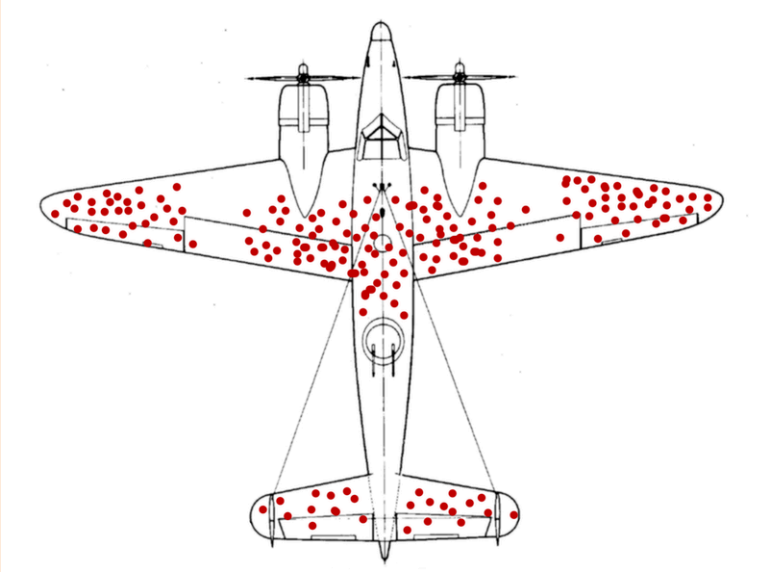
CPSC 340:
Machine Learning and Data Mining

L2Regularization

Bonus Slides

Related: Survivorship Bias

- Plotting location of bullet holes on planes returning from WW2:

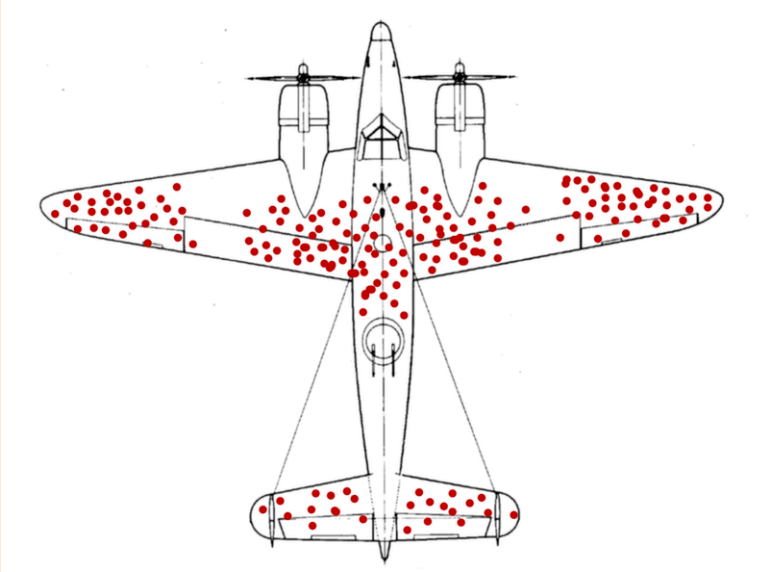


- Where are the “relevant” parts of the plane to protect?
  - “Relevant” parts are actually where there are no bullets.
  - Planes shot in other places did not come back (armor was needed).

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

Related: Survivorship Bias

- Plotting location of bullet holes on planes returning from WW2:

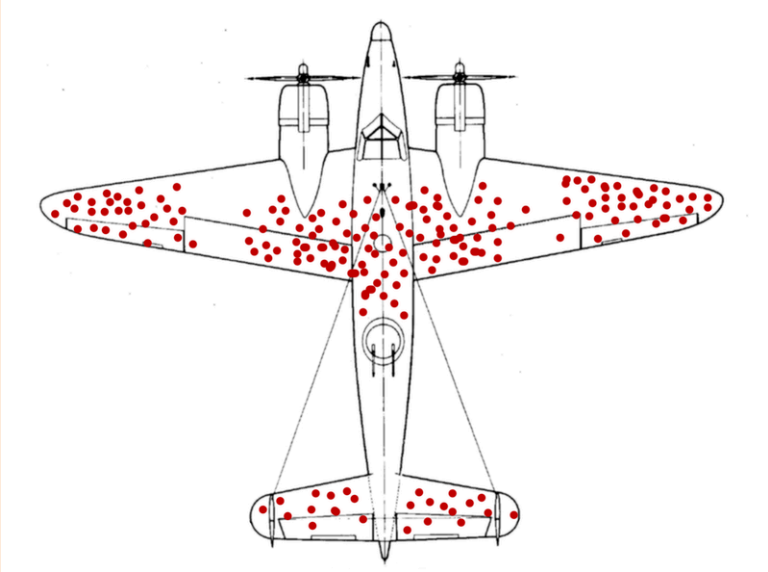


- This is an example of "survivorship bias":
  - Data is not IID because you only sample the "survivors".
  - Causes havoc for feature selection, and ML methods in general.

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

Related: Survivorship Bias

- Plotting location of bullet holes on planes returning from WW2:



- People come to wrong conclusions due to survivor bias all the time.
  - Article on “secrets of success”, focusing on traits of successful people.
    - But ignoring the number of non-super-successful people with the same traits.
  - Article hypothesizing about various topics (allergies, mental illness, etc.).

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

Why use L2-Regularization?

- It's a weird thing to do, but Mark says "always use regularization".
  - "Almost always decreases test error" should already convince you.

- But here are 6 more reasons:
  1. Solution 'w' is unique.
  2. X^TX does not need to be invertible (no collinearity issues).
  3. Less sensitive to changes in X or y.
  4. Gradient descent converges faster (bigger λ means fewer iterations).
  5. Stein's paradox: if $d \geq 3$, 'shrinking' moves us closer to 'true' w.
  6. Worst case: just set λ small and get the same performance.

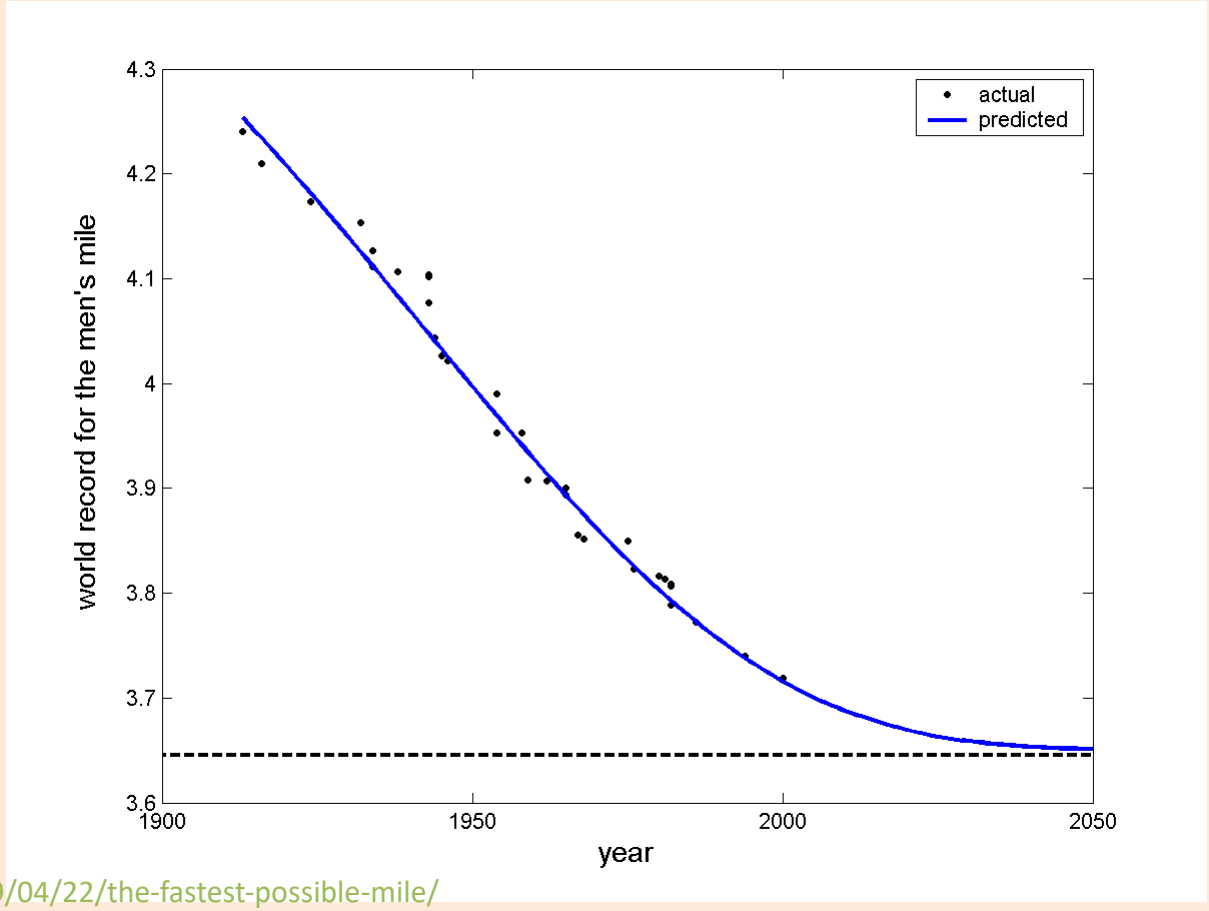
Regularizing the y-Intercept?

- Should we regularize the y-intercept?
- No! Why encourage it to be closer to zero? (It could be anywhere.)
  - You should be allowed to shift function up/down globally.
- Yes! It makes the solution unique and it easier to compute 'w'.
- Compromise: regularize by a smaller amount than other variables.

$$f(w, w_0) = \frac{1}{2}||Xw + w_0 - y||^2 + \frac{\lambda}{2}||w||^2 + \frac{\lambda_0}{2}w_0^2$$

Predicting the Future

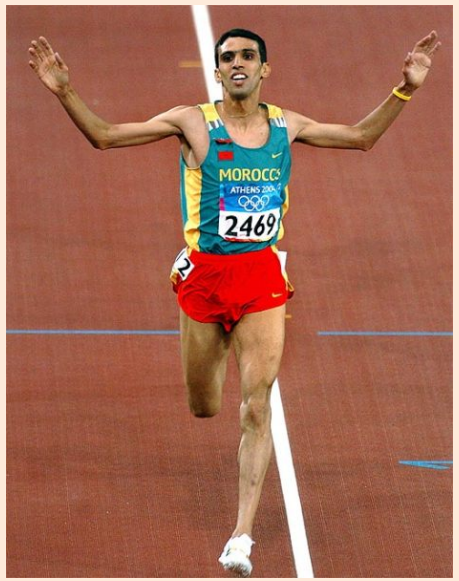
- In principle, we can use any features x_i that we think are relevant.
- This makes it tempting to use time as a feature, and predict future.

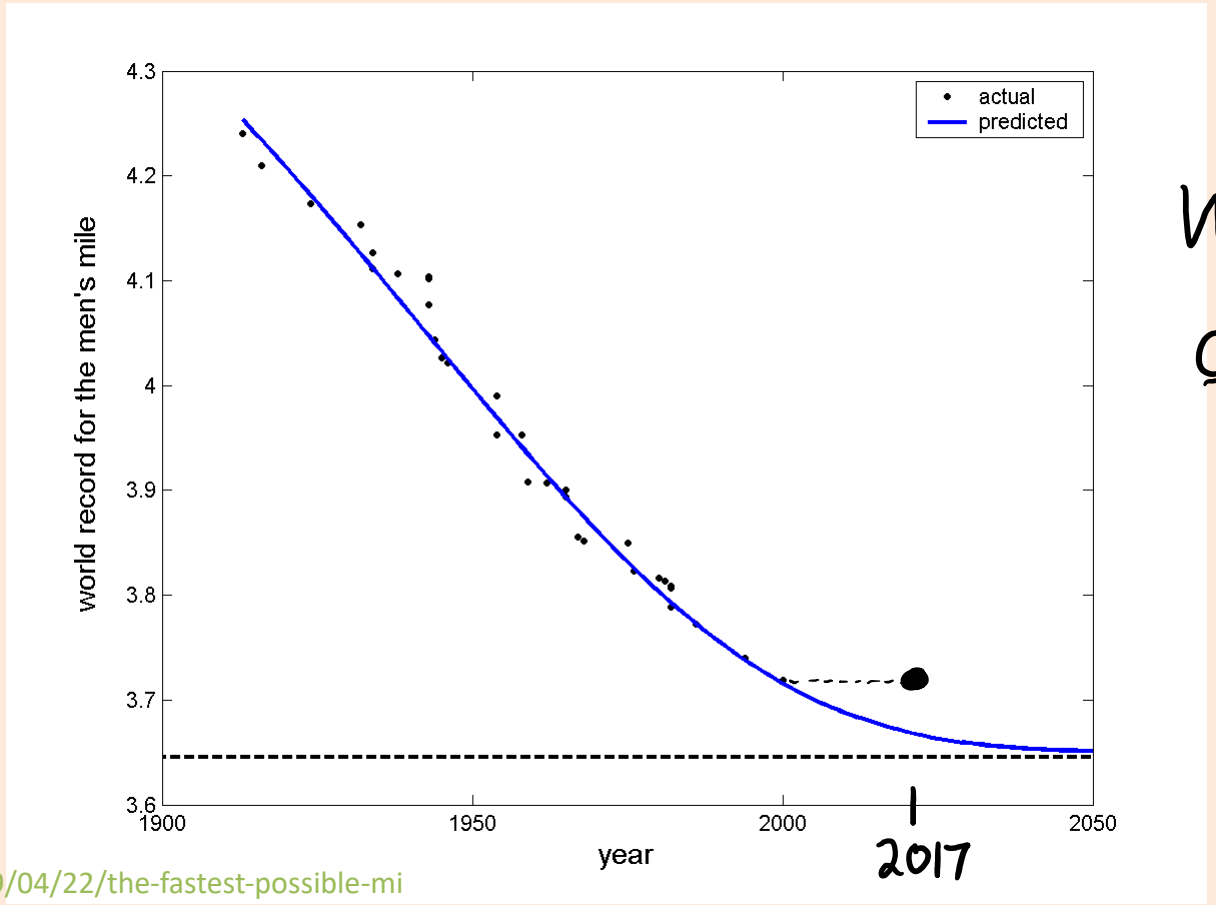


https://gravityandlevity.wordpress.com/2009/04/22/the-fastest-possible-mile/

Predicting the Future

- In principle, we can use any features x_i that we think are relevant.
- This makes it tempting to use time as a feature, and predict future.

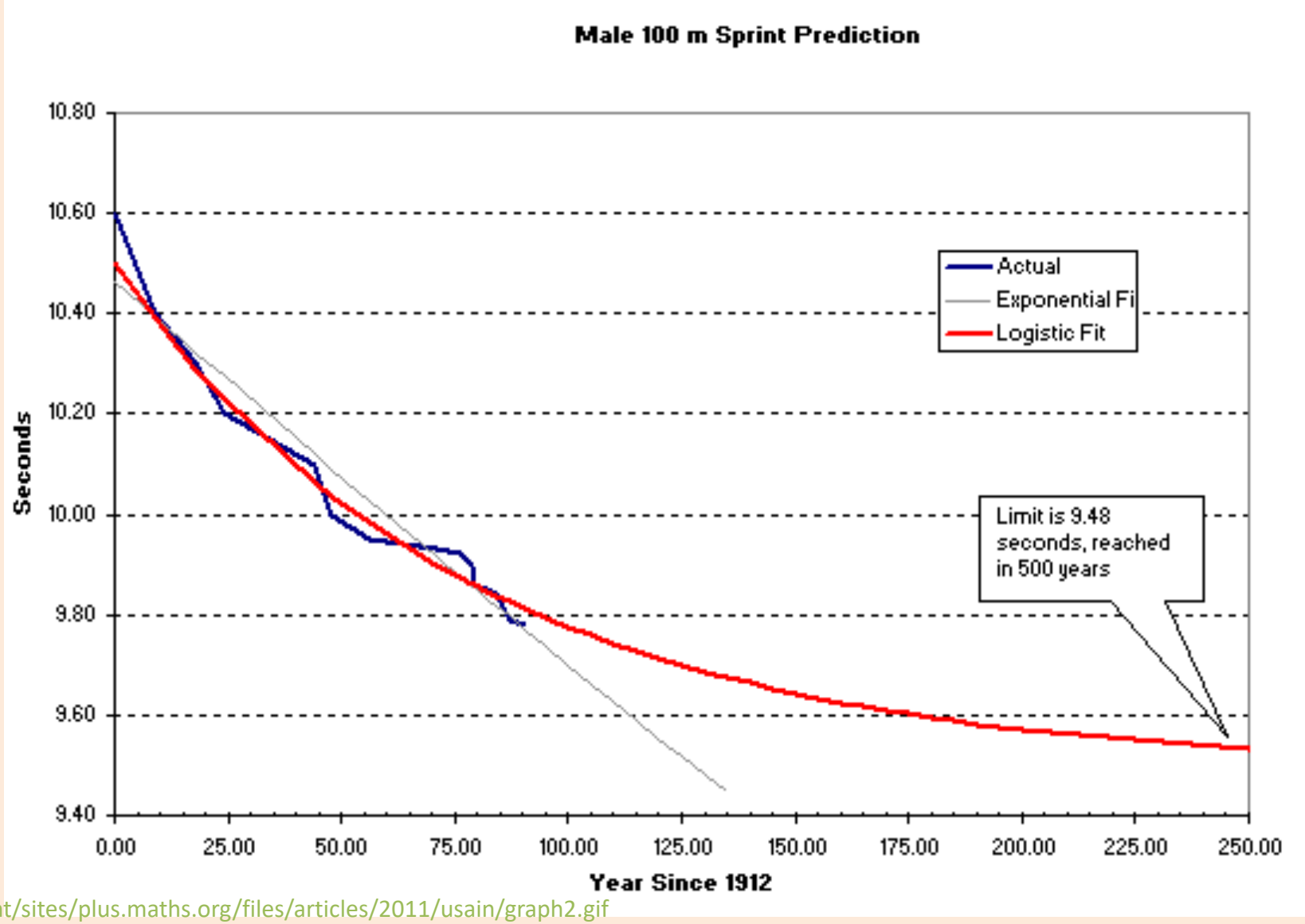




We need to be cautious about doing this.

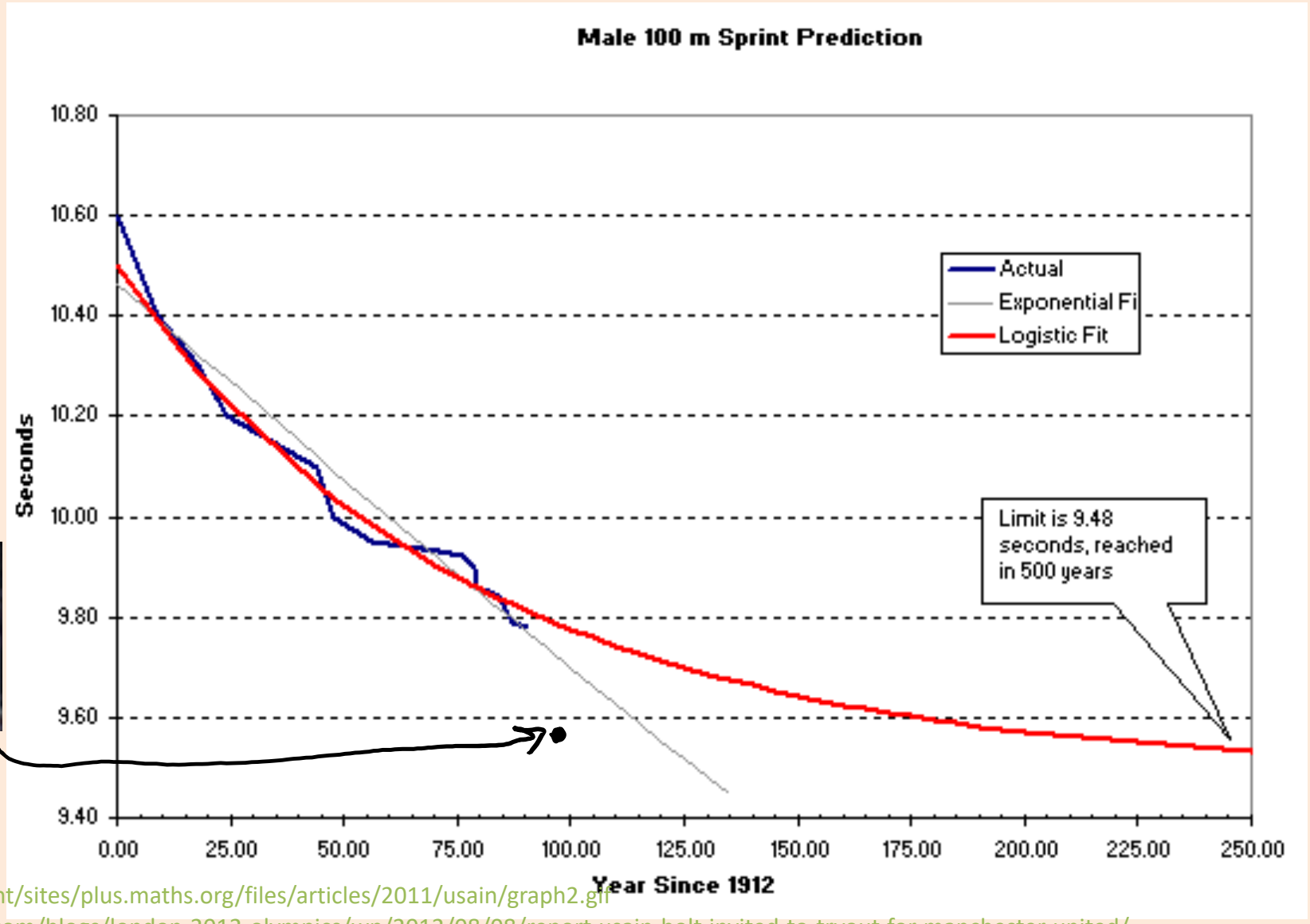
https://gravityandlevity.wordpress.com/2009/04/22/the-fastest-possible-mi
https://overthehillsports.wordpress.com/tag/hicham-el-guerrouj/le/

Predicting 100m times 400 years in the future?



https://plus.maths.org/content/sites/plus.maths.org/files/articles/2011/usain/graph2.gif

Predicting 100m times 400 years in the future?



9.58

https://plus.maths.org/content/sites/plus.maths.org/files/articles/2011/usain/graph2.gif

http://www.washingtonpost.com/blogs/london-2012-olympics/wp/2012/08/08/report-usain-bolt-invited-to-tryout-for-manchester-united/

Interpolation vs Extrapolation

- Interpolation is task of predicting “between the data points”.
  - Regression models are good at this if you have enough data and function is continuous.
- Extrapolation is task of prediction outside the range of the data points.
  - Without assumptions, regression models can be embarrassingly-bad at this.

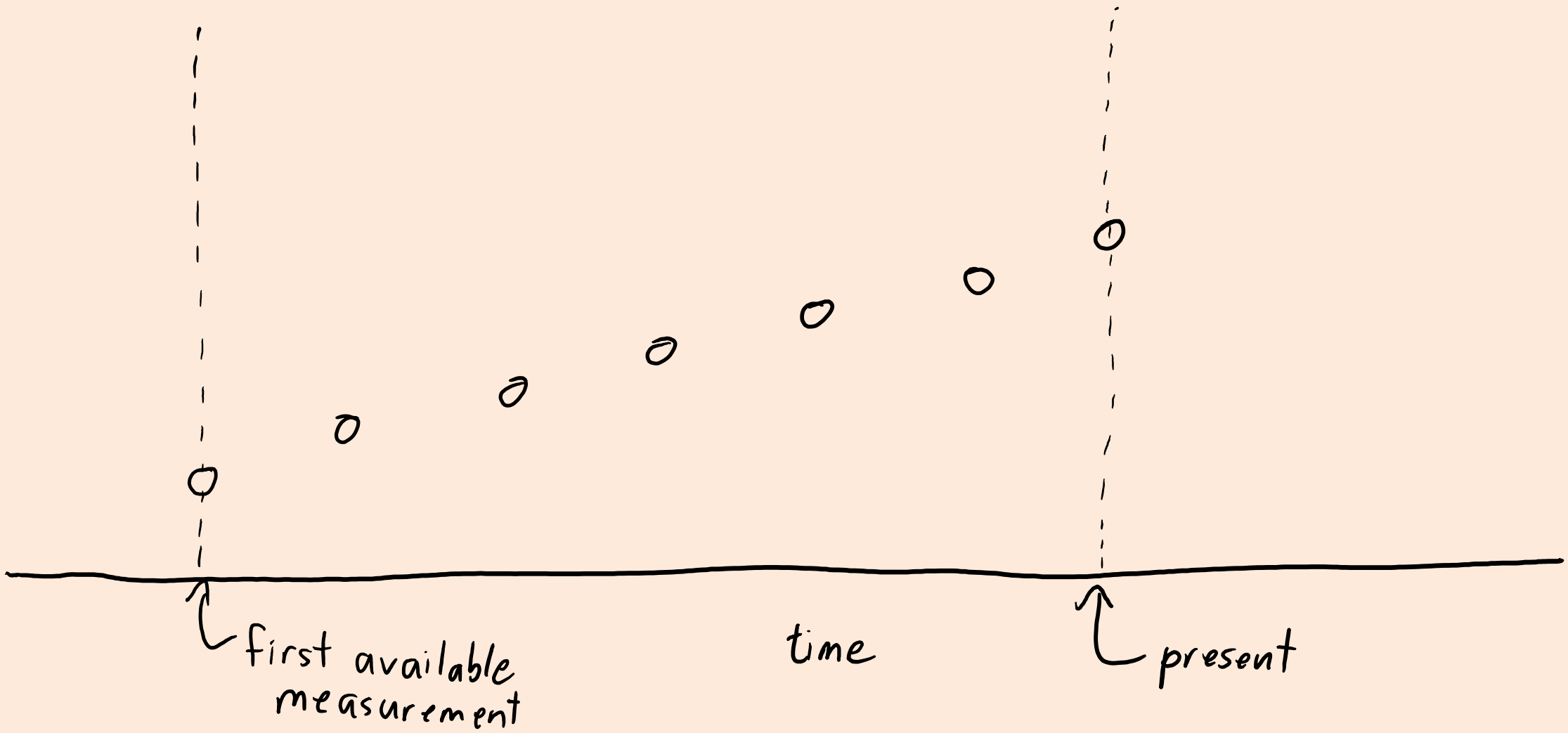
- If you run the 100m regression models backwards in time:
  - They predict that humans used to be really really slow!

- If you run the 100m regression models forwards in time:
  - They might eventually predict arbitrarily-small 100m times.
  - The linear model actually predicts negative times in the future.
    - These time traveling races in 2060 should be pretty exciting!

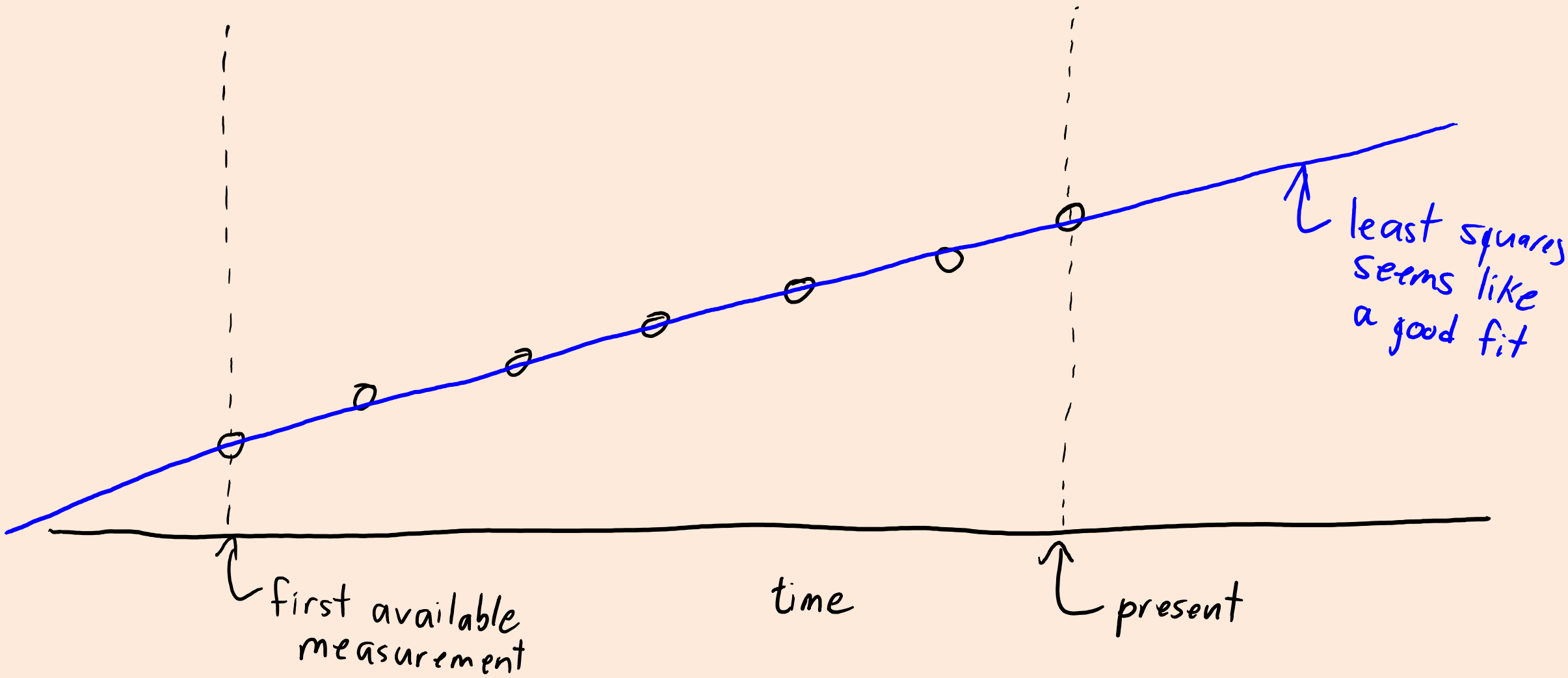
- Some discussion here:
  - http://callingbullshit.org/case_studies/case_study_gender_gap_running.html

https://www.smbc-comics.com/comic/rise-of-the-machines

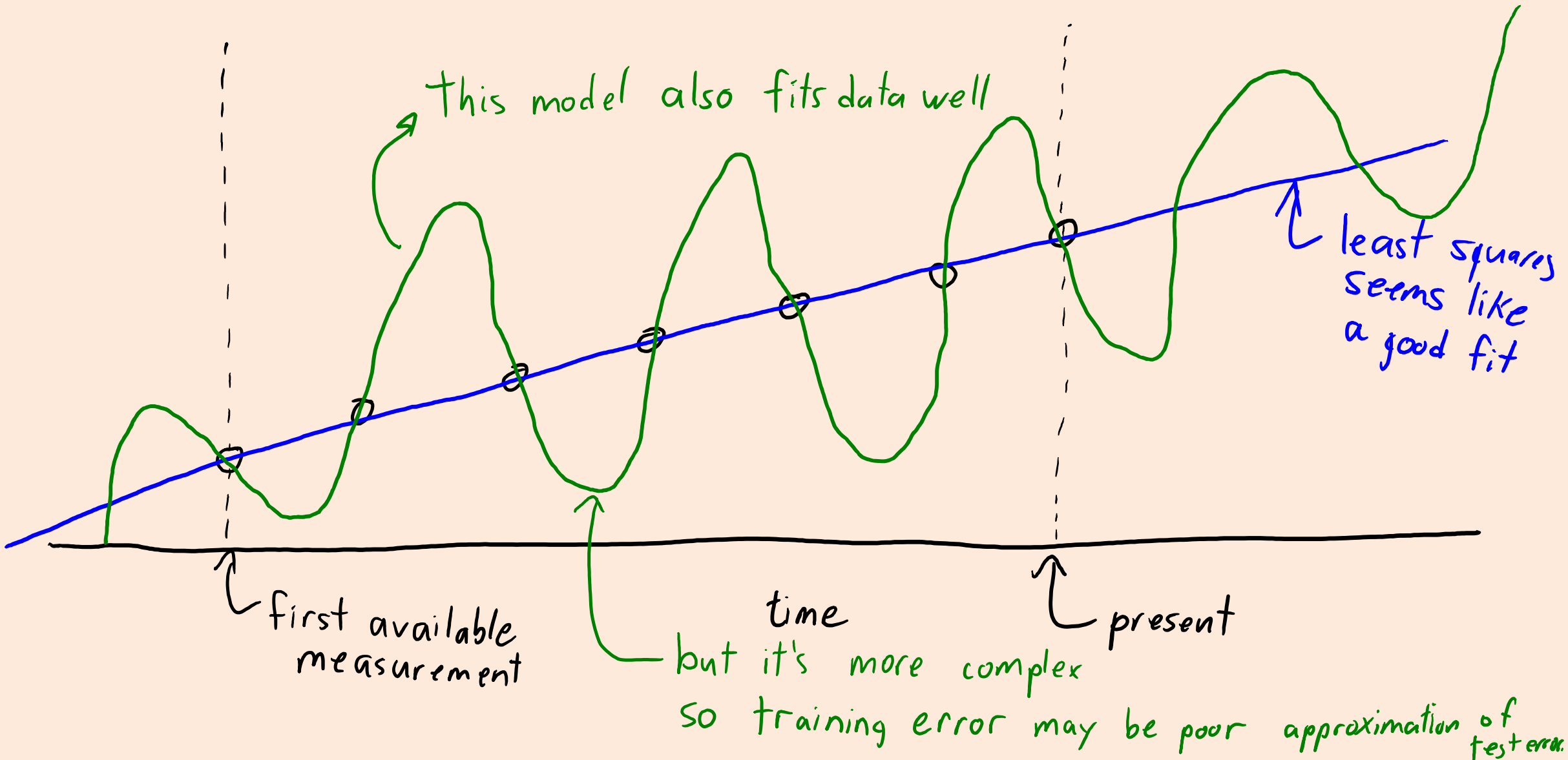
No Free Lunch, Consistency, and the Future



No Free Lunch, Consistency, and the Future

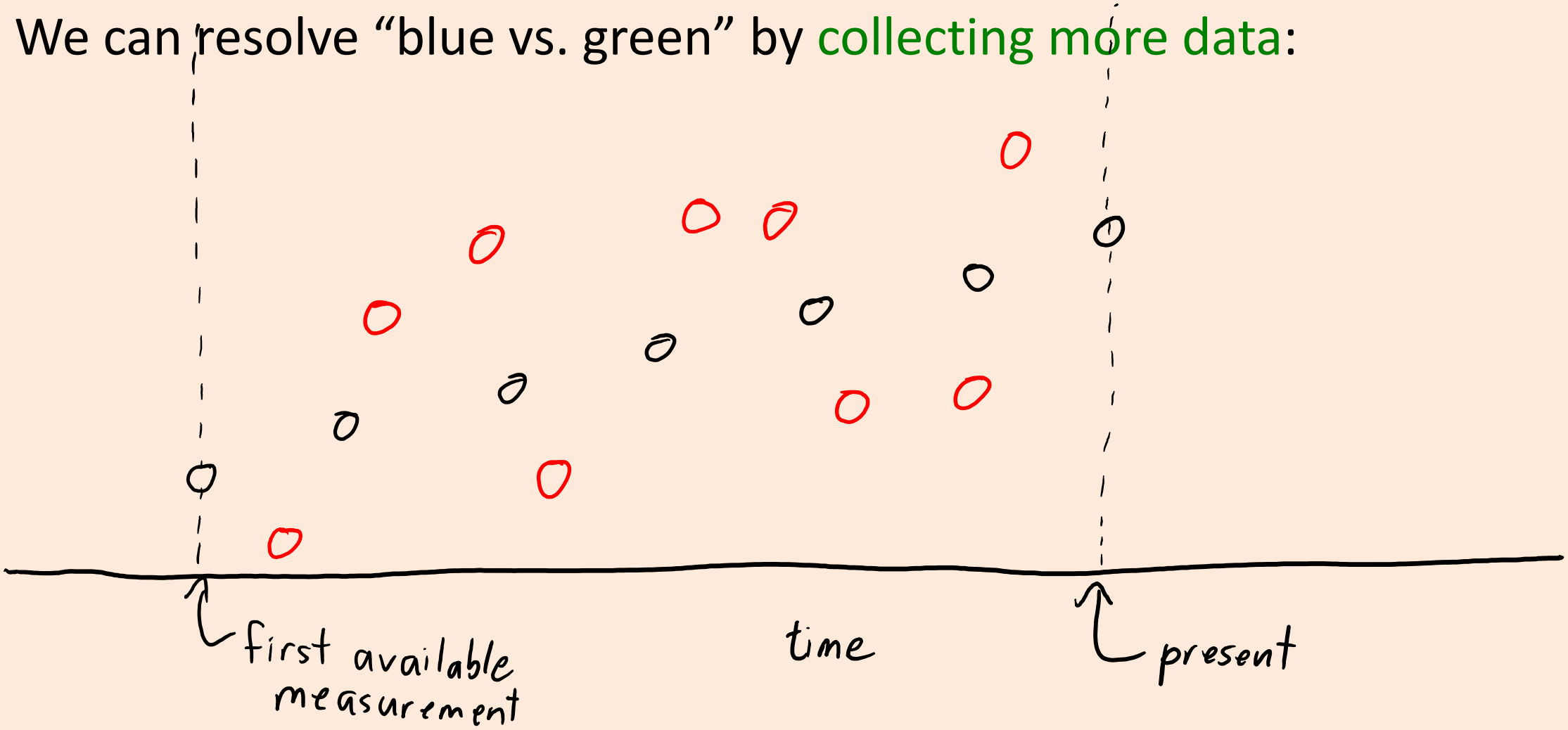


No Free Lunch, Consistency, and the Future

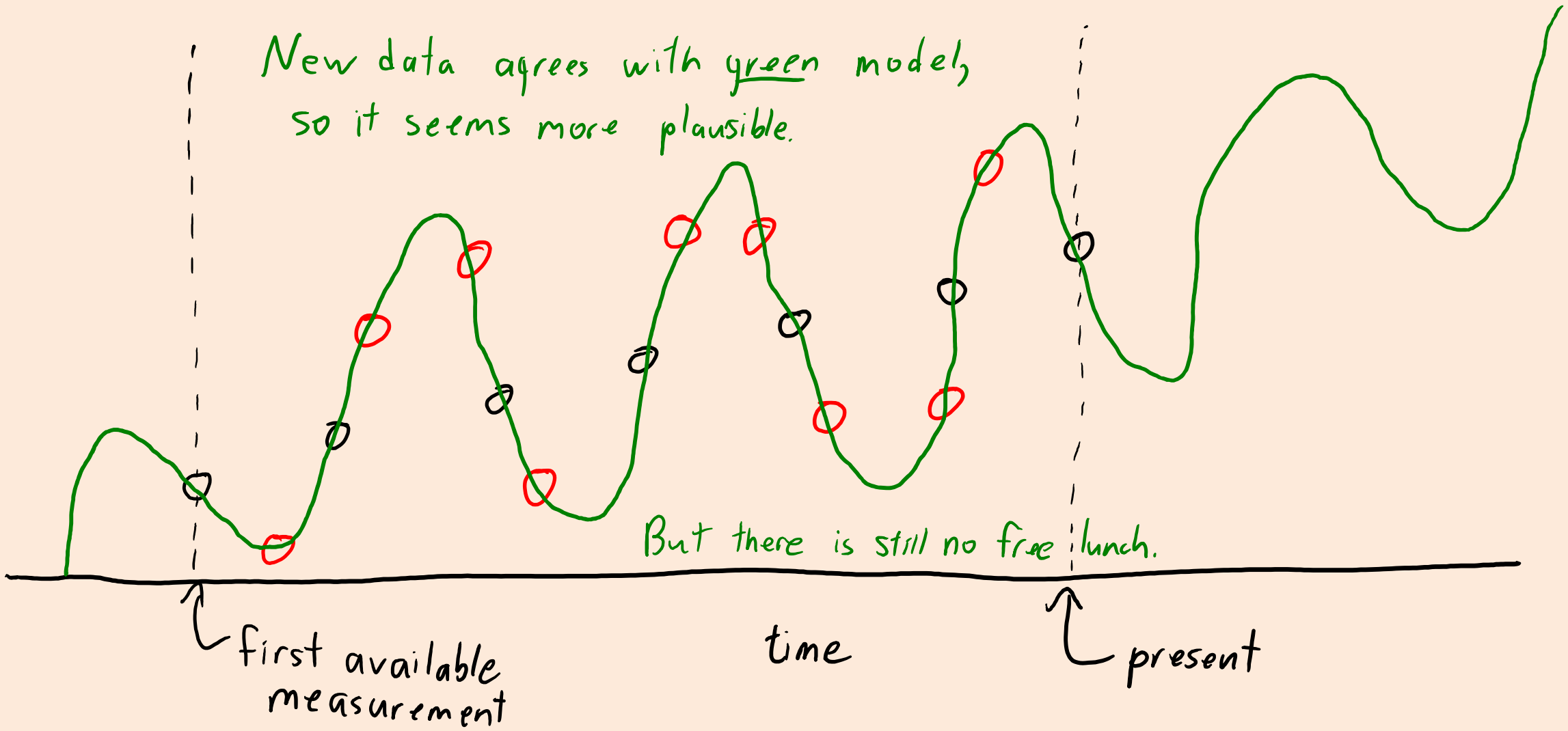


No Free Lunch, Consistency, and the Future

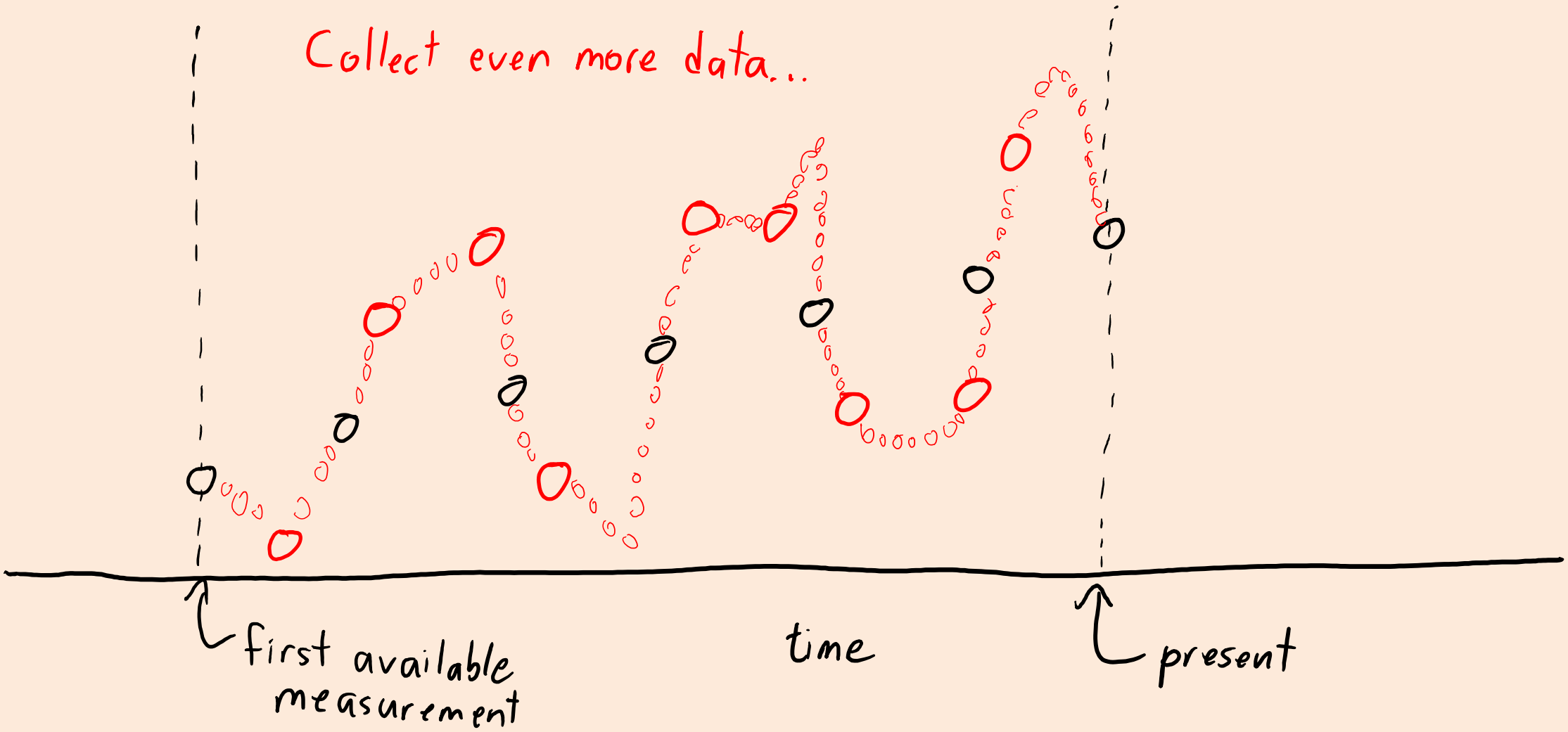
- We can resolve “blue vs. green” by collecting more data:



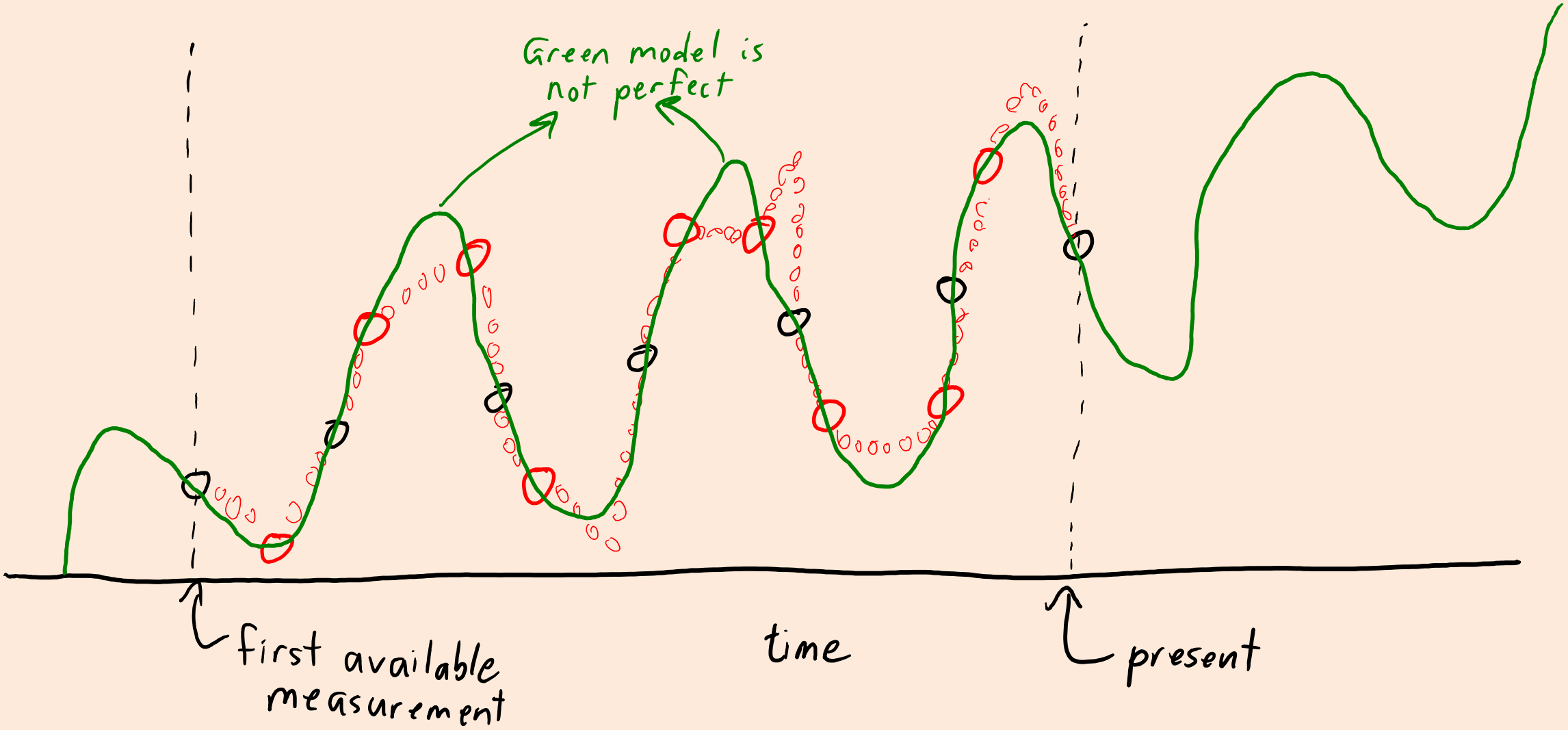
No Free Lunch, Consistency, and the Future



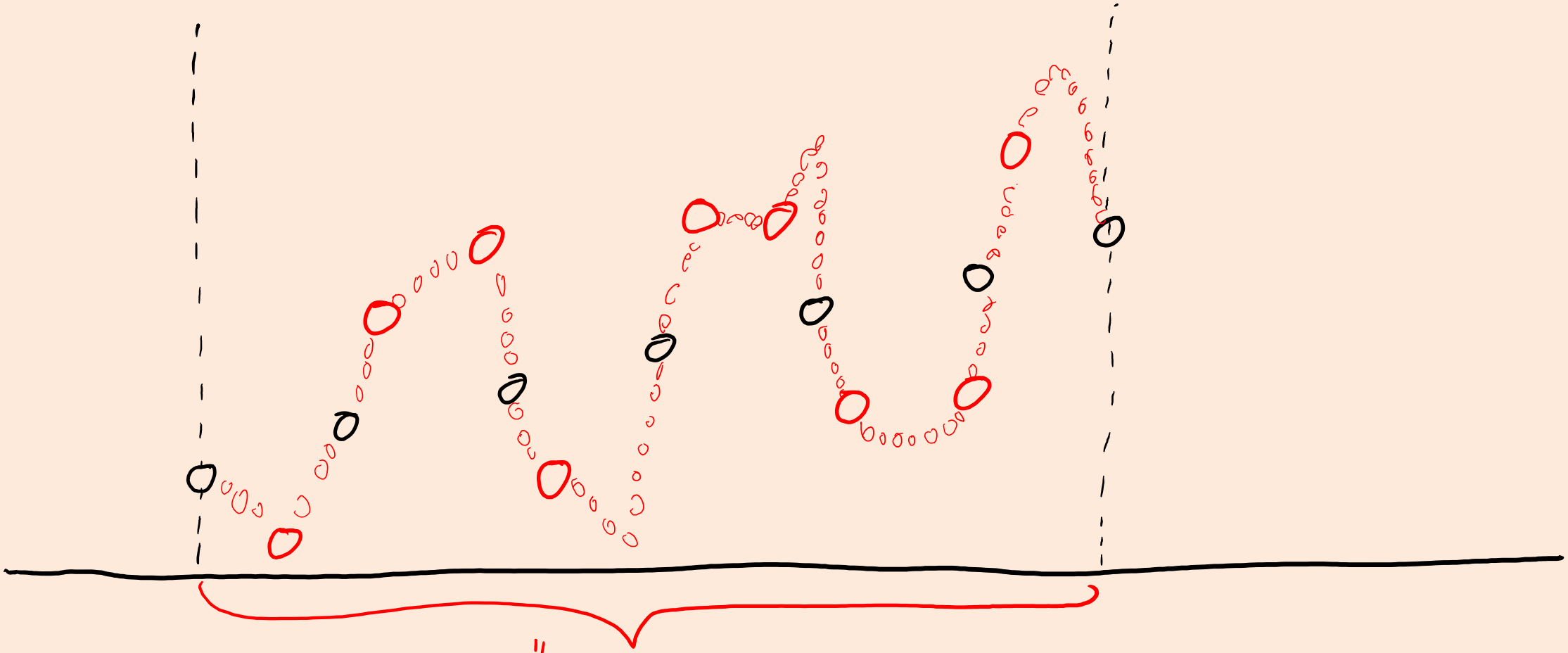
No Free Lunch, Consistency, and the Future



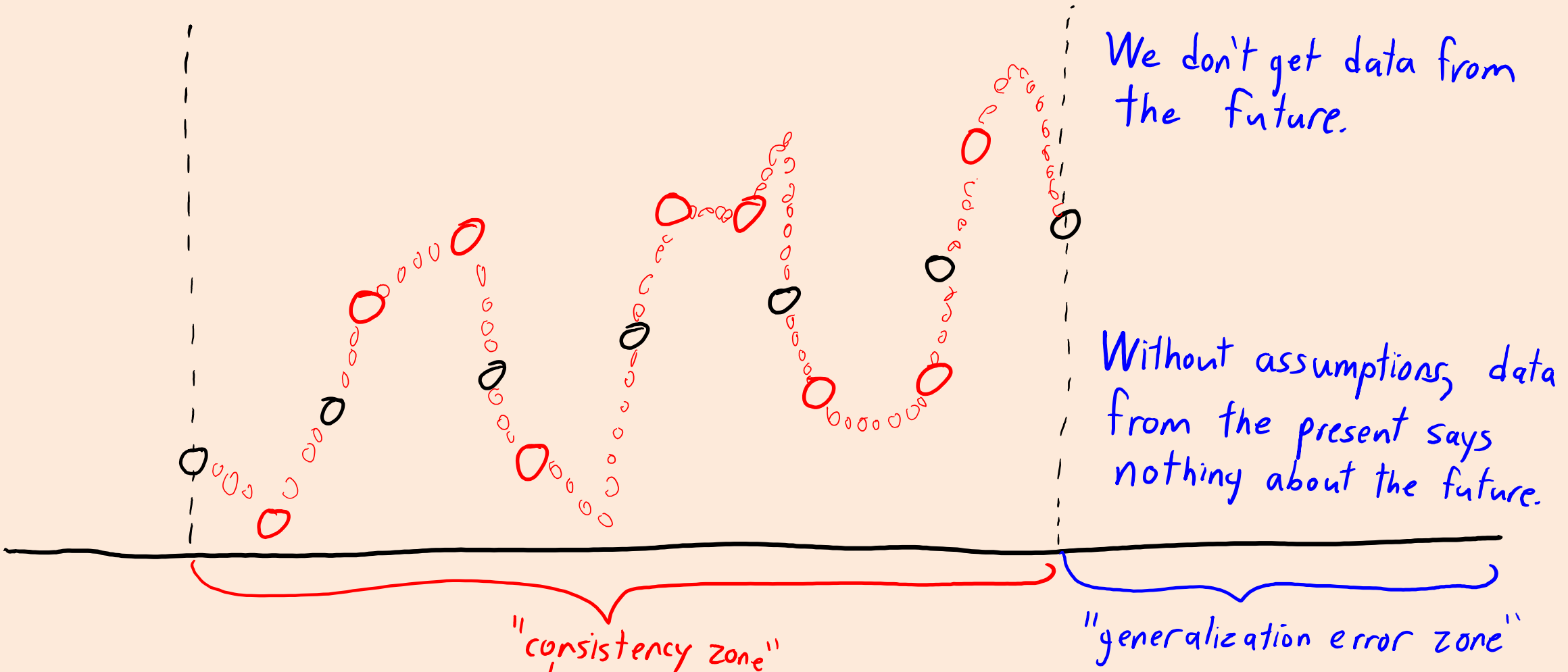
No Free Lunch, Consistency, and the Future



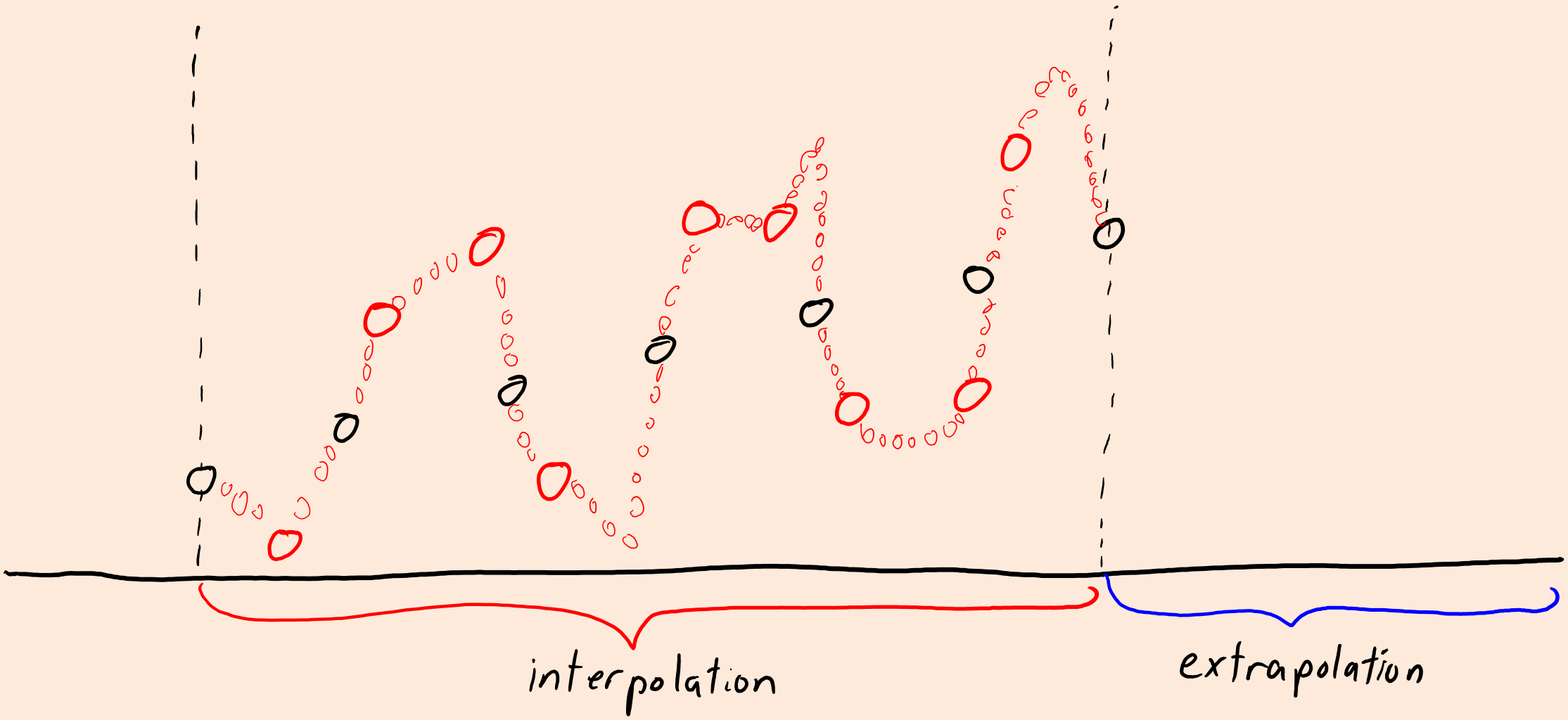
No Free Lunch, Consistency, and the Future



No Free Lunch, Consistency, and the Future

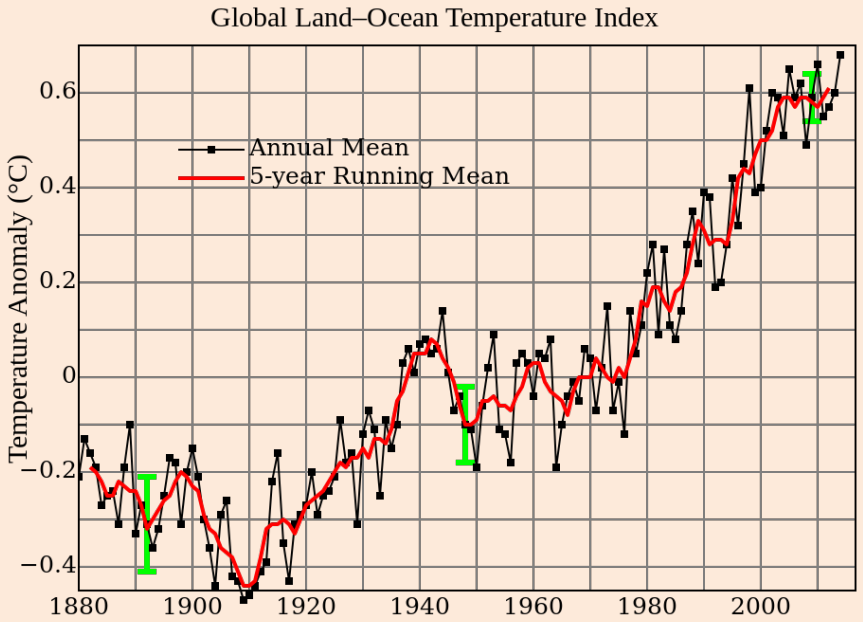


No Free Lunch, Consistency, and the Future



Discussion: Climate Models

- Has Earth warmed up over last 100 years? (Consistency zone)
  - Data clearly says “yes”.

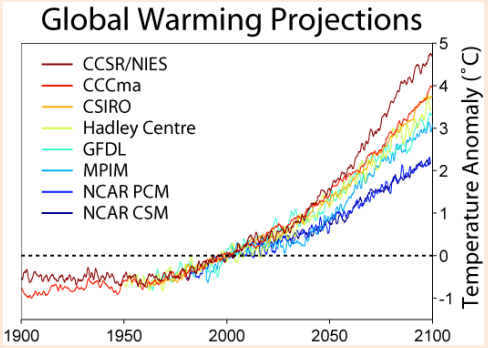


- Will Earth continue to warm over next 100 years? (generalization error)
  - We should be more skeptical about models that predict future events.

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

Discussion: Climate Models

- So should we all become global warming skeptics?
- If we average over models that overfit in *independent* ways, we expect the test error to be lower, so this gives more confidence:

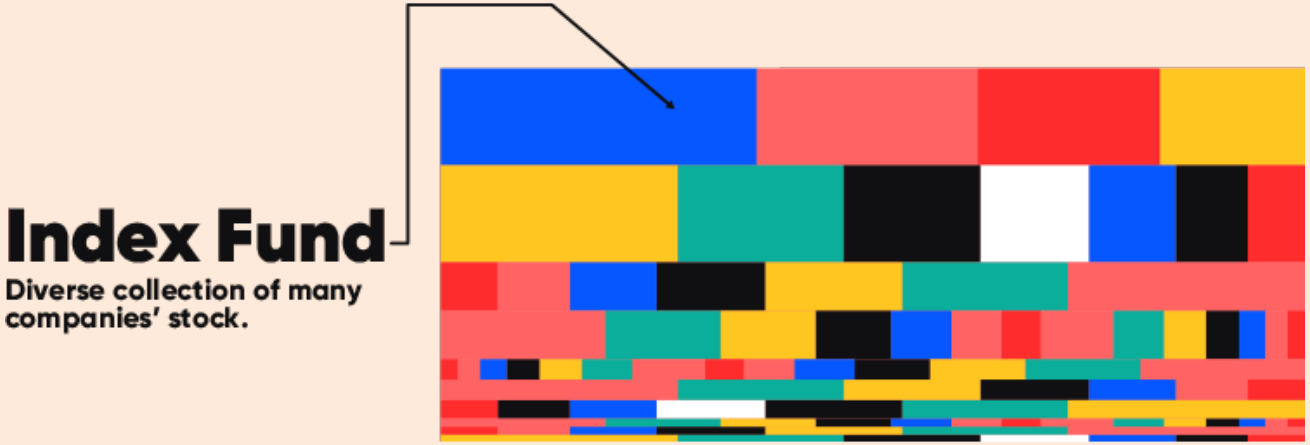


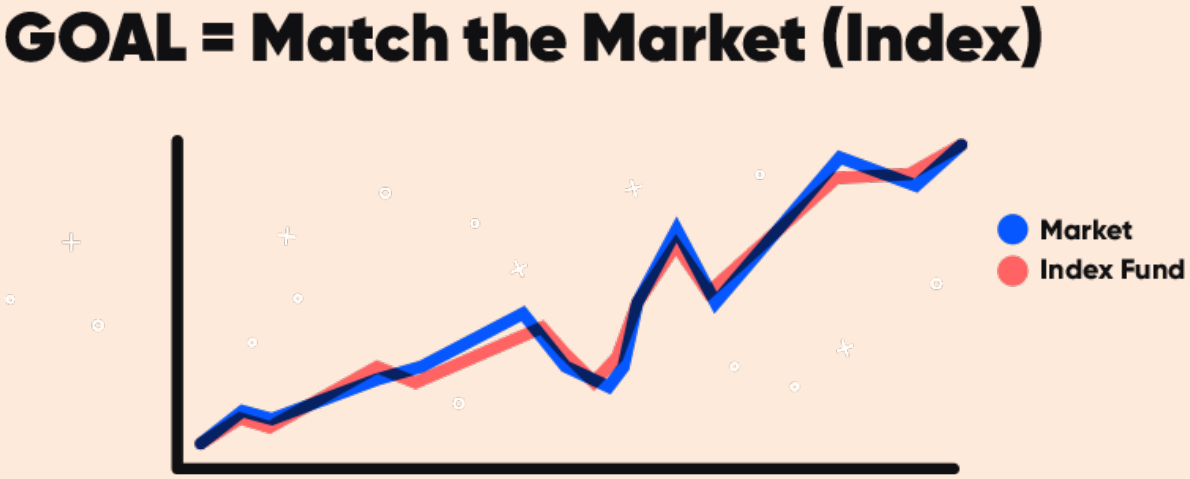
  - We should be skeptical of individual models, but agreeing predictions made by models with different data/assumptions are more likely be true.
- All the near-future predictions agree, so they are likely to be accurate.
  - And it's probably reasonable to assume fairly continuous change (no big "jumps").
- Variance is higher further into future, so predictions are less reliable.
  - Relying more on assumptions and less on data.

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

Index Funds: Ensemble Extrapolation for Investing

- Want to do extrapolation when investing money.
  - What will this be worth in the future?
- Index funds can be viewed as an ensemble method for investing.
  - For example, buy stock in top 500 companies proportional to value.
  - Tries to follow average price increase/decrease.





  - This simple investing strategy outperforms most fund managers.

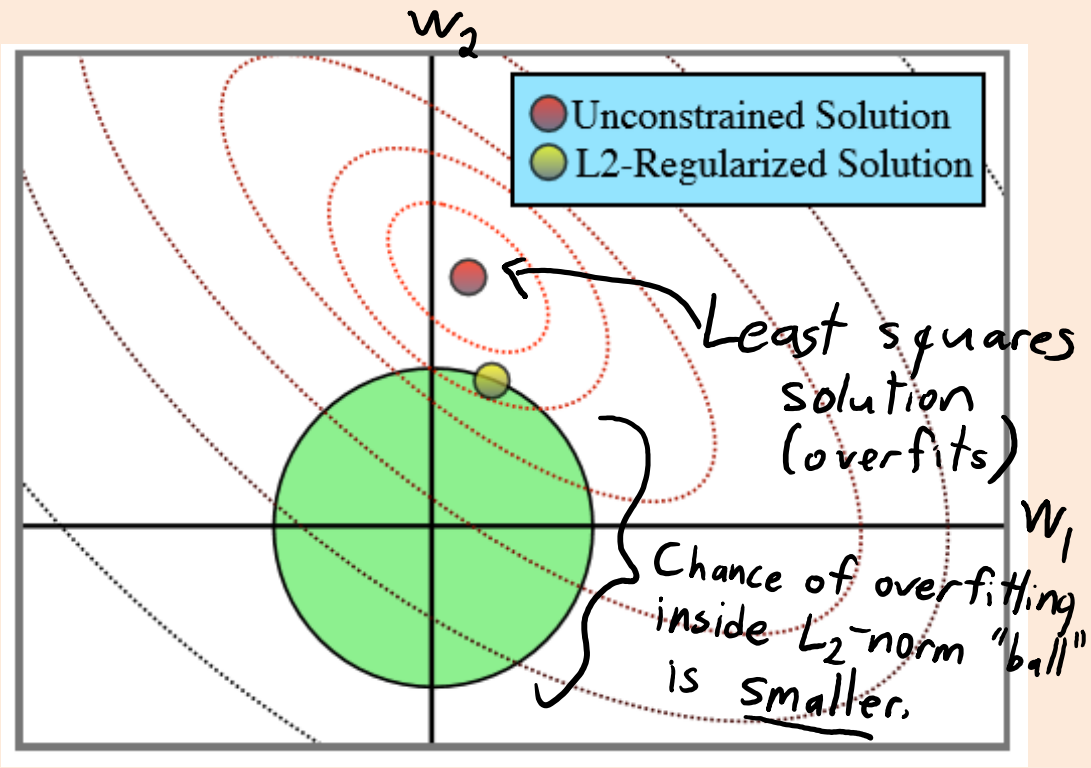
http://fibydesign.com/005-introduction-to-index-investing-stocks-index-funds-vtsax/

L2-Regularization

- Standard regularization strategy is L2-regularization:

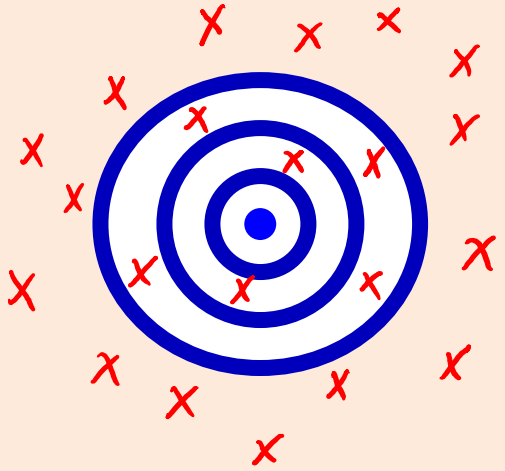
$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2 \quad \text{or} \quad f(w) = \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2$$

- Equivalent to minimizing squared error but keeping L2-norm small.



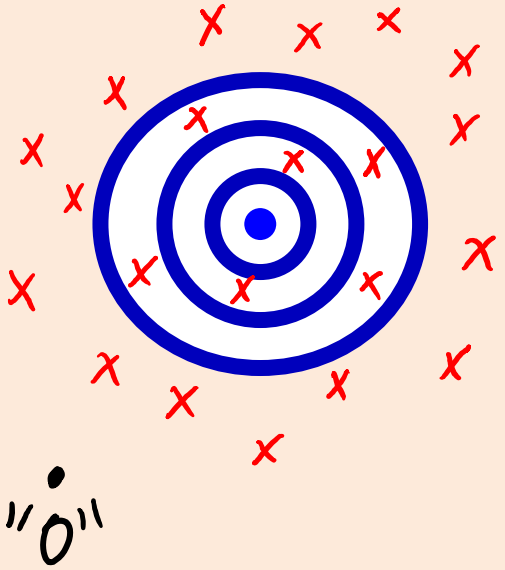
Regularization/Shrinking Paradox

- We throw darts at a target:
  - Assume we don't always hit the exact center.
  - Assume the darts follow a symmetric pattern around center.



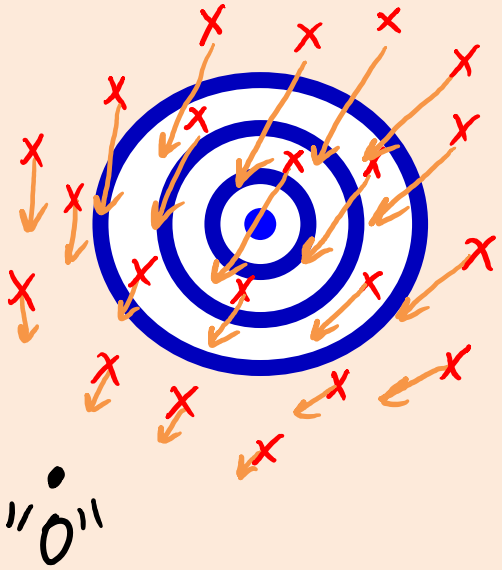
Regularization/Shrinking Paradox

- We throw darts at a target:
  - Assume we don't always hit the exact center.
  - Assume the darts follow a symmetric pattern around center.
- Shrinkage of the darts :
  1. Choose some arbitrary location '0'.
  2. Measure distances from darts to '0'.



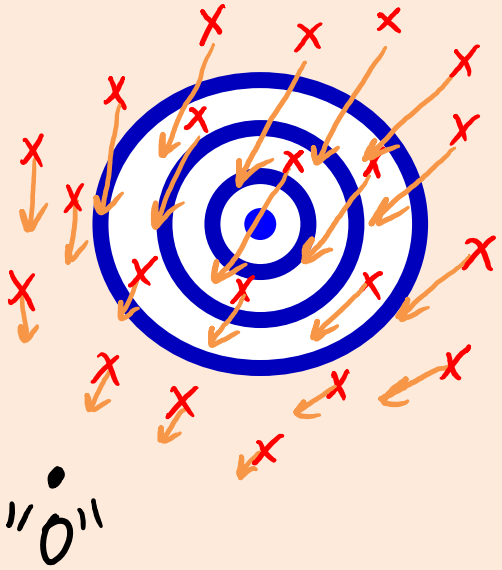
Regularization/Shrinking Paradox

- We throw darts at a target:
  - Assume we don't always hit the exact center.
  - Assume the darts follow a symmetric pattern around center.
- Shrinkage of the darts :
  1. Choose some arbitrary location '0'.
  2. Measure distances from darts to '0'.
  3. Move misses towards '0', by *small* amount proportional to distance from 0.
- If small enough, darts will be closer to center on average.



Regularization/Shrinking Paradox

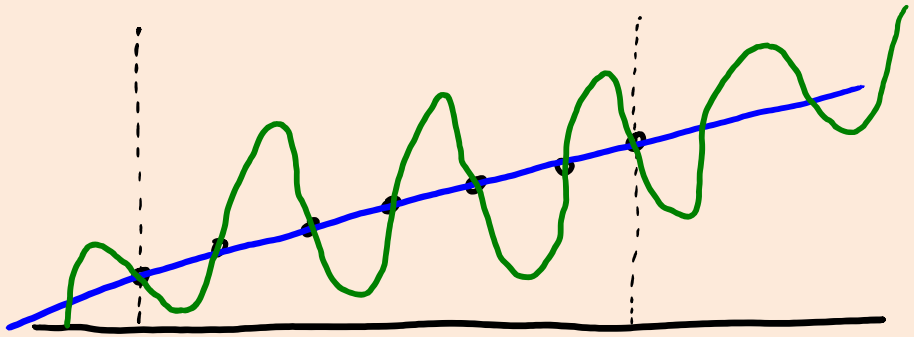
- We throw darts at a target:
  - Assume we don't always hit the exact center.
  - Assume the darts follow a symmetric pattern around center.
- Shrinkage of the darts :
  1. Choose some arbitrary location '0'.
  2. Measure distances from darts to '0'.
  3. Move misses towards '0', by *small* amount proportional to distance from 0.
- If small enough, darts will be closer to center on average.



Visualization of the related higher-dimensional paradox that the mean of data coming from a Gaussian is not the best estimate of the mean of the Gaussian in 3-dimensions or higher: https://www.naftaliharris.com/blog/steinviz

Ockham's Razor vs. No Free Lunch

- Ockham's razor is a problem-solving principle:
  - "Among competing hypotheses, the one with the fewest assumptions should be selected."
  - Suggests we should select linear model.



- Fundamental trade-off:
  - If same training error, pick model less likely to overfit.
  - Formal version of Occam's problem-solving principle.
  - Also suggests we should select linear model.
- No free lunch theorem:
  - There *exists possible datasets* where you should select the green model.