CSE 332: Data Abstractions

Lecture 3: Asymptotic Analysis

Ruth Anderson

Winter 2015

Announcements

- **Project 1** – phase A due Mon, phase B due Thurs
- **Homework 1** – due Friday

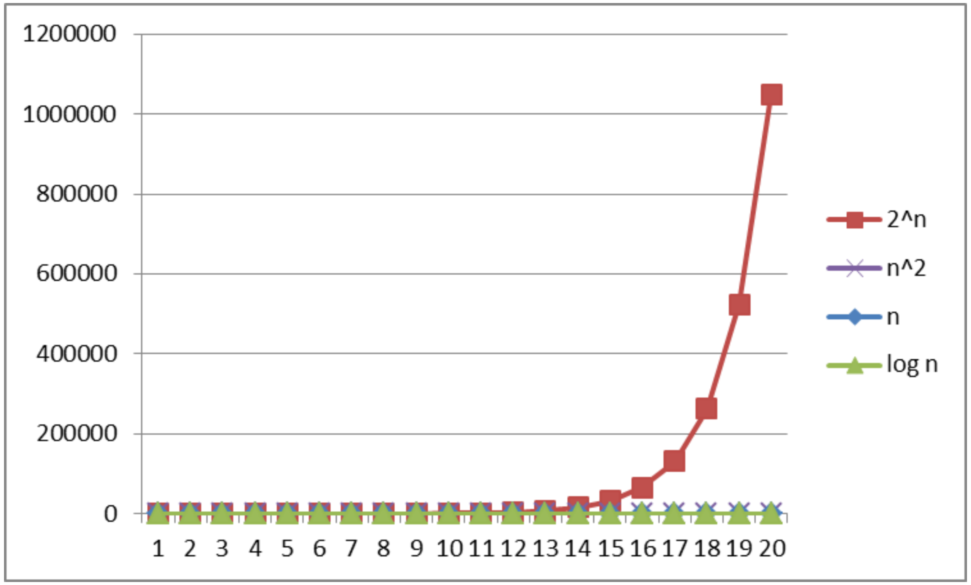
Today

- Analyzing code
- Big-Oh

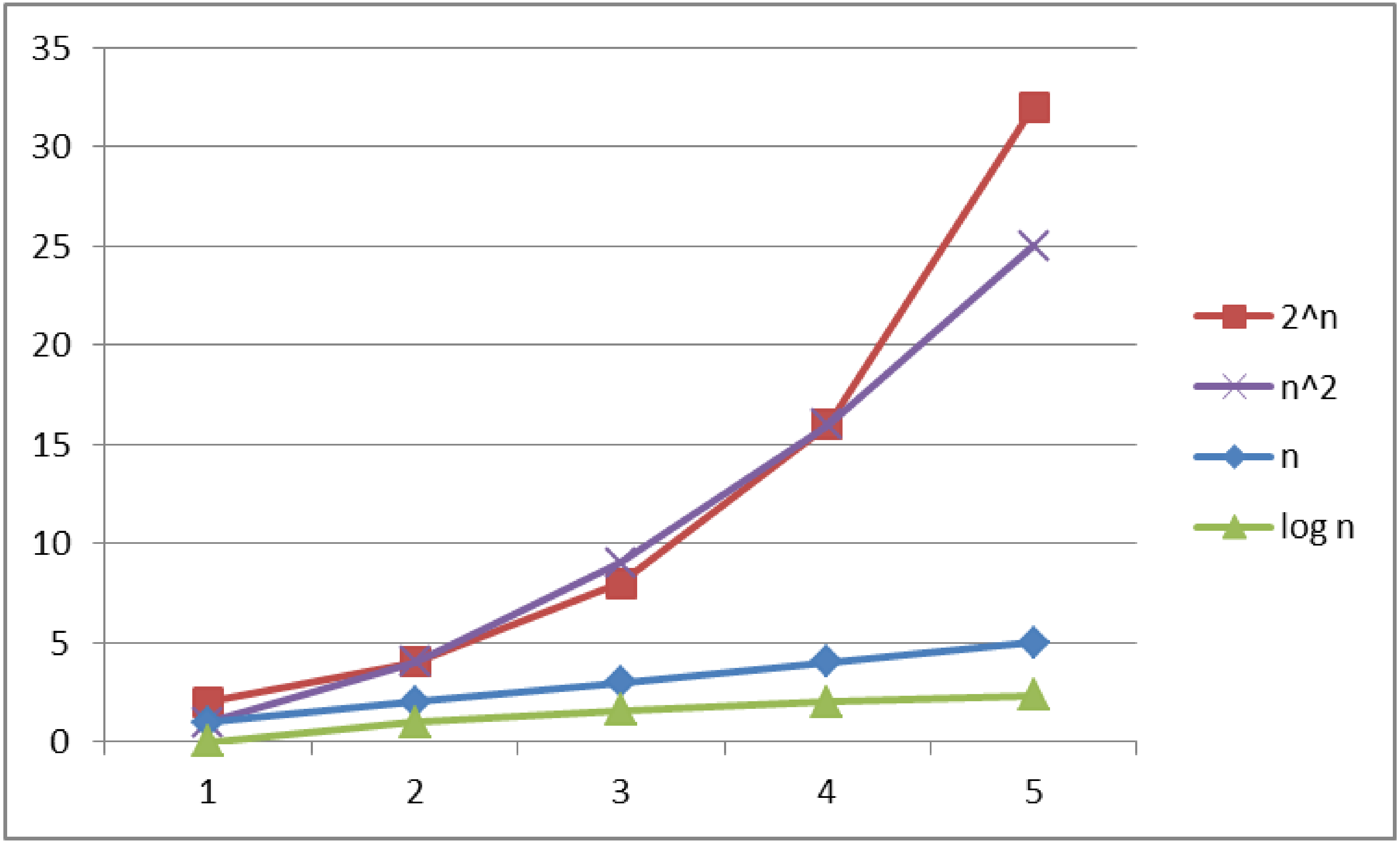
Logarithms and Exponents

- Since so much is binary in CS, **`log`** almost always means $\mathbf{log_2}$
- Definition: $\mathbf{log_2\ x = y}$ if $\mathbf{x = 2^y}$
- So, $\mathbf{log_2}$ 1,000,000 = "a little under 20"
- Just as exponents grow *very* quickly, logarithms grow *very* slowly

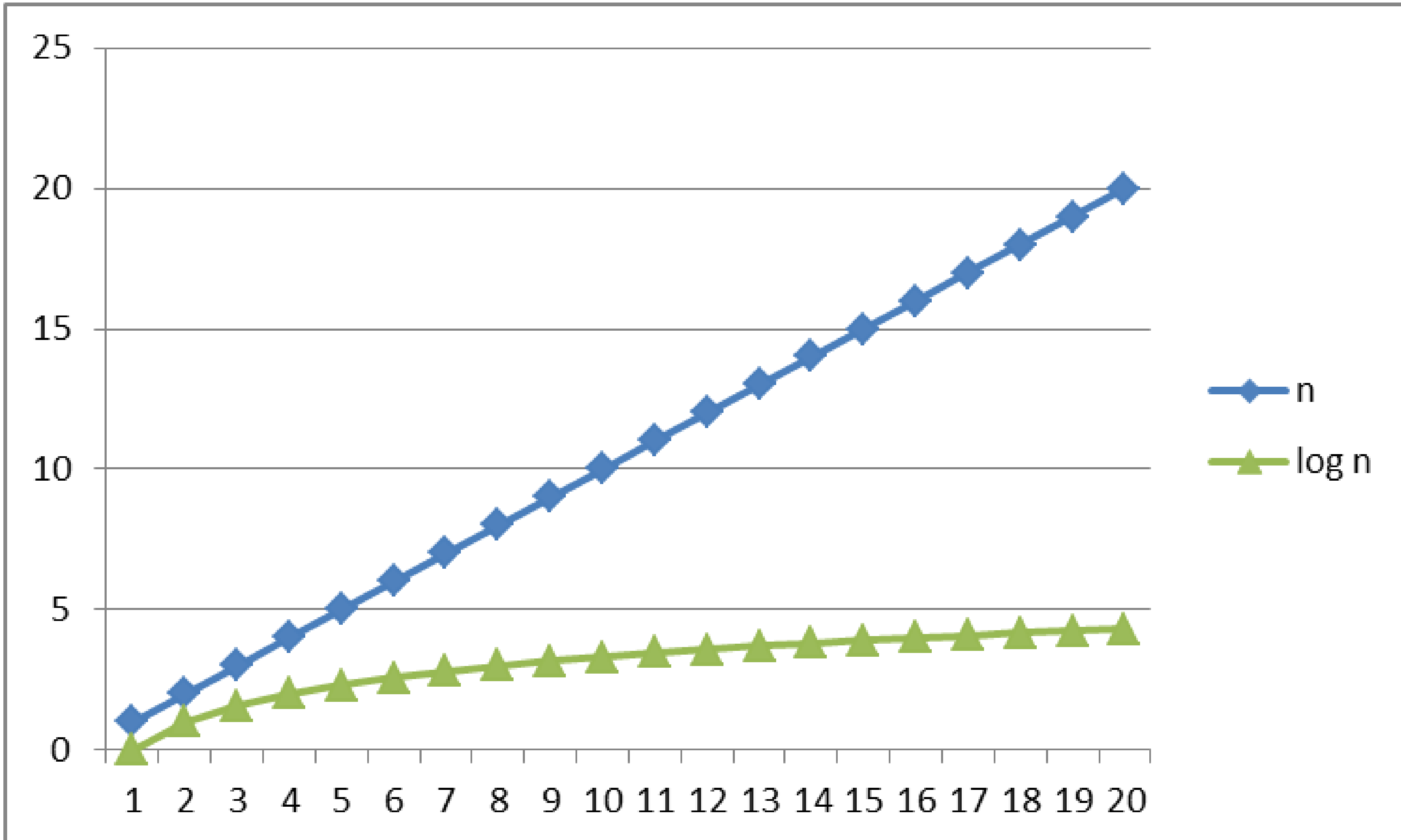
See Excel file for plot data – play with it!



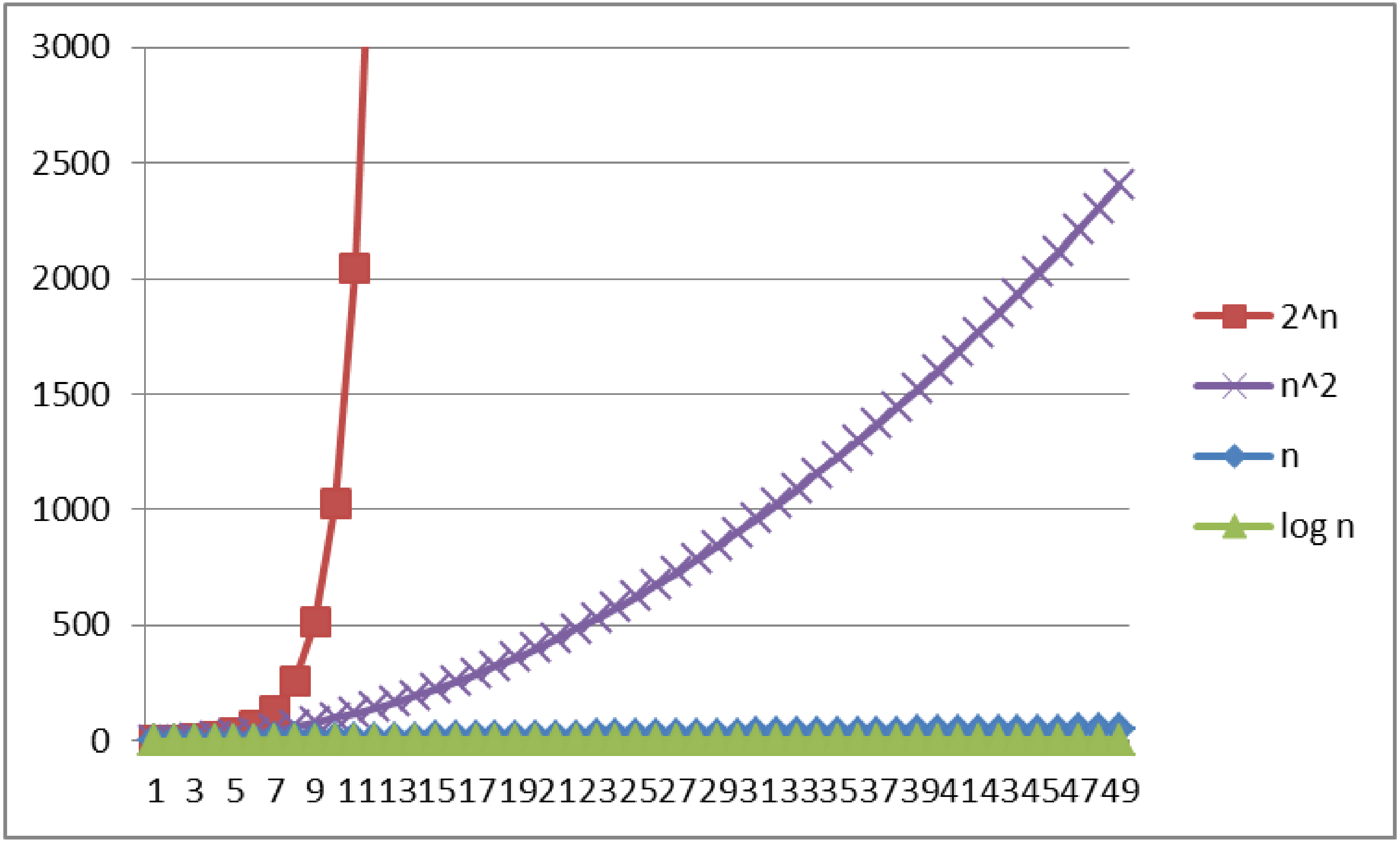
Logarithms and Exponents



Logarithms and Exponents



Logarithms and Exponents



Asymptotic notation

About to show formal definition, which amounts to saying:

1. Eliminate low-order terms
2. Eliminate coefficients

Examples:

- $4n + 5$
- $0.5n \mathbf{\texttt{log}}\ n + 2n + 7$
- $n^3 + 2^n + 3n$
- $n \mathbf{\texttt{log}}\ (10n^2)$

Examples

True or false?

1. $4+3n$ is $O(n)$
2. $n+2\log n$ is $O(\log n)$
3. $\log n+2$ is $O(1)$
4. n^{50} is $O(1.1^n)$

Notes:

- Do NOT ignore constants that are not multipliers:
  - n^3 is $O(n^2)$: FALSE
  - 3^n is $O(2^n)$: FALSE
- When in doubt, refer to the definition

Examples

True or false?

1. $4+3n$ is $O(n)$ — True
2. $n+2\log n$ is $O(\log n)$ — False
3. $\log n+2$ is $O(1)$ — False
4. n^{50} is $O(1.1^n)$ — True

Notes:

- Do NOT ignore constants that are not multipliers:
  - n^3 is $O(n^2)$: FALSE
  - 3^n is $O(2^n)$: FALSE
- When in doubt, refer to the definition

Big-Oh relates functions

We use O on a function f(n) (for example n^2) to mean *the set of functions with asymptotic behavior* ***less than or equal to*** f(n)

So ($3n^2$+17) **is in** $O(n^2)$

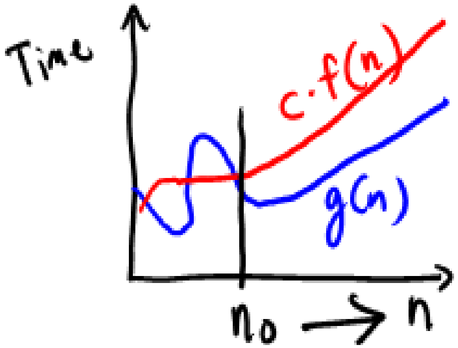
- $3n^2$+17 and n^2 have the same asymptotic behavior

Confusingly, we also say/write:

- ($3n^2$+17) **is** $O(n^2)$
- ($3n^2$+17) = $O(n^2)$

But we would never say $O(n^2)$ = ($3n^2$+17)

Formally Big-Oh



Definition: **g(*n*)** is in O(**f(*n*)**) iff there exist positive constants ***c*** and $\boldsymbol{n_0}$ such that

g(*n*) ≤ ***c*** **f(*n*)** for all $\boldsymbol{n} \geq \boldsymbol{n_0}$

To show **g(*n*)** is in O(**f(*n*)**), pick a ***c*** large enough to "cover the constant factors" and $\boldsymbol{n_0}$ large enough to "cover the lower-order terms"

- Example: Let **g(*n*)** $= 3n^2+17$ and **f(*n*)** $= n^2$

  c = 5 and $\boldsymbol{n_0}$ = 10 is more than good enough

This is "less than or equal to"

- So $3n^2+17$ is also $O(n^5)$ and $O(2^n)$ etc.

Using the definition of Big-Oh (Example 1)

For **g(n)** = 4n & **f(n)** = n^2, prove **g(n)** is in O(**f(n)**)

- A valid proof is to find valid **c** & $\mathbf{n_0}$
- When n=4, **g(n)** =16 & **f(n)** =16; this is the crossing over point
- So we can choose $\mathbf{n_0}$ = 4, and **c** = 1

- Note: There are many possible choices:
  ex: $\mathbf{n_0}$ = 78, and **c** = 42 works fine

The Definition: g(*n*) is in O(f(*n*)) iff there exist *positive* constants *c* and n_0 such that

$$g(n) \leq c\, f(n) \text{ for all } n \geq n_0.$$

Using the definition of Big-Oh (Example 2)

For **g(n)** = n^4 & **f(n)** = 2^n, prove **g(n)** is in O(**f(n)**)

- A valid proof is to find valid **c** & $\mathbf{n_0}$
- One possible answer: $\mathbf{n_0}$ = 20, and **c** = 1

The Definition: g(*n*) is in O(f(*n*)) iff there exist *positive* constants *c* and n_0 such that

$$g(n) \le c\, f(n) \text{ for all } n \ge n_0.$$

*What's with the **c**?*

- To capture this notion of similar asymptotic behavior, we allow a constant multiplier (called **c**)
- Consider:

  g(n) = 7n+5

  f(n) = n
- These have the same asymptotic behavior (linear), so **g(n)** is in O(**f(n)**) even though **g(n)** is always larger
- There is no positive n_0 such that **g(n)** ≤ **f(n)** for all n ≥ n_0
- The '**c**' in the definition allows for that:

  $g(n) \leq c\, f(n)$ for all $n \geq n_0$
- To prove **g(n)** is in O(**f(n)**), have **c** = 12, n_0 = 1

What you can drop

- Eliminate coefficients because we don't have units anyway
  - $3n^2$ versus $5n^2$ doesn't mean anything when we have not specified the cost of constant-time operations (can re-scale)

- Eliminate low-order terms because they have vanishingly small impact as n grows

- Do NOT ignore constants that are not multipliers
  - n^3 is not $O(n^2)$
  - 3^n is not $O(2^n)$

(This all follows from the formal definition)

Big Oh: Common Categories

From fastest to slowest

$O(1)$	constant (same as $O(k)$ for constant k)
$O(\log n)$	logarithmic
$O(n)$	linear
$O(n \log n)$	"$n \log n$"
$O(n^2)$	quadratic
$O(n^3)$	cubic
$O(n^k)$	polynomial (where is k is any constant > 1)
$O(k^n)$	exponential (where k is any constant > 1)

Usage note: "exponential" does not mean "grows really fast", it means "grows at rate proportional to k^n for some k>1"

More Asymptotic Notation

- **Upper bound**: O(**f(n)**) is the set of all functions asymptotically less than or equal to **f(n)**
  - g(n) is in O(**f(n)**) if there exist constants c and n_0 such that

    g(n) $\leq$ c **f(n)** for all $n \geq n_0$

- **Lower bound**: Ω(**f(n)**) is the set of all functions asymptotically greater than or equal to **f(n)**
  - g(n) is in Ω(**f(n)**) if there exist constants c and n_0 such that

    g(n) $\geq$ c **f(n)** for all $n \geq n_0$

- **Tight bound**: θ(**f(n)**) is the set of all functions asymptotically equal to **f(n)**
  - Intersection of O(**f(n)**) and Ω(**f(n)**) (use *different* c values)

Regarding use of terms

A common error is to say $O(\ f(n)\ )$ when you mean $\theta(\ f(n)\ )$

- People often say O() to mean a tight bound
- Say we have f(n)=n; we could say f(n) is in O(n), which is true, but only conveys the upper-bound
- Since f(n)=n is *also* $O(n^5)$, it's tempting to say "this algorithm is *exactly* $O(n)$"
- Somewhat incomplete; instead say it is $\theta(n)$
- That means that it is not, for example $O(\mathtt{log}\ n)$

Less common notation:

- "little-oh": like "big-Oh" but strictly less than
  - Example: sum is $o(n^2)$ but not $o(n)$
- "little-omega": like "big-Omega" but strictly greater than
  - Example: sum is $\omega(\mathtt{log}\ n)$ but not $\omega(n)$

What we are analyzing

- The most common thing to do is give an O or θ bound to the worst-case running time of an algorithm

- Example: True statements about binary-search algorithm
  - Common: $\theta(\mathbf{log}\ n)$ running-time in the worst-case
  - Less common: $\theta(1)$ in the best-case (item is in the middle)
  - Less common: Algorithm is $\Omega(\mathbf{log\ log}\ n)$ in the worst-case (it is not really, really, really fast asymptotically)
  - Less common (but very good to know): the find-in-sorted-array ***problem*** is $\Omega(\mathbf{log}\ n)$ in the worst-case
    - *No* algorithm can do better (without parallelism)
    - A ***problem*** cannot be $O(\mathrm{f}(n))$ since you can always find a slower algorithm, but can mean ***there exists*** an algorithm

Other things to analyze

- Space instead of time
  - Remember we can often use space to gain time

- Average case
  - Sometimes only if you assume something about the distribution of inputs
    - See CSE312 and STAT391
  - Sometimes uses randomization in the algorithm
    - Will see an example with sorting; also see CSE312
  - Sometimes an *amortized guarantee*
    - Will discuss in a later lecture

Summary

Analysis can be about:

- The problem or the algorithm (usually algorithm)
- Time or space (usually time)
  - Or power or dollars or …
- Best-, worst-, or average-case (usually worst)
- Upper-, lower-, or tight-bound (usually upper or tight)

Big-Oh Caveats

- Asymptotic complexity (Big-Oh) focuses on behavior for **<u>large *n*</u>** and is independent of any computer / coding trick
  - But you can "abuse" it to be misled about trade-offs
  - Example: $n^{1/10}$ vs. **`log`** *n*
    - Asymptotically $n^{1/10}$ grows more quickly
    - But the "cross-over" point is around $5 * 10^{17}$
    - So if you have input size less than 2^{58}, prefer $n^{1/10}$
- Comparing O() for ***<u>small n</u>*** values can be misleading
  - Quicksort: O(nlogn) (expected)
  - Insertion Sort: $O(n^2)$ (expected)
  - Yet in reality Insertion Sort is faster for small n's
  - We'll learn about these sorts later

Addendum: Timing vs. Big-Oh?

- At the core of CS is a backbone of theory & mathematics
  - Examine the algorithm itself, mathematically, not the implementation
  - Reason about performance as a function of n
  - Be able to mathematically prove things about performance
- Yet, timing has its place
  - In the real world, we do want to know whether implementation A runs faster than implementation B on data set C
  - Ex: Benchmarking graphics cards
  - We will do some timing in project 3 (and in 2, a bit)
- Evaluating an algorithm? Use asymptotic analysis
- Evaluating an implementation of hardware/software? Timing can be useful

Extra slides

Powers of 2

- A bit is 0 or 1
- A sequence of *n* bits can represent 2^n distinct things
  - For example, the numbers 0 through 2^n-1
- 2^{10} is 1024 ("about a thousand", kilo in CSE speak)
- 2^{20} is "about a million", mega in CSE speak
- 2^{30} is "about a billion", giga in CSE speak

Java: an **`int`** is 32 bits and signed, so "max int" is "about 2 billion"

a **`long`** is 64 bits and signed, so "max long" is 2^{63}-1

Therefore…

Could give a unique id to…

- Every person in the U.S. with 29 bits
- Every person in the world with 33 bits
- Every person to have ever lived with 38 bits (estimate)
- Every atom in the universe with 250-300 bits

So if a password is 128 bits long and randomly generated, do you think you could guess it?

Properties of logarithms

- **log(A*B) = log A + log B**
  - So $\log(N^k) = k \log N$
- **log(A/B) = log A – log B**
- $x = \log_2 2^x$
- **log(log x)** is written **log log x**
  - Grows as slowly as 2^{2^y} grows fast
  - Ex:

$$\log_2 \log_2 4\,billion \sim \log_2 \log_2 2^{32} = \log_2 32 = 5$$

- **(log x)(log x)** is written $\log^2 x$
  - It is greater than **log x** for all **x > 2**

Log base doesn't matter (much)

"Any base *B* log is equivalent to base 2 log within a constant factor"

- And we are about to stop worrying about constant factors!
- In particular, $\log_2 x = 3.22 \log_{10} x$
- In general, we can convert log bases via a constant multiplier
- Say, to convert from base A to base B:

$$\log_B x = (\log_A x) / (\log_A B)$$

Algorithm Analysis

As the "size" of an algorithm's input grows
(integer, length of array, size of queue, etc.):

- How much longer does the algorithm take (time)
- How much more memory does the algorithm need (space)

Because the curves we saw are so different, we often only care about "which curve we are like"

Separate issue: Algorithm *correctness* – does it produce the right answer for all inputs

- Usually more important, naturally

Example

- What does this pseudocode return?

```
x := 0;
for i=1 to N do
  for j=1 to i do
     x := x + 3;
return x;
```

- Correctness: For any N ≥ 0, it returns…

Example

- What does this pseudocode return?

```
x := 0;
for i=1 to N do
  for j=1 to i do
     x := x + 3;
return x;
```

- Correctness: For any N ≥ 0, it returns 3N(N+1)/2
- Proof: By induction on *n*
  - *P(n)* = after outer for-loop executes *n* times, **`x`** holds 3n(n+1)/2
  - Base: n=0, returns 0
  - Inductive: From *P(k)*, **`x`** holds 3k(k+1)/2 after *k* iterations. Next iteration adds 3(k+1), for total of 3k(k+1)/2 + 3(k+1) = (3k(k+1) + 6(k+1))/2 = (k+1)(3k+6)/2 = 3(k+1)(k+2)/2

Example

- How long does this pseudocode run?

```
x := 0;
for i=1 to N do
  for j=1 to i do
     x := x + 3;
return x;
```

- Running time: For any N ≥ 0,
  - Assignments, additions, returns take “1 unit time”
  - Loops take the sum of the time for their iterations

- So: 2 + 2*(number of times inner loop runs)
  - And how many times is that?

Example

- How long does this pseudocode run?

```
x := 0;
for i=1 to N do
  for j=1 to i do
     x := x + 3;
return x;
```

- How many times does the inner loop run?

Example

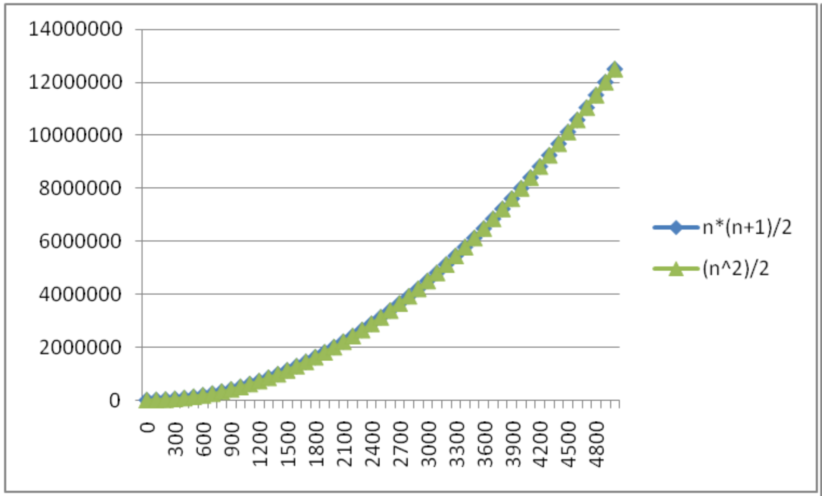
- How long does this pseudocode run?

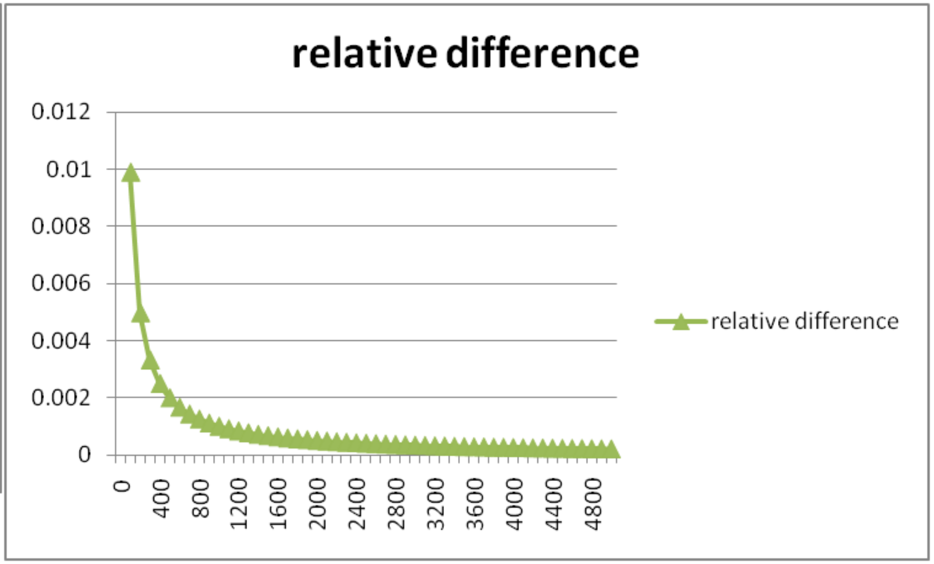
```
x := 0;
for i=1 to N do
  for j=1 to i do
     x := x + 3;
return x;
```

- The total number of loop iterations is N*(N+1)/2
  - This is a very common loop structure, worth memorizing
  - This is *proportional to* N^2 , and we say $O(N^2)$, "big-Oh of"
    - For large enough N, the N and constant terms are irrelevant, as are the first assignment and return
    - See plot… N*(N+1)/2 vs. just $N^2/2$

Lower-order terms don't matter

N*(N+1)/2 vs. just $N^2/2$

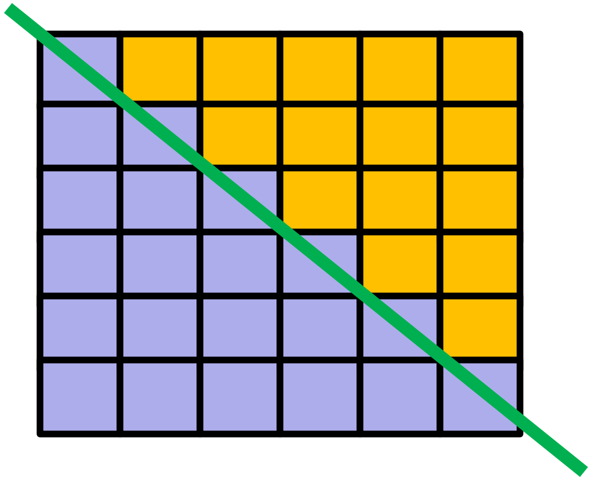




Geometric interpretation

$$\sum_{i=1}^{N} i = N*N/2+N/2$$

```
for i=1 to N do
  for j=1 to i do
     // small work
```



- Area of square: N*N
- Area of lower triangle of square: N*N/2
- Extra area from squares crossing the diagonal: N*1/2
- As N grows, fraction of “extra area” compared to lower triangle goes to zero (becomes insignificant)

Recurrence Equations

- For running time, what the loops did was irrelevant, it was how many times they executed.
- Running time as a function of input size n (here loop bound):

  $$T(n) = n + T(n-1)$$

  (and $T(0)$ = 2ish, but usually implicit that $T(0)$ is some constant)
- Any algorithm with running time described by this formula is $O(n^2)$
- "Big-Oh" notation also ignores the constant factor on the high-order term, so $3N^2$ and $17N^2$ and $(1/1000)\ N^2$ are all $O(N^2)$
  - As N grows large enough, no smaller term matters
  - Next time: Many more examples + formal definitions