DATA VISUALIZATION

Bùi Tiến Lên

01/01/2020

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

What is Visualization

Concept 1

Computer-based **visualization** (**vis**) systems provide visual representations of datasets designed to **help** people **carry out** tasks more effectively.

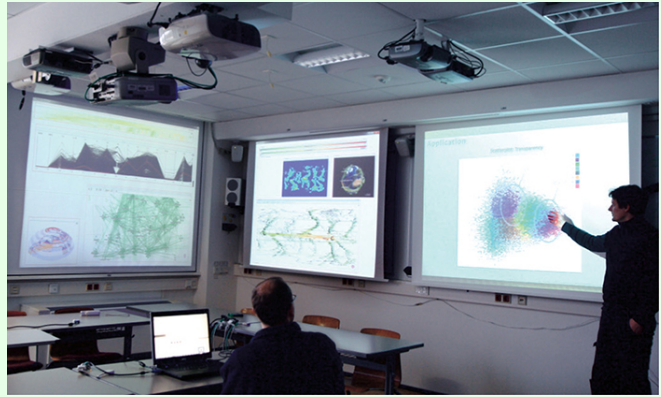
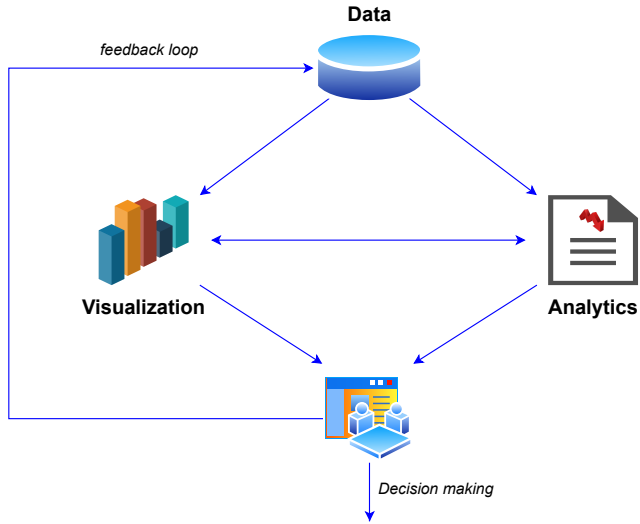


Diagram of Components

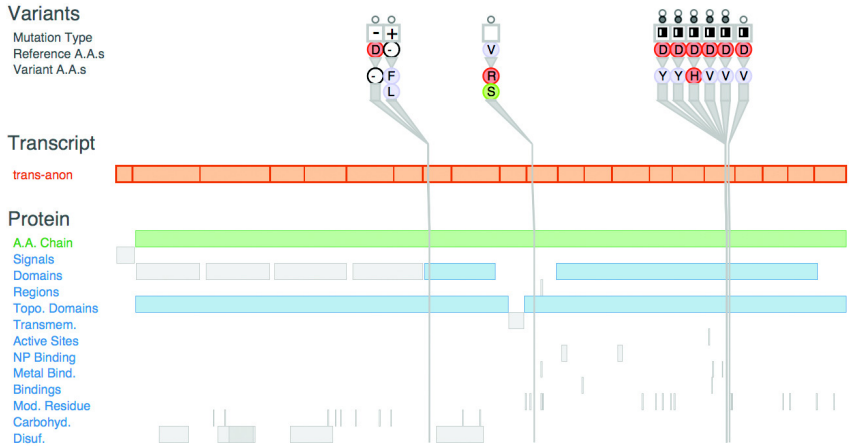


Why have a human in the loop?

- Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods
- Vis allows people to analyze data when they don't know exactly what questions they need to ask in advance
- Possibilities
  - long-term use for end users (e.g. exploratory analysis of scientific data)
  - presentation of known results
  - stepping stone to better understanding of requirements before developing models
  - help developers of automatic solution refine/debug, determine parameters
  - help end users of automatic solutions verify, build trust
- **Don't need** vis when fully automatic solution exists and is trusted

Why have a human in the loop? (cont.)

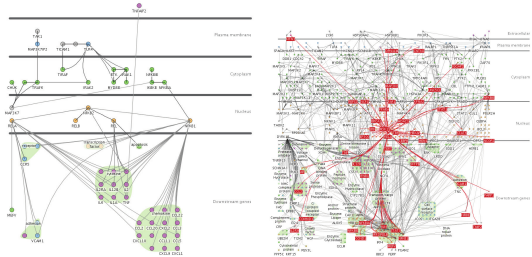
- The Variant View vis tool supports biologists in assessing the impact of genetic variants by speeding up the exploratory analysis process



Why Have a Computer in the Loop?

- By enlisting computation, we can build tools that allow people to explore or present large datasets that would be completely infeasible to draw by hand, thus opening up the possibility of seeing how datasets change over time.

The Cerebral vis tool captures the style of hand-drawn diagrams in biology textbooks with vertical layers that correspond to places within a cell where interactions between genes occur. (a) A small network of 57 nodes and 74 edges might be possible to lay out by hand with enough patience. (b) Automatic layout handles this large network of 760 nodes and 1269 edges and provides a substrate for interactive exploration: the user has moved the mouse over the MSK1 gene, so all of its immmediate neighbors in the network are highlighted in red.



(a) (b)

Why use an external representation?

- **External representation** (sometimes also called *external memory*): replace cognition with perception
- External representations augment human capacity by allowing us to surpass the limitations of our own internal cognition and memory.
- Diagrams can be designed to support perceptual inferences, which are very easy for humans to make.

Why depend on vision?

- **Human visual system** is high-bandwidth channel to brain
  - overview possible due to background processing
    - subjective experience of seeing everything simultaneously
    - significant processing occurs in parallel and pre-attentively
- **Sound**: lower bandwidth and different semantics
  - overview not supported
    - subjective experience of sequential stream
- **Touch/haptics**: impoverished record/replay capacity
  - only very low-bandwidth communication thus far
- **Taste, smell**: no viable record/replay devices

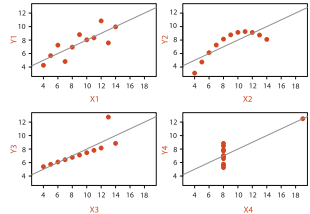
Why show the data in detail?

Summaries lose information

- confirm expected and find unexpected patterns
- assess validity of statistical model

Anscombe's Quartet: Raw Data

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

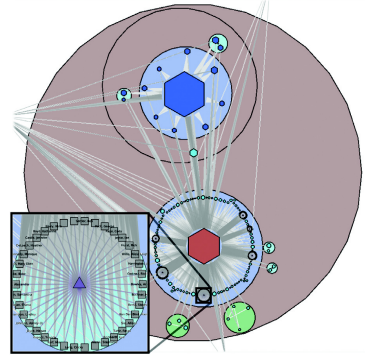


Why Use Interactivity?

- **Interactivity** is crucial for building vis tools that handle complexity.
- When datasets are large enough, the limitations of both people and displays preclude just showing everything at once; interaction where user actions cause the view to change is the way forward.

Why Is the Vis Idiom Design Space Huge?

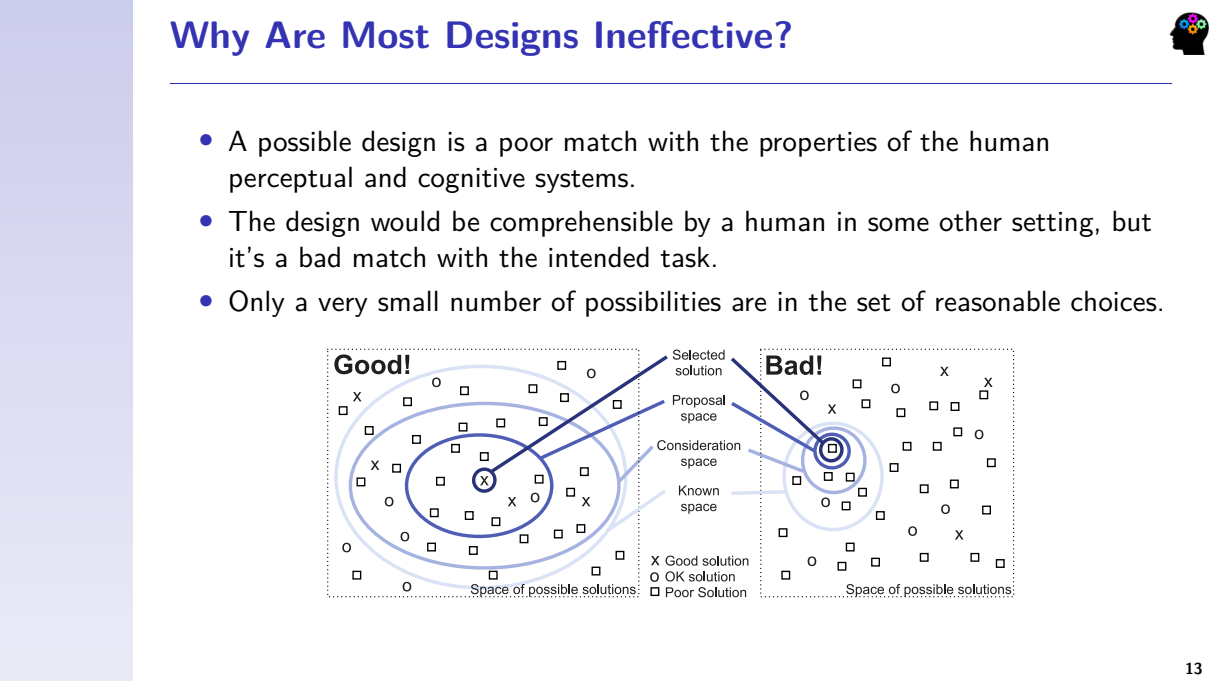
- A vis **idiom** is a distinct approach to creating and manipulating visual representations.
- There are many ways to create a **visual encoding** of data as a single picture.
- The design space of possibilities gets even bigger when we consider how to manipulate one or more of these pictures with **interaction**.



Why focus on tasks and effectiveness?

- **Tasks** serve as constraint on design (as does data)
  - idioms do not serve all tasks equally!
  - challenge: recast tasks from domain-specific vocabulary to abstract forms
- Most possibilities ineffective
  - validation is necessary, but tricky
  - increases chance of finding good solutions if we understand full space of possibilities
- What counts as effective?
  - novel: enable entirely new kinds of analysis
  - faster: speed up existing workflows

Why Are Most Designs Ineffective?



- A possible design is a poor match with the properties of the human perceptual and cognitive systems.
- The design would be comprehensible by a human in some other setting, but it's a bad match with the intended task.
- Only a very small number of possibilities are in the set of reasonable choices.

Why Is Validation Difficult?

The problem of **validation** for a vis design is difficult because there are so many questions that you could ask when considering whether a vis tool has met your design goals

- How do you know if it works?
- How do you argue that one design is better or worse than another for the intended users?
- For one thing, what does better mean?
- Do users get something done faster?
- Do they have more fun doing it?
- Can they work more effectively?
- What does effectively mean?
- How do you measure insight or engagement?

Why Is Validation Difficult? (cont.)

- What is the design better than?
- Is it better than another vis system?
- Is it better than doing the same things manually, without visual support?
- Is it better than doing the same things completely automatically?
- What sort of thing does it do better?
- How do you decide what sort of task the users should do when testing the system?
- Who is this user?
- An expert who has done this task for decades, or a novice who needs the task to be explained before they begin?
- Are they familiar with how the system works from using it for a long time, or are they seeing it for the first time?

Why Is Validation Difficult? (cont.)

A concept like faster might seem straightforward, but tricky questions still remain.

- Are the users limited by the speed of their own thought process, or their ability to move the mouse, or simply the speed of the computer in drawing each picture?
- How do you decide what sort of benchmark data you should use when testing the system?
- Can you characterize what classes of data the system is suitable for?
- How might you measure the quality of an image generated by a vis tool?
- How well do any of the automatically computed quantitative metrics of quality match up with human judgements?

Why Is Validation Difficult? (cont.)

Even once you limit your considerations to purely computational issues, questions remain.

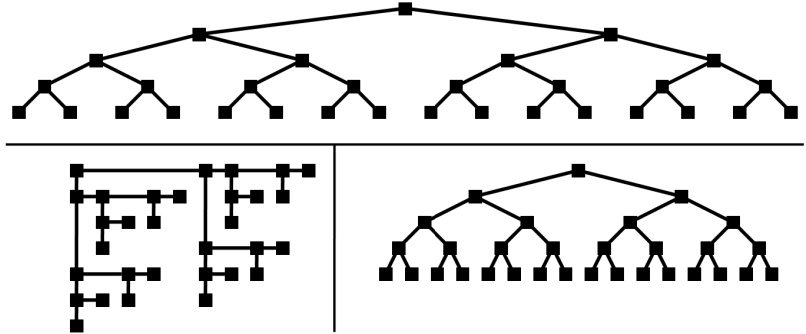
- Does the complexity of the algorithm depend on the number of data items to show or the number of pixels to draw?
- Is there a trade-off between computer speed and computer memory usage?

What resource limitations are we faced with?

- Computational limits
  - processing time
  - system memory
- Human limits
  - human attention and memory: **change blindness**
- Display limits
  - pixels are precious resource, the most constrained resource
  - **information density**: ratio of space used to encode info vs. unused whitespace $\rightarrow$ tradeoff between clutter and wasting space, find sweet spot between dense and sparse

What resource limitations are we faced with? (cont.)

- Low and high information density visual encodings of the same small tree dataset; nodes are the same size in each. (*top*) Low information density. (*left*) Higher information density, but depth in tree cannot be read from spatial position. (*right*) High information density, while maintaining property that depth is encoded with position.

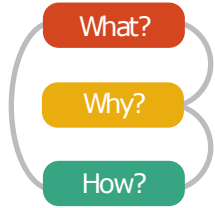


Why analyze?

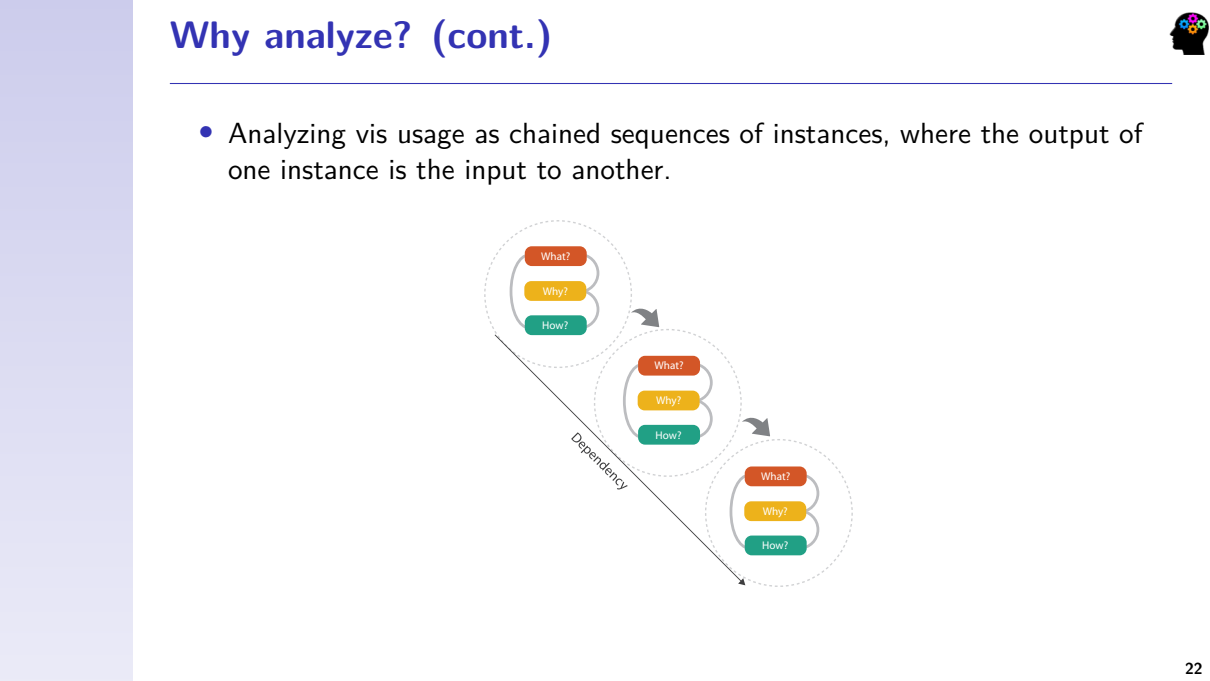
- Imposes structure on huge design space
  - scaffold to help we think systematically about choices
  - analyzing existing as stepping stone to designing new
  - most possibilities ineffective for particular task/data combination

Why analyze? (cont.)

- Three-part analysis framework for a vis instance: **why** is the task being performed, **what** data is shown in the views, and **how** is the vis idiom constructed in terms of design choices.



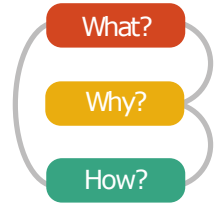
Why analyze? (cont.)



- Analyzing vis usage as chained sequences of instances, where the output of one instance is the input to another.

Analysis framework: Four levels, three questions

- **Domain situation**
  - who are the target users?
- **Abstraction**
  - translate from specifics of domain to vocabulary of vis
  - **what** is shown? **data abstraction**, often don't just draw what you're given: transform to new form
  - **why** is the user looking at it? **task abstraction**
- **Idiom**
  - **how** is it shown?
    - visual encoding idiom: how to draw
    - interaction idiom: how to manipulate
- **Algorithm**
  - efficient computation

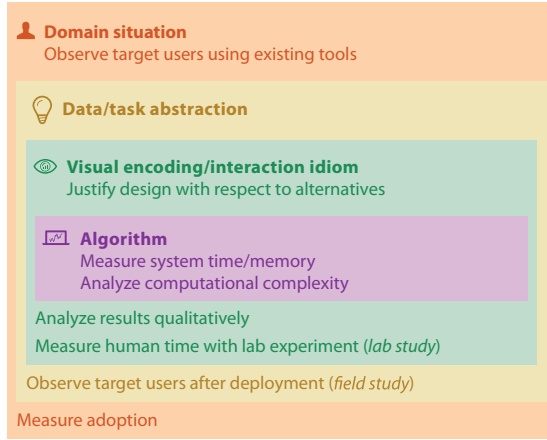


Why is validation difficult?

- different ways to get it wrong at each level

Why is validation difficult? (cont.)

- solution: use methods from different fields at each level



References

Goodfellow, I., Bengio, Y., and Courville, A. (2016).
Deep learning.
MIT press.

Munzner, T. (2014).
Visualization analysis and design.
CRC press.

Russell, S. and Norvig, P. (2016).
Artificial intelligence: a modern approach.
Pearson Education Limited.

Ward, M. O., Grinstein, G., and Keim, D. (2015).
Interactive data visualization: foundations, techniques, and applications.
CRC Press.