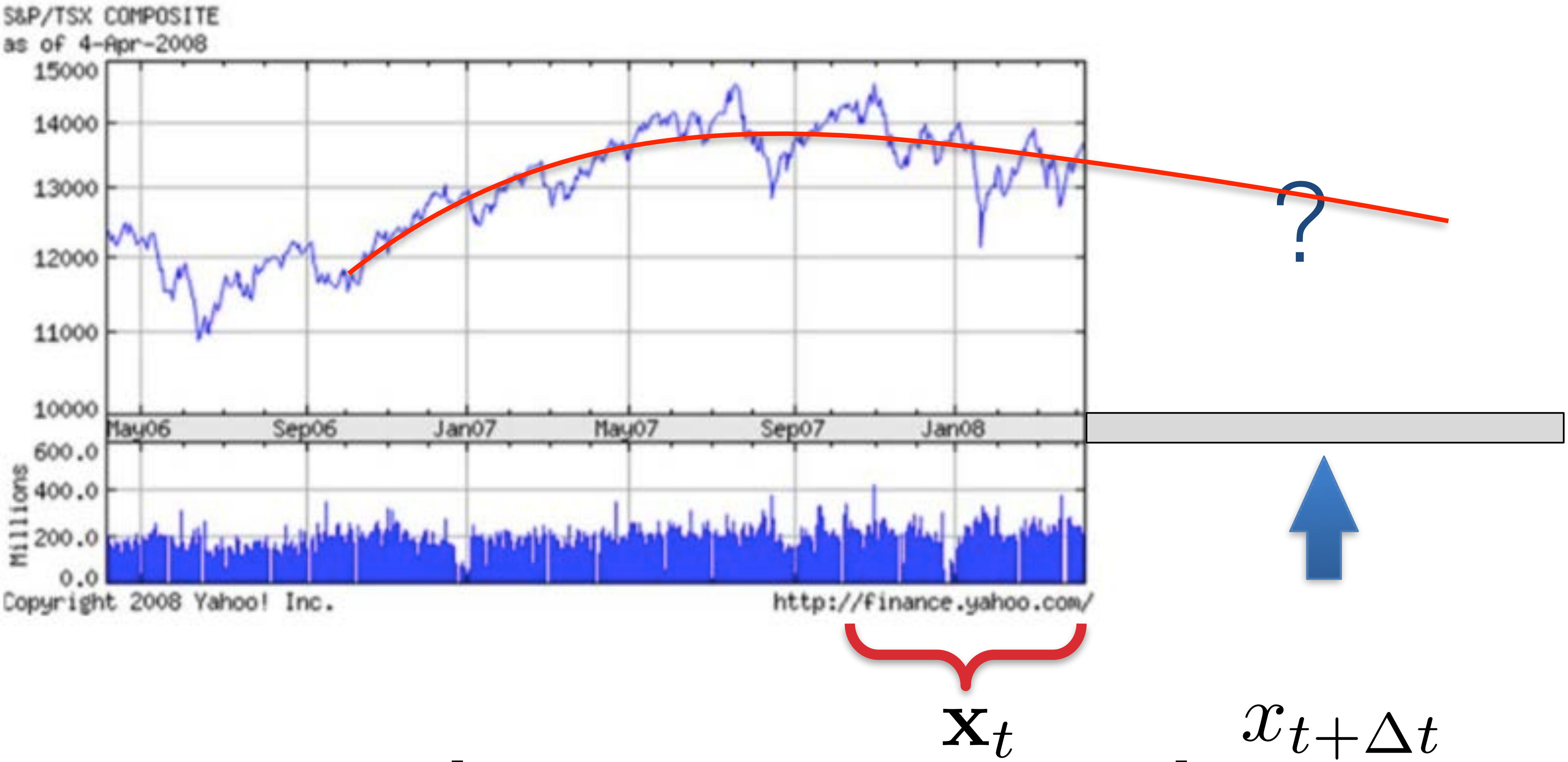
- True age 17, predicted age 16.48

"I can't sleep, but this time I have school tomorrow, so I have to try I guess. My parents got all pissed at me today because I forgot how to do the homework [...]. Really mad, I ended it pissing off my mom and [...] NOTHING! Damn, when I'm at my cousin's I have no urge to use the computer like I do here, […]"

- True age 70, predicted age 71.53

"[...] I was a little bit fearful of having surgery on both sides at once (reduction and lift on the right, tissue expander on the left) [...] On the good side, my son and family live near the plastic surgeon's office and the hospital, [...], at least from my son and my granddaughter […]"

Example: Stock Price Prediction



$$\mathbf{x}_t = [x_{t-T+1}, \ldots, x_{t-1}, x_t]$$

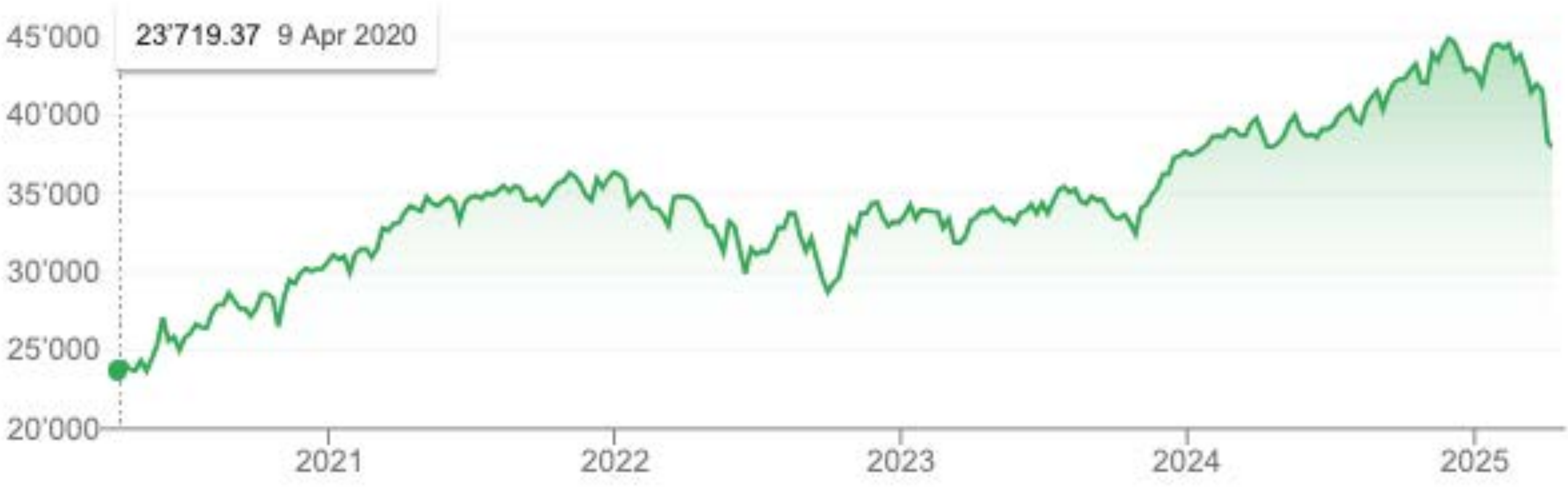
$$y(\mathbf{x}_t; \mathbf{w}) = x_{t+\Delta t}$$

—> Regression problem

But Be VERY Careful!

Dow Jones index from March 28th to April 4th 2025

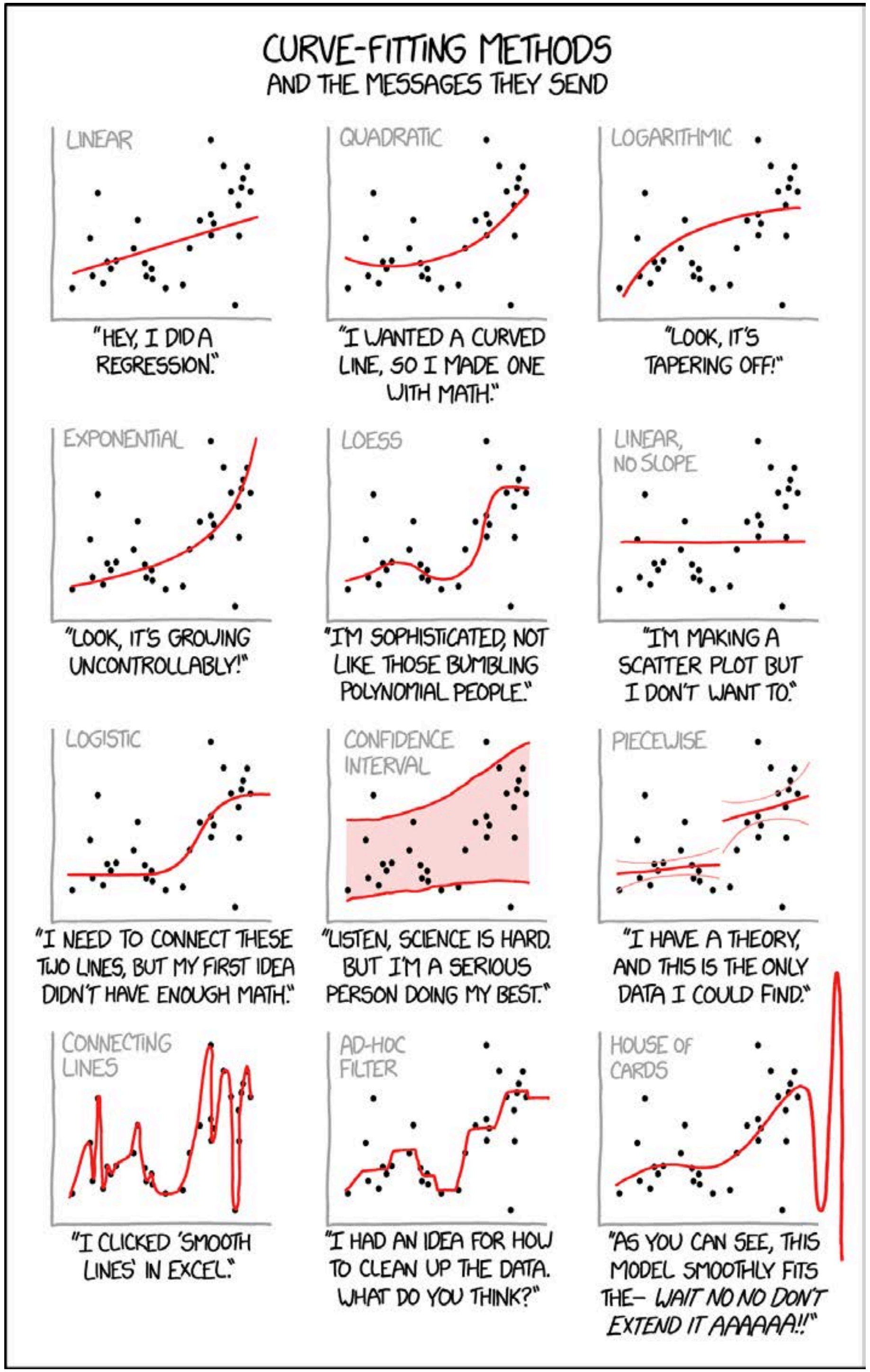
Optional: And how do you regress this?



Depending the time interval chosen, the results will be wildly different:

- Merton & Scholes got a Nobel prize in 1997 for determining the value of derivatives.
- Long-Term Capital Management crashed in 1998, arguably in part because it was using this model (improperly).

Optional: Words of Wisdom from XKCD



Never trust a statistic you have not faked yourself!

https://xkcd.com/2048/

EPFL