Research Methods

CSCI 8901:
Visualizing Your Research

Prof. Tim Wood
GWU
2021

Visuals Matter

Slides, Videos, Posters, Demos

Papers

- Diagrams
- Graphs
- Even fonts and formatting!

You want your work to look:

- Professional
- Attractive
- Memorable
- Informative

Let’s go back in time…

Tim Wood’s Thesis Defense

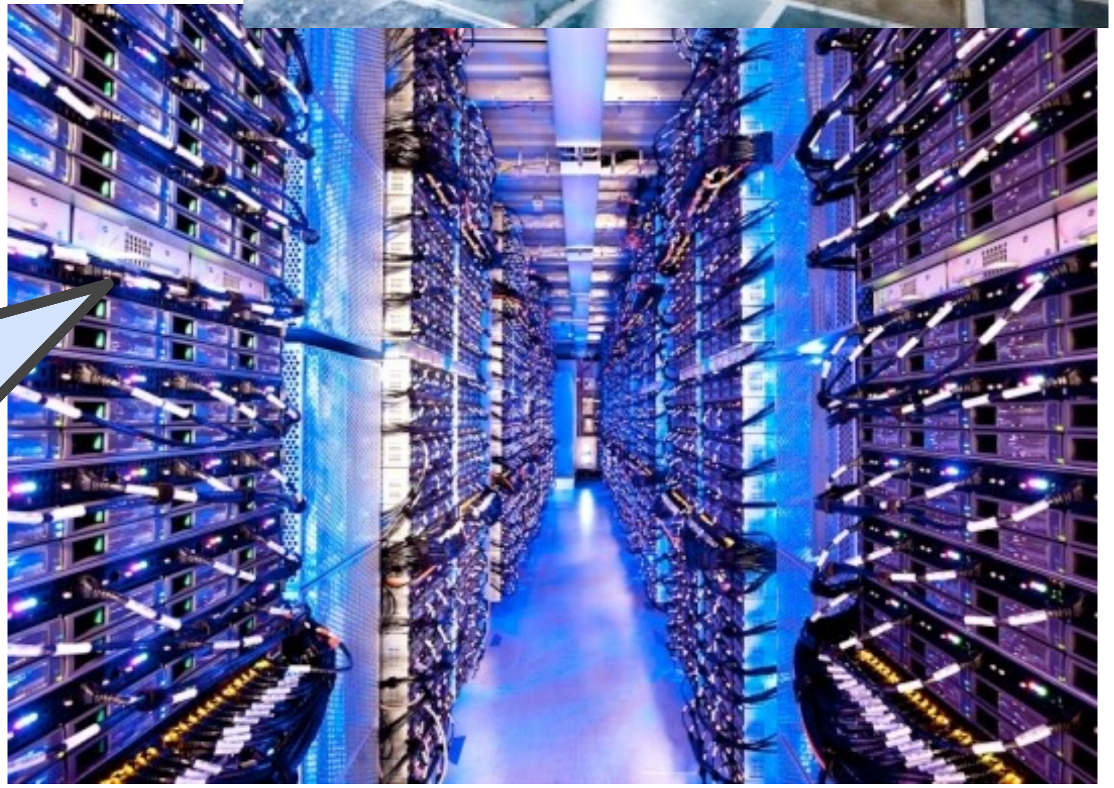
Sometime in April, 2011
Somewhere in farm country, Massachusetts

Data Centers

- **Infrastructure as a Service** clouds rent server and storage resources on demand

- Data Centers are BIG server farms
  - Clusters of 10,000s of servers
  - Growing to 100s of thousands

- Host many application types
  - Web servers, databases
  - Custom business apps
  - Search clusters, data mining

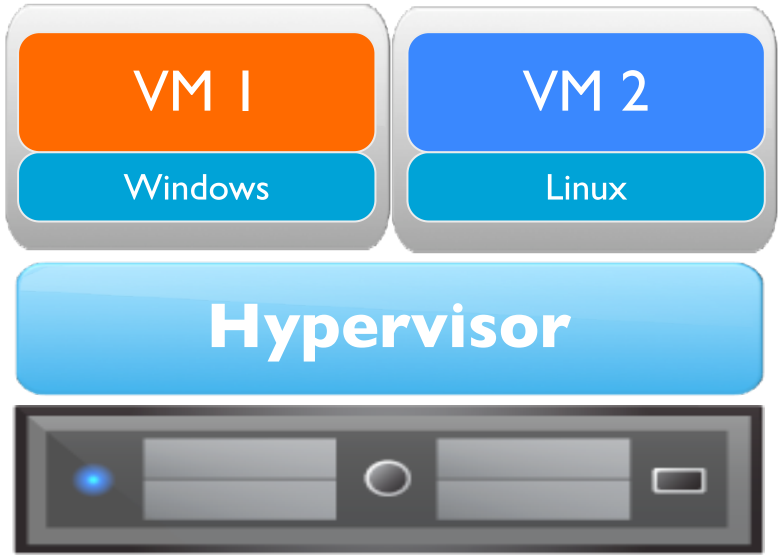


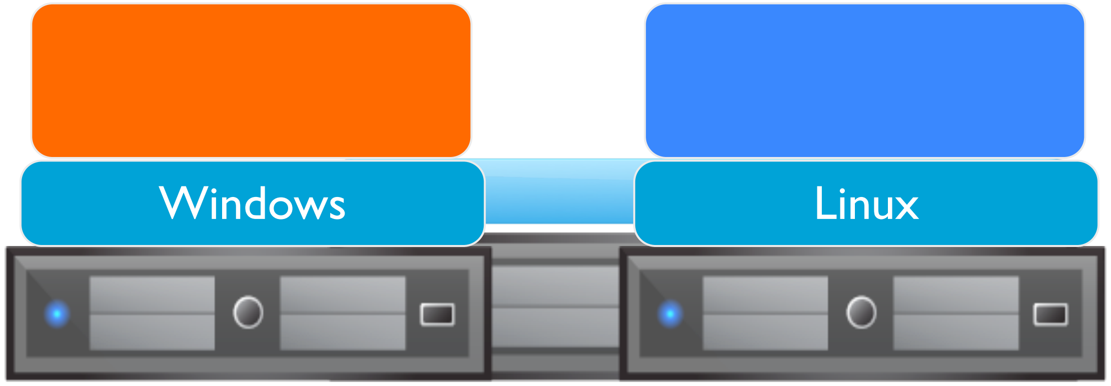
Challenges: large scale and dynamic workload fluctuations

Server Virtualization

- Data centers use virtualization to share physical resources and simplify automation

- Allows a server to be "sliced" into Virtual Machines
- VM has own OS/applications
- Rapidly adjust resource allocations
- VM migration within a LAN



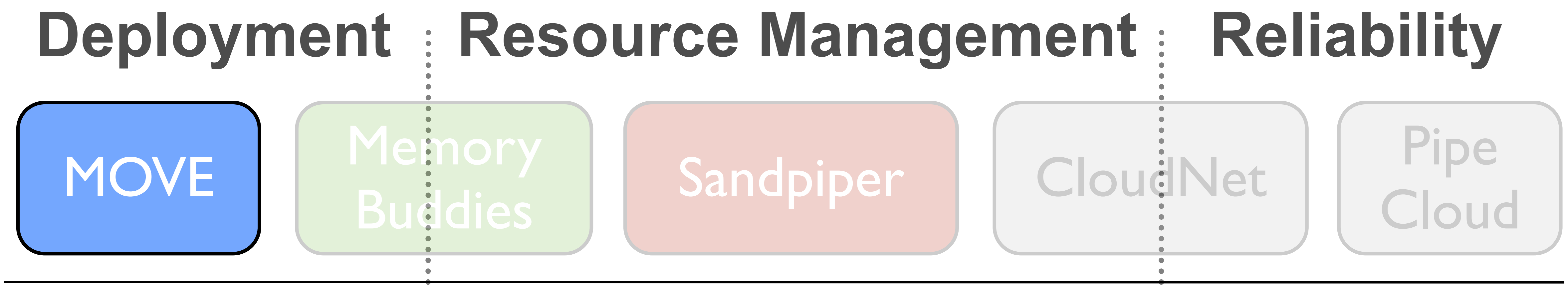


Within a Data Center

- How to transition applications to VMs and account for virtualization overheads?
  - **MOVE**: Modeling Overheads of Virtual Environments

- Where should VMs be placed to allow for the greatest level of server consolidation?
  - **Memory Buddies**: Memory sharing guided placement

- How to dynamically allocate VM resources to prevent server overload?
  - **Sandpiper**: Automated VM migration and resizing



Deployment

Resource Management

Reliability

MOVE

Memory Buddies

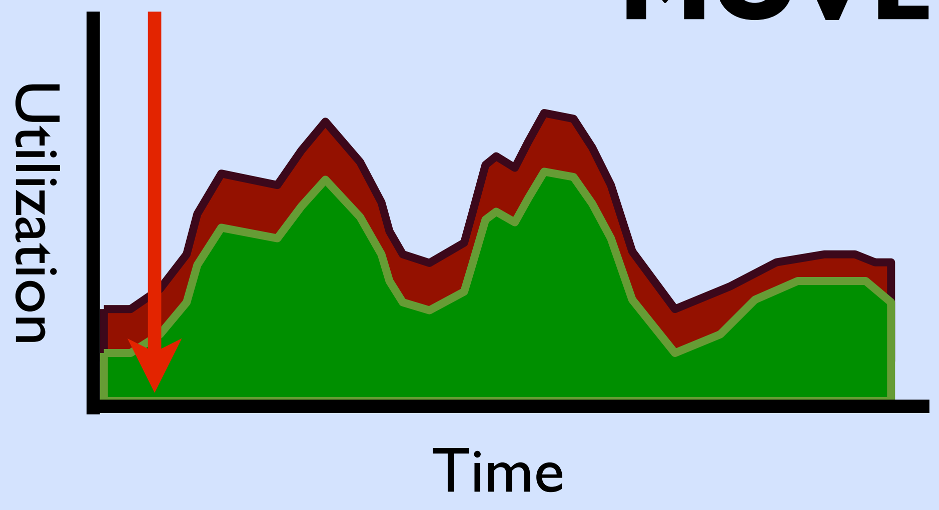
Sandpiper

CloudNet

Pipe Cloud

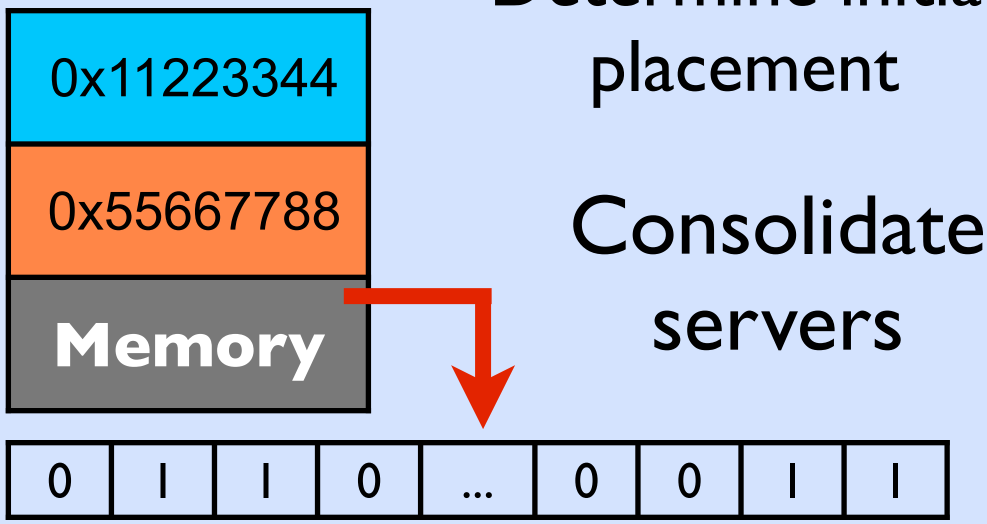
Within a Data Center

MOVE



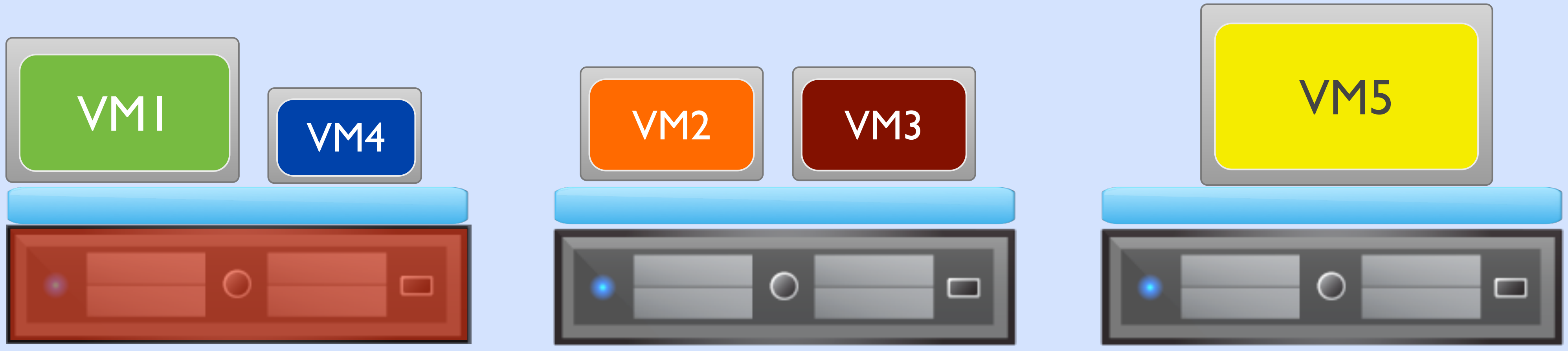
Predict initial resource requirements

Memory Buddies



Determine initial placement

Consolidate servers



Sandpiper

Balance load to prevent hotspots

What is good in these slides?

(or bad!)

(your ideas here)

What is good in these slides?

Good balance of text and visuals

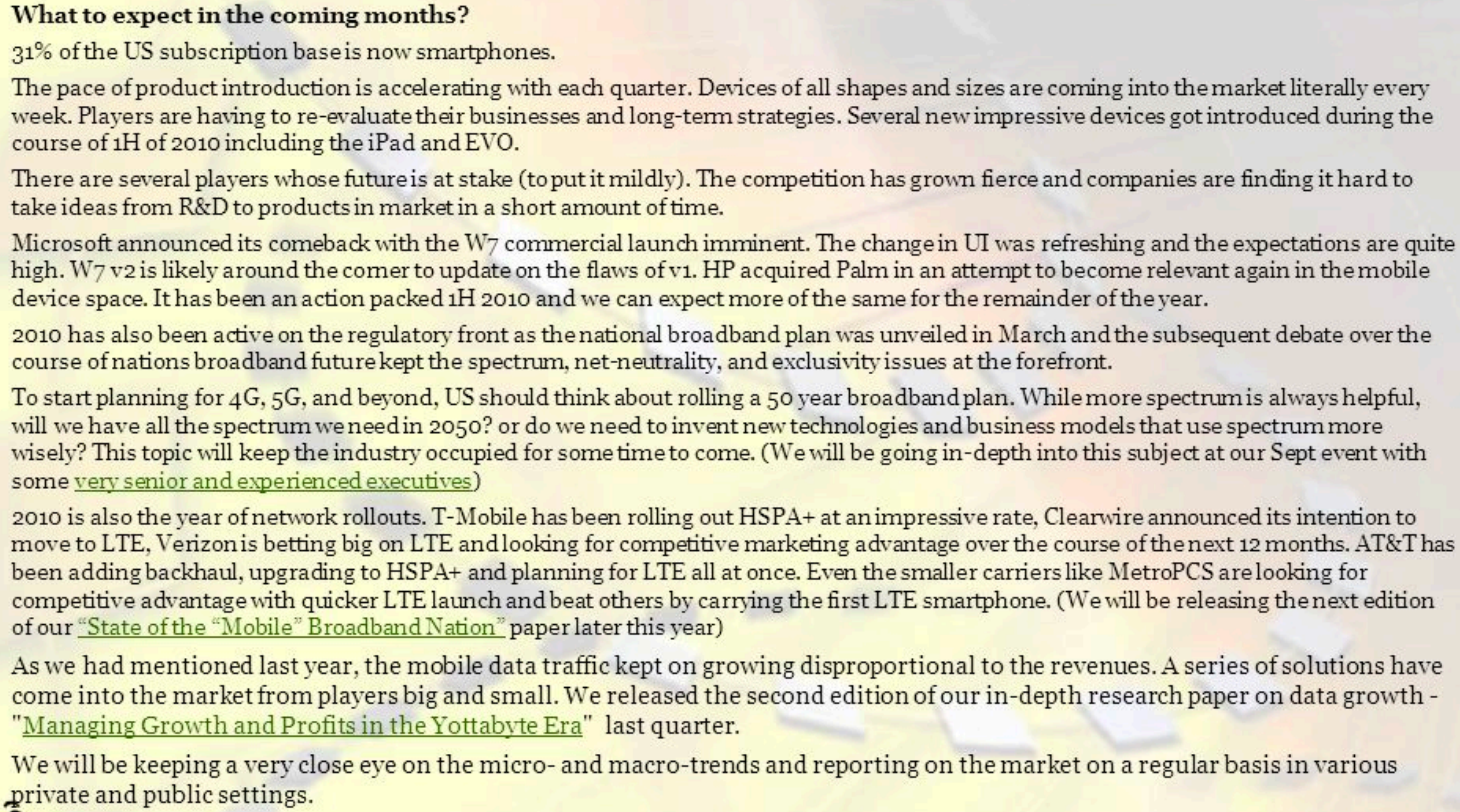
Nice high level overview of thesis

Good connections between components of thesis

Animation is a bit excessive

Slide Tips

US Wireless Market – Q2 2010 Update



What to expect in the coming months?

31% of the US subscription base is now smartphones.

The pace of product introduction is accelerating with each quarter. Devices of all shapes and sizes are coming into the market literally every week. Players are having to re-evaluate their businesses and long-term strategies. Several new impressive devices got introduced during the course of 1H of 2010 including the iPad and EVO.

There are several players whose future is at stake (to put it mildly). The competition has grown fierce and companies are finding it hard to take ideas from R&D to products in market in a short amount of time.

Microsoft announced its comeback with the W7 commercial launch imminent. The change in UI was refreshing and the expectations are quite high. W7 v2 is likely around the corner to update on the flaws of v1. HP acquired Palm in an attempt to become relevant again in the mobile device space. It has been an action packed 1H 2010 and we can expect more of the same for the remainder of the year.

2010 has also been active on the regulatory front as the national broadband plan was unveiled in March and the subsequent debate over the course of nations broadband future kept the spectrum, net-neutrality, and exclusivity issues at the forefront.

To start planning for 4G, 5G, and beyond, US should think about rolling a 50 year broadband plan. While more spectrum is always helpful, will we have all the spectrum we need in 2050? or do we need to invent new technologies and business models that use spectrum more wisely? This topic will keep the industry occupied for some time to come. (We will be going in-depth into this subject at our Sept event with some very senior and experienced executives)

2010 is also the year of network rollouts. T-Mobile has been rolling out HSPA+ at an impressive rate, Clearwire announced its intention to move to LTE, Verizon is betting big on LTE and looking for competitive marketing advantage over the course of the next 12 months. AT&T has been adding backhaul, upgrading to HSPA+ and planning for LTE all at once. Even the smaller carriers like MetroPCS are looking for competitive advantage with quicker LTE launch and beat others by carrying the first LTE smartphone. (We will be releasing the next edition of our "State of the "Mobile" Broadband Nation" paper later this year)

As we had mentioned last year, the mobile data traffic kept on growing disproportional to the revenues. A series of solutions have come into the market from players big and small. We released the second edition of our in-depth research paper on data growth - "Managing Growth and Profits in the Yottabyte Era" last quarter.

We will be keeping a very close eye on the micro- and macro-trends and reporting on the market on a regular basis in various private and public settings.

Limit your text (84 point)

Use large fonts (41 point)

- Not smaller than this (32 point)

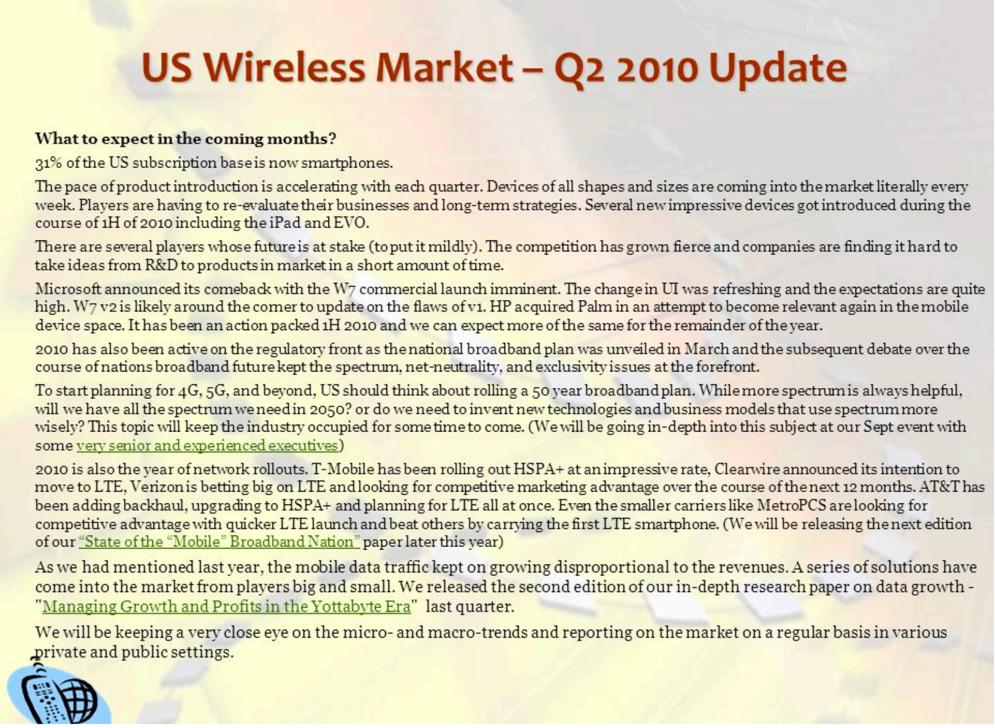
Use bullets, not paragraphs

- **Emphasize** your key points

Don't try to be exhaustive

- Unless the slides will be referred to later without your speech

Don't try to cram in too much content!



US Wireless Market – Q2 2010 Update

What to expect in the coming months?

31% of the US subscription base is now smartphones.

The pace of product introduction is accelerating with each quarter. Devices of all shapes and sizes are coming into the market literally every week. Players are having to re-evaluate their businesses and long-term strategies. Several new impressive devices got introduced during the course of 1H of 2010 including the iPad and EVO.

There are several players whose future is at stake (to put it mildly). The competition has grown fierce and companies are finding it hard to take ideas from R&D to products in market in a short amount of time.

Microsoft announced its comeback with the W7 commercial launch imminent. The change in UI was refreshing and the expectations are quite high. W7 v2 is likely around the corner to update on the flaws of v1. HP acquired Palm in an attempt to become relevant again in the mobile device space. It has been an action packed 1H 2010 and we can expect more of the same for the remainder of the year.

2010 has also been active on the regulatory front as the national broadband plan was unveiled in March and the subsequent debate over the course of nations broadband future kept the spectrum, net-neutrality, and exclusivity issues at the forefront.

To start planning for 4G, 5G, and beyond, US should think about rolling a 50 year broadband plan. While more spectrum is always helpful, will we have all the spectrum we need in 2050? or do we need to invent new technologies and business models that use spectrum more wisely? This topic will keep the industry occupied for some time to come. (We will be going in-depth into this subject at our Sept event with some very senior and experienced executives)

2010 is also the year of network rollouts. T-Mobile has been rolling out HSPA+ at an impressive rate, Clearwire announced its intention to move to LTE, Verizon is betting big on LTE and looking for competitive marketing advantage over the course of the next 12 months. AT&T has been adding backhaul, upgrading to HSPA+ and planning for LTE all at once. Even the smaller carriers like MetroPCS are looking for competitive advantage with quicker LTE launch and beat others by carrying the first LTE smartphone. (We will be releasing the next edition of our "State of the "Mobile" Broadband Nation" paper later this year)

As we had mentioned last year, the mobile data traffic kept on growing disproportional to the revenues. A series of solutions have come into the market from players big and small. We released the second edition of our in-depth research paper on data growth - "Managing Growth and Profits in the Yottabyte Era" last quarter.

We will be keeping a very close eye on the micro- and macro-trends and reporting on the market on a regular basis in various private and public settings.

Limit your text (ugly version)

Use large fonts

- Not smaller than this (32 point)

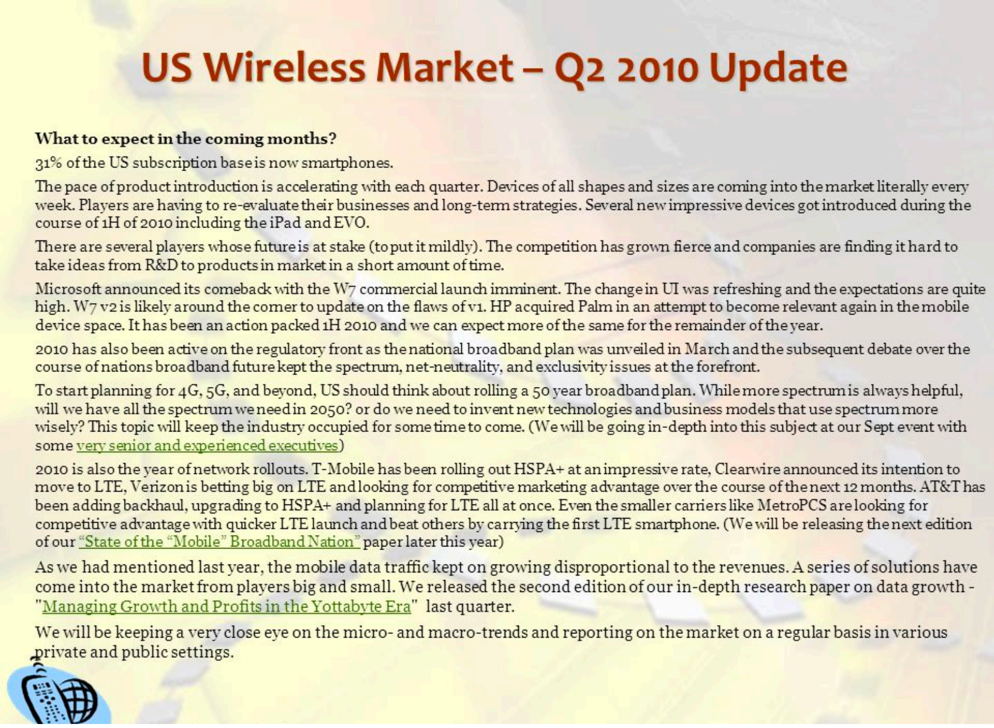
Use bullets, not paragraphs

- Emphasize your key points

Don’t try to be exhaustive

- Unless the slides will be referred to later without your speech

Don’t try to cram in too much content!



US Wireless Market – Q2 2010 Update

What to expect in the coming months?

31% of the US subscription base is now smartphones.

The pace of product introduction is accelerating with each quarter. Devices of all shapes and sizes are coming into the market literally every week. Players are having to re-evaluate their businesses and long-term strategies. Several new impressive devices got introduced during the course of 1H of 2010 including the iPad and EVO.

There are several players whose future is at stake (to put it mildly). The competition has grown fierce and companies are finding it hard to take ideas from R&D to products in market in a short amount of time.

Microsoft announced its comeback with the W7 commercial launch imminent. The change in UI was refreshing and the expectations are quite high. W7 v2 is likely around the corner to update on the flaws of v1. HP acquired Palm in an attempt to become relevant again in the mobile device space. It has been an action packed 1H 2010 and we can expect more of the same for the remainder of the year.

2010 has also been active on the regulatory front as the national broadband plan was unveiled in March and the subsequent debate over the course of nations broadband future kept the spectrum, net-neutrality, and exclusivity issues at the forefront.

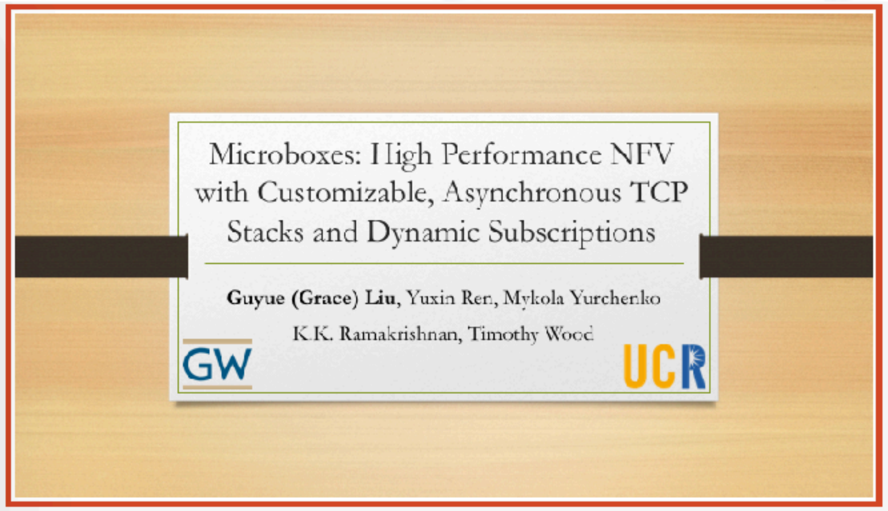
To start planning for 4G, 5G, and beyond, US should think about rolling a 50 year broadband plan. While more spectrum is always helpful, will we have all the spectrum we need in 2050? or do we need to invent new technologies and business models that use spectrum more wisely? This topic will keep the industry occupied for some time to come. (We will be going in-depth into this subject at our Sept event with some very senior and experienced executives)

2010 is also the year of network rollouts. T-Mobile has been rolling out HSPA+ at an impressive rate, Clearwire announced its intention to move to LTE, Verizon is betting big on LTE and looking for competitive marketing advantage over the course of the next 12 months. AT&T has been adding backhaul, upgrading to HSPA+ and planning for LTE all at once. Even the smaller carriers like MetroPCS are looking for competitive advantage with quicker LTE launch and beat others by carrying the first LTE smartphone. (We will be releasing the next edition of our “State of the “Mobile” Broadband Nation” paper later this year)

As we had mentioned last year, the mobile data traffic kept on growing disproportional to the revenues. A series of solutions have come into the market from players big and small. We released the second edition of our in-depth research paper on data growth - "Managing Growth and Profits in the Yottabyte Era" last quarter.

We will be keeping a very close eye on the micro- and macro-trends and reporting on the market on a regular basis in various private and public settings.

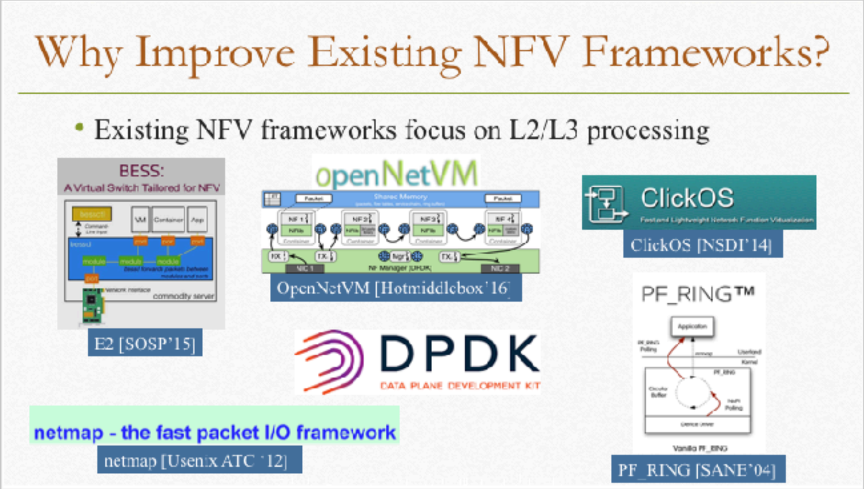
Mix Text and Images



Microboxes: High Performance NFV with Customizable, Asynchronous TCP Stacks and Dynamic Subscriptions

Guyue (Grace) Liu, Yuxin Ren, Mykola Yurchenko
K.K. Ramakrishnan, Timothy Wood

1



Why Improve Existing NFV Frameworks?

- Existing NFV frameworks focus on L2/L3 processing

E2 [SOSP'15]

OpenNetVM [Hotmiddlebox'16]

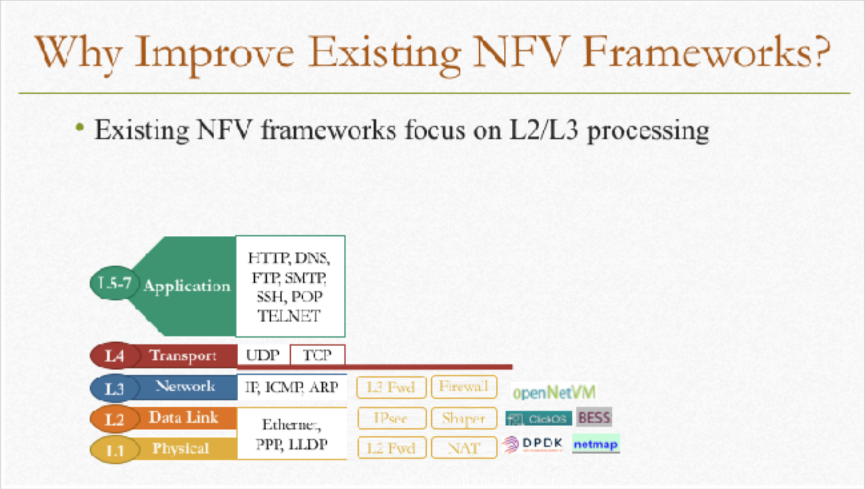
ClickOS [NSDI'14]

netmap [Usenix ATC '12]

PF_RING [SANE'04]

BESS:
A Virtual Switch Tailored for NFV
openNetVM
ClickOS
DPDK
DATA PLANE DEVELOPMENT KIT
netmap - the fast packet I/O framework
PF_RING™

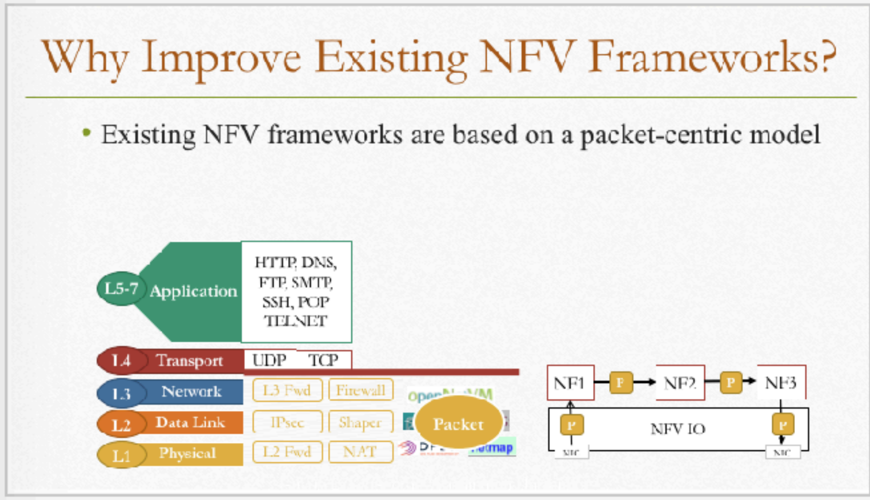
2



Why Improve Existing NFV Frameworks?

- Existing NFV frameworks focus on L2/L3 processing

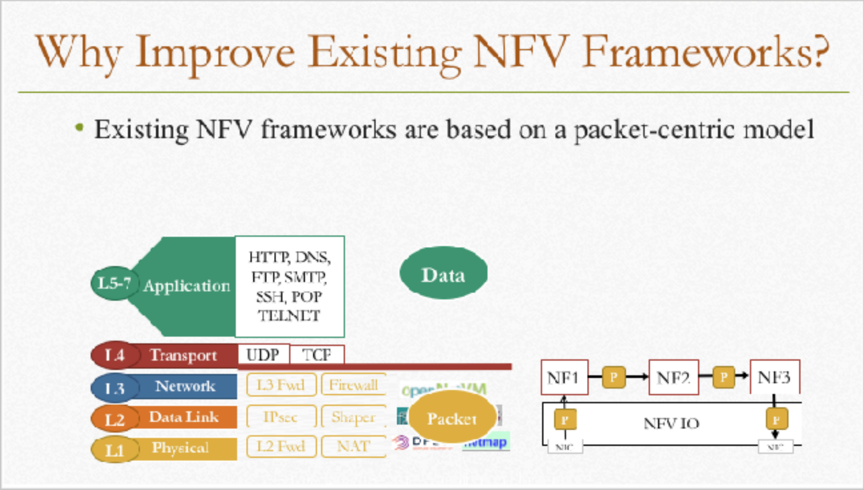
3



Why Improve Existing NFV Frameworks?

- Existing NFV frameworks are based on a packet-centric model

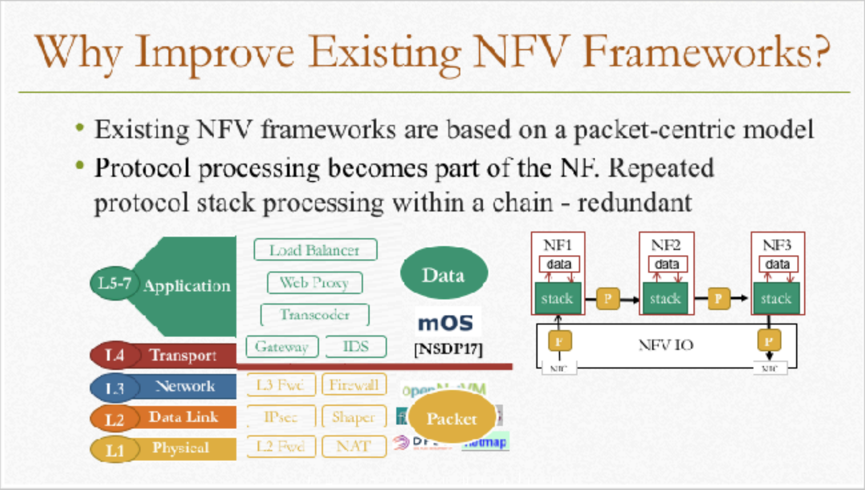
4



Why Improve Existing NFV Frameworks?

- Existing NFV frameworks are based on a packet-centric model

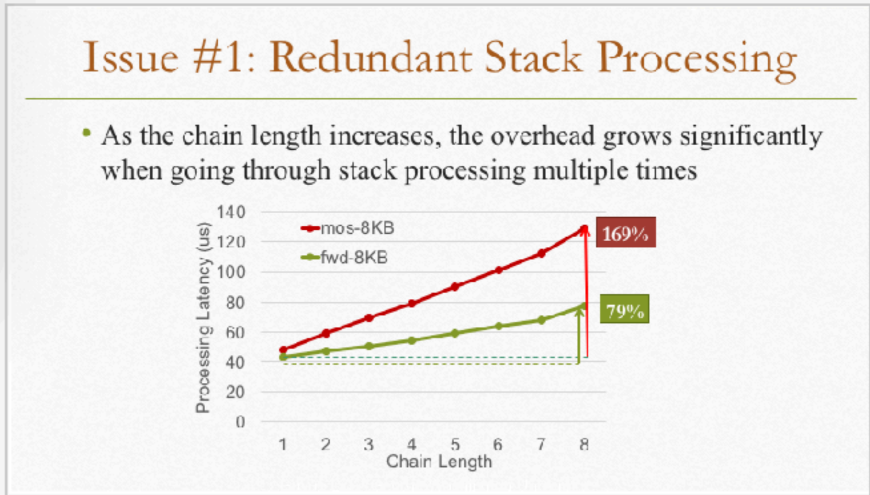
5



Why Improve Existing NFV Frameworks?

- Existing NFV frameworks are based on a packet-centric model
- Protocol processing becomes part of the NF. Repeated protocol stack processing within a chain - redundant

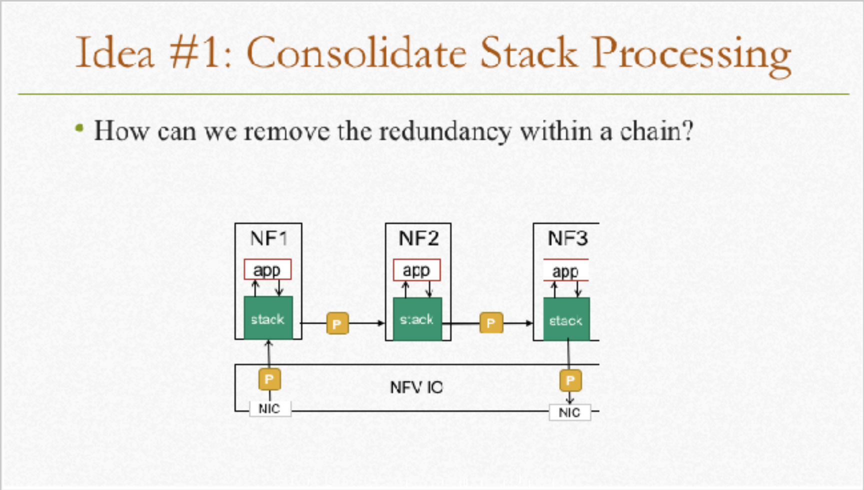
6



Issue #1: Redundant Stack Processing

- As the chain length increases, the overhead grows significantly when going through stack processing multiple times

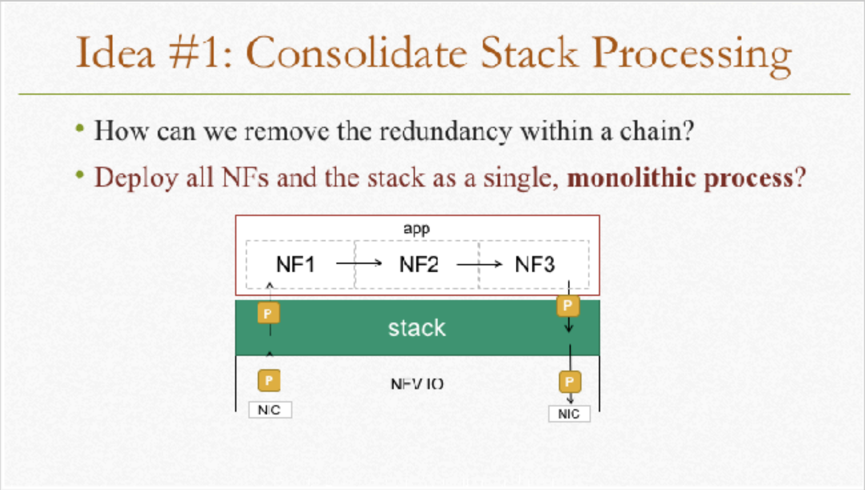
7



Idea #1: Consolidate Stack Processing

- How can we remove the redundancy within a chain?

8



Idea #1: Consolidate Stack Processing

- How can we remove the redundancy within a chain?
- Deploy all NFs and the stack as a single, **monolithic process**?

9

USE A HIERARCHY

- This is text
  - This is also text
    - This is even more text
  - Why are they all the same size?

Use a hierarchy

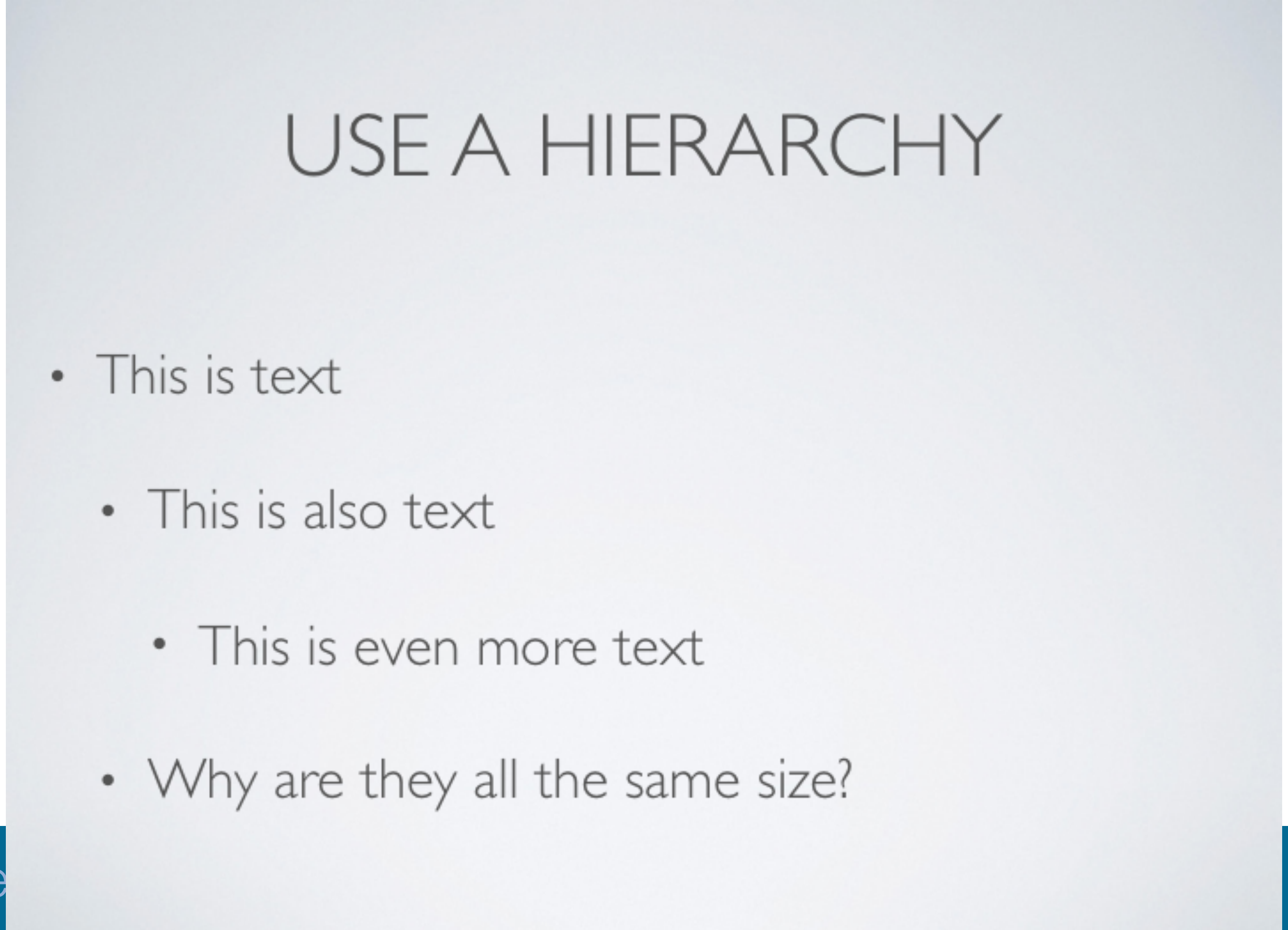
This is text

- This is also text
    - This is even more text
- Note that they are not the same size!

White space is important, but don't go overboard

- And sub bullets with a smaller font size help viewers focus on key points

Make your own template and keep improving it!



USE A HIERARCHY

- This is text
    - This is also text
        - This is even more text
    - Why are they all the same size?

GW PPT Template

This is where I put my content

Here is more content

Wow, this is just awful.

Why is the bar so big at the bottom?

I have so little useful space and it is poorly laid out.

THE GEORGE
WASHINGTON
UNIVERSITY
WASHINGTON, DC

My Template

Text that is reasonably large

- Sub bullets that are smaller
  - Sub-sub bullets that are even smaller, although I rarely use them
  - (Mainly so I can add spacing more flexibly)

Large “before paragraph” spacing so bullets aren’t too tight and smaller line spacing so you can fit denser text when needed (try to avoid multi-line)

A useful footer with your name and affiliation

- Always include the slide number in corner!

Minimal background images

Optional: school / lab logos

Pop up boxes to emphasize key points!

Should we animate?

Text that is reasonably large

- Sub bullets that are smaller
  - Sub-sub bullets that are even smaller, although I rarely use them
  - (Mainly so I can add spacing more flexibly)

Large “before paragraph” spacing so bullets aren’t too tight and smaller line spacing so you can fit denser text when needed (try to avoid multi-line)

A useful footer with your name and affiliation

- Always include the slide number in corner!

Minimal background images

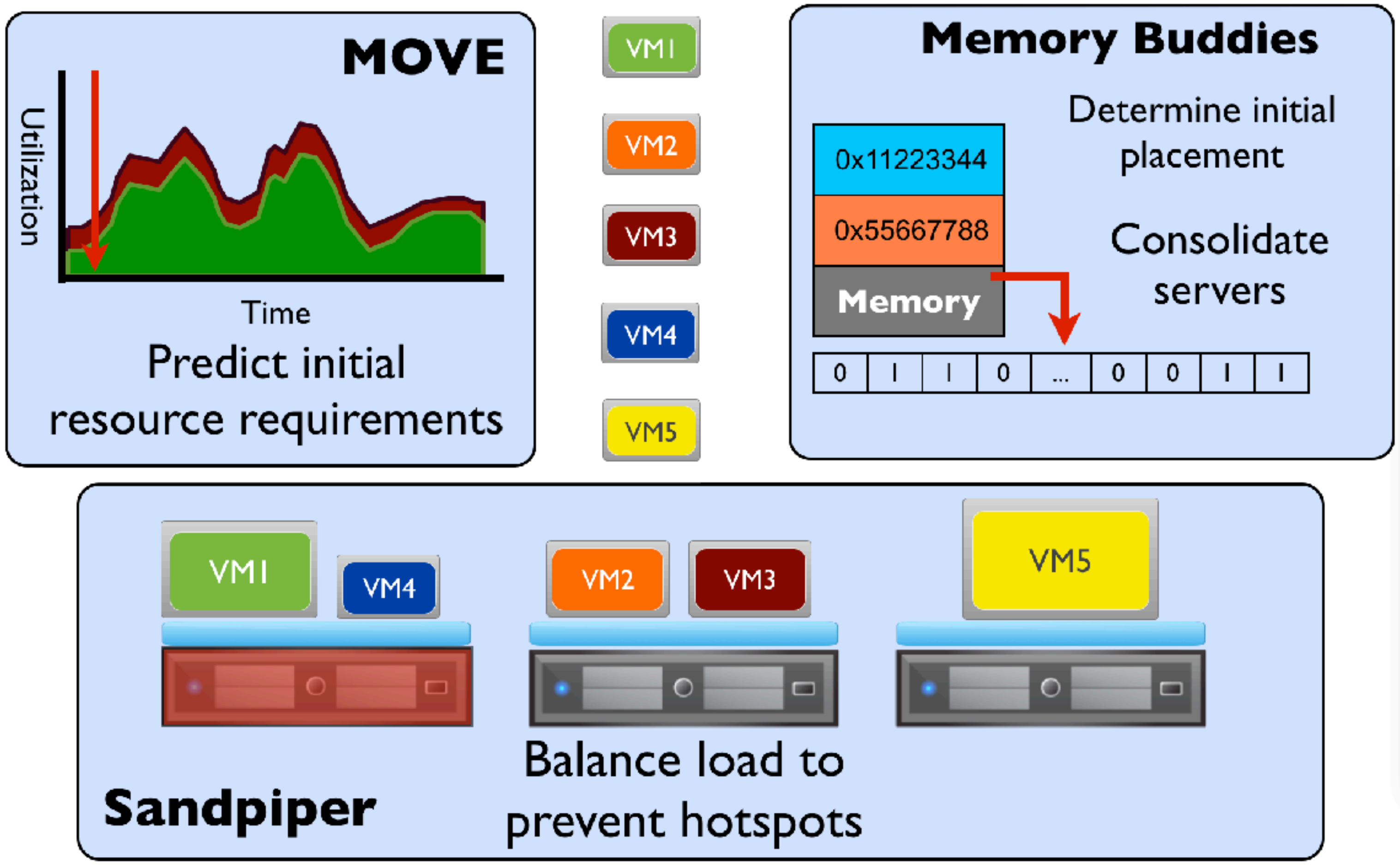
Optional: school / lab logos

It depends! Usually only if ~3 bullets on slide

Animations

Useful, but “expensive” to create

Can be distracting if overused



Suggestion: only use animation for emphasizing most important points

- And you only have at most 3 of those, right?

Color Inspiration

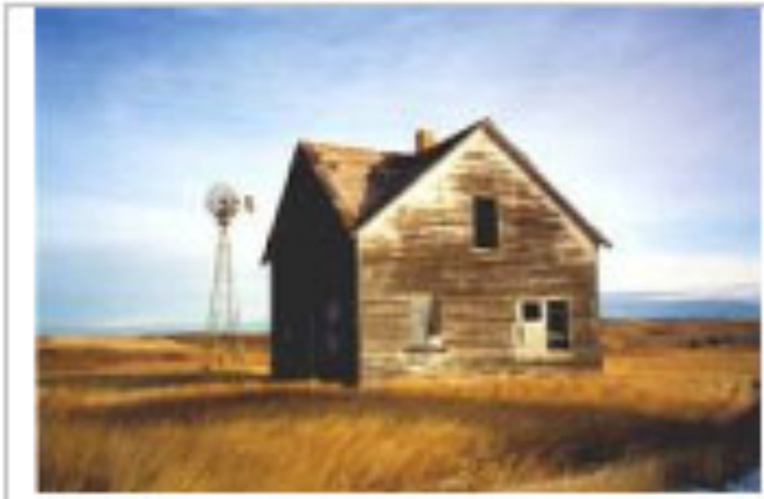
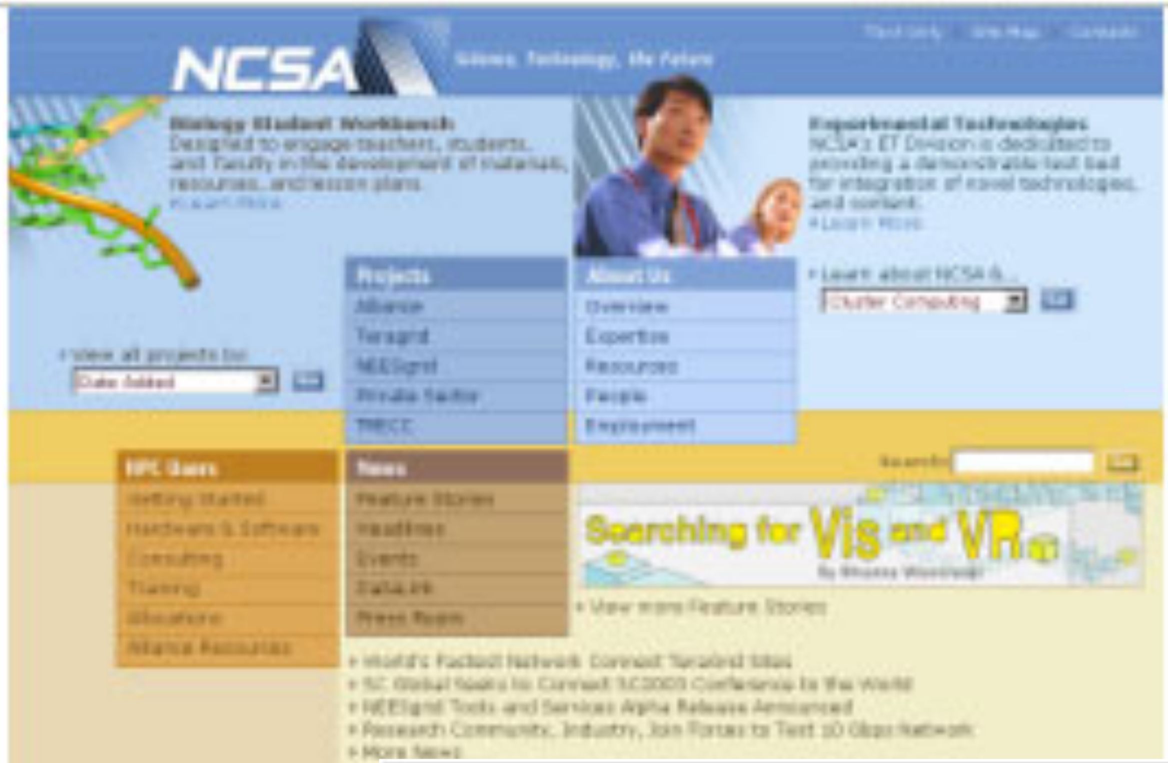


Fig 2



Text Only | Site Map | Contacts

Science, Technology, the Future

Biology Student Workbench
Designed to engage teachers, students, and faculty in the development of materials, resources, and lesson plans.
» Learn More

Experimental Technologies
NCSA's ET Division is dedicated to providing a demonstrable test bed for integration of novel technologies and content.
» Learn More

» View all projects by: Date Added

Projects
- Alliance
- Teragrid
- NEESgrid
- Private Sector
- TRECC

About Us
- Overview
- Expertise
- Resources
- People
- Employment

» Learn about NCSA & ... Cluster Computing

HPC Users
- Getting Started
- Hardware & Software
- Consulting
- Training
- Allocations
- Alliance Resources

News
- Feature Stories
- Headlines
- Events
- Datalink
- Press Room

Search for:

Searching for Vis and VR

» View more Feature Stories

- » World's Fastest Network Connect Teragrid Sites
- » SC Global Seeks to Connect SC2003 Conference to the World
- » NEESgrid Tools and Services Alpha Release Announced
- » Research Community, Industry, Join Forces to Test 10 Gbps Network
- » More News

Fig 3

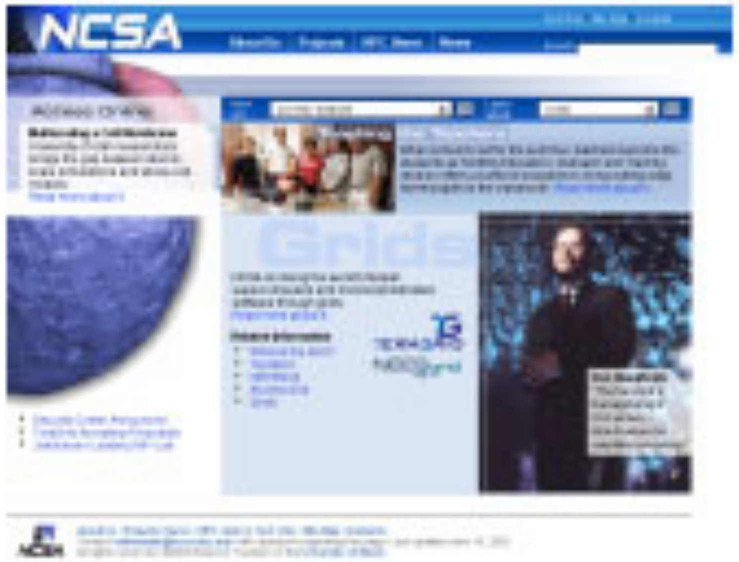


Fig 4

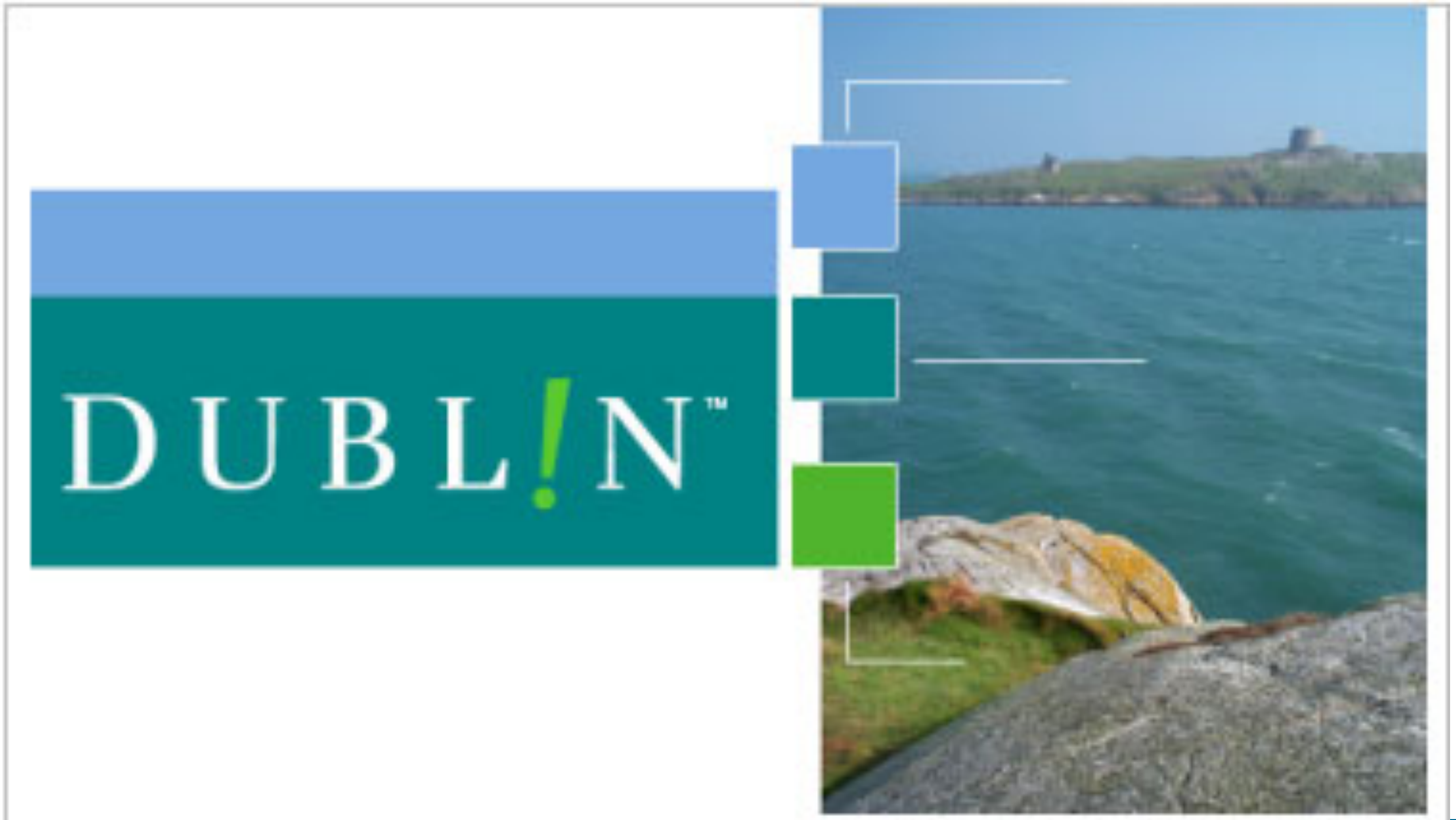
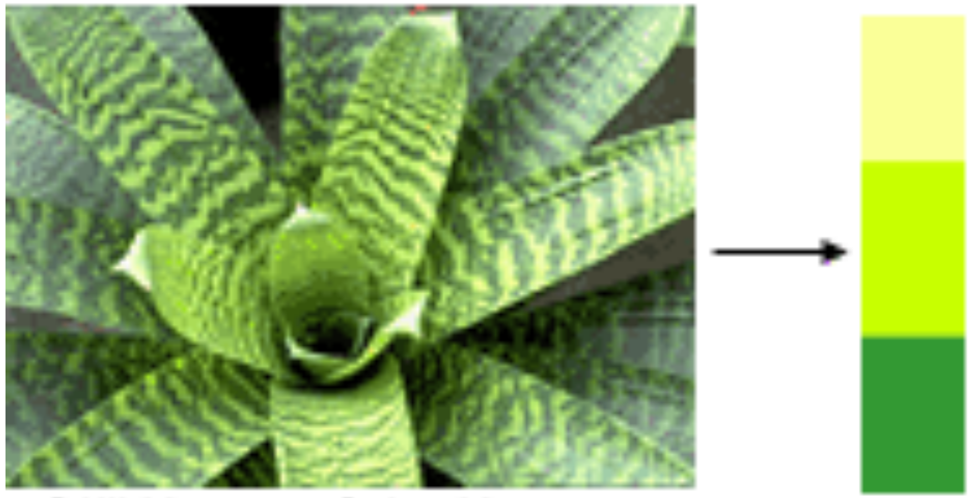
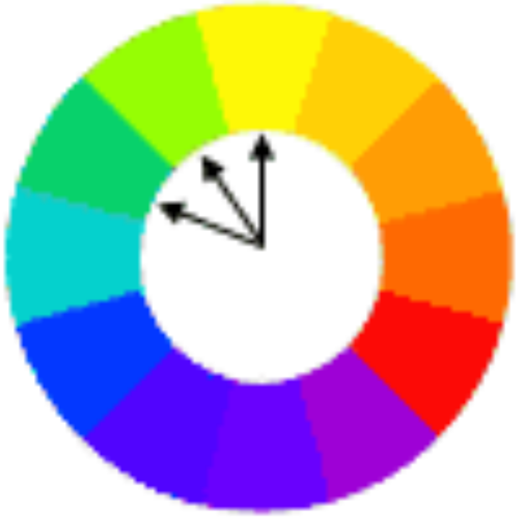


Fig 7

From Boxes and Arrows

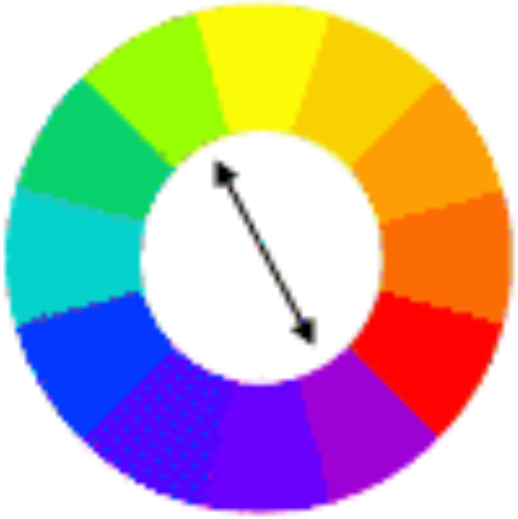
Color

Related colors

©Jill Morton - Color Matters

Complementary colors



©Jill Morton - Color Matters

Limit the number of colors

Max per display: 4

Max across entire app: 7

Font/Background Color

White background with a black font is easier to read

Black background with white font can look childish

Other colors may not have enough contrast or could look strange depending on the projector

Font/Background Color

White background with a black font is easier to read

Black background with white font can look childish

Other colors may not have enough contrast or could look strange depending on the projector

Font/Background Color

White background with a black font is easier to read

Black background with white font can look childish

Other colors may not have enough contrast or could look strange depending on the projector

Fonts

Know the difference between:

Serif fonts: easier to read in print

- Times New Roman

Sans-Serif fonts: more modern on screen

- Arial, Helvetica

Monospaced fonts: only for code

- `Courier`

Never use Comic Sans!

Ted Talk style?

Should we mimic TED talk slide style?

PREPARING TO FAIL

Photo: Blair Harkness

PEOPLE

CODE

INFRASTRUCTURE

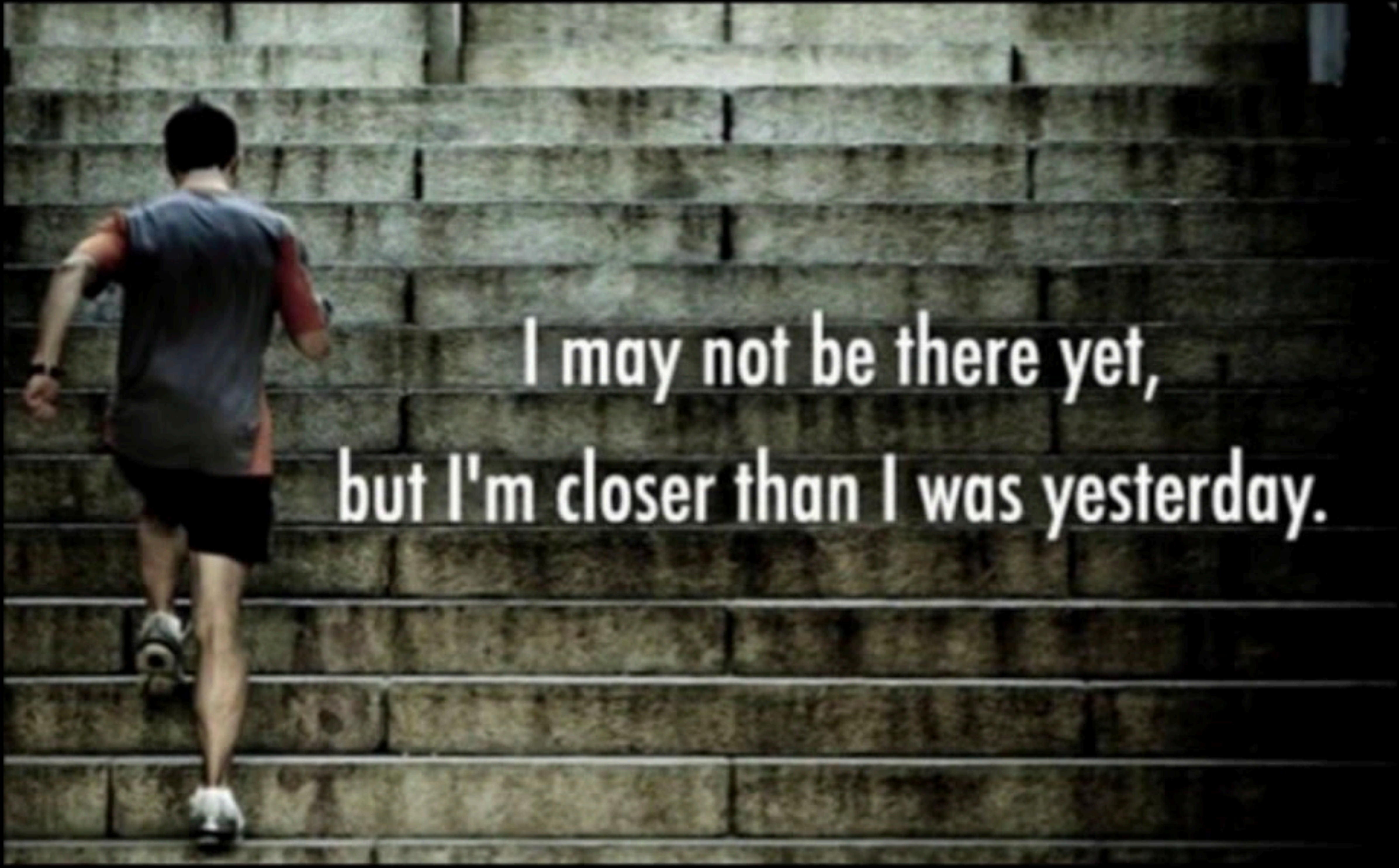
INFRASTRUCTURE

“Success is stumbling from failure to failure with no loss of enthusiasm.”

WINSTON CHURCHILL

REDUNDANCY

MULTIPLE HARD DRIVES


I may not be there yet,
but I'm closer than I was yesterday.

TED Slide Style

Don’t use this for a technical talk

TED is great inspiration for **speaking** style

- but the slide format is mainly relevant for “motivational” talks

Similarly, much of the advice for making great slides online is not relevant!

- They are for a business audience!

How many slides?

I typically aim for ~1 minute per slide

Varies depending on the depth of information per slide and whether you use “real” animation or multi-slide animation

Know your audience!

Slide format will be very different…

Class

- lots

Talk a

- Pre

Pitchi

- Foc

Talk a

- The

P != NP

Each company / org will have its own “culture"

P != NP

Blah blah

I proved it using the X conjecture

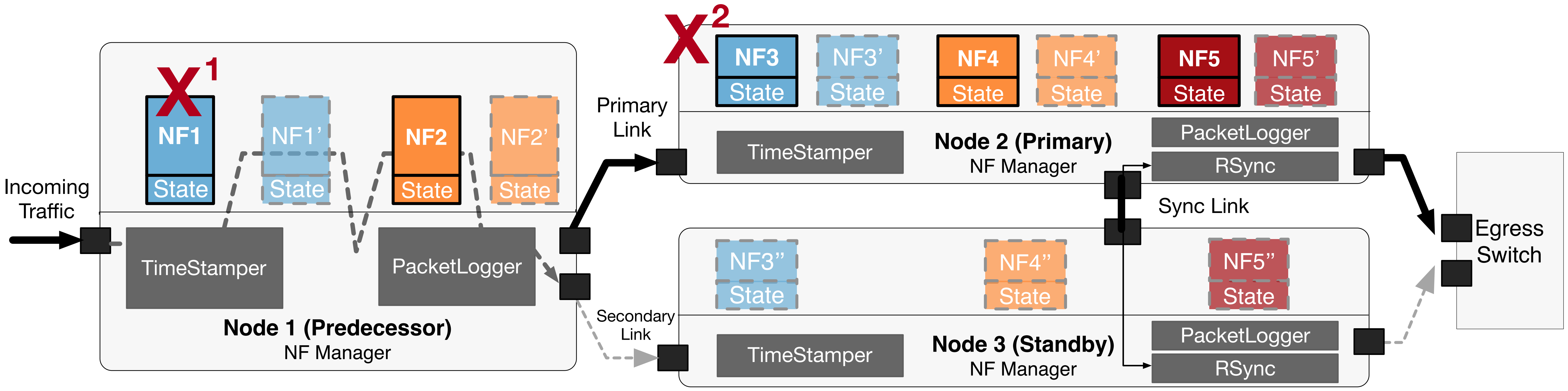
Blah blah

Diagrams and Graphs Tips

Diagrams

Visual representations of your algorithm, system, or approach are always helpful

- Make the paper easier to understand
- Break up large chunks of text

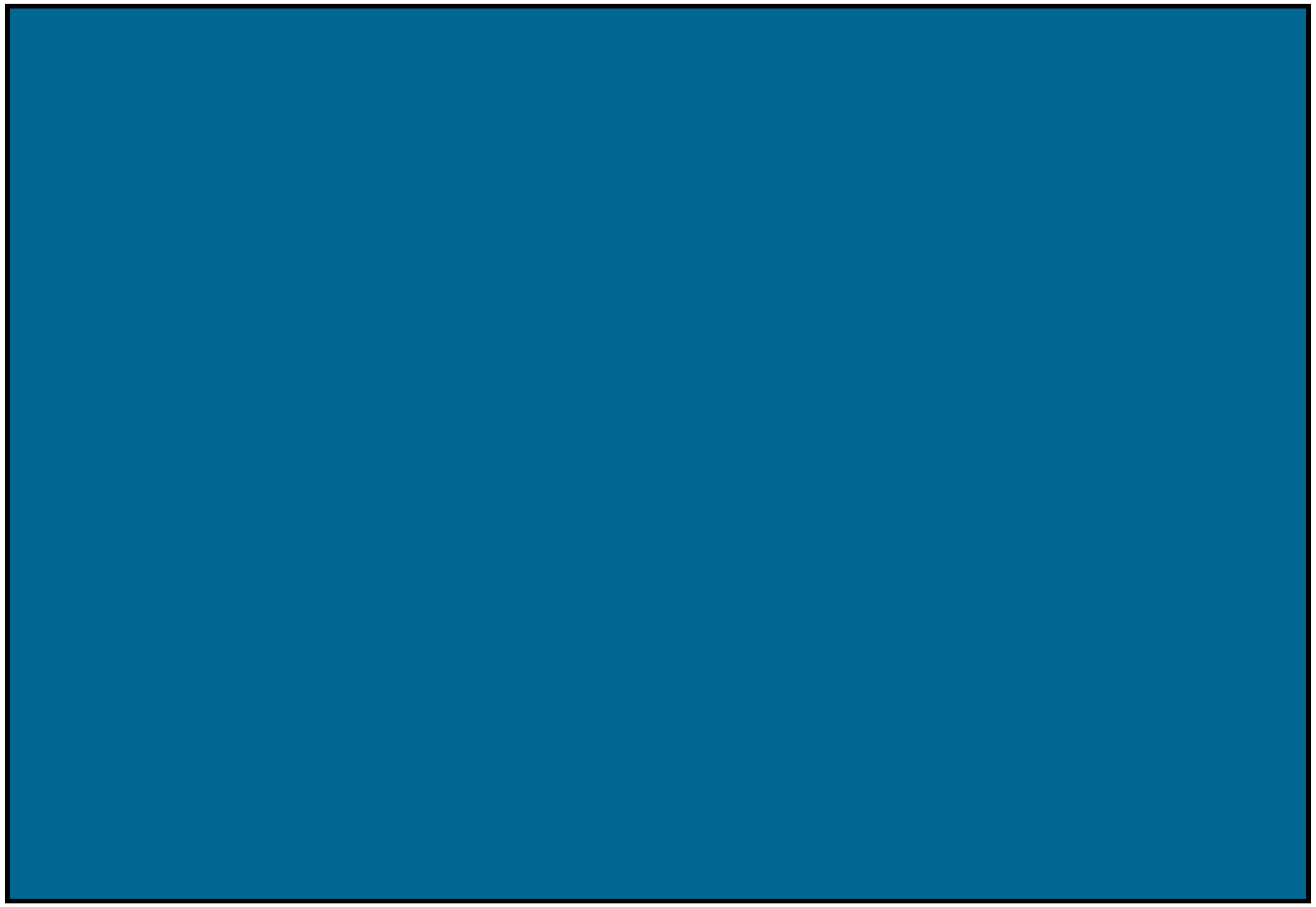


Find a tool that works for you

- My lab used Omnigraffle (mac only), switching to **diagrams.net**
- Use something consistently so you become more efficient

Bad Diagrams

Is this a good system diagram?



Common Problems

Useless content

Bad color choices

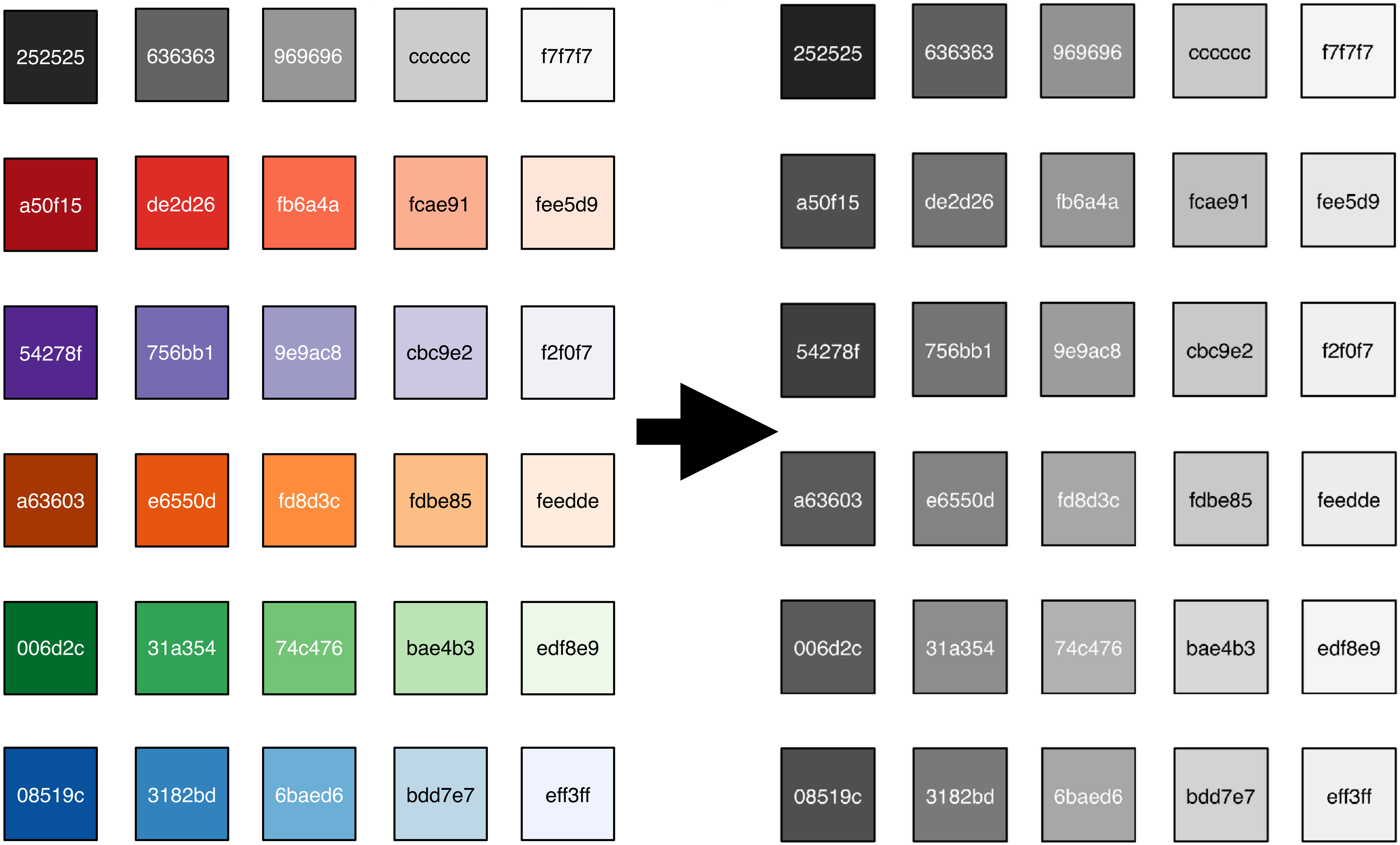
- Indistinguishable, childish, etc

Fonts need to be bigger!

No caption to explain

- I prefer useful captions instead of “title” captions

Color to BW

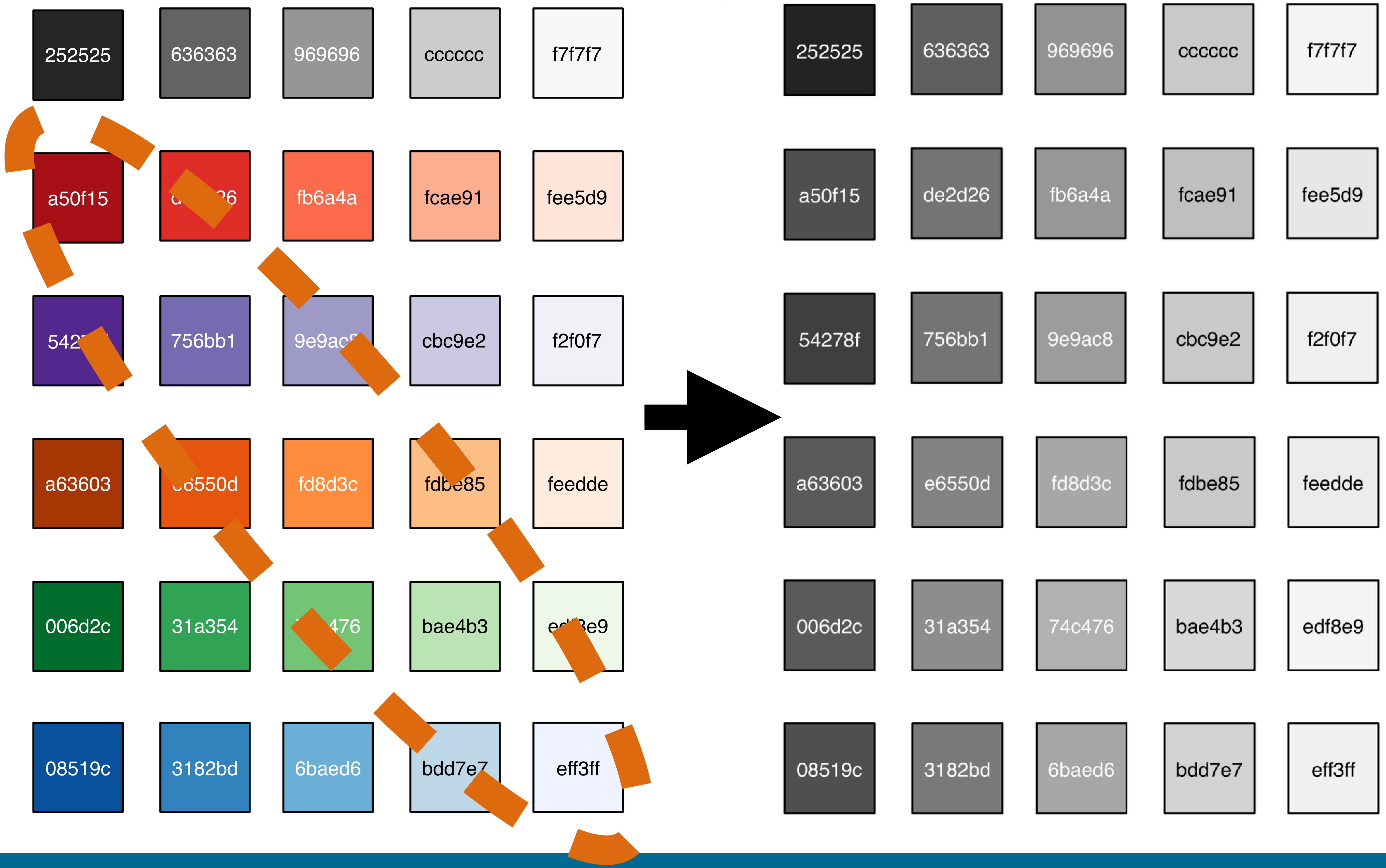


252525	636363	969696	cccccc	f7f7f7
a50f15	de2d26	fb6a4a	fcae91	fee5d9
54278f	756bb1	9e9ac8	cbc9e2	f2f0f7
a63603	e6550d	fd8d3c	fdbe85	feedde
006d2c	31a354	74c476	bae4b3	edf8e9
08519c	3182bd	6baed6	bdd7e7	eff3ff

→

252525	636363	969696	cccccc	f7f7f7
a50f15	de2d26	fb6a4a	fcae91	fee5d9
54278f	756bb1	9e9ac8	cbc9e2	f2f0f7
a63603	e6550d	fd8d3c	fdbe85	feedde
006d2c	31a354	74c476	bae4b3	edf8e9
08519c	3182bd	6baed6	bdd7e7	eff3ff

Color to BW



252525	636363	969696	cccccc	f7f7f7
a50f15	de2d26	fb6a4a	fcae91	fee5d9
54278f	756bb1	9e9ac8	cbc9e2	f2f0f7
a63603	e6550d	fd8d3c	fdbe85	feedde
006d2c	31a354	74c476	bae4b3	edf8e9
08519c	3182bd	6baed6	bdd7e7	eff3ff

252525	636363	969696	cccccc	f7f7f7
a50f15	de2d26	fb6a4a	fcae91	fee5d9
54278f	756bb1	9e9ac8	cbc9e2	f2f0f7
a63603	e6550d	fd8d3c	fdbe85	feedde
006d2c	31a354	74c476	bae4b3	edf8e9
08519c	3182bd	6baed6	bdd7e7	eff3ff

Relationships

Most good diagrams focus on the relationships between different components/concepts

A good diagram should help a reader use the "black box" technique to filter out aspects of the work that are less important to them

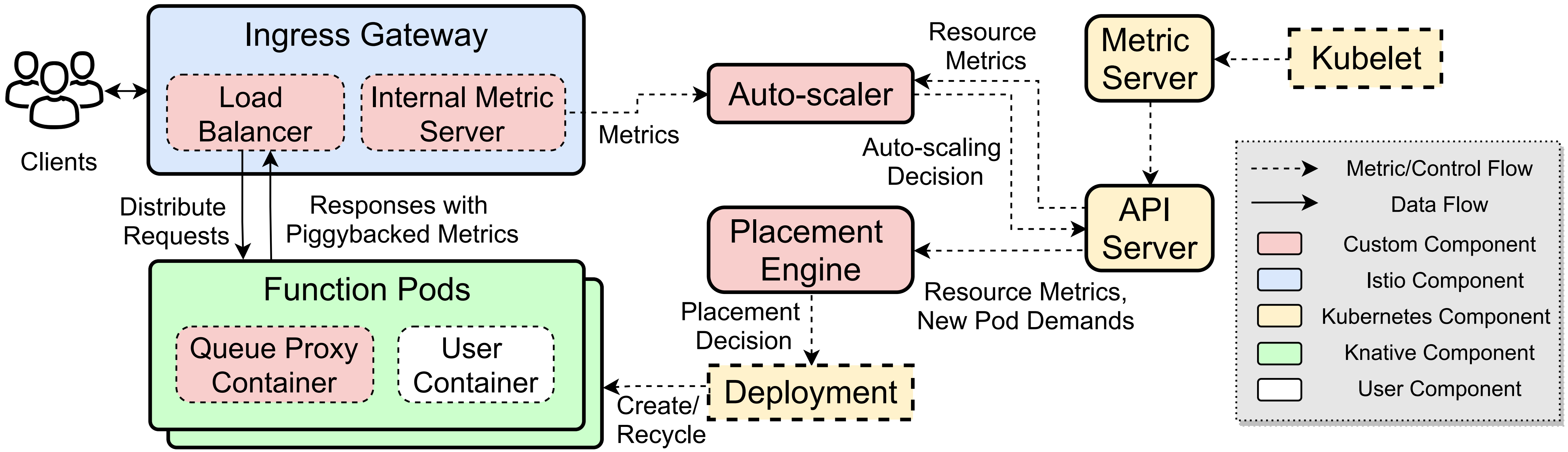
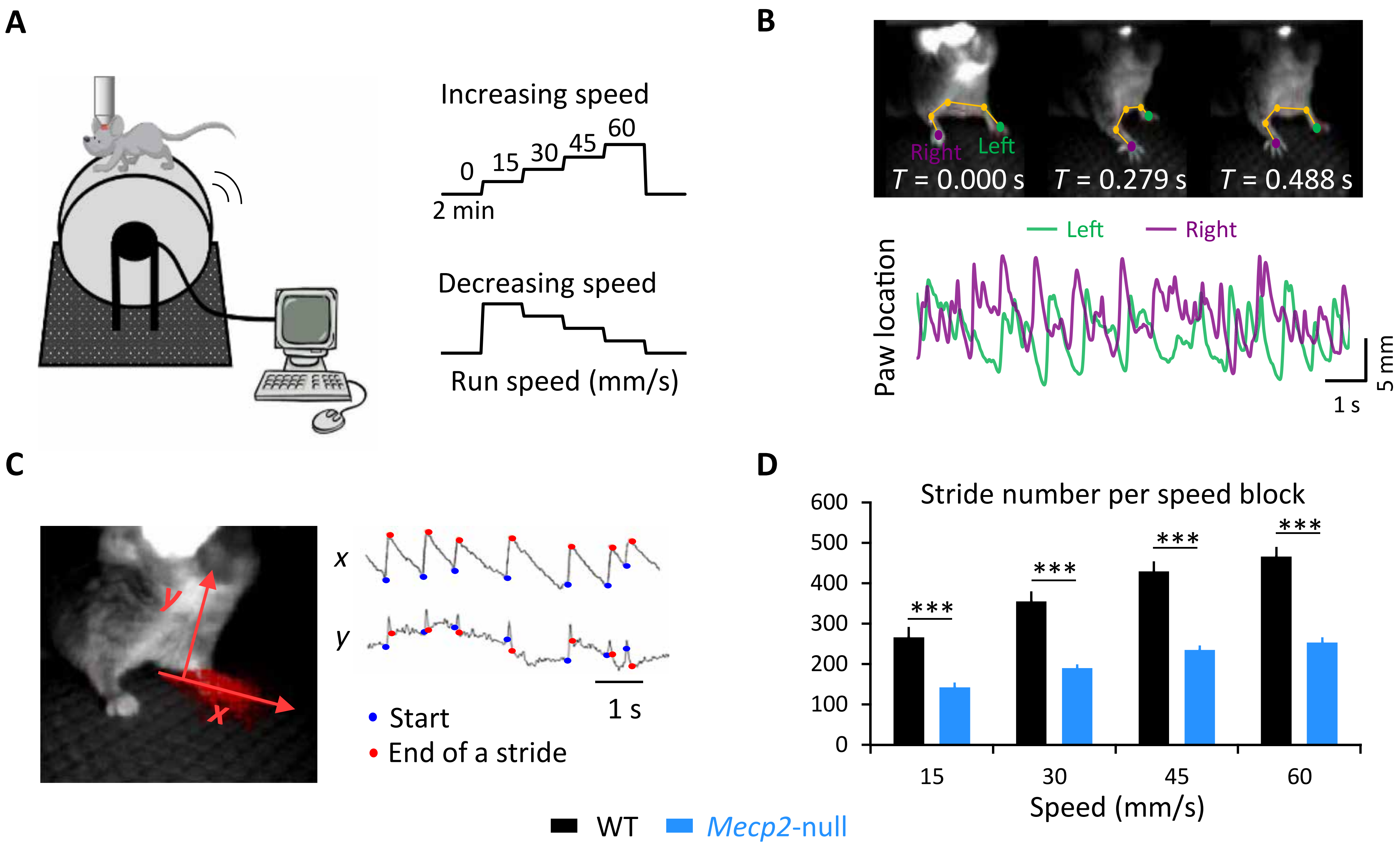


Figure 1: Mu Overview.

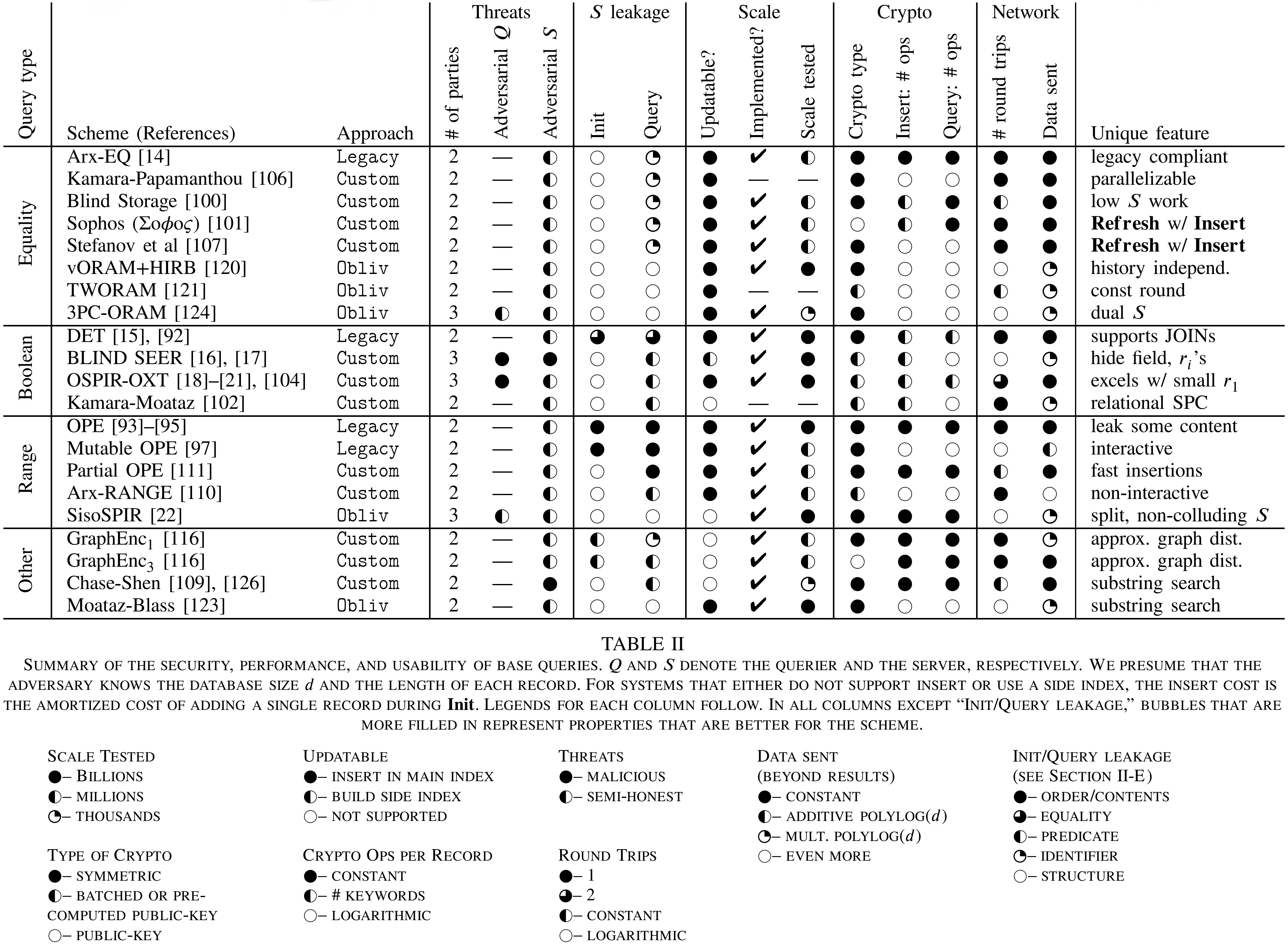
Diagrams vs Results

Diagrams can explain experimental setup



From "Motor training improves coordination and anxiety in symptomatic *Mecp2*-null mice despite impaired functional connectivity within the motor circuit" by Simha and friends

Diagrams of Results



			Threats			S leakage		Scale			Crypto			Network		
Query type	Scheme (References)	Approach	# of parties	Adversarial Q	Adversarial S	Init	Query	Updatable?	Implemented?	Scale tested	Crypto type	Insert: # ops	Query: # ops	# round trips	Data sent	Unique feature
Equality	Arx-EQ [14]	Legacy	2	—	◐	○	◔	●	✔	◐	●	●	●	●	●	legacy compliant
	Kamara-Papamanthou [106]	Custom	2	—	◐	○	◔	●	—	—	●	○	○	●	●	parallelizable
	Blind Storage [100]	Custom	2	—	◐	○	◔	●	✔	◐	●	◐	●	◐	●	low S work
	Sophos (Σοφος) [101]	Custom	2	—	◐	○	◔	●	✔	◐	○	◐	●	●	●	**Refresh** w/ **Insert**
	Stefanov et al [107]	Custom	2	—	◐	○	◔	●	✔	◐	●	○	○	●	●	**Refresh** w/ **Insert**
	vORAM+HIRB [120]	Obliv	2	—	◐	○	○	●	✔	●	●	○	○	○	◔	history independ.
	TWORAM [121]	Obliv	2	—	◐	○	○	●	—	—	◐	○	○	◐	◔	const round
	3PC-ORAM [124]	Obliv	3	◐	◐	○	○	●	✔	◔	●	○	○	○	◔	dual S
Boolean	DET [15], [92]	Legacy	2	—	◐	◕	◕	●	✔	●	●	◐	◐	●	●	supports JOINs
	BLIND SEER [16], [17]	Custom	3	●	●	○	◐	◐	✔	●	◐	◐	○	○	◔	hide field, r_i's
	OSPIR-OXT [18]–[21], [104]	Custom	3	●	◐	○	◐	●	✔	●	◐	◐	◐	◕	●	excels w/ small r_1
	Kamara-Moataz [102]	Custom	2	—	◐	○	◐	○	—	—	◐	◐	○	●	◔	relational SPC
Range	OPE [93]–[95]	Legacy	2	—	◐	●	●	●	✔	●	●	●	●	●	●	leak some content
	Mutable OPE [97]	Legacy	2	—	◐	●	●	●	✔	◐	●	○	○	○	◐	interactive
	Partial OPE [111]	Custom	2	—	◐	○	●	●	✔	◐	●	●	●	◐	●	fast insertions
	Arx-RANGE [110]	Custom	2	—	◐	○	◐	●	✔	◐	◐	○	○	●	○	non-interactive
	SisoSPIR [22]	Obliv	3	◐	◐	○	○	○	✔	●	●	●	●	○	◔	split, non-colluding S
Other	GraphEnc$_1$ [116]	Custom	2	—	◐	◐	◔	○	✔	◐	●	●	●	●	◔	approx. graph dist.
	GraphEnc$_3$ [116]	Custom	2	—	◐	◐	◐	○	✔	◐	○	●	●	●	●	approx. graph dist.
	Chase-Shen [109], [126]	Custom	2	—	●	○	◐	○	✔	◔	●	●	●	◐	●	substring search
	Moataz-Blass [123]	Obliv	2	—	◐	○	○	●	✔	●	●	○	○	○	◔	substring search

TABLE II

Summary of the security, performance, and usability of base queries. Q and S denote the querier and the server, respectively. We presume that the adversary knows the database size d and the length of each record. For systems that either do not support insert or use a side index, the insert cost is the amortized cost of adding a single record during **Init**. Legends for each column follow. In all columns except "Init/Query leakage," bubbles that are more filled in represent properties that are better for the scheme.

Scale Tested
●– Billions
◐– Millions
◔– Thousands

Type of Crypto
●– Symmetric
◐– Batched or pre-computed public-key
○– Public-key

Updatable
●– Insert in main index
◐– Build side index
○– Not supported

Crypto Ops per Record
●– Constant
◐– # keywords
○– Logarithmic

Threats
●– Malicious
◐– Semi-honest

Round Trips
●– 1
◕– 2
◐– Constant
○– Logarithmic

Data Sent
(beyond results)
●– Constant
◐– Additive polylog(d)
◔– Mult. polylog(d)
○– Even more

Init/Query Leakage
(see Section II-E)
●– Order/contents
◕– Equality
◐– Predicate
◔– Identifier
○– Structure

From "Sok: Cryptographically protected database search" by Yerukhimovich and friends

Condensing Information

A diagram is a way of extracting out the key concepts of a complex system or phenomenon

If you don't know how to draw it, you probably don't fully understand it yourself!

Practice sketching out your systems/algorithms/concepts as diagrams to help in conversations with your advisor/collaborators

- Communication is much more efficient with a whiteboard!
- Don't let your advisor do all the drawing!

Let's make a diagram…

Goals:

Clean, consistent shapes

Clear connections between components

Useful colors

Looks good in print

Running Experiments

I ran an experiment where I evaluated the effectiveness of my new operating systems, TimOS

When running BenchmarkX, TimOS had a score of 250,000!

Am I done? What do you need to know to interpret these results?

An experiment is a sample

The result of an experiment is not "truth"

It is a **sample**, drawn from a **population** of possible results for your combination of Task, System, and Environment

- In some cases one experiment gives you many samples, e.g., accuracies of 10,000 images classified, response time of 10M requests, etc

Data Analysis

You need to look closely at your data!

Is it consistent?

- Always repeat experiment where possible

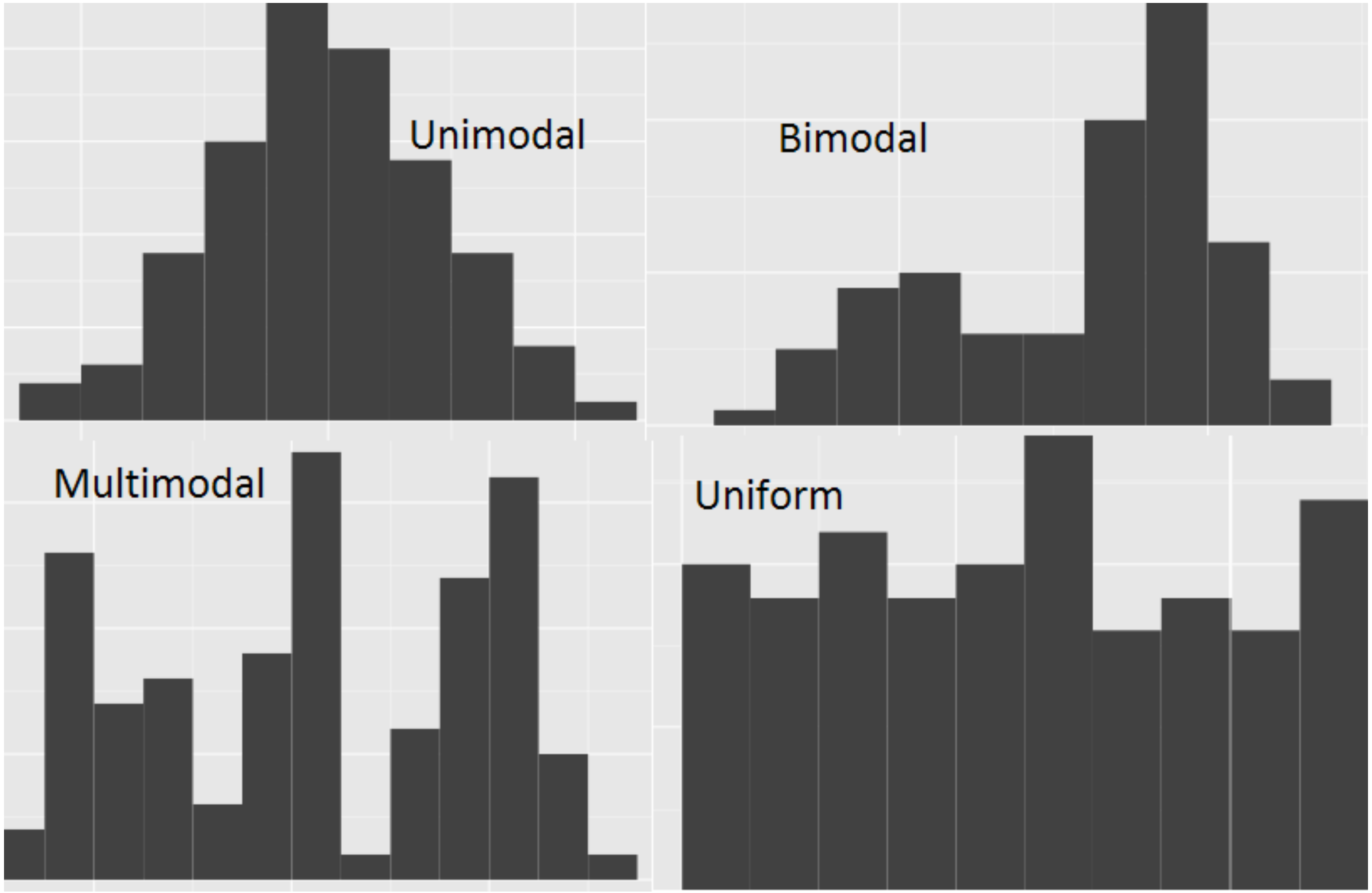
Does it match your expectation?

- You should have had a mental model for the expected result

Detailed Statistics

Average is not enough

Often you need to understand the **distribution**

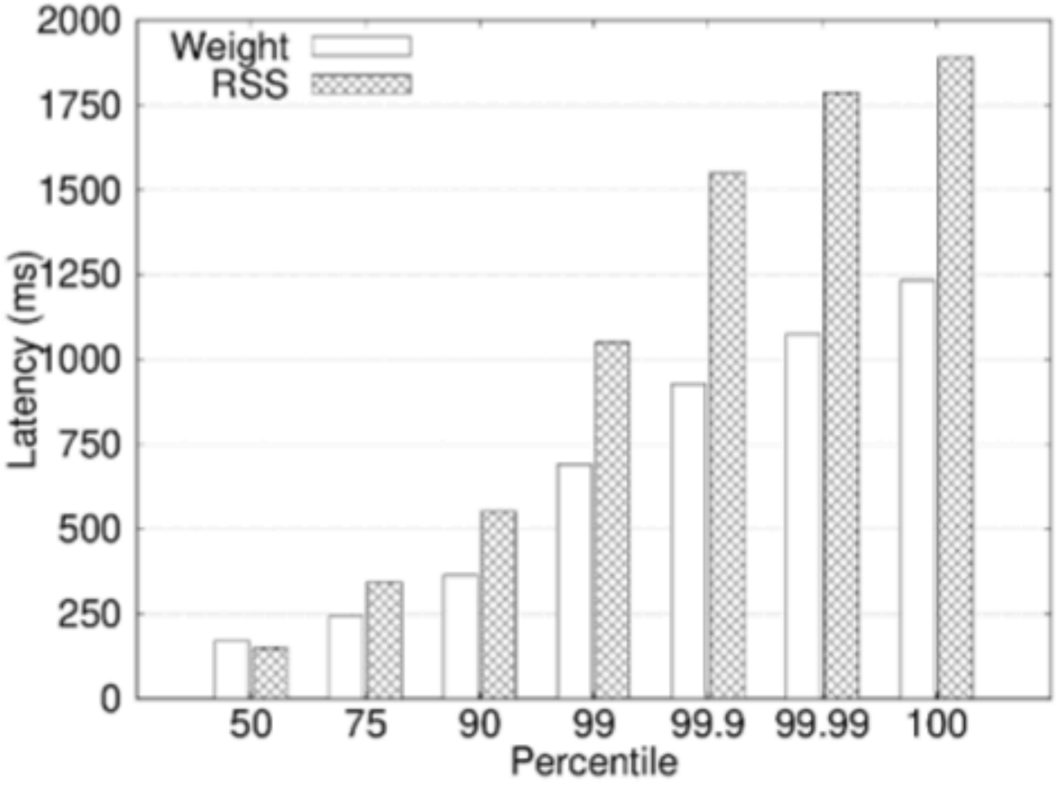


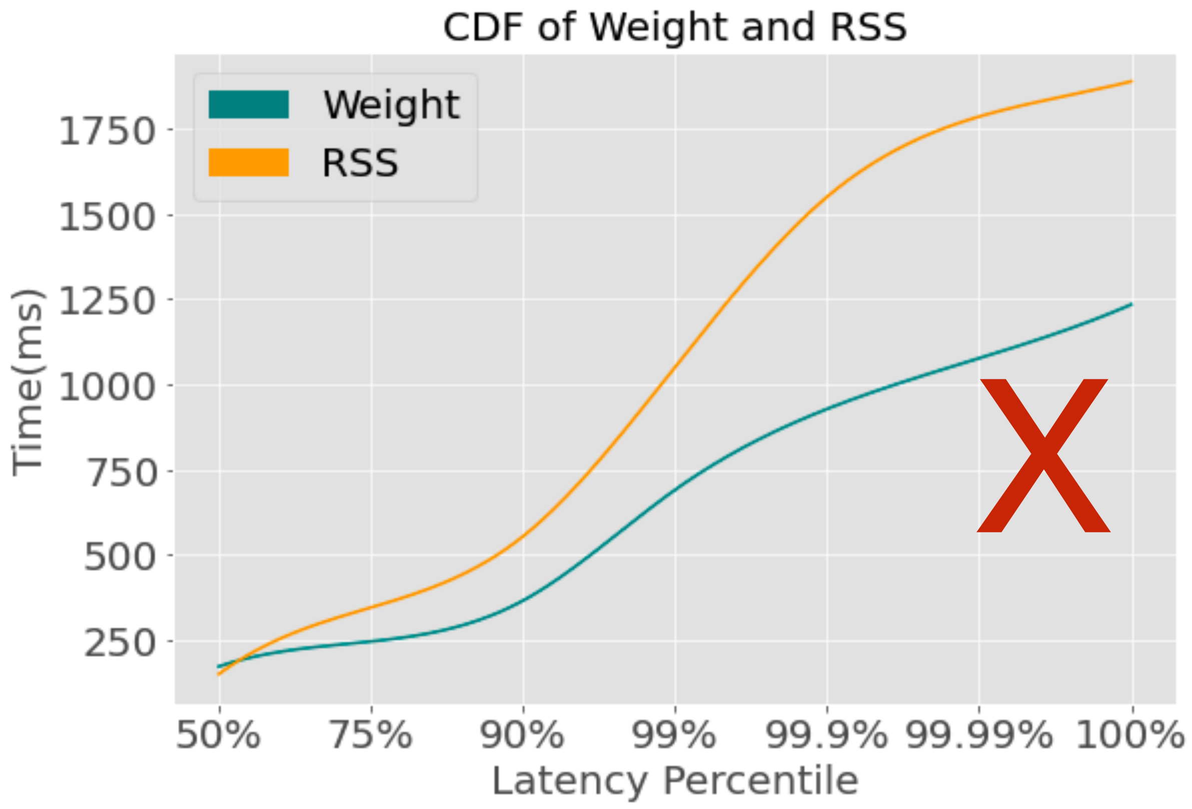
Bar vs Line Graphs

Bar graphs let you break the "axis must be consistently spaced" rule

Or use bars when measurements along x-axis are not ordered (categories of values, experiment #, etc)

Whenever you connect the dots, think about their meaning




CDF of Weight and RSS

More Bars

For % of whole, why not use pie chart?

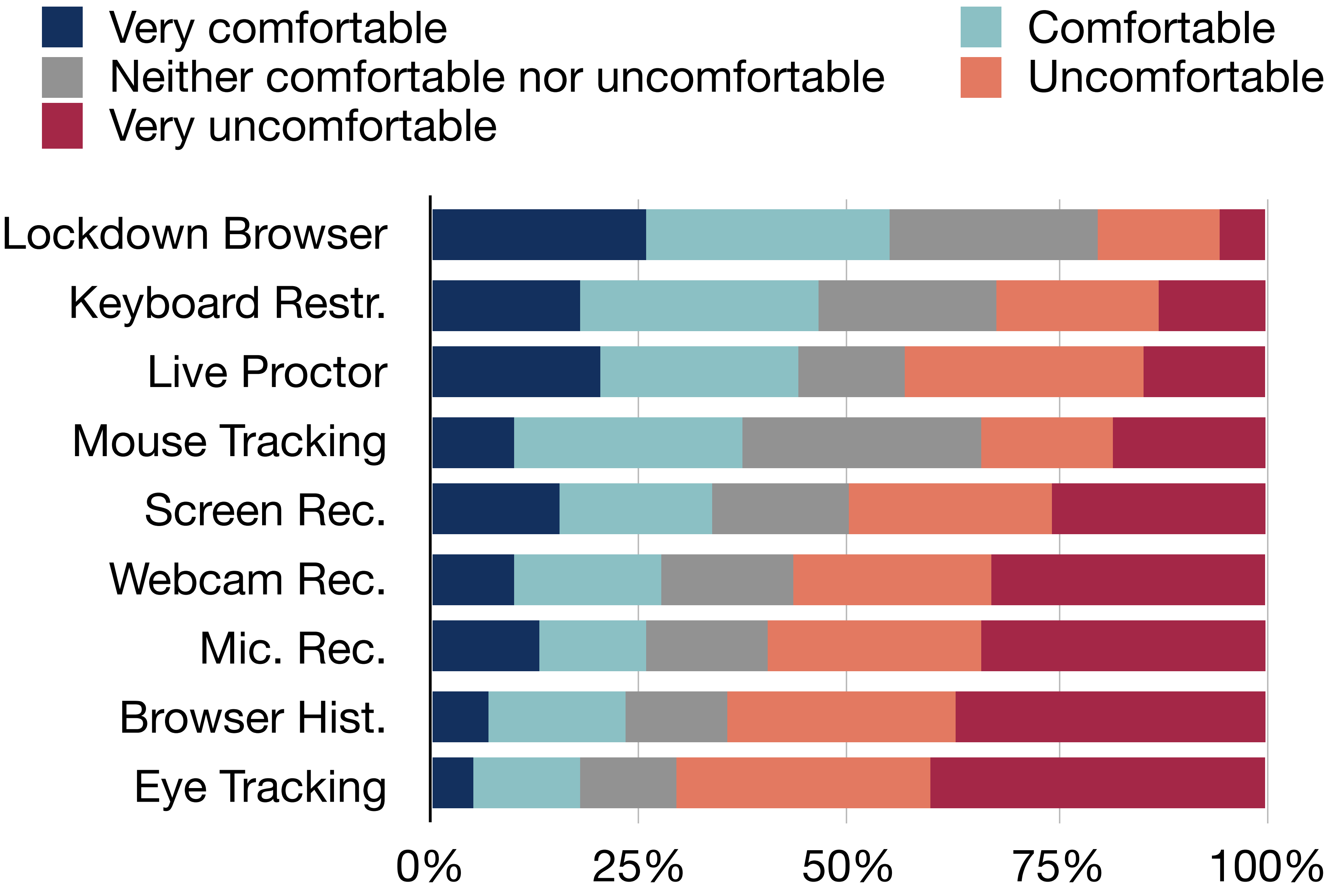


Figure 7: Comfort with monitoring types (**Q28**).

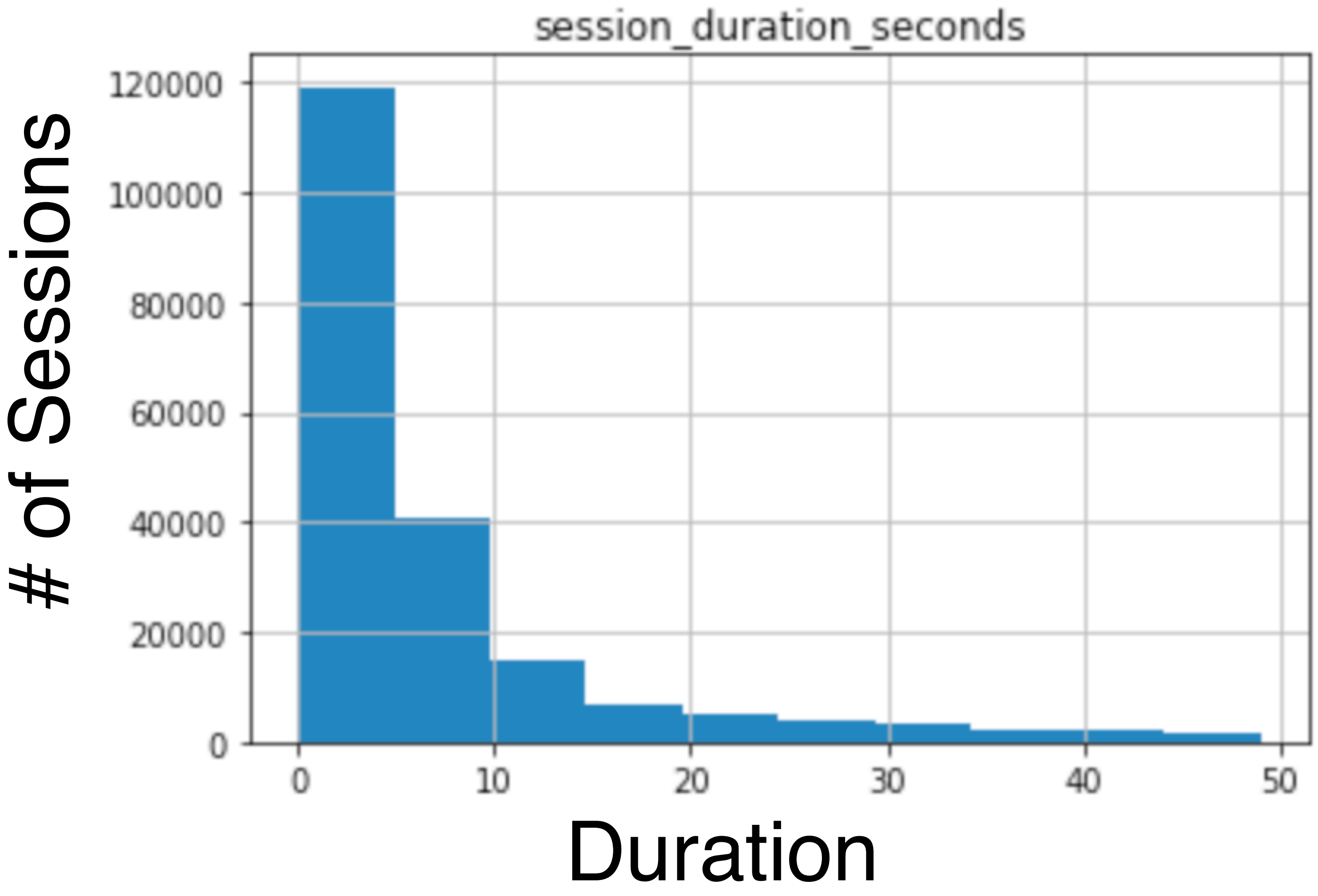
From "Examining the Examiners: Students' Privacy and Security Perceptions of Online Proctoring Services", by Aviv and friends

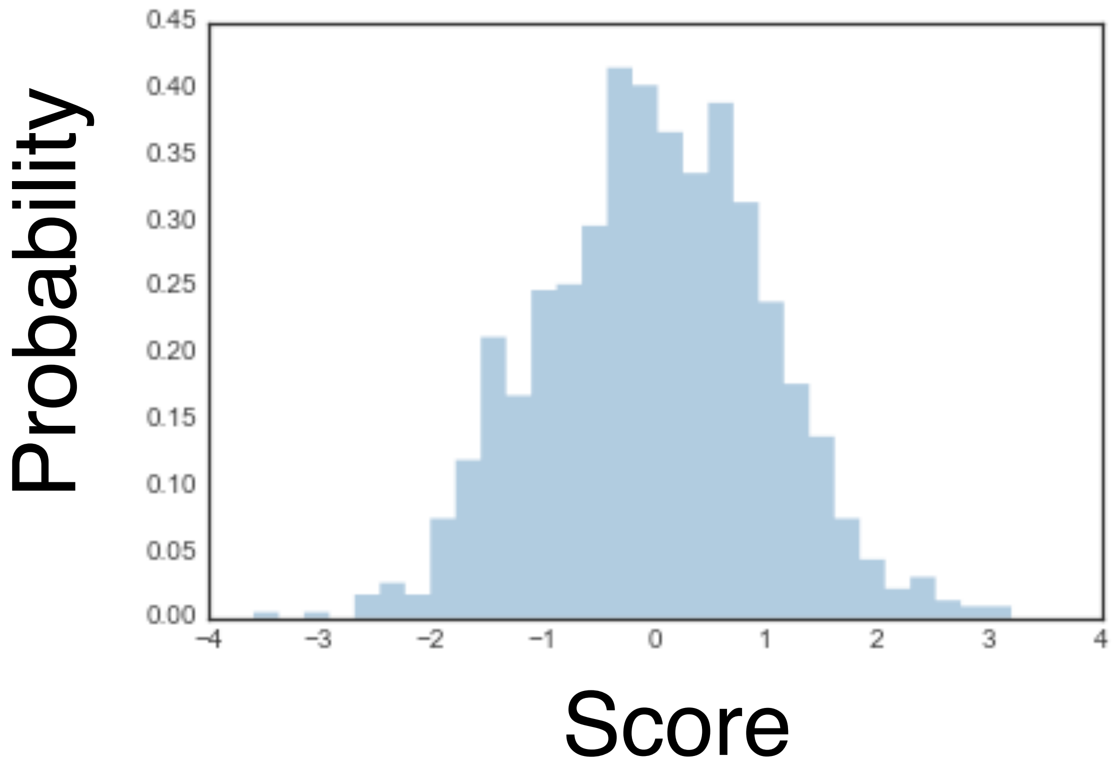
Histogram or PDF

How many samples occurred within each bucket range?

Or, what is the probability of hitting a certain range?

- Normalized histogram or continuous PDF





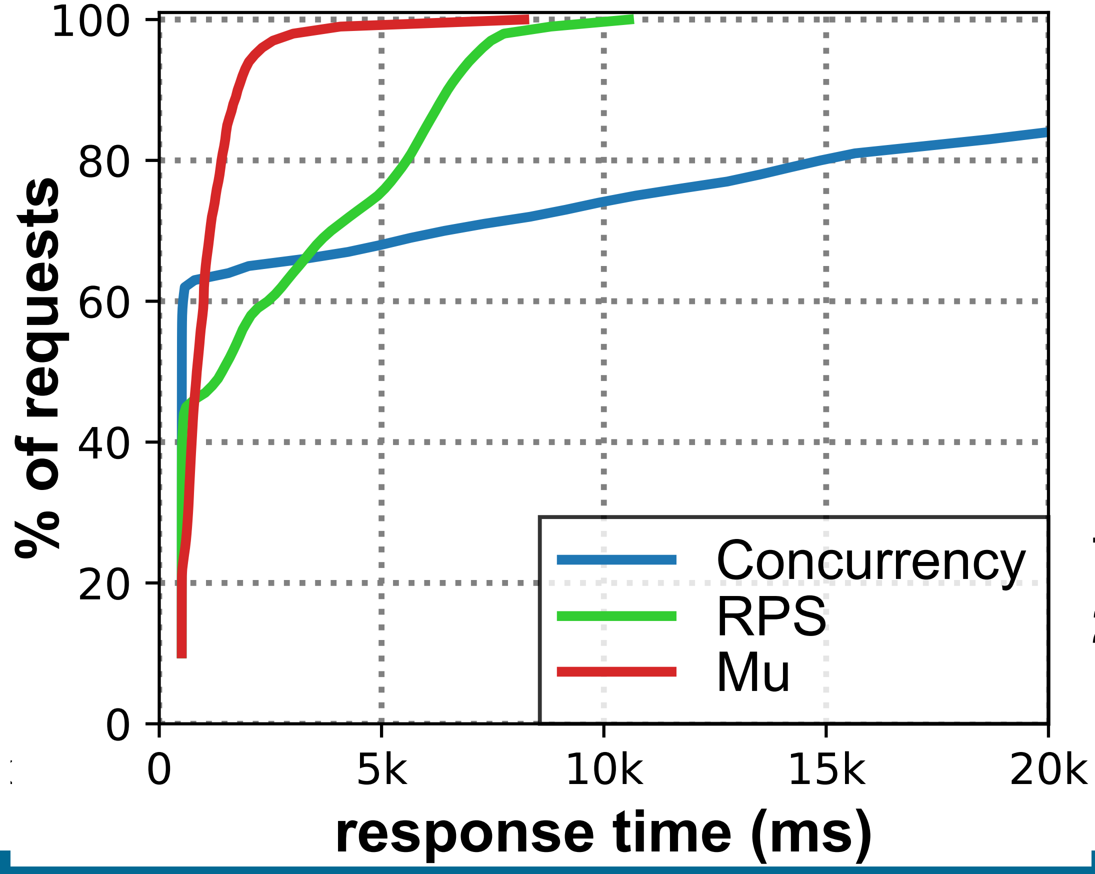
CDF

Cumulative Distribution Function

What percent of samples had this score or lower

How to read?

- Find quartiles and 99 percentile on y-axis and go right until you hit the line, then down to the x-axis



Avg:
7765
2605
1020

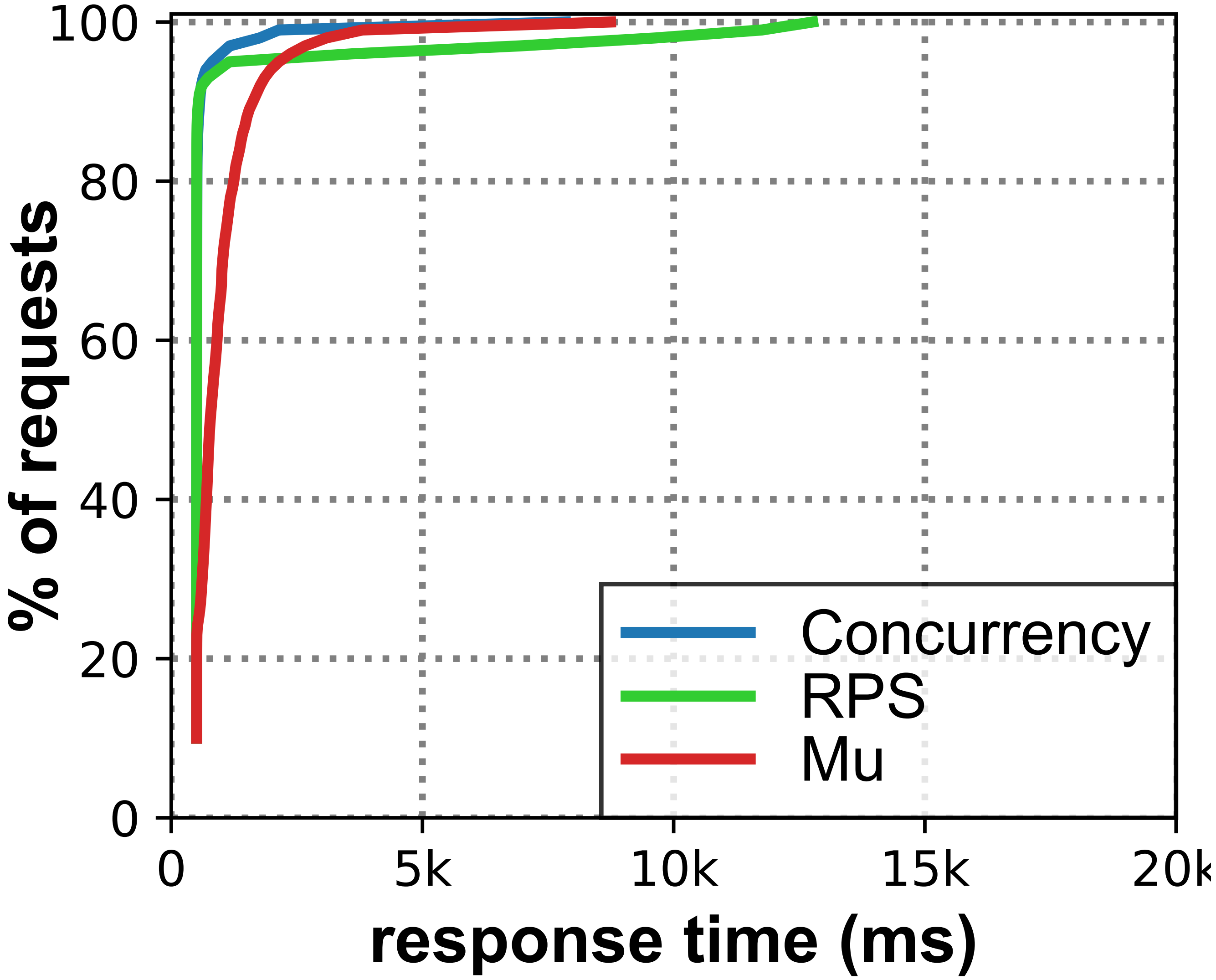
When to use CDF/PDF

Use when outliers/tail are important

Histogram is often easier to read

But, comparing multiple histograms takes a lot of space

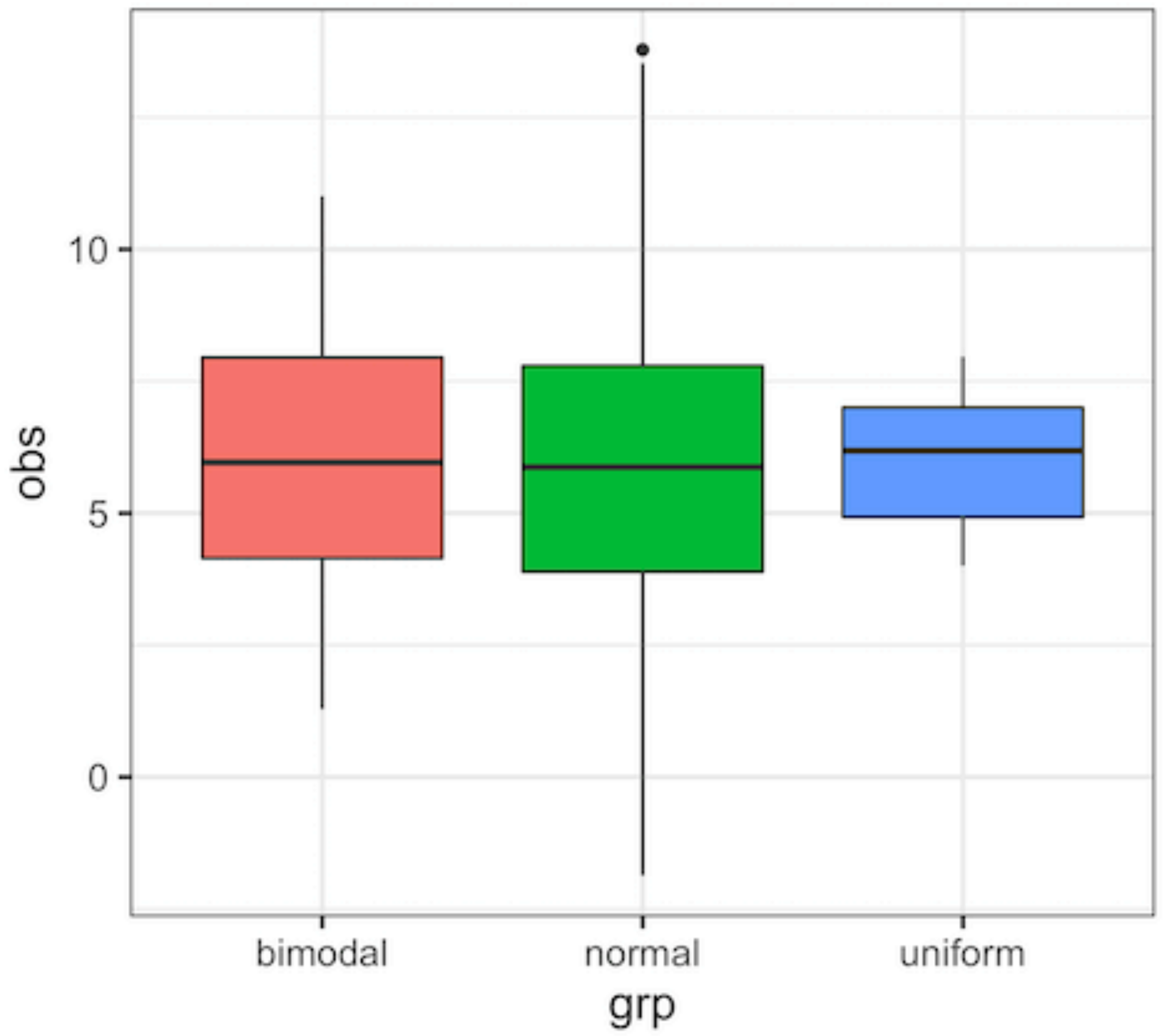
CDF is useful when comparing multiple distributions in one plot



Avg:
588
880
952

Box-Plots

Quickly display summary statistics and allow easy comparison



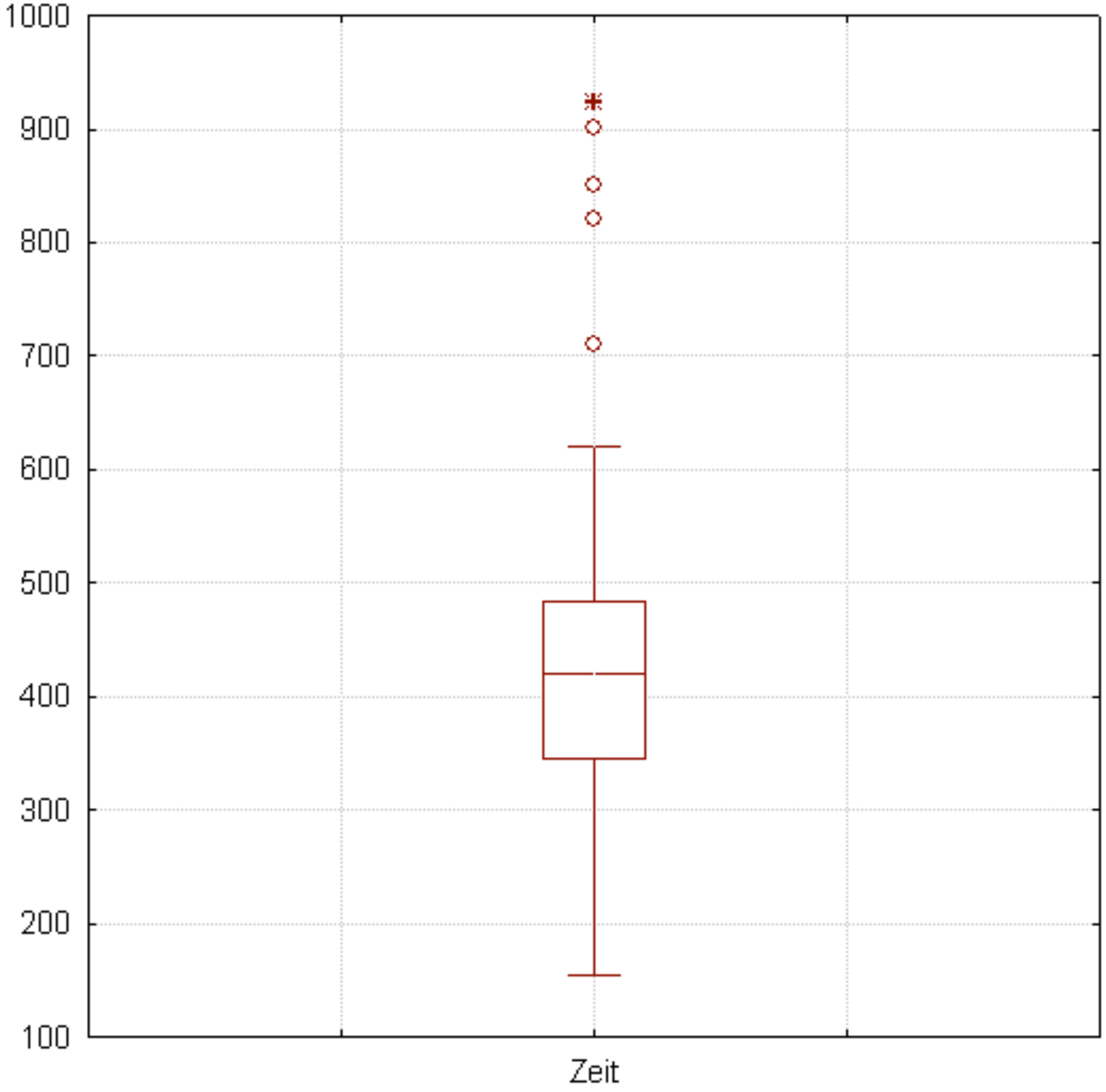
Whiskers: Min & Max
center line: Median (50%ile)
top/bottom: 75/25th %ile

Boxplot Variations

Whiskers aren't always max/min

Might show another statistics such as 2nd and 98th percentile, with outlier points explicitly shown

Paper should specify what the whiskers show

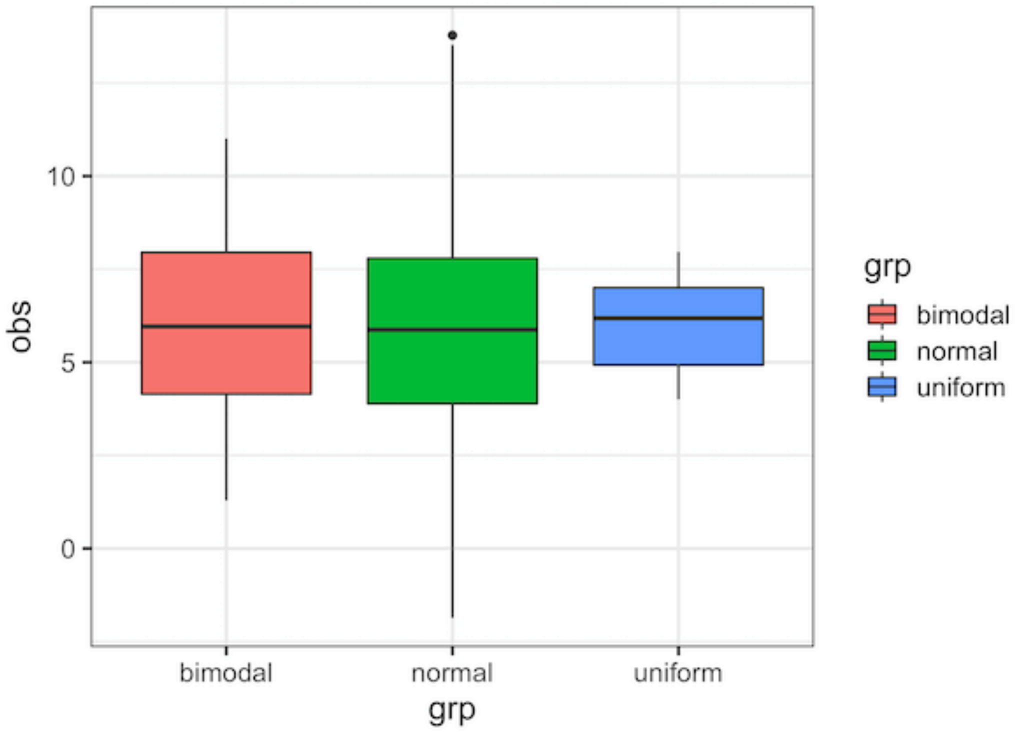


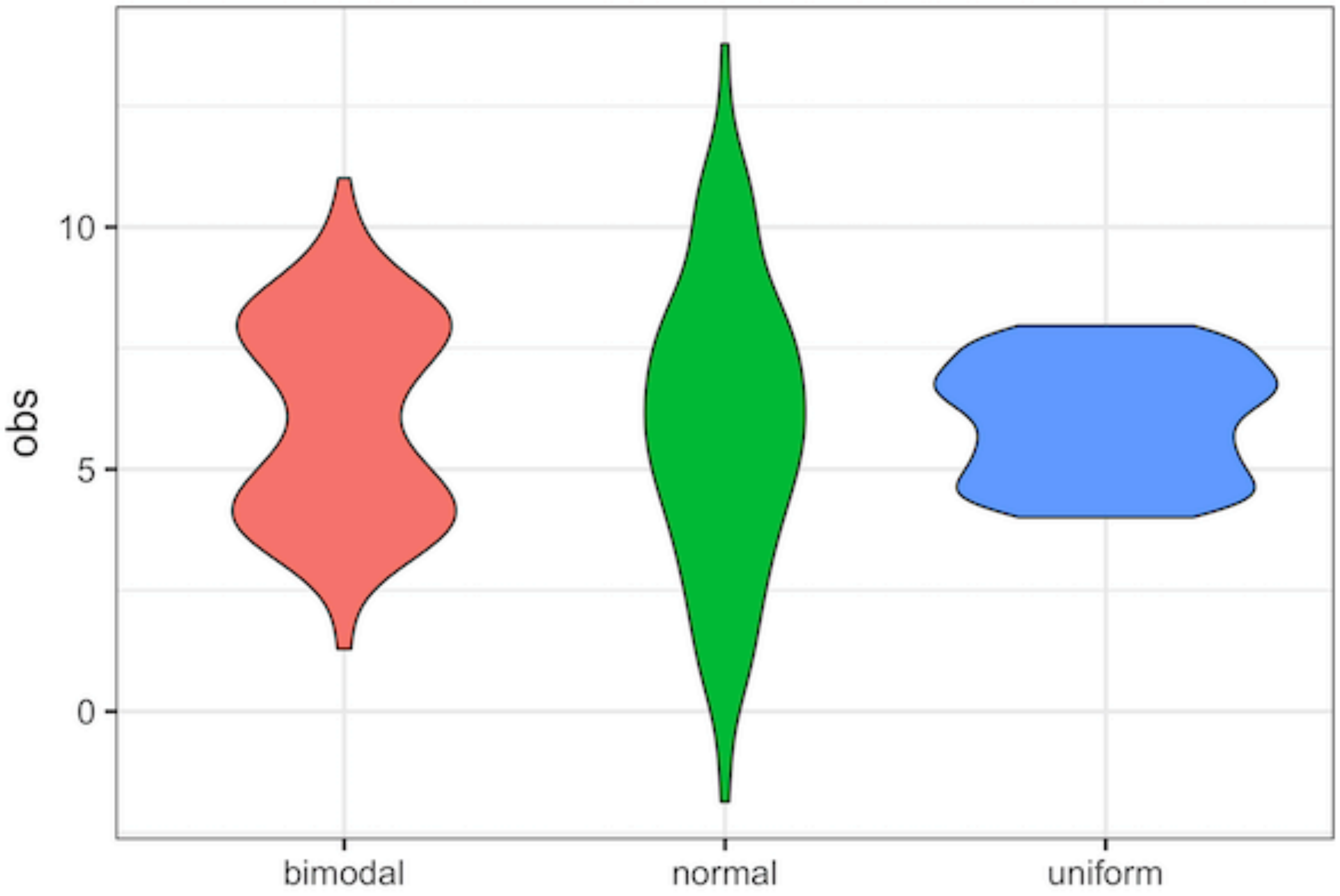
Violin Plots

Like a box plot, but shows the full distribution

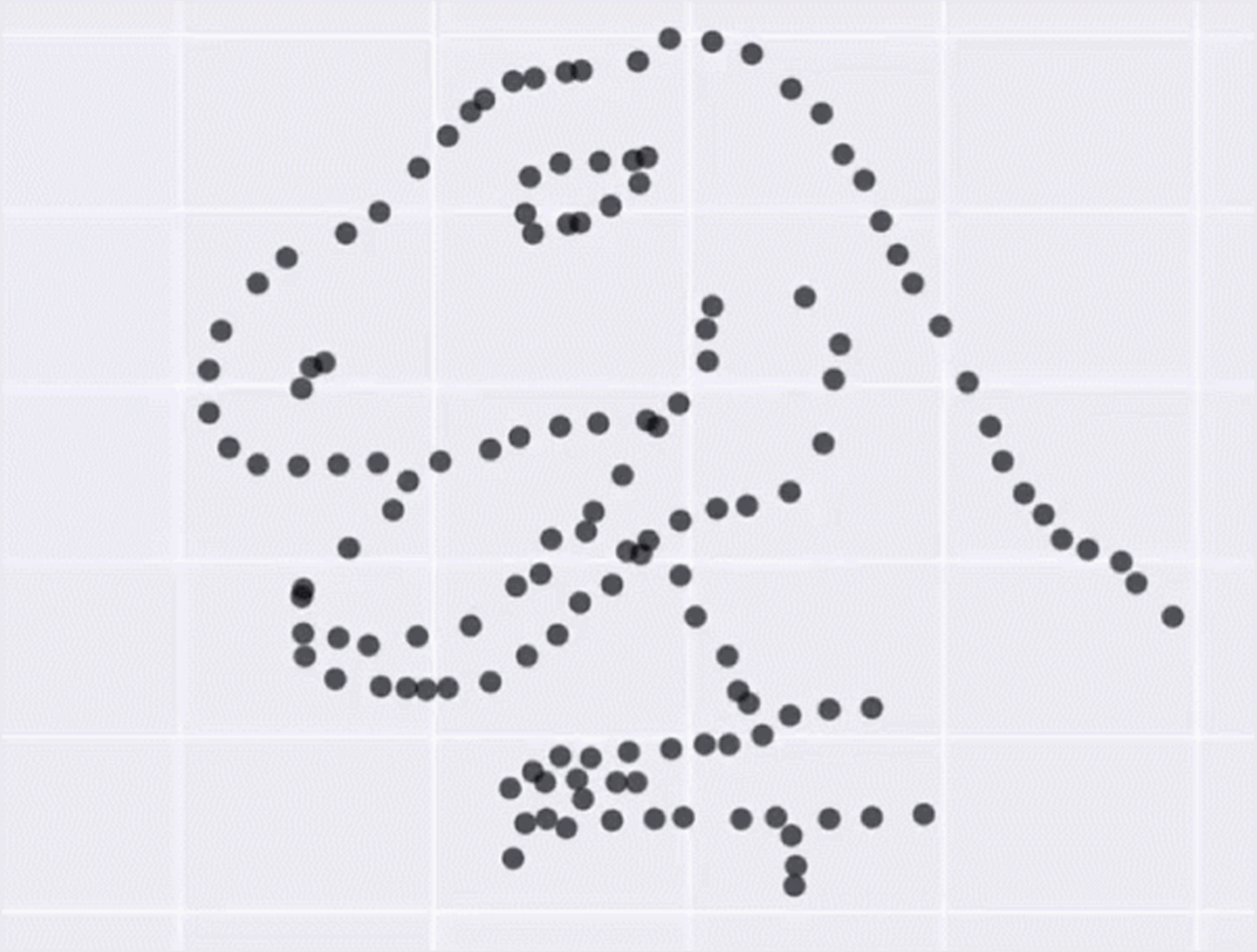
More descriptive

Useful if comparisons have a fundamentally different distribution



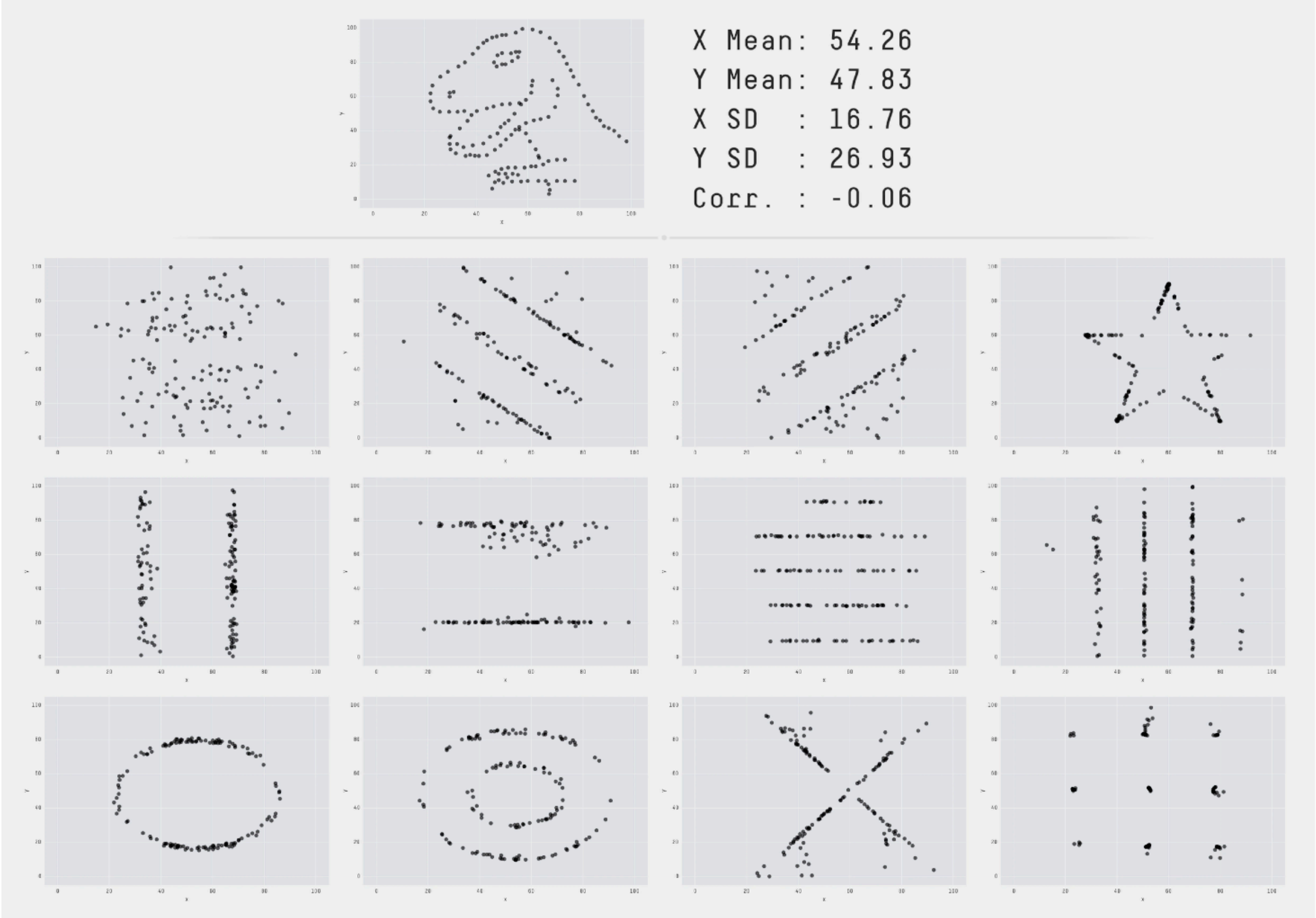


???



Same Stats, Different Graphs

https://www.autodesk.com/research/publications/same-stats-different-graphs



```
X Mean: 54.26
Y Mean: 47.83
X SD  : 16.76
Y SD  : 26.93
Corr. : -0.06
```

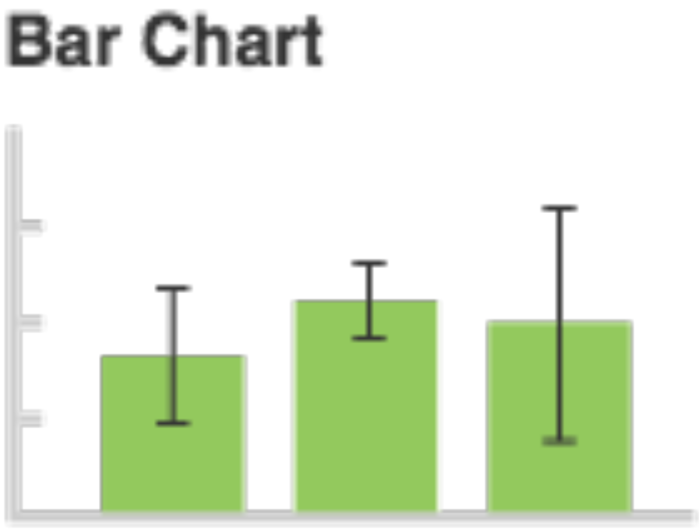
Graphs

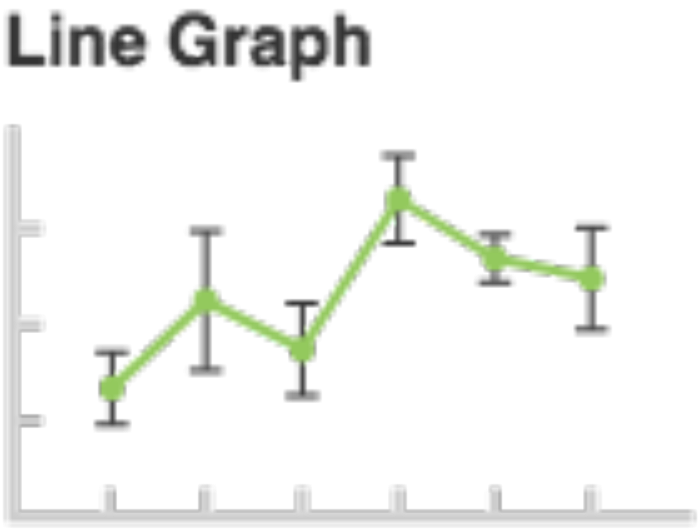
Stick with “standard” graph types

- Unless you have a good reason
- Standards: Scatter plot, line plot, bar plot
- Avoid pie charts and “infographic” look

Include error bars!

- Use standard deviation
  or confidence interval





Every graph in this paper was a radar chart… odd

Graphs

What tool do you use to make plots?

- gnuplot, matplotlib, seaborn, tableau, matlab, R…

Avoid tools like Excel

- Most papers I review with Excel graphs I reject (not usually because of the graphs, but it is a sign of amateur-ness)
- OK for initial data exploration by yourself

Use the same tool as your lab-mates

- Have lab scripts for making beautiful graphs

Jupyter Notebooks

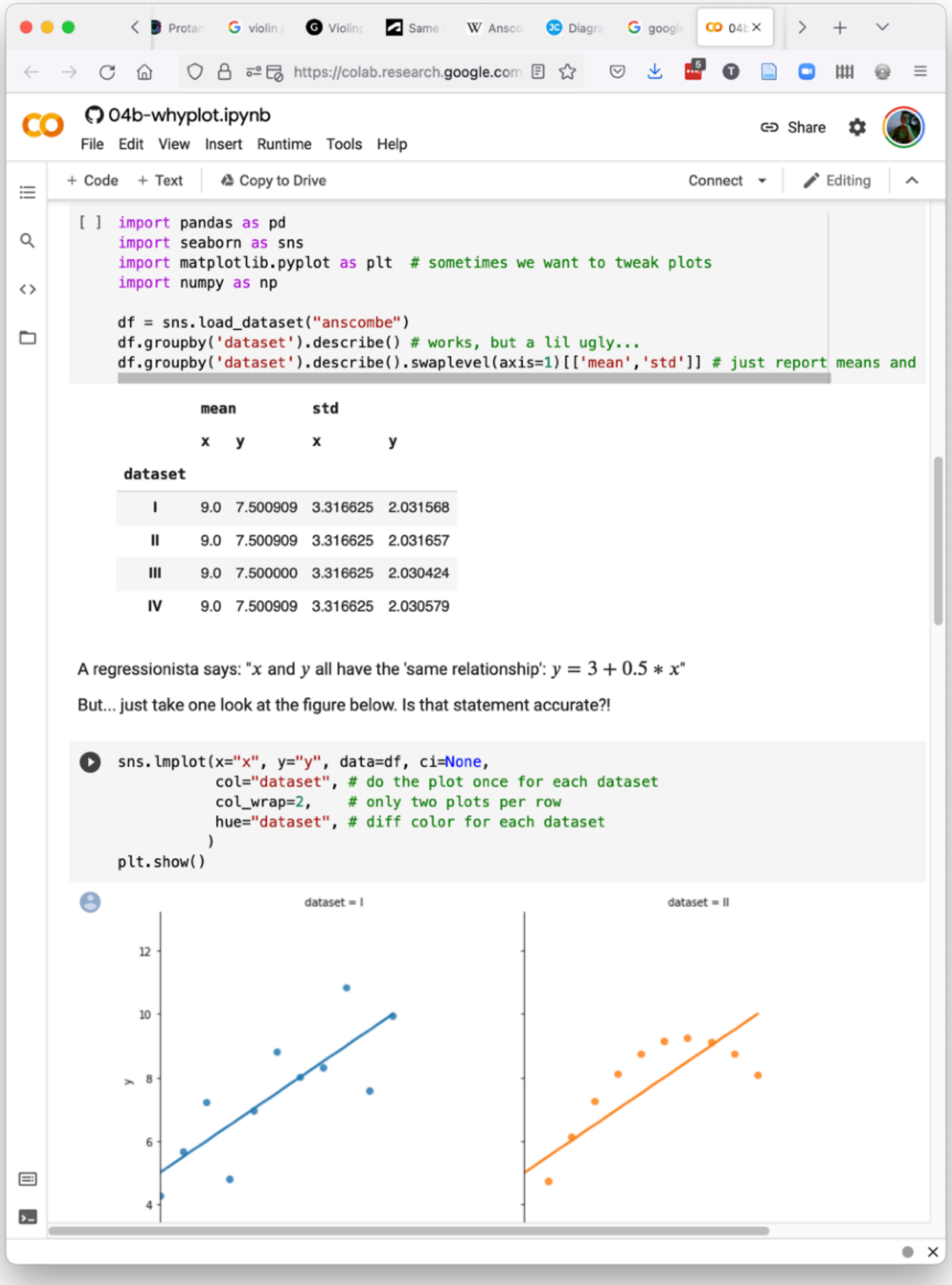
This is the "right" way to do it

Browser-based (or VS Code extension!) editor and python runtime

Parse, analyze, and visualize data all in one place

Easy to share with others

- Github displays rendered notebooks
- Google Collab allows shared notebooks / execution environments



```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt  # sometimes we want to tweak plots
import numpy as np

df = sns.load_dataset("anscombe")
df.groupby('dataset').describe() # works, but a lil ugly...
df.groupby('dataset').describe().swaplevel(axis=1)[['mean','std']] # just report means and
```

	mean		std	
	x	y	x	y
dataset				
I	9.0	7.500909	3.316625	2.031568
II	9.0	7.500909	3.316625	2.031657
III	9.0	7.500000	3.316625	2.030424
IV	9.0	7.500909	3.316625	2.030579

A regressionista says: "x and y all have the 'same relationship': $y = 3 + 0.5 * x$"

But... just take one look at the figure below. Is that statement accurate?!

```
sns.lmplot(x="x", y="y", data=df, ci=None,
           col="dataset", # do the plot once for each dataset
           col_wrap=2,    # only two plots per row
           hue="dataset", # diff color for each dataset
          )
plt.show()
```

Scripting Experiments

More experiments is more data

More data is more information

More information is more value

- As long as you have the time to process it

You should run as many experiments as possible

- Uses scripts to automate gathering AND analyzing data as much as possible
- Try to make general purposes scripts you can reuse in the future

You will need to repeat experiments, so make it easy!

Basic Shell Scripting

Often output of experiments is messy

Bash scripts that combine basic Unix utilities are usually the fastest way to pre-process your data

Or you can do this all in Python, but may be simpler as part of your experiment running scripts

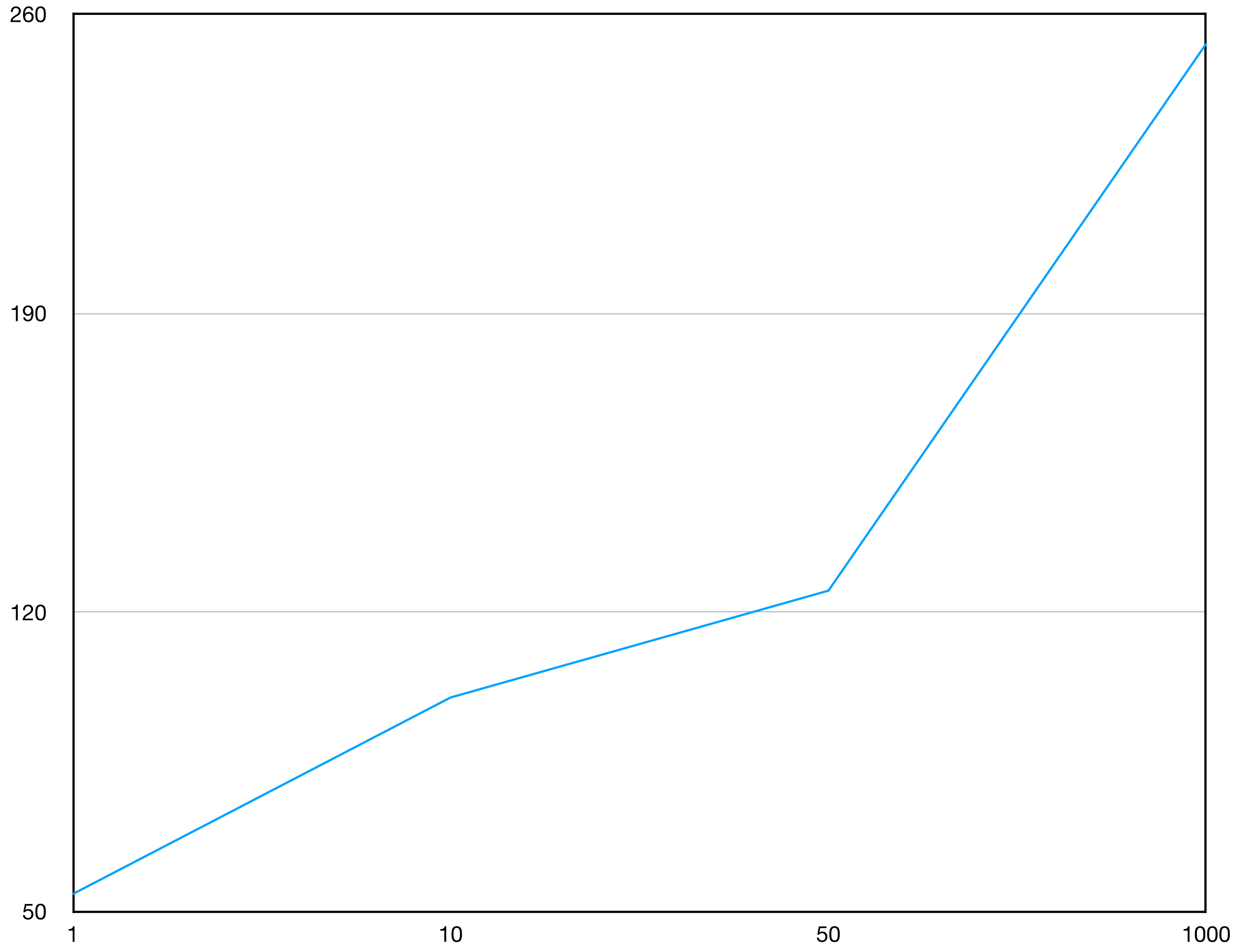
```
grep “exp1” file.data | awk ‘{print $3}’ | sort -n | tail
```

sed - replace

awk - processing rows / columns in a file

grep - filtering

What is wrong with this?



What is wrong with this?

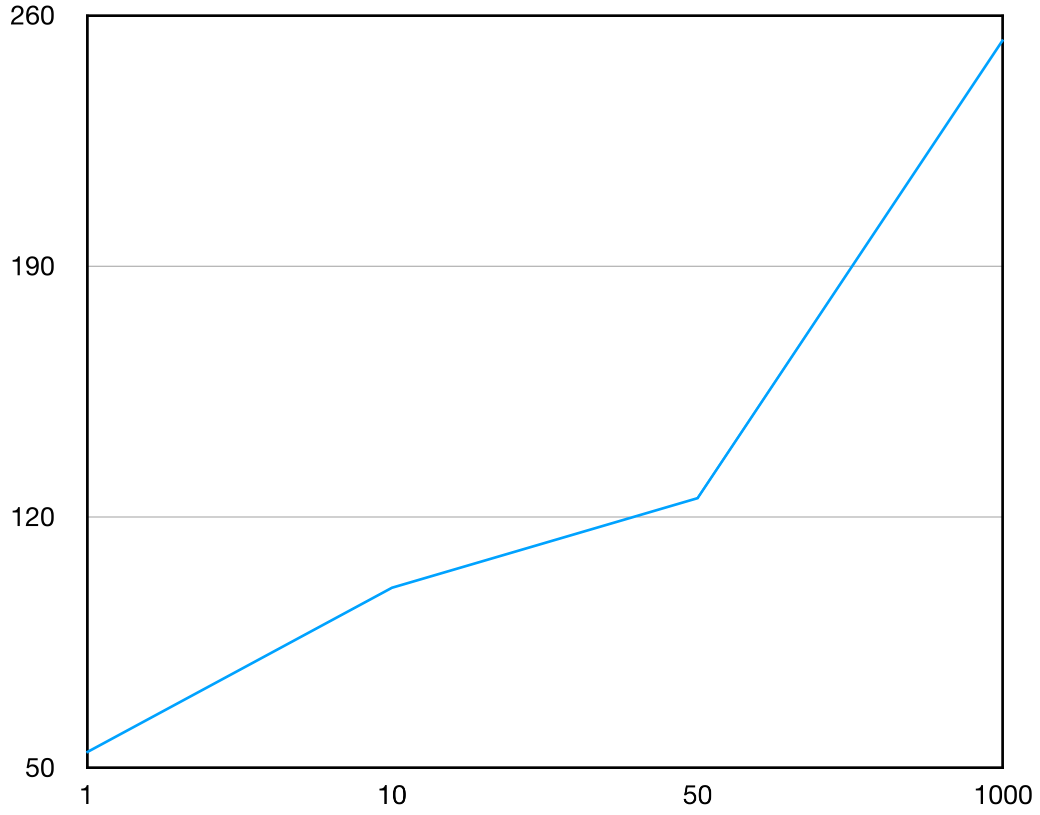
Needs axis labels

No error bars

Irregular x-axis

Non-zero y-axis

Font sizes too small



Aspect ratio may be awkward in paper layout

- Typically either want 2 square images or 1 wide image for 2-column layout

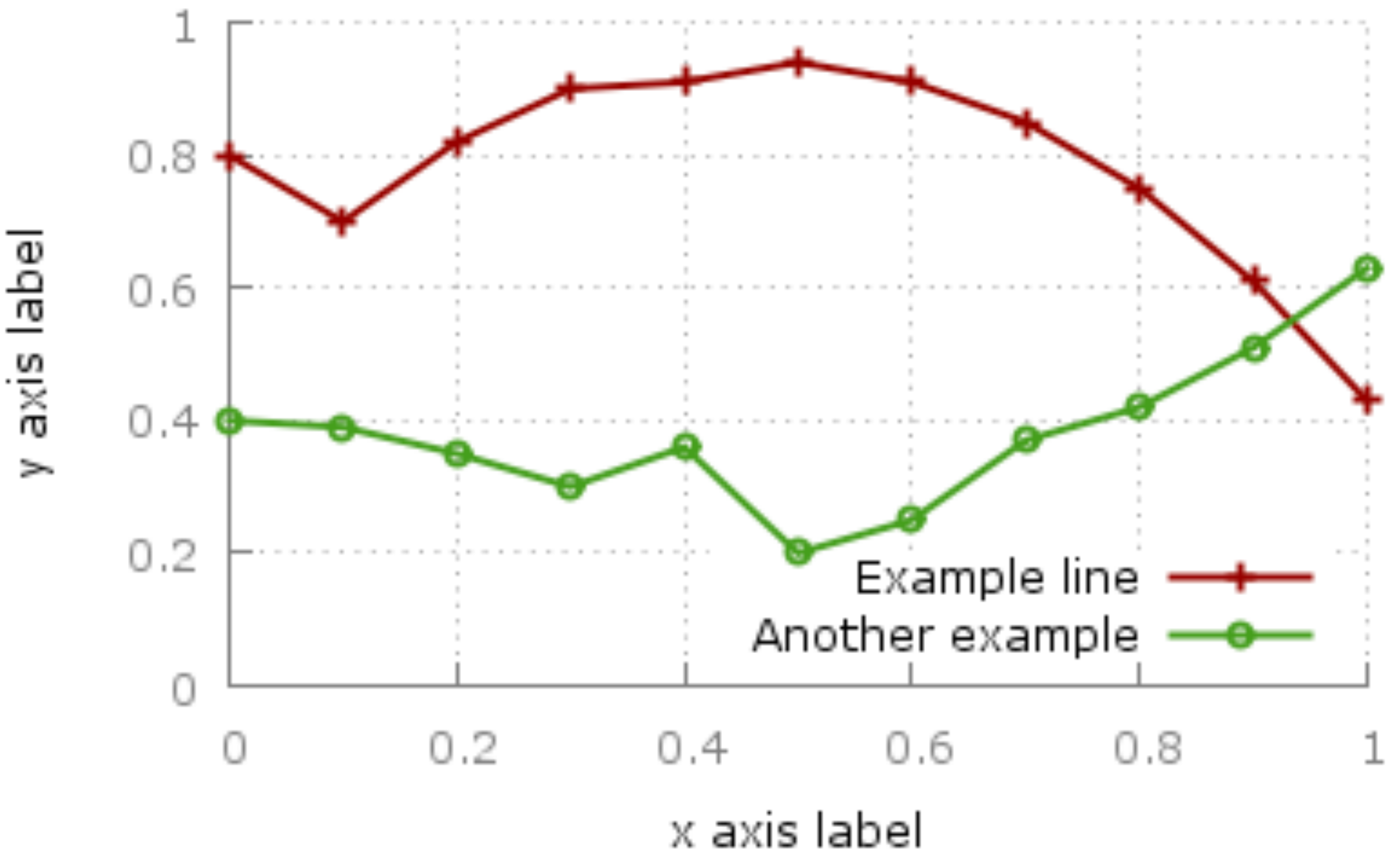
Graph Tips

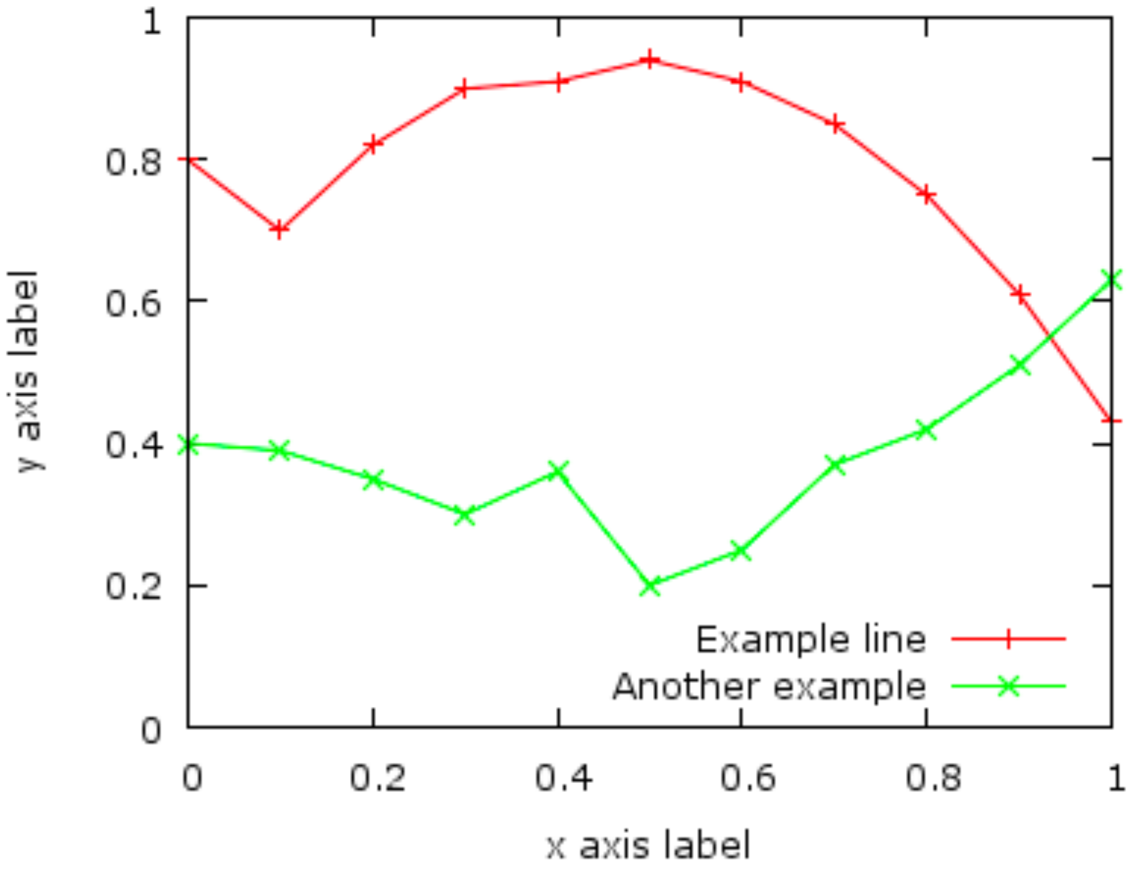
Use:

- Thick lines
- Very large fonts
- Axis labels
- Wide format

Avoid:

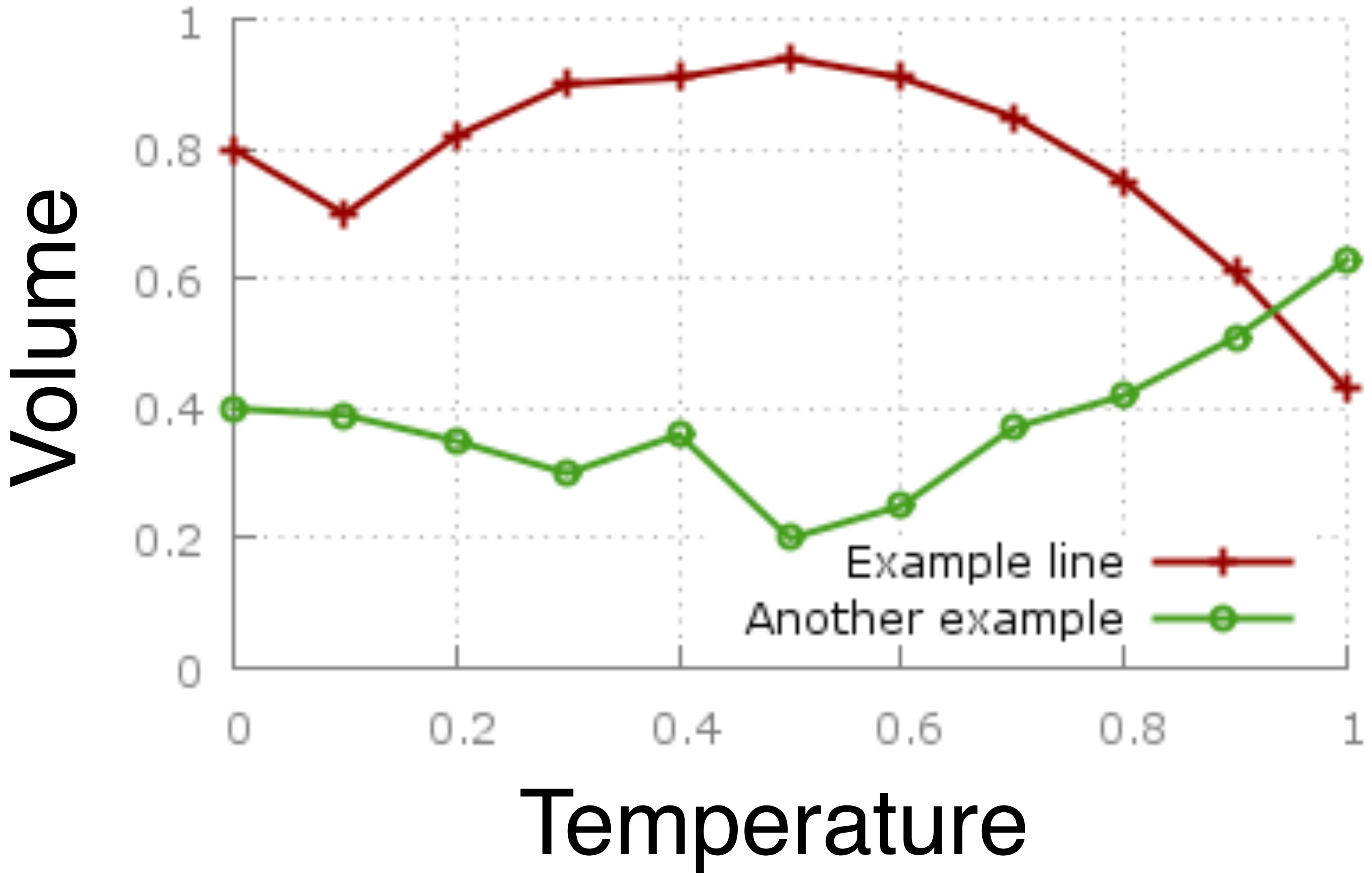
- Similar colors (check BW!)
- Non-0 starting axes
- Titles (if you have caption)
- Square aspect ratio





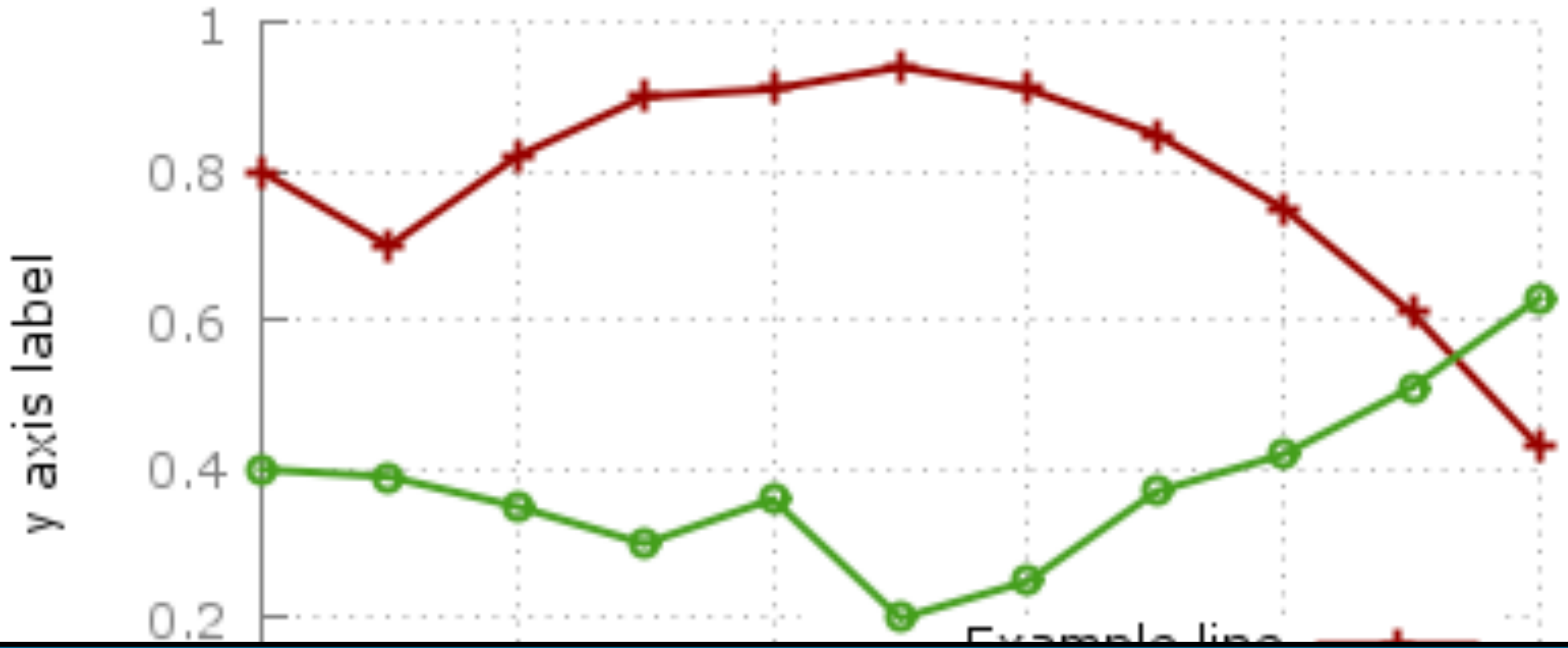
Line Graphs…

My high school science teacher would (correctly) fail me for making this graph… why?



Line Graphs…

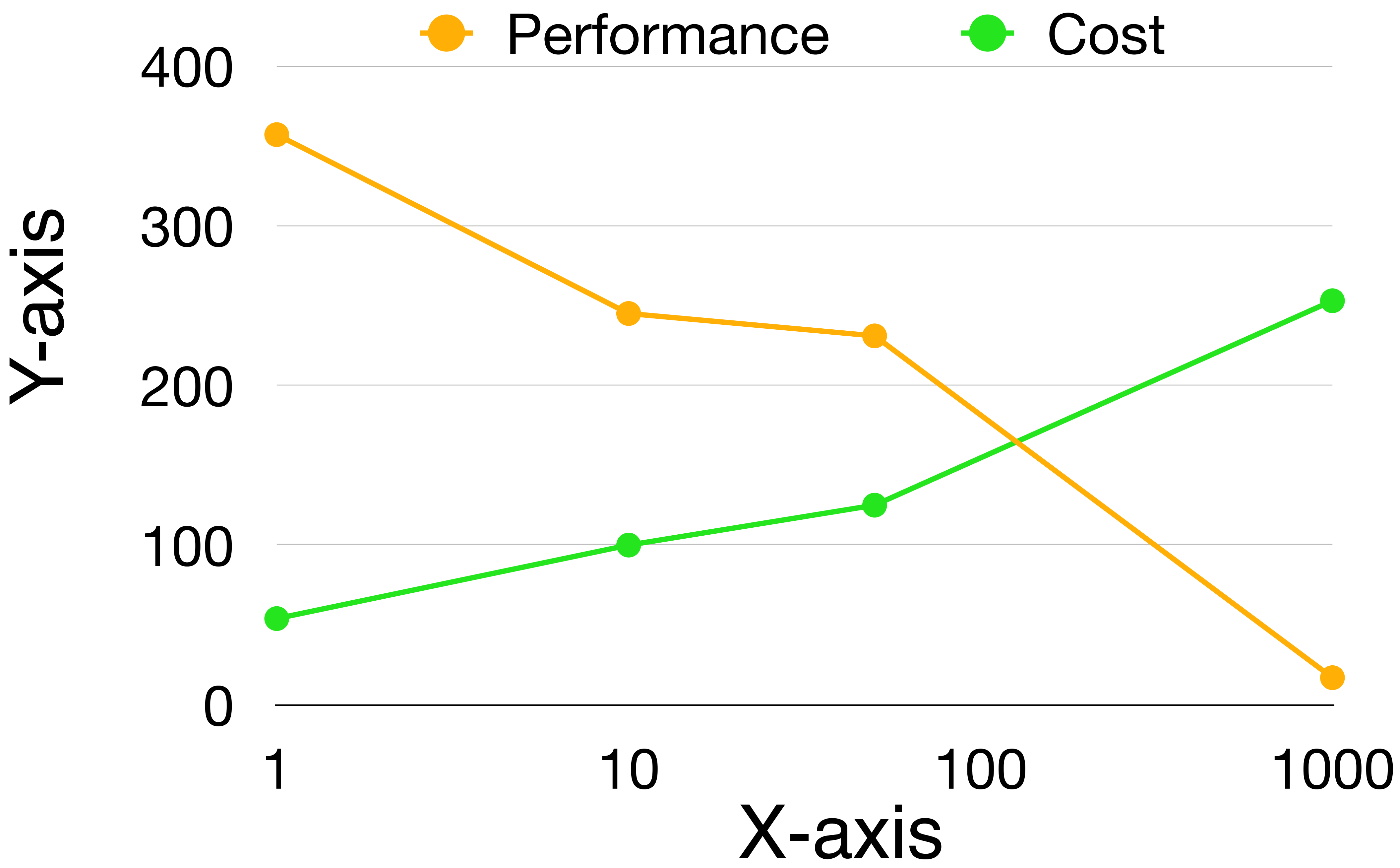
My high school science teacher would (correctly) fail me for making this graph… why?



"Connect the dots" is really unscientific!
Trend lines are way better!

But this is what the community expects!

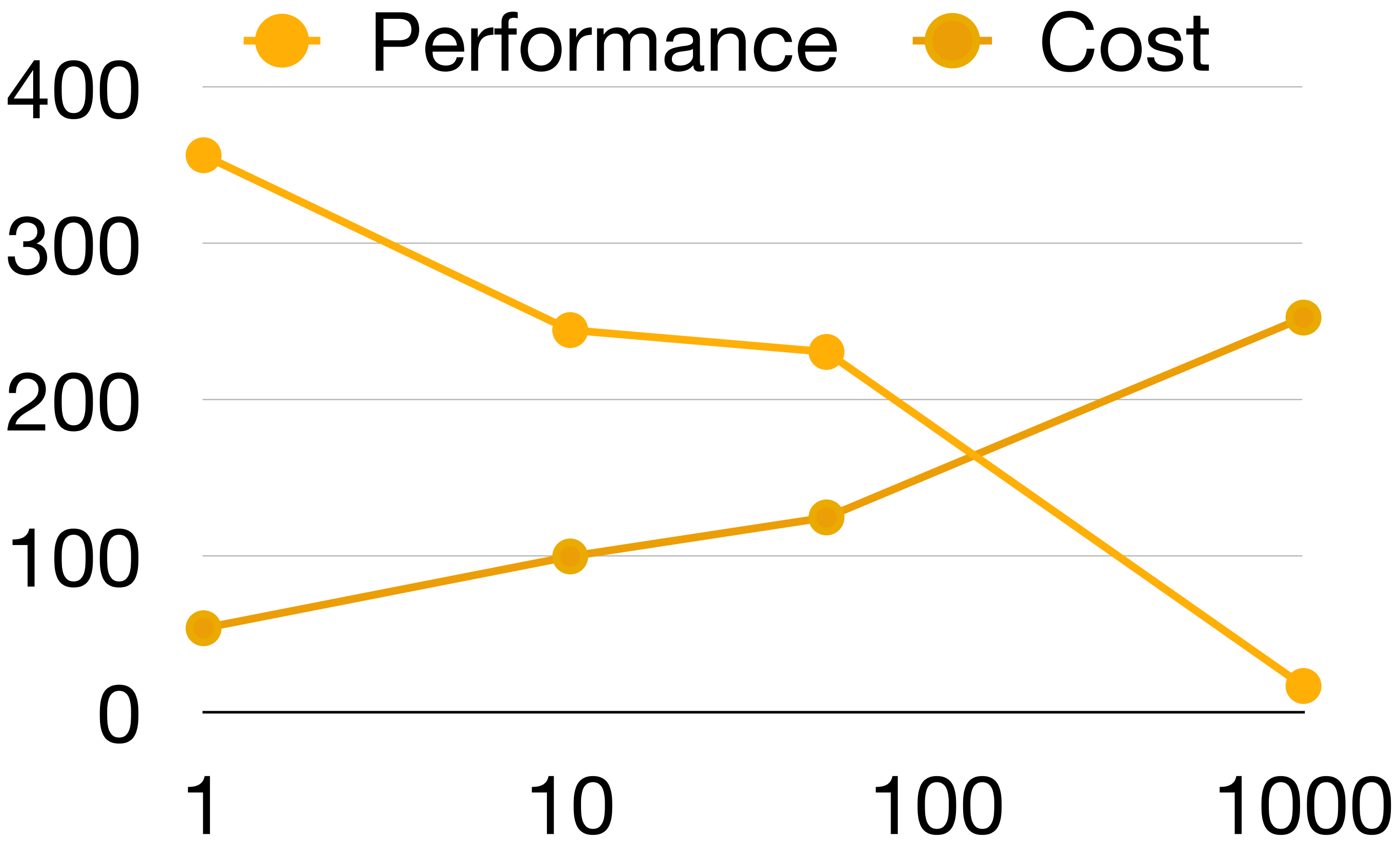
Problem with this?

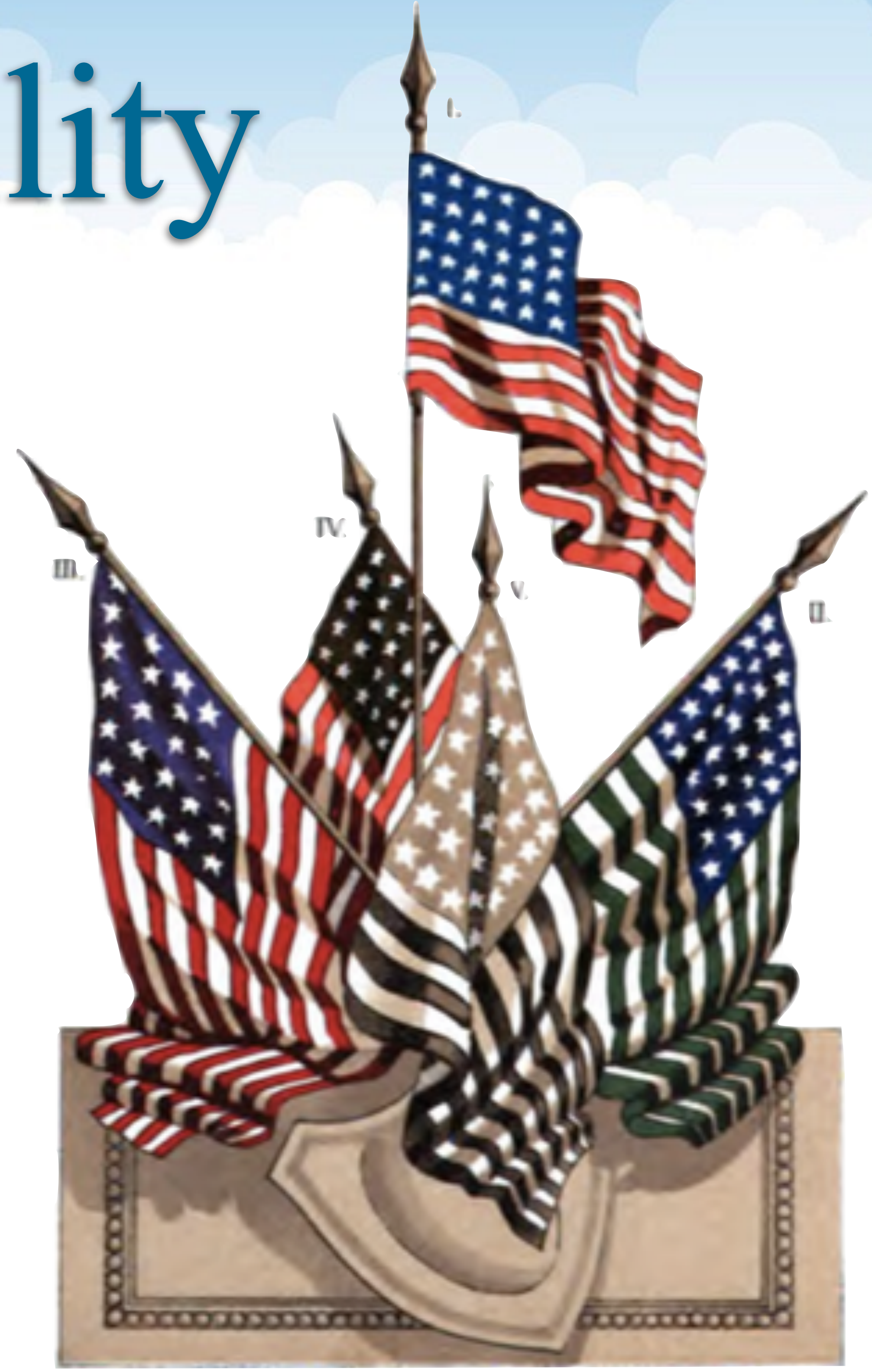


Accessibility

Color blindness

- 1 in 12 men, 1 in 200 women





THE AMERICAN FLAG
AS SEEN
I.- BY MOST PEOPLE
II.- BY RED-BLIND PERSONS
III.- BY GREEN-BLIND PERSONS
IV.- BY VIOLET-BLIND PERSONS
V.- BY TOTALLY COLOR-BLIND PERSONS

Accessibility

Color blindness

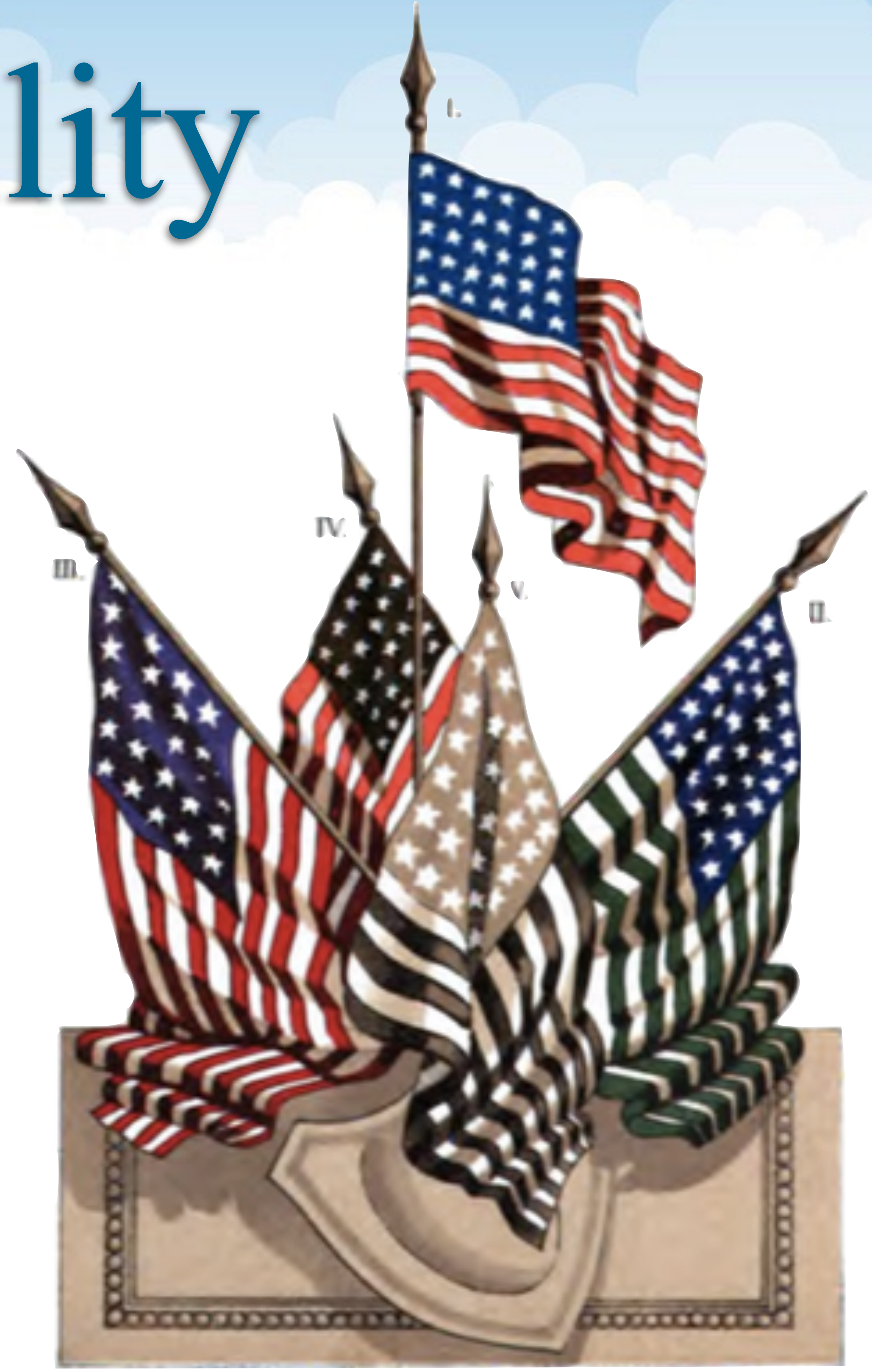
- 1 in 12 men, 1 in 200 women

Blindness

- about 8 million people in the US have a visual disability

Many types of disabilities to be aware of: sight, sound, touch, mobility

Remember these!



THE AMERICAN FLAG
AS SEEN
I.- BY MOST PEOPLE
II.- BY RED-BLIND PERSONS
III.- BY GREEN-BLIND PERSONS
IV.- BY VIOLET-BLIND PERSONS
V.- BY TOTALLY COLOR-BLIND PERSONS

Visual Recipes

Slides:

- Mixture of text and images
- Keep bullets simple, fonts clear
- Use animation sparingly for emphasis

Diagrams:

- Pick the right level of abstraction; focus on relationships
- Use LARGE fonts!
- Be sure colors work in B&W

Graphs:

- Use easy to understand plot types
- Use thick lines and be sure they are distinguishable
- Use LARGE fonts!
- Be sure colors work in B&W