Data Visualization
W8-2

Some suggestions for your final presentation / report

- Some quantitative evidence for motivation
- Time
- Novelty

Quiz

- What do you find interesting in today's VotW?
- What is the main advantage of Kernel density estimation (KDE) compared to histograms?
- In KDE, what are the choices you have? And how do they affect the resulting KDE?

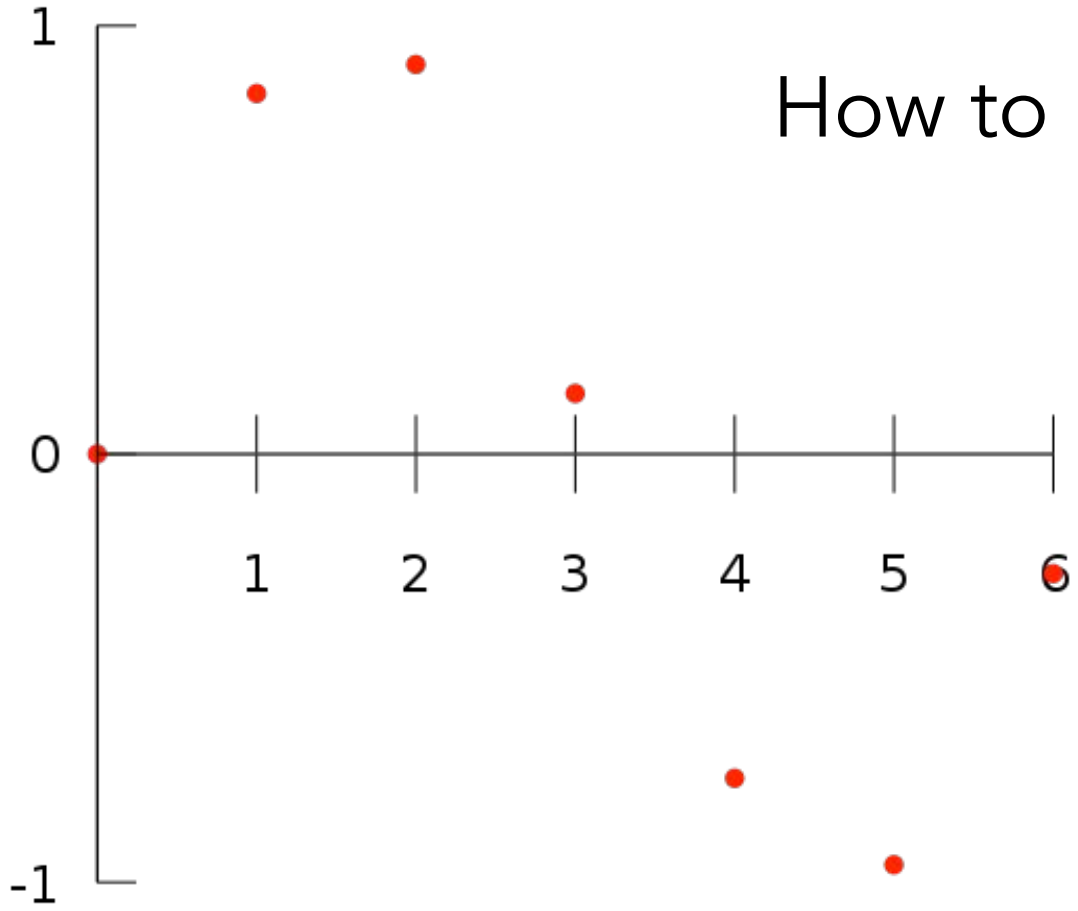
Interpolation, Extrapolation, and Regression

Interpolation

"How can we fill the gaps between data points?"

Interpolation

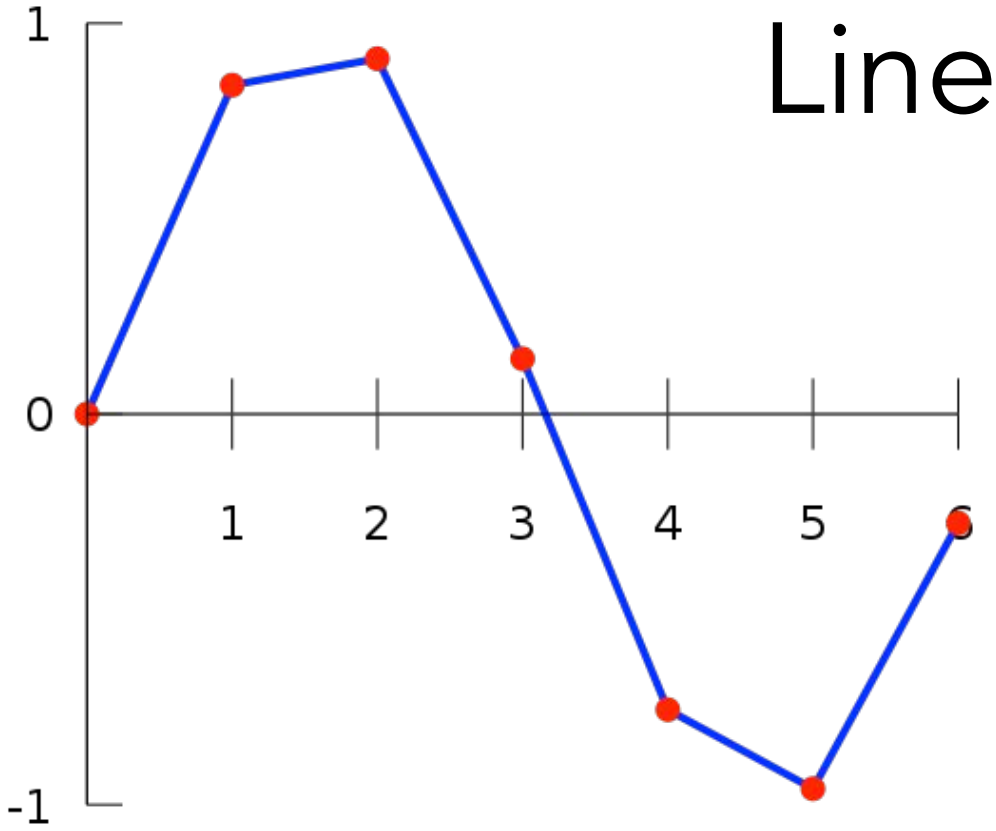
"Let's connect the dots."



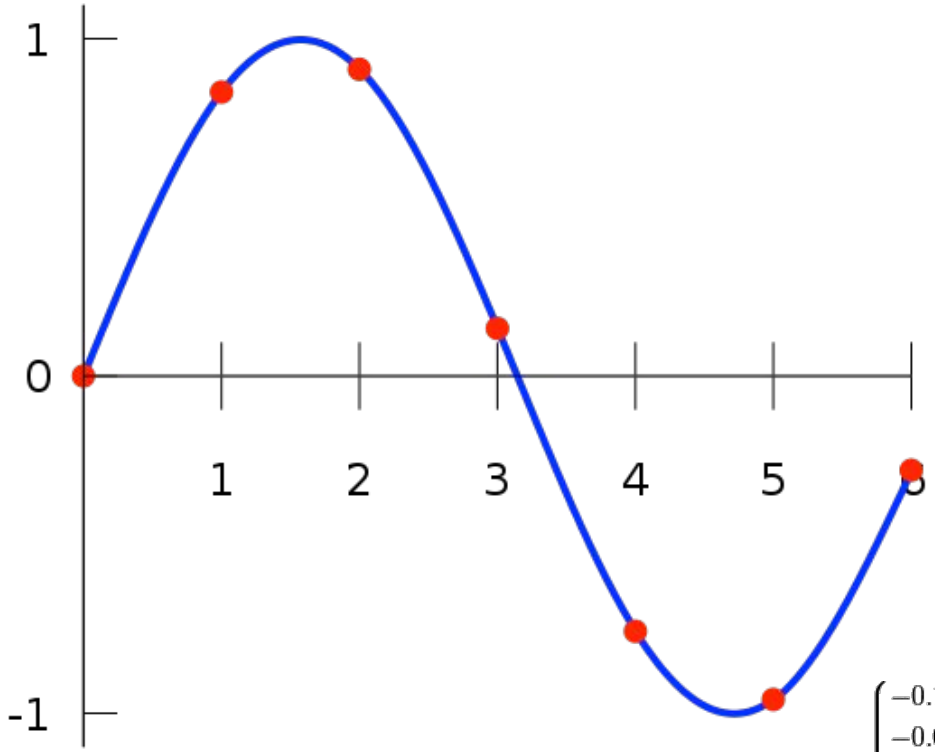
How to interpolate?

http://en.wikipedia.org/wiki/Interpolation

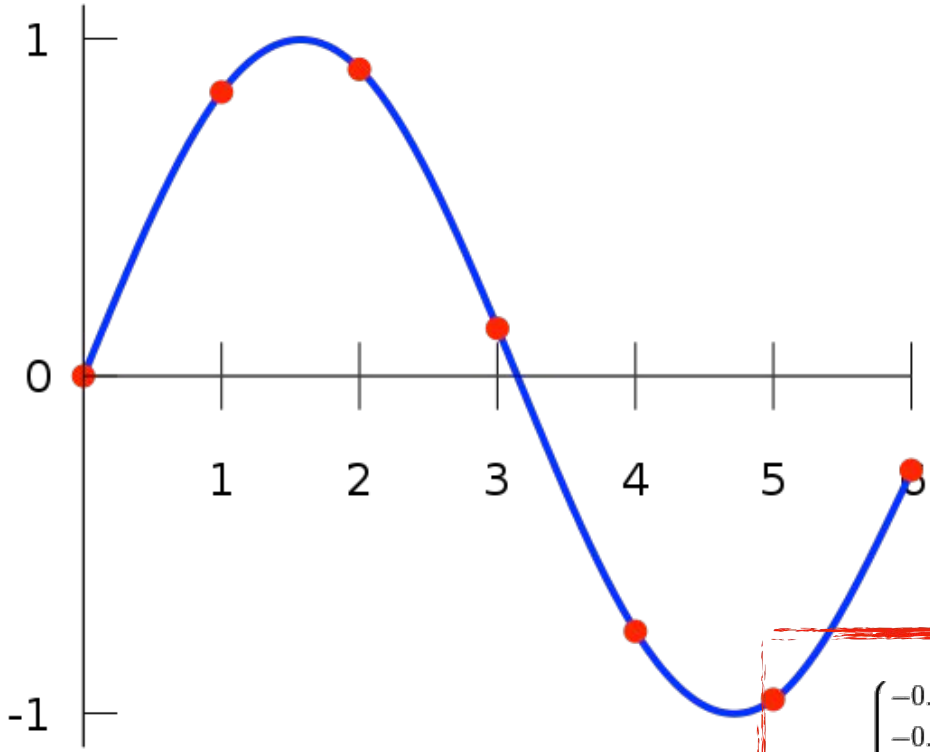
Linear



Splines



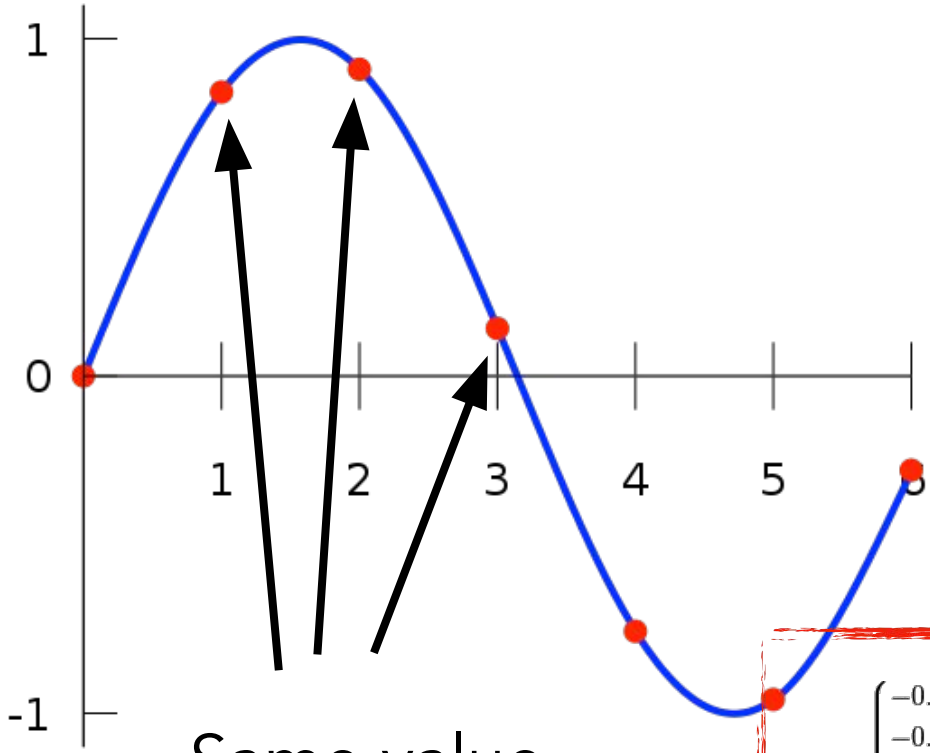
$$f(x) = \begin{cases} -0.1522x^3 + 0.9937x, & \text{if } x \in [0,1], \\ -0.01258x^3 - 0.4189x^2 + 1.4126x - 0.1396, & \text{if } x \in [1,2], \\ 0.1403x^3 - 1.3359x^2 + 3.2467x - 1.3623, & \text{if } x \in [2,3], \\ 0.1579x^3 - 1.4945x^2 + 3.7225x - 1.8381, & \text{if } x \in [3,4], \\ 0.05375x^3 - 0.2450x^2 - 1.2756x + 4.8259, & \text{if } x \in [4,5], \\ -0.1871x^3 + 3.3673x^2 - 19.3370x + 34.9282, & \text{if } x \in [5,6]. \end{cases}$$



Splines

$$f(x) = \begin{cases} -0.1522x^3 + 0.9937x, & \text{if } x \in [0,1], \\ -0.01258x^3 - 0.4189x^2 + 1.4126x - 0.1396, & \text{if } x \in [1,2], \\ 0.1403x^3 - 1.3359x^2 + 3.2467x - 1.3623, & \text{if } x \in [2,3], \\ 0.1579x^3 - 1.4945x^2 + 3.7225x - 1.8381, & \text{if } x \in [3,4], \\ 0.05375x^3 - 0.2450x^2 - 1.2756x + 4.8259, & \text{if } x \in [4,5], \\ -0.1871x^3 + 3.3673x^2 - 19.3370x + 34.9282, & \text{if } x \in [5,6]. \end{cases}$$

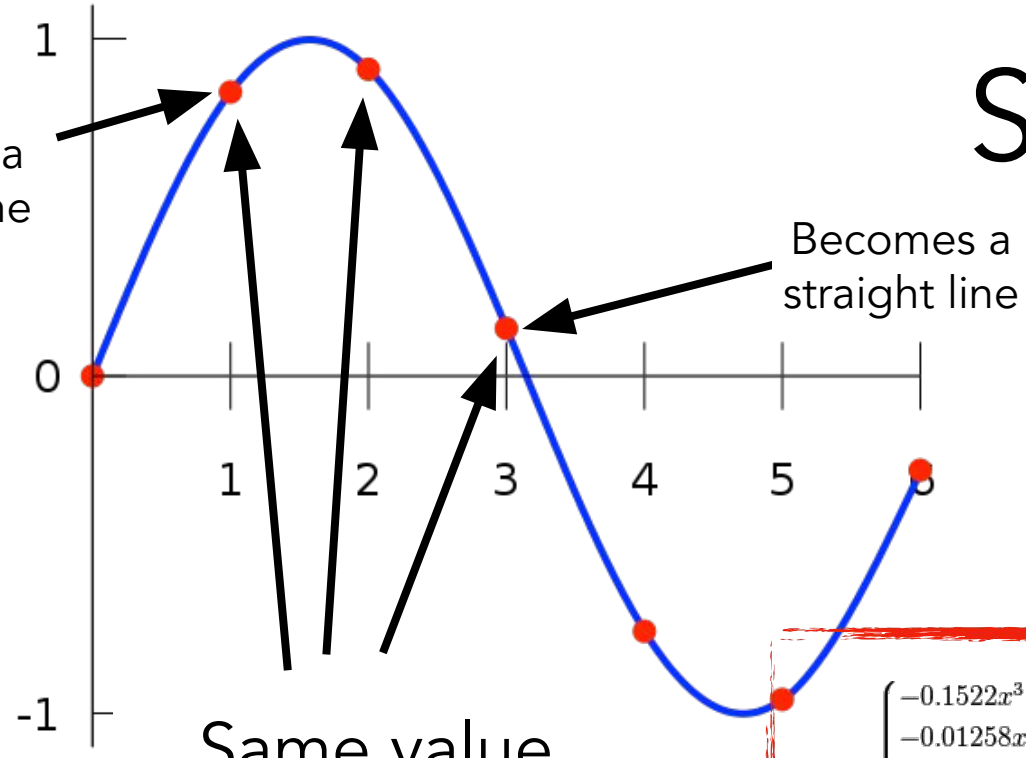
Splines



Same value
Same f′ and f″

$$f(x) = \begin{cases} -0.1522x^3 + 0.9937x, & \text{if } x \in [0,1], \\ -0.01258x^3 - 0.4189x^2 + 1.4126x - 0.1396, & \text{if } x \in [1,2], \\ 0.1403x^3 - 1.3359x^2 + 3.2467x - 1.3623, & \text{if } x \in [2,3], \\ 0.1579x^3 - 1.4945x^2 + 3.7225x - 1.8381, & \text{if } x \in [3,4], \\ 0.05375x^3 - 0.2450x^2 - 1.2756x + 4.8259, & \text{if } x \in [4,5], \\ -0.1871x^3 + 3.3673x^2 - 19.3370x + 34.9282, & \text{if } x \in [5,6]. \end{cases}$$

Splines



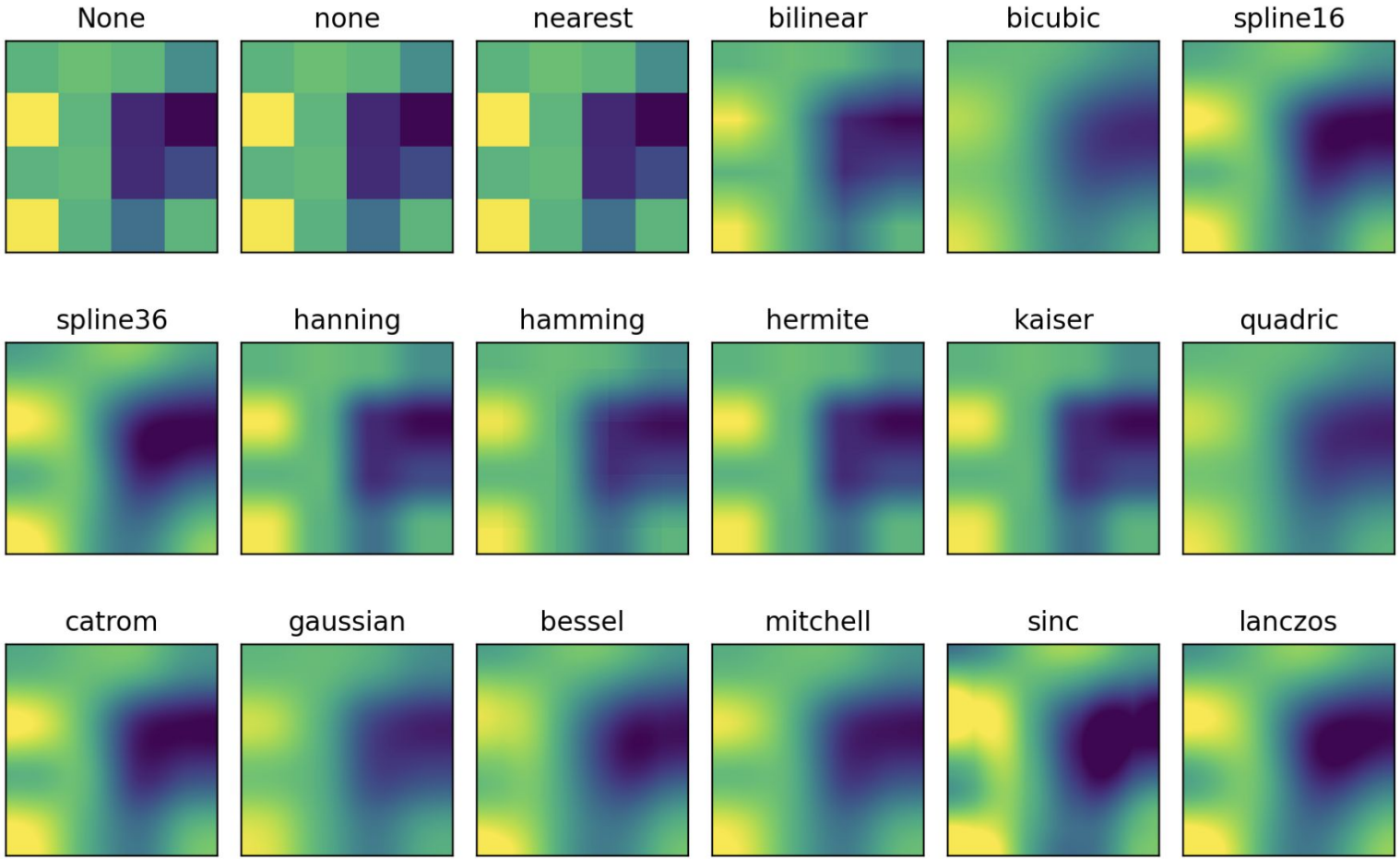
Becomes a straight line

Becomes a straight line

Same value
Same f′ and f″

$$f(x) = \begin{cases} -0.1522x^3 + 0.9937x, & \text{if } x \in [0,1], \\ -0.01258x^3 - 0.4189x^2 + 1.4126x - 0.1396, & \text{if } x \in [1,2], \\ 0.1403x^3 - 1.3359x^2 + 3.2467x - 1.3623, & \text{if } x \in [2,3], \\ 0.1579x^3 - 1.4945x^2 + 3.7225x - 1.8381, & \text{if } x \in [3,4], \\ 0.05375x^3 - 0.2450x^2 - 1.2756x + 4.8259, & \text{if } x \in [4,5], \\ -0.1871x^3 + 3.3673x^2 - 19.3370x + 34.9282, & \text{if } x \in [5,6]. \end{cases}$$

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



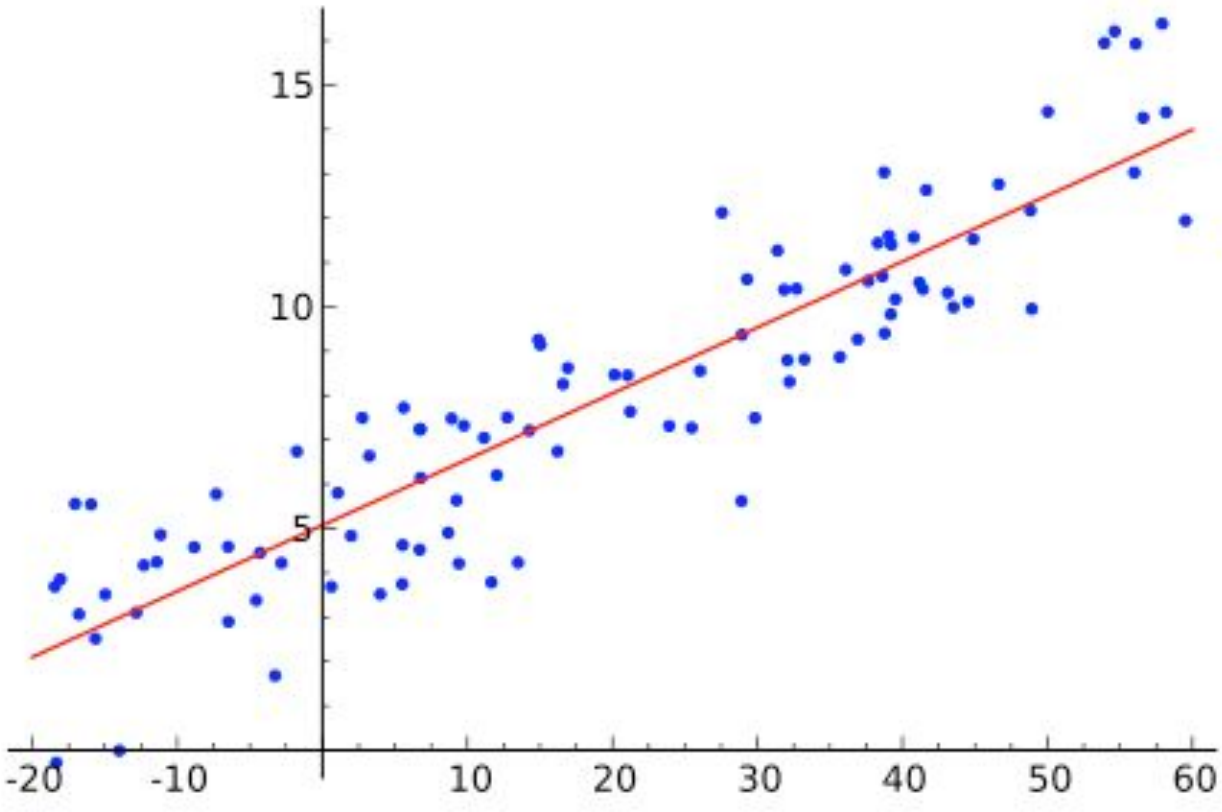
https://matplotlib.org/stable/gallery/images_contours_and_fields/interpolation_methods.html

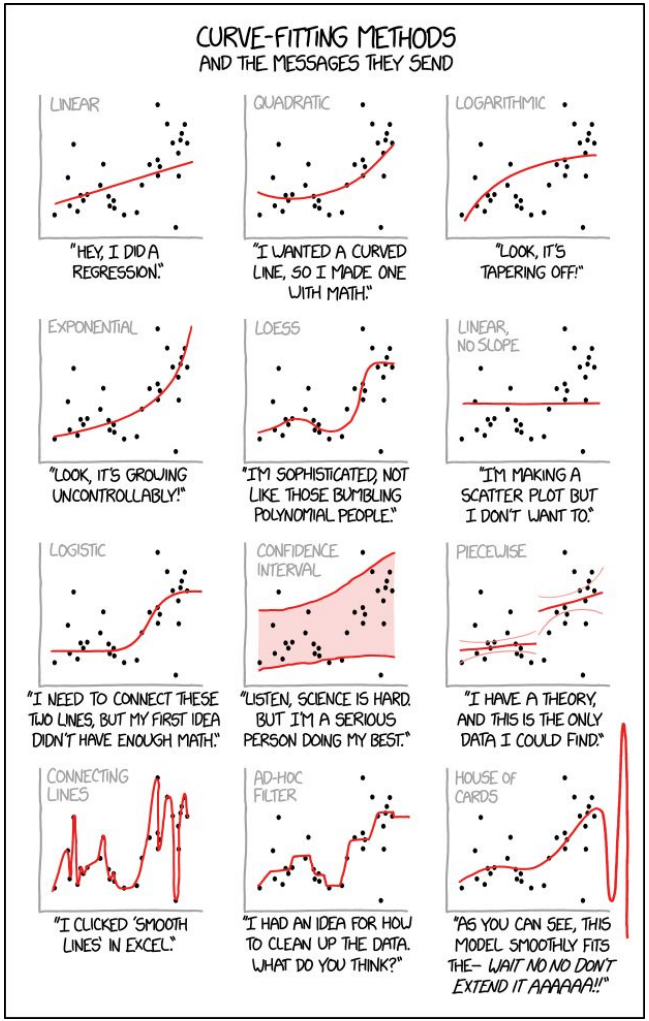
Regression

Ok. Interpolation “connects” the dots. But what if we have some noise?

How can we visualize the **trends**?

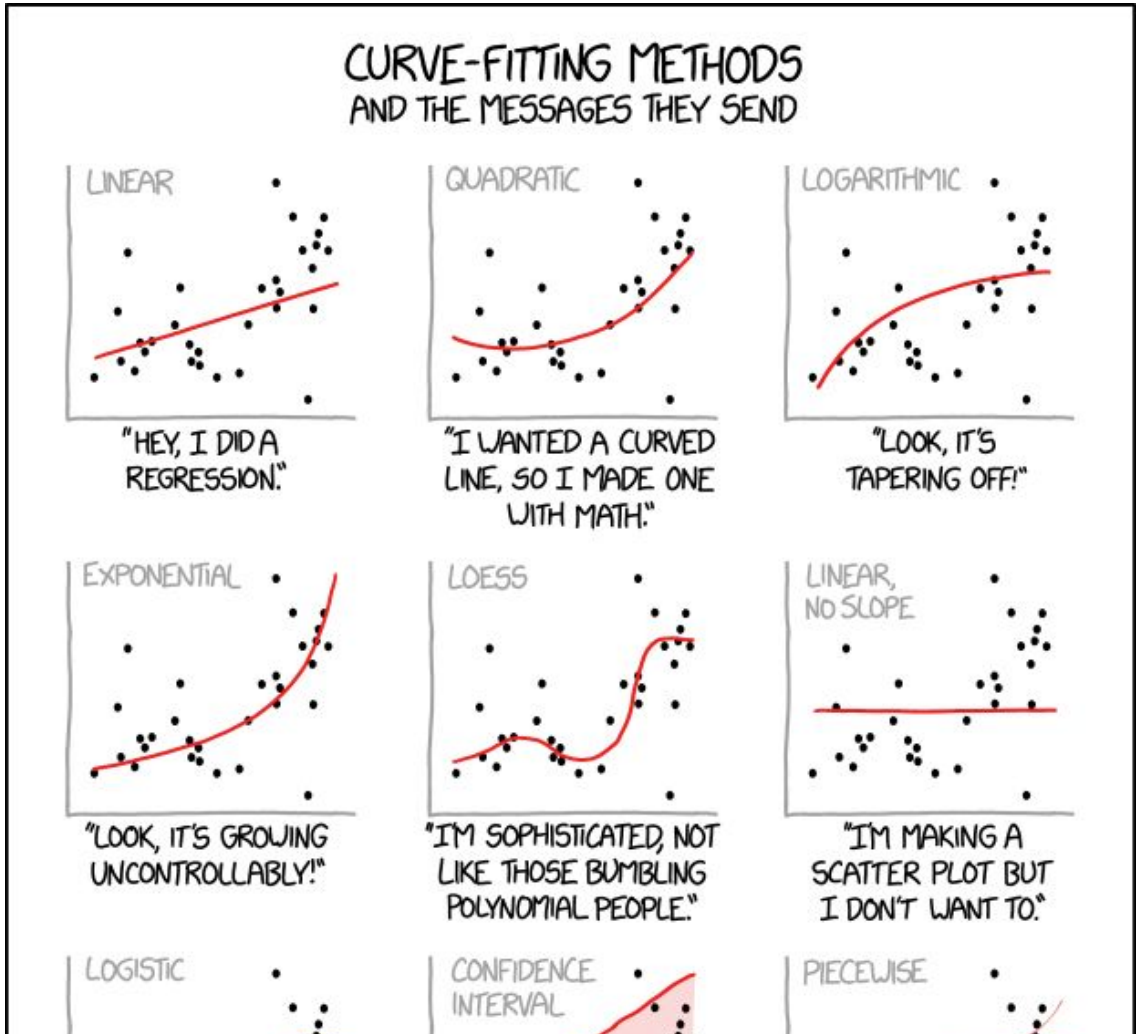
Regression: We assume a ‘model’ and find the best fit of the parameters of the model to the data.



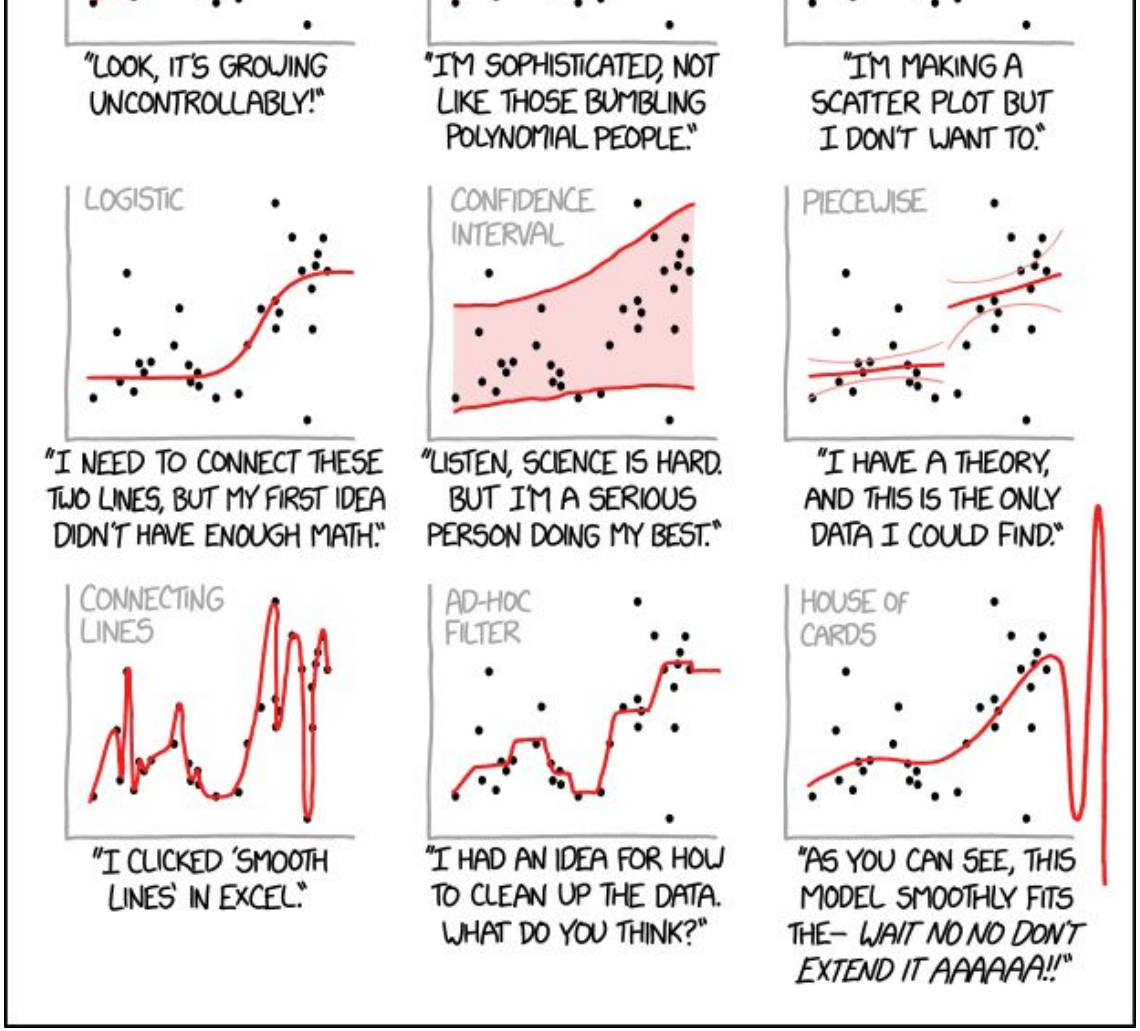


When taking a parametric approach, our "**assumptions**" (or simply **models**) may skew how data gets interpreted.

https://www.xkcd.com/2048/



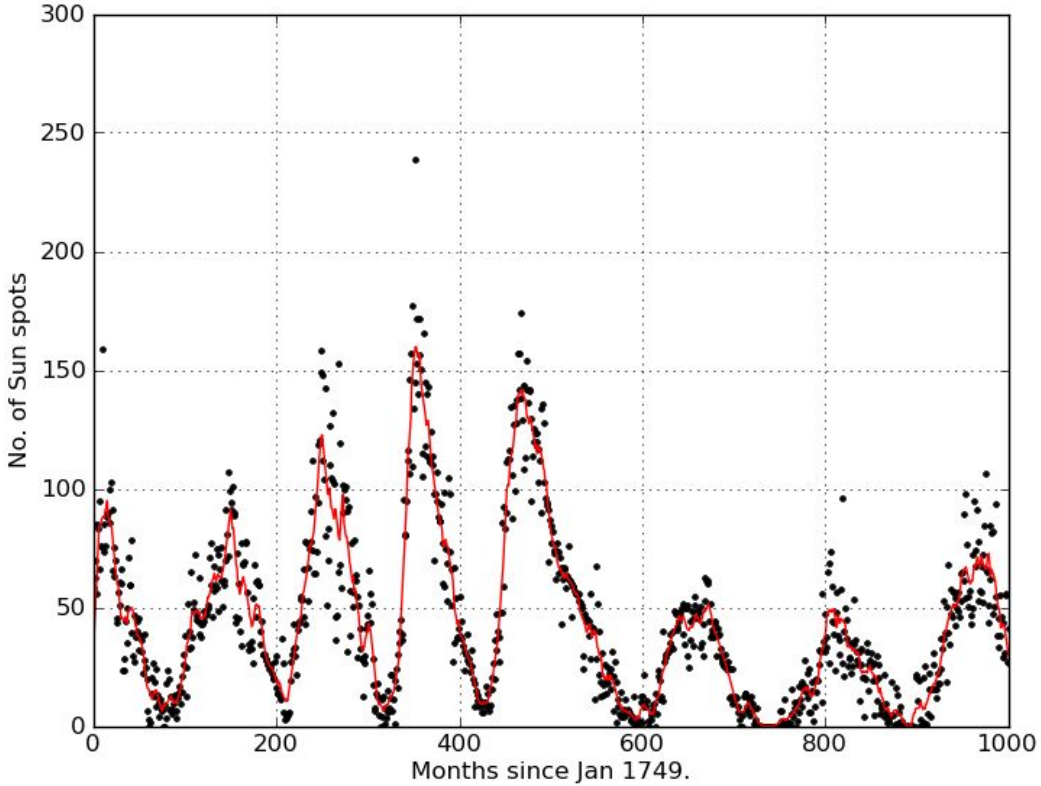
https://www.xkcd.com/2048/



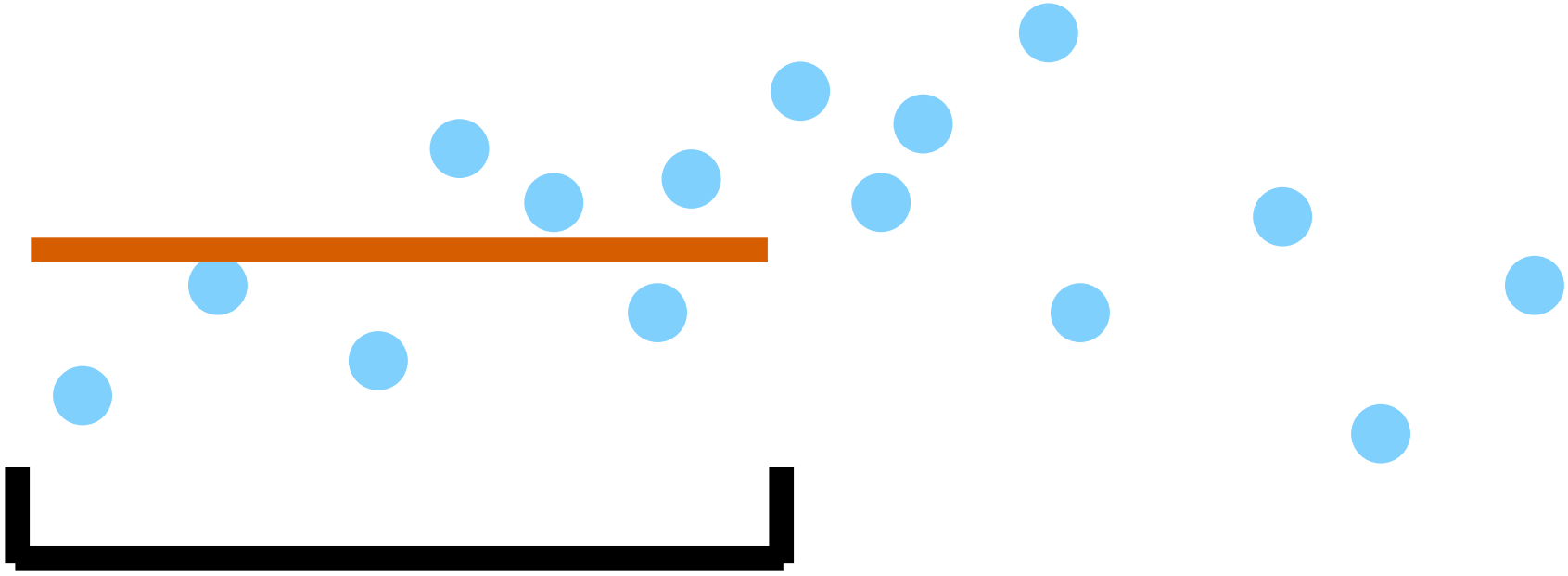
https://www.xkcd.com/2048/

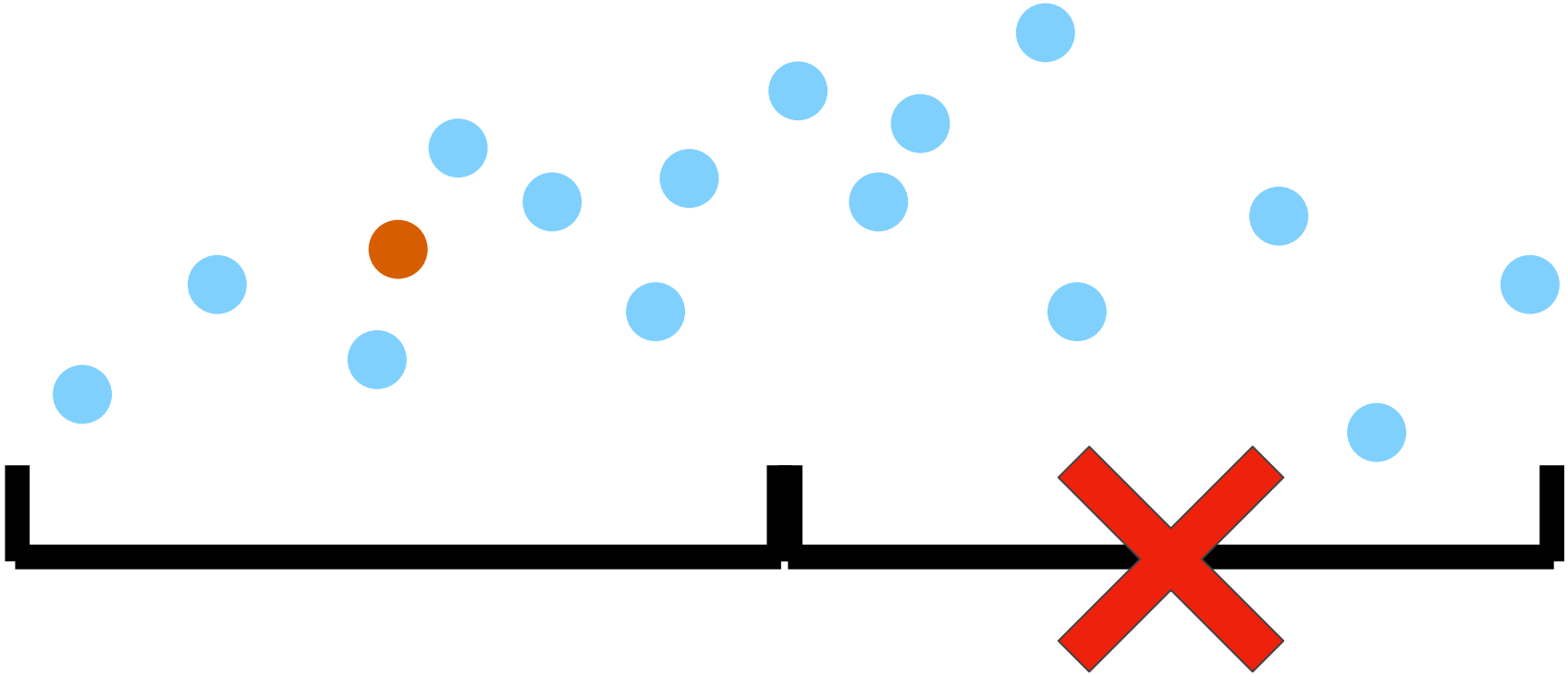
What would be non-parametric (data-centric) version?

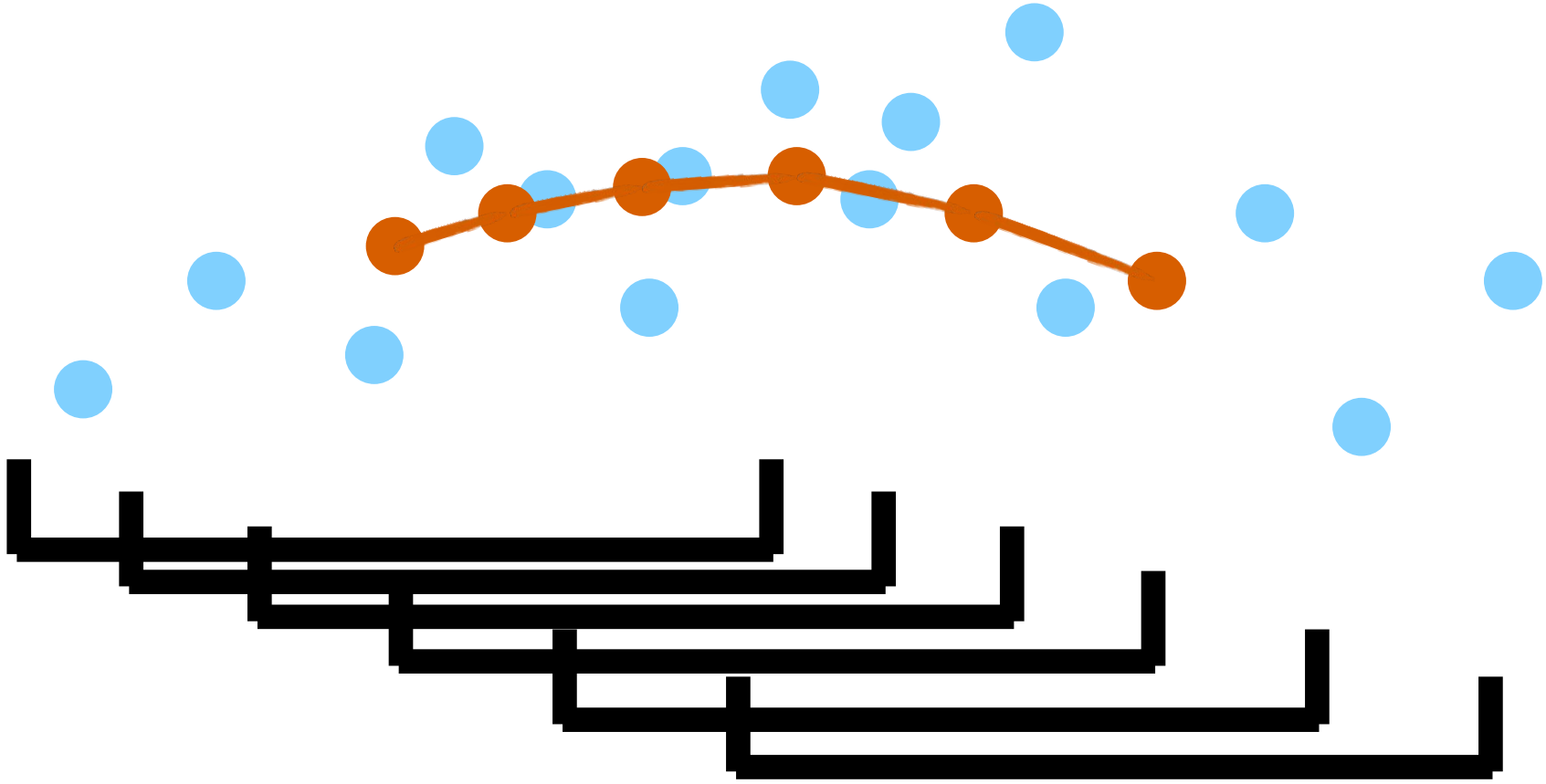
Moving average: a non-parametric approach



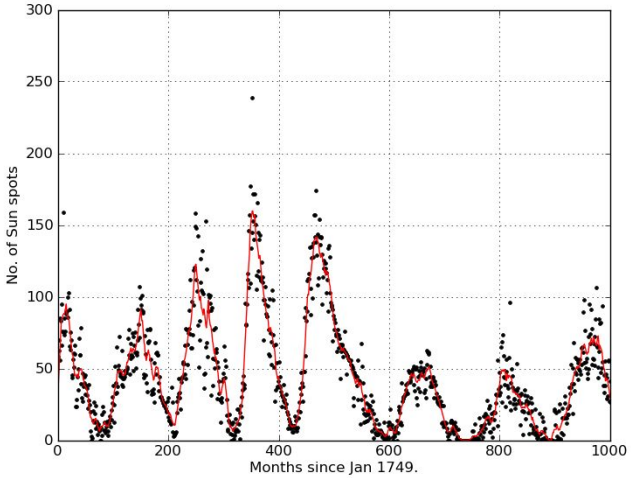
https://stackoverflow.com/questions/11352047/finding-moving-average-from-data-points-in-python

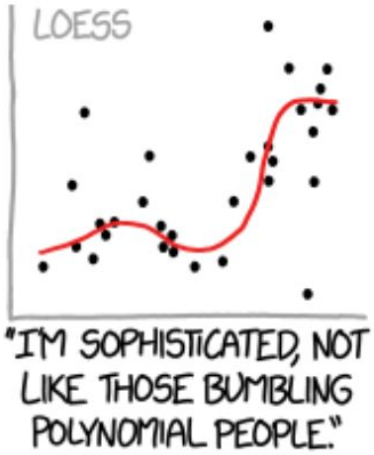






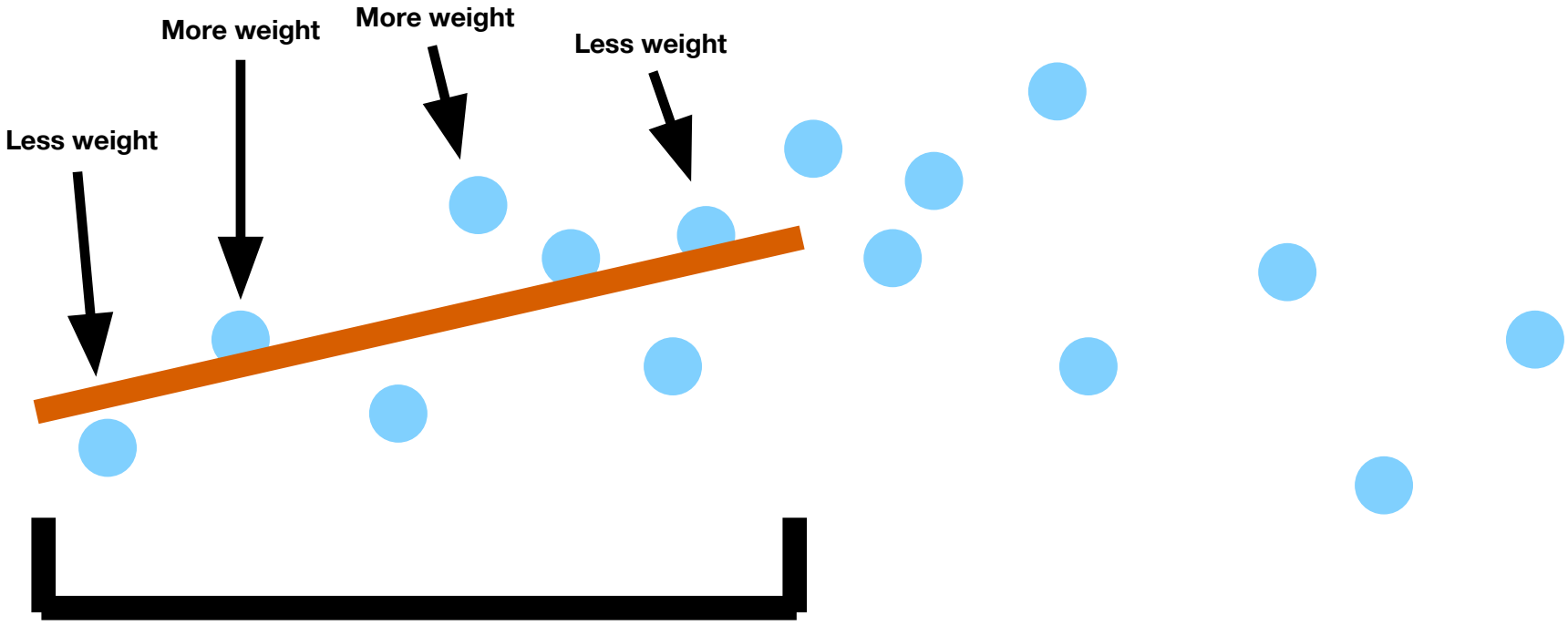
Moving average vs. LOESS (locally estimated scatterplot smoothing)

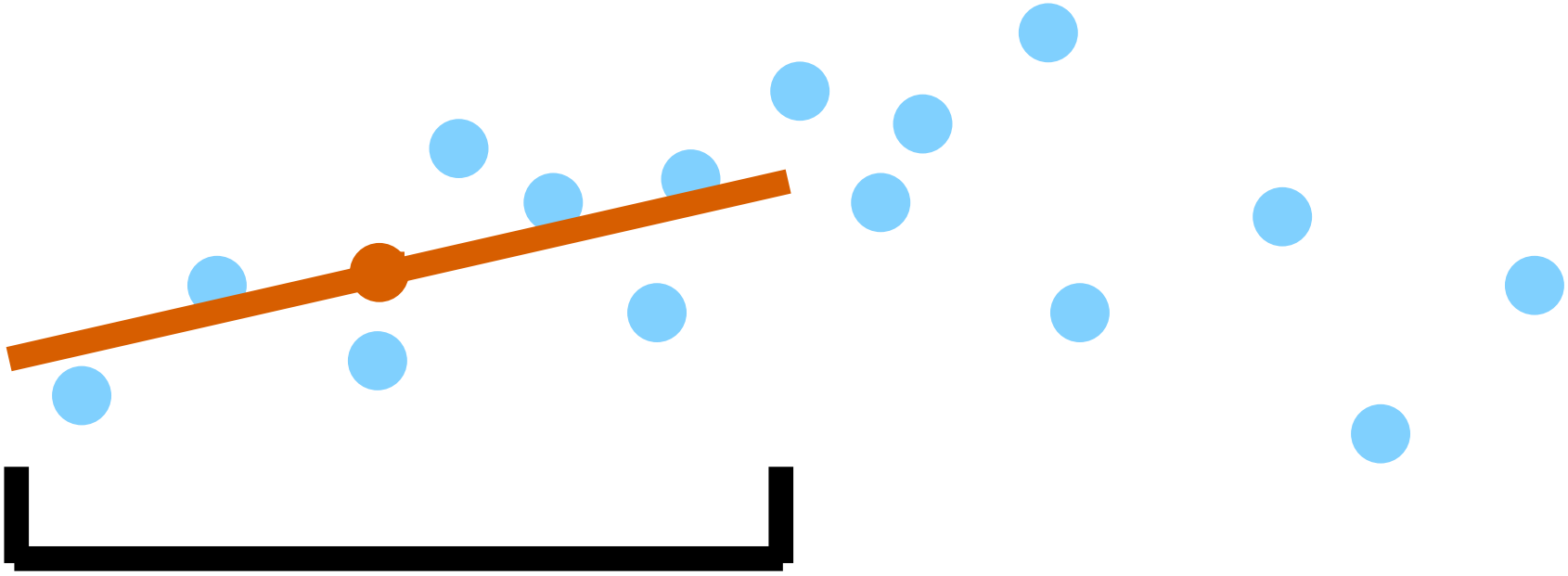


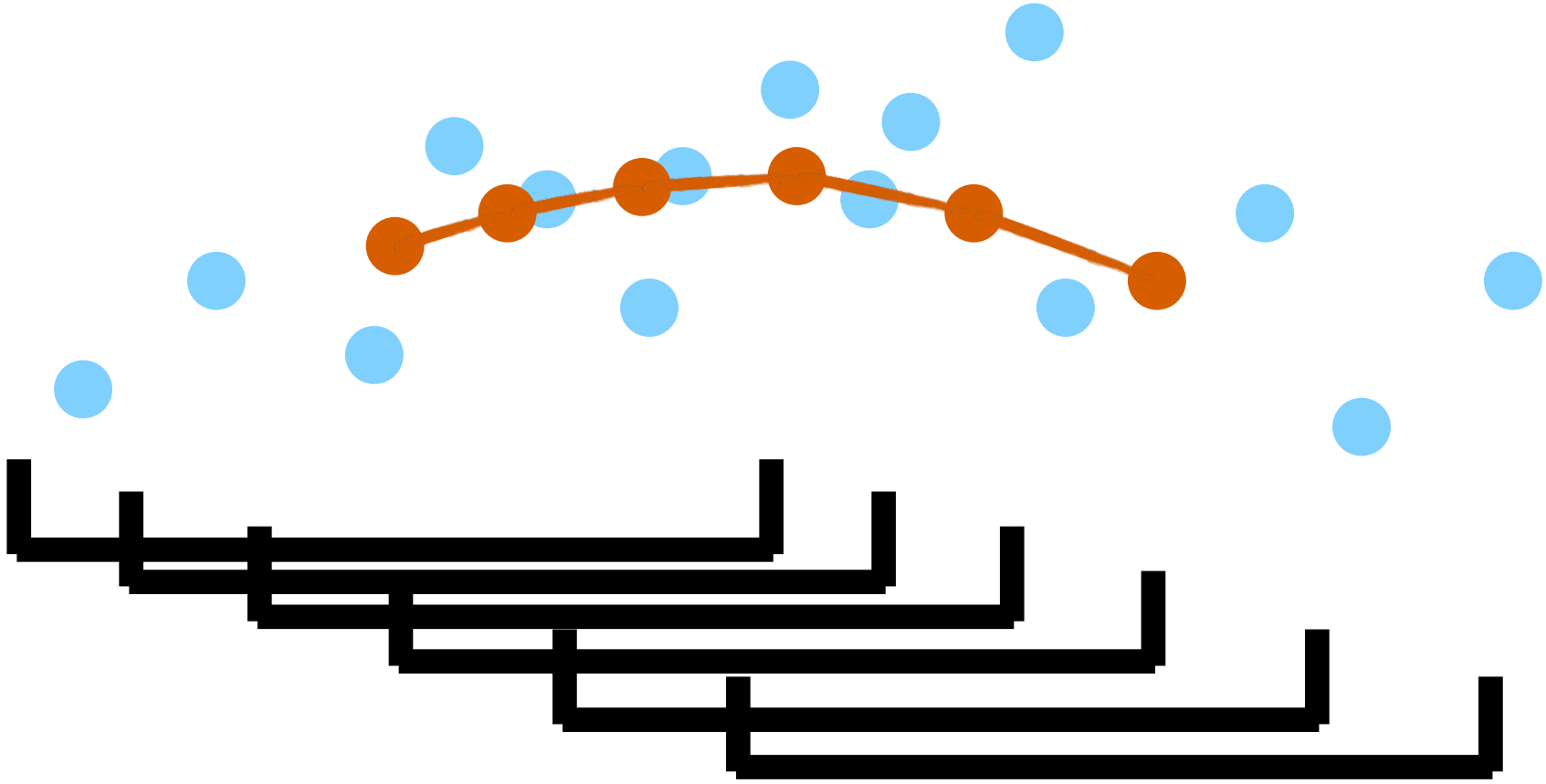


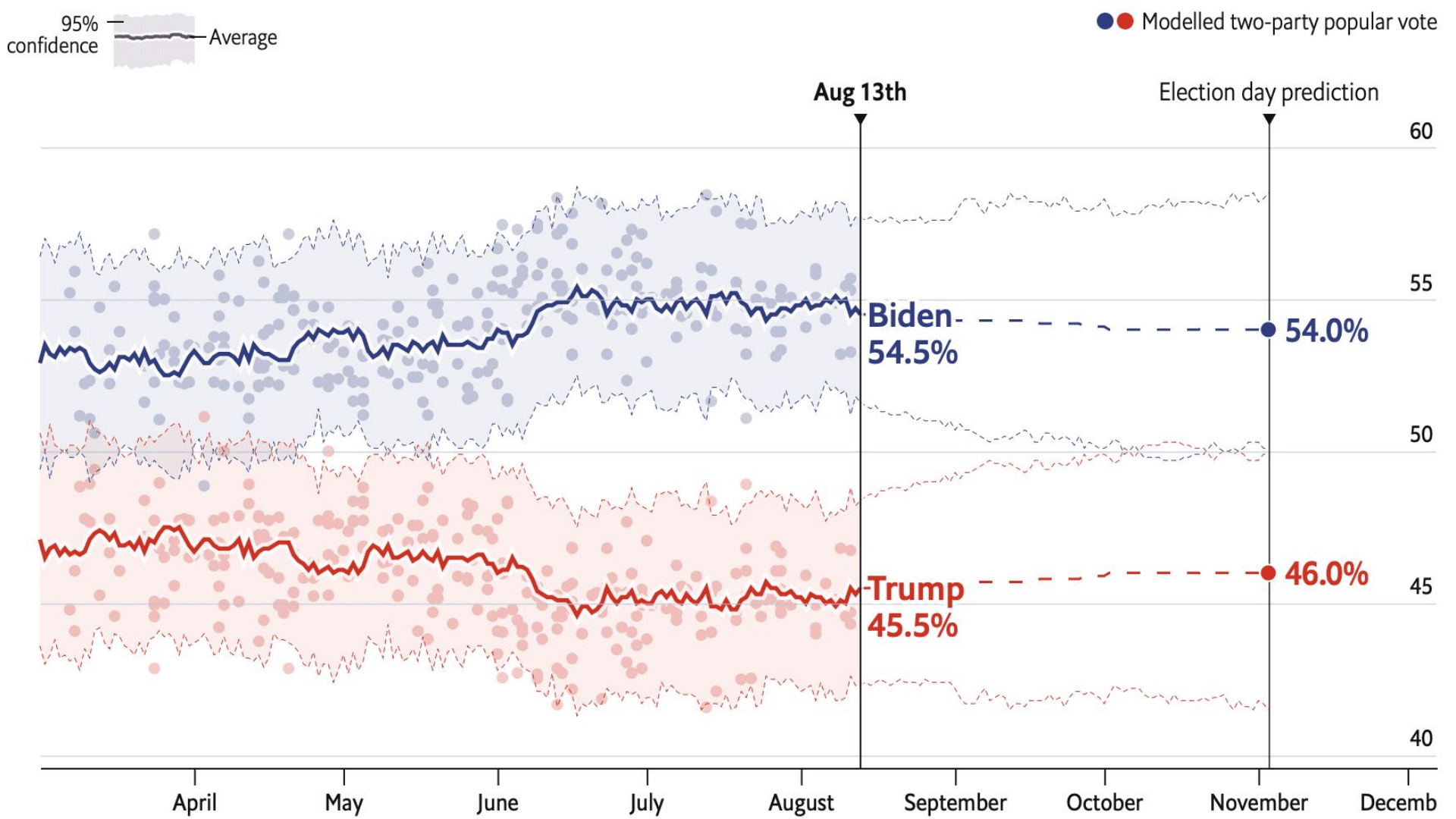
LOESS/LOWESS (local regression)

- LOESS (locally estimated scatterplot smoothing)
- Instead of taking the **average**, perform a (locally weighted) **regression.**


Less weight
More weight
More weight
Less weight



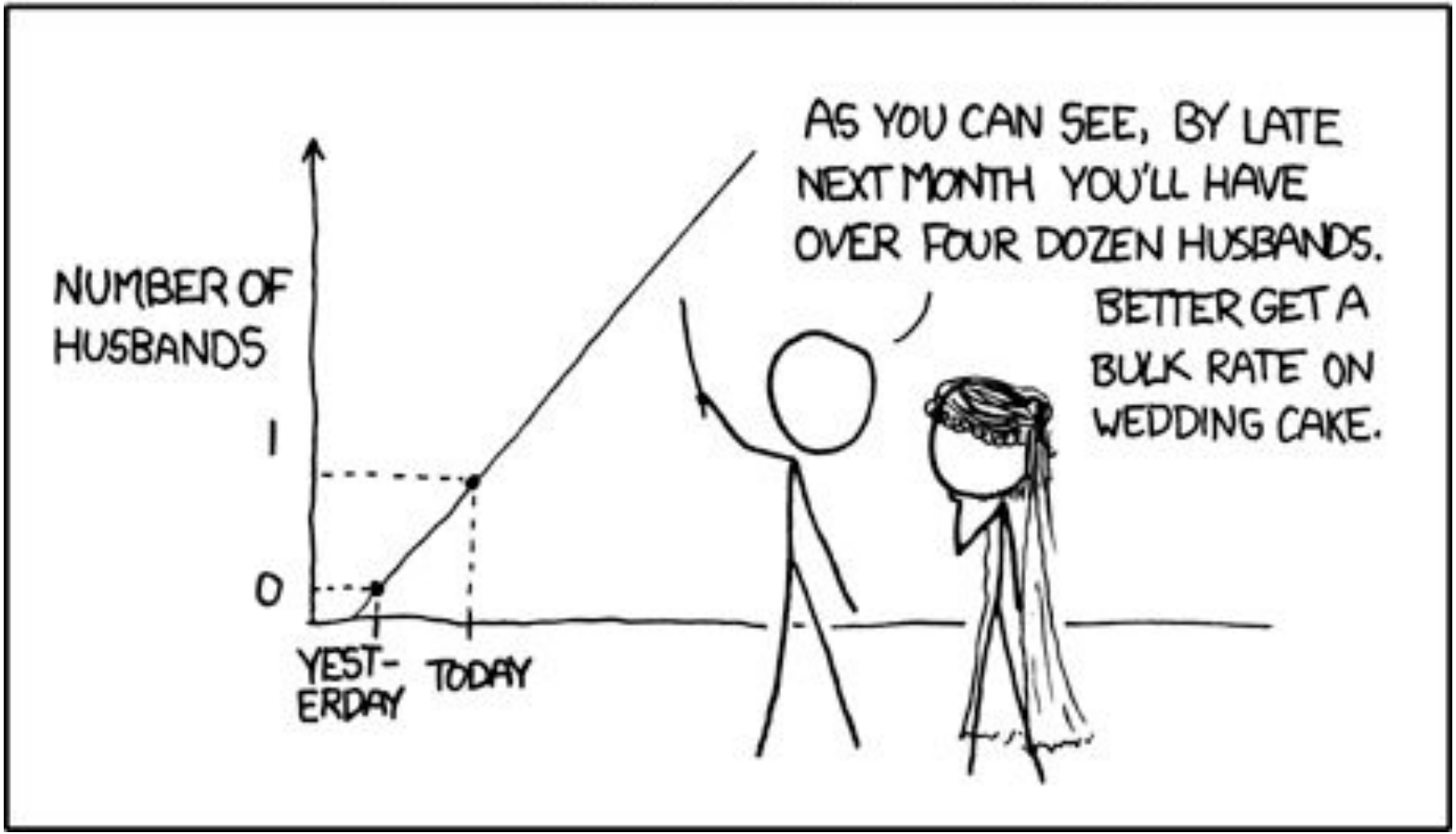



95% confidence
Average
Modelled two-party popular vote
Aug 13th
Election day prediction
60
55
50
45
40
Biden
54.5%
54.0%
Trump
45.5%
46.0%
April
May
June
July
August
September
October
November
Decemb

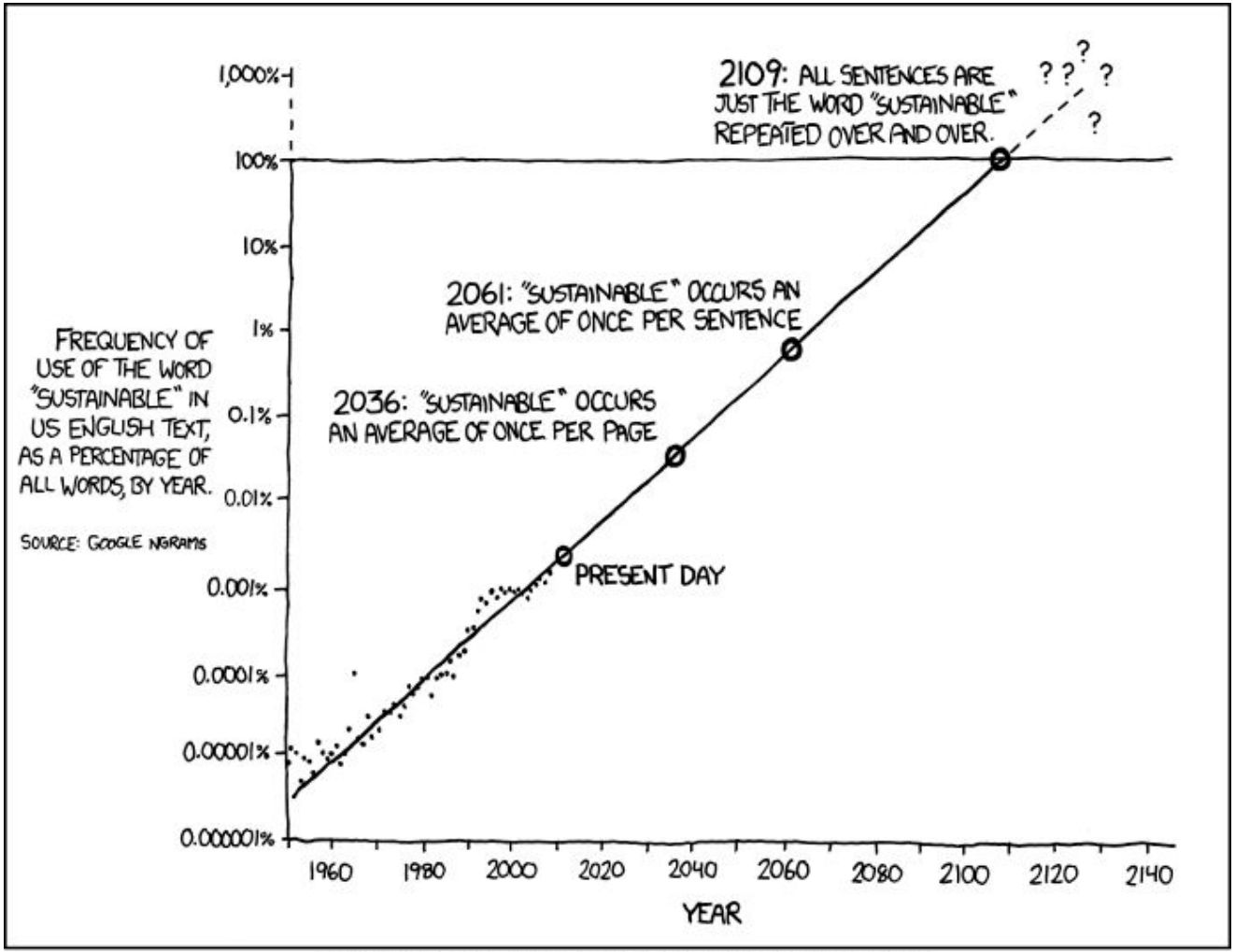
Extrapolation

Let’s extend the trend we have seen.
(Dangerous!)

MY HOBBY: EXTRAPOLATING



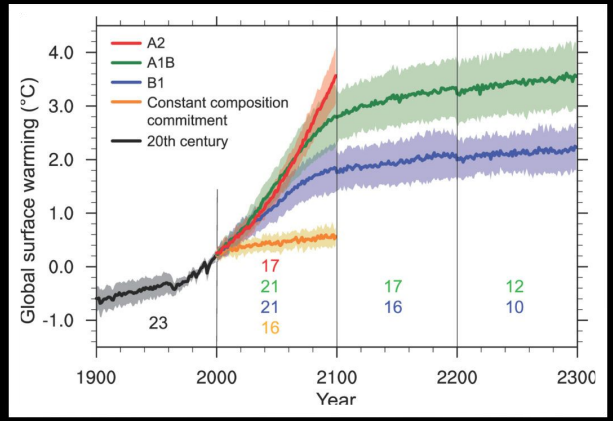
xkcd

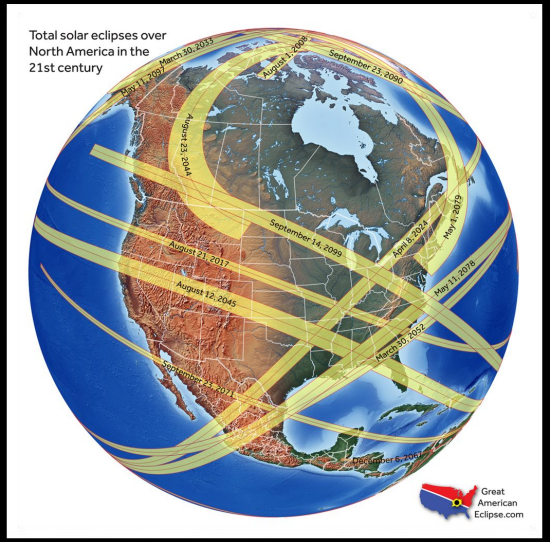


THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

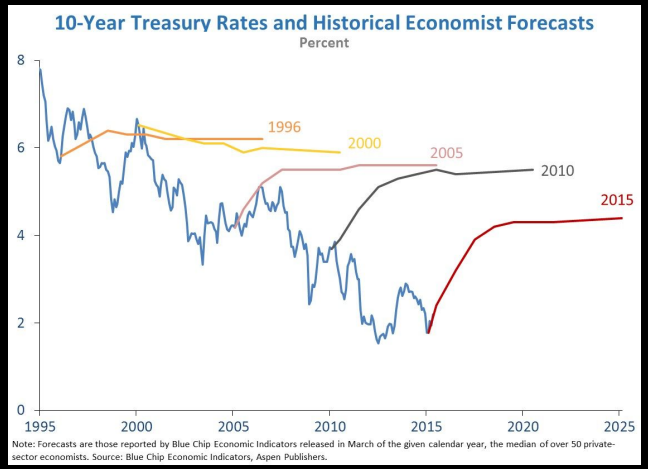
xkcd

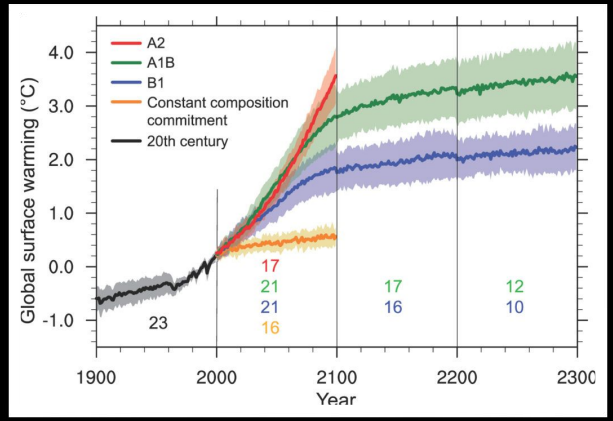
Extrapolation? Prediction?

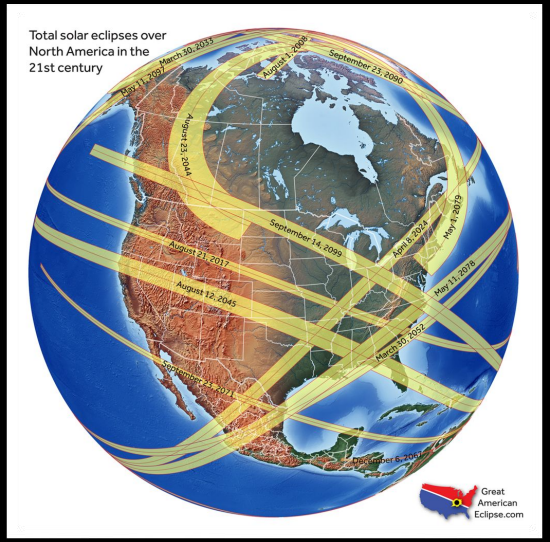




Extrapolation? Prediction?







There is a huge gamut from crazy extrapolations to extremely accurate predictions, largely depending on our understanding of underlying mechanisms.