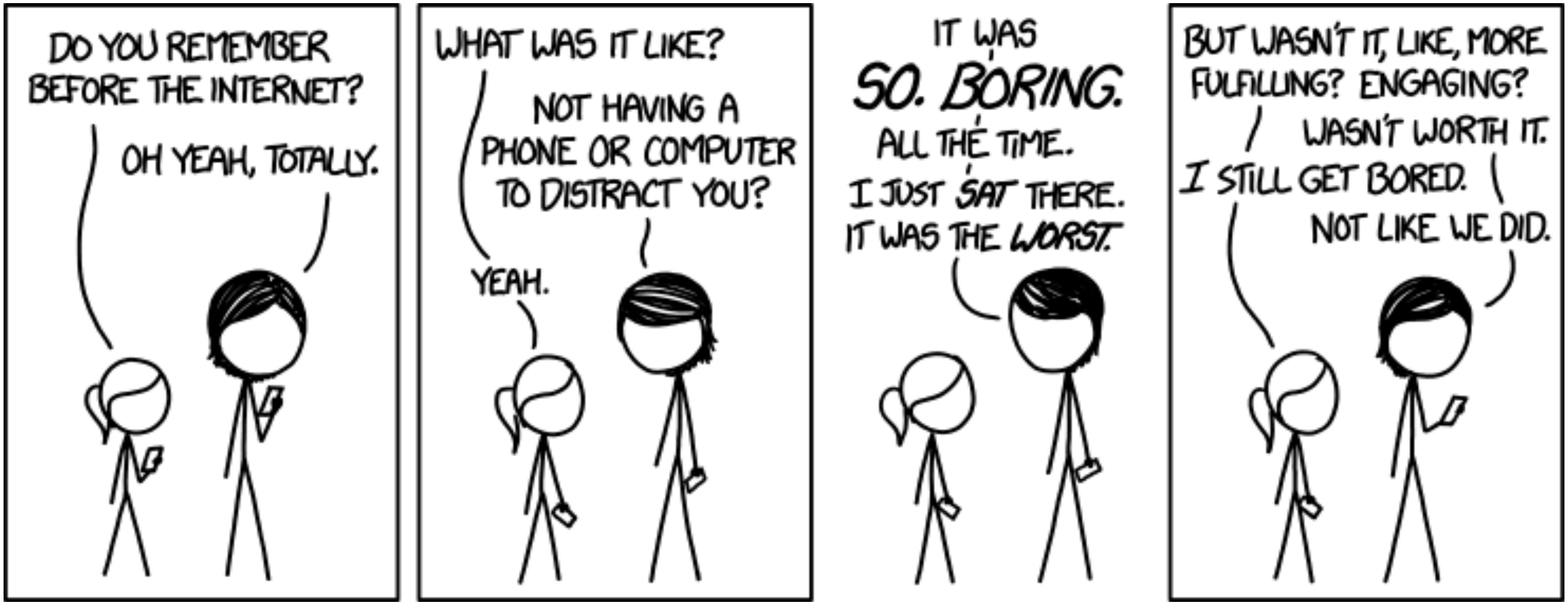
Final Exam Review



Comic: xkcd.com

Schedule

- Today is the last lecture!
- No discussion this week
- Teaching staff review session:
  Friday, April 21 7pm-8:30pm, IOE 1610
- Final exam time:
  Tuesday, April 25 7pm-9pm, various rooms
- Alternate exam details: contact us, or expect email shortly

Final Exam Locations

Room	First uniqname in room	Last uniqname in room
CHRYSLER 220	aanant	dyiming
CHRYSLER 133	eanndale	hiimeggr
STAMPS AUDITORIUM	hkardos	wanhongz
BEYSTER 1670	wenctsai	zzkwx

Policies

- Closed book
- Closed notes
- One "cheat sheet"
  - 8.5"x11", double-sided
- No calculators or electronics
  - None needed
- Given under the engineering honor code

Format

- 60% multiple choice
  - **REMEMBER #2 PENCIL!**
- 40% code and free response

Study materials

- Practice exams
- Lecture notes
- Lecture slides
- Discussion slides

- Everything is posted on Google Drive

Topics

- Everything we have covered this semester
- Heavy emphasis on new material since midterm

Topics since the midterm

- Google File System (overlap, w/midterm)
- MapReduce
- OS and Parallelism
- Cloud
- DNS and CDNs
- Replication and Scaling
- Information Retrieval
- Machine learning
- Recommender systems
- Auctions
- Ethics
- The Dark Web

Google File System: Question

- What's an atomic operation?
- What's an example of an atomic GFS operation? Why do we need it?
- How are atomic operations implemented in GFS?
- Why not make all GFS operations atomic?

Google File System: Answer

- What's an atomic operation?
  - An atomic operation appears to the rest of the system as if it happen instantaneously
- What's an example of an atomic GFS operation?
  - File creation
  - Avoid two files with the same name
- How are atomic operations implemented?
  - Serialize through the master node
- Why not make all GFS operations atomic?
  - It's slow

MapReduce: Question

- Write map and reduce functions for generating a reverse web-link graph
  - Input (URL, content)
  - Output: (target_URL, list(source_URL))
  - Assume you have a extract_urls() function

MapReduce: Answer

- Write map and reduce functions for generating a reverse web-link graph
  - Input (URL, content)
  - Output: (target_URL, list(source_URL))
  - Assume you have a extract_urls() function

```
map(url, content):
  source = url
  for target in extract_urls(content):
    EmitIntermediate(target, source)

reduce(target, list):
  Emit(target, list)
```

Information Retrieval: Question

- Explain why 0.85 might be a reasonable value for d. What change in user behavior would account for a rise in d? What about a decline? What would you expect to happen for surfers on mobile phones?

- Give several reasons that PageRank might be a misleading guide to page popularity.

Information Retrieval: Answer

- Explain why 0.85 might be a reasonable value for d. What change in user behavior would account for a rise in d? What about a decline? What would you expect to happen for surfers on mobile phones?
- d is roughly the percentage of the time that people click, rather than type something into the URL. If d goes up, it might mean that people are lazier about typing. If d goes down, then it means people are newly enthusiastic about entering URLs. Because text entry on a phone is relatively difficult, we'd expect d to go up.

Information Retrieval: Answer

- Give several reasons that PageRank might be a misleading guide to page popularity.
- Links are now generated by social and technical mechanisms, and are no longer a reliable "vote for quality"
- Spammers have learned to create link farms, again undermining the informative quality of the hyperlink.
- Client technologies like AJAX have rendered traditional surfing somewhat obsolete. People don't click like they used to.

OS and Parallelism: Question

- What is the CAP theorem?
- Give an example that could cause a problem with each element of the theorem

OS and Parallelism: Answer

- What is the CAP theorem? Give examples.
- Pick two:
- Consistency
  - Web caching, DNS - might not be consistent if an address is stale
- Availability
  - When someone posts a review on TripAdvisor other users can’t see the review for at least a day
- Partitioning
  - Single node MySQL server - don’t need to worry about partitioning, so it will always be available and consistent as long as the server doesn’t go down

Replication and Scaling: Question

- How could we fix these problems? (from previous slide):
  - DNS lookup could return a server that is down
  - Caching could lead to server overload
    - If there is server overload you could add more servers but people would have cached the old ones so this would not help (unless time-to-live of cache is up)
  - Doesn't take into account how long a user will be on a website which can affect server load, or how much work the server is doing at the moment

Replication and Scaling: Question

- How could we fix these problems?
- Load balancer
- Better yet, round robin DNS to multiple load balancers

DNS and CDNs: Question

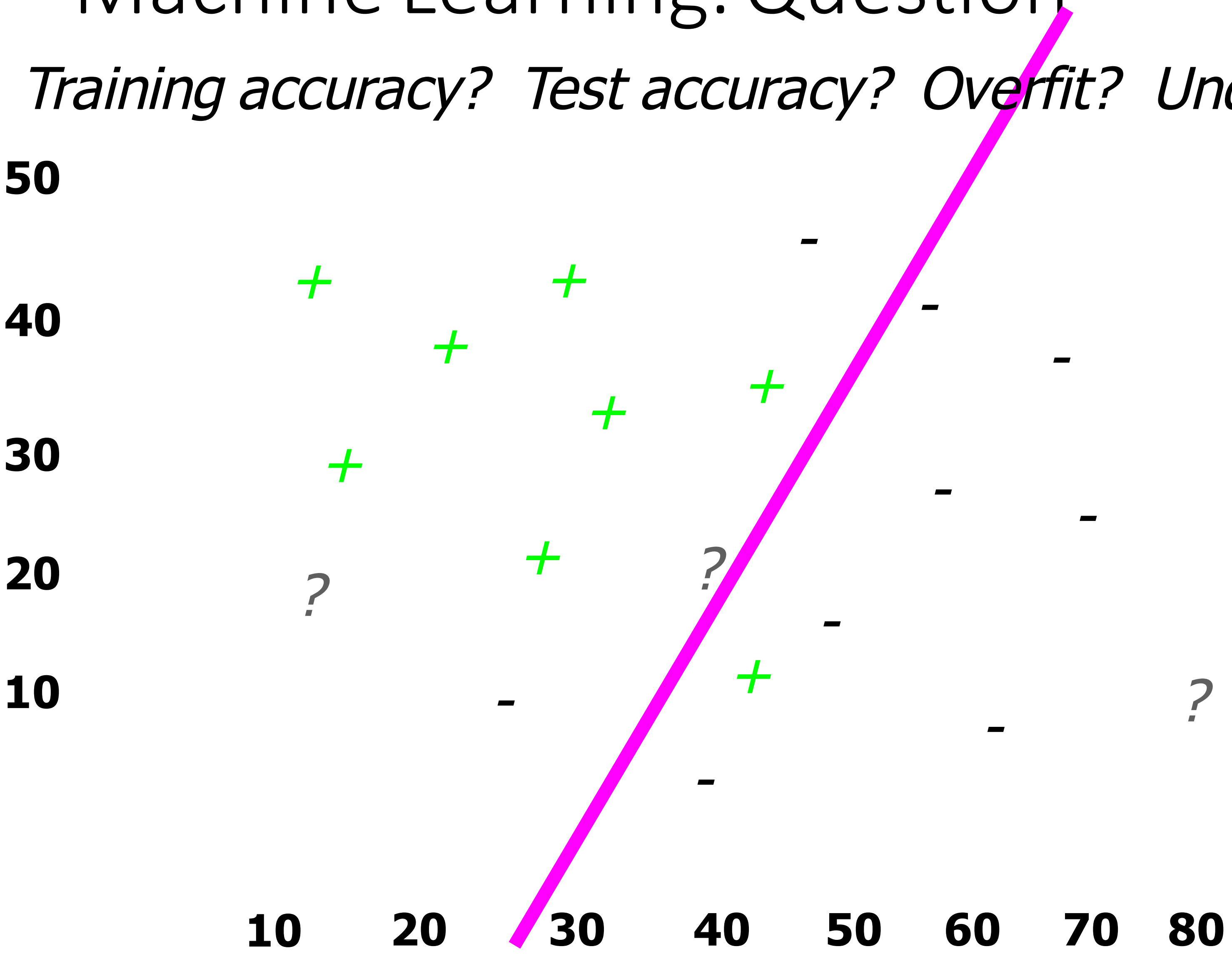
- What are some downsides to Round Robin DNS?

DNS and CDNs: Answer

- What are some downsides to Round Robin DNS?
- DNS lookup could return a server that is down
- Caching could lead to server overload
  - If there is server overload you could add more servers but people would have cached the old ones so this would not help (unless time-to-live of cache is up)
- Doesn't take into account how long a user will be on a website which can affect server load, or how much work the server is doing at the moment

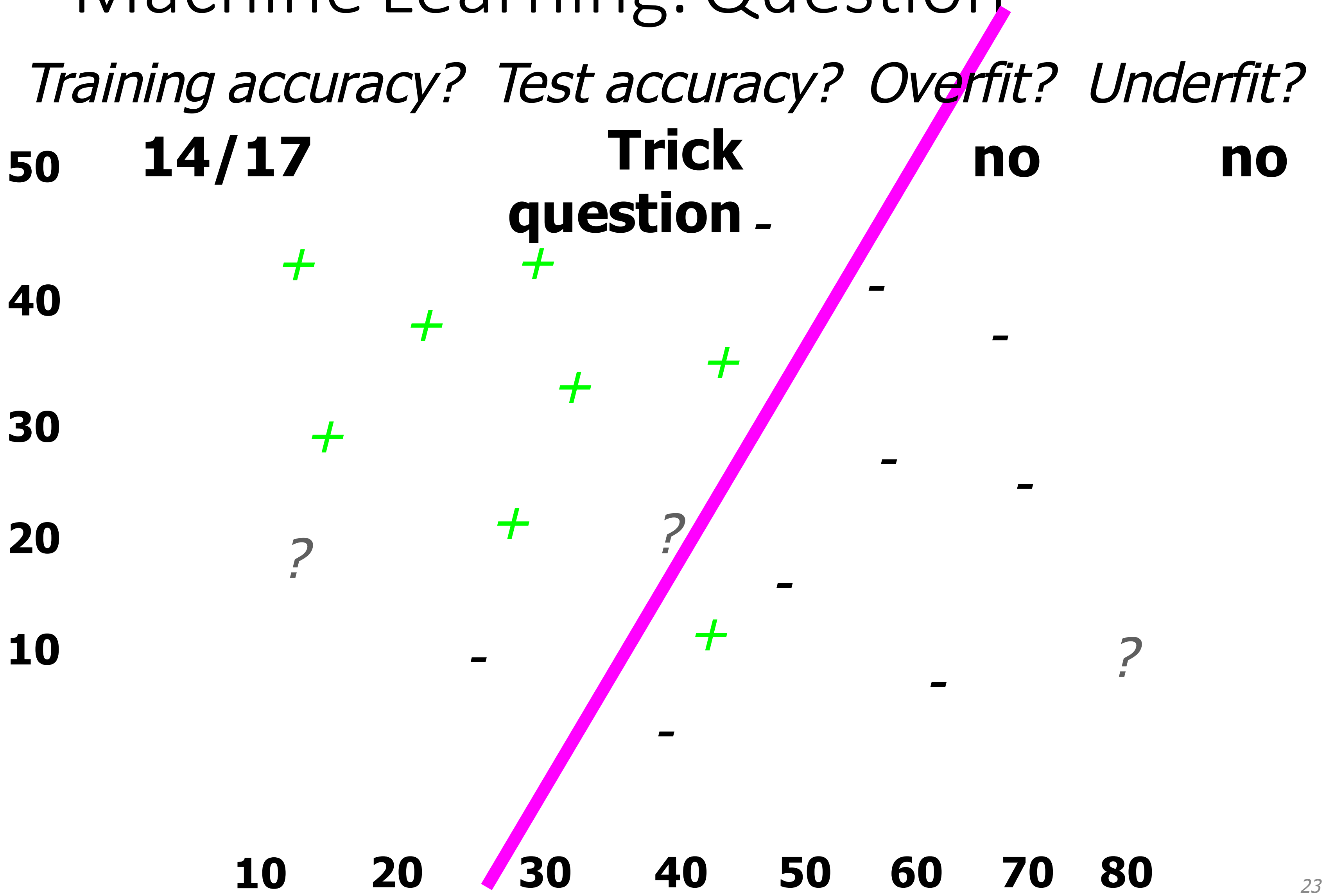
Machine Learning: Question

Training accuracy? Test accuracy? Overfit? Underfit?



Machine Learning: Question

Training accuracy?	*Test accuracy?*	*Overfit?*	*Underfit?*
14/17	**Trick question**	**no**	**no**



Auctions: Question

- Recall that "Sniping" is the Ebay practice of increasing one's bid dramatically right before the end of the auction. Recall that Ebay auctions are similar to sealed-bid second-price auctions (aka, Vickrey auctions).
- Why is sniping nonsensical, from the perspective of standard auction/game-theoretic analysis of Vickrey auctions?
- Give at least one possible reason why sniping is nonetheless observed in practice. For each reason, what part of the analysis of a Vickrey auction is flawed?

Auctions: Answer

- Answer: Why is sniping nonsensical?
- A Vickrey auction bidder should always bid their true valuation. The temporal ordering of the bid shouldn't matter. If the bidder has bid the true valuation, and is currently not the winner of the auction, then entering a higher last-minute bid would result in negative utility for the bidder.

Auctions: Answer

- Why does sniping happen in practice?
- Bidders do not actually know the best valuation of an object, and that the auction process itself reveals the true valuation. In this case, it makes sense to snipe because a late bidder will have more information about the true valuation.
- By not entering a bid earlier in the auction, the bidder can deny other participants the informational benefit of the bid. So even if the bidder knows his own valuation perfectly and acts accordingly, it can make sense to engage in sniping if other bidders do not.

Auctions: Answer

- Why does sniping happen in practice?
- People snipe to enforce bidding strategies that are not possible using Ebay's tools, such as wanting to win "M of N" auctions.
- There are also behavioral theories about people becoming so excited about the auction that they misjudge their own true valuation of the object, etc.

Recommender Systems: Question

- Why is movie-recommendation is different from Web-page recommendation? Give a few reasons.

Recommender Systems: Answer

- Why is movie-recommendation is different from Web-page recommendation? Give a few reasons.
- Many relevant features are hard to extract from movies (acting, cinematography, fun, etc.)
- People's preferences differ, so a standard training set will not necessarily be helpful. Making matters worse, most individuals will not add much personalized training data.

Ethics: Question

- After the GermanWings plane crash in 2015, where a mentally ill pilot intentionally crashed his aircraft in the Alps killing himself and all on board, there was widespread criticism of German privacy laws, which shielded the pilot's medical record (and mental illness) from his employer (Lufthansa).
- Based on the data ethics framework discussed in class, why is this criticism not justified?

Ethics: Answer

- Two reasons:
- 1. Improper repurposing of data (medical record), to determine fitness-to-fly, without notice to employee (the pilot).
- 2. Data validity. If pilots know that their medical record will be used to determine fitness-to-fly, they will either avoid seeking medical care, or try to obtain it "outside the system", to avoid having troublesome entries in their record. We will end up with a system where pilots do not get good treatment for certain conditions, such as some mental illnesses, and the data doesn't show this illness anyway.