bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

Learn how to distribute your work using this licence

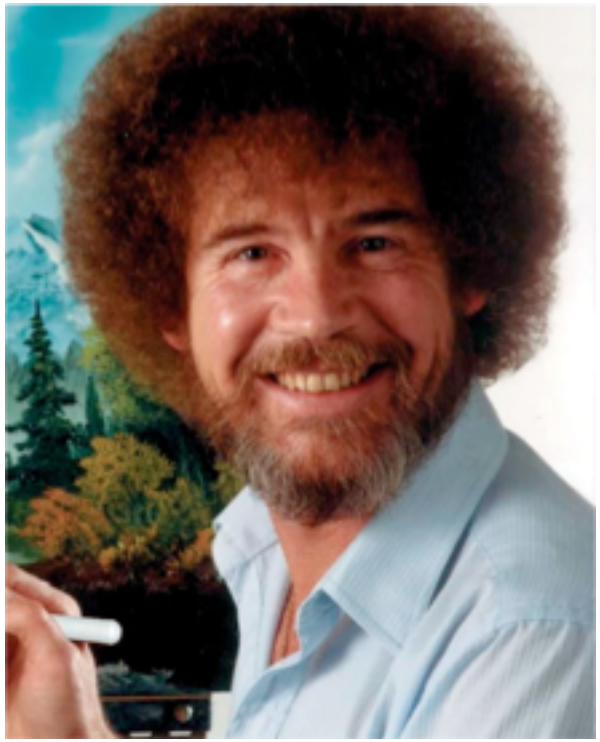
Module 7
Data Visualization

Anamaria Crisan

Infectious Disease Genomic Epidemiology

July 10-12, 2018

bioinformatics.ca

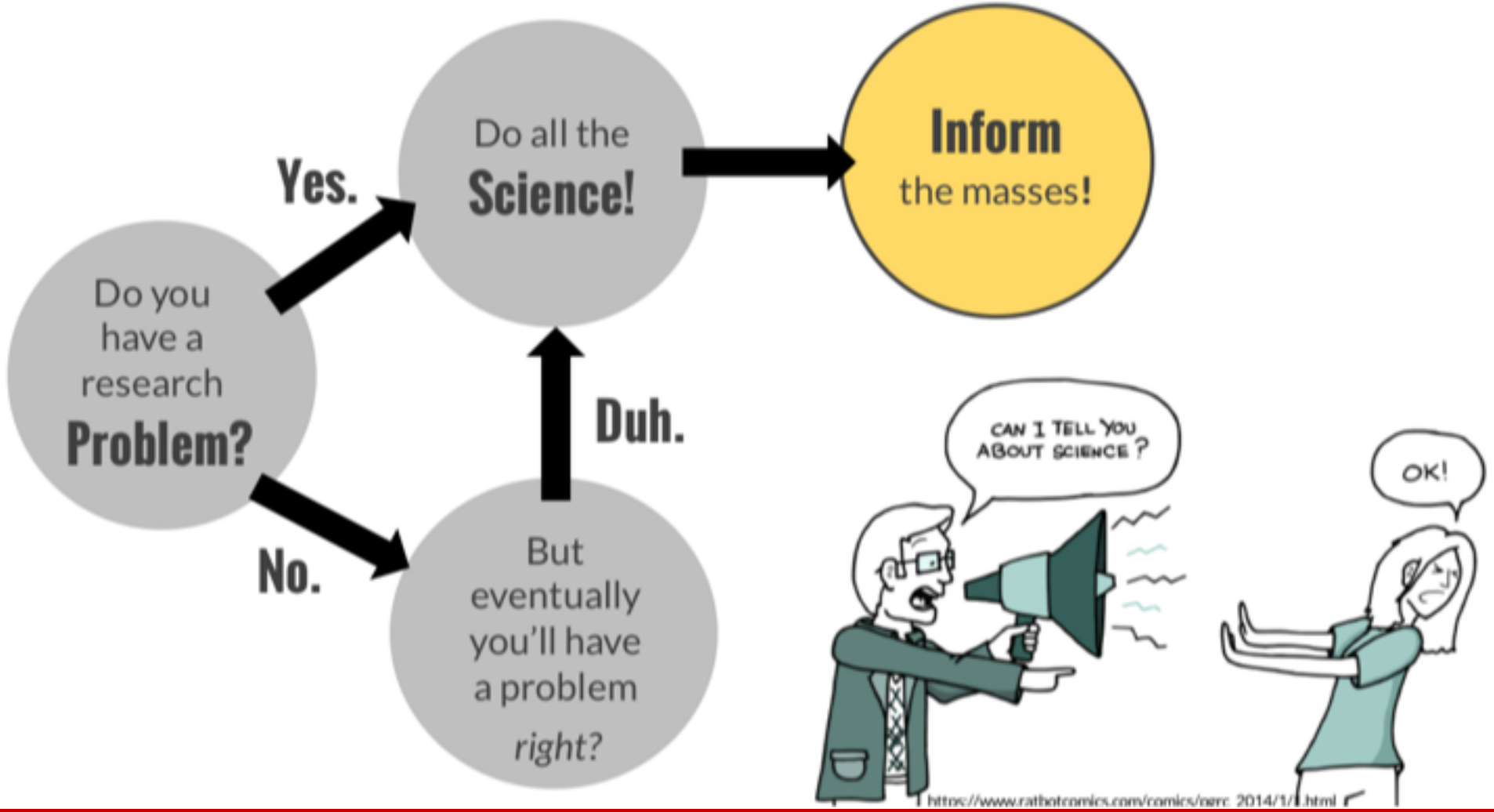


Learning Objectives of Module

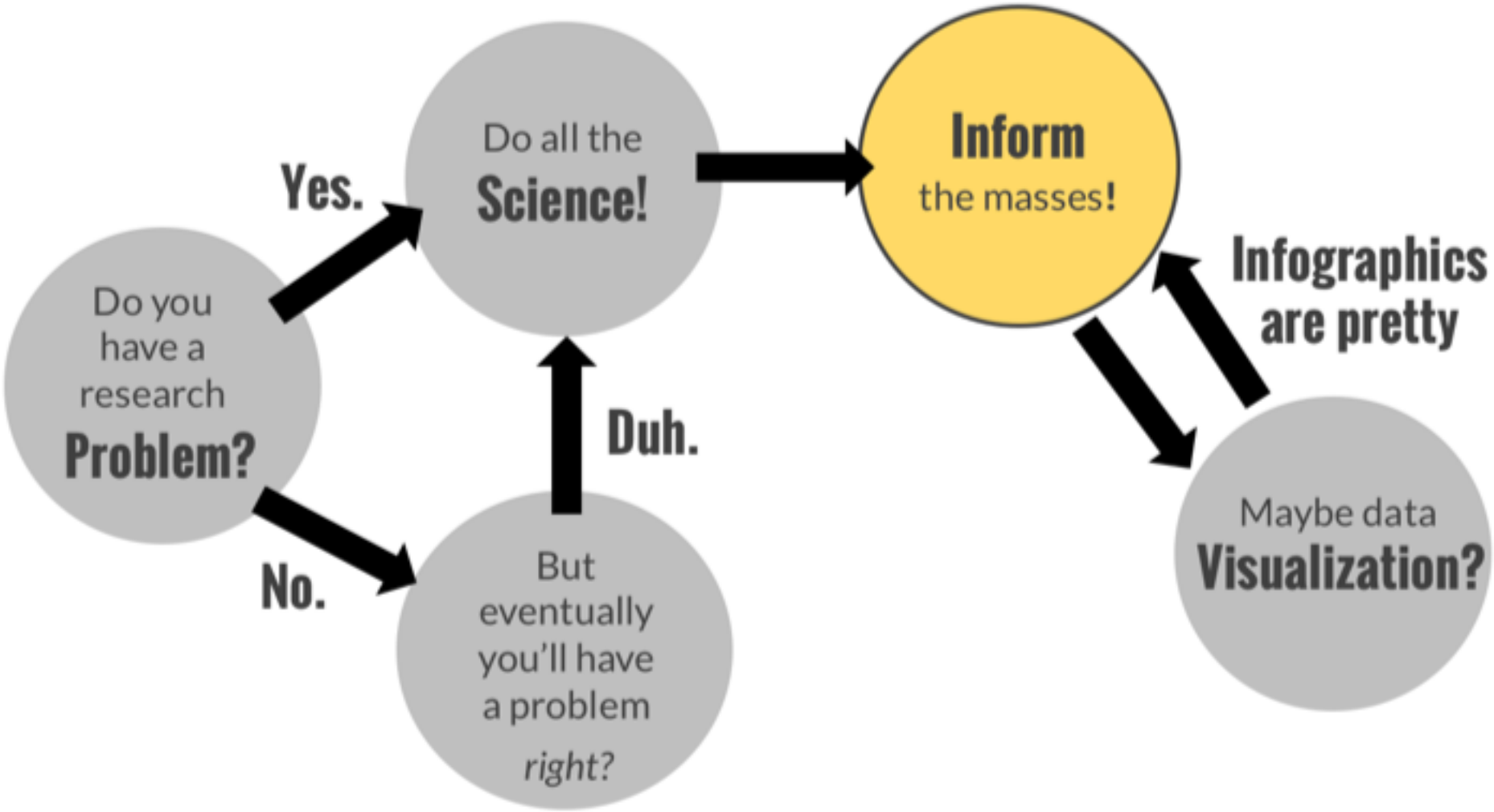
- Understand the process of encoding and decoding data that is visualized
- How to think systematically about data visualizations
- Know what a visualization design space is and how to reason between different visualization design choices

Why should we visualize data?

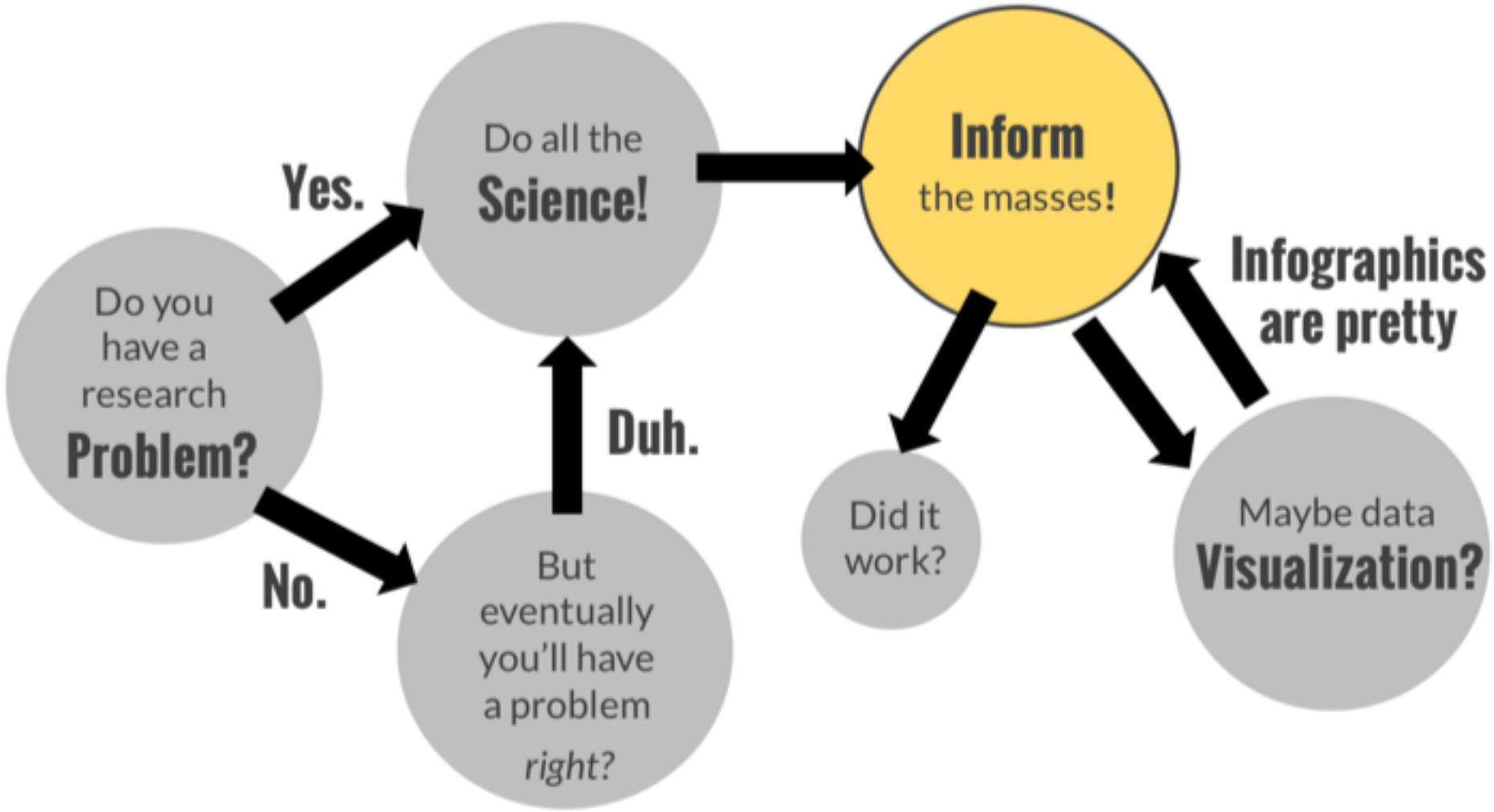
Data visualization in infectious disease GenEpi



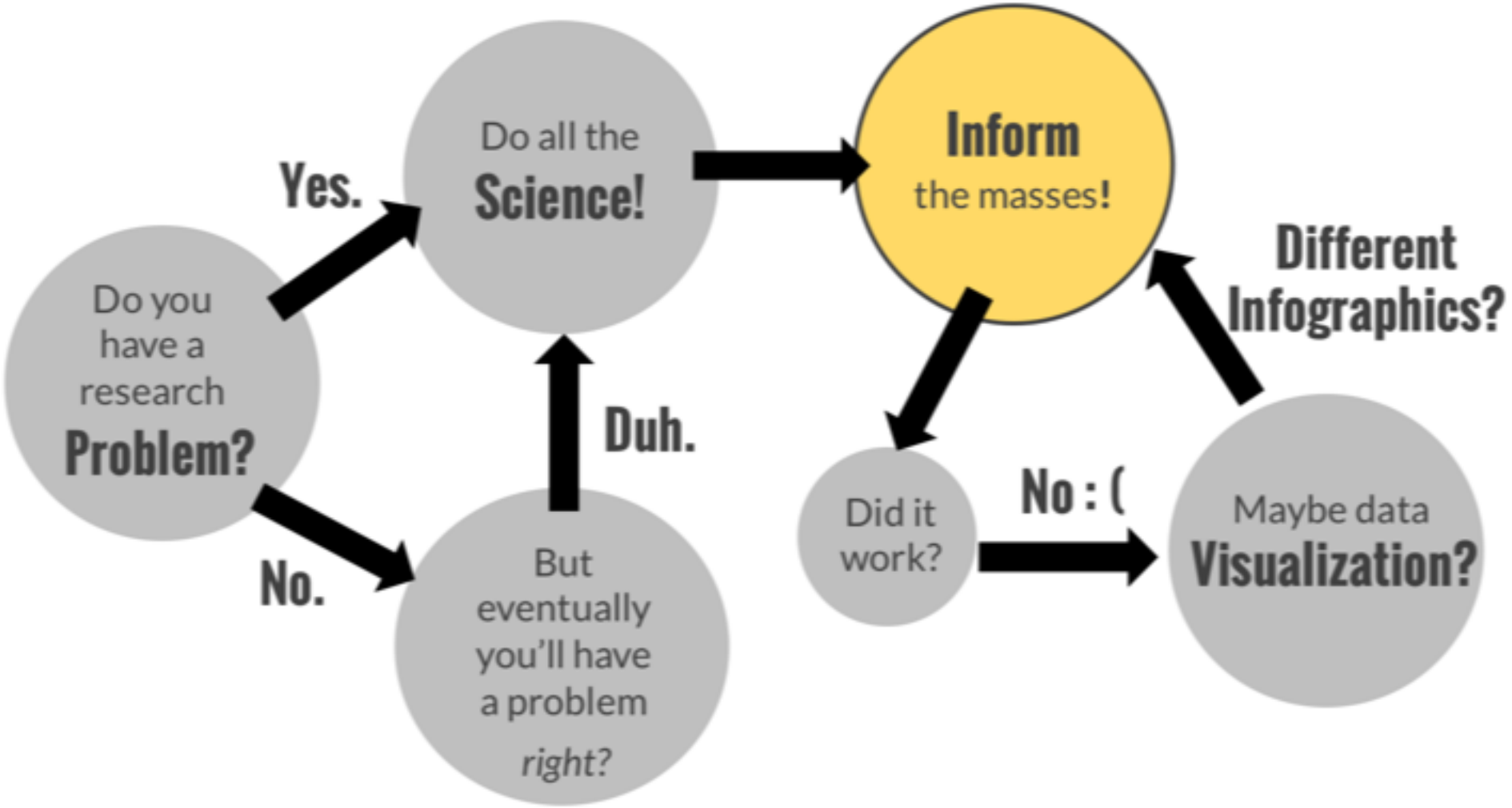
Data visualization in infectious disease GenEpi



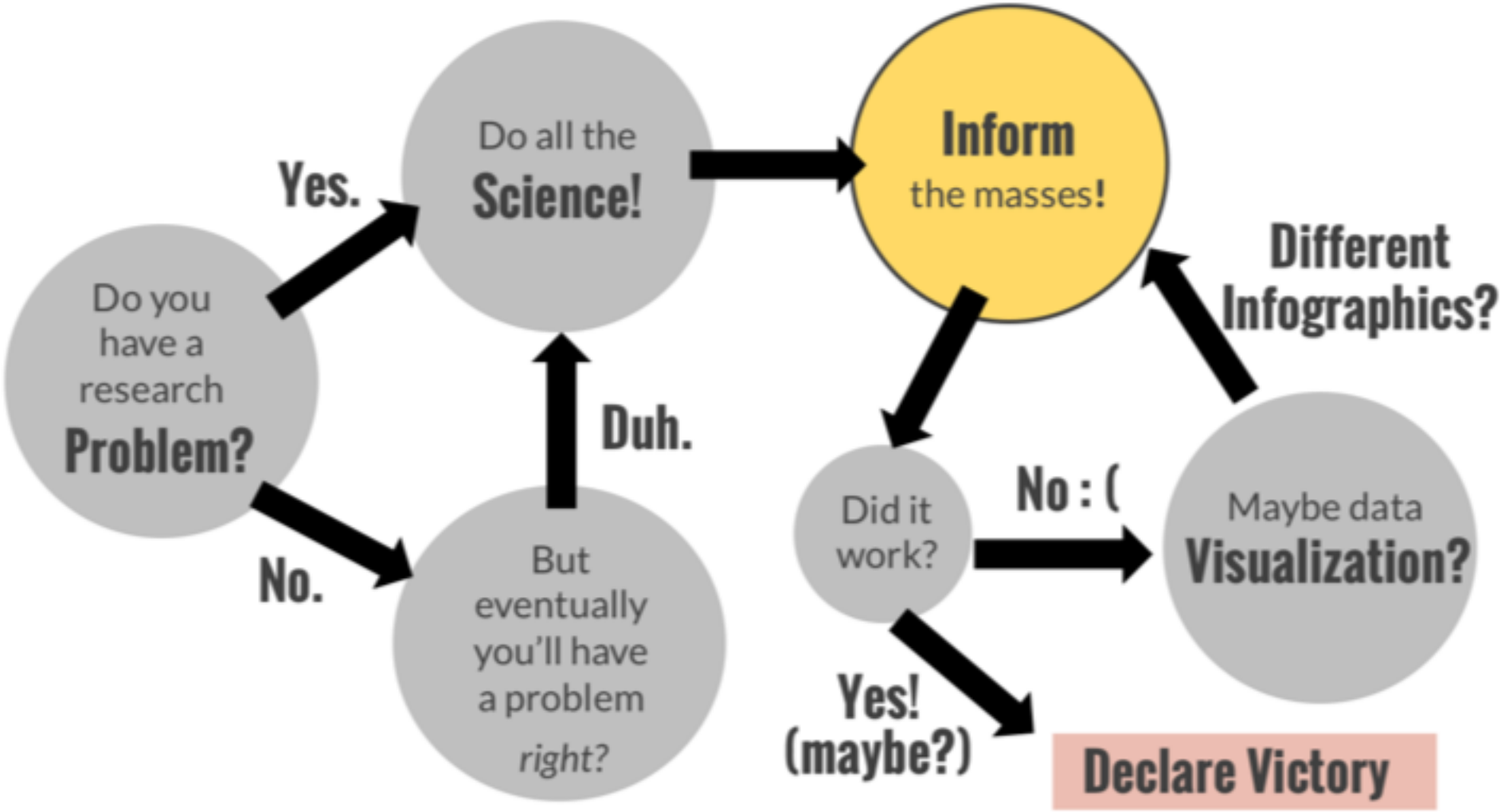
Data visualization in infectious disease GenEpi



Data visualization in infectious disease GenEpi

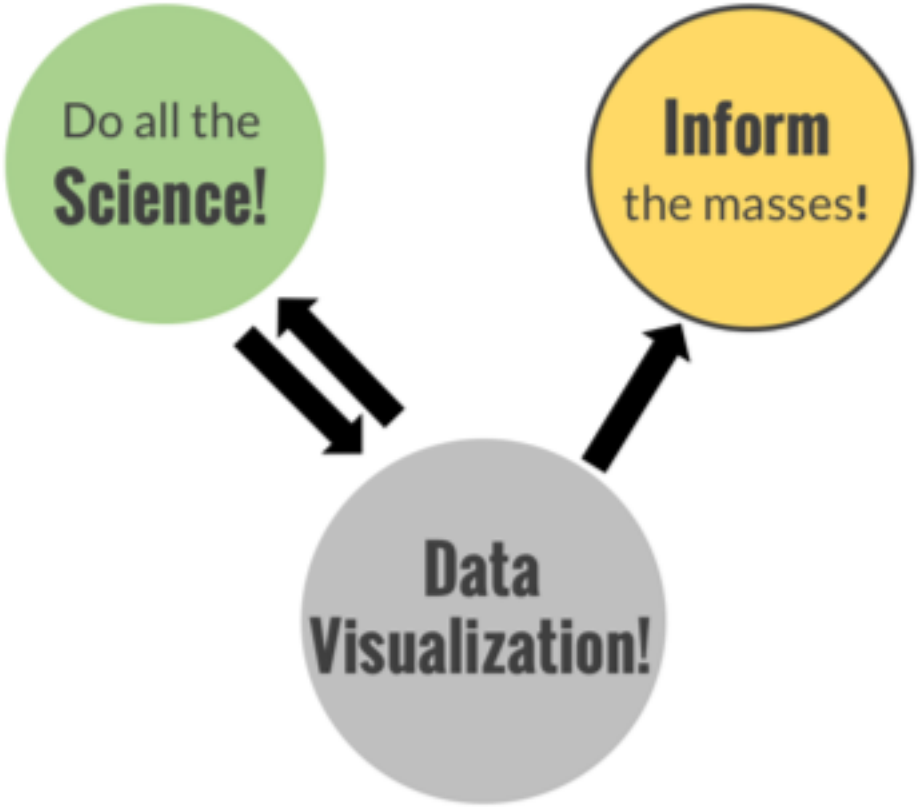


Data visualization in infectious disease GenEpi



Missed opportunity #1 : Getting to know your data

Use datavis to get to know your data!

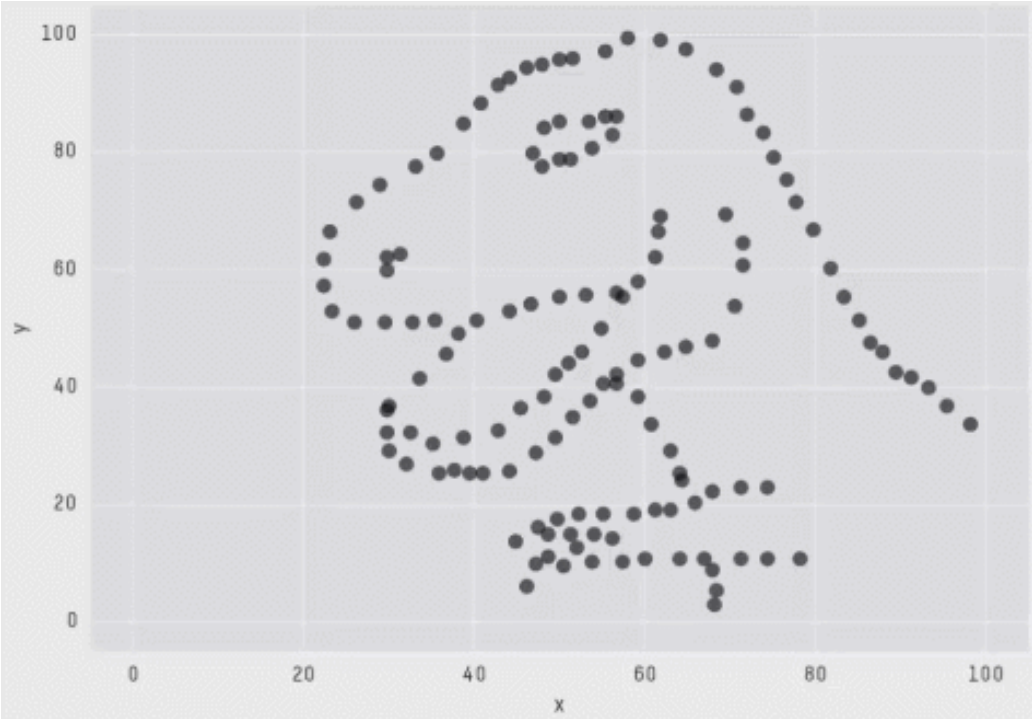


Missed opportunity #1 : Getting to know your data

```
X Mean:  54.2659224
Y Mean:  47.8313999
X SD   :  16.7649829
Y SD   :  26.9342120
Corr.  :  -0.0642526
```

Missed opportunity #1 : Getting to know your data

There could be dinosaurs in your data!



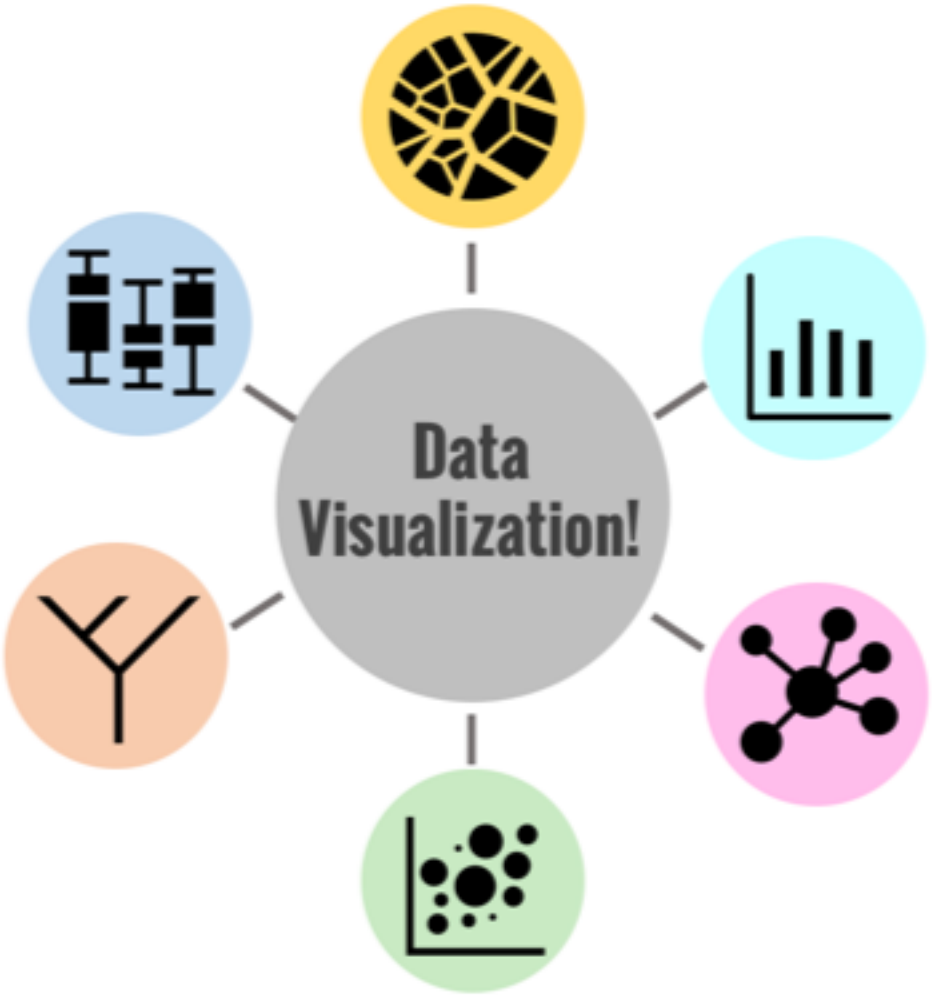
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Autodesk Research (2017). https://www.autodeskresearch.com/publications/samestats

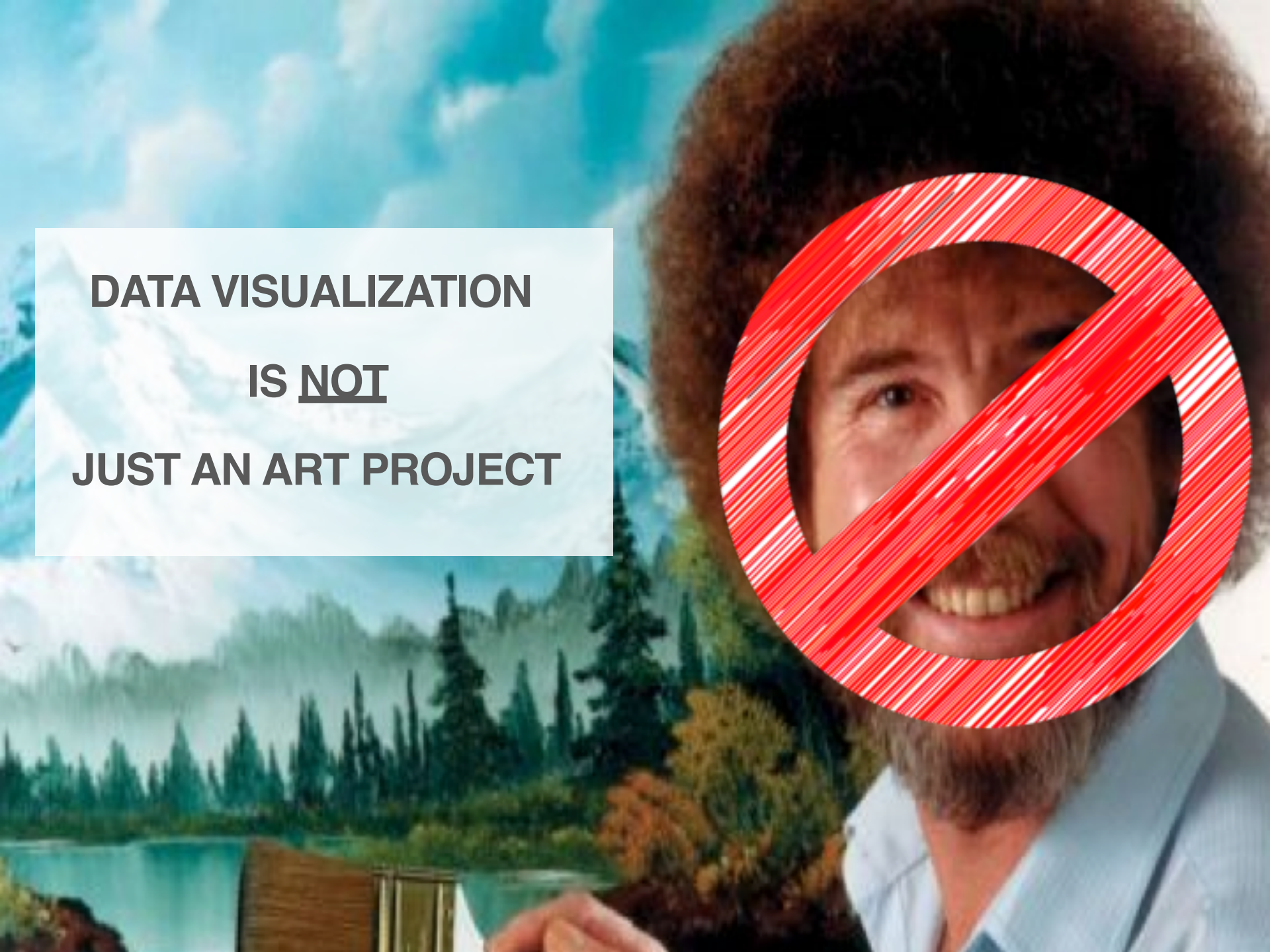
Missed opportunity #2 : Trying different visualizations!

Selecting the appropriate data visualization is challenging!

This is what we're going to be talking about today!



What is data visualization, really?


DATA VISUALIZATION
IS NOT
JUST AN ART PROJECT

How we'll talk about data visualization

There are two aspects of visualizations to think about:

How do you make a visualization?

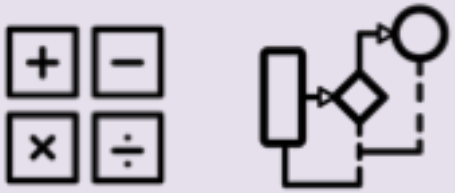
How do you choose the right visualization?

Data Visualization: there's more than meets the eye

Data visualization is not just a graphic design project

Human Perception & Cognition

Computer Graphics

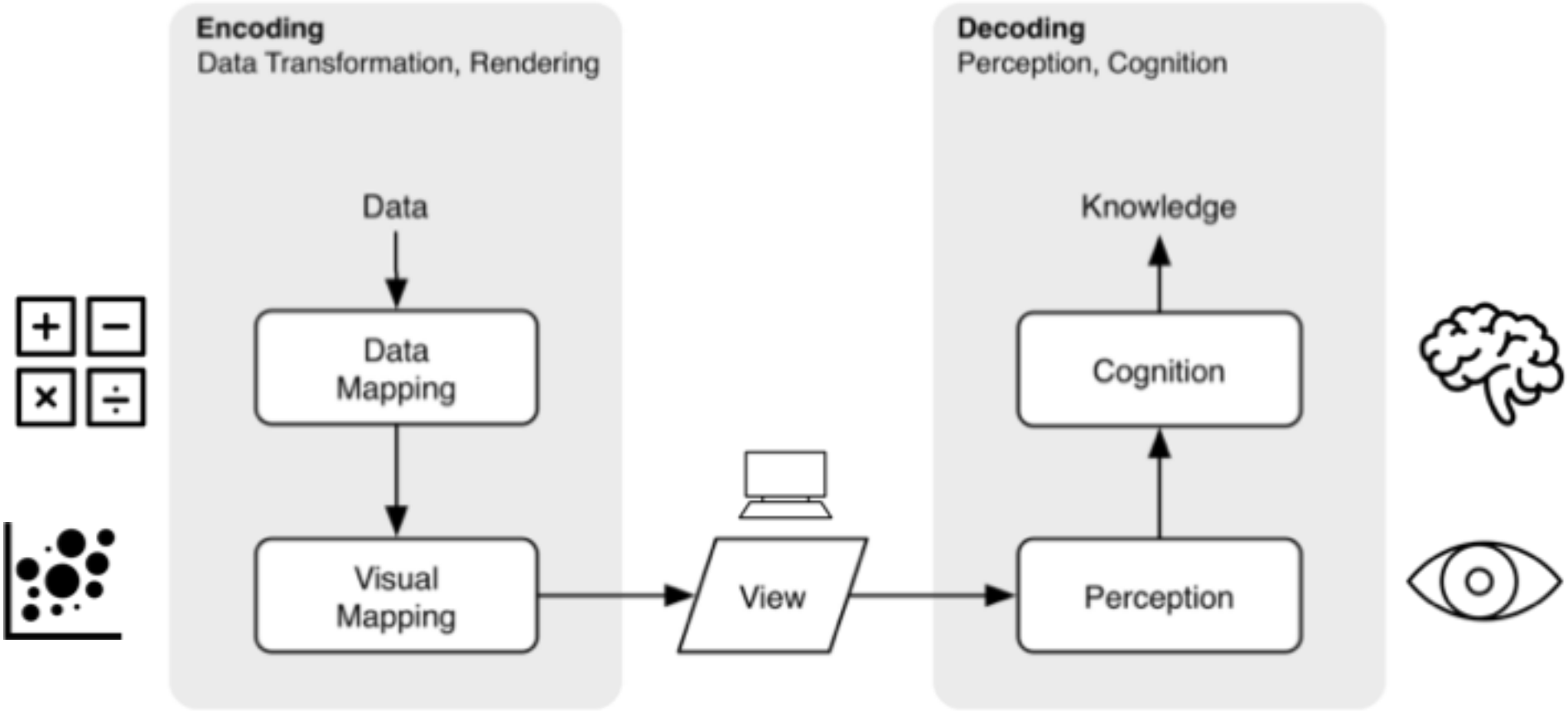


Data Analysis

Visualization Design

Encoding data so others can decode it later!

You should be aware that all these factors are in play



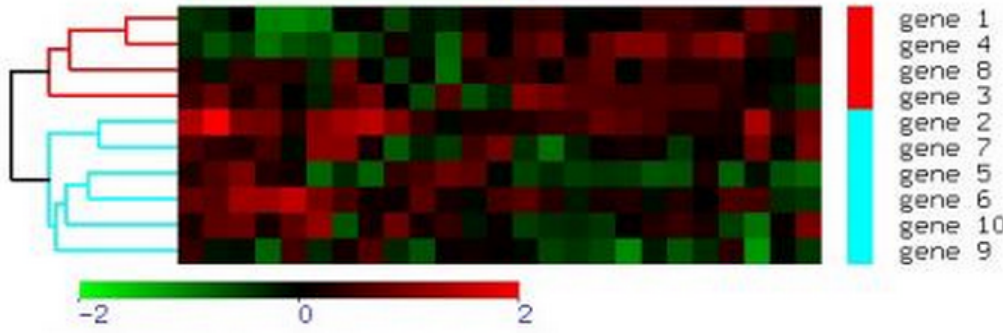
R. Kosara (EagerEyes) – https://eagereyes.org/basics/encoding-vs-decoding

A Small Digression

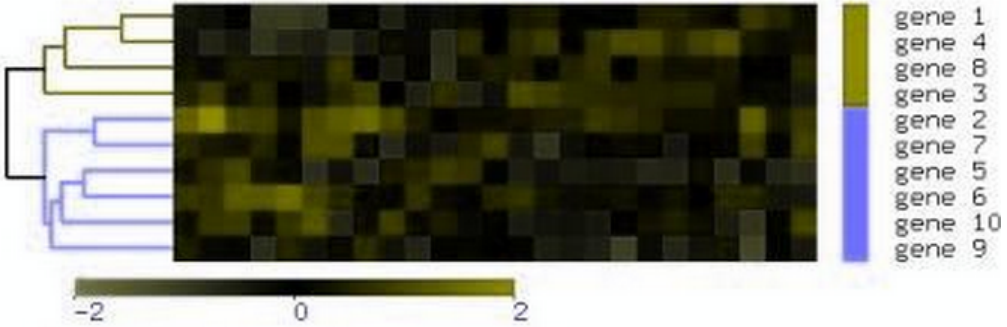
Examples of cognition and perception in practice

Example 1: A Heat map

Non-colour blind individual



Colour blind individual



Example 2: The Dress

Colour Blind Simulator: http://www.color-blindness.com/coblis-color-blindness-simulator/

Examples of cognition and perception in practice

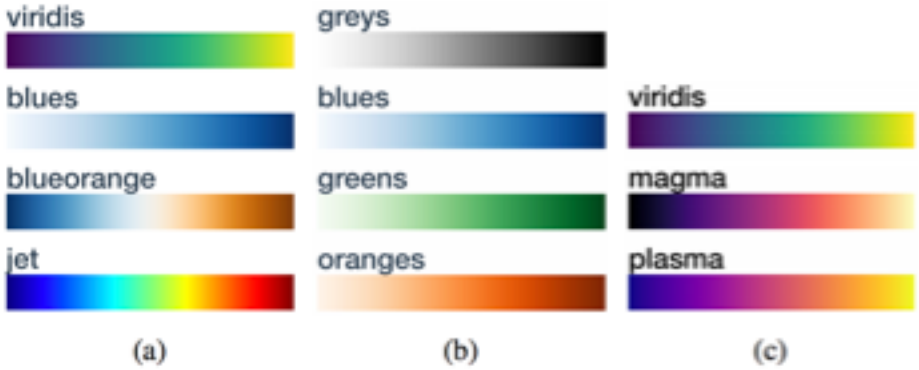


Figure 1: **Colormaps under study.** We evaluate four single-hue, three perceptually-uniform multi-hue, a diverging, and a rainbow colormap(s). We divide them into (a) assorted, (b) single-hue and (c) multi-hue groups, with two colormaps repeated across groups for replication.

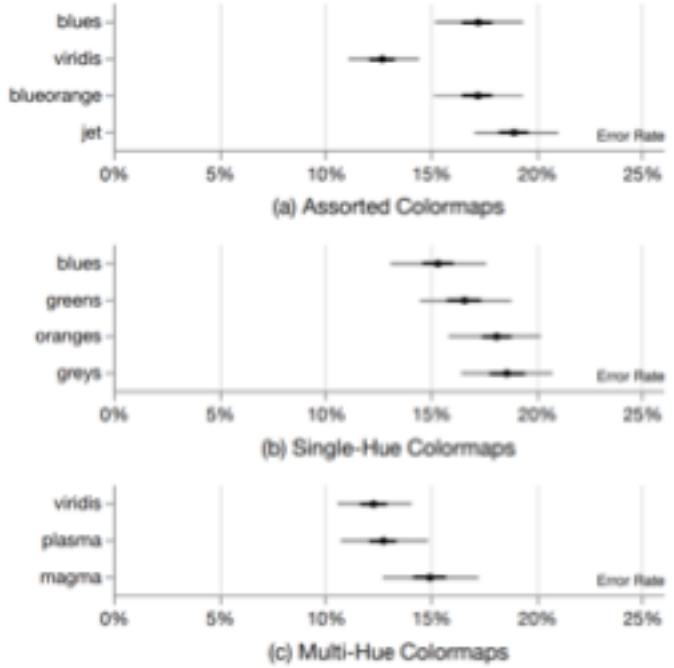
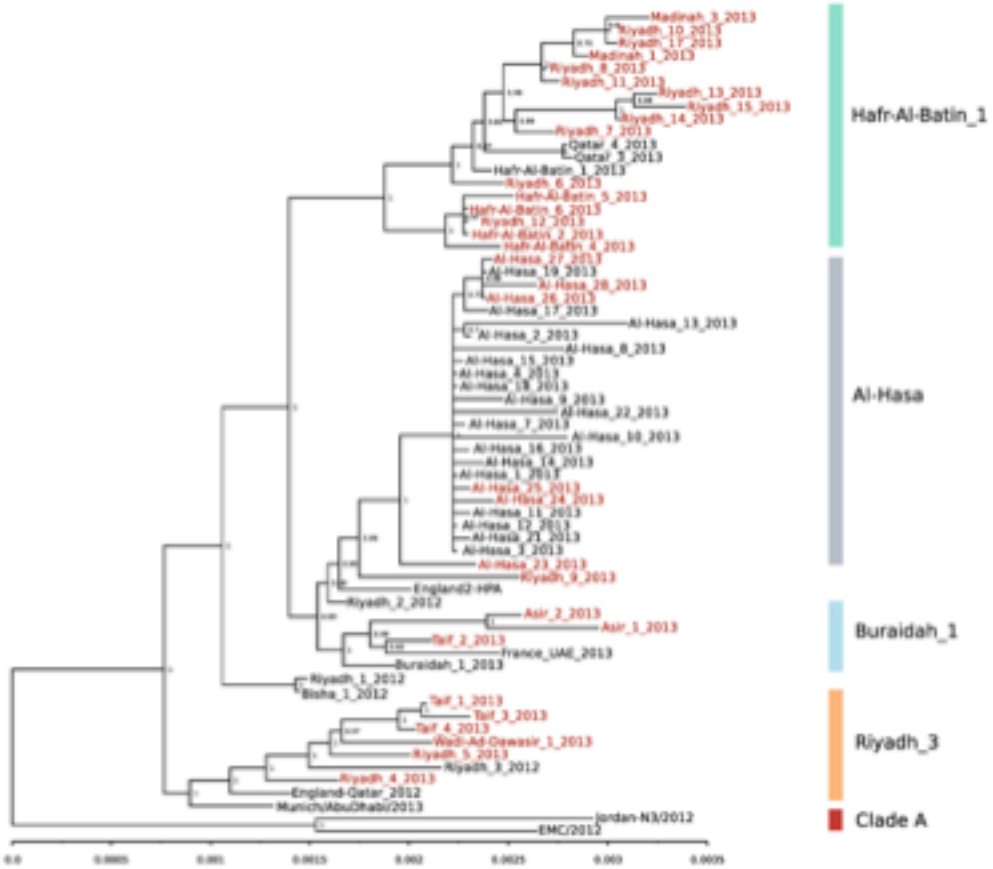


Figure 4: **Error rate by colormap for each study.** Plots depict bootstrapped means, with 50% (thick) and 95% (thin) CIs. (a) Assorted colormaps. *Viridis* excels in accuracy while *jet* is the most error-prone. (b) Single-hue colormaps. Though slightly faster, *blues* and *greens* have overlapping confidence intervals with the slower colormaps, *oranges* and *greys*. (c) Multi-hue colormaps. Multi-hue colormaps have comparable accuracy within group. The per-colormap average error rate of *magma* is higher as it contains degenerate cases.

Liu et al. (2018) - Somewhere Over the Rainbow: An Empirical Assessment of Quantitative Colormaps

And… we're back!

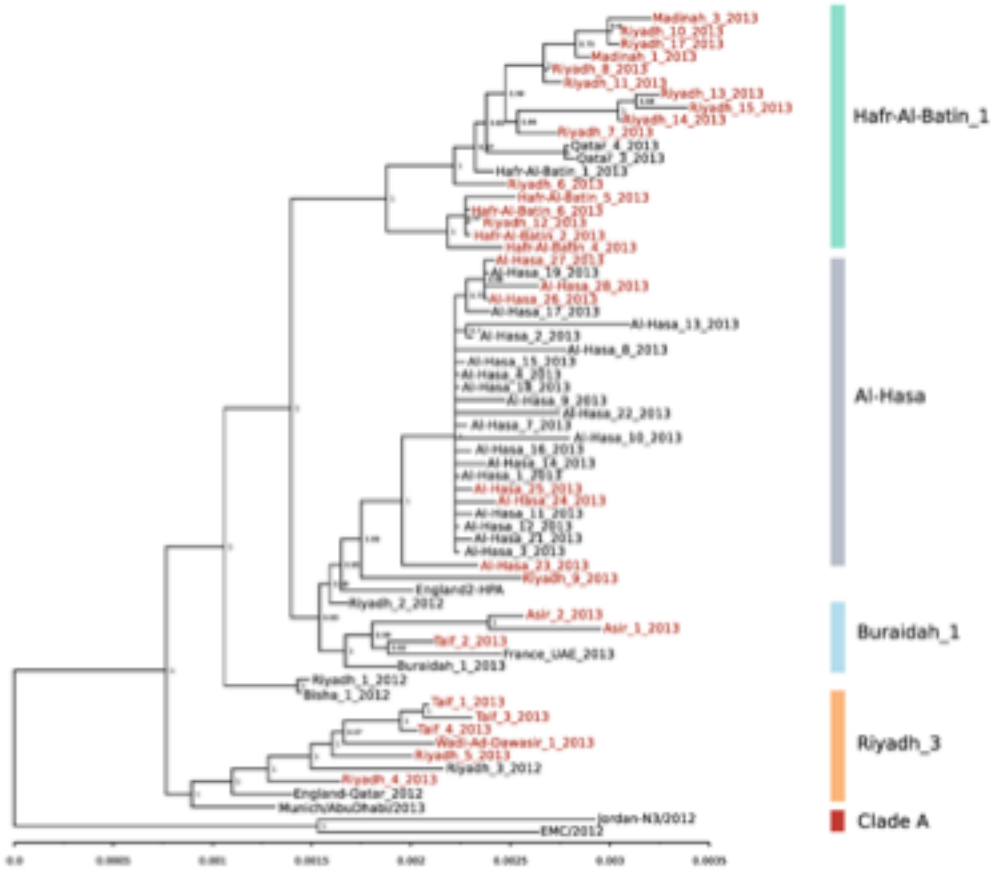
Worked example: encoding and decoding data in IDGE



Let's talk through encoding and decoding data!

Worked example: encoding and decoding data in IDGE

Worked example: encoding and decoding data in IDGE



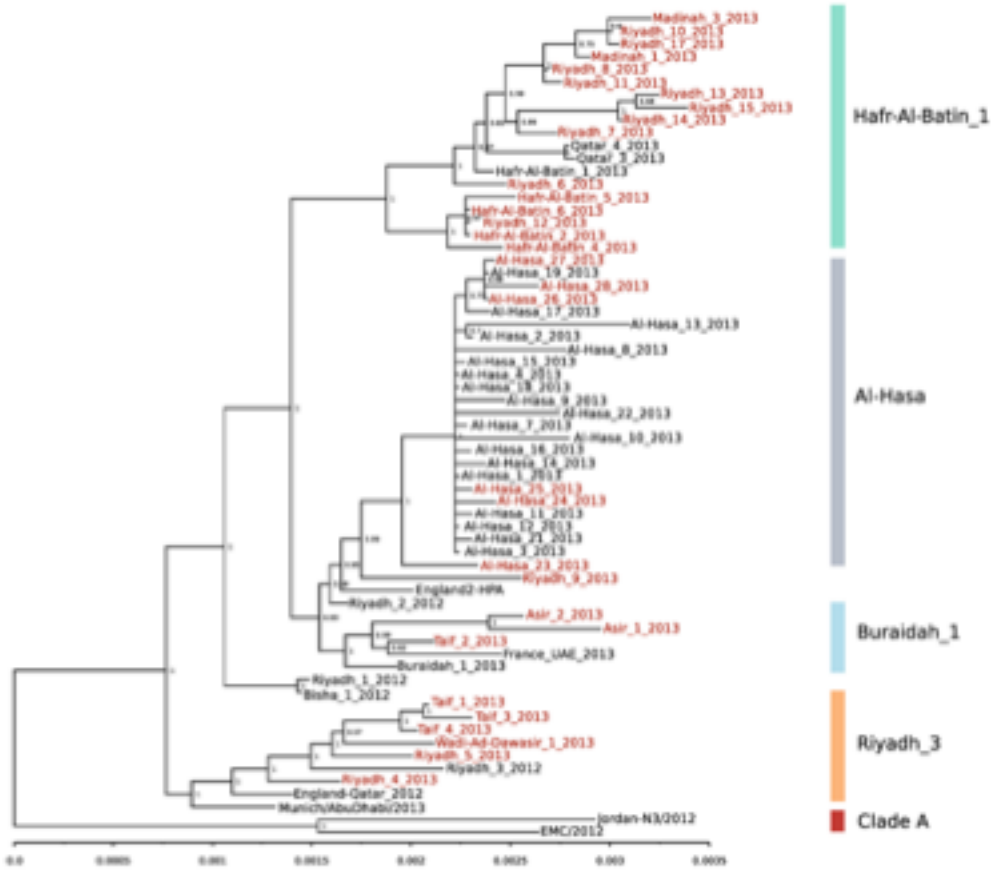
Data

- Individual Cases
  - Location
  - Date
  - Virus clade
- Genomic
  - Sequence data

Data Mapping

- Genomic sequence data to phylogeny
- Phylogeny to clades

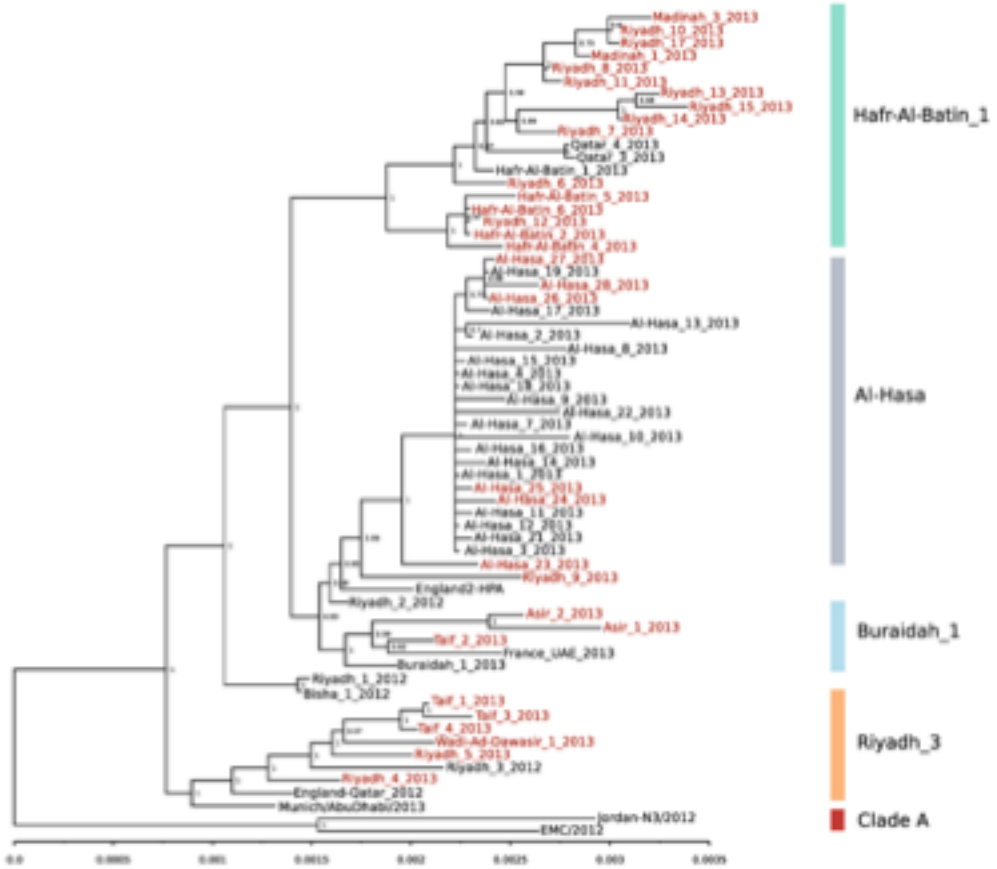
Worked example: encoding and decoding data in IDGE



Visual Mapping:

- Visualization is a phylogenetic tree
- Each case is **a leaf node**
  - New data = red text
- Text shows timing of case infection
- Text shows case city
- Each clade is marked by a thick, colour coded line

Worked example: encoding and decoding data in IDGE



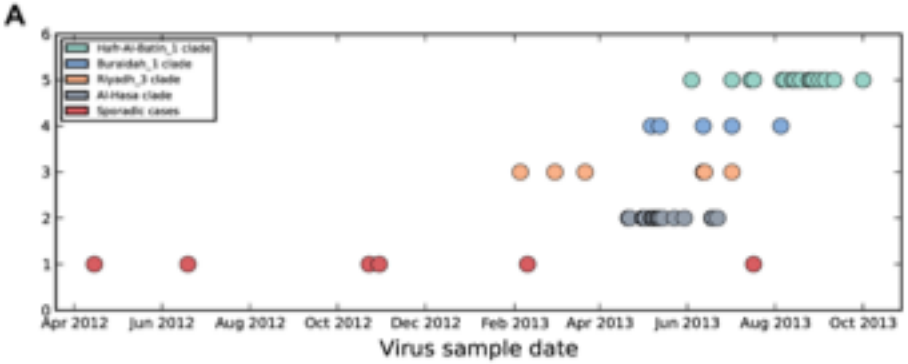
Perception & Cognition:

- Easy to see colours
- Can understand "relatedness" between samples
- *HARD* to understand location AND time
- *High cognitive effort to read text!*

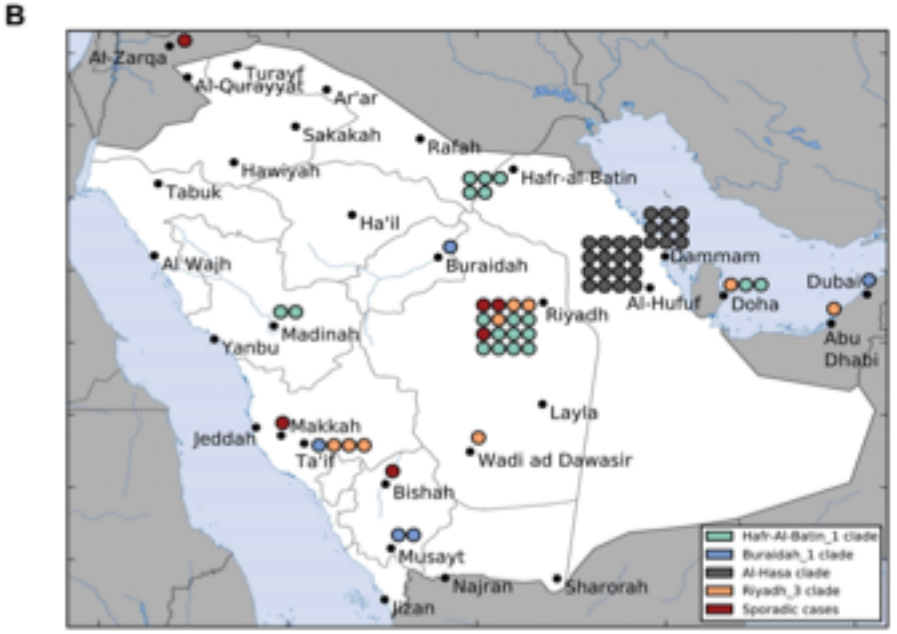
Can we do better?

(actually, the authors did in their own paper)

Worked example: encoding and decoding data in IDGE

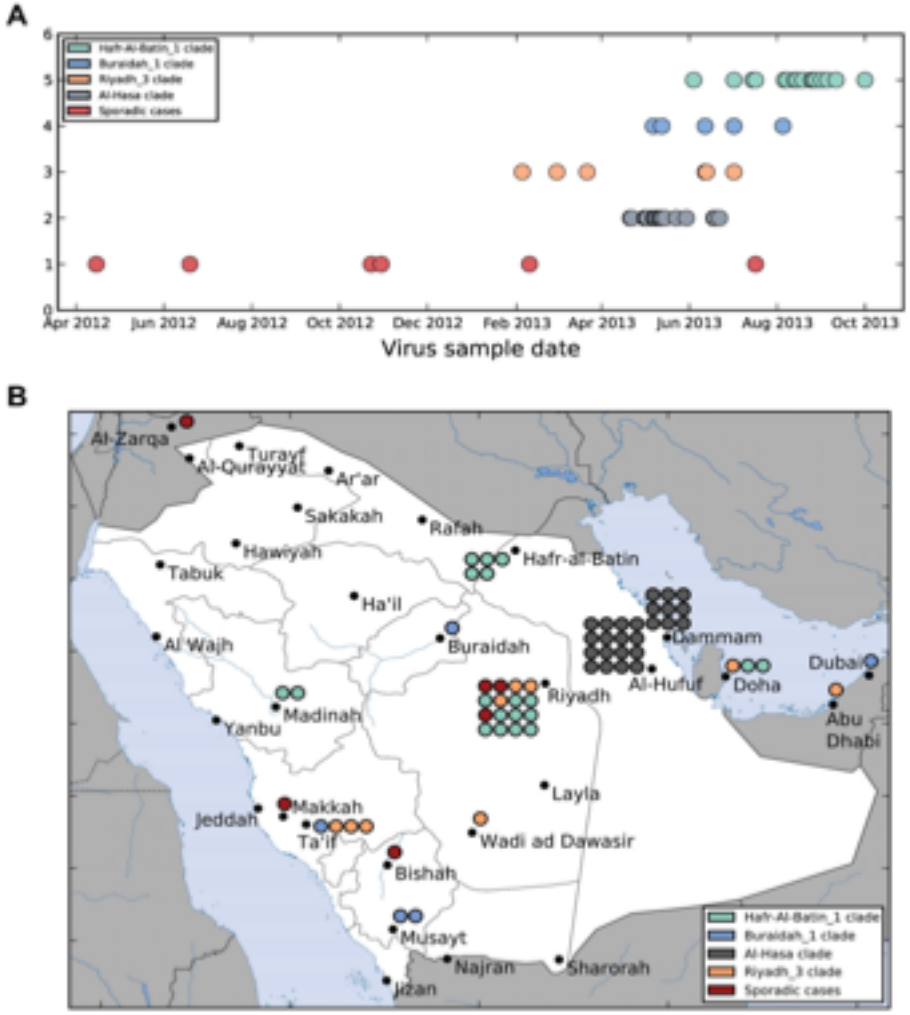


Timeline



Geographic Map

Worked example: encoding and decoding data in IDGE



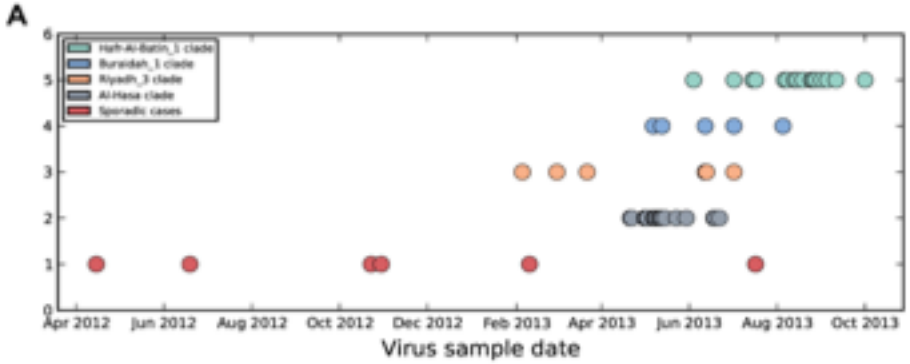
Data:

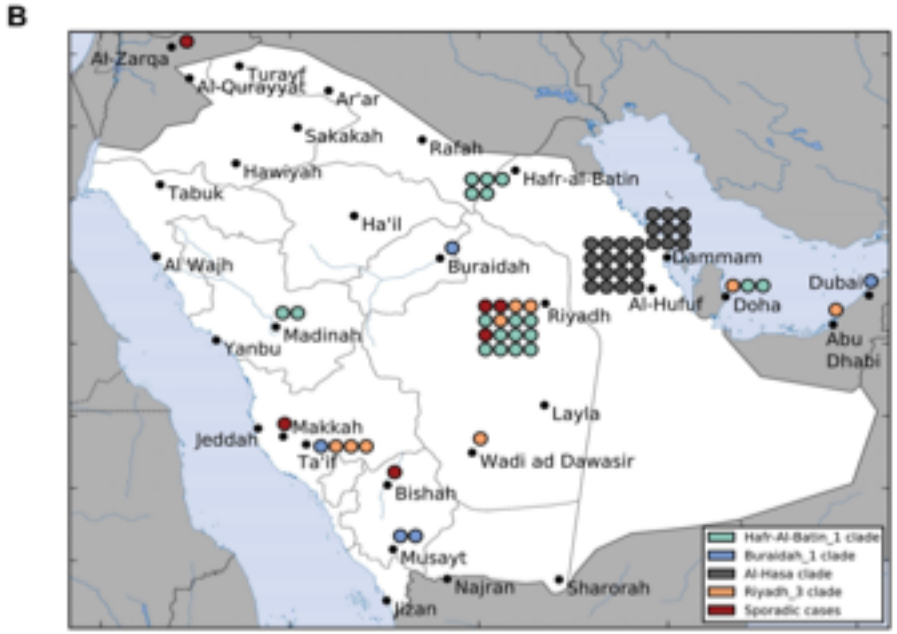
- Individual Cases
  - Location
  - Date
  - Virus clade
- Geographic Context
  - Cities (lat, long)
  - Geographic boundaries

Data Mapping

- Genomic data to phylogeny
- Phylogeny to clades

Worked example: encoding and decoding data in IDGE

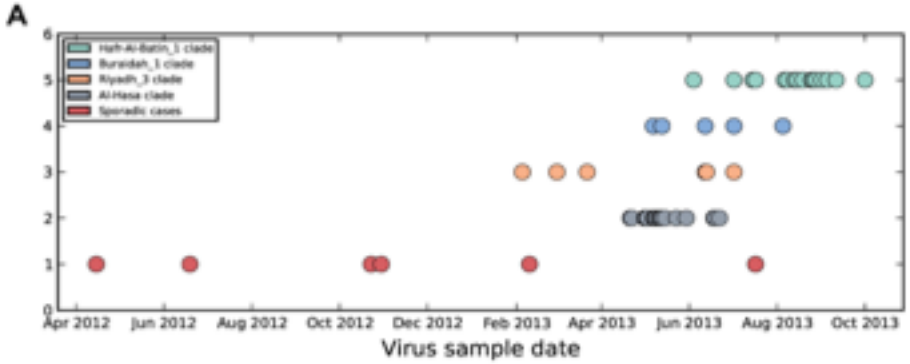


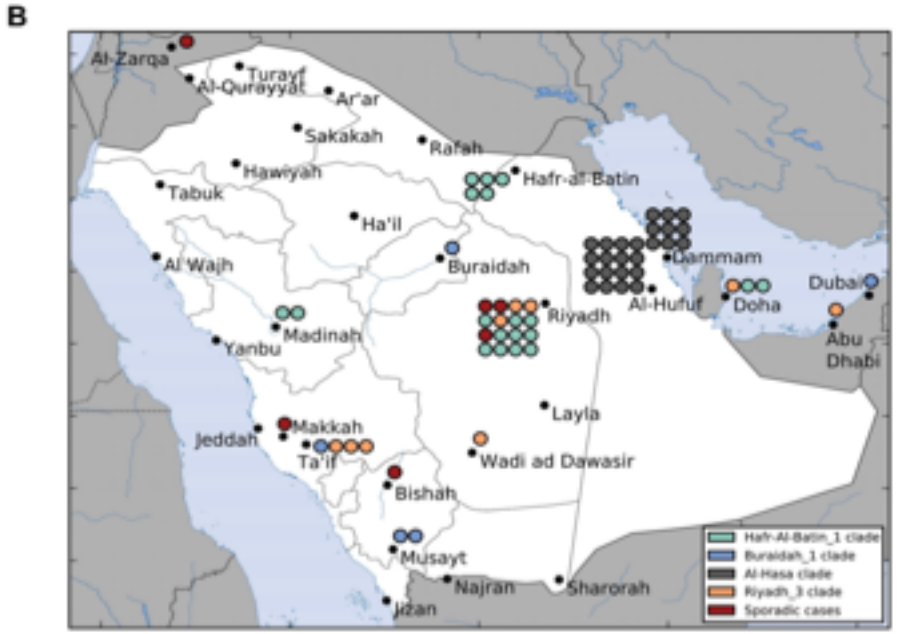


Visual Mapping:

- Each case is **a point**
  - Colour points by clade
- Show case on timeline
- Show case on geographic map
- Each city is a **point** accompanied by text
  - *Warning! City points vs case points!*

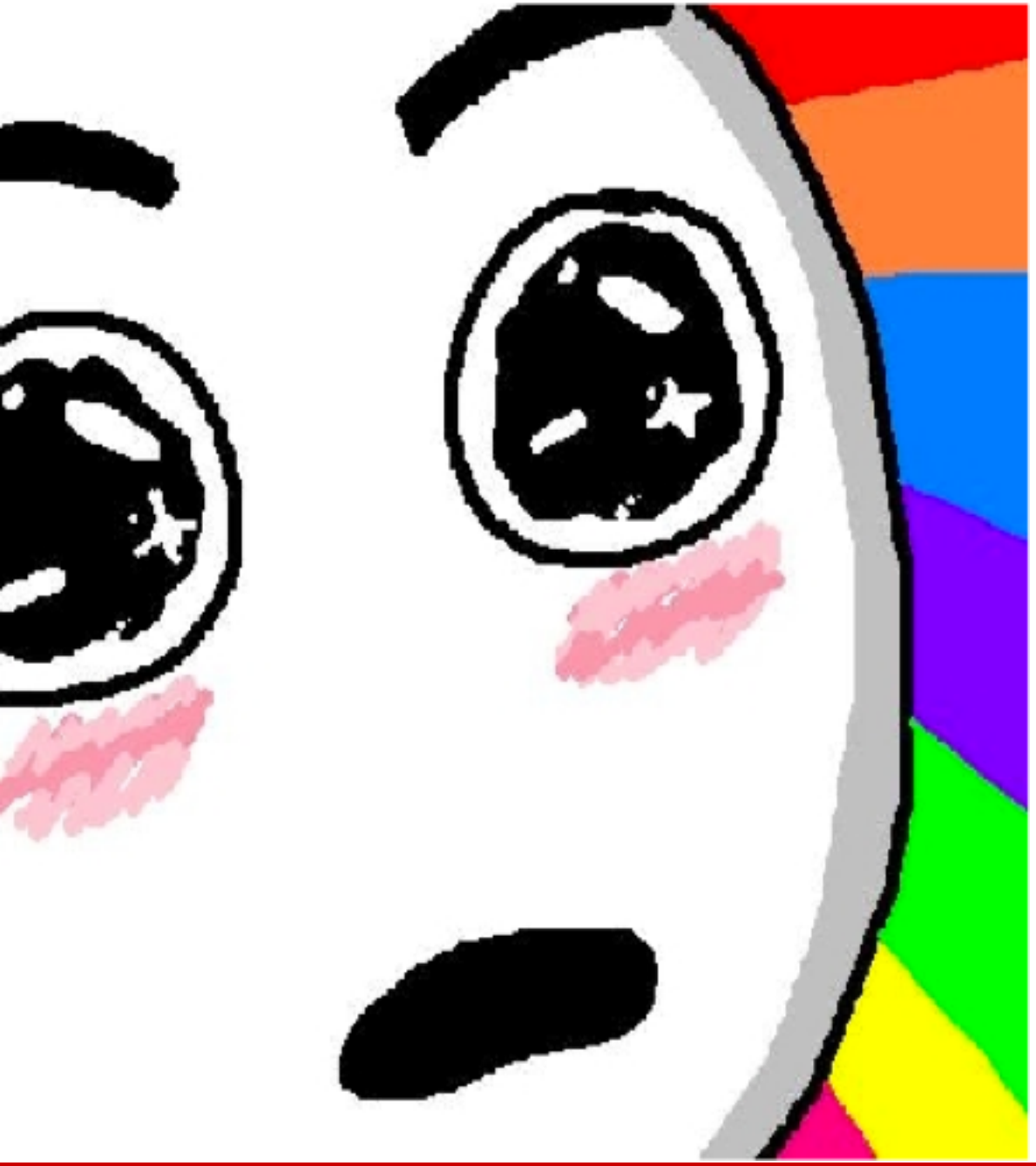
Worked example: encoding and decoding data in IDGE





Perception & Cognition

- Can see time, geography, and location!
- Can't see the exact phylogenetic relationships (there are still clades!)
- *Lower cognitive* effort! No reading necessary!



We did better!

Can we do this good
consistently?

How should we visualize data ?

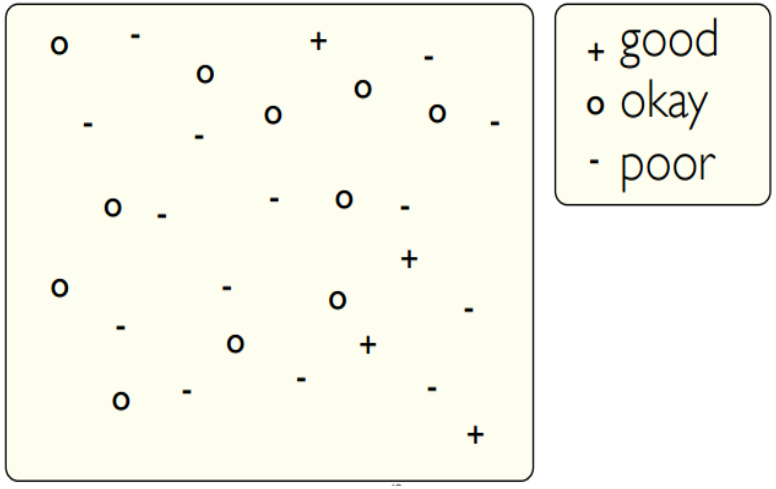
Notes on what follows in this section:

- The content in this section is to help you *make* and *appraise* data visualizations
  - Use the concepts here to talk to you friends about data vis!
  - You don't need to become an expert in all the things described here, but you should be aware of them
- I'm transplanting content from infovis (a field of study in computer science)
  - Infovis is a young and evolving field of study
  - I've summarized that I think it most useful for you to know

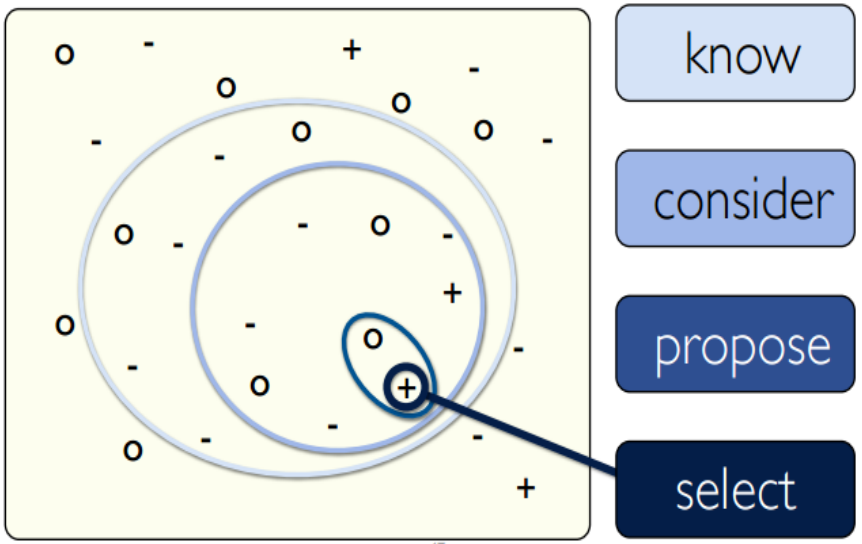
The BIG picture – thinking about design spaces

- Design spaces are made of visualization design choices
- Design choices have varying utility (+ 0 -)

A design space



Searching through a design space

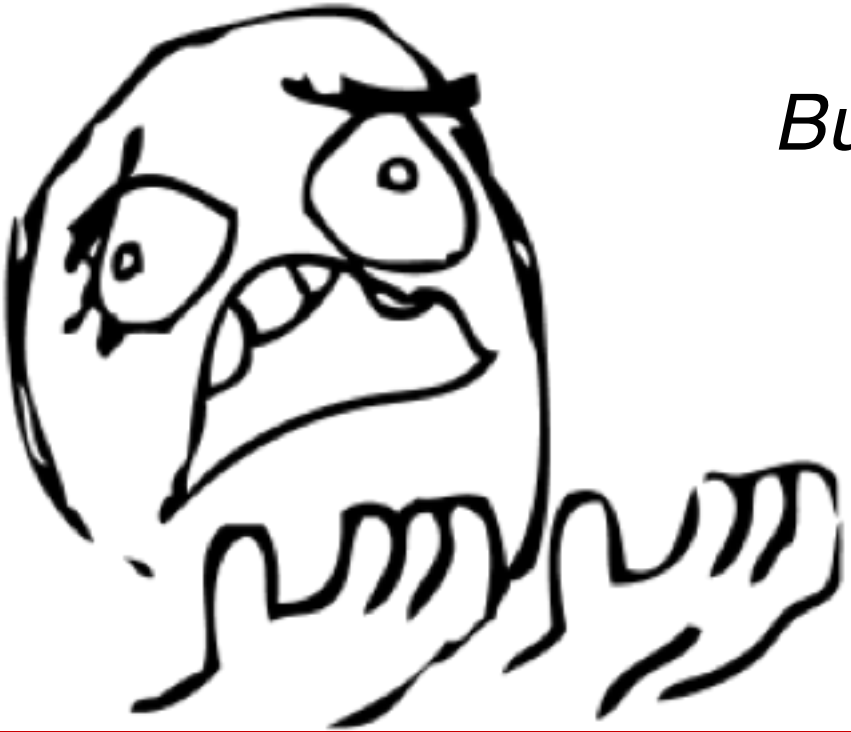


You actually already think about design spaces!

- All of the chairs below have different designs
- All chairs can be used for a common task : sitting
- But – fundamentally, different chairs are suited for different tasks

Terrible office chairs (-)

Suitable office chairs (+, 0)



But pictures are not chairs!

Give up there's no hope

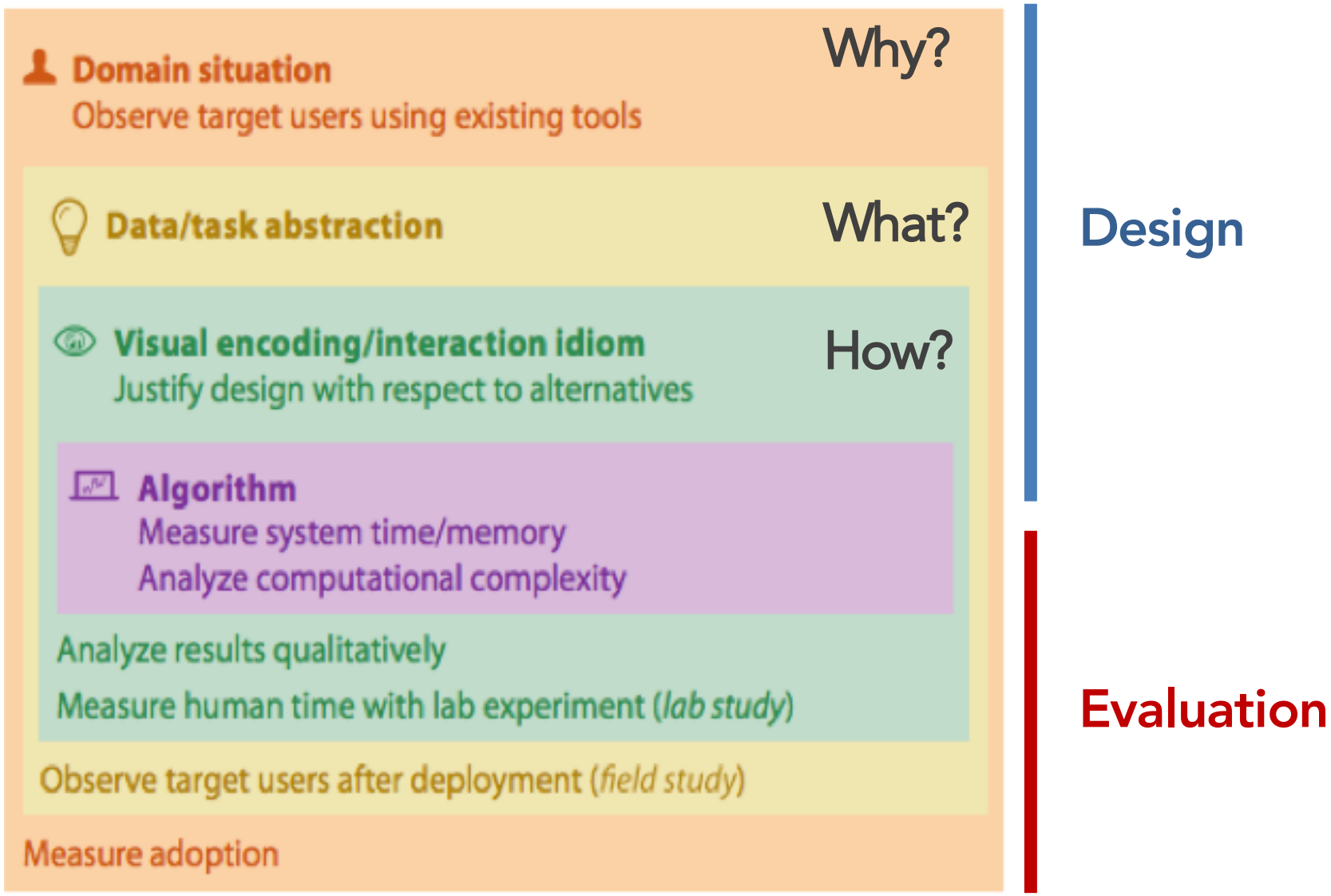
Wait until AI solves the problem

Just come up with something

Think differently and more systematically about data visualization

(if you don't know what this is, it is the 'expanding mind' meme)

Systematic thinking: the layers of a data visualization



T. Munzner (2009) - A nested model for visualization design and validation

Systematic thinking: the layers of a data visualization

Why? (Motivation)

Why do you need to visualize data?

How will you, or others, use the visualization?

Systematic thinking: the layers of a data visualization

Why? (Motivation)

Why do you need to visualize data?

How will you, or others, use the visualization?

What? (Data & Tasks)

What kind of data is being visualized?

What tasks are performed with the data?

Systematic thinking: the layers of a data visualization

Why? (Motivation)

Why do you need to visualize data?

How will you, or others, use the visualization?

What? (Data & Tasks)

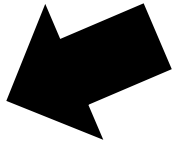
What kind of data is being visualized?

What tasks are performed with the data?

How? (Visual & Interactive Design)

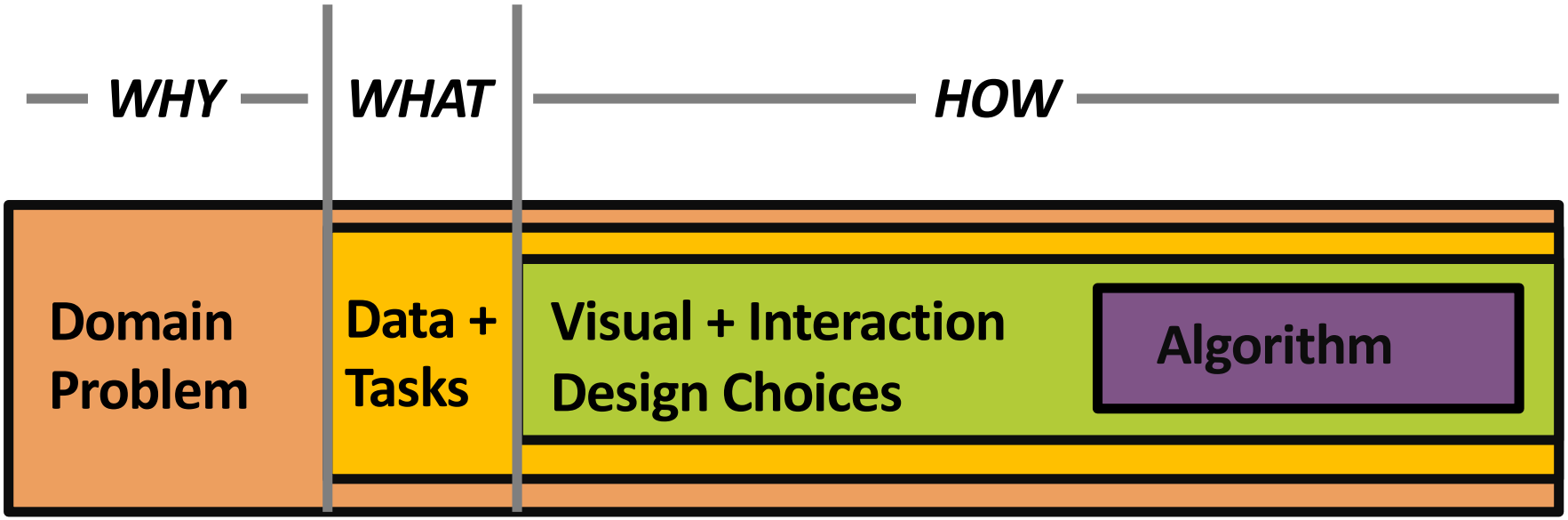
How do you make the visualization?

Is it the right visualization?



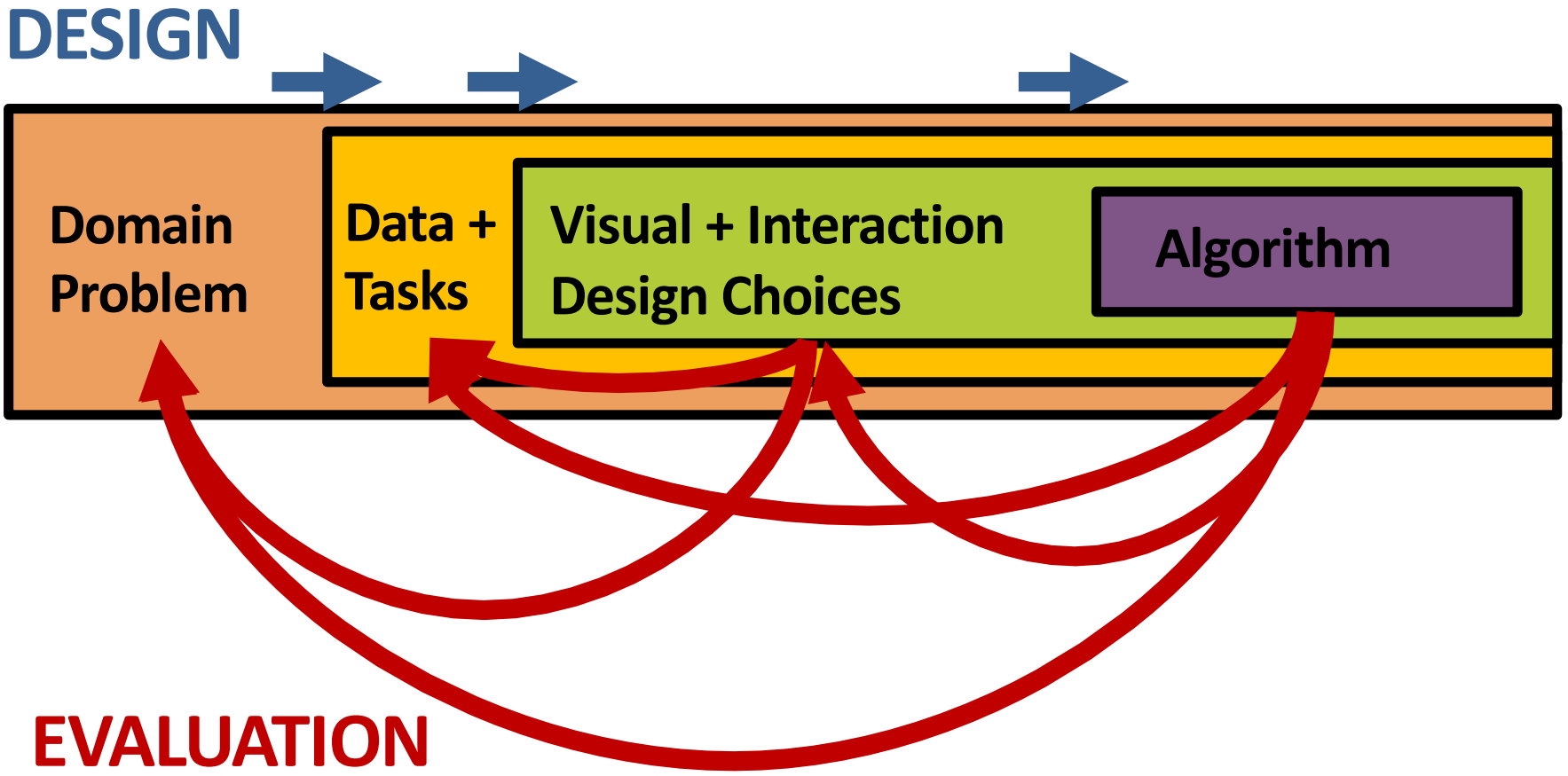
People tend to jump to this level and ignore *why* and *what*

Systematic thinking in *action*

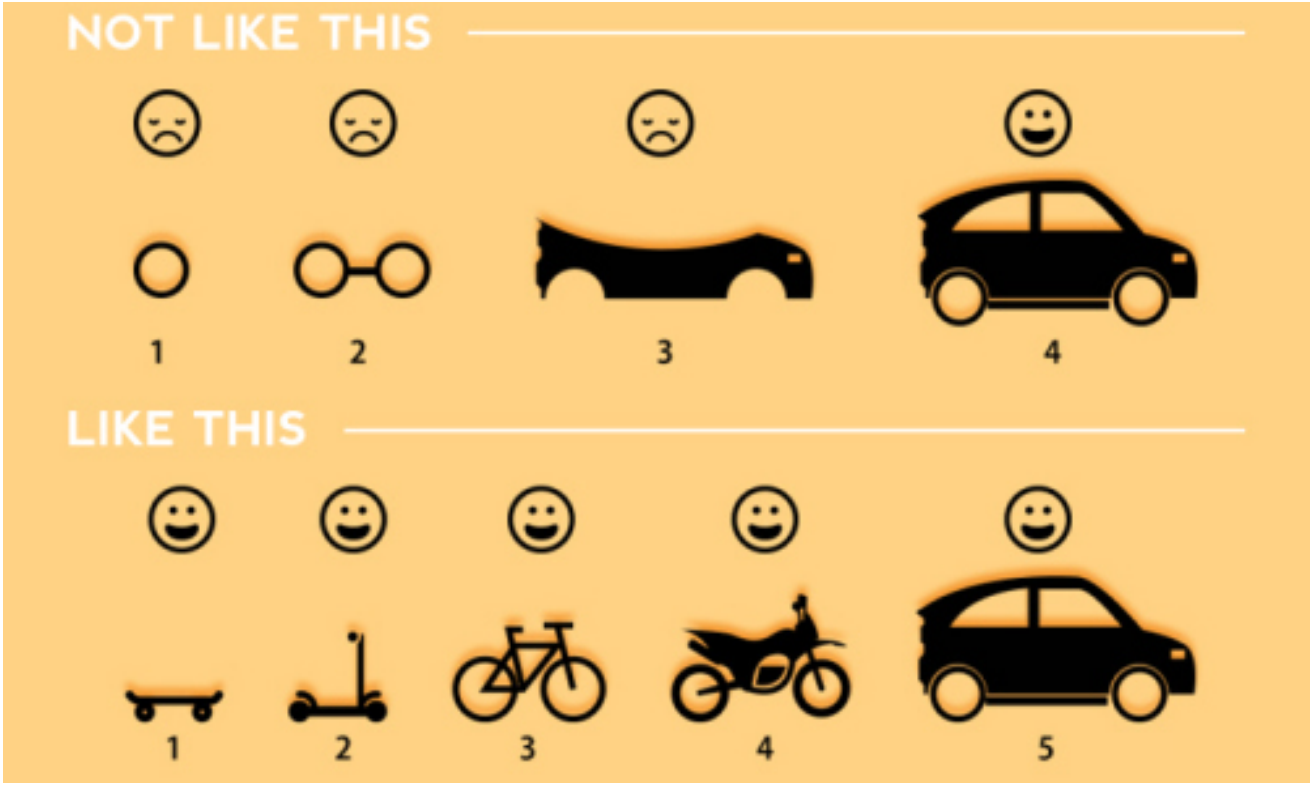


T. Munzner (2009) - A nested model for visualization design and validation

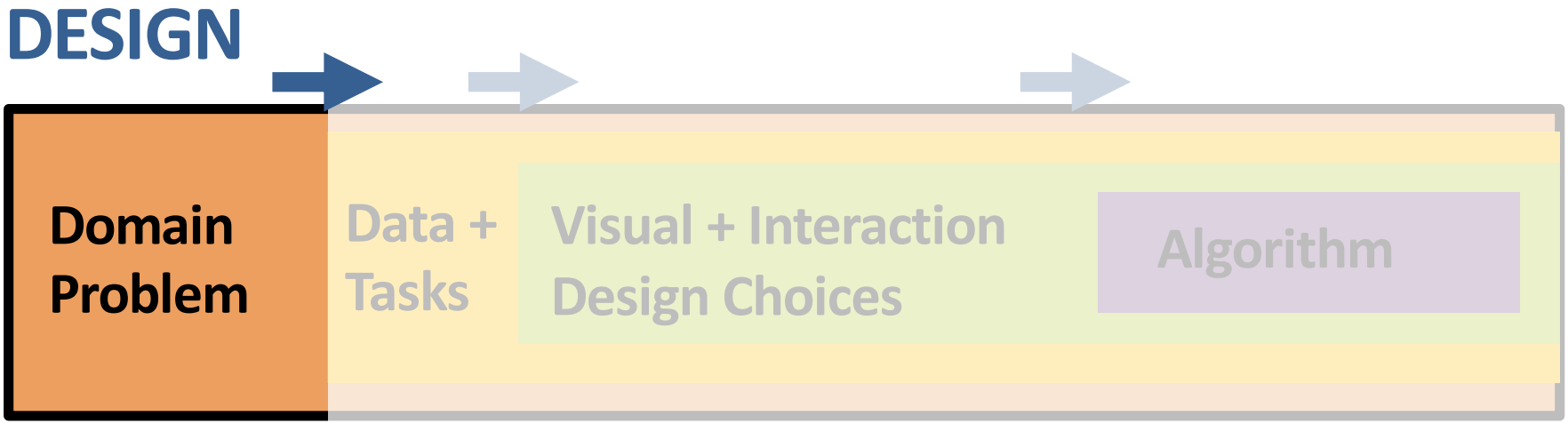
Use an **iterative** process



Use an **iterative** process



An iterative approach to development allows us to get feedback before committing to ineffective design choices



1. **Why** is data visualization needed? What problem is data visualization solving? For whom?

Stakeholders and their different data needs

- **Multidisciplinary decision making teams**
  - More data & diverse data types = more informed decision making
  - BUT – different stakeholder abilities to interpret data & different needs

Medical Health Officers

Clinicians

Nurses

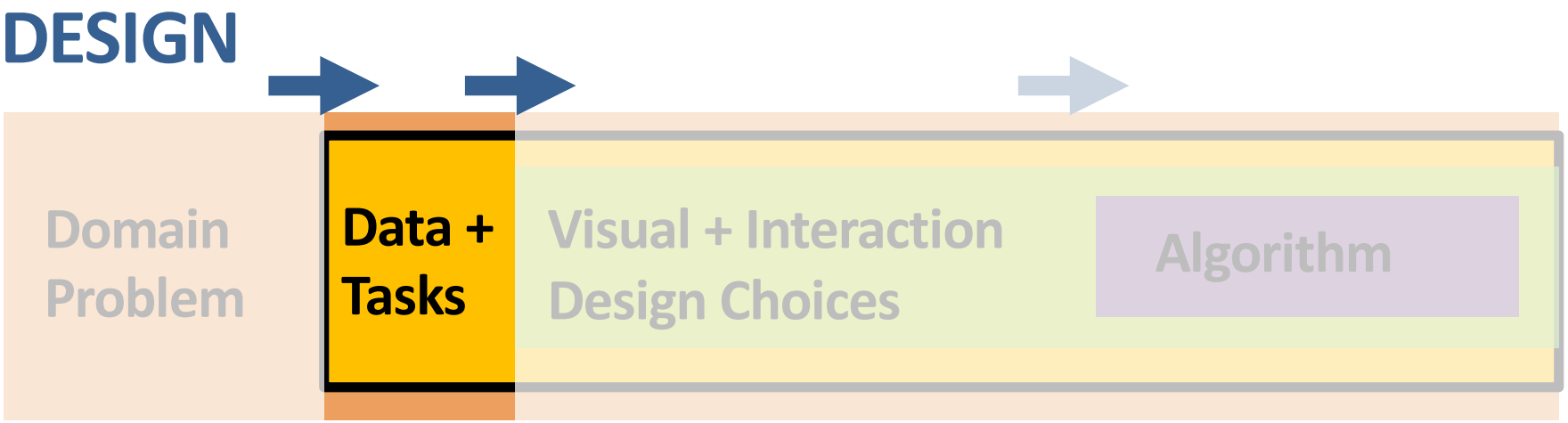
Researchers

Community Leaders

Patients

Politicians

Systematic thinking in *action*



2. **What** data should be visualized (is it available?)
3. **What** is the data used for ? [**tasks**]

M. Sedlmair et al. (2012) - A design study methodology : lessons from the trenches

An example of data and tasks

Data

	WGS equivalent	DIAGNOSIS TASKS				TREATMENT TASKS			SURVEILLENCE TASKS					TOTAL SCORE
		Diagnose Latent TB	Diagnose Active TB	Reactive vs New Acuqistion	Characterize Transmission Risk	Choose Meds	Choose Tx Duration	Assess Response to Tx	Guide Contact Tracing	Report to Public Health	Define a Cluster	Connect case to Existing Cluster	Guide Public Health Response	
Patient Identifier	Same	3	3	3	3	3	3	3	2	1	1	1	1	26
Sample Collection Date	Same	3	3	2	3	3	3	3	1	1	1	1	1	24
Patient Prior TB Results	Same	3	2	3	3	3	3	3	1	1	1	0	1	23
Speciation	Speciation	1	3	2	3	3	3	3	2	1	1	1	1	23
Sample Type (sputum, fine needle aspirate)	Same	2	3	2	3	3	3	3	1	1	1	0	1	22
Culture results	WGS data	1	3	2	3	3	3	3	2	1	1	0	1	22
Sample Collection Site (lymph node, blood draw etc.)	Same	2	3	2	3	3	3	3	1	1	0	0	1	21
Acid Fast Bacilli Smear	Speciation	2	3	2	3	2	3	3	1	1	1	0	1	21
Resistotype	Predicted DST	0	2	3	1	3	3	2	2	1	1	1	1	19
Phenotype DST	Predicted DST*	0	2	3	2	3	3	2	1	1	1	0	1	18
Chest x-ray	NA	3	3	2	3	0	2	3	1	0	0	0	0	17
Report Releate Date	Same	2	2	1	2	2	2	2	1	0	1	0	1	15
Requester IDs	Same	2	2	2	2	2	2	2	1	0	0	0	0	15
Interpretation or comments from reviewer	Same	2	2	1	2	2	2	3	1	0	0	0	0	15
Predicted DST	Predicted DST	0	2	2	1	3	3	2	1	0	1	0	0	15
MIRU-VNTR	SNPs	0	2	3	1	1	1	1	1	1	1	1	1	13
Cluster Assignment	Cluster Assignemnt	0	2	2	1	1	1	0	1	1	1	1	1	11
SNP/variant disance	SNPs	0	1	2	1	1	1	0	1	1	1	1	1	10
Phylogenetic Tree	Phylogenetic Tree	0	2	1	1	1	1	0	1	0	1	1	1	9
Reviewer ID	Same	1	1	1	1	1	1	1	1	0	0	0	0	8
TST results	Speciation**	3	1	1	1	0	0	0	1	0	0	0	0	7
IGRA results	Speciation**	3	1	1	1	0	0	0	1	0	0	0	0	7
Lab QC	WGS Speciffic	0	1	2	1	1	1	0	1	0	0	0	0	7
Spoligotype	SNPs	0	1	1	1	0	0	0	0	0	0	0	0	3
RFLP	SNPs	0	1	1	1	0	0	0	0	0	0	0	0	3

Degree of consensus: 3 (High) | 2 (Some) | 1 (Low) | 0 (V. L ow)

A Crisan (2018) – Evidence Base Design and Analysis of a whole genome sequence clinical report....

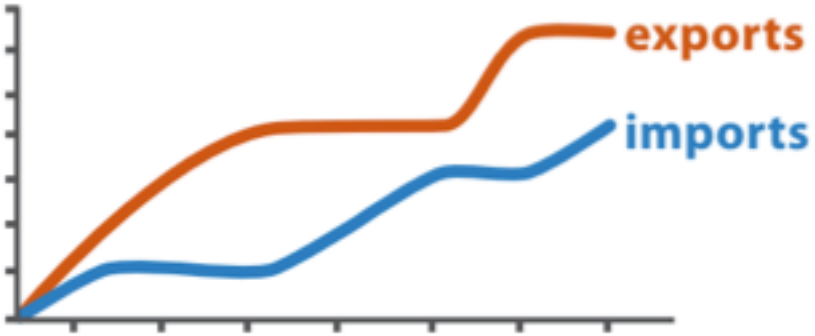
Some extra thoughts about data

Note! Don't just use *raw* data, *derive* new data

(you actually should already know this, since phylogenies are derived from raw genomic data)

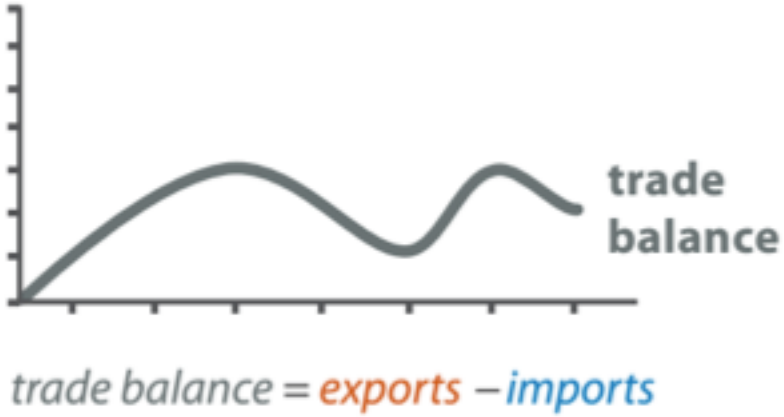
Raw data

Data exactly as it is in the spreadsheet



Derived data

Performing a calculation on the data



T. Munzner (2014) – Visualization Design and Analysis

Some extra thoughts about data

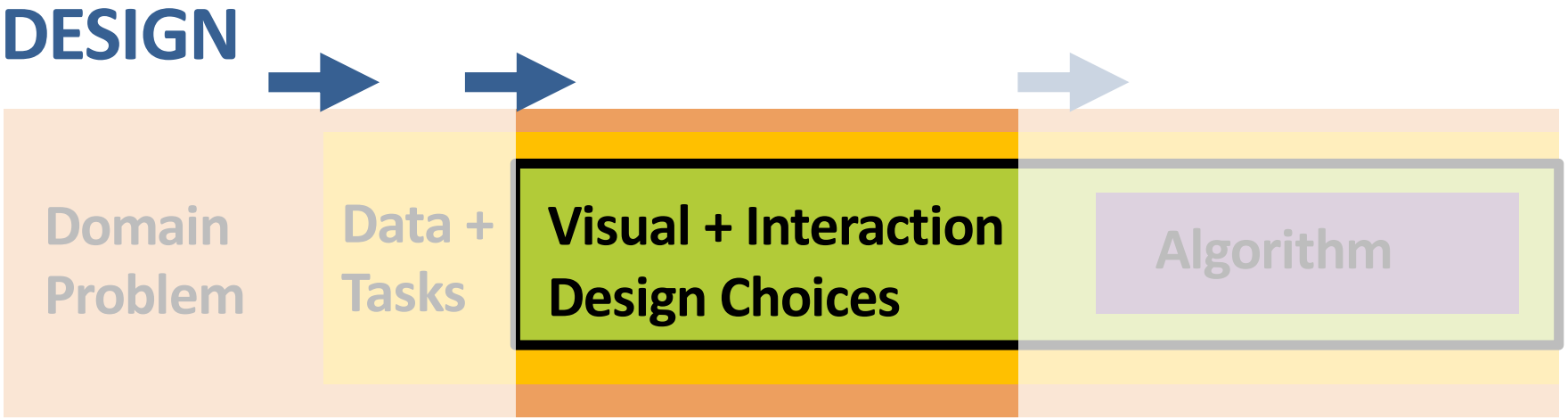
Don't just *let the data drive,* it could be **wrong**

Example:

Using (raw) absolute counts, when you actually need to derive rates

https://xkcd.com/1138/

Systematic thinking in *action*



4. Explore if other visualizations have addressed this **problem** and set of **tasks & data**
5. Implement **your own solution**

M. Sedlmair et al. (2012) - A design study methodology : lessons from the trenches

A Small Digression

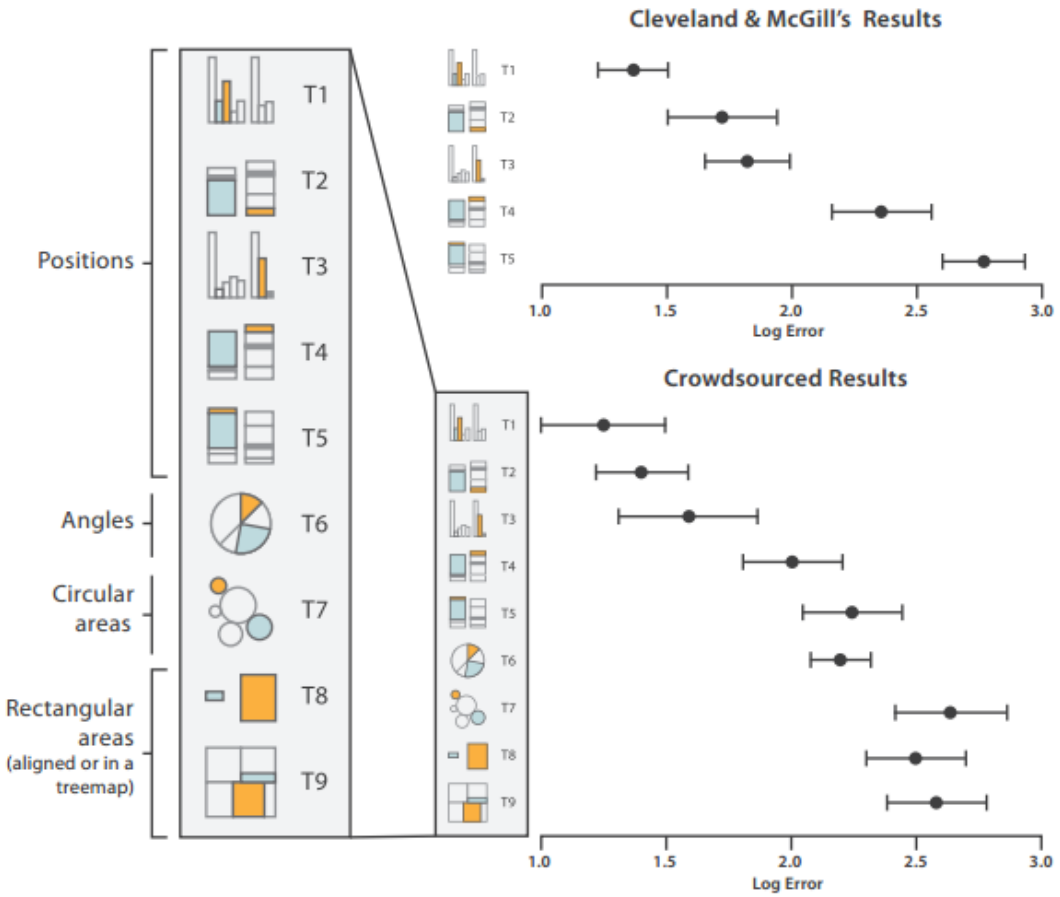
Marks & Channels : building blocks of data visualizations

Mark:
Basic Graphical Element
(basic building block)

Channel:
Controls the
appearance of marks

T. Munzner (2014) – Visualization Design and Analysis

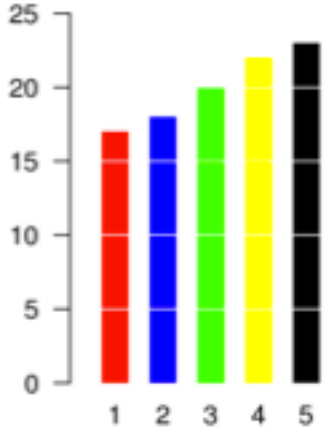
Channels vary in their effectiveness



Example

Pie Chart
Angle & Area

Bar Chart
Position Common Scale



J. Heer (2010) – Crowdsourcing Graphical Perception: Using Mechanical Turk ……

Marks & Channels : ggplot2 example

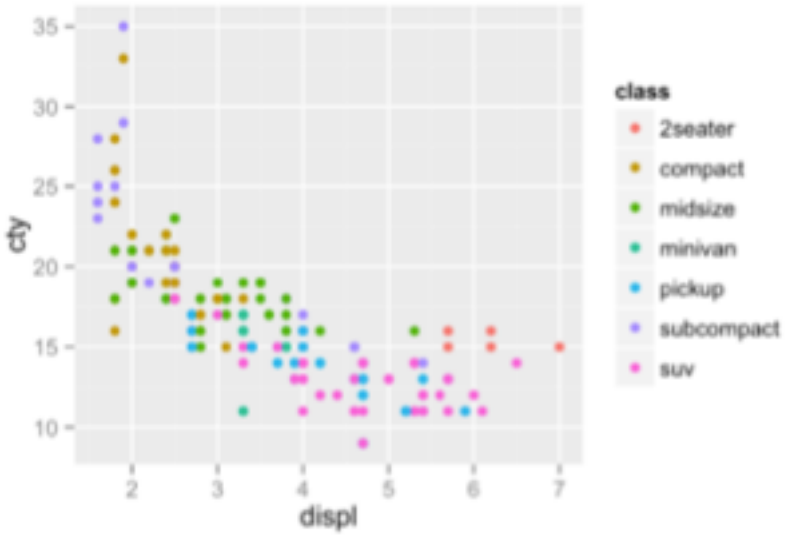
Data is *visually encoded* via marks and channels

Channel: *Position*

Channel: *Colour*

```
ggplot (data = mpg, aes( x= display, y = cty, colour = class)) +
        geom_point( )
```

Mark: *Point*

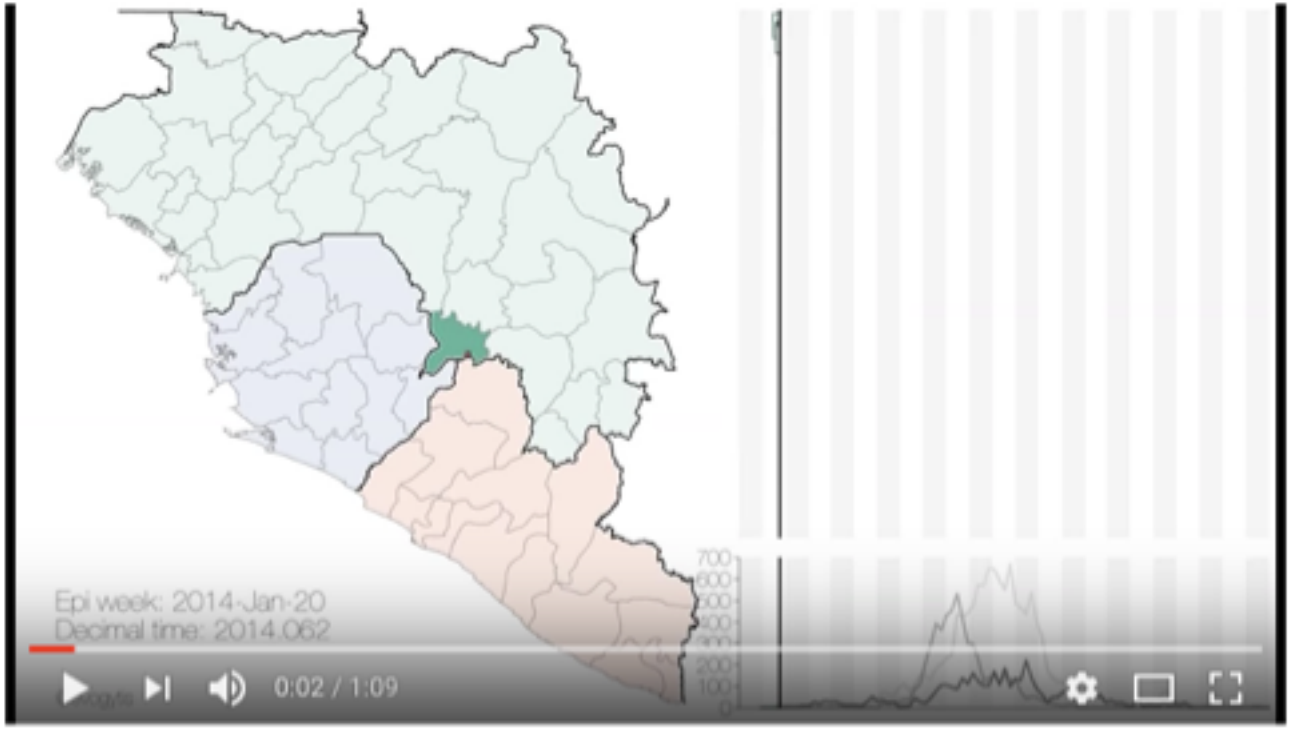


Note: Generally in ggplot2 aesthetics refer to channels and geoms refer to marks, but there are complex geoms that aren't simple marks but chart types (i.e. geom_density) and there are aesthetics that have little to do with the visual channels directly (i.e. group)

https://rpubs.com/hadley/ggplot-intro

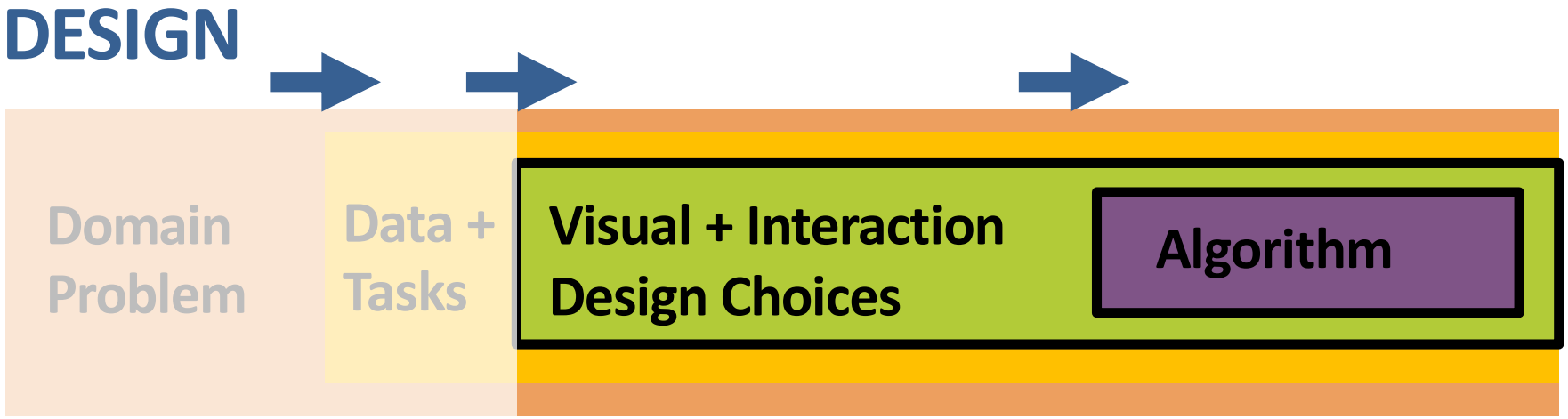
And… we're back!

There are also *interactions!*



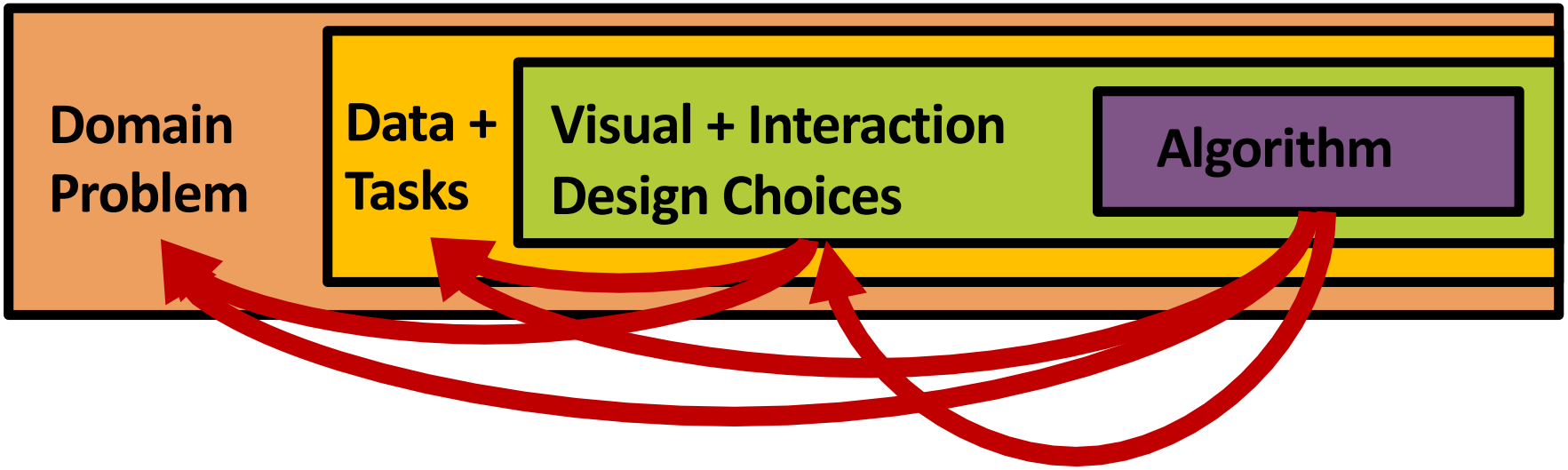
Virus genomes reveal factors that spread and sustained the Ebola epidemic

https://www.youtube.com/watch?v=j4Ut4krp8GQ



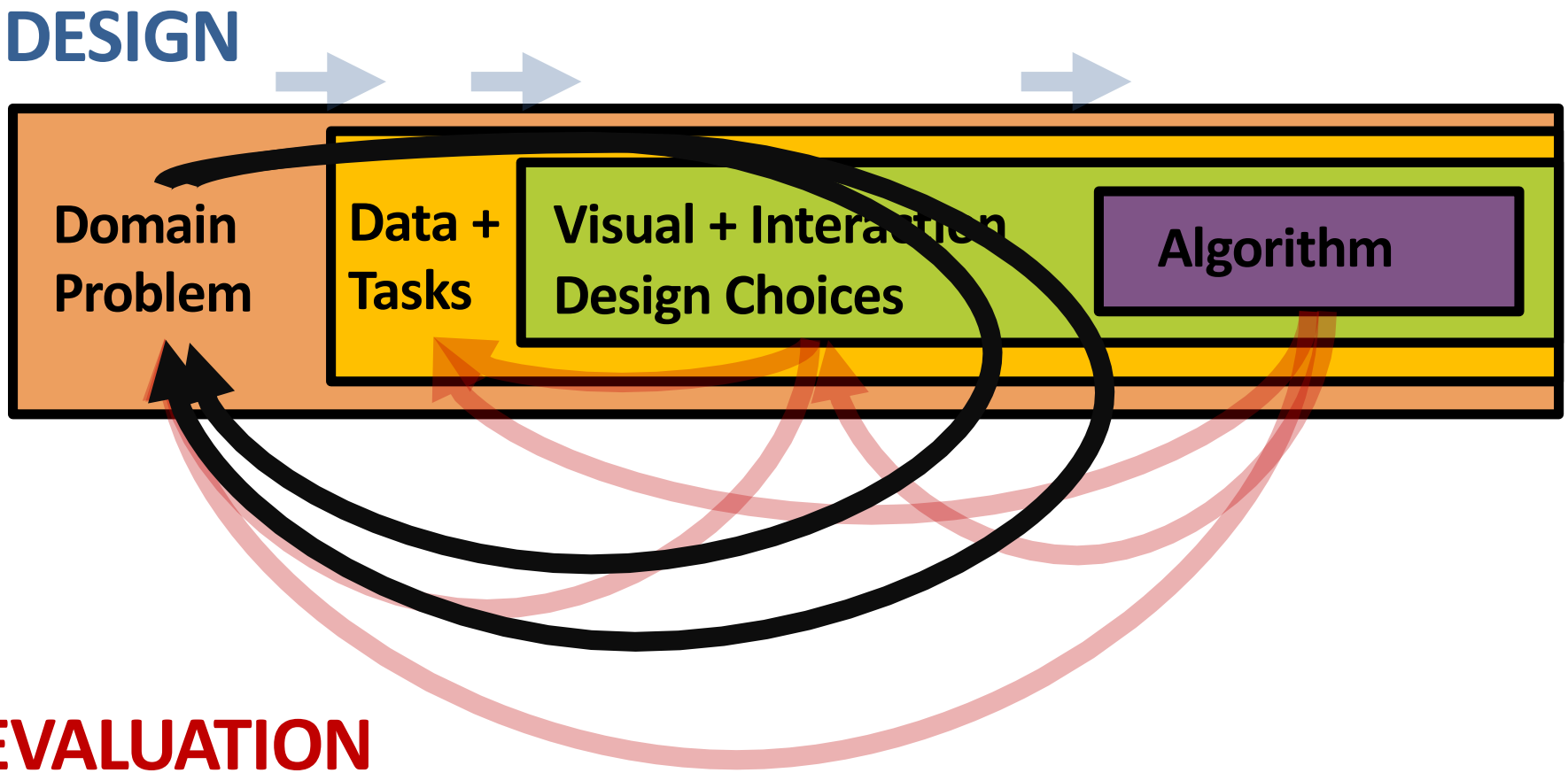
4. Explore if other visualizations have addressed this **problem** and set of **tasks & data**
5. Implement **your own solution** (part or all of that solution could be a new algorithm)

EVALUATION



6. Test **multiple alternatives** (including new ones you develop) with stakeholders
7. Gather **qualitative & quantitative** evaluation data

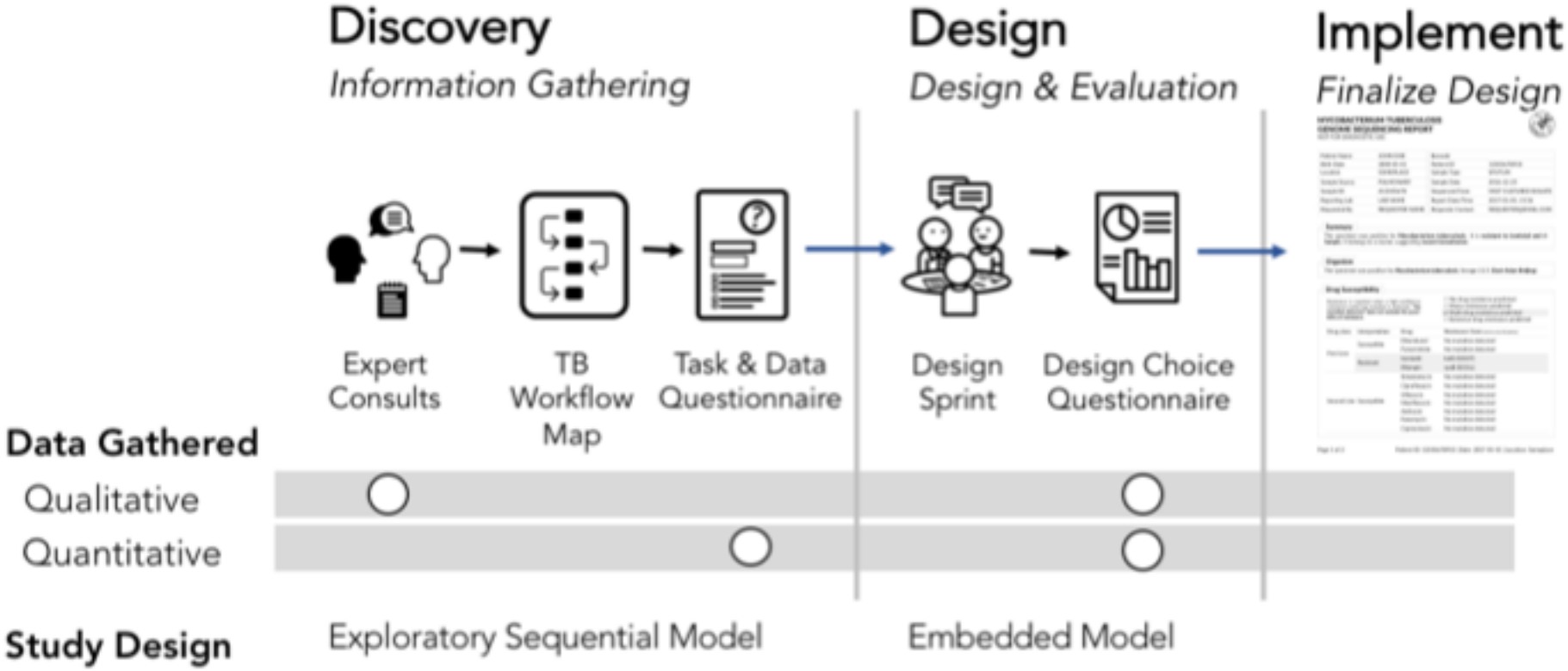
Systematic thinking in *action*



Systematic thinking in *action*

Design	Evaluation
Why?	Does the visualization address the the intended need?
What?	Are you using the right data, or *deriving* the right data?
How?	Are the visual & interactive choices appropriate for the data and tasks? If interactive / computer based, is the visualization easy to use and reliable (*i.e.* doesn't crash all the time)

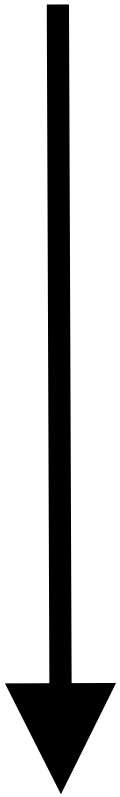
Example: An evaluation of design choices in clinical reports

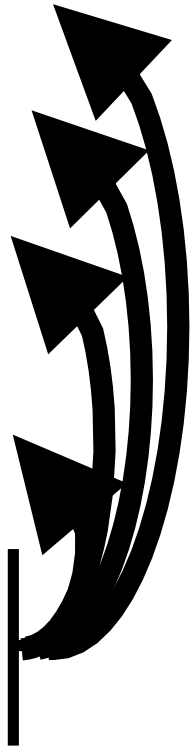


https://peerj.com/articles/4218/

Systematic thinking in *action*

1. What **problem** is data visualization solving? For whom?
2. **What** data should be visualized (is it available)?
3. **What** is the data used for ?[**tasks**]
4. Explore if other visualizations have addressed this **problem** and set of **tasks & data**
5. Implement **your own solution** (vis and/or algorithm)
6. Test **multiple alternatives** (including new ones you develop) with stakeholders
7. Gather **qualitative & quantitative** evaluation data





M. Sedlmair et al. (2012) - A design study methodology : lessons from the trenches

How should we visualize data in *infectious disease GenEpi??*

Introducing GEViT

- **GEviT** = Genomic Epidemiology Visualization Typology
  - A way to describe data visualization for analysis
  - Organizes qualitative descriptors into a typology
- What does GEviT do and not do?

GEviT provides a base

- Deliverables :
  1. Typology
  2. Interactive Gallery

GEviT does not evaluate

- Massive undertaking that would take many years
- Needs GEViT to conduct evaluations

https://doi.org/10.1101/325290

What is GEViT?

- GEViT is a **why-what-how** typology
  - Uses methods from qualitative methods & infovis research
- **Why** are data being visualized?
  - *i.e. show transmission in a hospital*
- **What** data are being visualized?
  - *i.e. patient location, duration in hospital, test outcomes, SNPs, clusters*
- **How** are data being visualized?
  - *i.e. timeline, phylogenetic*
  - *GEViT breaks down data visualizations into basic chart types, chart combinations and chart enhancements*

The GEViT Gallery

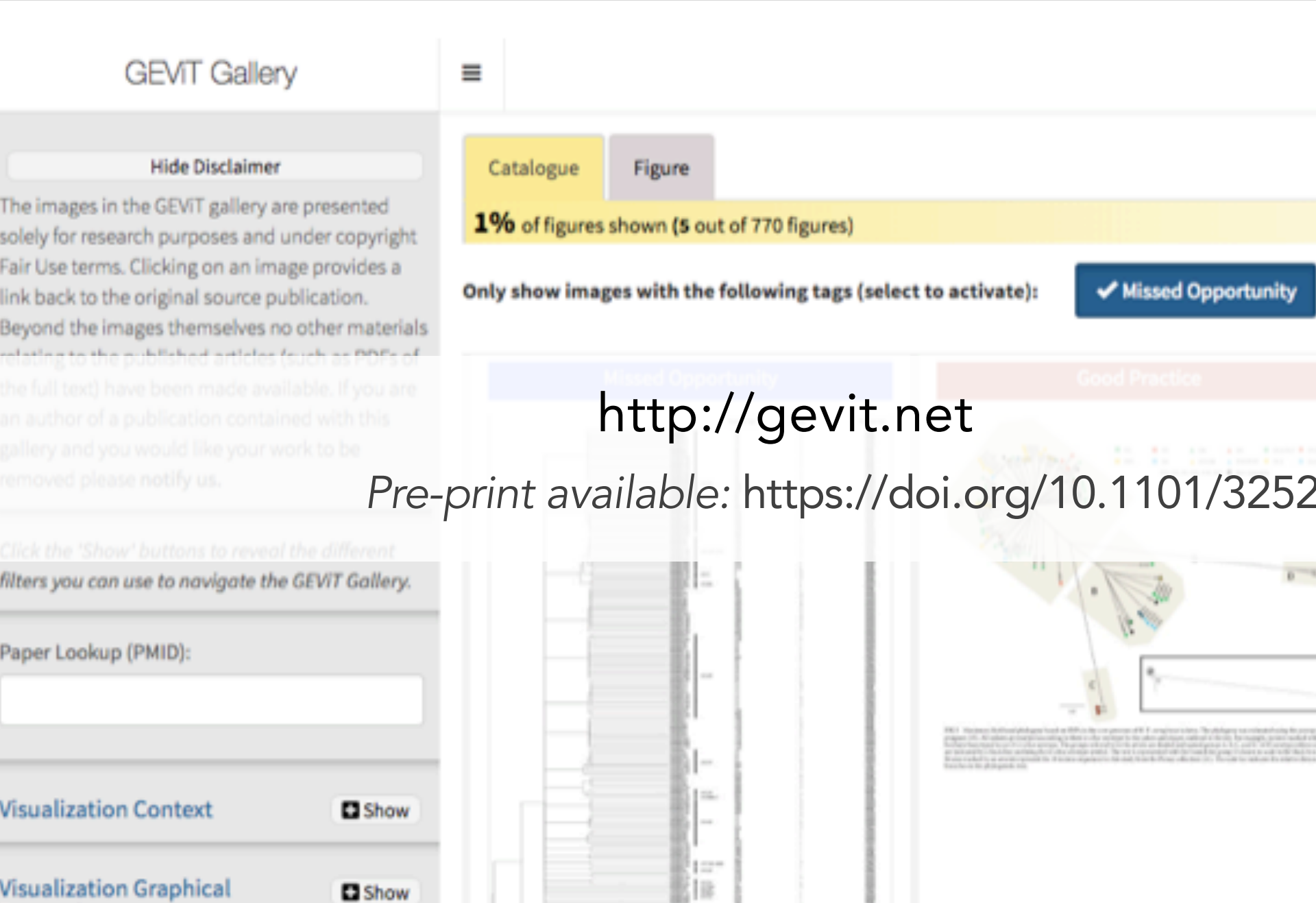
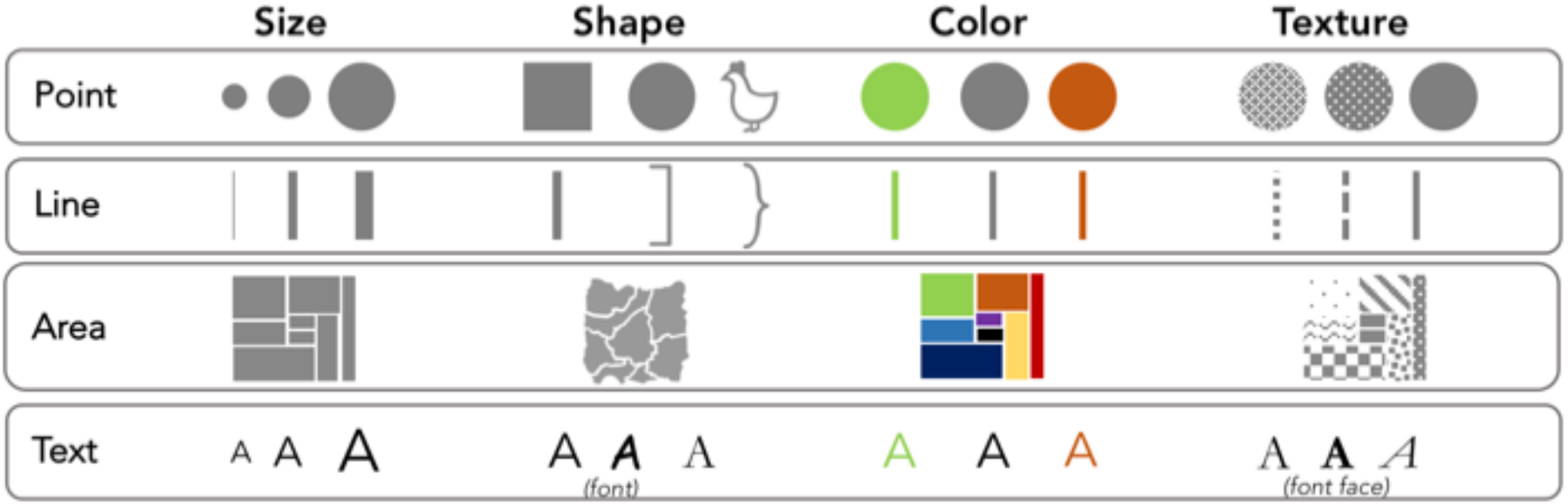


Chart Types : the building block of a data visualization

- Self explanatory what chart types are…
- In our study, we identified seven classes of charts:
  - Common Statistical Charts
  - Relational
  - Temporal
  - Spatial
  - Trees
  - Genomic
  - Other
- Each class of chart also has *special chart types*

Chart Enhancements: a way to overlay additional metadata



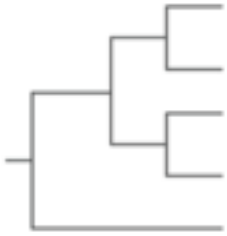
Special Cases : Containment Mark; Connection Mark; Glyph

Chart Enhancements: a way to overlay additional metadata

Structured Enhancements

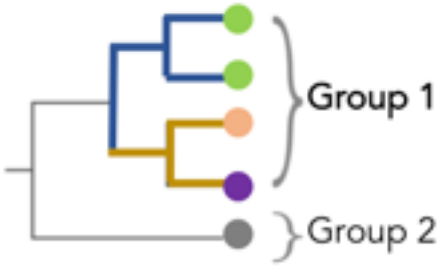
Unstructured Enhancements

Base Chart
Tree



Re-encode Marks
Line - color

Add Marks
Point, line, & text



Add Annotation
Arrow, text

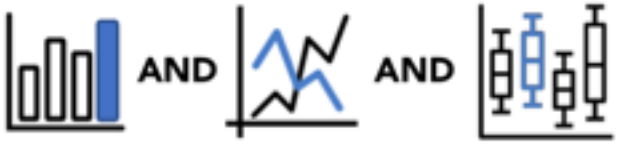
Node-link

Point - size

Glyph - Pie Chart

Containment Mark

Chart Combinations: data from different perspectives

Combination Type	# of chart types	# of charts	Linkage type	*Example*
Simple	1	1	NA	
Composite	Many	1	Spatially Aligned	
Small Multiples	1	Many	Chart Type & Data	
Many Type *Linked*	Many	Many	Visual, but not spatial	
Many Type General	Many	Many	NA	
Complex Combination	Many	Many	Context dependent	

How GEViT can help you

- Let's you explore an IDGE visualization design space
  - Get ideas for your data visualization needs
  - See if different bioinformatics tools can help you create the kinds of data visualizations you need
  - Free your data from text labels and tables!
- Let us know what you think! GEViT will continue to evolve with your feedback!

What have you (hopefully) learned?

Learning Objectives of Module

- Understand the process of encoding and decoding data that is visualized
- How to think systematically about data visualizations
- Know what a visualization design space is and how to reason between different visualization design choices

We are on a Coffee Break & Networking Session