CPSC 340:
Machine Learning and Data Mining

Regularization

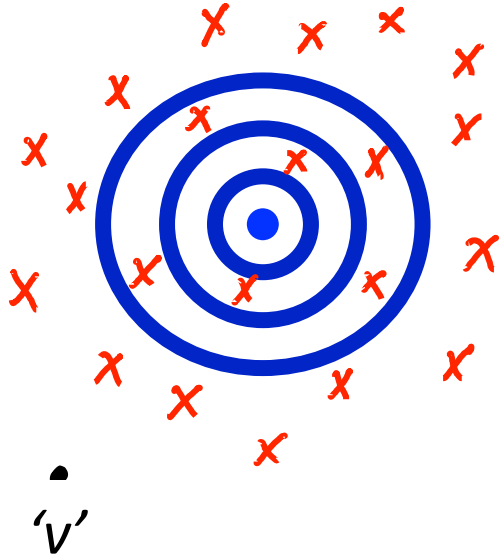
BONUS SLIDES

“Shrinking” intuition

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

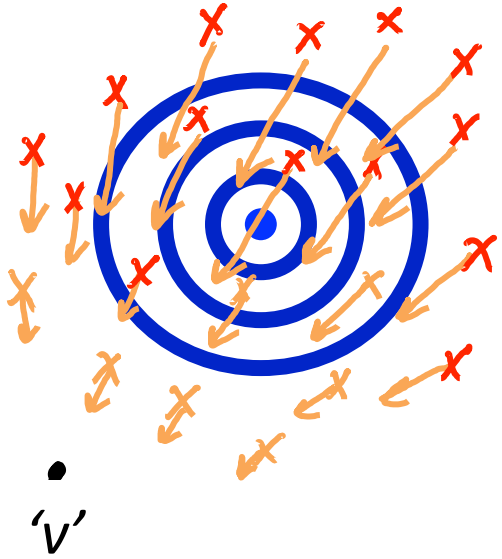
“Shrinking” intuition

- 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 ‘v’.
  2. Measure distances from darts to ‘v’.



"Shrinking" intuition

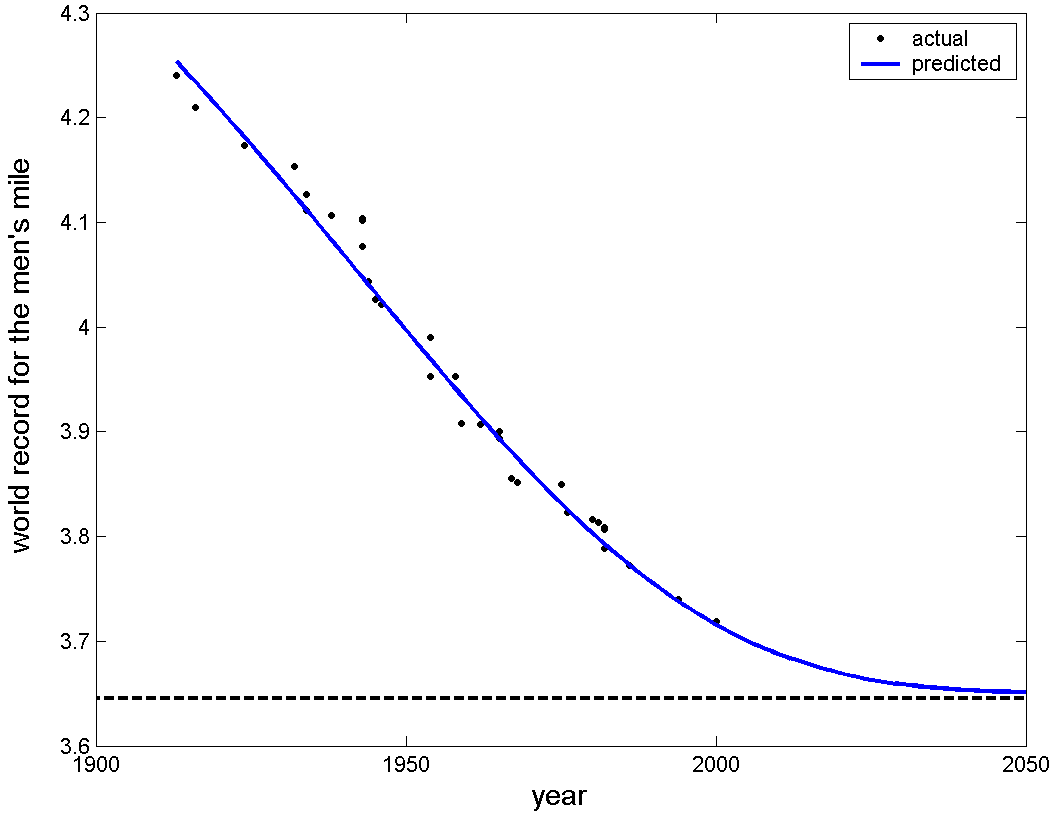
- We throw darts at a target:
  - Assume we don't always hit the exact center.
  - Assume the darts follow a symmetric pattern around center.
- Shrinkage of the darts towards 'v':
  1. Choose some arbitrary location 'v'.
  2. Measure distances from darts to 'v'.
  3. Move misses towards 'v', by *small*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



- If small enough, darts will be closer to center on average.
- In standard L2-regularization, our location 'v' is not arbitrary (it's w=0).
  - The shrinking works even better if 'v' is closer to the centre.
  - But we don't know where the centre is (this is the "true" w).
  - We pick 'v' as the origin because of the Occam's razor ideas (simpler models).

Bonus Slide: Predicting the Future

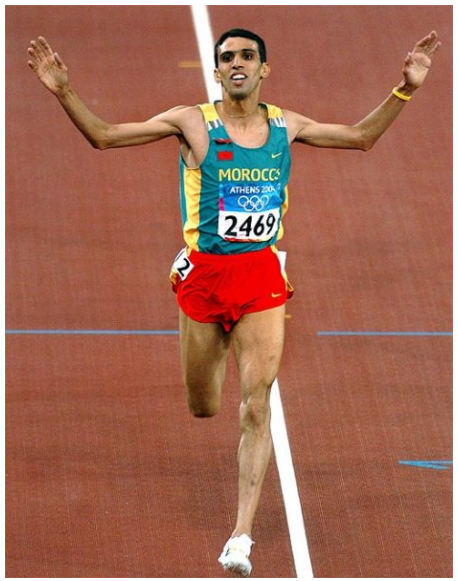
- In principle, we can use any features x_i that we think are relevant.
- This makes it tempting to use time as a feature, and predict future.

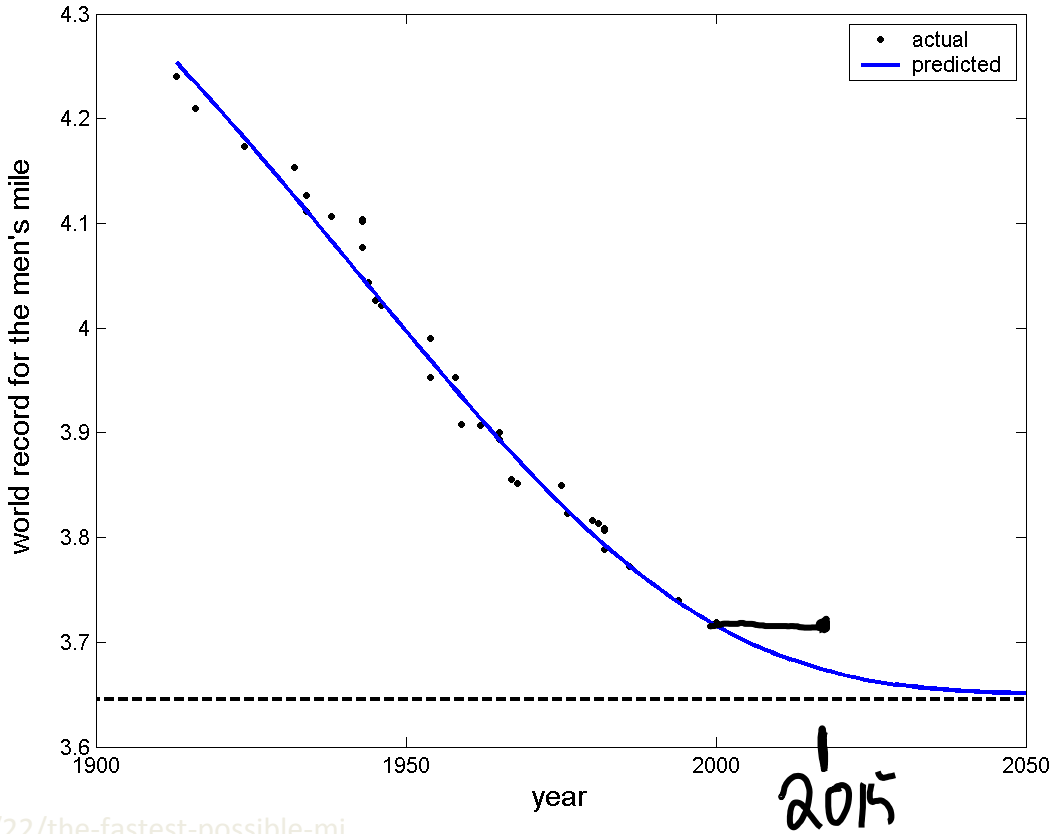


https://gravityandlevity.wordpress.com/2009/04/22/the-fastest-possible-mile/

Bonus Slide: 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.

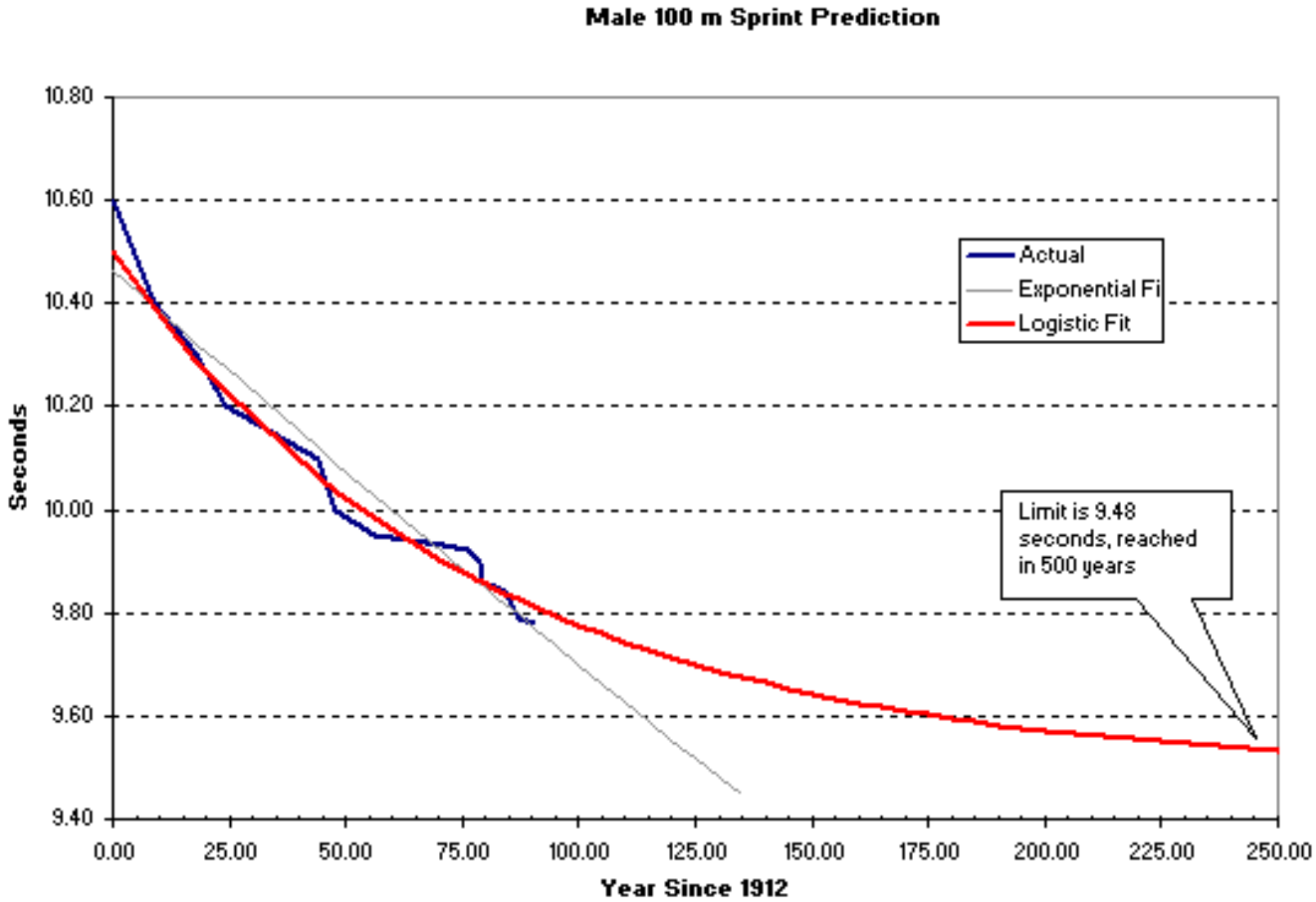




We need to be cautious about doing this.

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

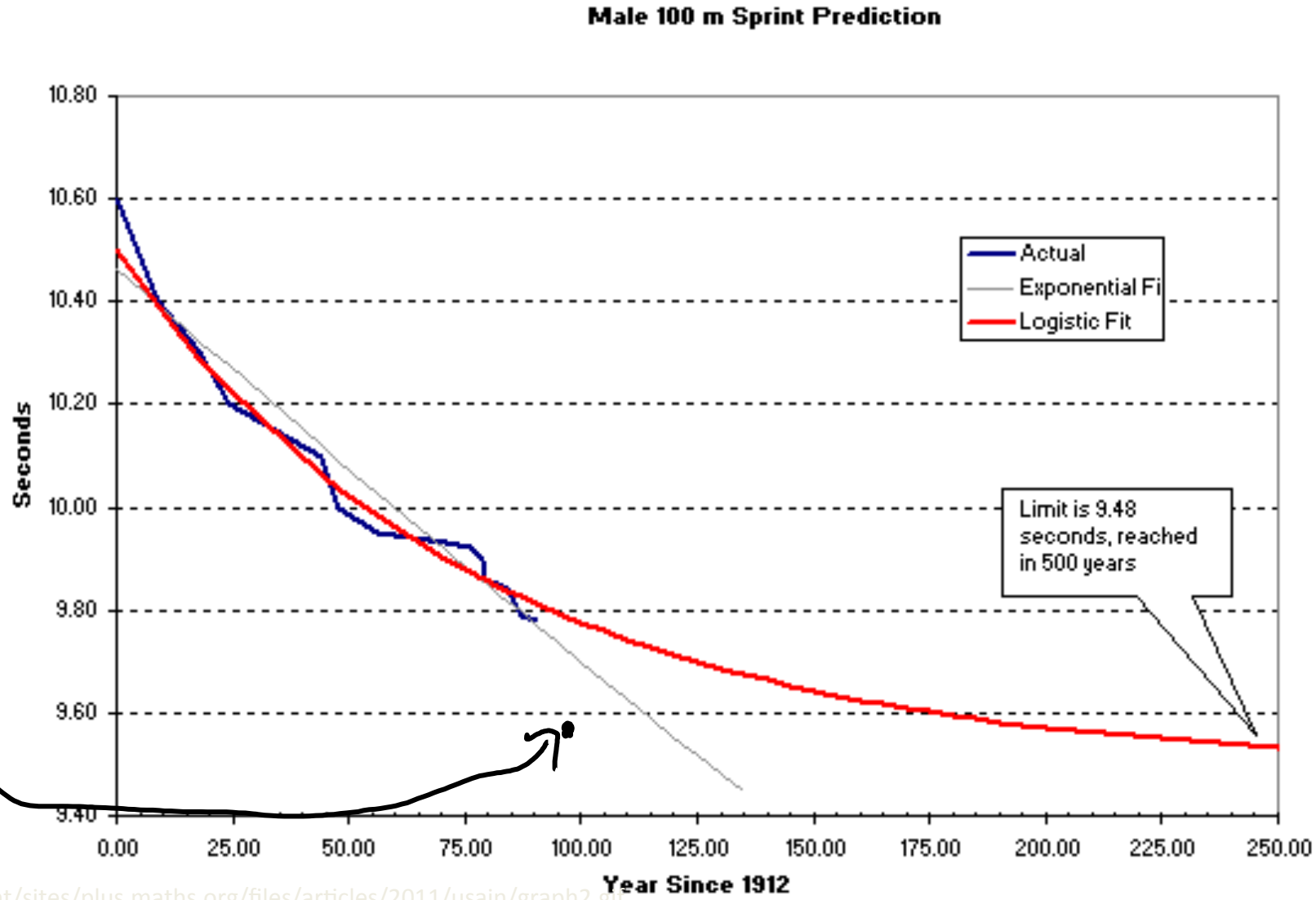
Bonus Slide: Predicting 100m times 400 years in the future?



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

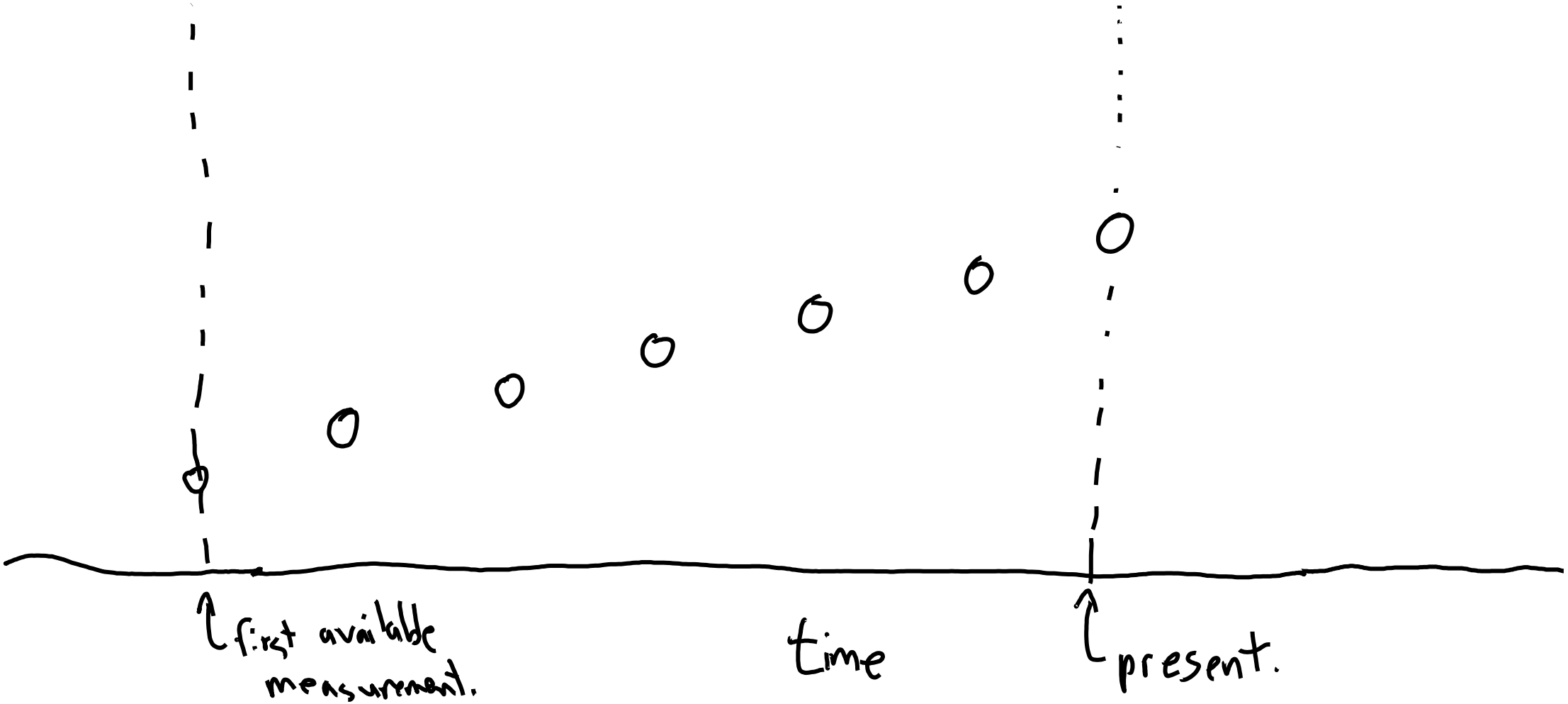
Bonus Slide: Predicting 100m times 400 years in the future?

9.58

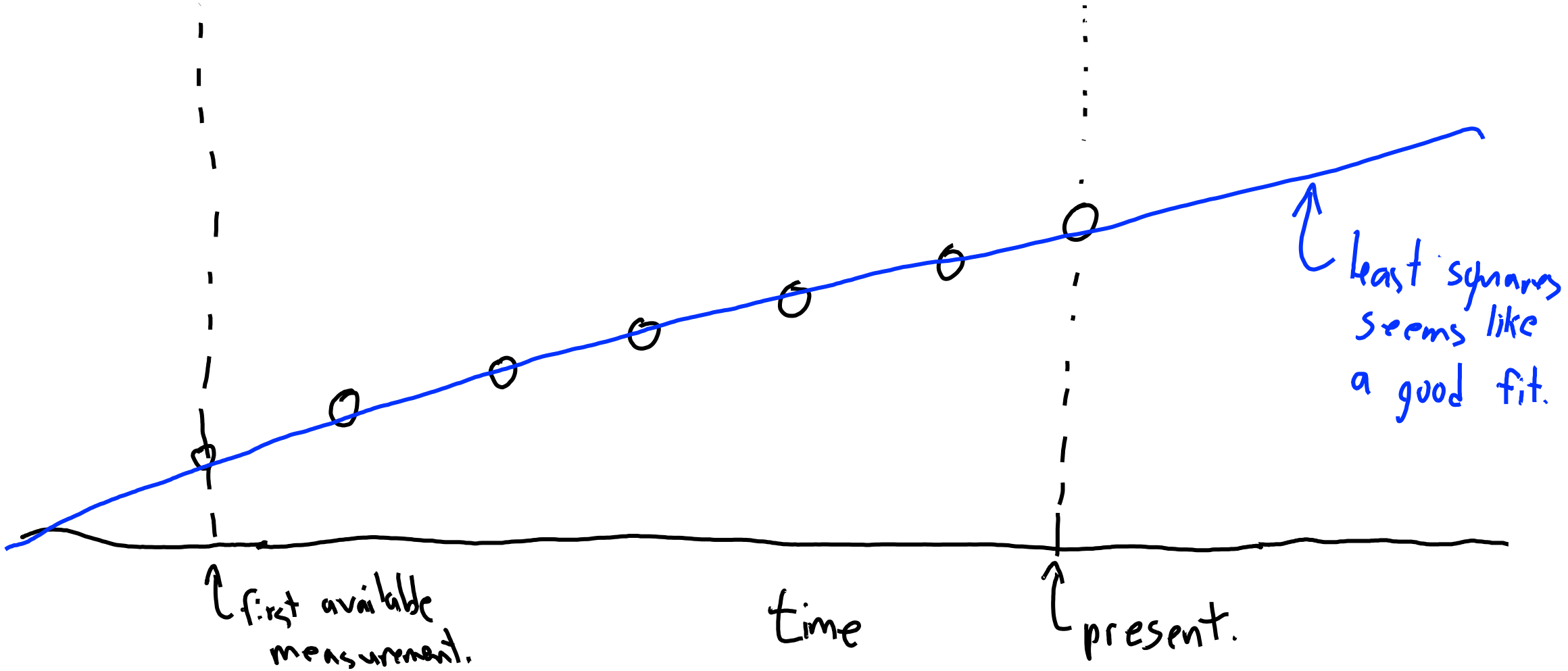


https://plus.maths.org/content/sites/plus.maths.org/files/articles/2011/usain/graph2.gif
http://www.washingtonpost.com/blogs/london-2012-olympics/wp/2012/08/08/report-usain-bolt-invited-to-tryout-for-manchester-united/

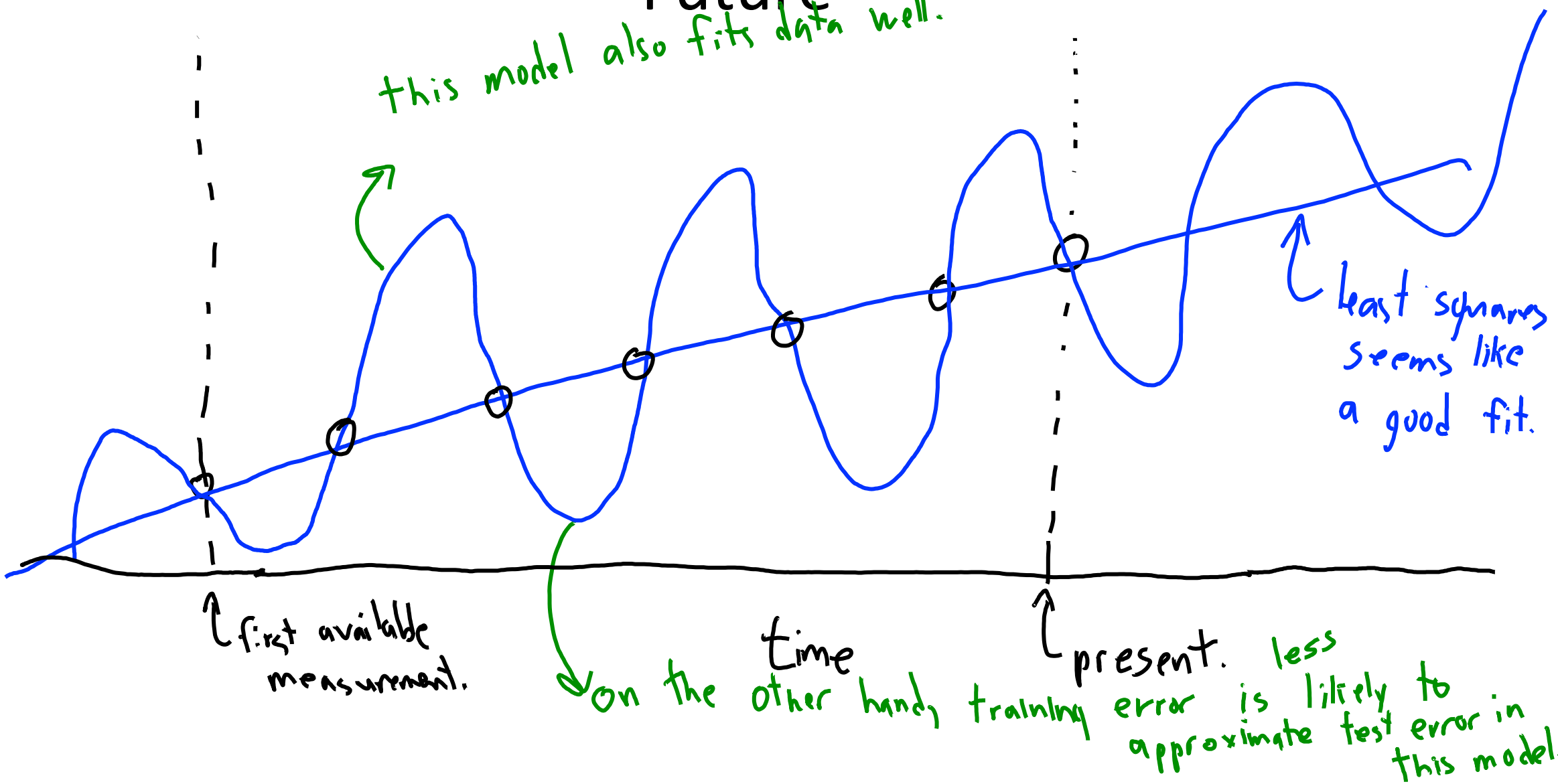
Bonus Slide: No Free Lunch, Consistency, and the Future



Bonus Slide: No Free Lunch, Consistency, and the Future

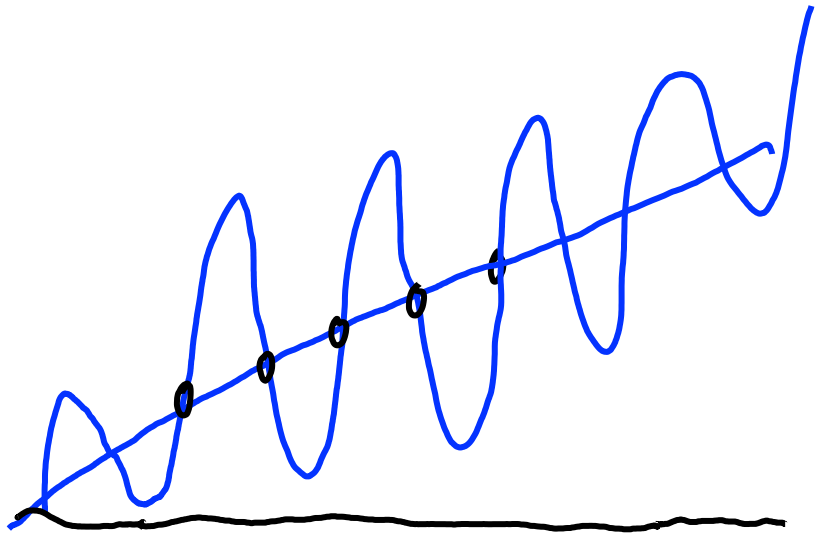


Bonus Slide: No Free Lunch, Consistency, and the Future

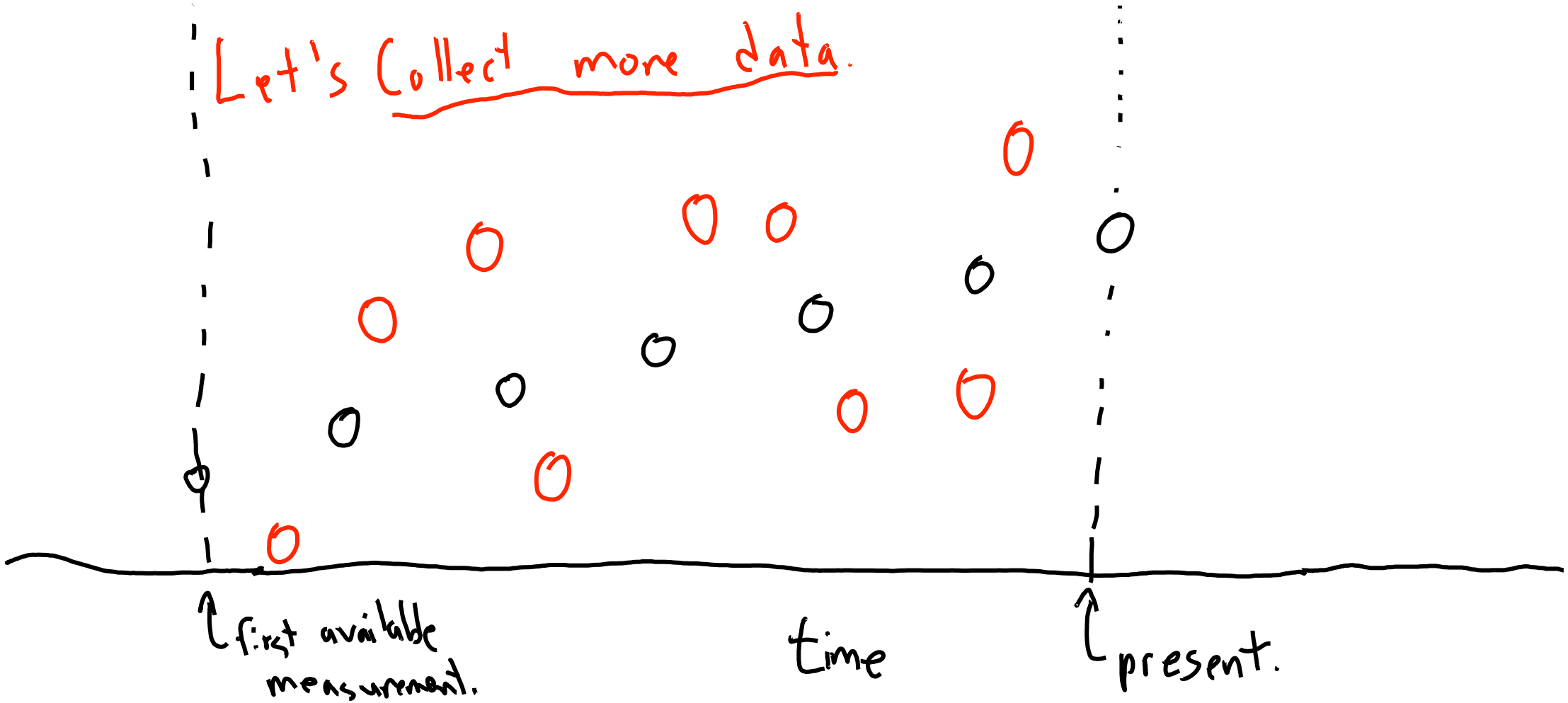


Bonus Slide: Ockham's Razor vs. No Free Lunch

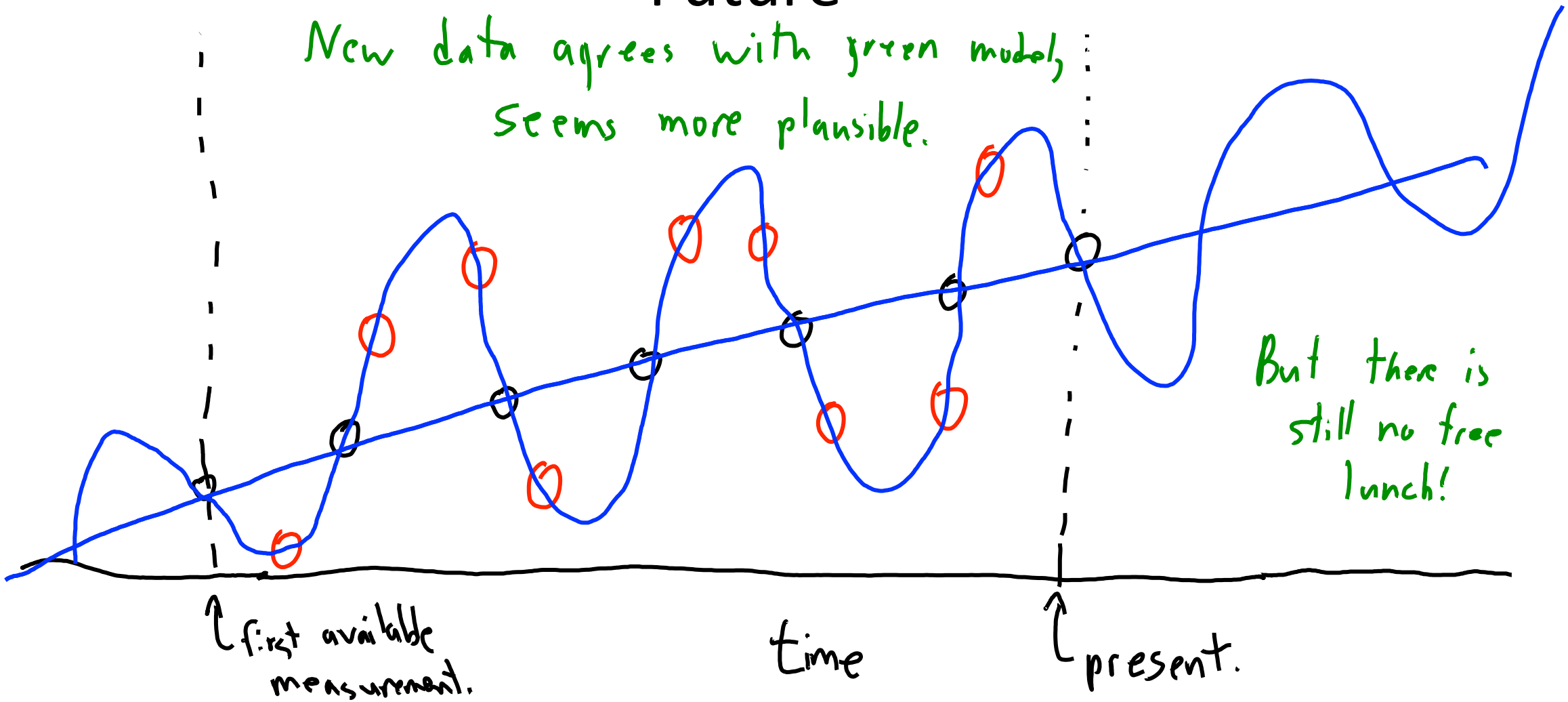
- Ockham's razor is a problem-solving principle:
  - "Among competing hypotheses, the one with the fewest assumptions should be selected."
  - Suggests we should select linear model.
- Fundamental theorem of ML:
  - If training same error, pick model less likely to overfit.
  - Formal version of Occam's 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.



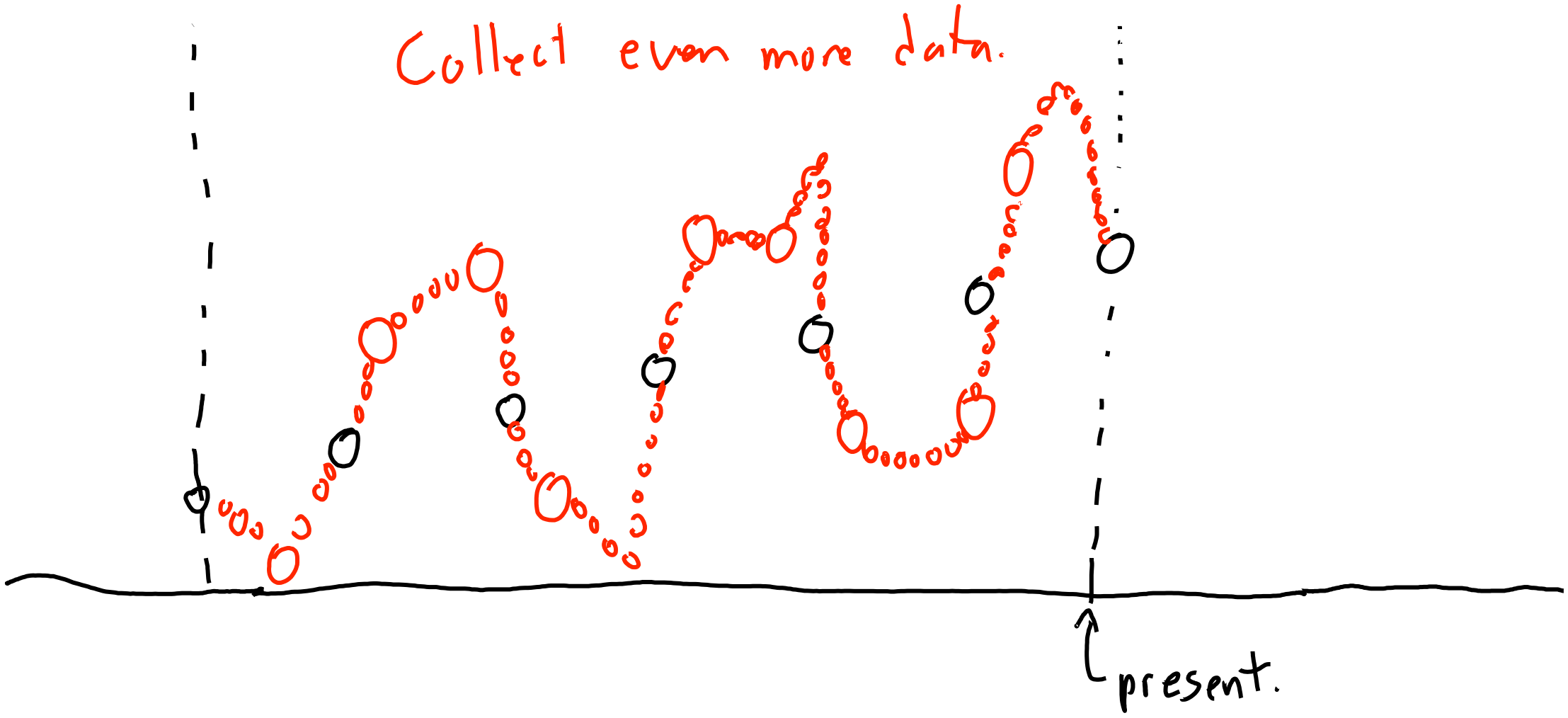
Bonus Slide: No Free Lunch, Consistency, and the Future



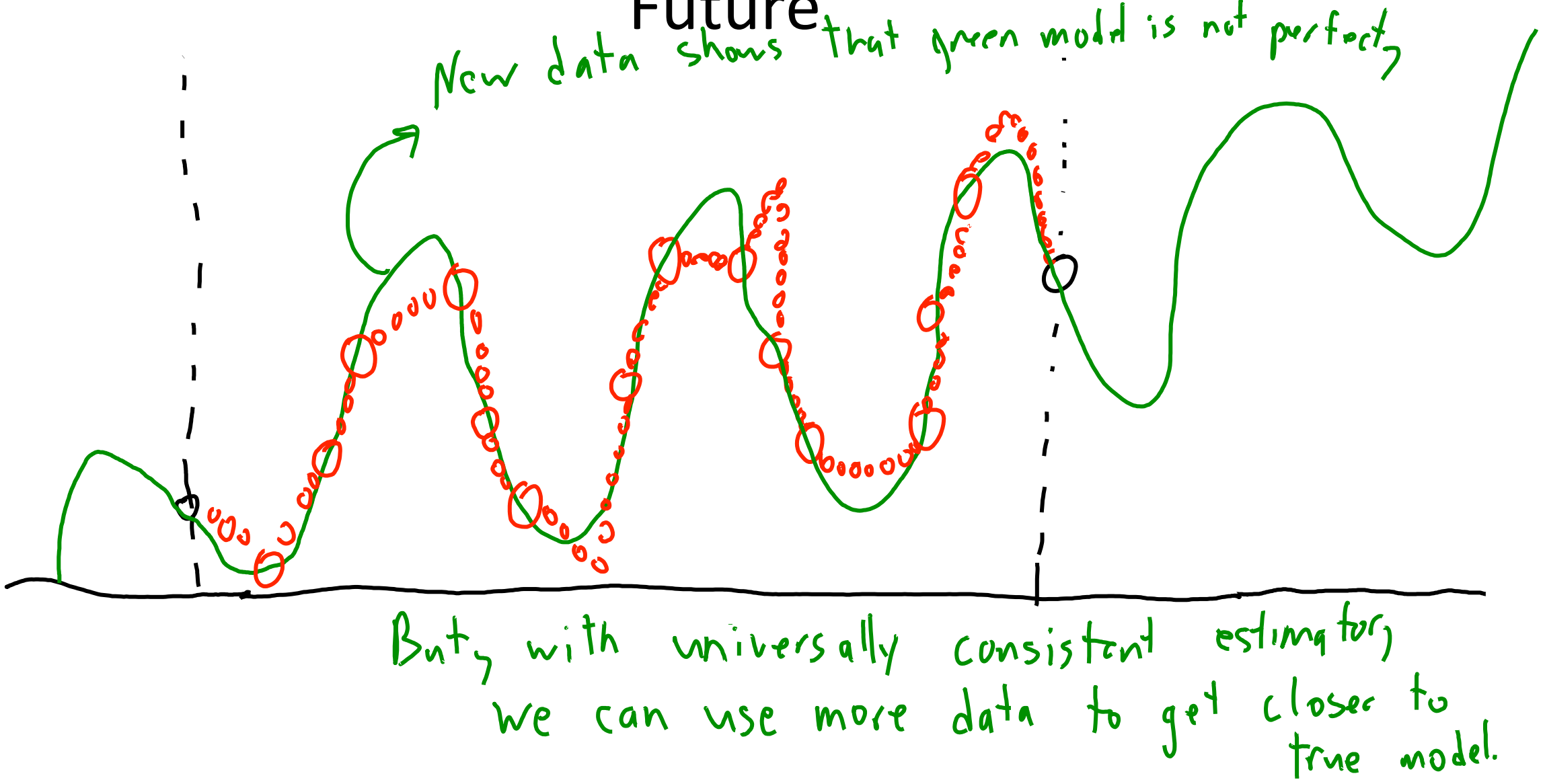
Bonus Slide: No Free Lunch, Consistency, and the Future



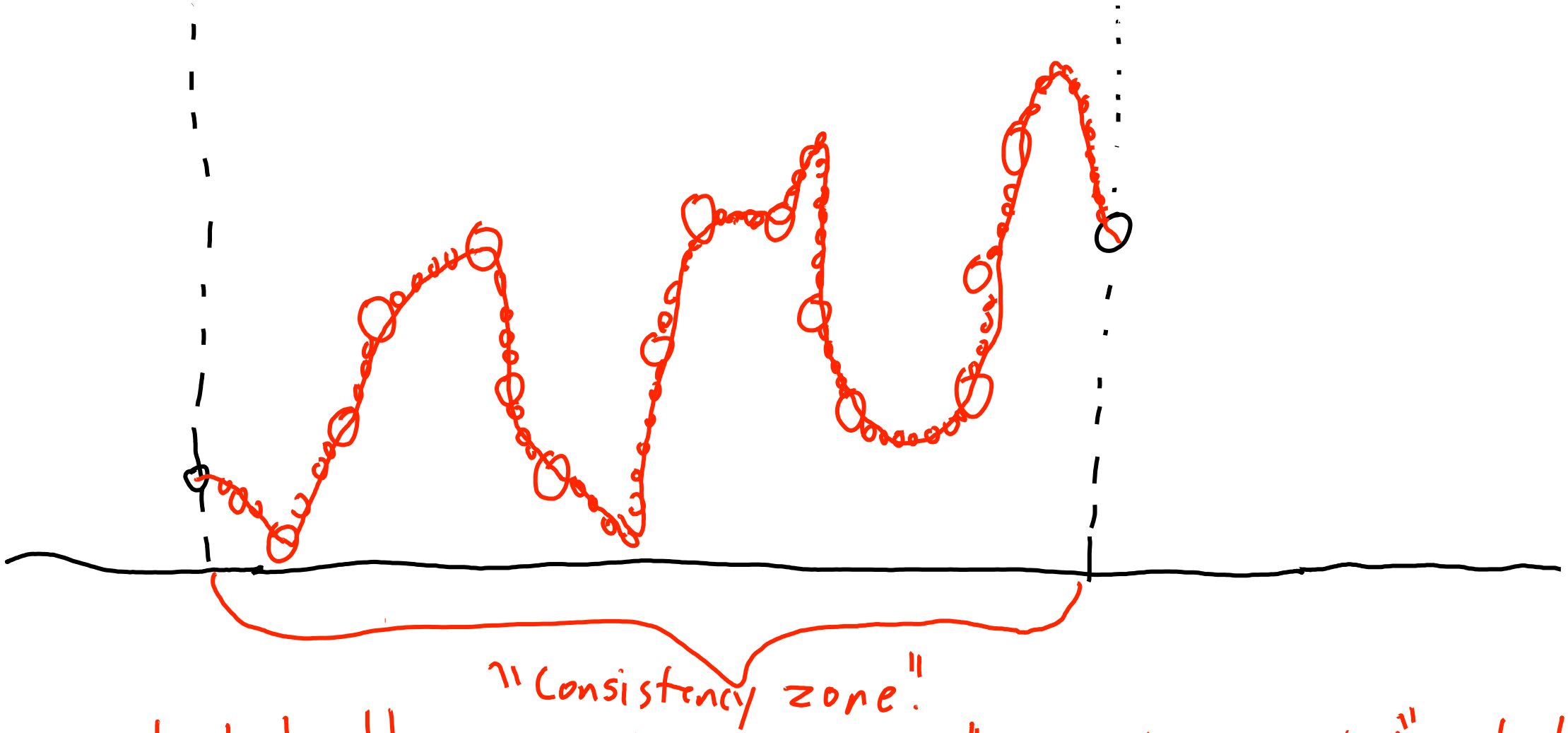
Bonus Slide: No Free Lunch, Consistency, and the Future



Bonus Slide: No Free Lunch, Consistency, and the Future

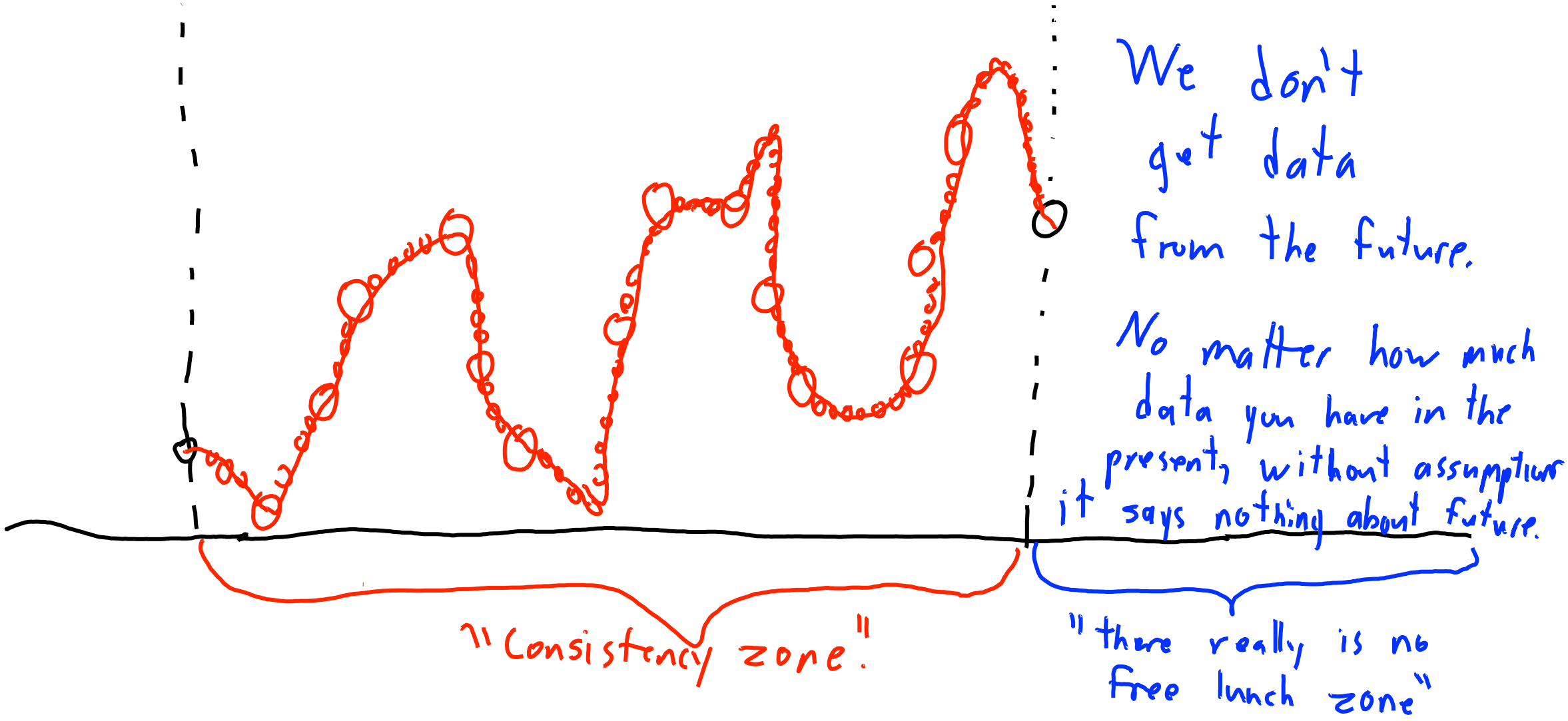


Bonus Slide: No Free Lunch, Consistency, and the Future

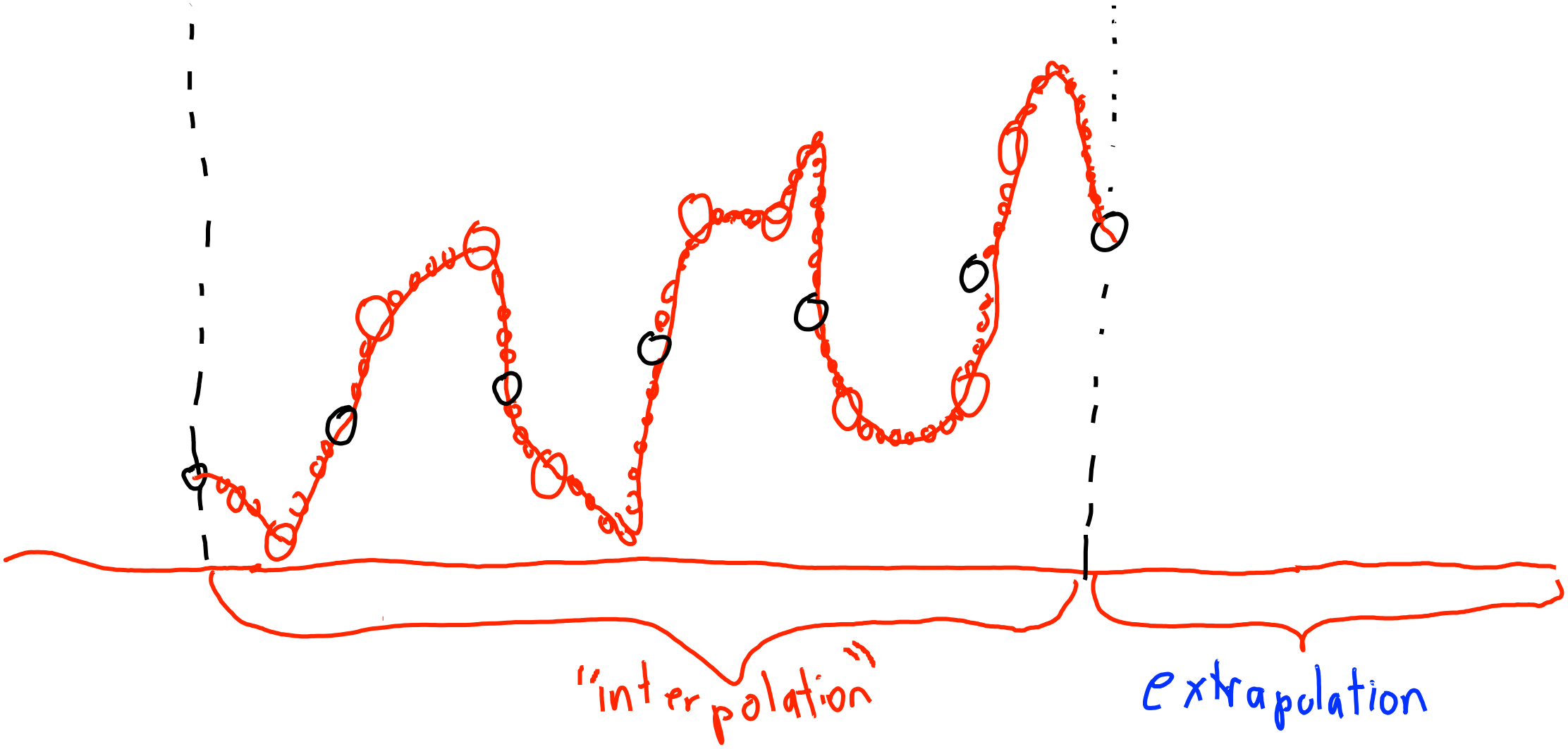


Converge to best model as $n \to \infty$, if we use a "universally consistent" method.

Bonus Slide: No Free Lunch, Consistency, and the Future

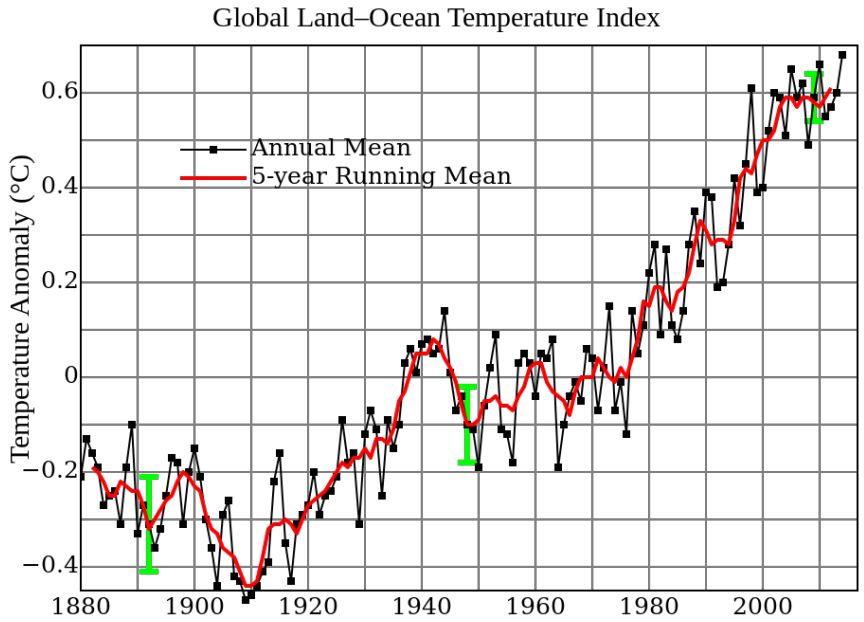


Bonus Slide: No Free Lunch, Consistency, and the Future



Bonus Slide: Application: Climate Models

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

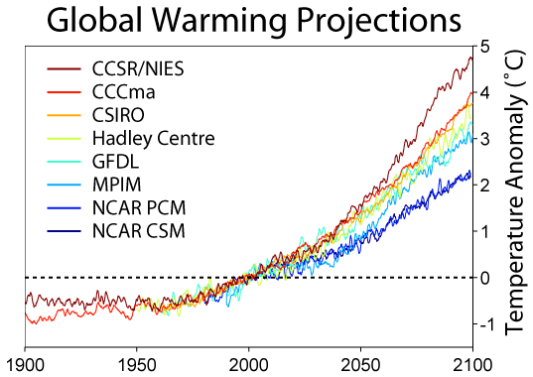


- Will Earth continue to warm over next 100 years? (Really NFL zone)
  - We should be more skeptical about models that predict future events.

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

Bonus Slide: Application: 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:

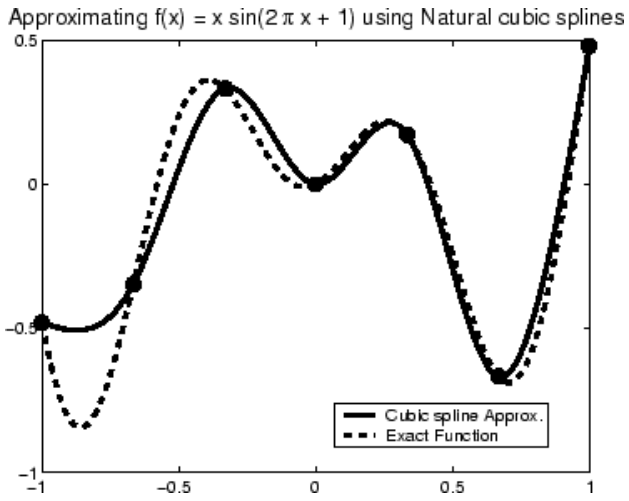


  - We should be skeptical of individual models, but agreeing predictions made by models with different data/assumptions are more likely be true.
- If all near-future predictions agree, they are likely to be accurate.
- As we go further in the future, variance of average will be higher.

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

Bonus Slide: Splines in 1D

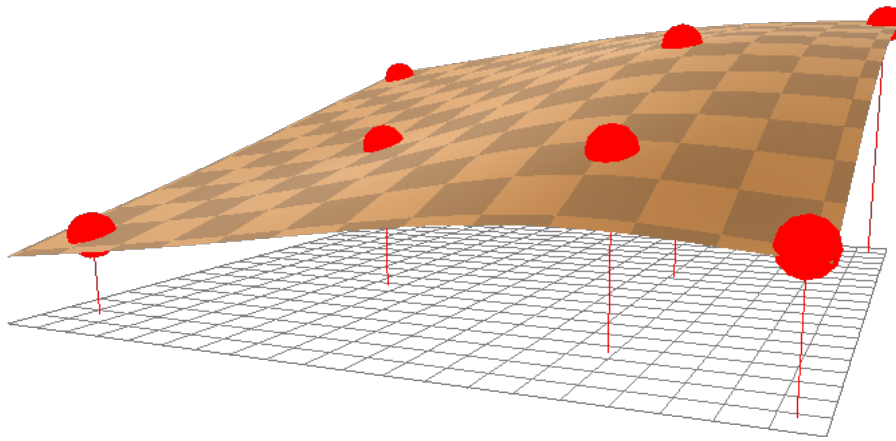
- For 1D interpolation, alternative to polynomials/RBFs are splines:
  - Use a polynomial in the region between each data point.
  - Constrain some derivatives of the polynomials to yield a unique solution.
- Most common example is cubic spline:
  - Use a degree-3 polynomial between each pair of points.
  - Enforce that f'(x) and f''(x) of polynomials agree at all point.
  - "Natural" spline also enforces f''(x) = 0 for smallest and largest x.
- Non-trivial fact: natural cubic splines are sum of:
  - Y-intercept.
  - Linear basis.
  - RBFs with $g(\alpha) = \alpha^3$.
    - Different than Gaussian RBF because it *increases with distance*.



http://www.physics.arizona.edu/~restrepo/475A/Notes/sourcea-/node35.html

Bonus Slide: Spline in Higher Dimensions

- Splines generalize to higher dimensions if data lies on a grid.
  - For more general ("scattered") data, there isn't a natural generalization.
- Common 2D "scattered" data interpolation is thin-plate splines:
  - Based on curve made when bending sheets of metal.
  - Corresponds to RBFs with $g(\alpha) = \alpha^2 \log(\alpha)$.
- Natural splines and thin-plate splines: special cases of "polyharmonic" splines:
  - Less sensitive to parameters than Gaussian RBF.



http://step.polymtl.ca/~rv101/thinplates/