Visual analytics

Information Visualization, 2020/21

Jan Zahálka

jan.zahalka@bohem.ai

Ω

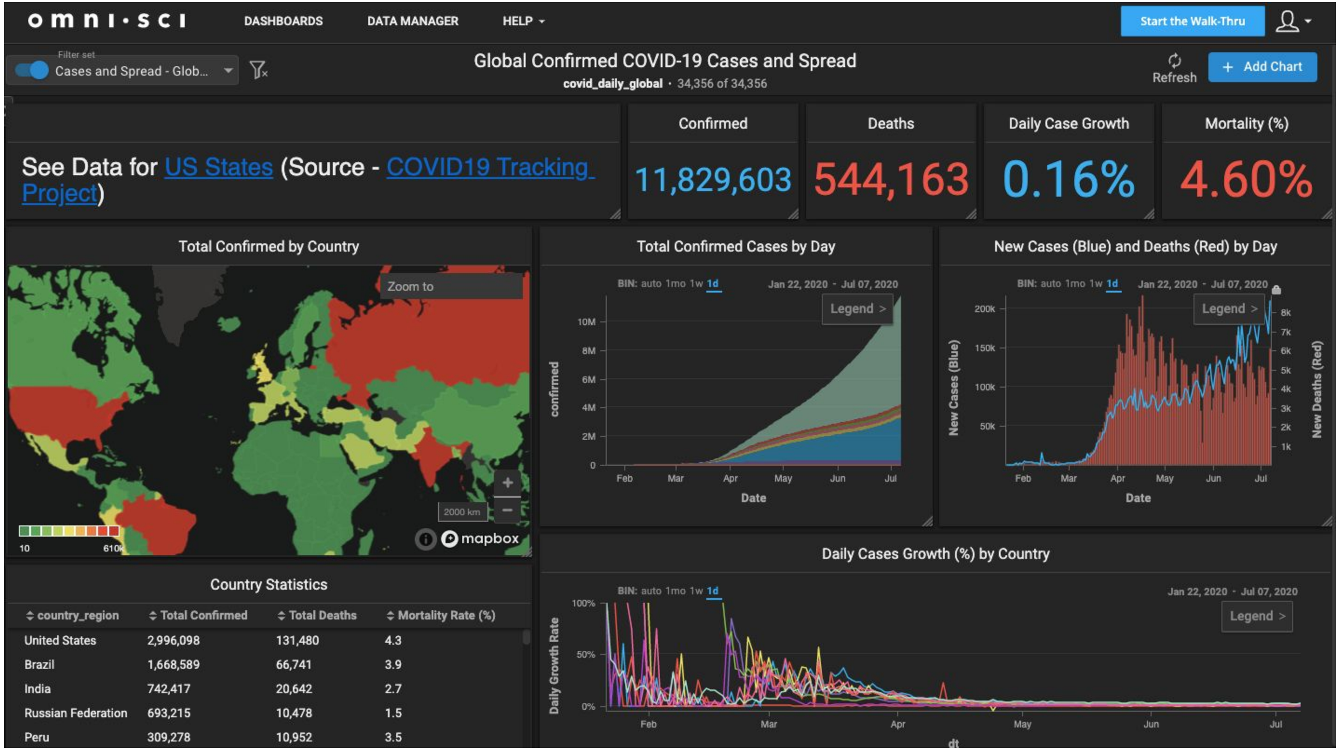
Today

- Visual analytics theory and motivation
- Designing models to accompany our visualization

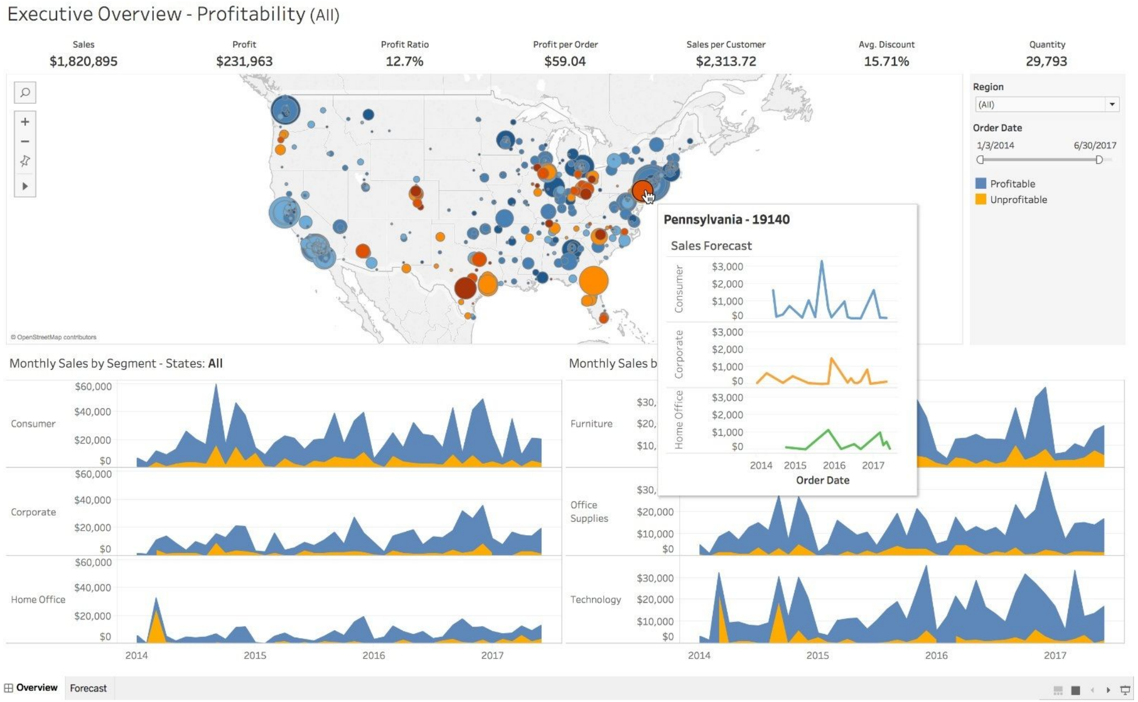
Recall lecture 1: Visual analytics

- The science of **analytical reasoning** facilitated by **interactive visual interfaces**
- In civil terms:
  - A domain expert (e. g., a scientist or a police investigator) wants to solve a problem (e. g., investigate a suspect's seized computer or the incidence of a disease in a population)
  - The solution comes from analyzing a **large, complex dataset** which cannot be feasibly analyzed by normal means
  - Visual analytics builds a **system** that allows the expert to analyze the data **iteratively** and **interactively**
    - **Iteratively**: it takes time and a gradual approach to grapple with the data
    - **Interactively**: static visualizations don't cut it, the expert has to perform many subtasks to progress, hands-on approach helps understanding

Visual analytics example: OmniSci



Visual analytics example: Tableau



Probably the closest to the concept of “general VA system”

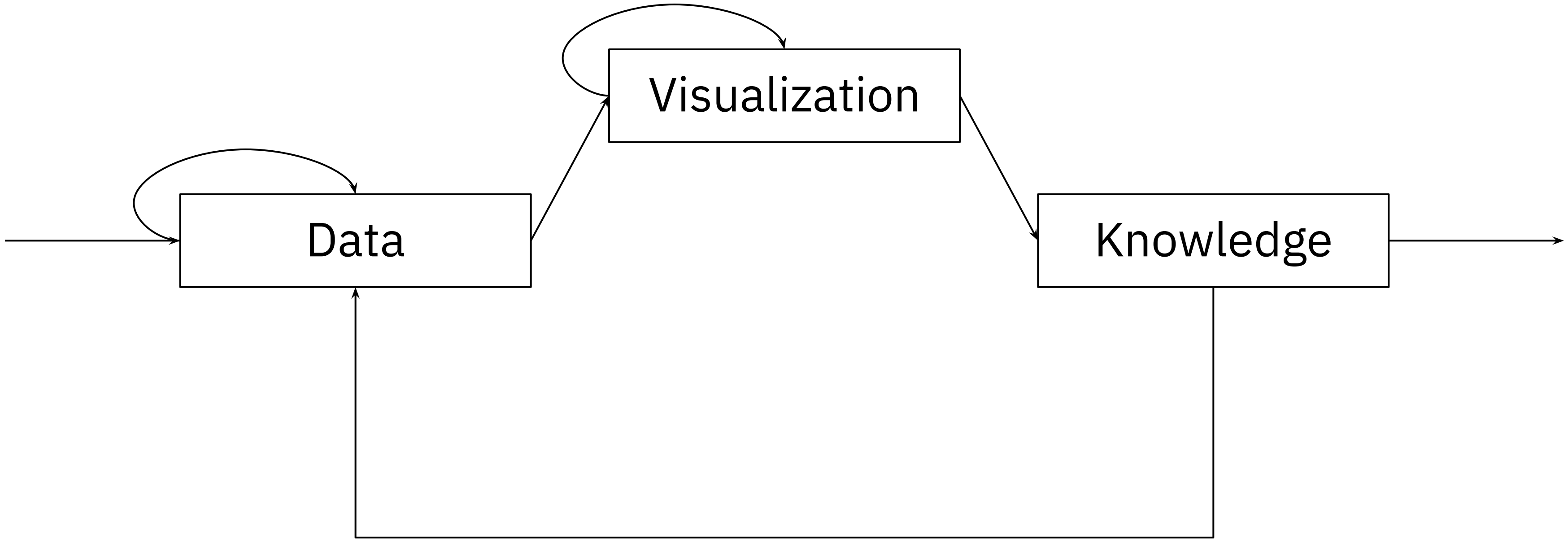
Visual analytics: Typical aspects

- At a glance:
  - **Dashboard-style interface**
  - Multiview design **very important**
  - Individual **views are often basic charts/plots** or heavily utilize them
- A true visual analytics tool goes deeper:
  - An interactive, intelligent **model** of the data that truly **assists the user**
  - Tight integration of **visualization** and the **model** through solid **interaction design**

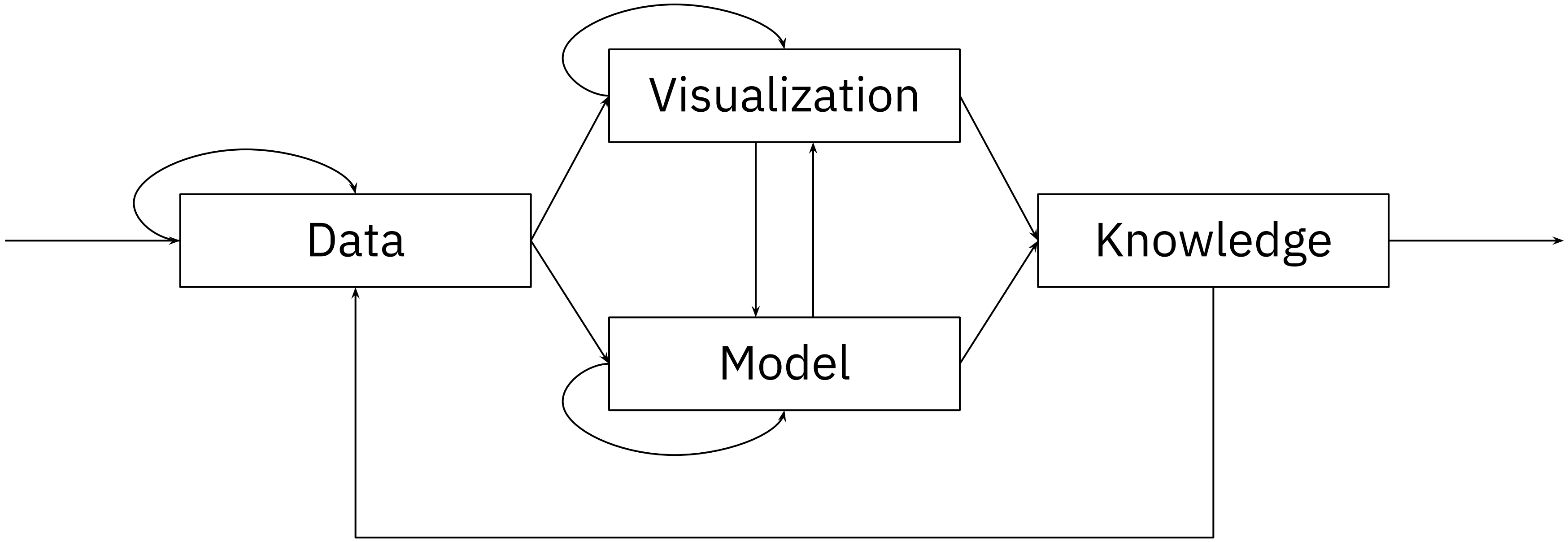
Evolution: From data mining to knowledge disc.

[Fayyad96]

Recall lecture 1: InfoVis pipeline

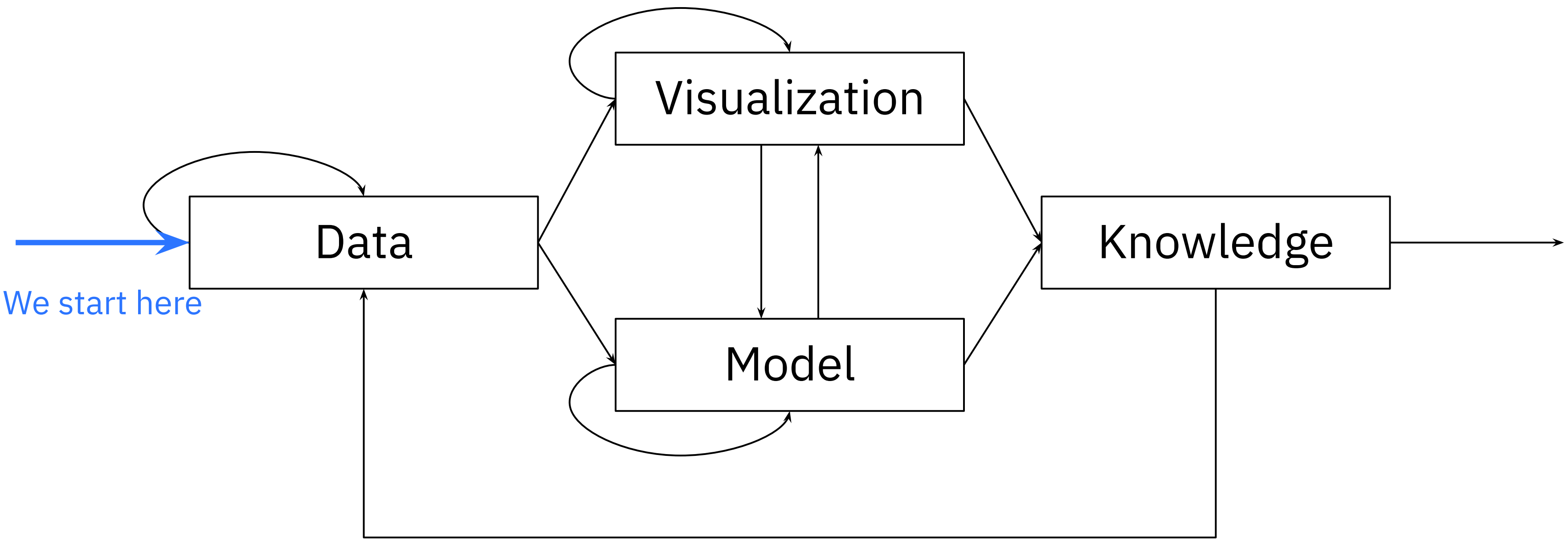


Recall lecture 1: Visual analytics pipeline



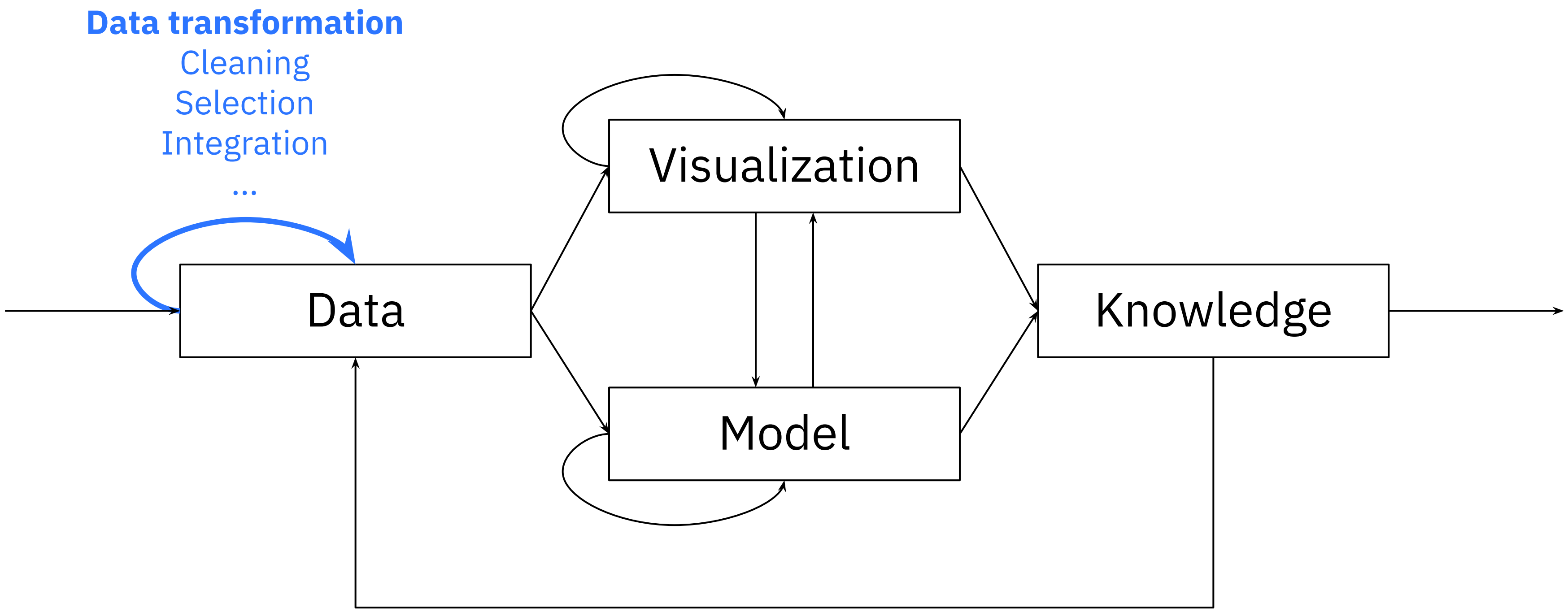
[Keim08]

Visual analytics pipeline



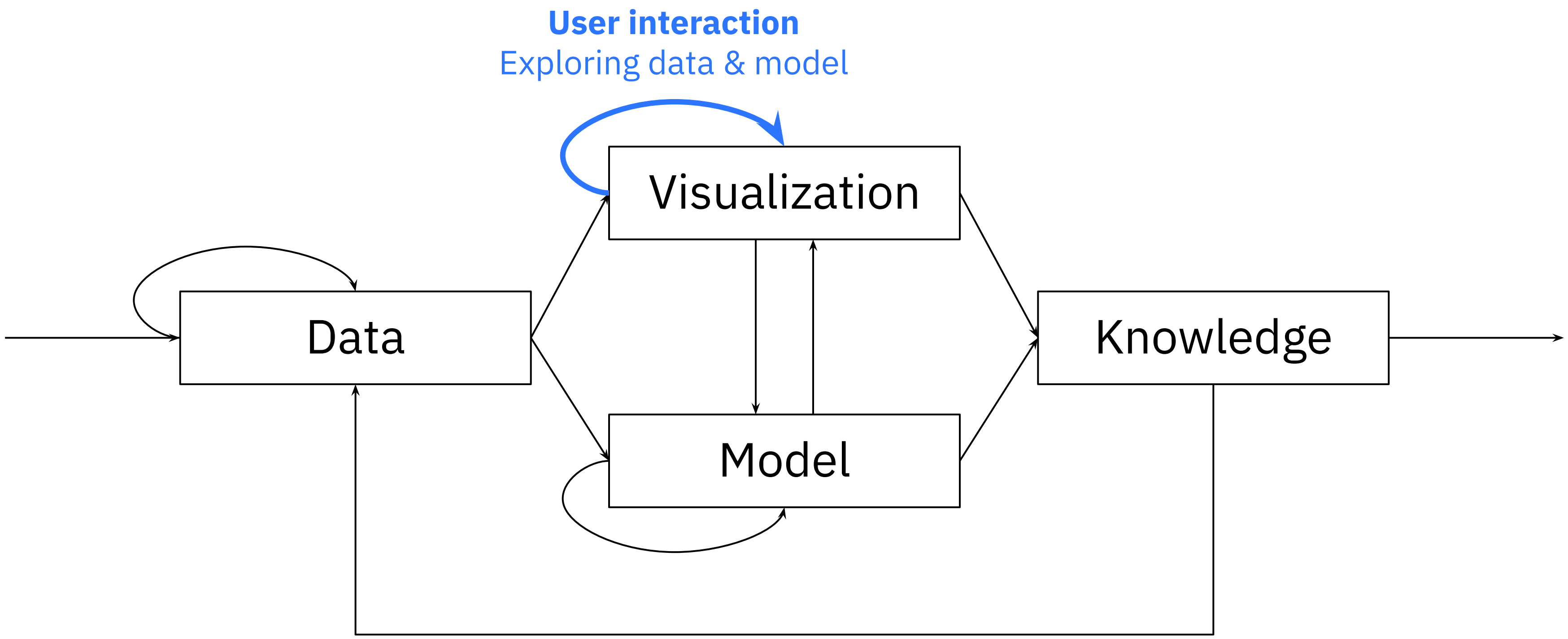
[Keim08]

Visual analytics pipeline



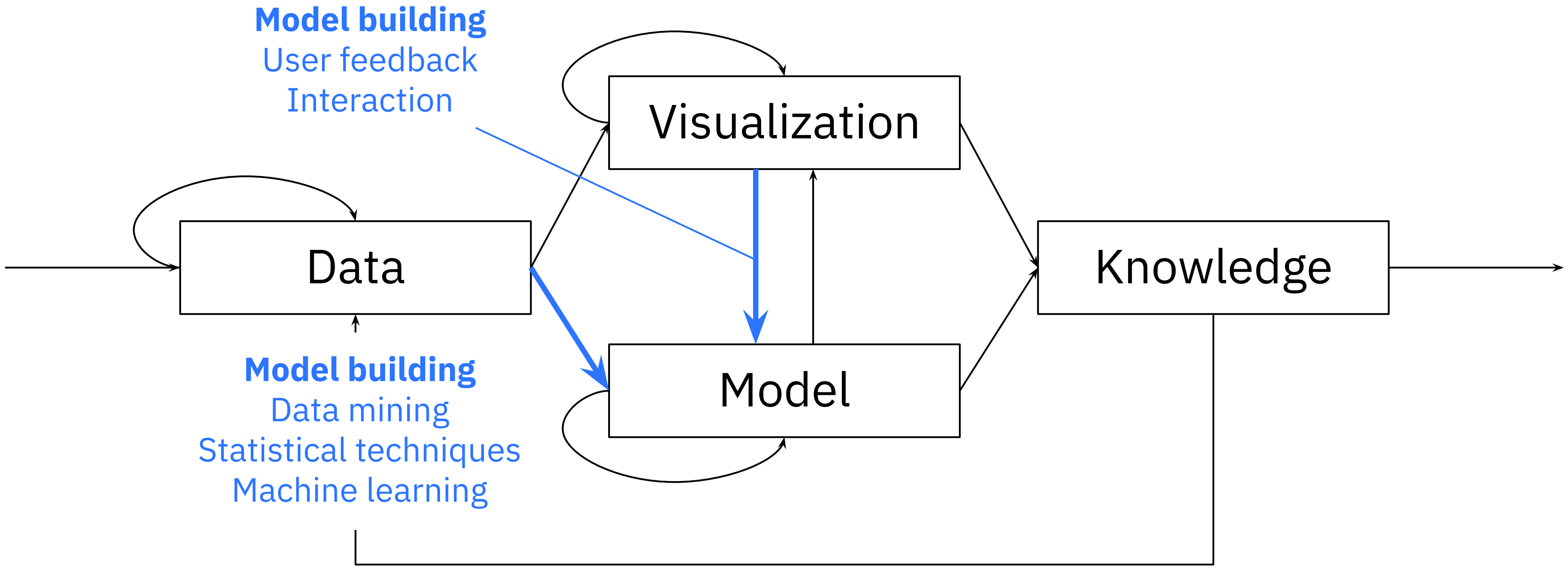
[Keim08]

Visual analytics pipeline



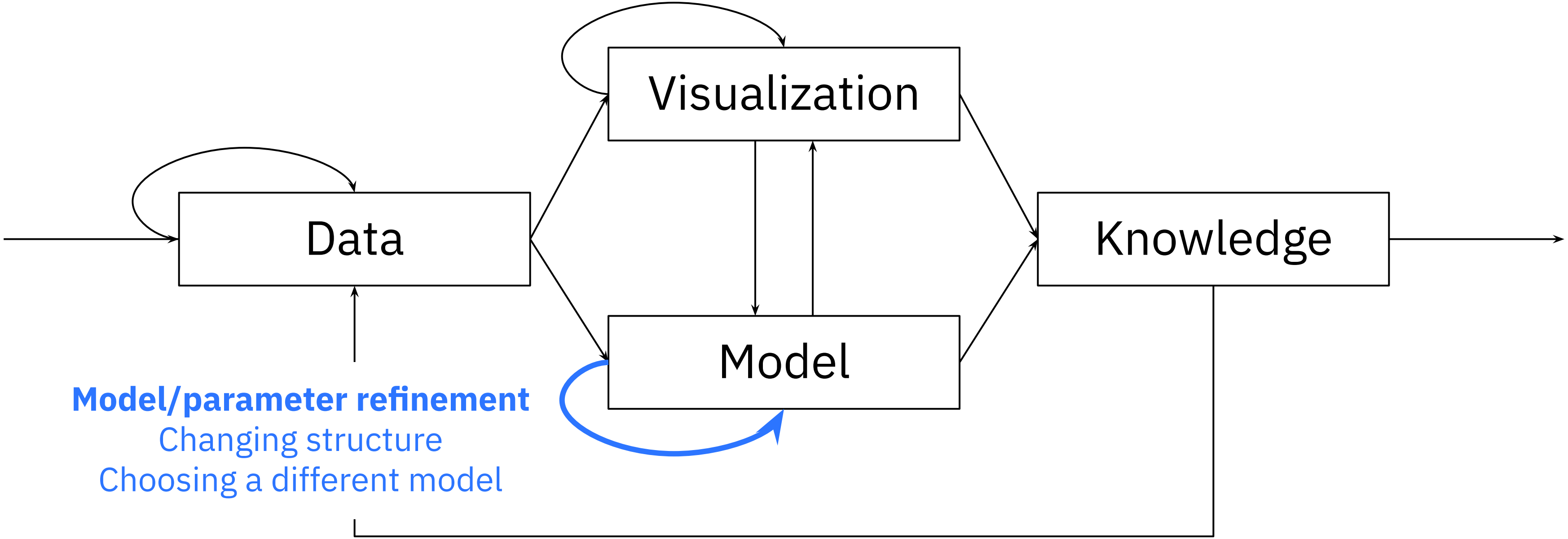
[Keim08]

Visual analytics pipeline



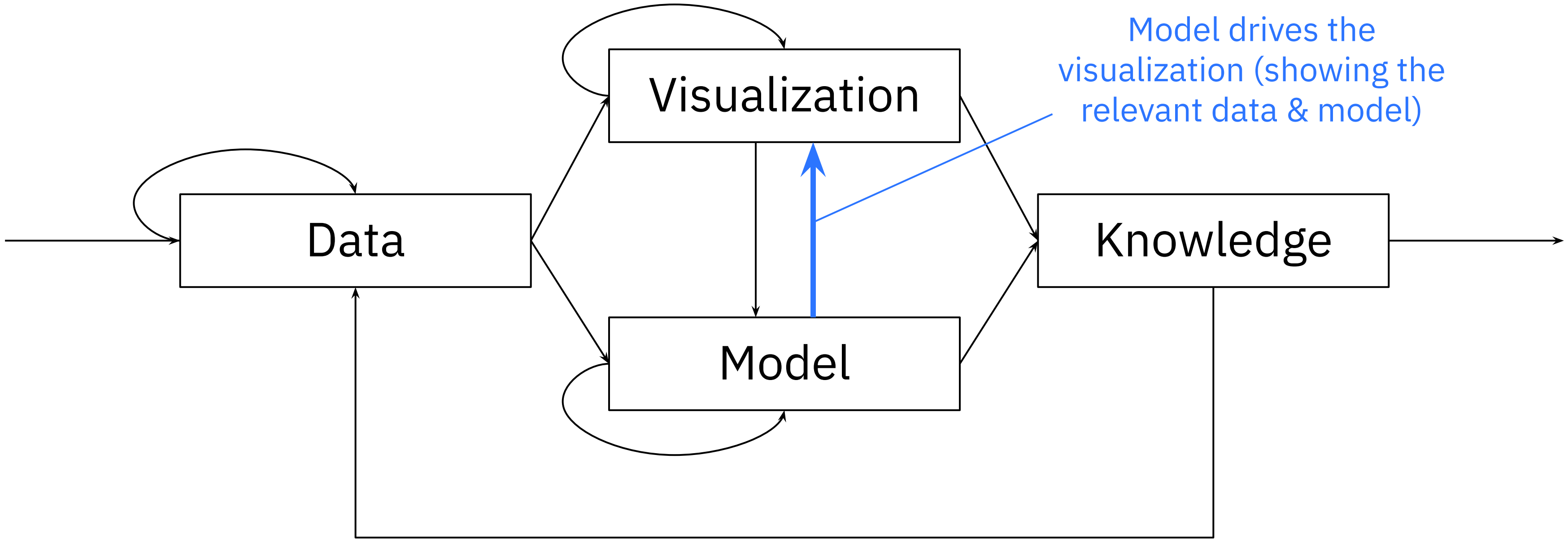
[Keim08]

Visual analytics pipeline



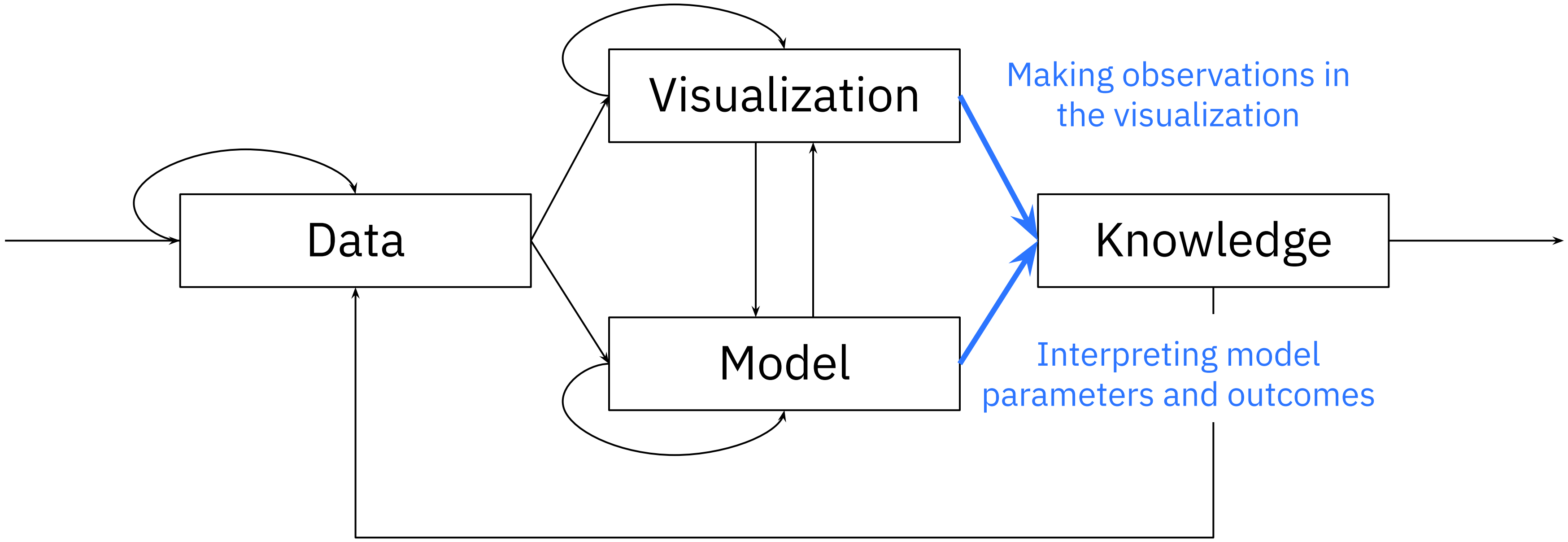
[Keim08]

Visual analytics pipeline



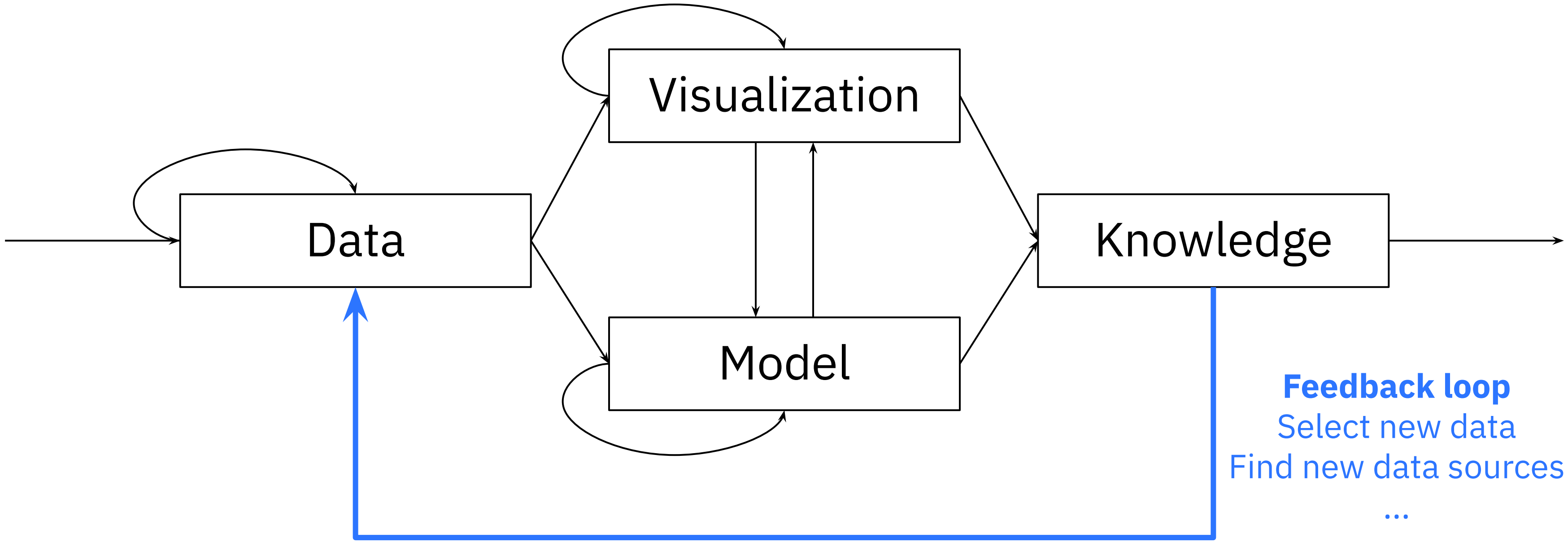
[Keim08]

Visual analytics pipeline



[Keim08]

Visual analytics pipeline

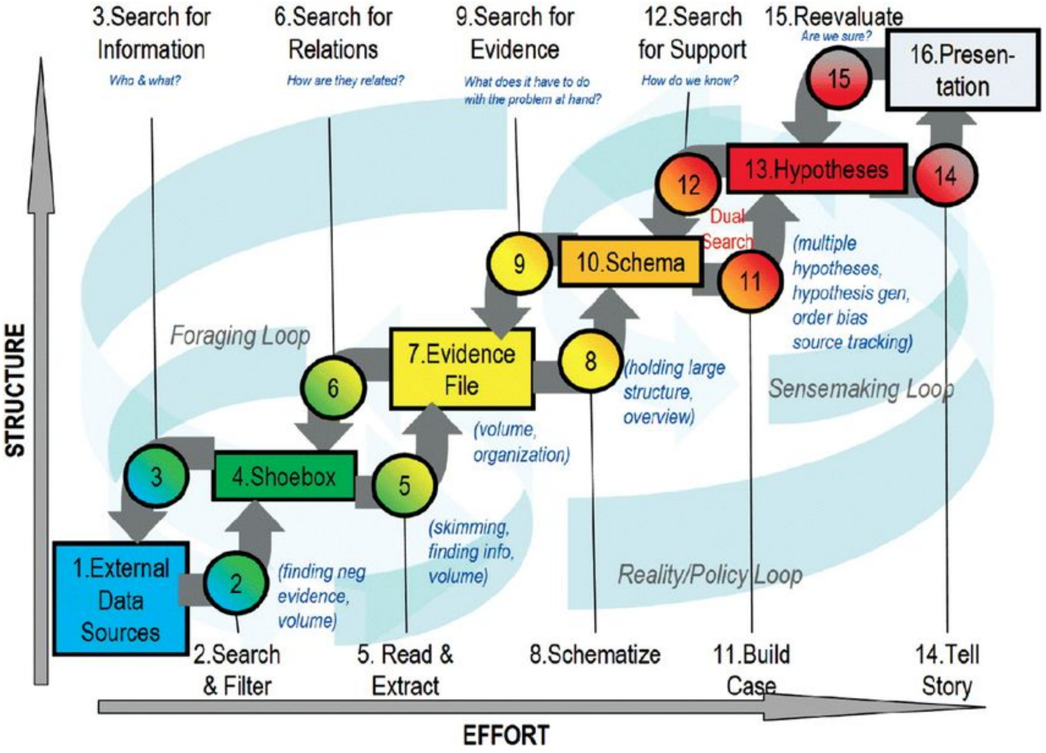


[Keim08]

Visual analytics pipeline

- **System-centric** overview of key components of visual analytics
- Let’s add **human reasoning**

Sensemaking process



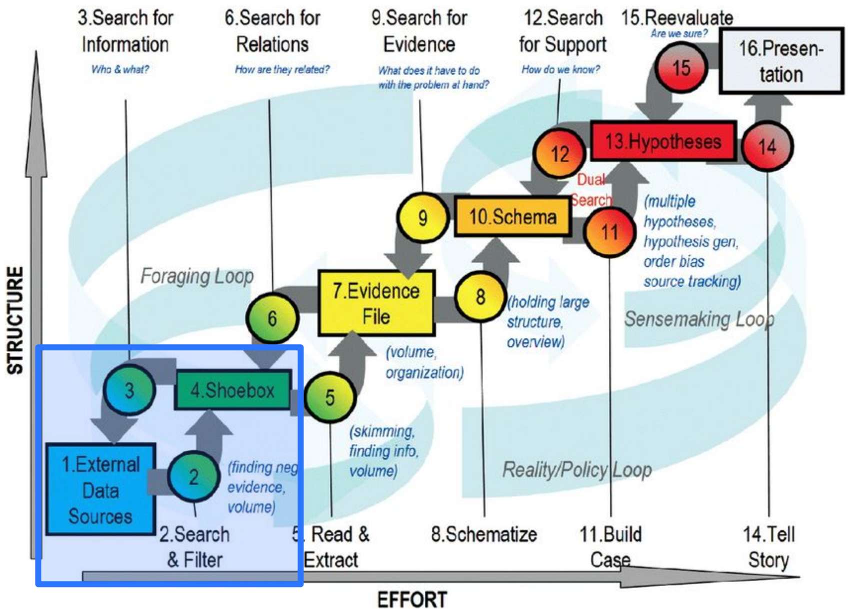
[Pirolli05]

Sensemaking

- **Sensemaking** – Structuring unknown data into a framework enabling us to comprehend, understand, explain, attribute, extrapolate, and predict
- The loops:
  - **Foraging loop** – Seeking information, searching and filtering it, reading and extracting it
  - **Sensemaking loop** – Iterative development of a mental model (conceptualization) that best fits the evidence
  - **Reality/policy loop** – Putting the findings in real-world context

Sensemaking

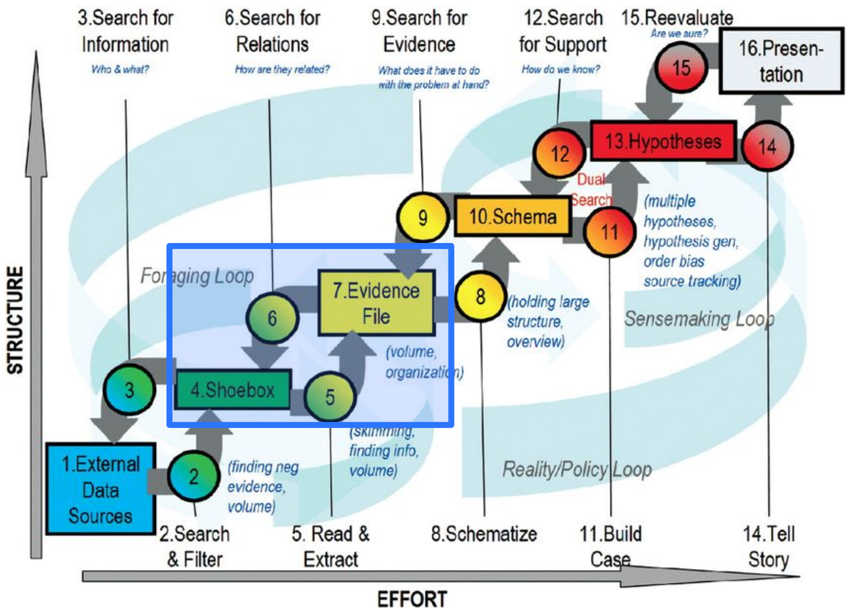
- Models:
  - **External data sources** – Self-explanatory
  - **Shoebox** – Unstructured storage of data filtered based on (rough) relevance
- Processes:
  - **Search & filter** (bottom-up) – Filter the unstructured data and put the potentially relevant instances into the shoebox
  - **Search for information** (top-down) – New hypotheses at higher levels might drive search for new data



[Pirolli05]

Sensemaking

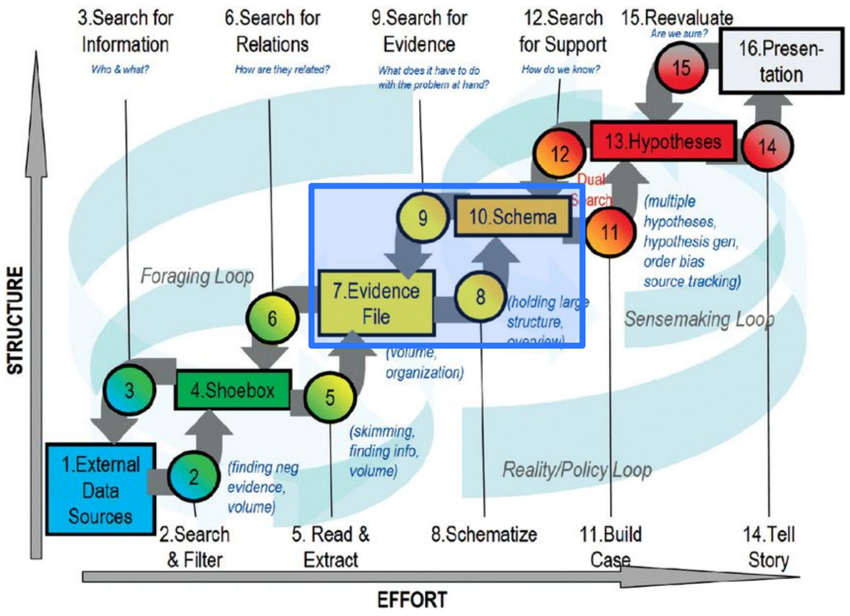
- Models:
  - **Shoebox**
  - **Evidence file** – Snippets extracted from the information in the shoebox, either confirming (hence "evidence") or leading to hypothesis (and thus insight)
- Processes:
  - **Read & extract** (bottom-up) – Placing relevant data items into the evidence file (secondary, more detailed stage of filtering)
  - **Search for relations** (top-down) – Information in evidence file might suggest new patterns or even hypotheses



[Pirolli05]

Sensemaking

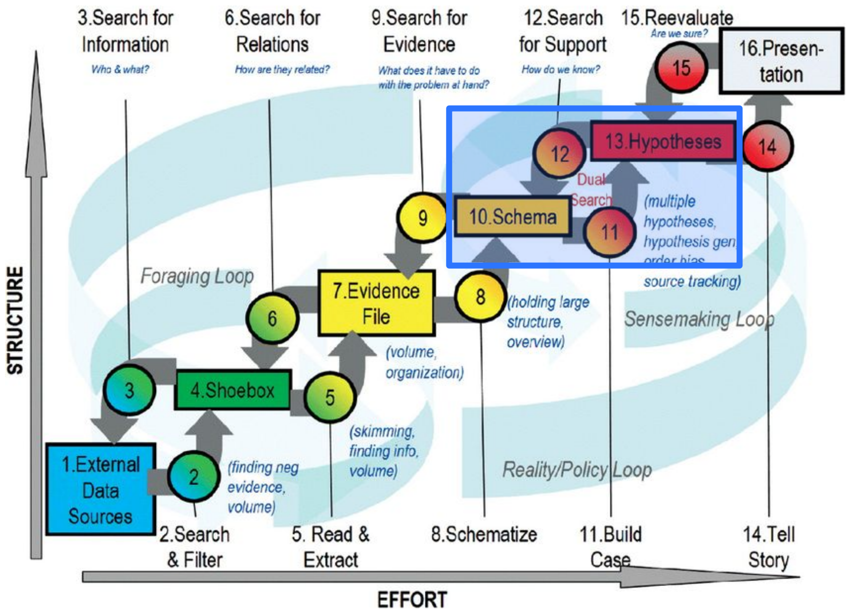
- Models:
    - **Evidence file**
    - **Schema** – A *structured*, well-organized collection of information. Wide range of forms: from a thought through (preliminary) visualizations to curated datasets being stored and documented
- Processes:
    - **Schematize** (bottom-up) – Putting structure on the relevant information
    - **Search for evidence** (top-down) – New hypotheses might drive search for new evidence to support them



[Pirolli05]

Sensemaking

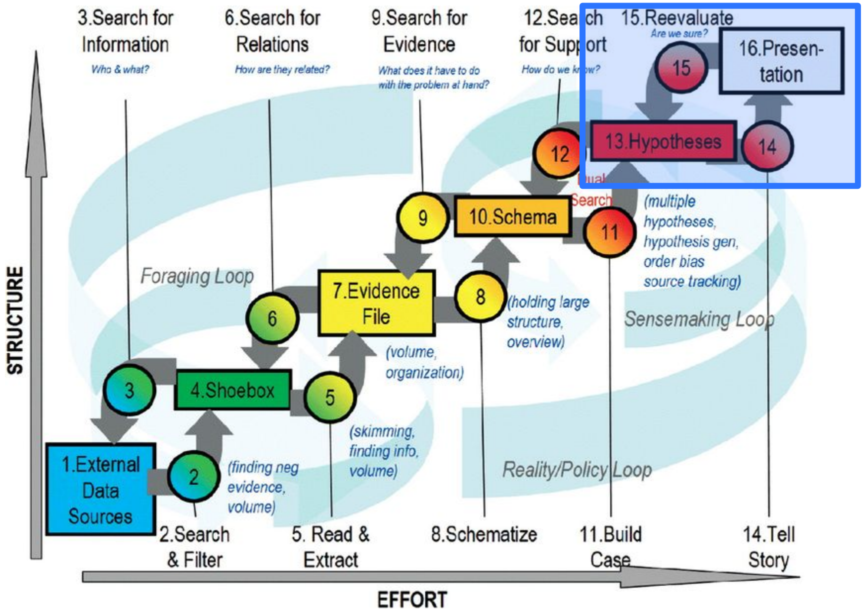
- Models:
  - **Schema**
  - **Hypotheses** – (Tentative) representations of the conclusions about the data
- Processes:
  - **Build case** (bottom-up) – A theory is formalized based on the schema to form/support hypotheses
  - **Search for support** (top-down) – Reevaluation of theories leads to reexamination of the schema



[Pirolli05]

Sensemaking

- Models:
  - **Hypotheses**
  - **Presentation** – The outcome of your work ("deliverable" could be a better word)
- Processes:
  - **Tell story** (bottom-up) – A deliverable is built based on the hypotheses and conclusion
  - **Reevaluate** (top-down) – Consumer feedback often leads to reevaluation of hypotheses or new hypotheses



[Pirolli05]

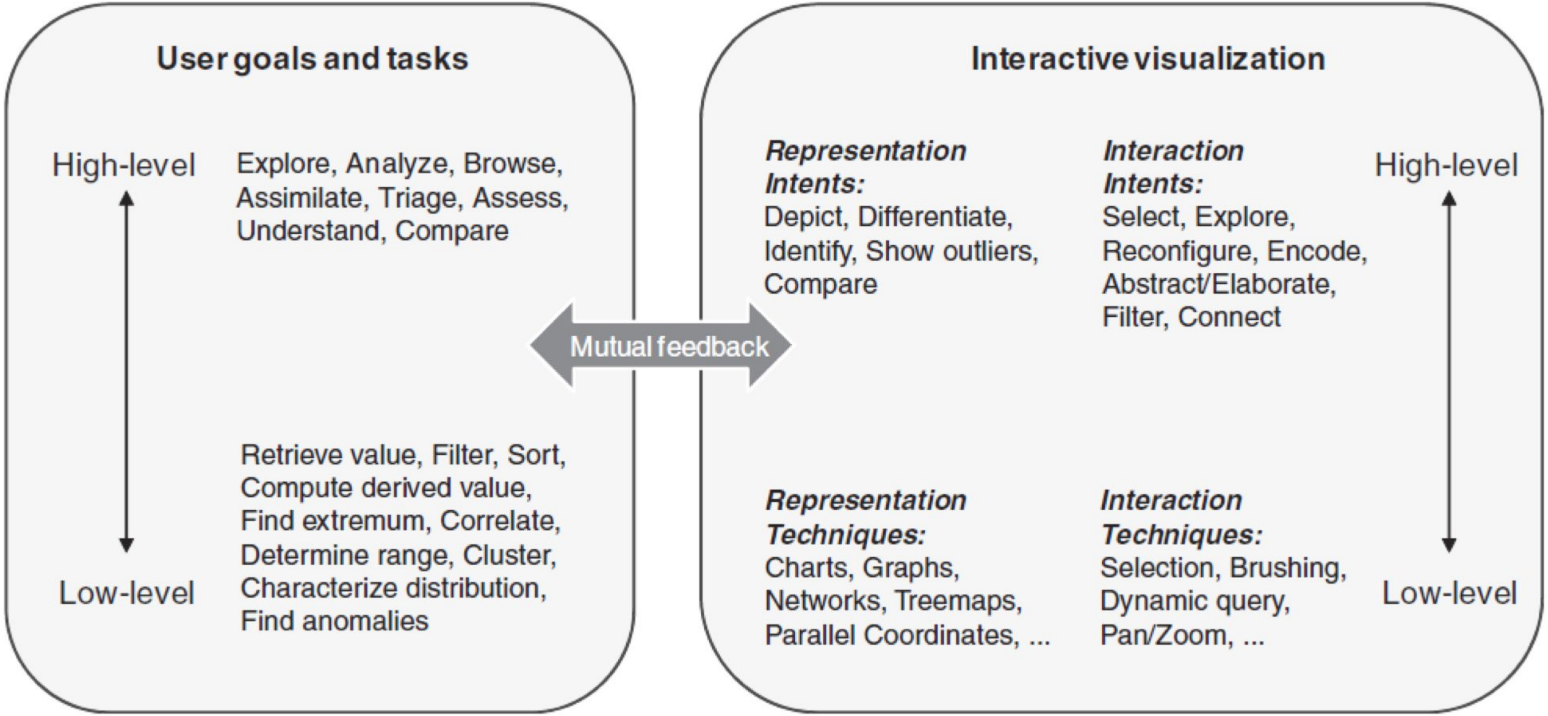
Sensemaking

- At a glance, the sensemaking diagram could give the idea of a **waterfall**
  - External data → Shoebox → Evidence file → Schema → Hypotheses → Presentation
  - You only go back in the case of a mistake
- On the contrary: **loops are an essential, constructive part of the process**
  - The users can loop freely as per their needs
  - “Top-down and bottom-up processes are invoked in an opportunistic mix” [Pirolli05]
- … and they drive the entire **analytics process**

Machine model vs. mental model

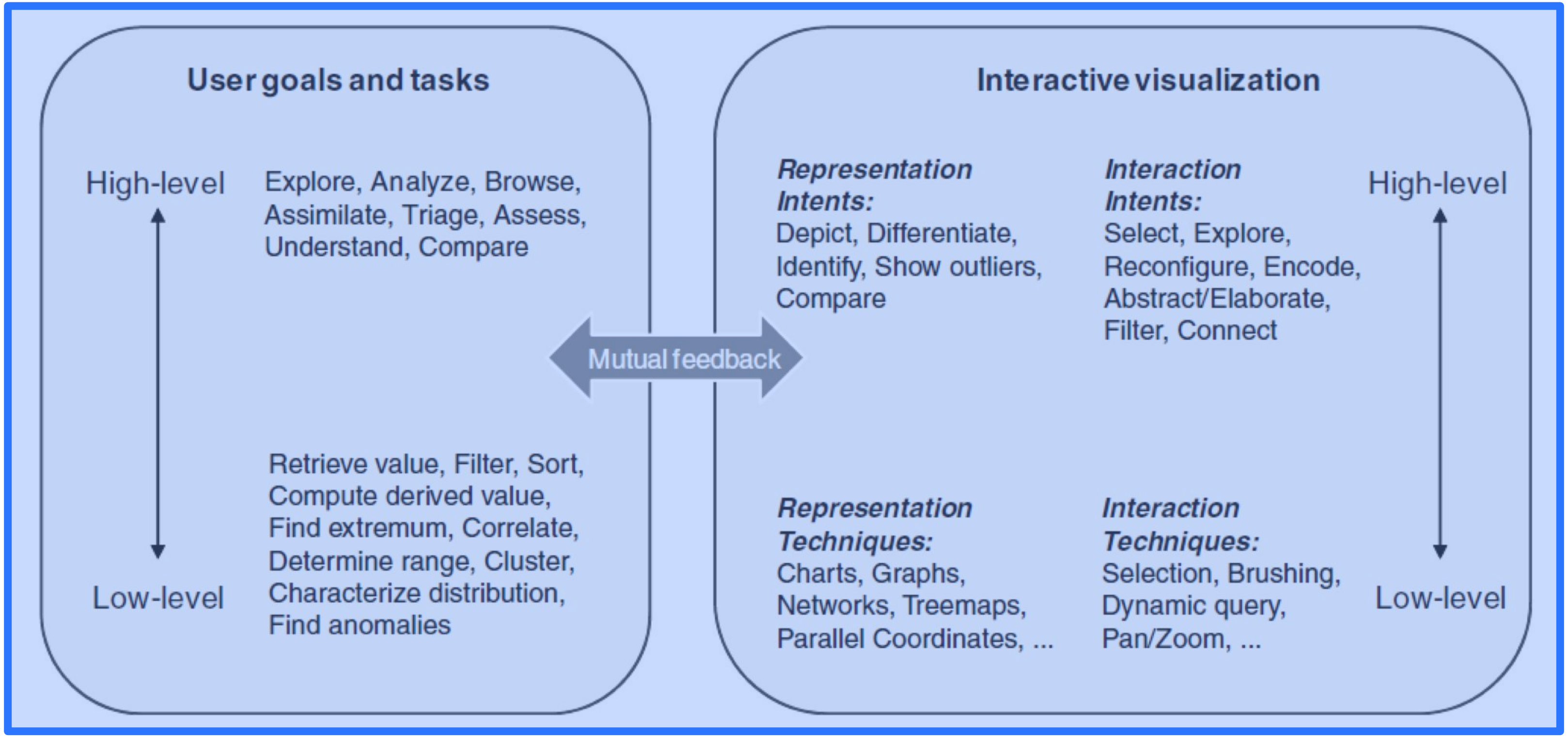
- Discern between the **machine model** as a component of the visual analytics system supporting analytics [Keim08], and the **user mental model of the data** [Pirolli05]
- Both are called just “model” in the literature unfortunately
- Careful about the **context**

Tasks, interactions and sensemaking



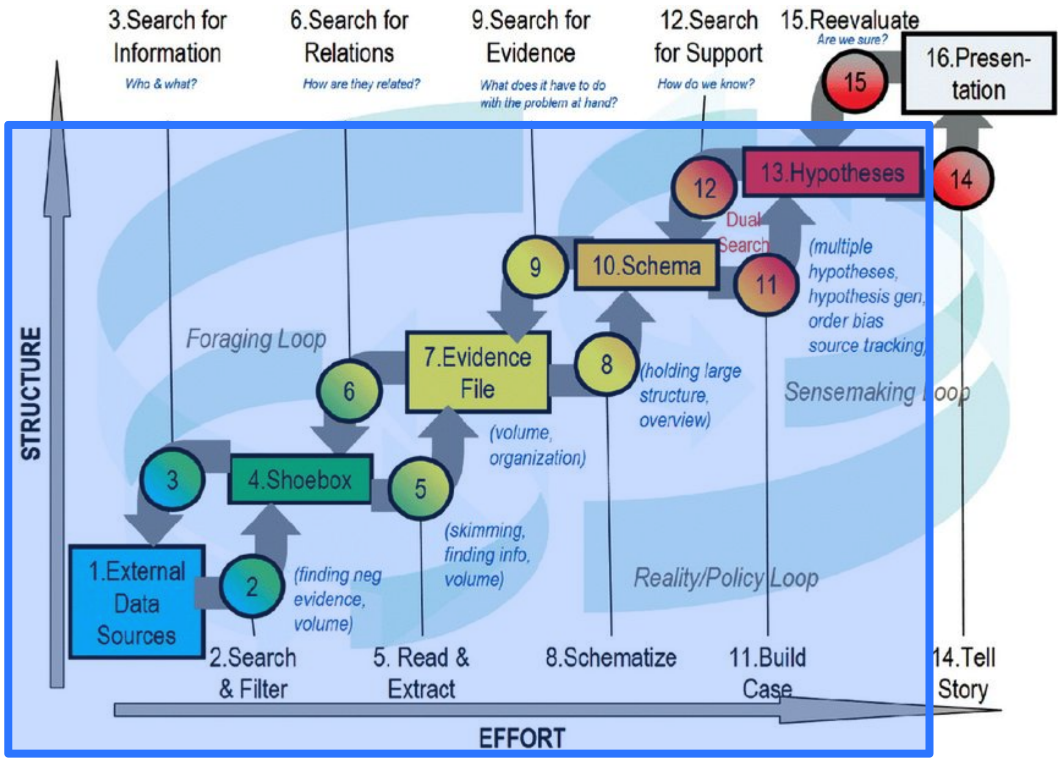
Which **components of the science of interaction** [Pike09] are relevant for visual analytics?

Tasks, interactions and sensemaking



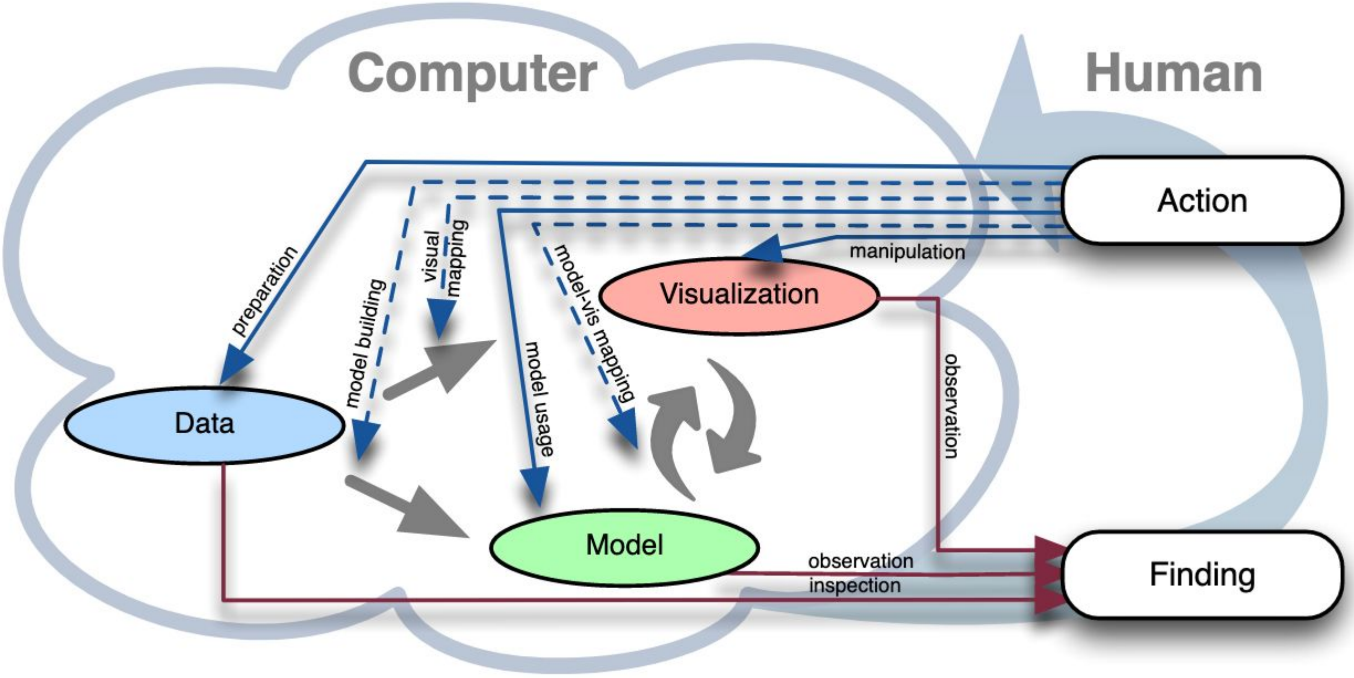
All of them...

Tasks, interactions, and sensemaking



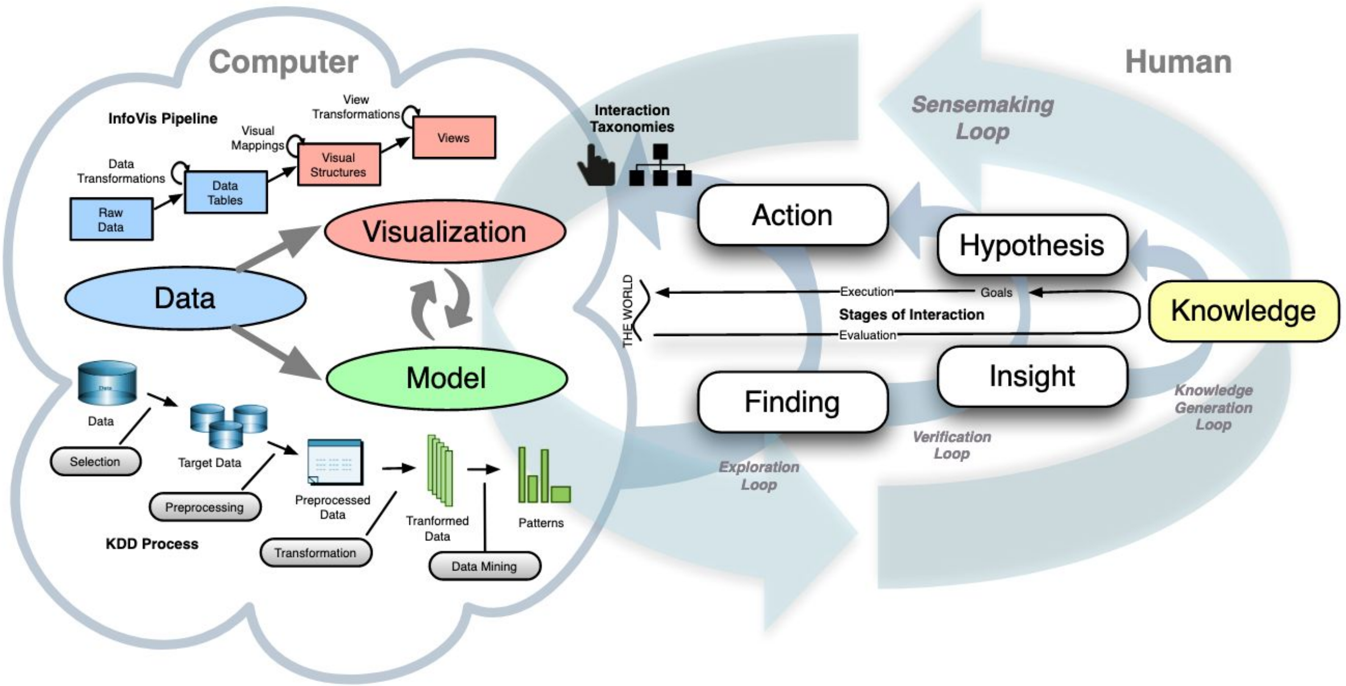
… because **visual analytics** needs to **support all this**

VA pipeline + human: Expanding "Knowledge"



[Sacha14]

VA pipeline + human: Fully expanded



A comprehensive VA model integrating several core, previously isolated ones [Sacha14]

VA theory: Usefulness?

- So, **visual analytics theory = adding more and more arrows to InfoVis** with each successive paper?
  - Certainly seemed that way to me when I started delving into it myself
  - And I didn’t even know [Sacha14] back then…
- **Is it really useful?**
  - The users can keep going between phases as they please
  - Vague terms such as **insight** or **knowledge** that are difficult to quantify
  - The advice seems to be vague as well: “do whatever the user might need, and connect everything with everything”

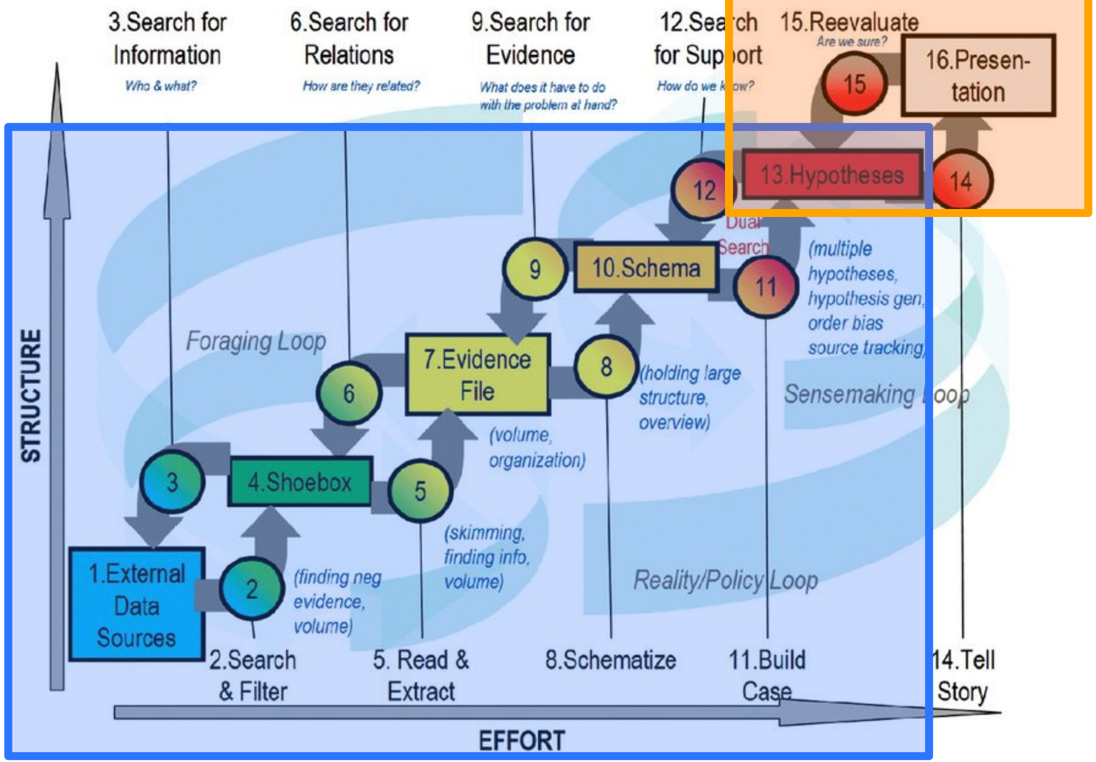
VA theory: Usefulness

- I’d say **it is useful**, even if it might need a second glance
- Gets you into the right **design mindset**
- Visual analytics indeed is **highly multidisciplinary**, involves (elements of):
  - InfoVis to design nice visualizations
  - Data science to design models that support analytics
  - Software engineering: getting good reqs from the user, VA systems are complex code-wise
  - High-performance computing when tackling large datasets
  - Empathy & communication: the ability to think like the users and empower them with analytics in their own domain

VA theory: Design takeaways

- In VA, **“making challenging data accessible”** is a perfectly valid objective
- Contrast with pure InfoVis: would be a **Q-type error** there (trifecta checkup)
  - Because your visualization doesn’t make a point then, it just shows the data
  - Visual analytics is all about **allowing the users** (analyst) **to come to correct points on their own**, we’re not telling a story
- If you still want to use trifecta in visual analytics, I’d say **Q in VA means “sufficient interactive support for meaningful high-level tasks”**

VA theory: Usefulness

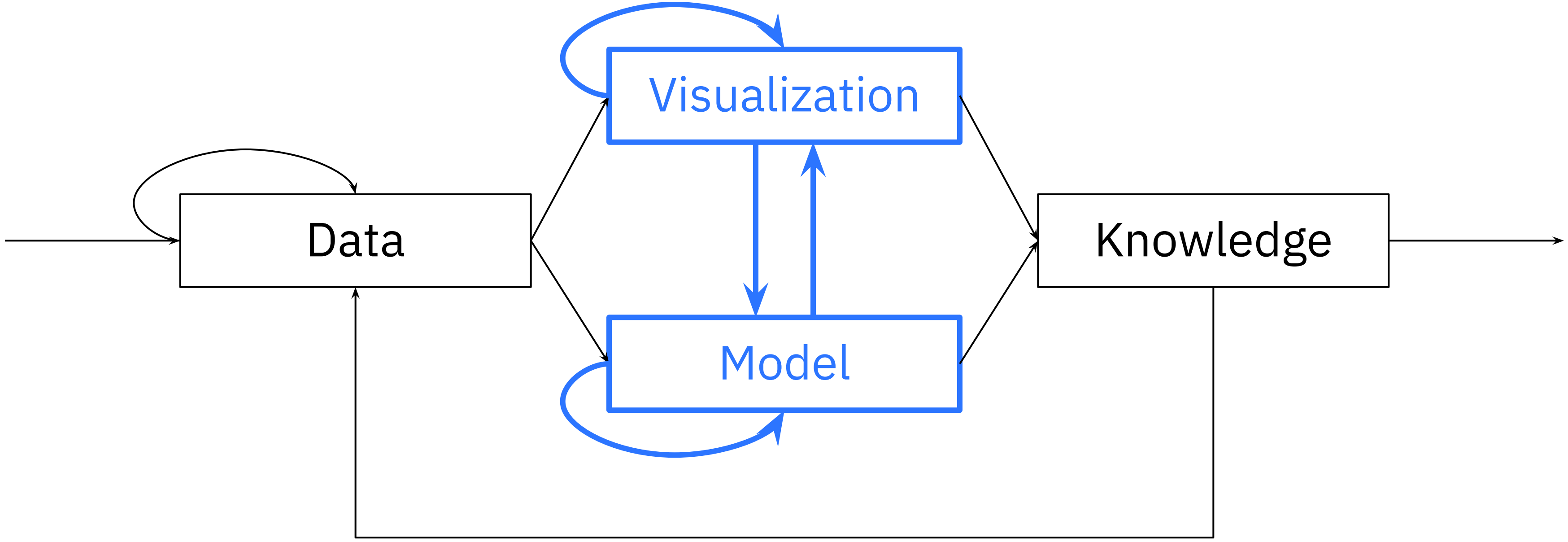


Visual analytics vs. **the final visualization** (simplified, there are analytic visualizations too)

VA theory: Design takeaways

- Fully supporting all the connections is **very challenging**
  - This is a **widely-recognized** challenge by the community
  - Not all VA systems support everything, not even all **successful** VA systems
- Visual analytics systems are **domain-specific** and **task-specific**
  - There is no single VA system best across all domains that work with data
  - Also, none really supports all fathomable high-level analytic tasks
  - **Tableau** is the closest to a “general VA system”, but even that is not the standard
- Both these aspects **simplify the problem**

VA theory: Design takeaways



The **blue components** are core to any true VA system

VA theory: Design takeaways

- We already know how to create **interactive, multi-view visualizations** – and that's exactly the "Visualization" in the VA pipeline
- We also know how to **evaluate visualization**, and that theory applies to VA too
  - **Insight-based evaluation** especially useful!
  - Remember: **Q in trifecta checkup** becomes "did we support the user in formulating and supporting their own hypotheses?"

VA theory: Design takeaways

- **Sensemaking** [Pirolli05] is a useful decomposition of different stages from raw data to crisp hypotheses
    - Helps identifying the key high- and low-level tasks
    - And designing the interactions accordingly
- The right side of the **fully expanded VA pipeline** [Sacha14] conceptualizes **user behaviour**
    - You can – and should – “roleplay” as the user throughout all stages of design
    - This version of the pipeline gives you a schema for that

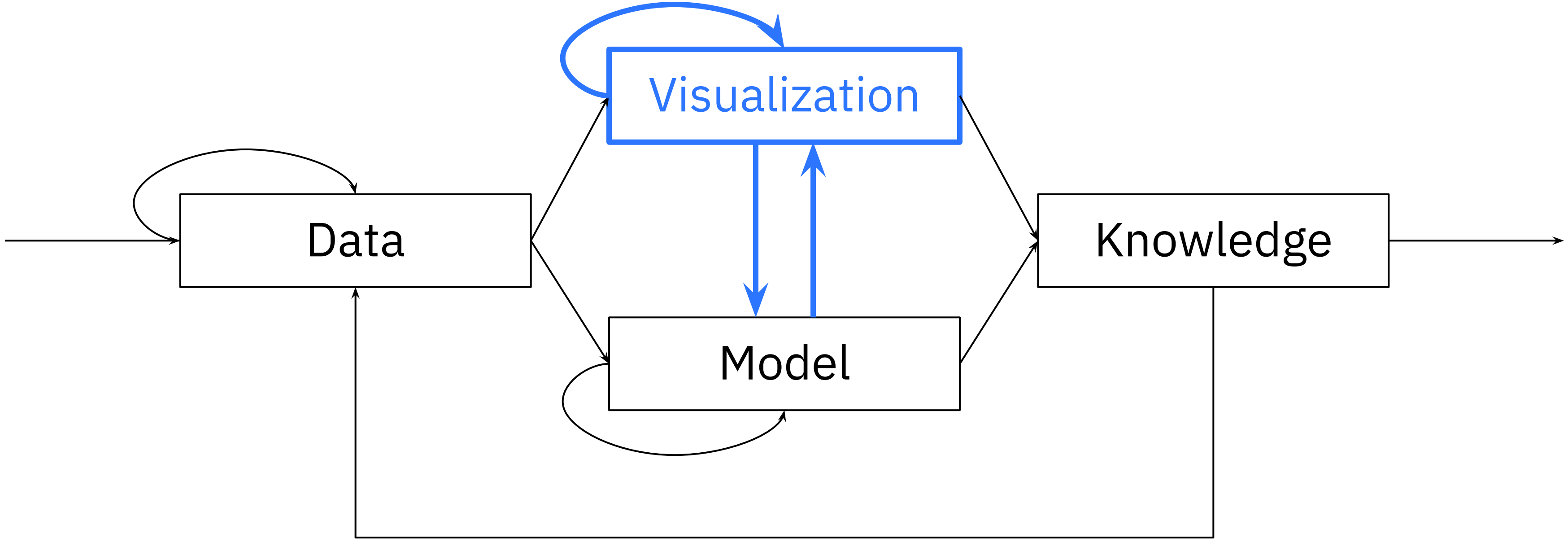
Visual analytics model

- How do we go about supporting an interactive visualization with a **model**?
  - The key missing ingredient so far

Integrating visualization and model

- The **model** should be **designed alongside the visualization**
  - Think of the **high-level task(s)** and your **data**
  - Design the visualization & interactions
  - If done systematically, this defines the **interface between visualization and model** and the requirements for it
  - Design the model
  - **Iterate** on this as you progress from design to implementation – waterfall planning never works

Integrating visualization and model



Systematic, thoughtful, and iterative design of visualization & interaction takes care of **this**

Visual analytics: Good software practices

- Shift the heaviest load to **preprocessing** (the loop from “Data” to itself), construct the model to maximize **lookup** operations
  - These don’t hamper interactivity
- **Model in the backend** – ideally all data ops should be performed here, with efficient communication with the frontend
- **Visualization in the frontend** – Visualization just shows the data digested by the model, plus rudimentary interactions
- **Keep the state** of the system as synchronized as possible
  - You’ll save yourself a lot of headaches

Model: Supporting interaction

- **Select** – Highlight the item(s) in the visualization and keep state (will be used in conjunction with the other interactions)
- **Explore** – Call on the **model** to show something else than what's on the screen/been seen in near past
  - In practice: the inverse of filter and/or random(ized) selection
- **Reconfigure** – In the visualization if trivial (just reshuffling the display), rely on the **model** to pull up data that are not on the screen

Model: Supporting interaction

- **Encode** – The **model** provides efficient data structure if the encoding is different, the visualization rerenders the data
- **Abstract/elaborate** – In all but trivial dataset cases, the zoom hierarchy must be precomputed and fetched from the **model**

Model: Supporting interaction

- **Filter** – Relies on an index which is again part of the **model**
  - For tabular data, a simple DB query might just do
- **Connect** – With good frontend design, can be taken care of (mostly) in the frontend
  - Views being able to access the selected data items from a frontend variable and not having to ask the model all the time what is selected
  - Efficient command to highlight specific IDs within the data structures across the views

Model: Supporting interaction

- 3 interactions rely on the model **heavily**, and present a **computational challenge** in a live user session (they hinge on dynamic user choice):
    - Explore
    - Filter
    - Abstract/elaborate
- 3 interactions need the model to supply **efficient data structures**:
    - Reconfigure
    - Encode
    - Connect
- The need for **tight integration between model and visualization** clear

VA Model: Modelling techniques

- A nice **survey** from a visual analytics perspective: [Endert17]
- Overviews all **key modelling** approaches beyond rudimentary statistical techniques
  - **Note**: The survey mentions “machine learning techniques”, but that is not a precise term.
  - Hence the term “modelling techniques” we use in the lecture

VA Model: Modelling technique categories

- **Modify parameters (MP)**
  - The user directly manipulates the model parameters through the visualization
  - The more populous category across all techniques
  - Pros: easier to implement, exact meaning
  - Cons: requires stats/machine learning knowledge from the user, non-intuitive
- **Define analytical expectations (DAE)**
  - The user interacts within the domain of expertise (using domain knowledge), the model behaves semantically: translating between the user's language and the ML/stats language
  - Fewer approaches exist
  - Pros: meaningful and intuitive to the user, no or little knowledge of stats/ML required
  - Cons: difficult to implement, knowledge gap between the developer and the user

[Endert17]

VA Model: Modelling technique table

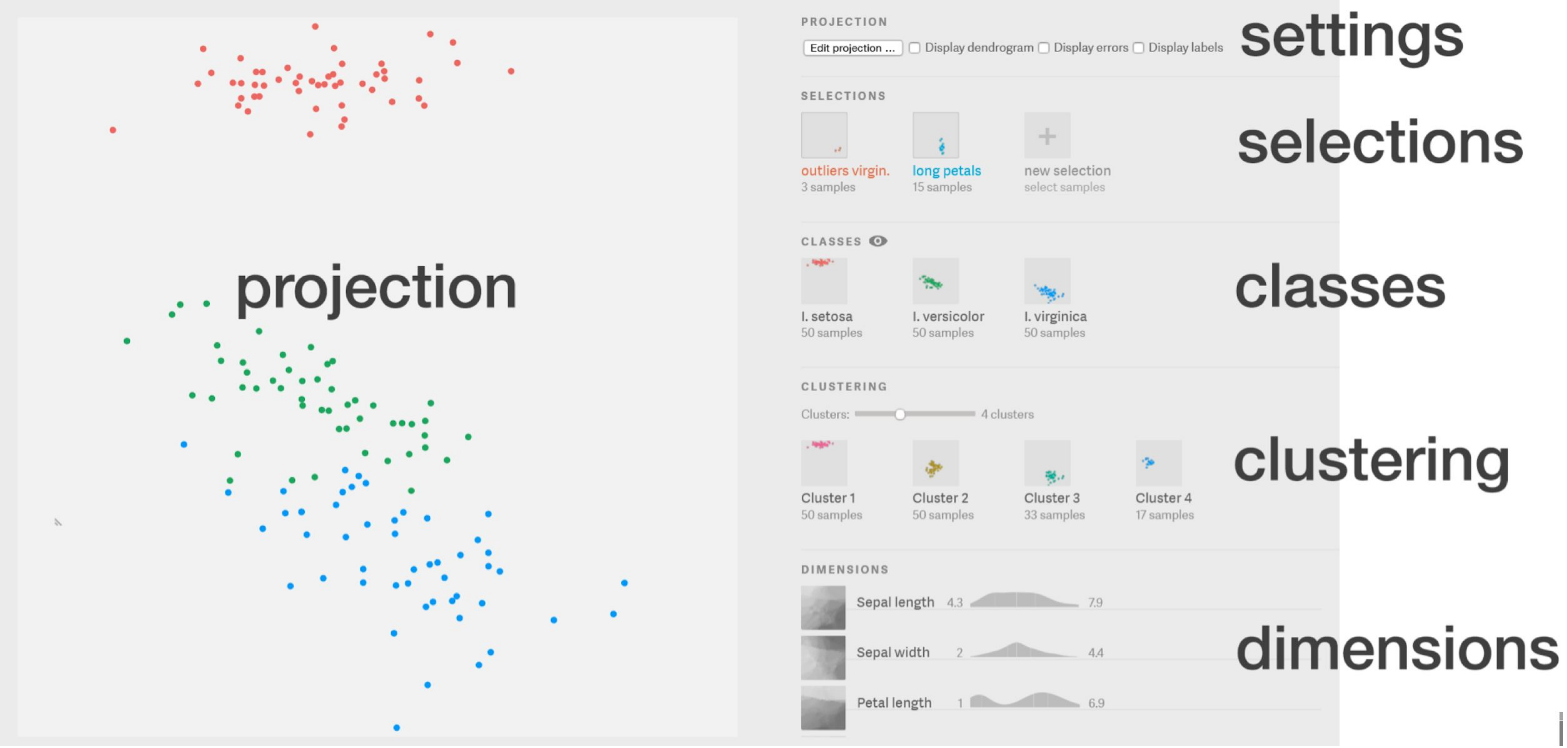
	Modify Parameters & Computation Domain	Define Analytical Expectations
Dimension Reduction	[JJ09], [FJA*11], [FWG09], [SDMT16], [WM04], [NM13], [TFH11], [TLLH12], [JBS08], [ADT*13], [JZF*09]	[EHM*11], [EBN13], [BLBC12], [HBM*13], [GNRM08], [IHG13], [KP11], [PZS*15], [KCPE16], [KKW*16]
Clustering	[Kan12], [RPN*08], [SBTK08], [RK04], [SS02], [LSS*12], [LSP*10], [TLS*14], [TPRH11a], [AW12], [RPN*08], [HSCW13], [TPRH11b], [PTRV13], [HHE*13], [WTP*99], [YNM*13], [SGG*14]	[HOG*12], [CP13], [BDW08], [CCM08], [BBM04], [ABV14], [KKP05], [KK08]
Classification	[PES*06], [MK08], [MBD*11], [vdEvW11], [CLKP10], [KPB14], [AAB*10], [AAR*09], [KGL*15]	[Set09], [SK10], [BKSS14], [PSPM15]
Regression	[PBK10], [MP13], [MME*12], [TLLH12], [KLG*16]	[MGJH08], [MGS*14] [LKT*14] [YKJ16]

[Endert17]

VA Model: Modelling techniques

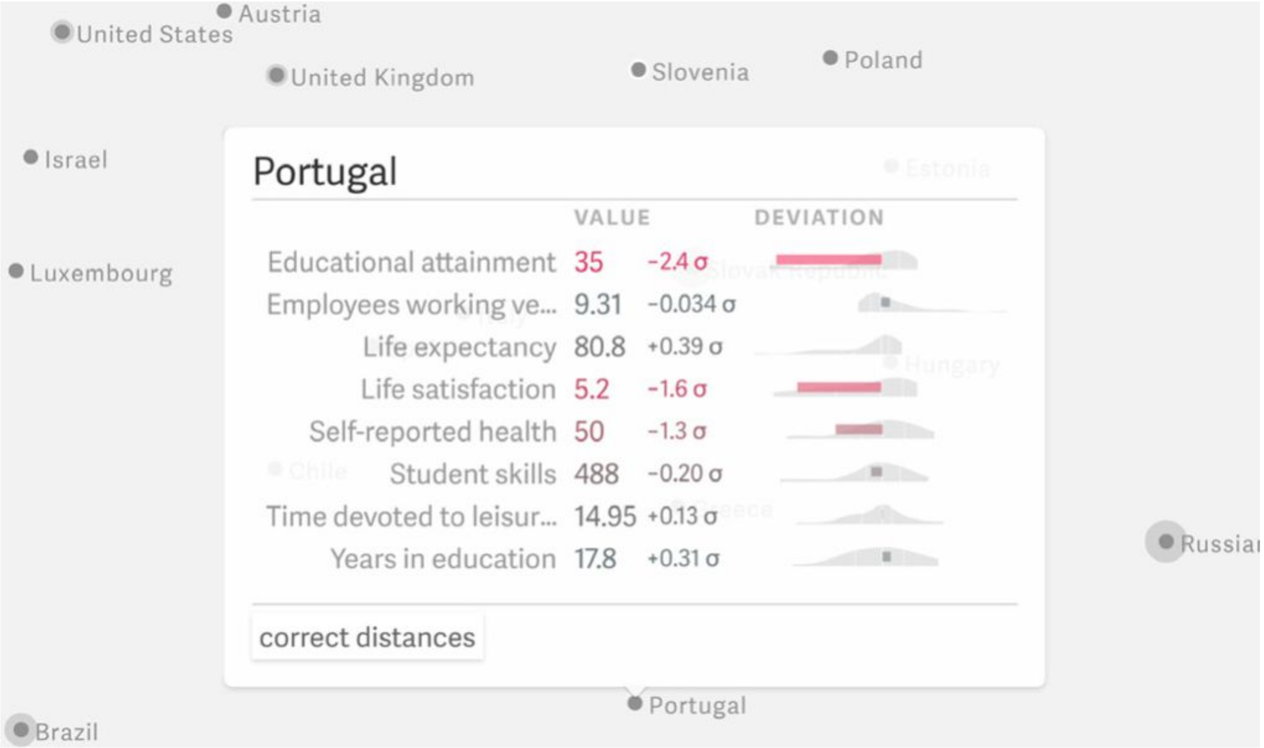
- **Dimensionality reduction**
  - Motivated in last lecture: enables visualization of n-D data where $n > 3$
  - PCA, MDS, ISOMAP, (t-)SNE, UMAP
  - Approaches exist for both categories (MP & DAE)
  - Unfortunately, no interactive methods for the top dim-reduction performers (t-SNE, UMAP)
- **Clustering**
  - Unsupervised learning, automatically find groups of data items close to each other (clusters)
  - k-means, spectral clustering...
  - The most populated out of the model categories (possibly due to the base algorithms being quite mature)
  - Clustering on the whole, while still very useful, is being overtaken by the modern dim-reduction methods in general ML applications

Dimensionality reduction (MP): Example



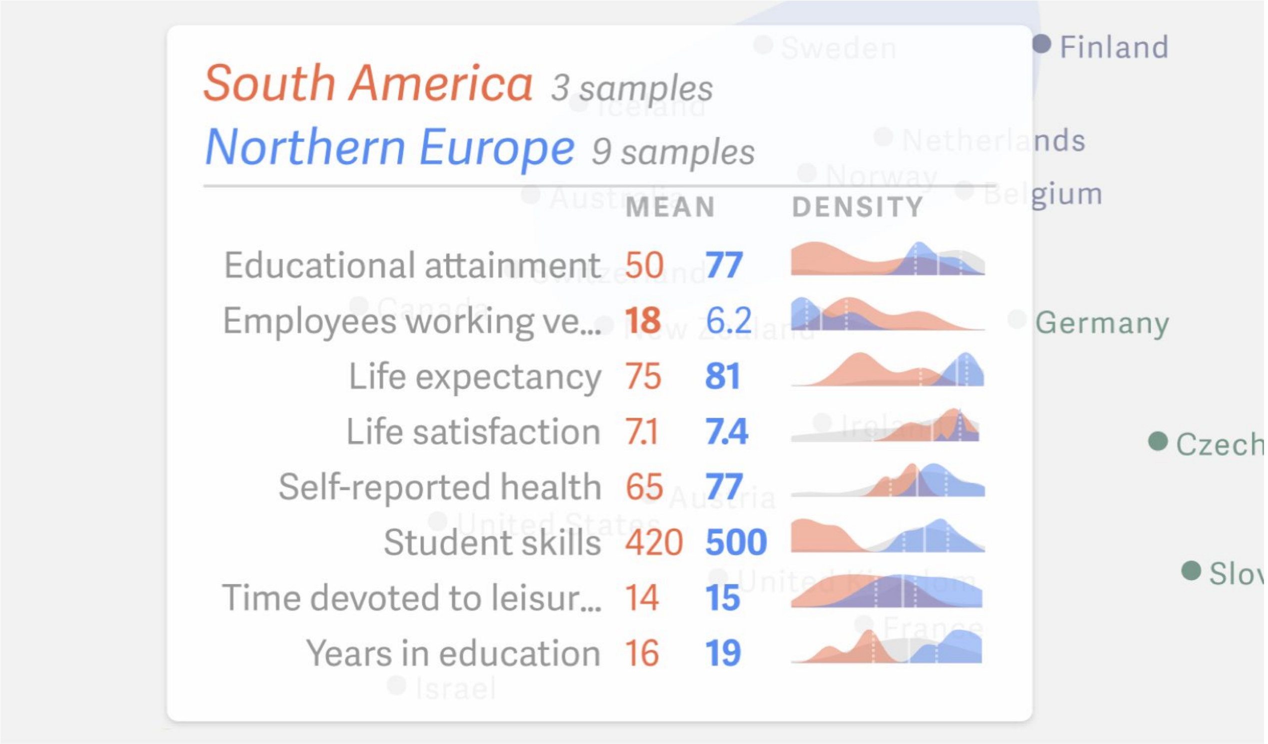
Multiview: central view with the projection, side panels for control

Dimensionality reduction (MP): Example



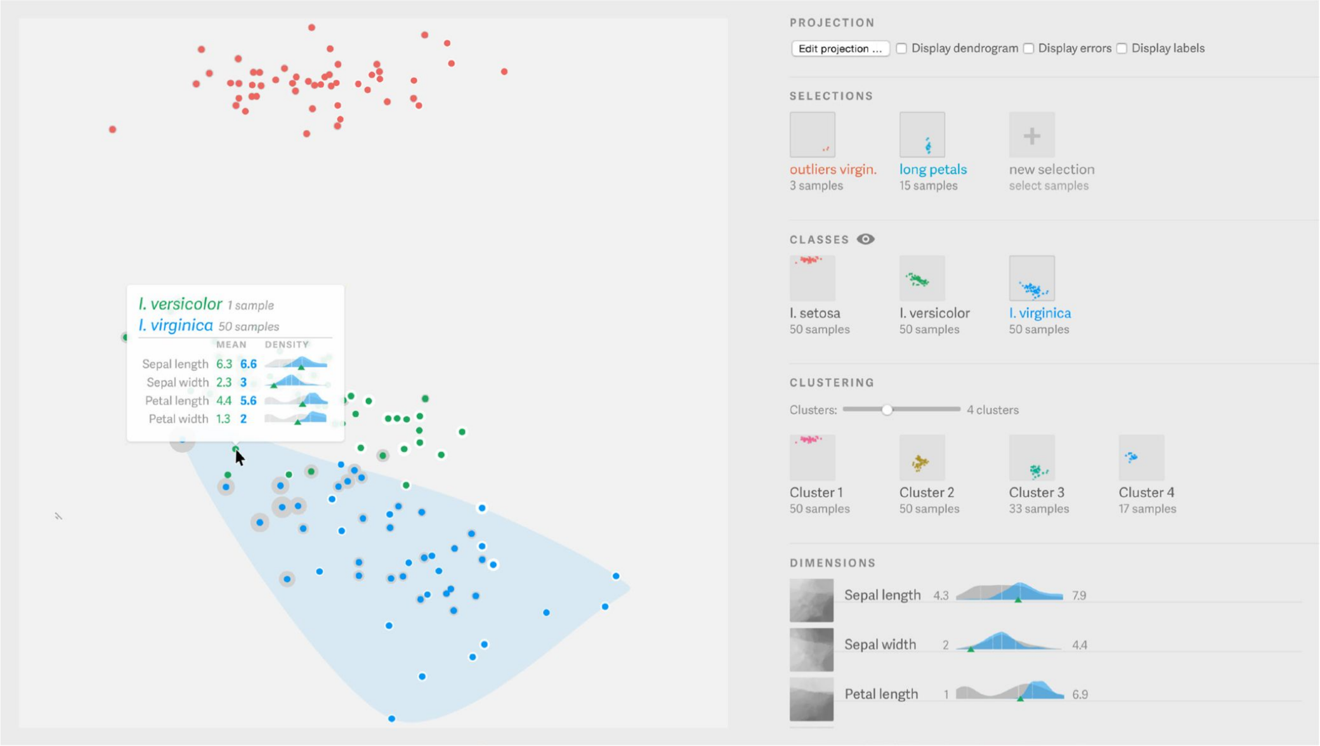
Tooltip: Statistical summary of samples in a category

Dimensionality reduction (MP): Example



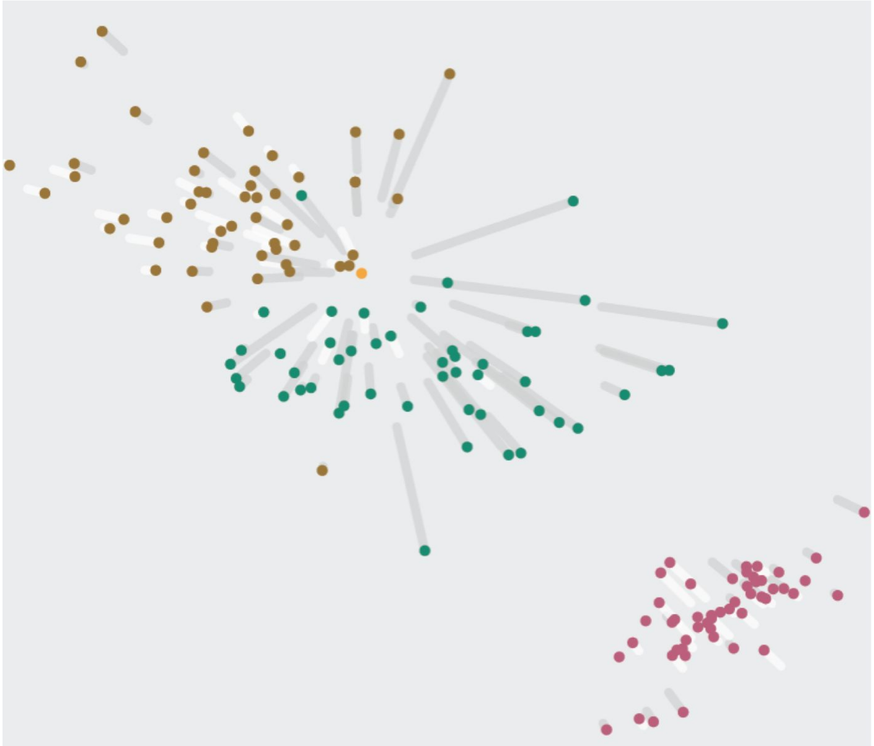
Select 2 groups to compare them visually

Dimensionality reduction (MP): Example



Comparing a sample with class. Shows value, distribution, and distortions (grey = close, white = far)

Dimensionality reduction (MP): Example

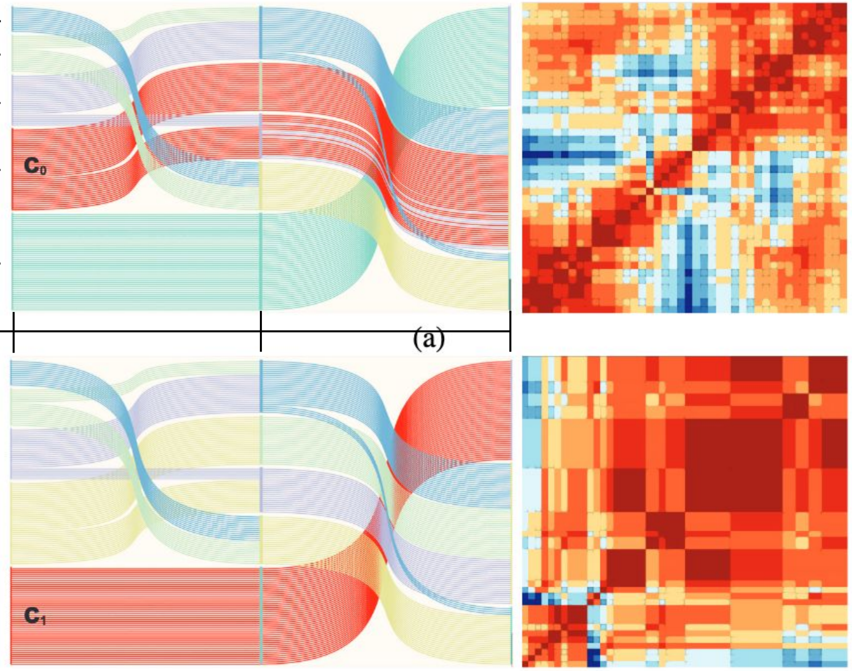


Projection errors corrected for the orange sample: grey trace: farther in high-dim space, white: closer

Clustering (DAE): Example

Clusters

Clustering 1, 2, 3



(a)

(b)

“The blue cluster and half of the green cluster from clustering 1 is merged into the yellow cluster in clustering 2.”

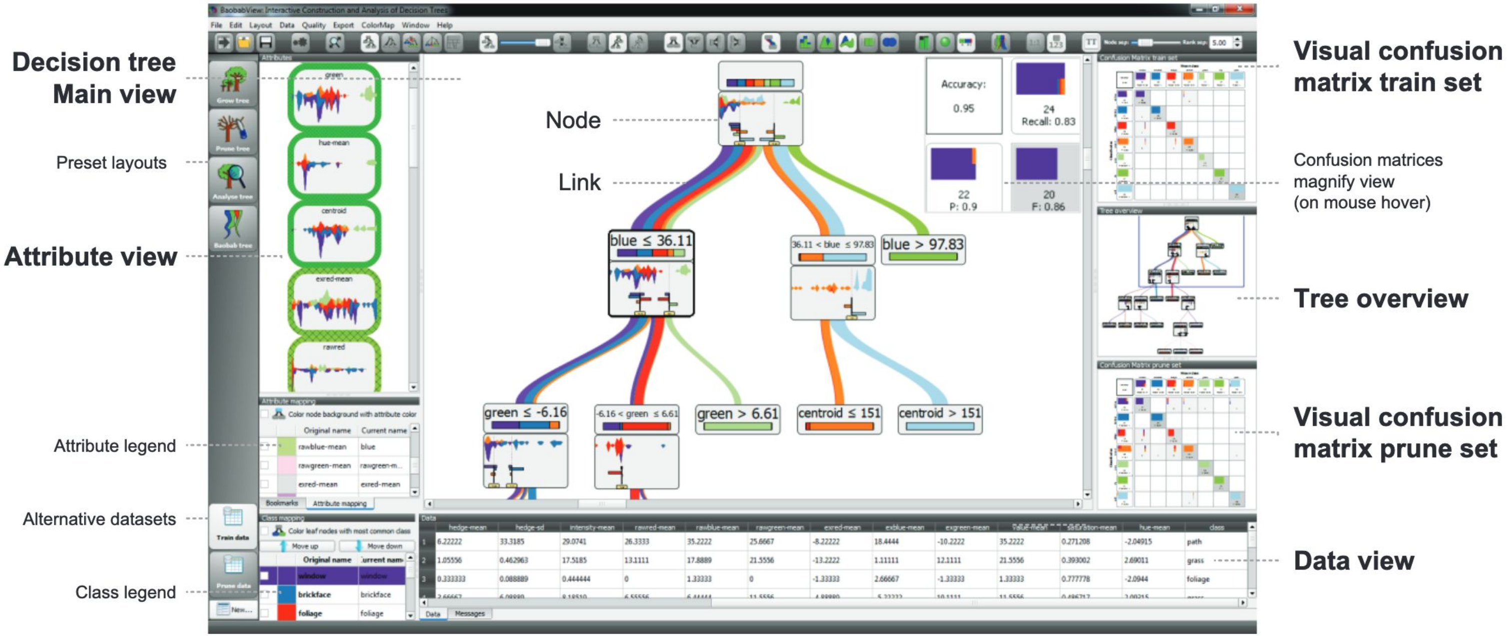
Cluster heatmaps
The redder, the closer the points are to each other, the bluer, the more distant. Red rectangles surrounded by blue around the diagonal = strong clusters.

[Turkay11]

VA Model: Modelling techniques

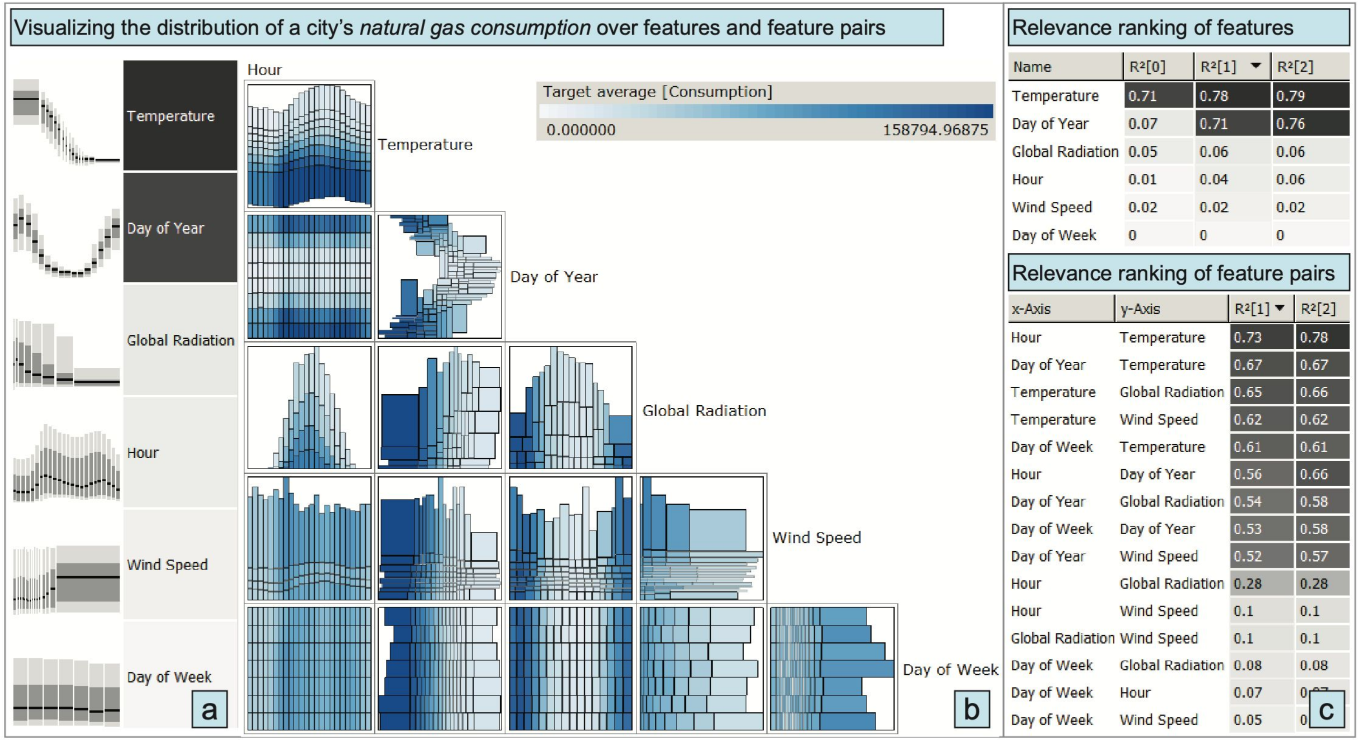
- **Classification**
  - Supervised learning: Data instances belong to categories called classes, the ML model tries to learn these classes. Then it is able to assign an unlabelled data instance to correct class
  - MP: prevalent in VA, techniques to construct classifiers in the UI, which then shows how well the data is categorized
  - DAE: again a smaller group, despite very good support for this in ML theory: semi-supervised learning, interactive learning, relevance feedback, active learning
- **Regression**
  - Supervised learning: “Continuous classification” – we don’t predict a class label, but a continuous variable. Used also to fit a trend line through the data.
  - Again, techniques for both MP and DAE,

Classification (MP): Example



BaobabView [VanDenElzen11]

Regression (MP): Example



[Mühlbacher13], video: https://youtu.be/e88dMUbbSSw

VA modelling techniques: Design takeaways

- A **plethora of techniques**
- Adding interactivity and dynamics to modelling state of the art is **difficult** though
  - Almost all techniques are static, “precompute once”
  - For example, interactive deep nets still a very open challenge
  - ML researchers rarely think about interactivity “natively”
- That’s why the truly interactive techniques seem to “lag behind” the ML state of the art by ~3-5 years at least
  - The technique has to mature before it can be optimized
- You can still rely on **state of the art in the precomputing phase** and then add interactivity on top

No ML in the model = No VA?

- Is a **system without ML** in the model actually a **true VA system**?
  - E. g., how about Tableau? Isn't that just a multiview visualization, even if analytic?
- **It could be** – as long as it the model:
  - Drives the visualization and is driven by the visualization
  - Assists the users in **gaining understanding**, showing what's relevant to them at a given time
- Example: the dimensionality reduction (MP) example, slides 53 – 57
  - No ML, just statistical summaries, yet clearly supports analytics on the dim-reduced data

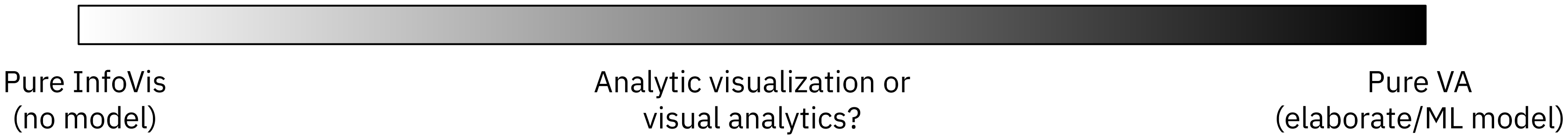
Pure InfoVis
(no model)

Analytic visualization or
visual analytics?

Pure VA
(elaborate/ML model)

No ML in the model = No VA?

- InfoVis & VA approaches seem to occupy a **continuous axis** between:
  - Pure InfoVis – no model involved, just a visualization
  - Pure VA – An elaborate model involving advanced techniques such as AI that clearly supports analytics
- **Gray zone** – is it an analytic visualization, or a visual analytics system?
  - A system with multiple connected data views, with solid interaction design, but light on the “backend calls”



No ML in the model = No VA?

- **Tough to decide** – no crisp, standard checklist to judge authoritatively
- Also, InfoVis + VA is **one community** scientifically
    - So the need for crisp boundaries is not very high
- My opinion: it's good that it's a continuous axis, allows for a wider palette of approaches with fewer formal exclusions of otherwise interesting ones
- **Tableau** is an example of a system in the gray zone

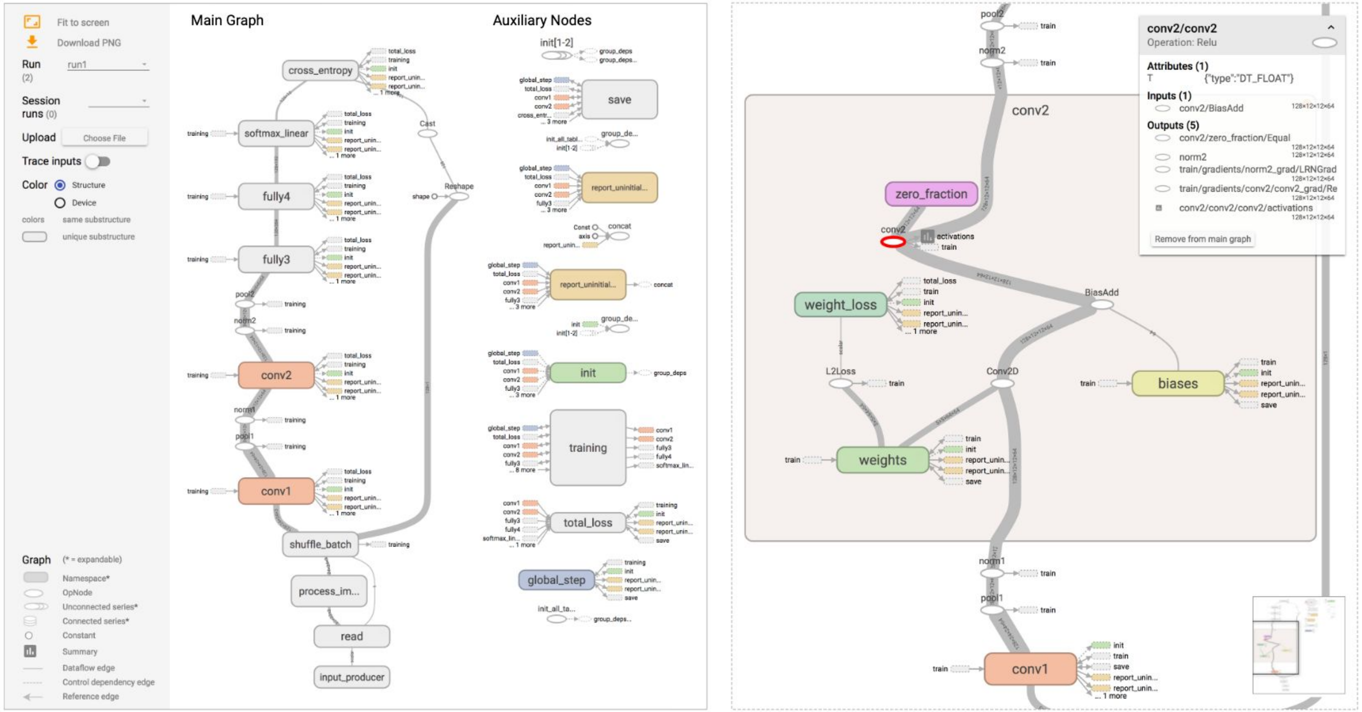
Pure InfoVis
(no model)

Analytic visualization or
visual analytics?

Pure VA
(elaborate/ML model)

Ω

Examples: VAST best paper 2017



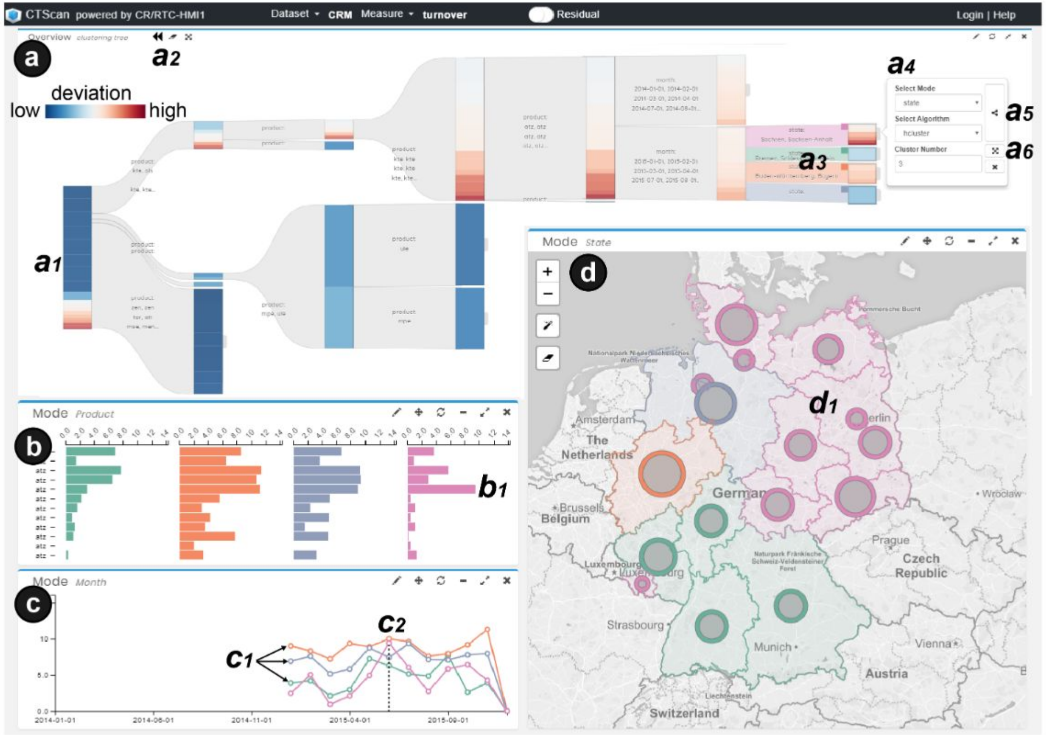
(a) (b)

TensorBoard [Wongsuphasawat18], video: https://vimeo.com/232930758

Examples: VAST best paper 2017

- More of a nice **visualization** than a true visual analytics system IMHO
  - Visualizes a deep net – that is a ML model, but from the PoV of the VA system, it's data
- Still:
  - Allows in-depth inspection of an arbitrary deep net
  - "Trace input" adds an analytics dimension to understand the model
- Nice example of multi-view **parsimony**

Examples: VAST best paper 2018

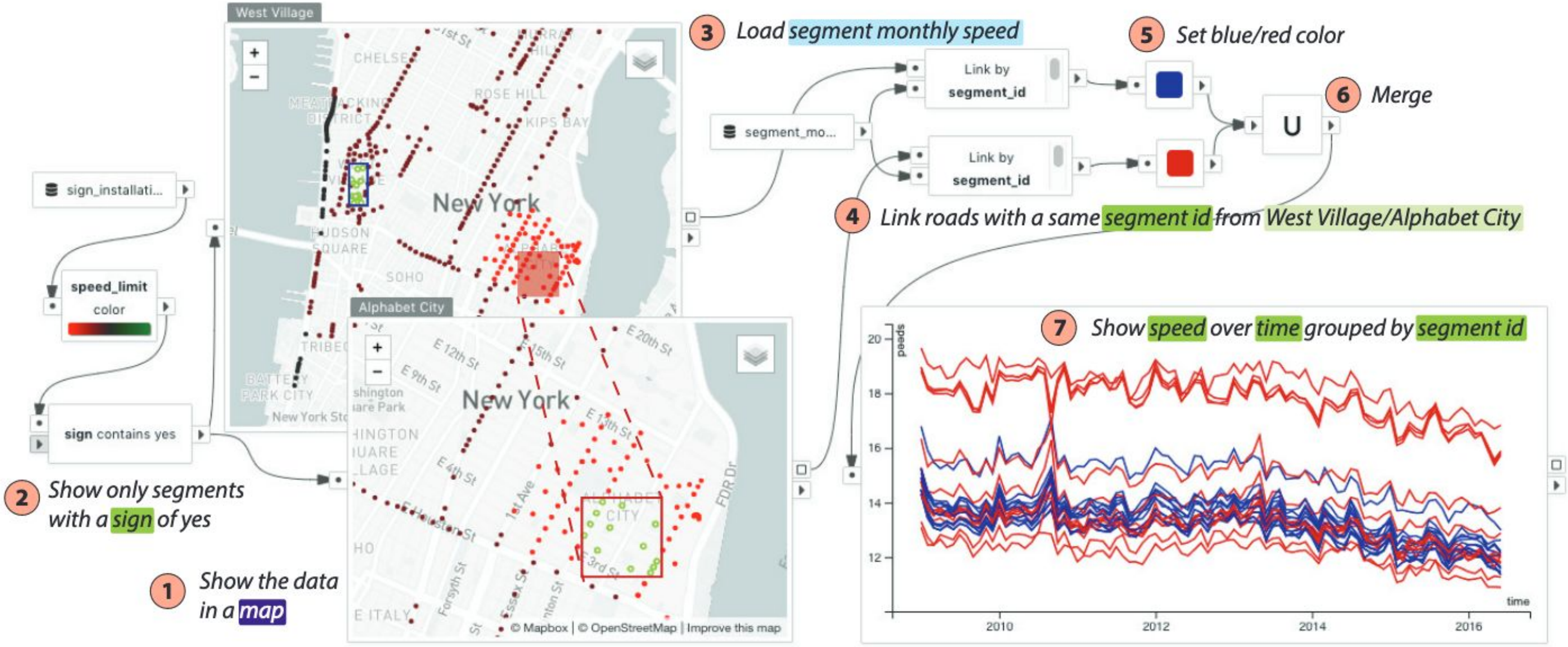


TPFlow [Liu19], video: https://youtu.be/oPZ1Xi-Ed6k

Examples: VAST best paper 2018

- A clever model: tensor-like processing of **spatiotemporal data**
- A masterclass in **multiview & interaction design**
  - Views make sense
  - They are well connected
  - Individual visualizations are appropriate
  - Packed with meaningful interaction

Examples: VAST best paper 2019



FlowSense [Yu20], teaser: https://vimeo.com/360154533

Examples: VAST best paper 2019

- The model is based on **natural language**
  - Write a query such as “draw mpg and cylinders in a scatterplot”
  - The model will parse the query and draw the plot
- Great technique against a cluttered UI
- Visualization: whatever the user wants it to be
  - The system makes sure to adhere to proper practices, such as labelling etc.

Examples: VAST best paper 2020

VATLD [Gou21], teaser: https://youtu.be/NmtAQBrSNrM

Examples: VAST best paper 2020

- VATLD = A **V**isual **A**nalytics system to assess, understand and improve **T**raffic **L**ight **D**etection
- Model: **representation learning** (extracts useful data semantics) + **semantic adversarial learning** (visual summarization)
- Good multiview design on top of the model, incl. **multimedia data** (images)

Conclusion

- Ingredients for a visual analytics system:
  - Fundamentally solid **visualization**
  - **Multiview design**
  - Meaningful support for **interactions**
  - A machine model that takes care of:
    - **Data representations** for the **visualizations**
    - Efficiently searchable/filterable representation(s) to support **filtering/exploring**
    - Hierarchical representation(s) for **(semantic) zooming**
    - Some/all of the above will highly probably require machine learning

Conclusion

- Try to support **all stages of sensemaking**
- Make the **model as transparent and understandable** for the users as possible
- Put yourself in the user’s shoes as you design
- [Endert17] provides a good overview of modelling techniques

References: Optional reading

- **[Endert17]** A. Endert et al.: The state of the art in integrating machine learning into visual analytics: Integrating machine learning into visual analytics. Computer Graphics Forum, 36 (4), March 2017.
- **[Fayyad96]** U. Fayyad et al.: From data mining to knowledge discovery in databases. AI Mag., vol. 17, pp. 37 – 54, 1996.
- **[Gou21]** L. Gou et al.: VATLD: A visual analytics system to assess, understand and improve traffic light detection. IEEE TVCG, 27 (2), pp. 261 – 271, February 2021.
- **[Keim08]** D. Keim et al.: Visual Analytics: Definition, Process, and Challenges. In Information Visualization, Lecture Notes in Computer Science, vol. 4950, Springer, Berlin.
- **[Liu19]** D. Liu et al.: TPFlow: Progressive partition and multidimensional pattern extraction for large-scale spatio-temporal data analysis. IEEE TVCG, 25 (1) pp. 1 – 11, January 2019.
- **[Mühlbacher13]** T. Mühlbacher and H. Piringer: A partition-based framework for building and validating regression models. IEEE TVCG, 19 (12), pp. 1962 – 1971, December 2013.
- **[Pike09]** W. A. Pike et al.: The science of interaction. Information Visualization, 8 (4), 2009.

Reference: Optional reading

- **[Pirolli05]** P. Pirolli and S. Card: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In Proc. Int Conf Int Analysis, January 2005.
- **[Sacha14]** D. Sacha et al.: Knowledge generation model for visual analytics. IEEE TVCG, 20 (12), pp. 1604 – 1613, December 2014.
- **[Turkay11]** C. Turkay et al.: Integrating cluster formation and cluster evaluation in interactive visual analysis. In Proc. Spring Conf on Computer Graphics, 2011.
- **[VanDenElzen11]** S. van den Elzen and J. J. van Wijk: BaobabView: Interactive construction and analysis of decision trees. In Proc. IEEE VAST, 2011.
- **[Wongsuphasawat18]** K. Wongsuphasawat et al.: Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. IEEE TVCG, 24 (1), pp. 1 – 12, January 2018.
- **[Yu20]** B. Yu and C. T. Silva: FlowSense: A natural language interface for visual data exploration within a dataflow system. IEEE TVCG, 26 (1), pp. 1 – 11, January 2020.