The Hall of Shame

James S. Plank
CS494 / CS594

Part of this class is about giving technical presentations, and an important part of giving a technical presentation is the preparation of your slides. You need to make your slides readable, uncluttered, and easy to parse.

There is a tendency to be lazy: To simply cut and paste a graphic from your technical paper, or to shove your data into Excel and then paste the ensuing graph into powerpoint. You need to resist this temptation.

These slides are gleaned from a few sources: Mostly, they are student presentation slides from earlier semesters, but some are from other sources. **They illustrate what you are not to do**. You don't want your slide here.
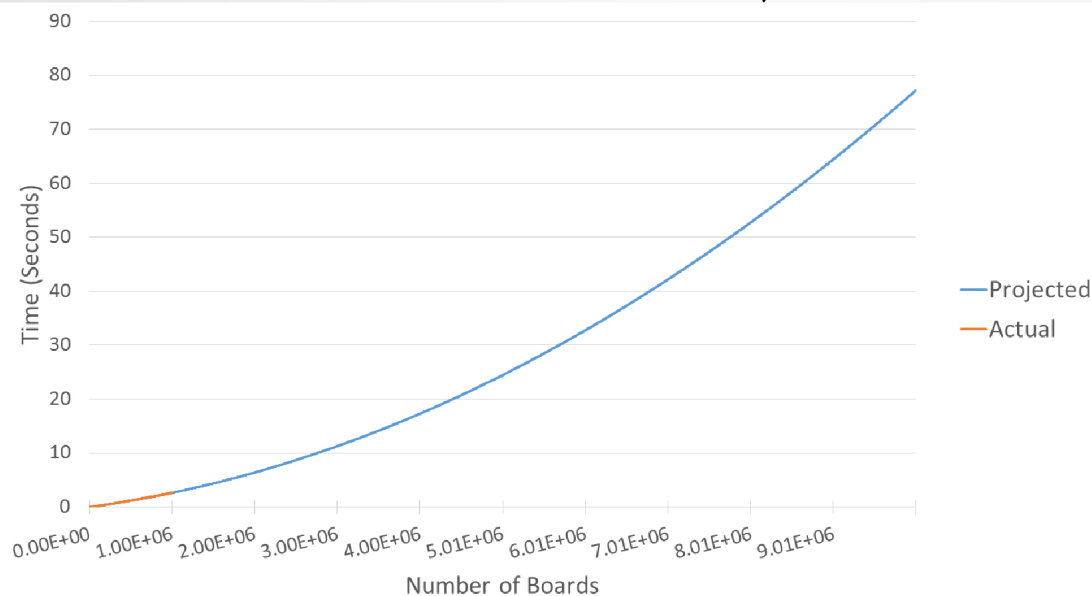
Problems:
- Font sizes in the graph too small.
- X axis labels are hard to read – 100,000, 200,000, … would be easier.
- X axis labels are nonsensical: 5.01E+06?

Running Time:

- Inserting into the multimap is $O(n * \log(n))$.

- CoverHole is $O(n)$

- So the program is $O(n^2 * \log(n))$.

- My solution reaches 2 seconds at 700,000

Intel i7-2760QM
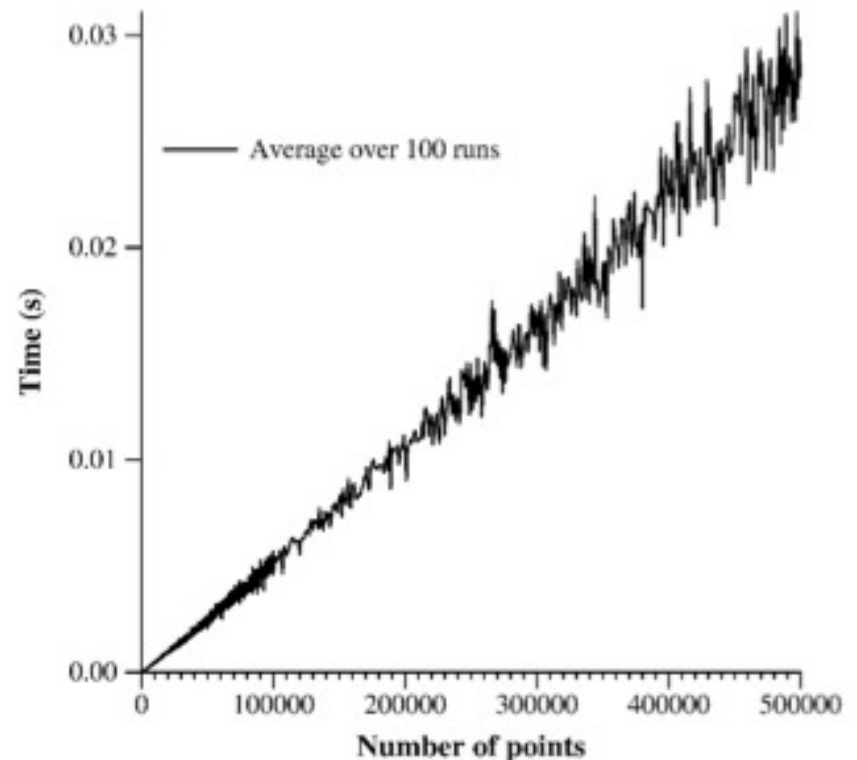2.4 GHz
16 GB Ram
No Optimizations

9

Source: CS494 talk in 2015.

Performance

65.800						
49.350						
32.900						
16.450						
00.000						

0   2   4   6   8   10   12
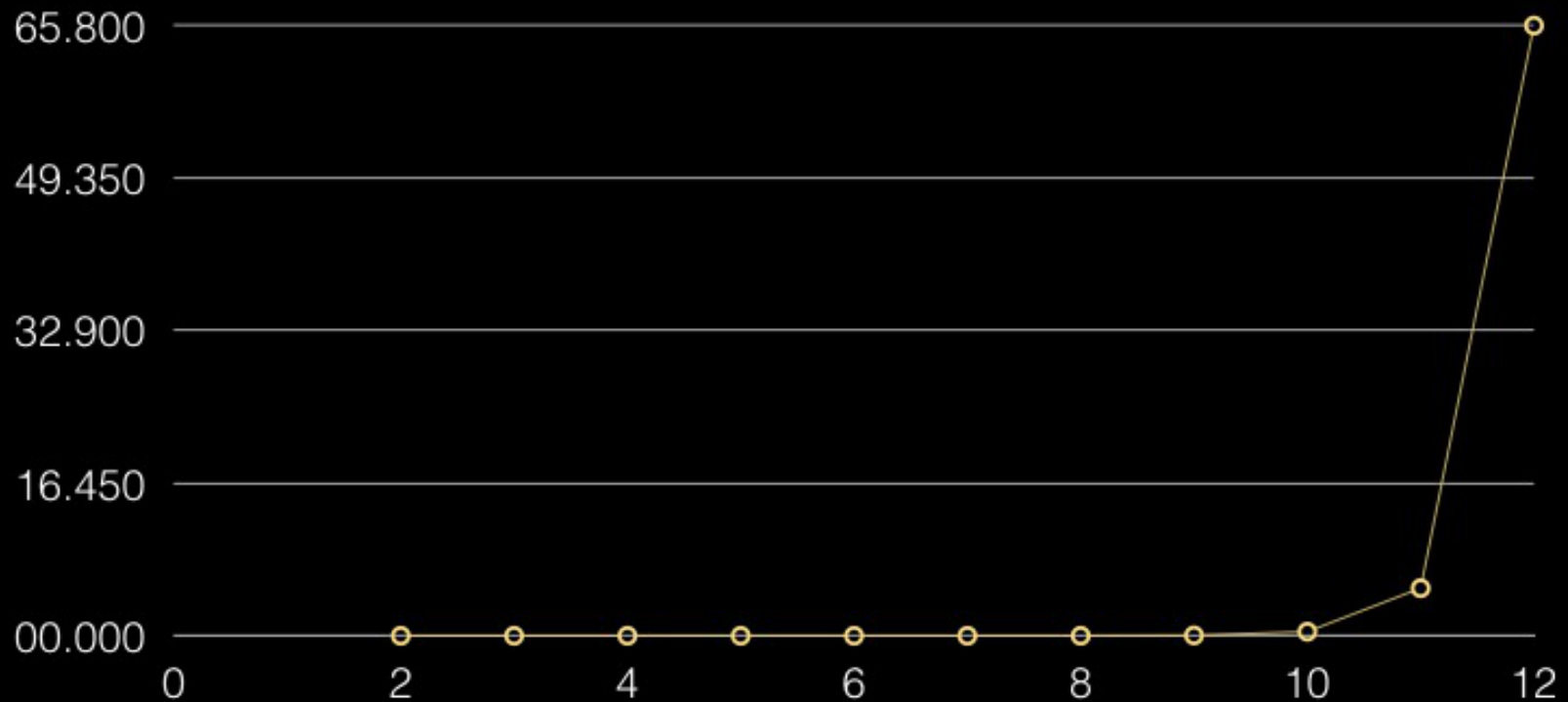
Source: CS494 talk in 2015.

Problems:
- Why use a log axis to show a linear curve?
- Grid lines look really cluttered and junky.
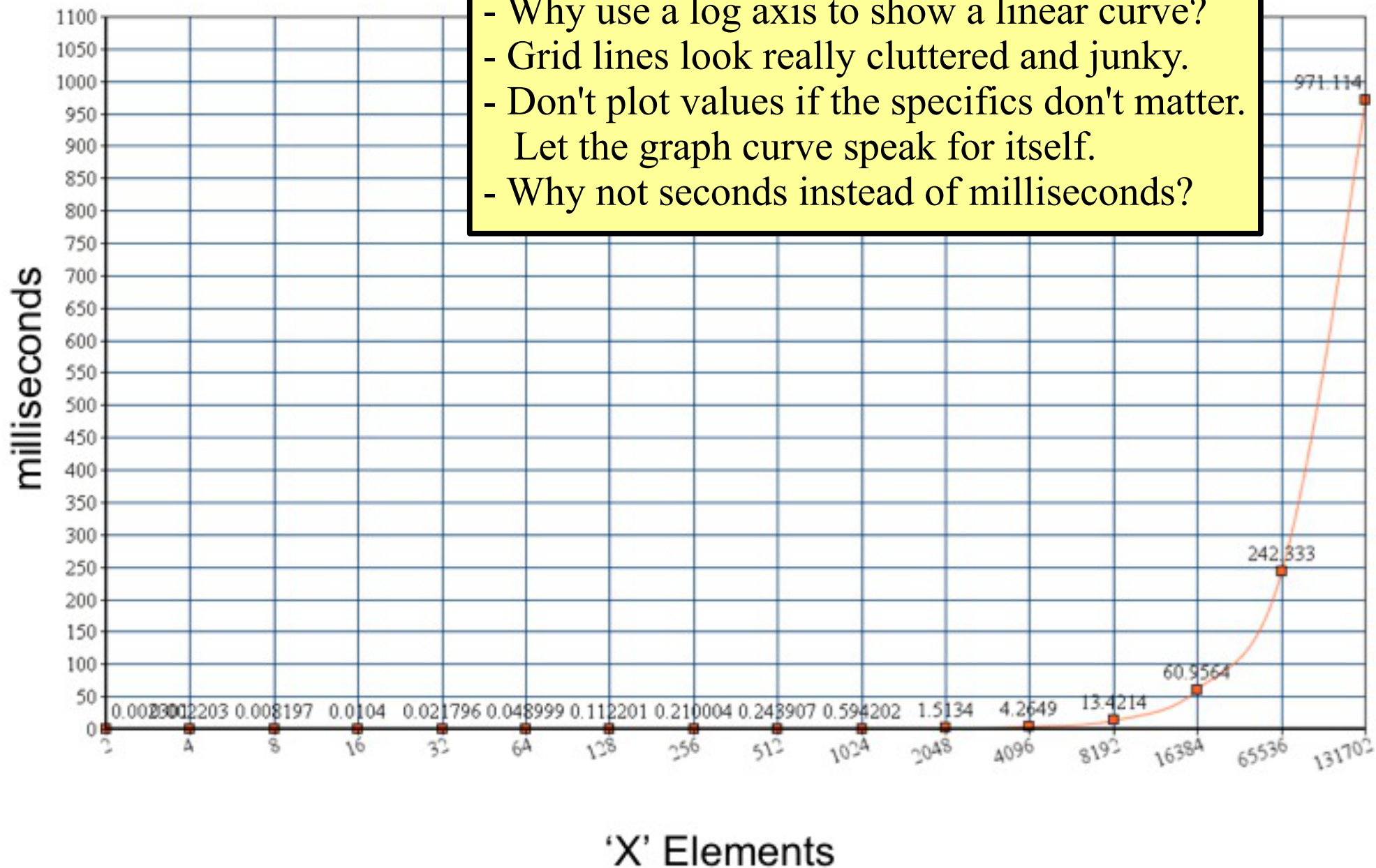- Don't plot values if the specifics don't matter.
  Let the graph curve speak for itself.
- Why not seconds instead of milliseconds?

Source: CS494 talk in 2015.

Timing

Source: CS494 talk in 2015.

Problems:
- OMG – what are those values on the X axis?
- Why a legend for only one curve?

Experiment

MacBook Pro
2.3 GHz Intel Core i7

Time (seconds)

6
5
4
3
2
1
0

Unoptimized

2
53,890,313
112,536,241
172,974,047
234,578,599
297,027,277
360,174,929
423,880,819
488,071,841

Divisor

Spike at ~130 million iterations

Source: Reddit post after the 2017 eclipse
(https://www.reddit.com/r/dataisbeautiful/comments/6v8a9m/oc_my_eyes_hurt/)
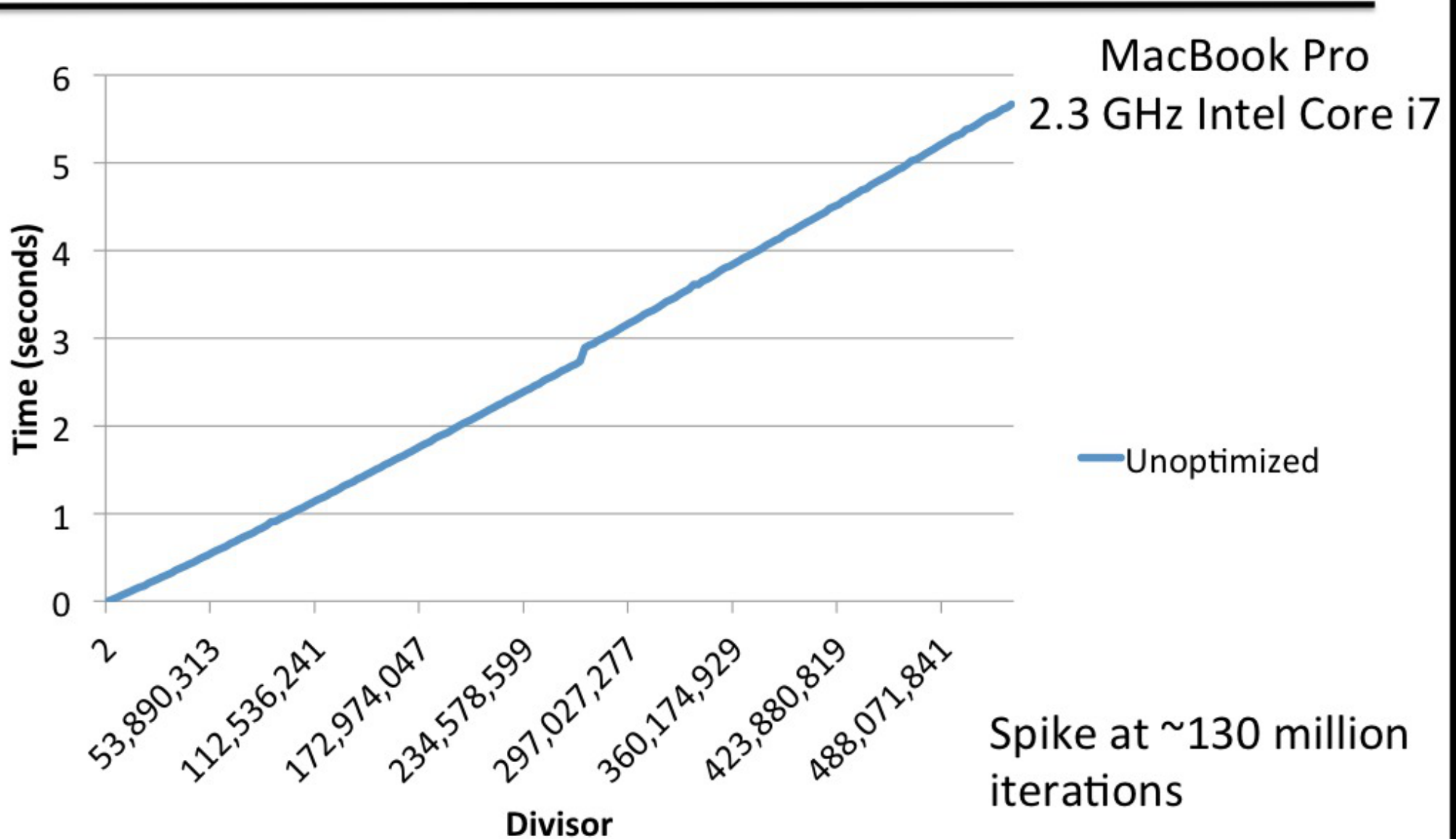

Search Term Popularity in New York

Problems:
- Y axis labels too small and don't have units.
- If you are plotting percentages, and the values can't go over 100%, don't have 120 on the graph.
- Lazy x axis labels (9:00, 10:00, etc)

Source: Google Trends

—"solar eclipse"   —"my eyes hurt"

Running Time, co

In practice, the algorithm t

on most sets.

- Generating new elements will over-power

validation most of the time.

Running Time

Green line: N-K = 2
Blue Line: N-K = 0

MacBook Pro

2.9 Ghz
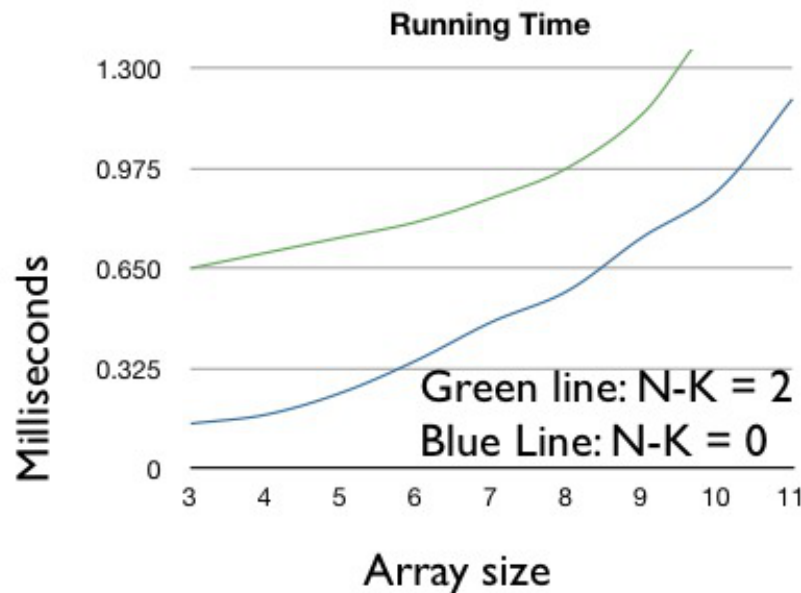
Problems:
- Fonts too small, again.
- Unnatural y axis labels, again.
- Why write out "green line" and
  "blue line". Just use a legend.
- Center the axis units.

Algorithm

- Run through the vector of badgers
  - If badger[i] > 200, increment answer
  - Else, add −badger[i] to a multiset
- While there are still badgers:
  - If the weight of the current bin is 0, pull the first badger off the set and delete it
  - If the weight is greater than 200, increment answer and set bin weight to 0
  - Else we call *lower_bound()* on −(300 - the current weight of the bin)
    - If *lower_bound()* returns end, we are either out of elements or the elements left are too big
    - Otherwise we add the badger's weight and delete It from the set

Performance

Hydra machine   O(nlog(n))

Time (seconds) — y-axis: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120

x-axis: 50,000 | 100,000 | 500,000 | 5,000,000 | 10,000,000 | 20,000,000 | 30,000,000 | 40,000,000 | 50,000,000 | 60,000,000
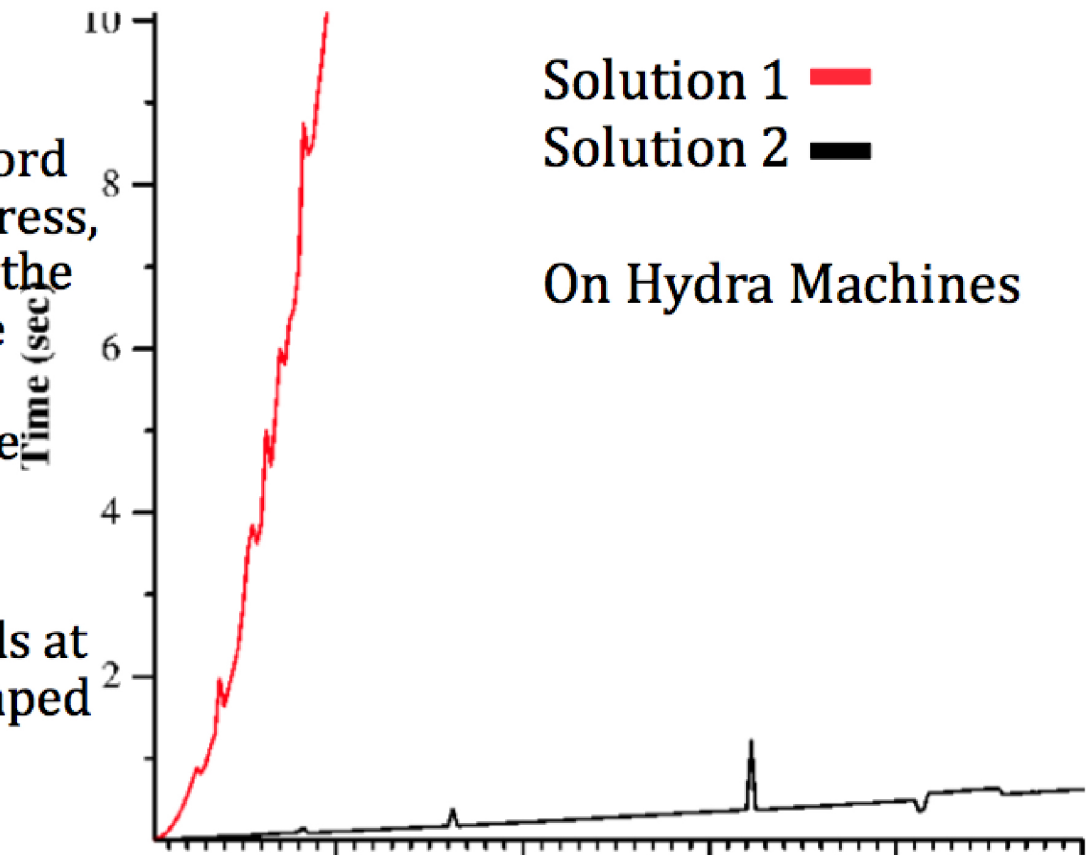
n (size of vectors)
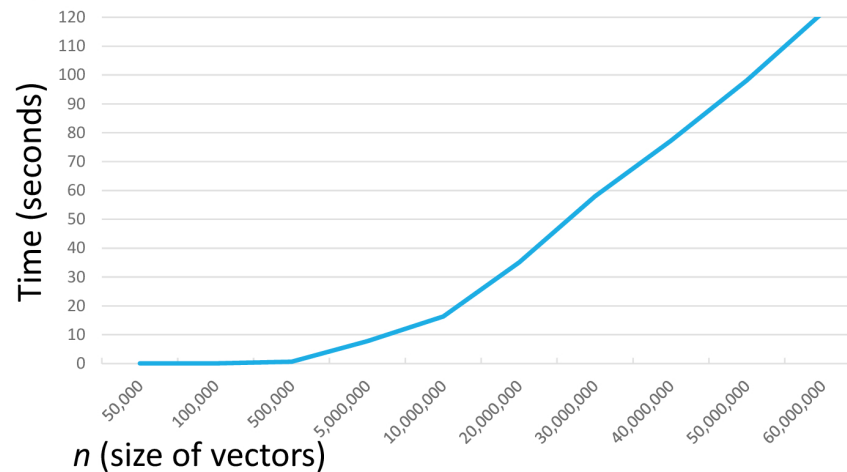
Handles up to *n* = 2^26 before inconsistently returning the correct value. Returns *n* = 50 (constraint) in negligible time (~0 seconds).

- A clear screen-shot disaster.
- No x axis values or units.
- Y axis units/label colliding with the text.
- Too many words in the first bullet.
- I wrote jgraph; While I'm happy that you're
  using it, please don't incorporate it into a disaster...

Running Time cont.

- Tests were run with each Address containing 1 Keyword shared by the previous Address, 1 other Keyword shared by the next Address and 50 unique Keywords. The initial Dangerous Keyword was the "next" Keyword on the last Address. Threshold was 1.

- Solution 2 took ~2.5 seconds at 17000 Addresses, then dumped core.

Solution 1 ▬
Solution 2 ▬

On Hydra Machines

Graph

- Fonts too small. Again.
- OMG, look at that x axis – the labels run into each
   other, and it's not linear or logarithmic...
- Values are unnecessary, and clutter up the graph.
- Ditto the lines in the graph.

- Fonts too small. Again.
- Why is -0.1 on an axis of running times?
- Use 0 on the y axis.
- "2*O(n)" and "O(n-1)" do not make your
  CS302 professor happy.
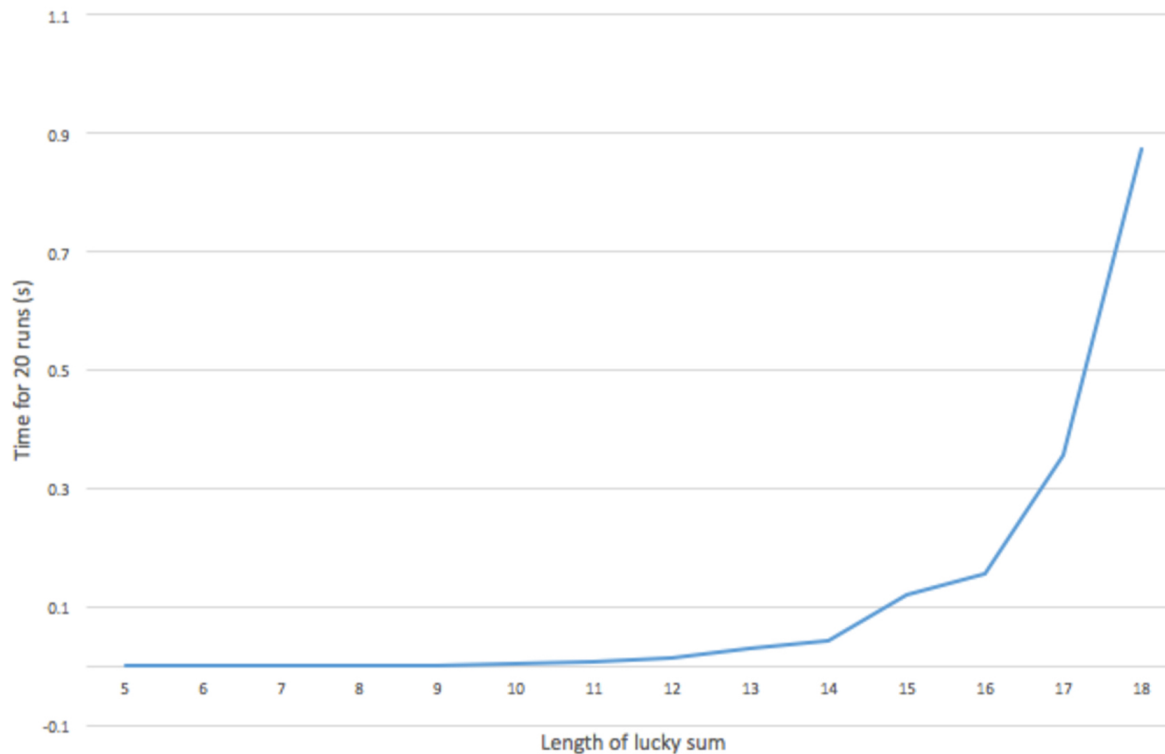
- 2*O(n) first recursive function
- O(n-1) second
- O(n²) overall

Run on 2.5GHz i5-3210m

Why do we need mo... computers?

- **TOO MUCH STUFF ON THE SLIDE.**
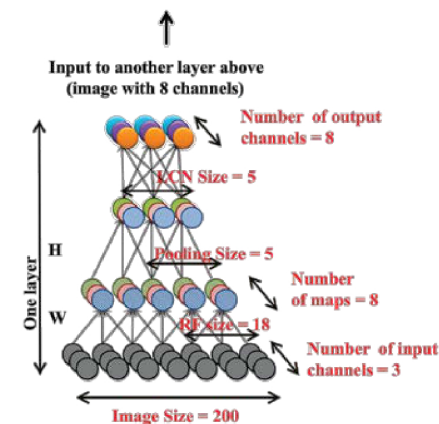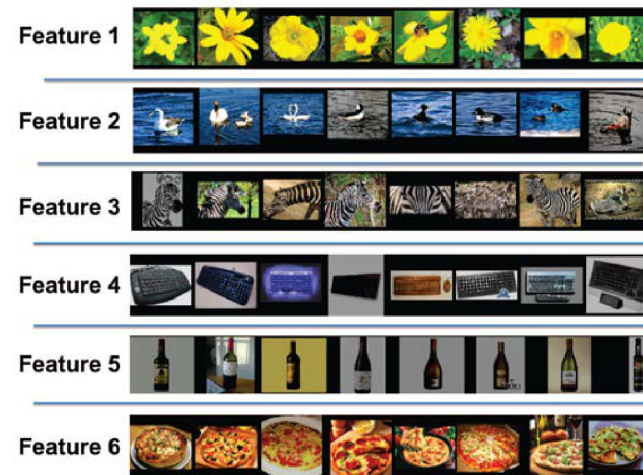- Unreadable pictures.

- **Google Deep Learning Study**
  - **16000 core, 1000 machine GPU cluster**
  - **Trained on 10 million 200x200 pixel images**
  - **Training required 3 days**
  - **Training dataset size: no larger than what can be trained in 1 week**
- **What would they like to do?**
  - **~2 billion photos uploaded to internet per day (2014)**
  - **Can we train a deep net on <u>one day of image data</u>?**
  - **Assume 1000x1000 nominal image size, linear scaling (both assumptions are unrealistically optimistic)**
  - *Requires 5 ZettaIPS to train in 3 days (ZettaIPS=10^{21} IPS; ~5 billion modern GPU cores)*
  - **World doesn't produce enough power for this!**
  - **Data is increasing exponentially with time**
- **Need >10^{16}-10^{18} instruction-per-second on 1 IC**
  - **Less than 10 fJ per instruction energy budget**

Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Feature 6

Input to another layer above
(image with 8 channels)

Number of output channels = 8

LCN Size = 5

H

Pooling Size = 5

Number of maps = 8

One layer

W

RF Size = 18

Number of input channels = 3

Image Size = 200

Q. Le, IEEE ICASSP 2013