

Advanced Data Visualization

CS 6965

Spring 2018

Prof. Bei Wang Phillips

University of Utah



Lecture 09

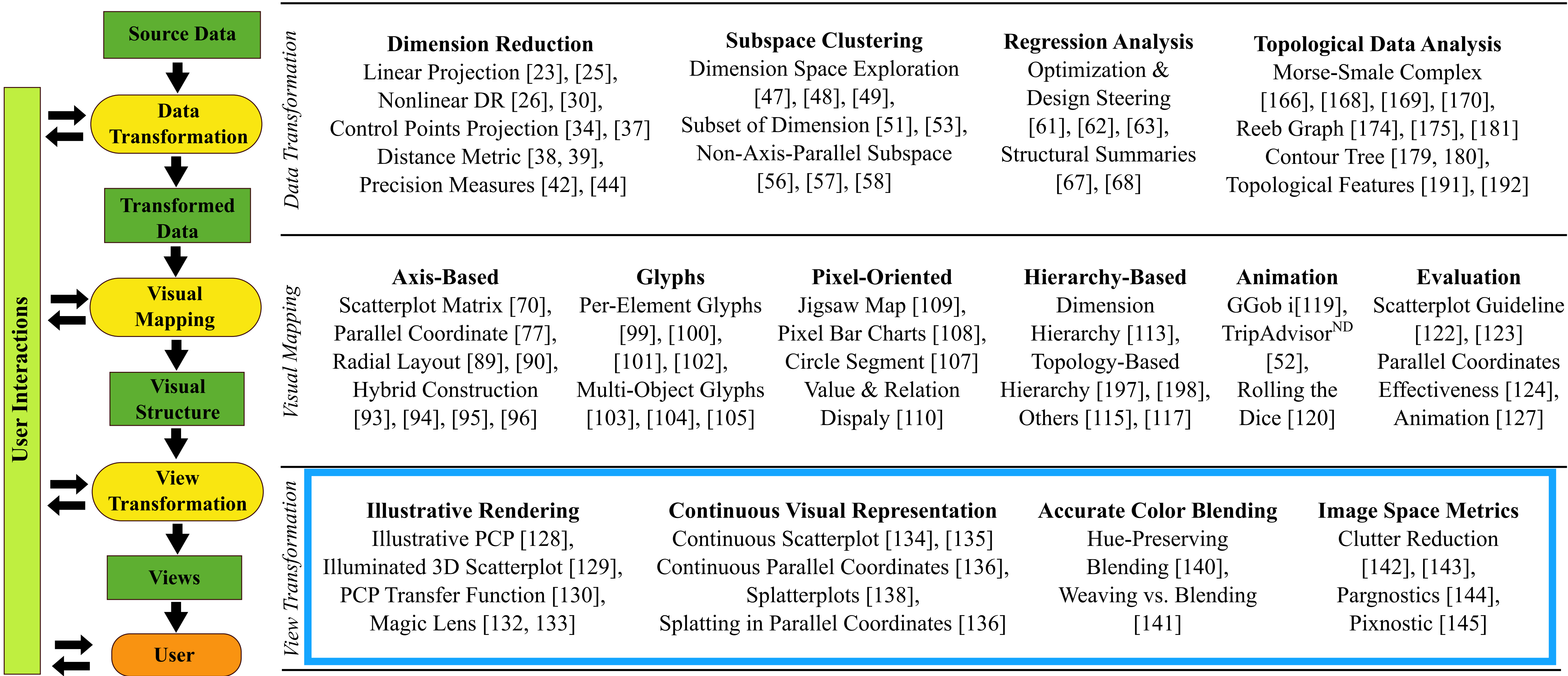
Project 2 Clarification

- Provide at least minimal analysis and visualization capabilities of the given data set
- Think about what visual mapping/view transformation you might employ (base line, map view; others, scatter plots, PCP, etc...)
- To challenge yourself: interactivity? (at least given examples of what interaction you might design)

View Transformation User Interaction Decision Tree



HD



Visualization pipeline for HD data

View Transformation

The rendering process that generates images in the screen space...
dictate what we ultimately see on the screen...

View Transformation

- Illustrative rendering: achieving a specific visual style by applying custom rendering algorithms
- Continuous visual representation: limitations for discrete representations — visual clutter and computational cost
- Accurate color blending

Illustrative Rendering

Illustrative PCP

- achieving a specific visual style by applying custom rendering algorithms

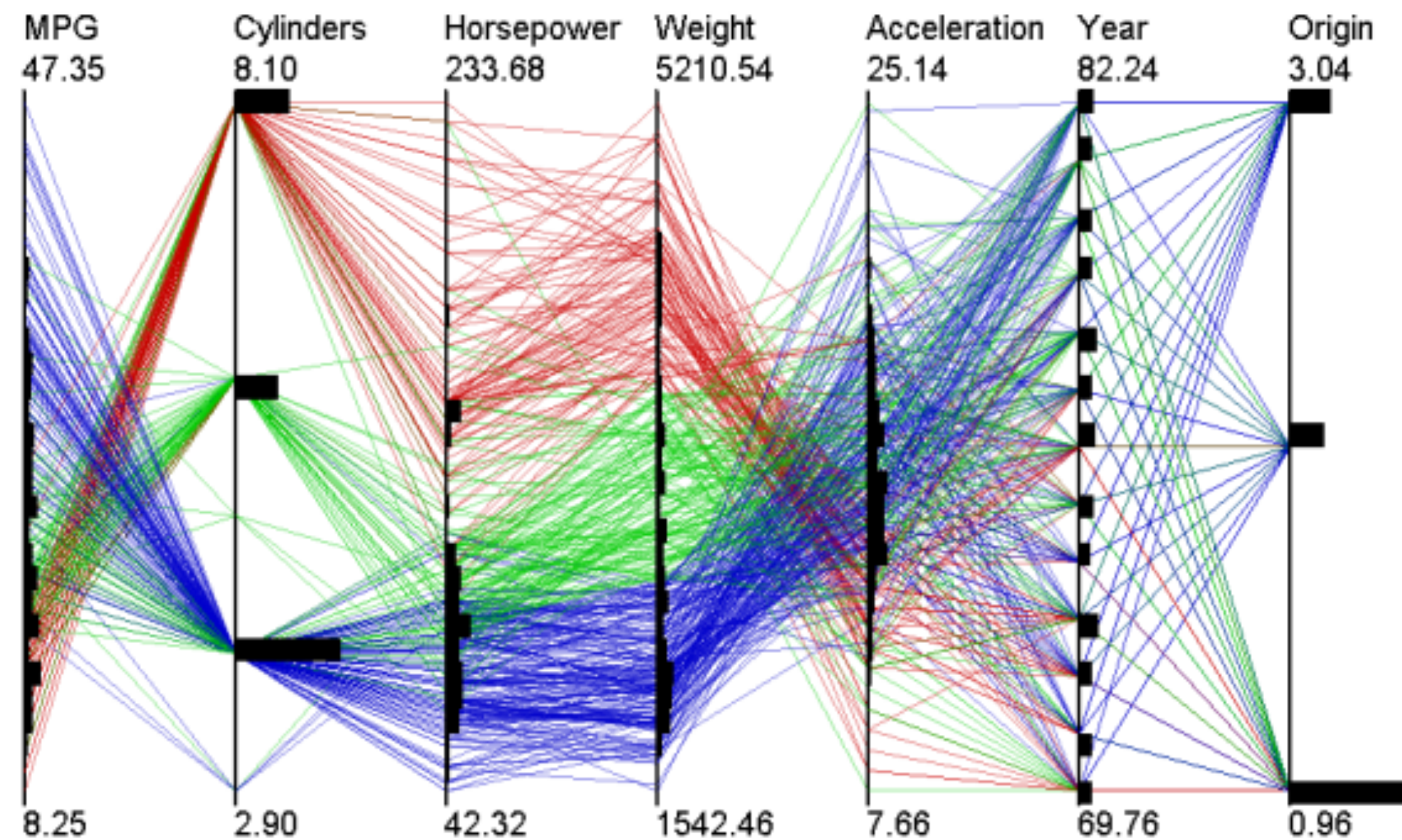
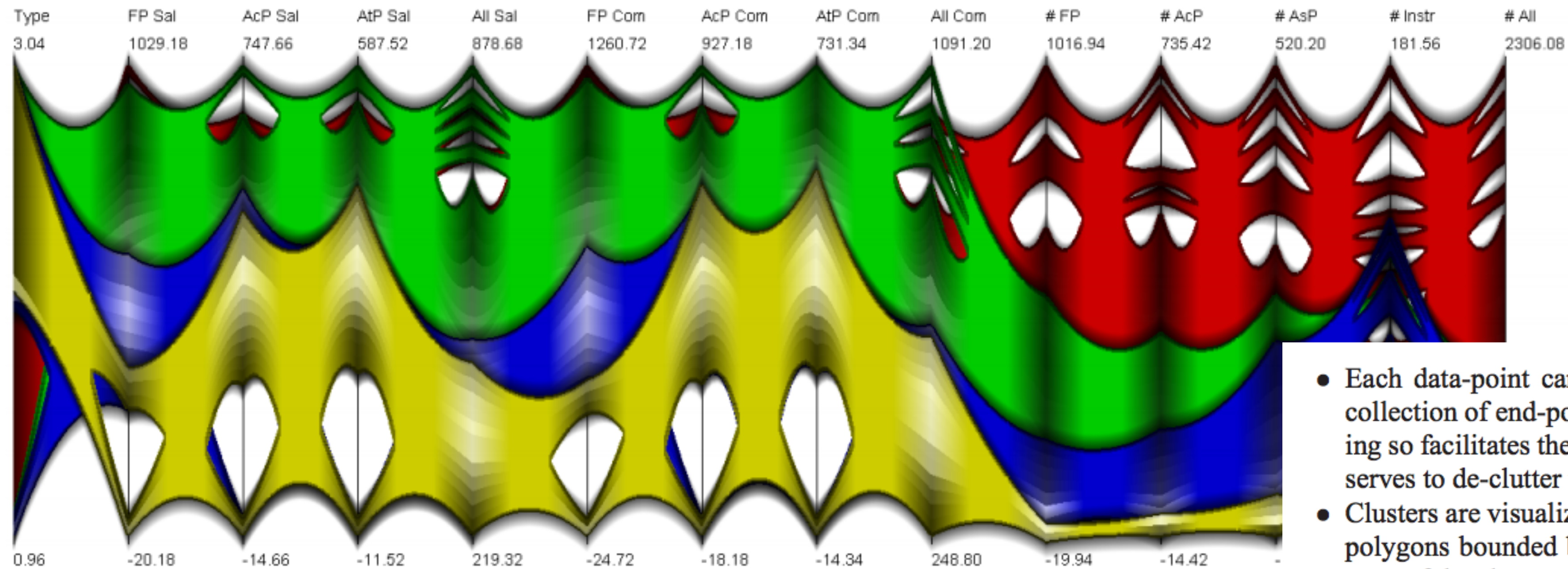


Figure 1: *Traditional parallel coordinates visualization (color-coded by cluster) of the 392-point, seven-dimensional “cars” dataset. Point distributions along axes are given by histogram bars. All datasets visualized in this paper are courtesy of the XmdvTool home page (davis.wpi.edu/~xmdv).*



- Each data-point can be rendered as a polycurve, i.e., a collection of end-point interpolating B-spline curves. Doing so facilitates the creation of edge bundles [Hol06] and serves to de-clutter the visualization.
- Clusters are visualized as a collection of semi-transparent polygons bounded by spline curves, which show the extents of the clusters and which can be scaled to control the screen area they consume. Higher cluster opacities correspond with clusters containing more points.
- The distribution of the data can be viewed at different levels of detail by displaying the clusters in a branched, tree-like manner.
- A density plot that conveys the distribution of the lines or curves between axes can be used to show correlations between axes.
- The distributions along individual axes are shown as faded quadrilateral strips and provide per-cluster histograms of the dataset for each dimension.
- Silhouettes, shadows and halos not only assist the eye in distinguishing between overlapping clusters, but also provide an interesting artistic effect.

Figure 2: An example of illustrative parallel coordinates (IPC) showing some of the major features of the approach. Shading and opacity are used extensively in IPC to convey information.

Illustrative PCP

Enhancing visual patterns in PCP, e.g. line density, etc.

Illustrative Scatterplots

- Classify points based on the Eigen analysis of the covariance matrix
- Give the user the opportunity to see effects such as planarity and linearity when visualizing dense scatterplots

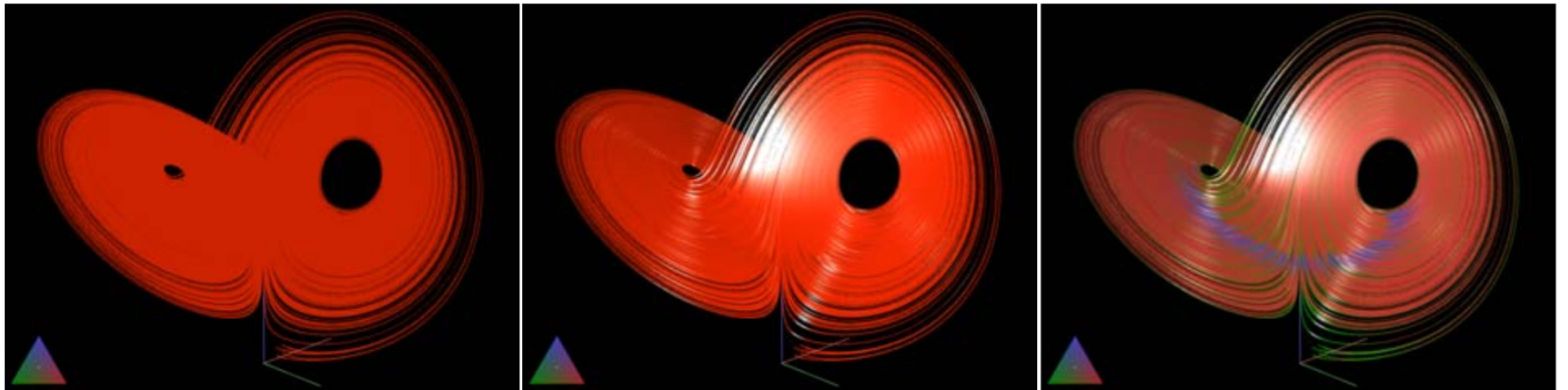


Figure 1: A Lorenz attractor. Left: traditional 3D scatterplot; middle: illuminated scatterplot; right: linear, planar, and spherical structures highlighted through mapping to green, red, and blue colors respectively. The base colors are chosen to have equal intensity.

Focal area highlighting: TableLens

- Visualizing large tables and reduce clutter in the views

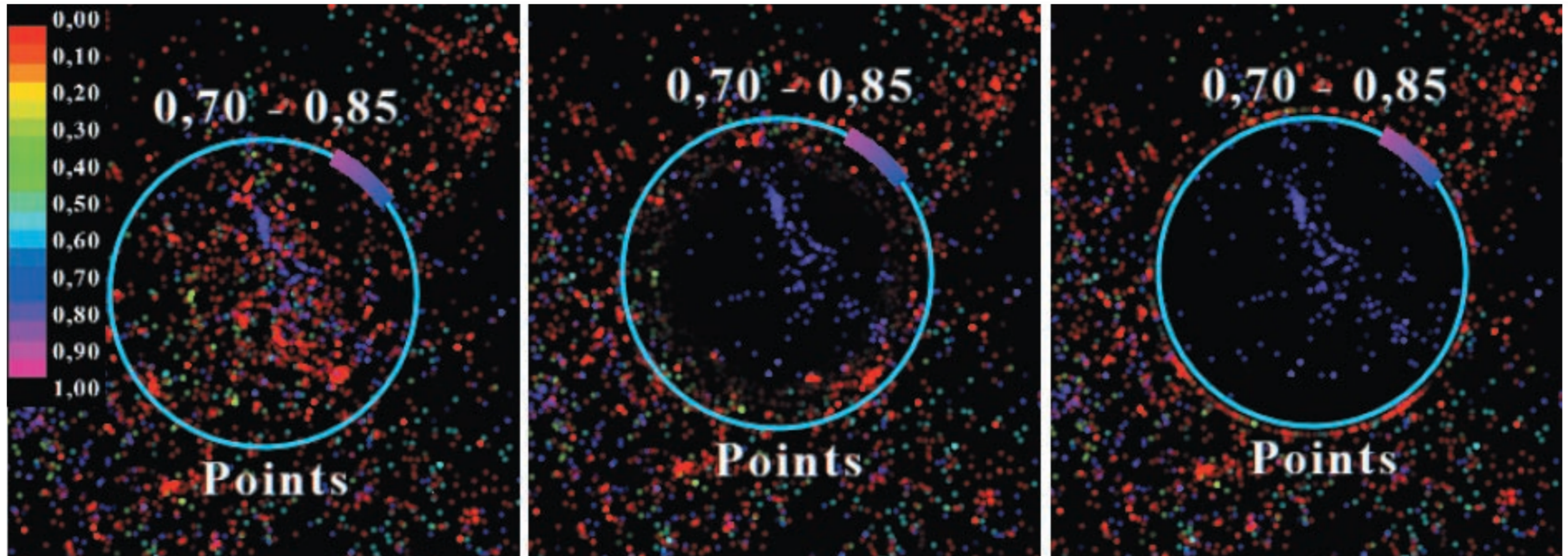
		G					H				
	4					G4		H4			
	5					G5		H5			
	6					G6		H6			

Figure 1: The Table Lens Focal Technique.

Focal area highlighting: MoleView

- MoleView: visualizing scatterplots and graphs
- Using semantic lens to focus on area of interest
- Keeping the in-focused data unchanged while simplifying or deforming the rest of the data to maintain context

HurterTeleaErsoy2011



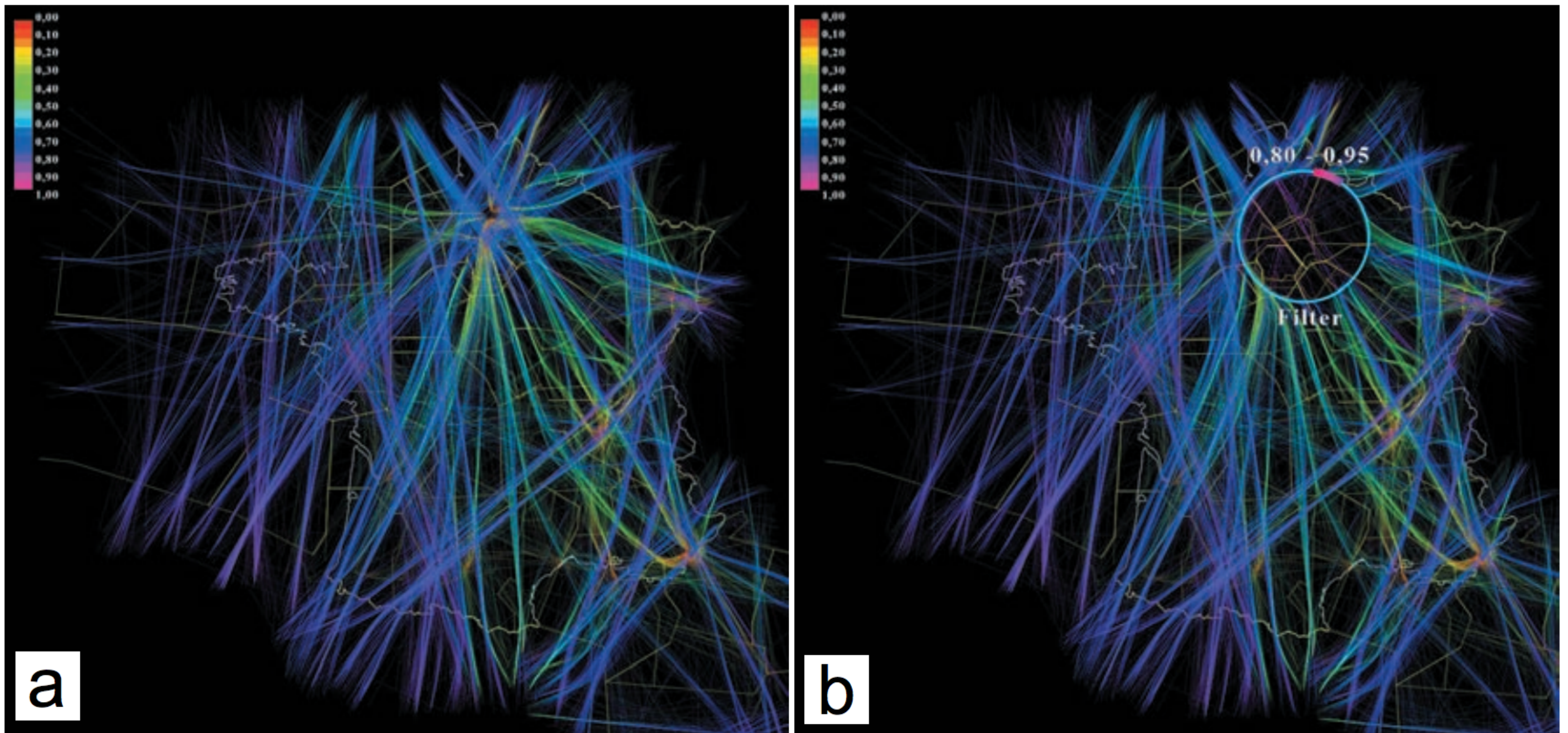


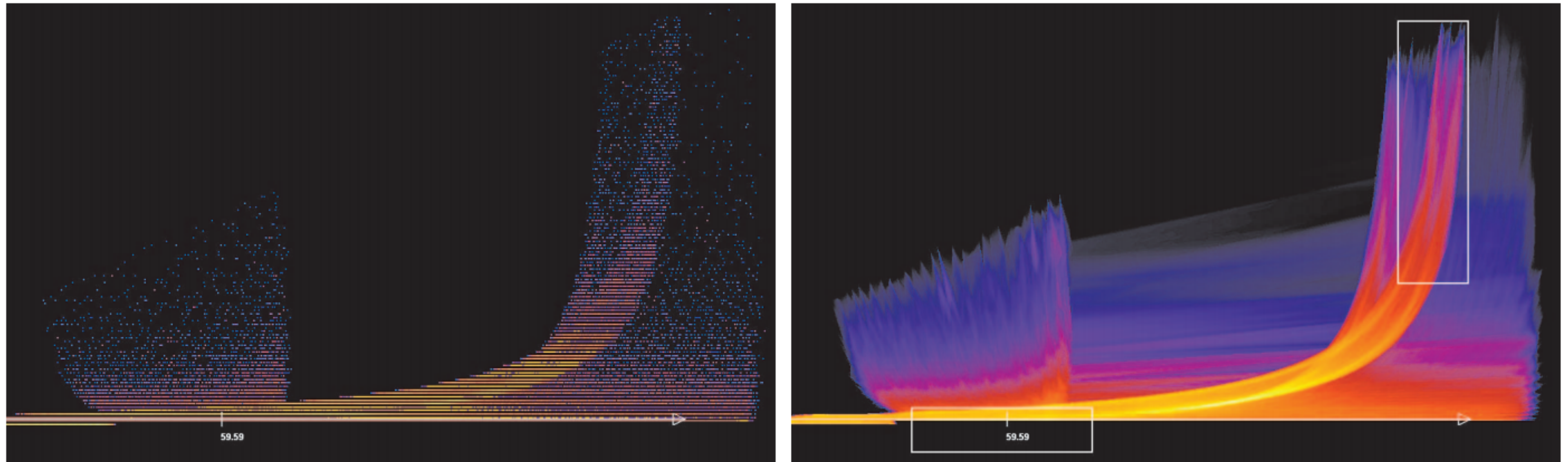
Fig. 4. Flight trails dataset (a) and element-based MoleView lens (b)

E.g. find flights with a certain altitude over a given spatial region

HurterTeleaErsoy2011

Continuous visual representation

Continuous scatterplot



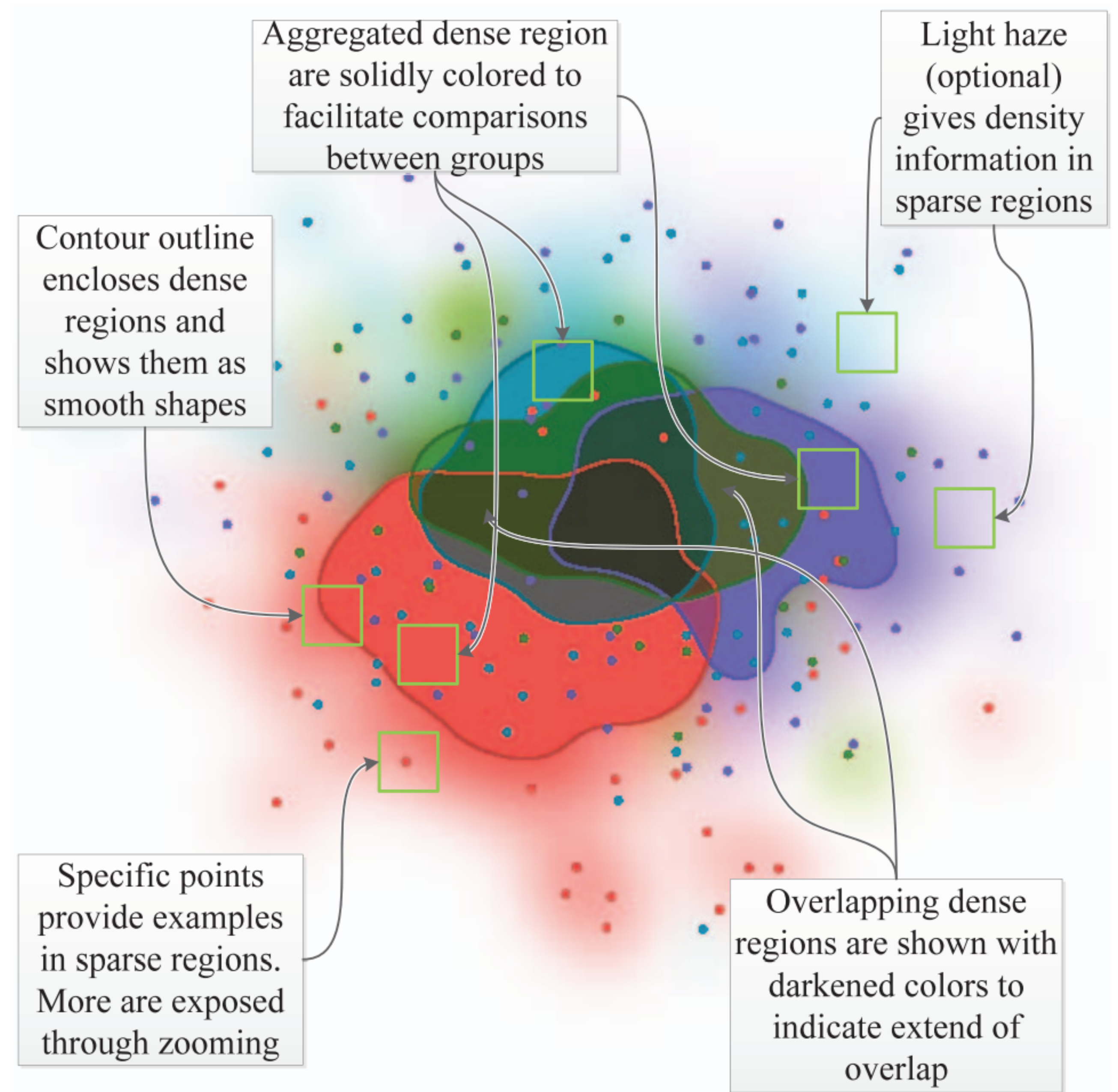
Takes into account the varying size and shape of grid cells by computing gradients within cells.

Continuous scatterplot



- Automatically groups dense regions into an abstract contour
- Renders the rest of the area using selected representatives, thus preserving the visual cue for outliers.

MayorgaGleicher2013



Accurate Color Blending

Hue-preserving color-blending

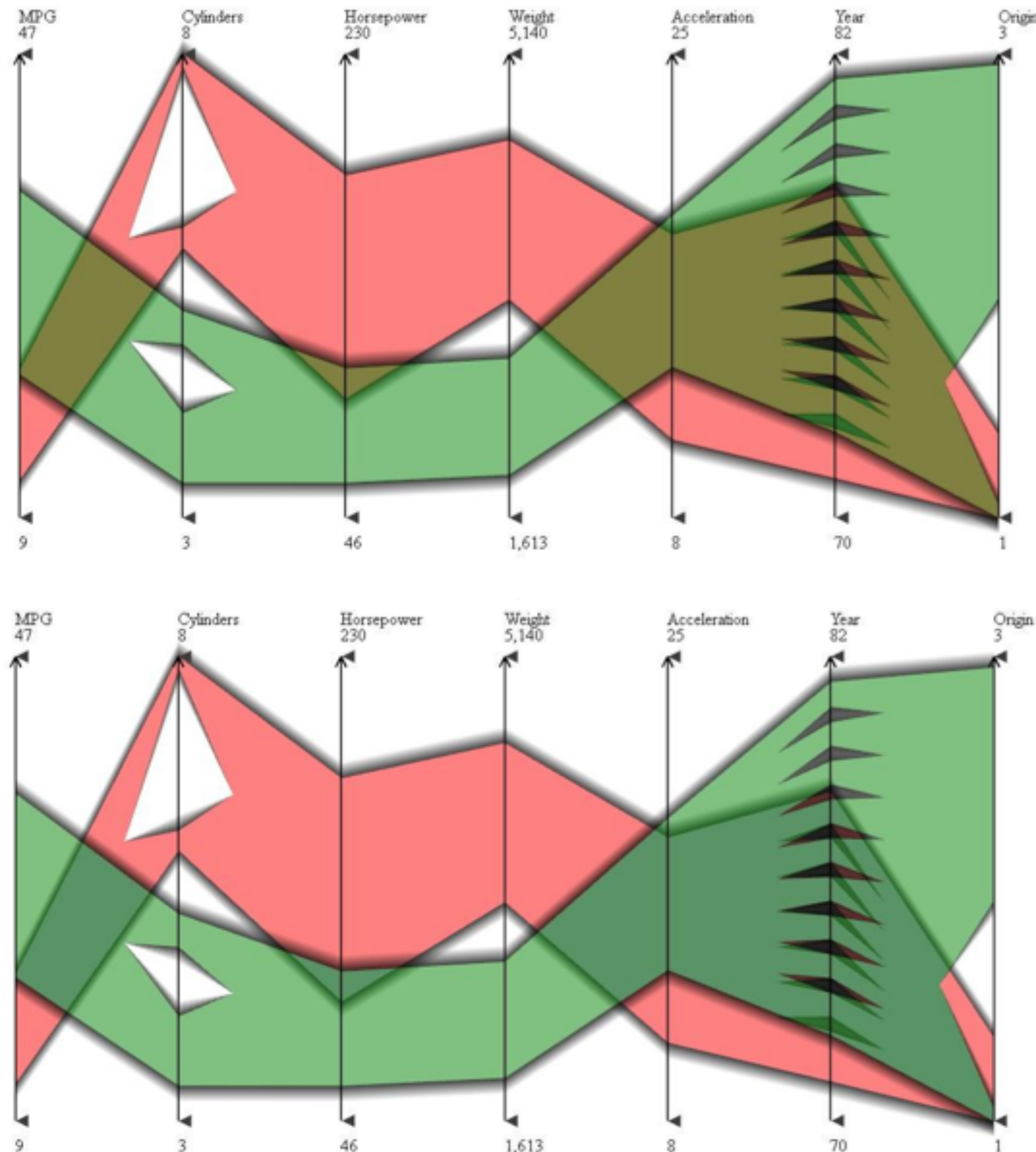


Fig. 9. A two-layer IPC with a green layer on top and red layer in the background: (top) blending using alpha-compositing; (bottom) blending using the data-driven blending operator.

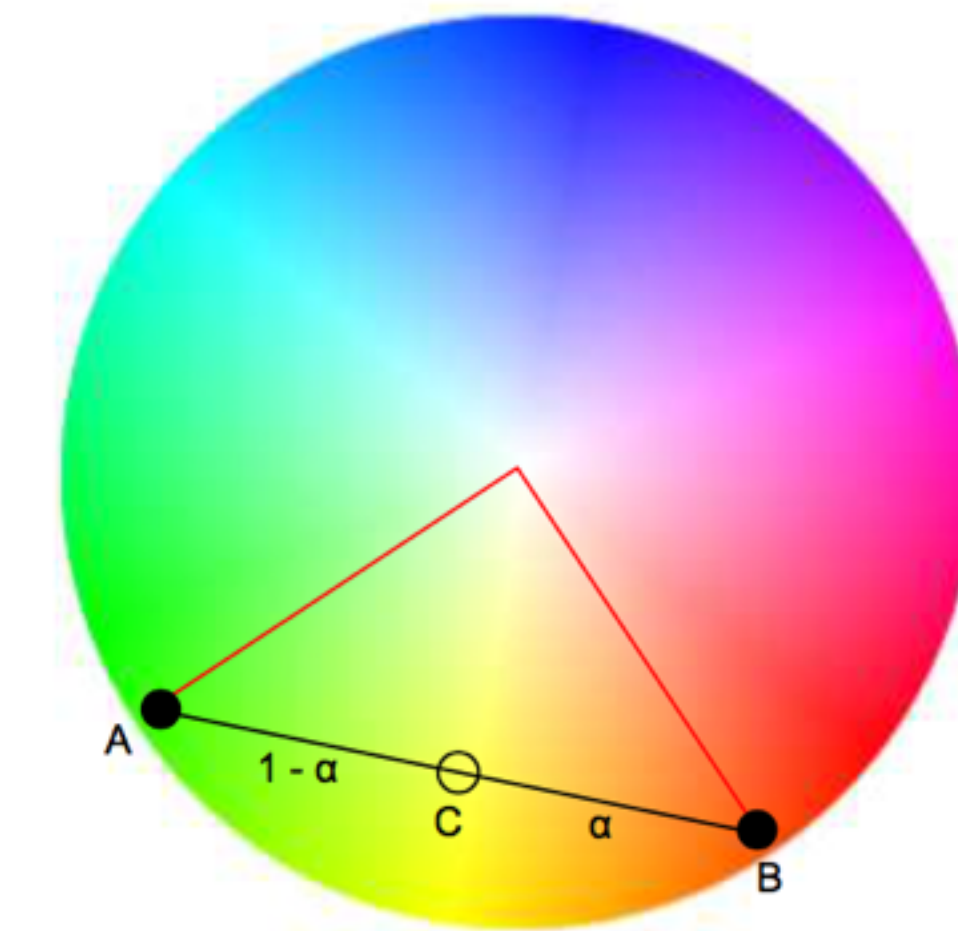
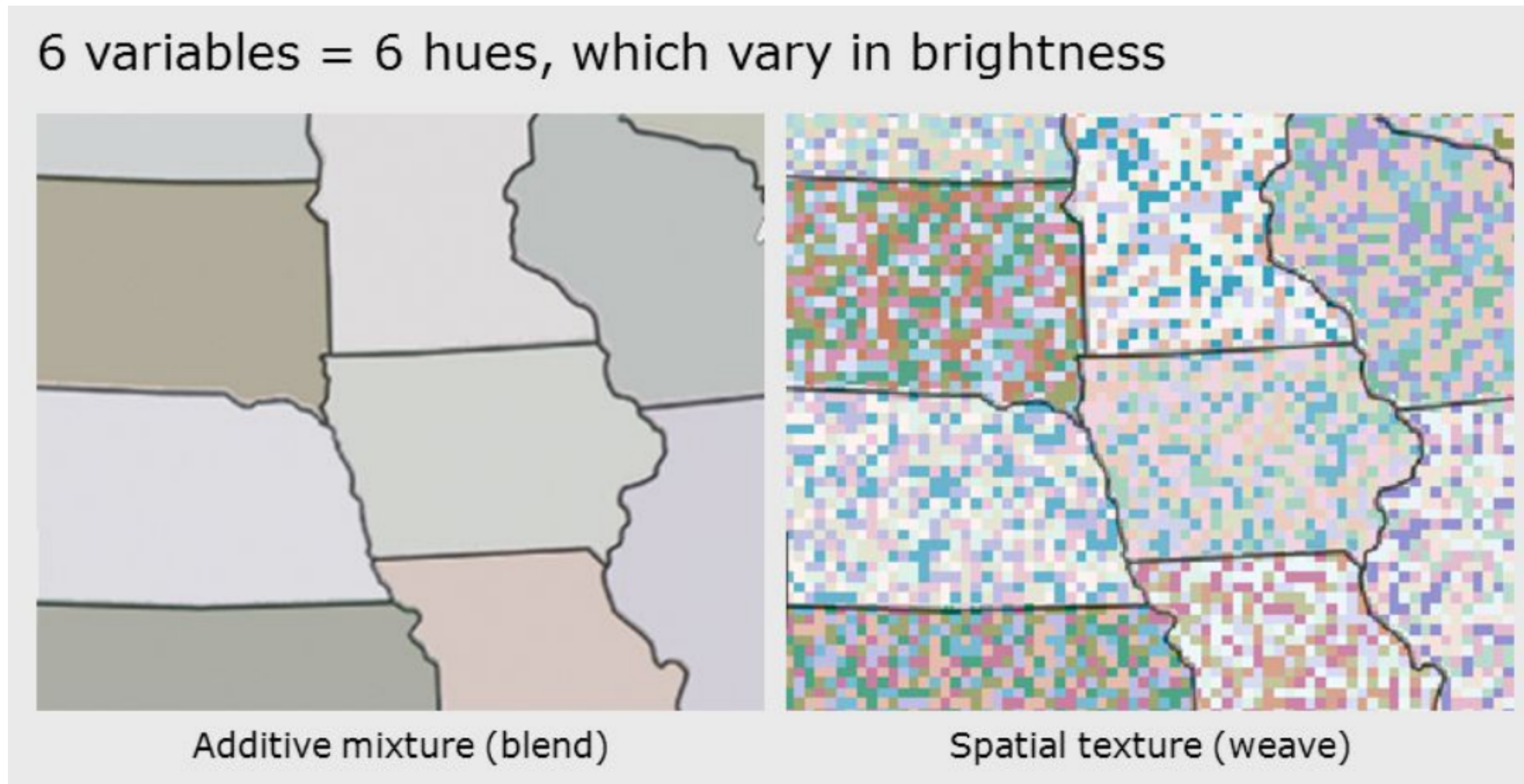


Fig. 1. Alpha-compositing. Two colors A and B create a false color C in a slice of the HSV representation of the $sRGB$ color space of equal luminance (color circle). The hue denotes an angle in the color circle.

Weaving vs. Blending

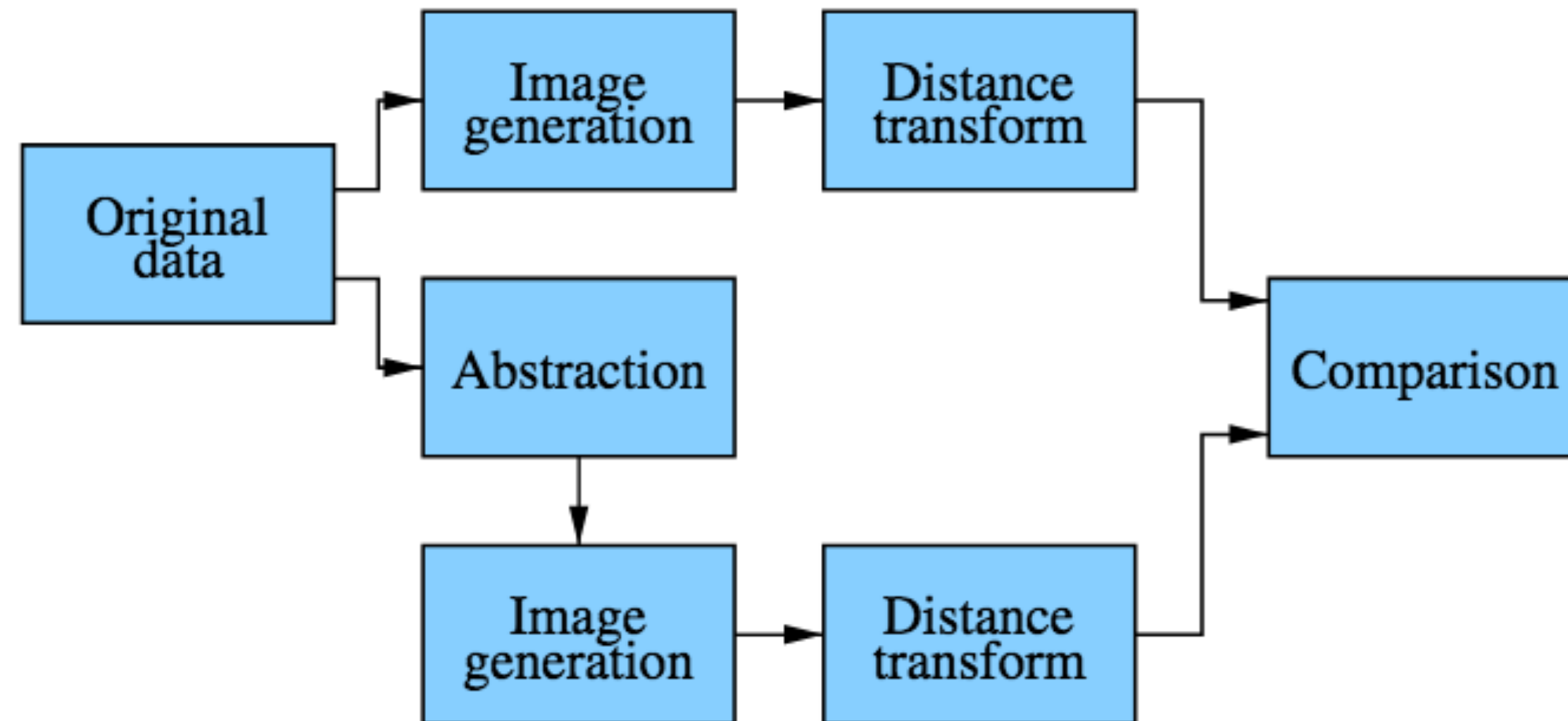


- Color weaving allows users to better infer the value of individual components; however, as the number of components increases, the advantage of color weaving diminishes.

Hagh-ShenasKimInterrante2007

Image Space Metrics

Screen Space Quality Metrics



- Measure for clutter reduction
- Based on distance transformation

Figure 2: *The screen space quality method consists of the following steps: (1) the original data set is abstracted, (2) the graphical representations of the original and abstracted data sets are rendered as images, (3) the images are transformed using distance transforms, and (4) a comparison function gives the quality value.*

Screen Space Quality Metrics

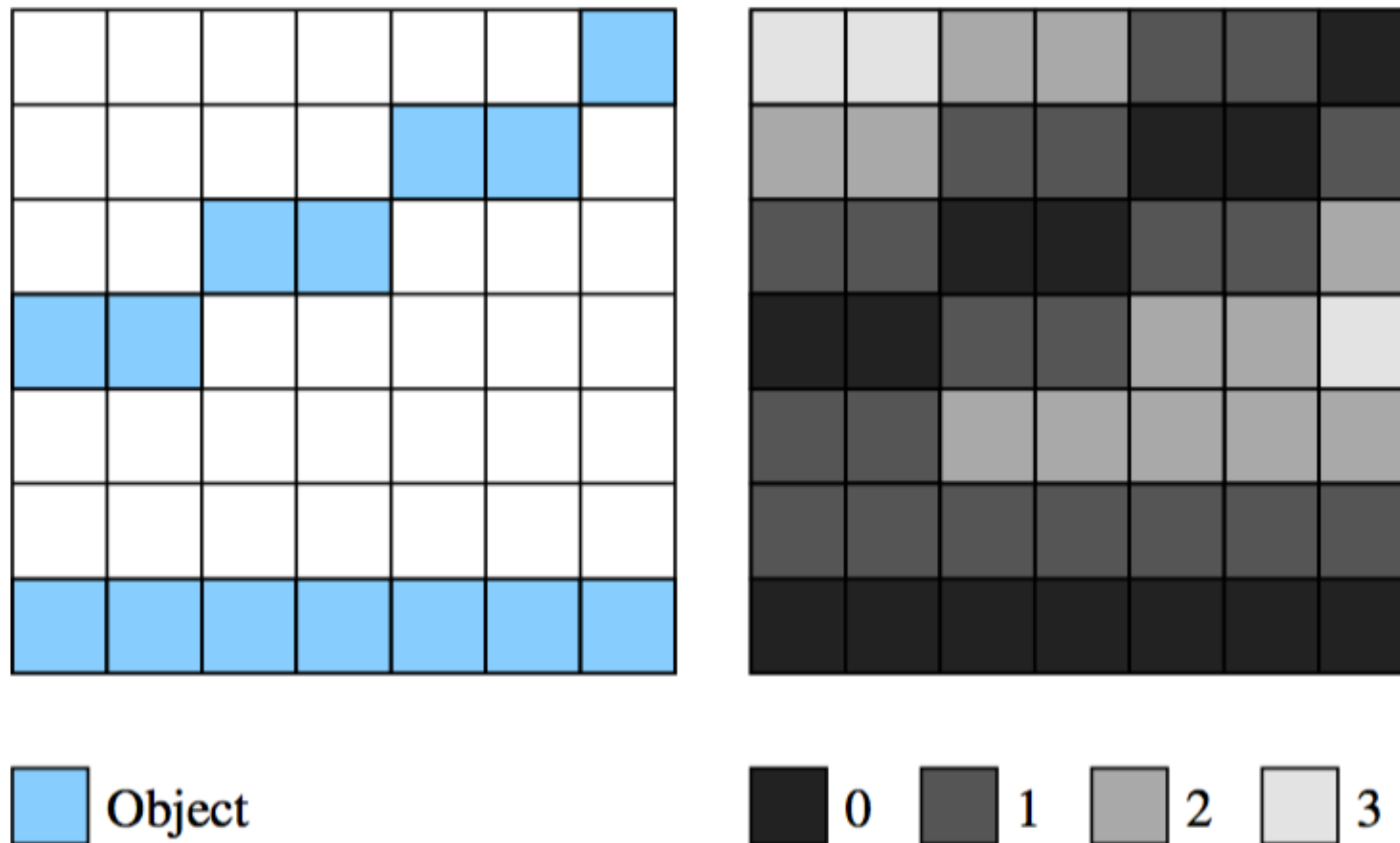


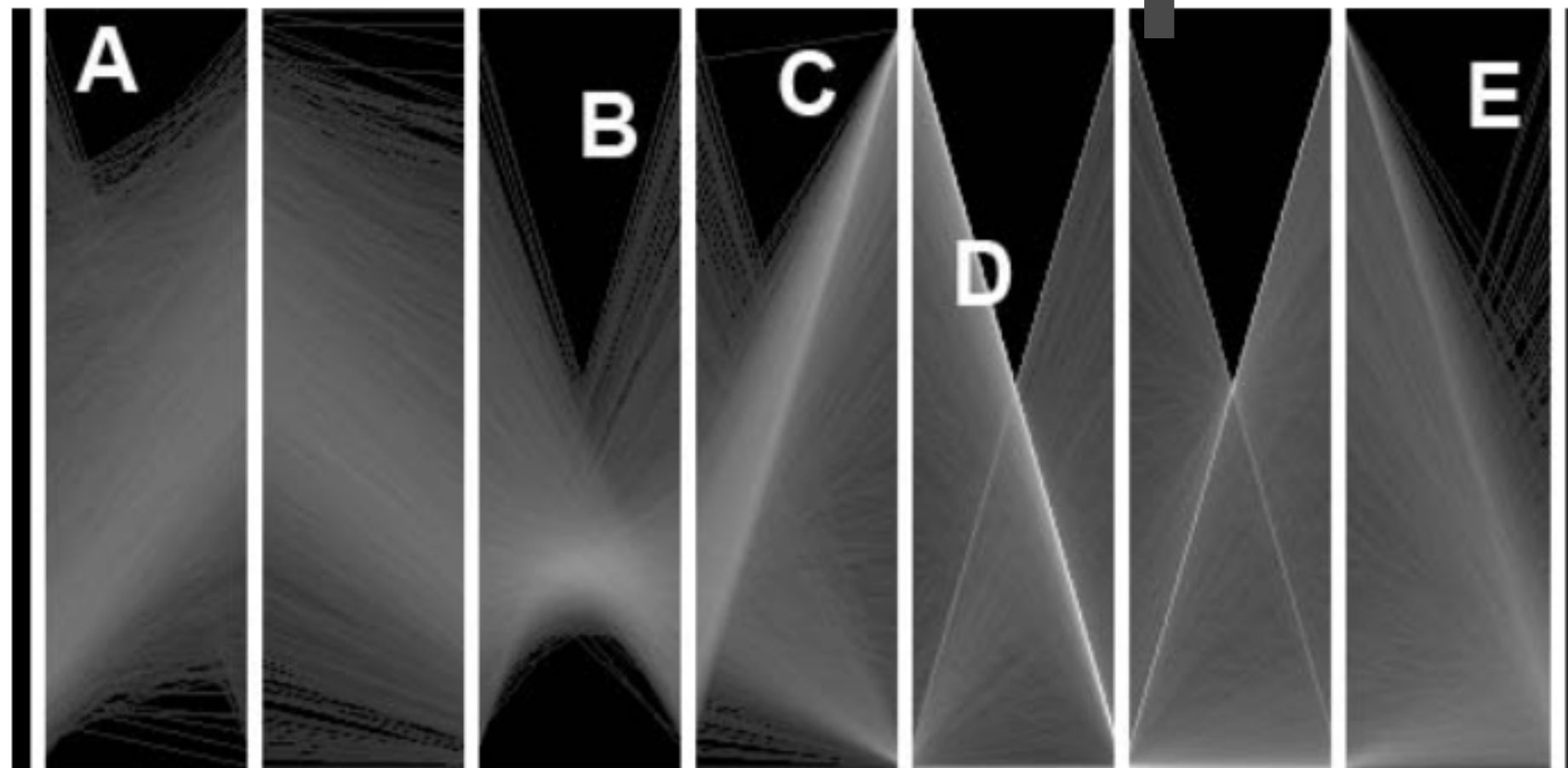
Figure 4: An illustration of the distance transform used for parallel coordinates. The left figure shows a graphical representation consisting of two lines, each resulting in several objects (in blue). The right figure shows the corresponding distance map where each pixel describes the vertical distance to the closest object, colour coded from black (a distance of 0) to light grey (a distance of 3).

1. an abstracted version of the original data set is created
2. the graphical representations of the original and abstracted data sets are rendered as two, equally sized images (I_O and I_A).
3. the images are transformed using a distance transform, ϕ_D , according to

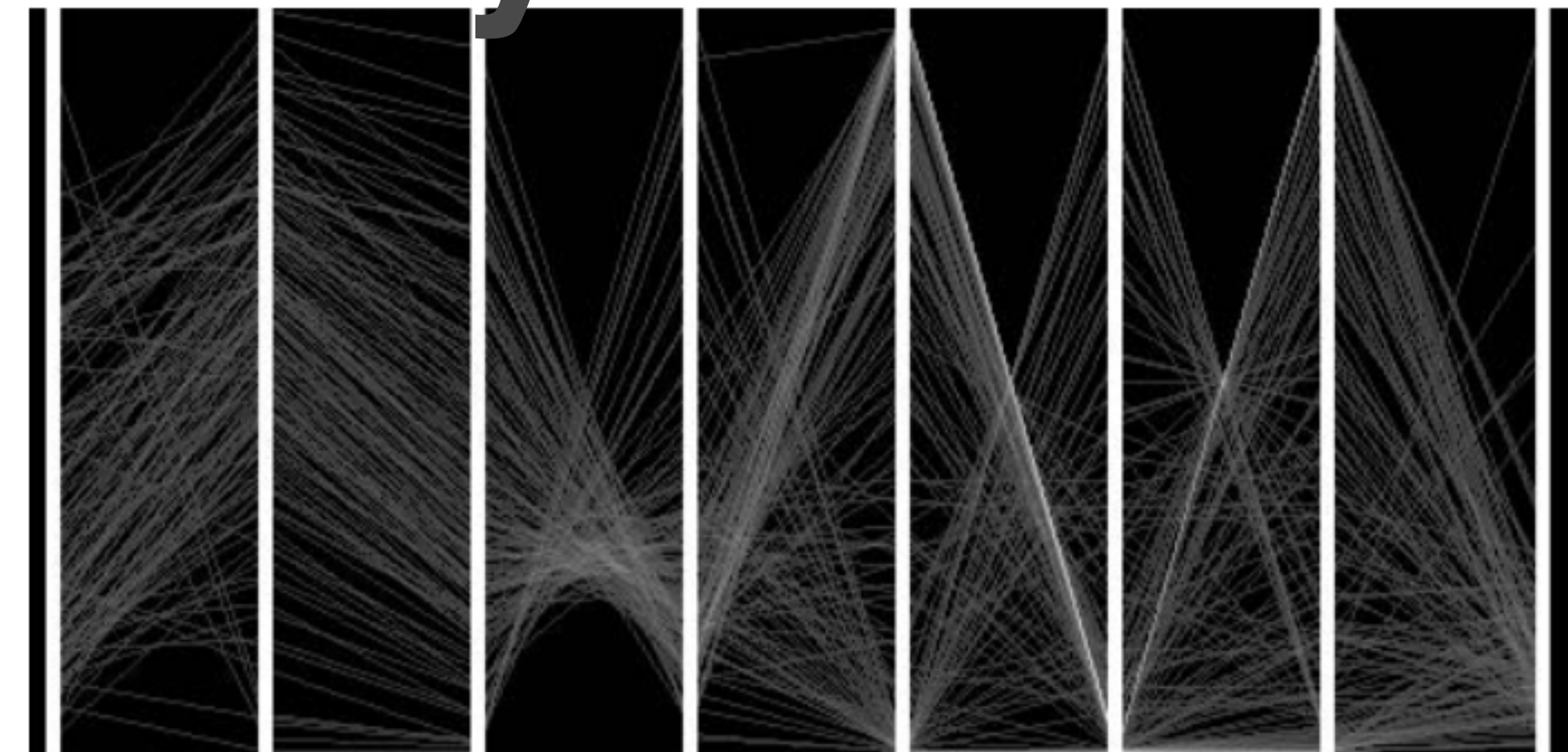
$$\begin{aligned} D_O &= \phi_D(I_O) \\ D_A &= \phi_D(I_A) \end{aligned} \quad (1)$$

- where D_O and D_A are new images of the same size as I_O and I_A representing the corresponding distance maps
4. the similarity between the distance maps is calculated as $s = \Psi(D_O, D_A)$, where Ψ is a comparison function

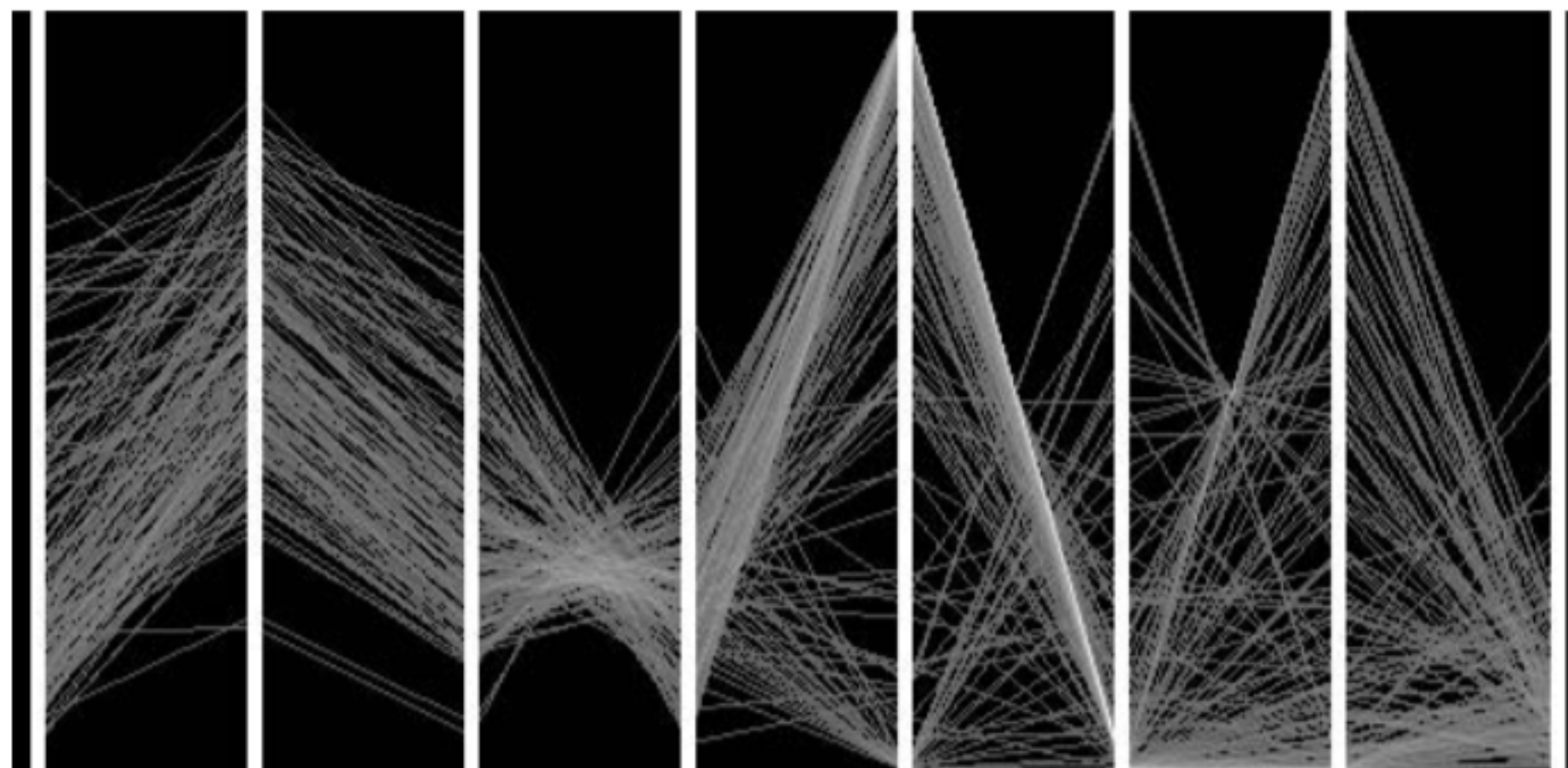
Screen Space Quality Metrics



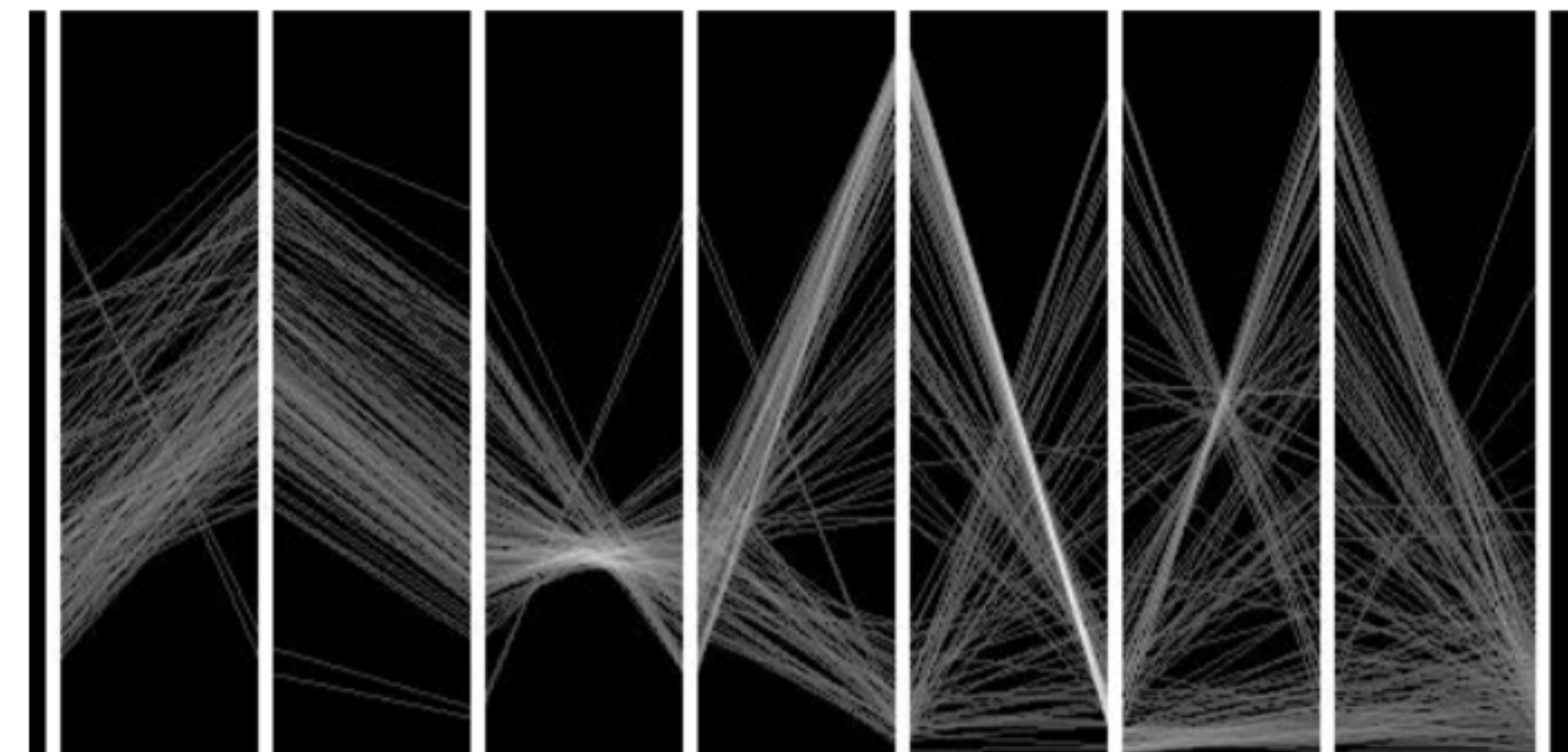
(a) Original data set comprising 7342 items.



(b) Targeting a visual quality of 0.90 retains 155 items.



(c) Using 155 randomly picked items results in a visual quality of 0.74.



(d) Using the K-means algorithm to construct 155 clusters results in a visual quality of 0.76.

Figure 8: Producing an abstraction of the household economics data set (a) by (b) targeting a visual quality of 0.90 using sampling retains 155 items. Structures are preserved in **B–E** and outliers are revealed in **A**. Randomly picking 155 items with no quality control (c) only preserves structures in **D**. Using 155 cluster centroids (d) also preserves structures in **A** and **E**.

Pixnostic

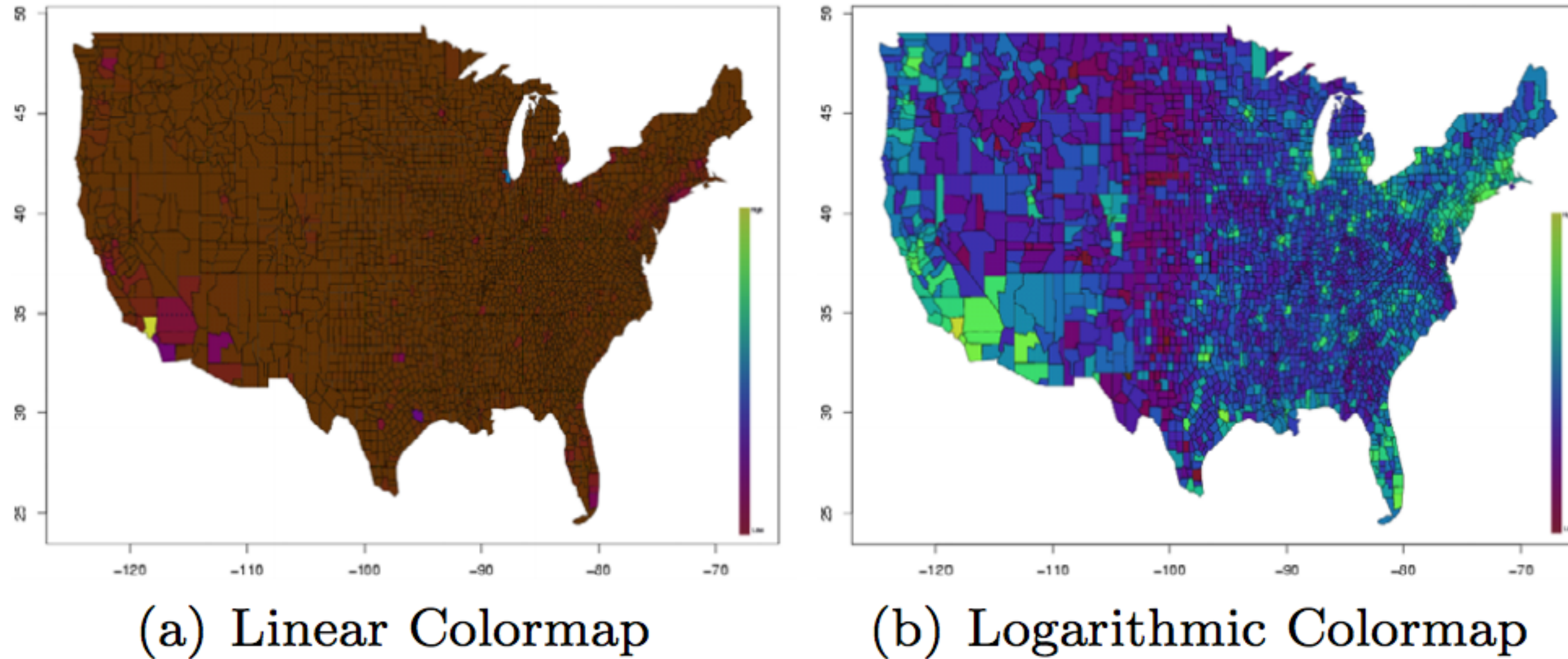


Figure 1: *A typical application scenario* – The visual analysis of a census data set involves different normalizations to a color scale; Although both visualizations are based on exactly the same input data, the right figure provides more insight since a logarithmic color scale is more suitable for the underlying data distribution

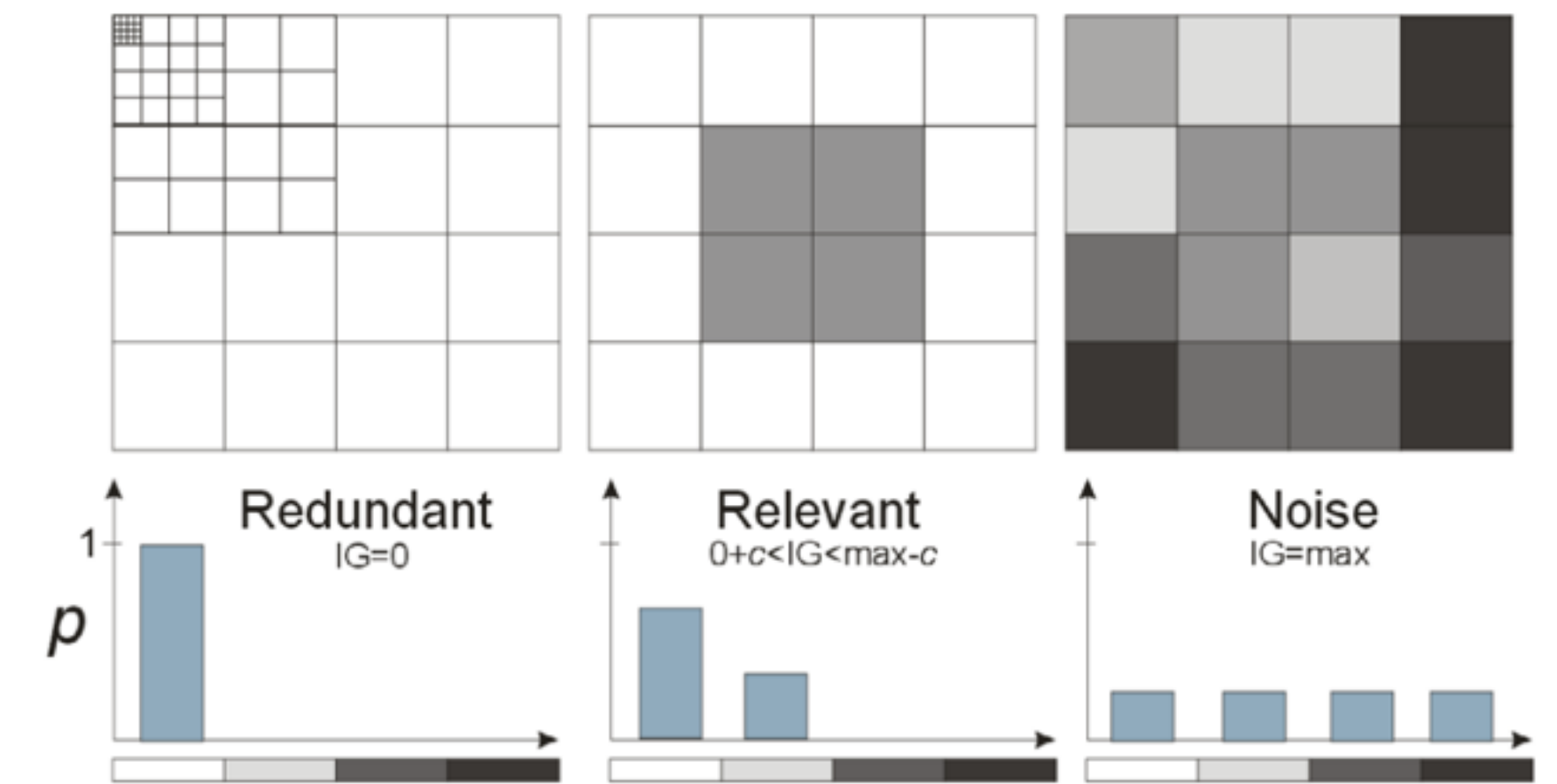
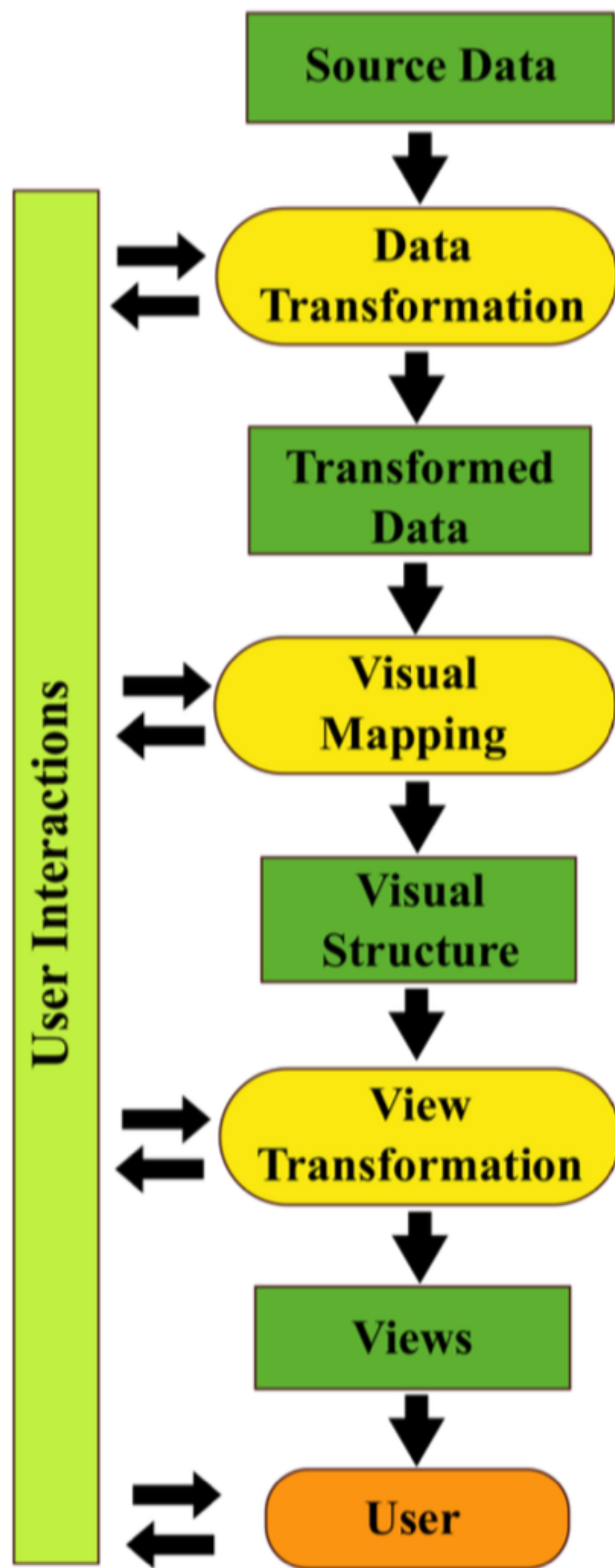


Figure 4: Information content (IG) of different gray level images. From an analyst's point of view, interesting images should have an Information content in a certain range c between 0 and IG_{max}

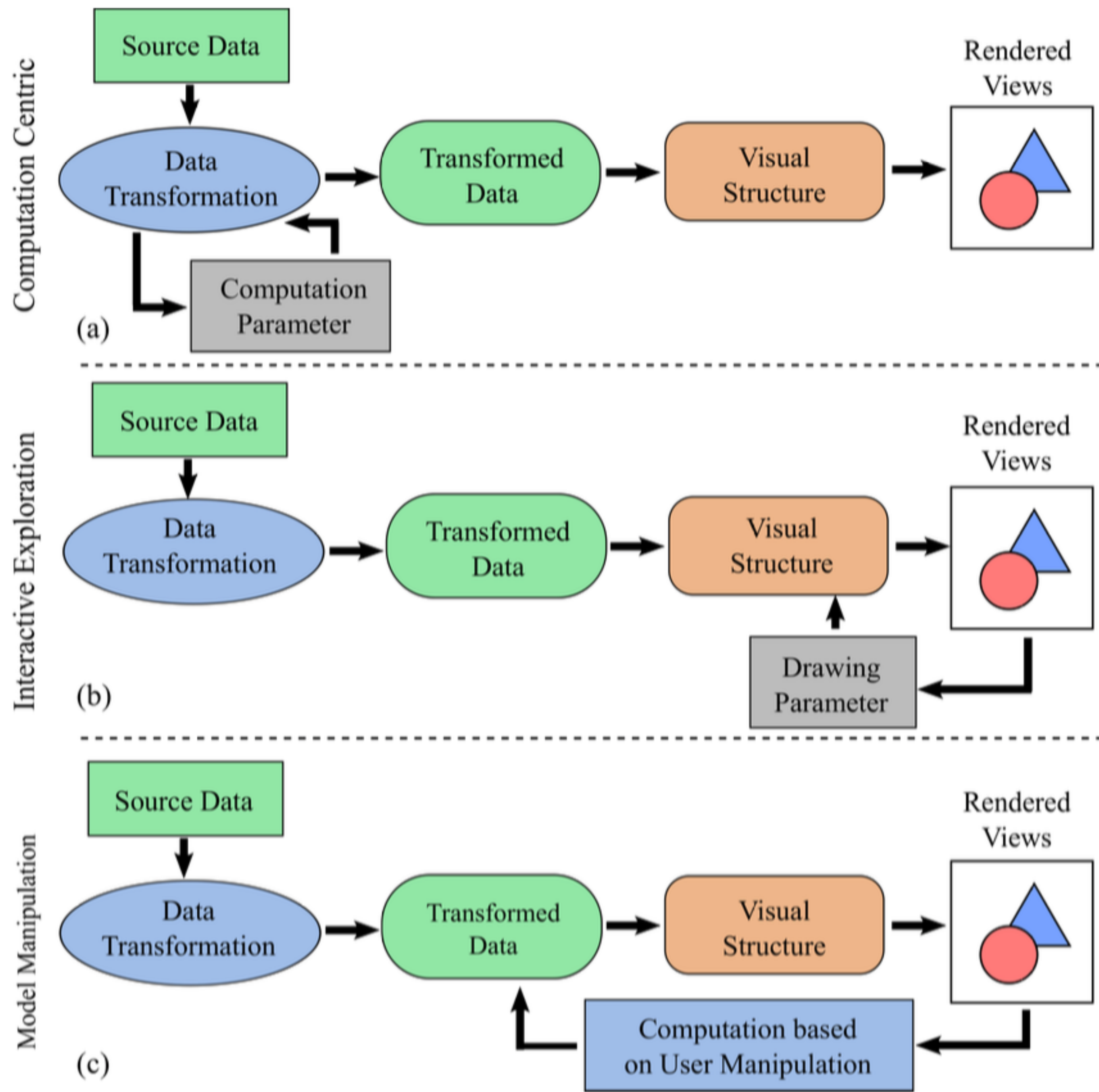
- Rank interestingness for pixel based visualization such as pixel bar charts

User Interactions



Visualization pipeline for high-dim data

Interactions are integrated into each processing stage

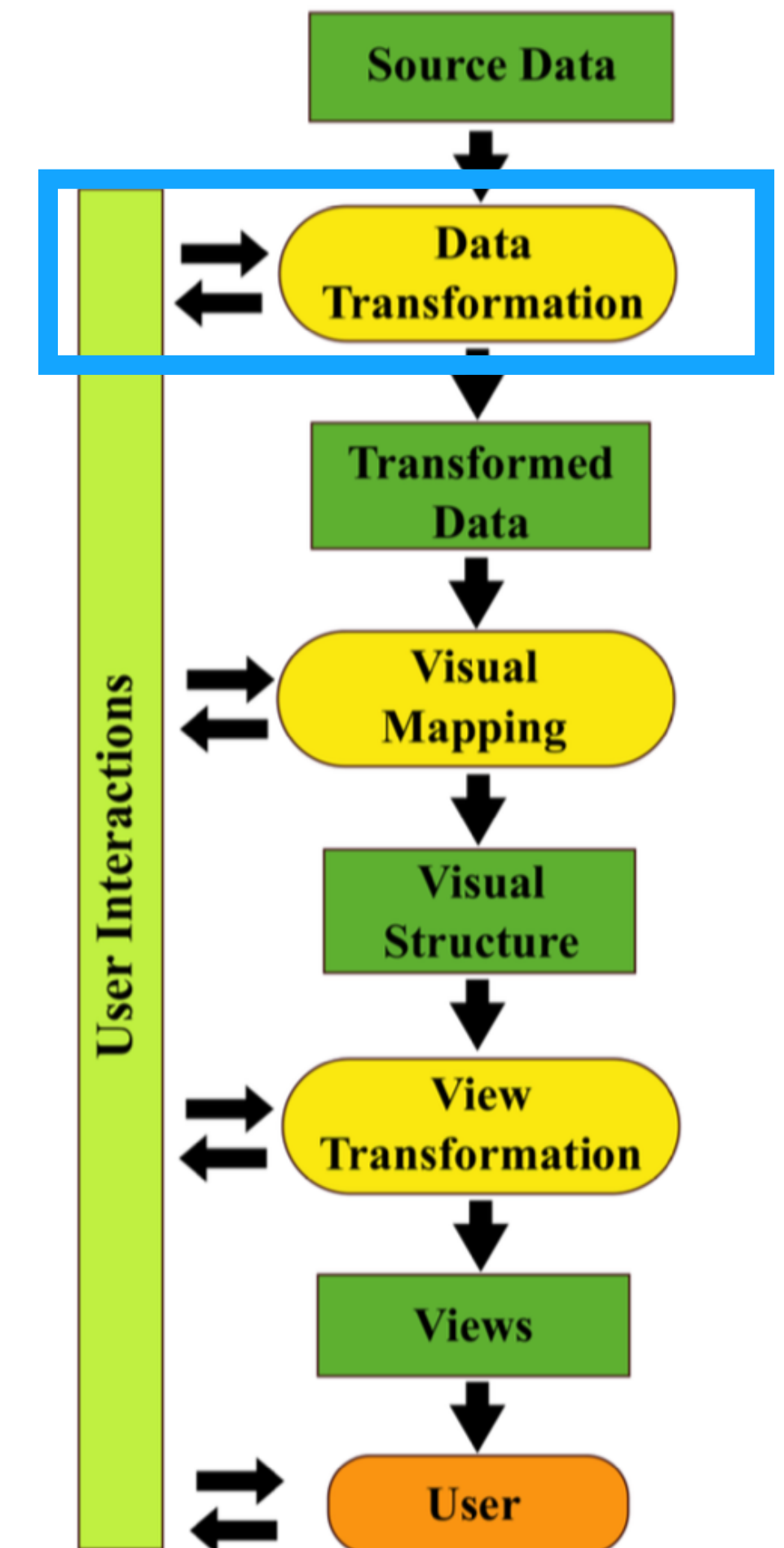
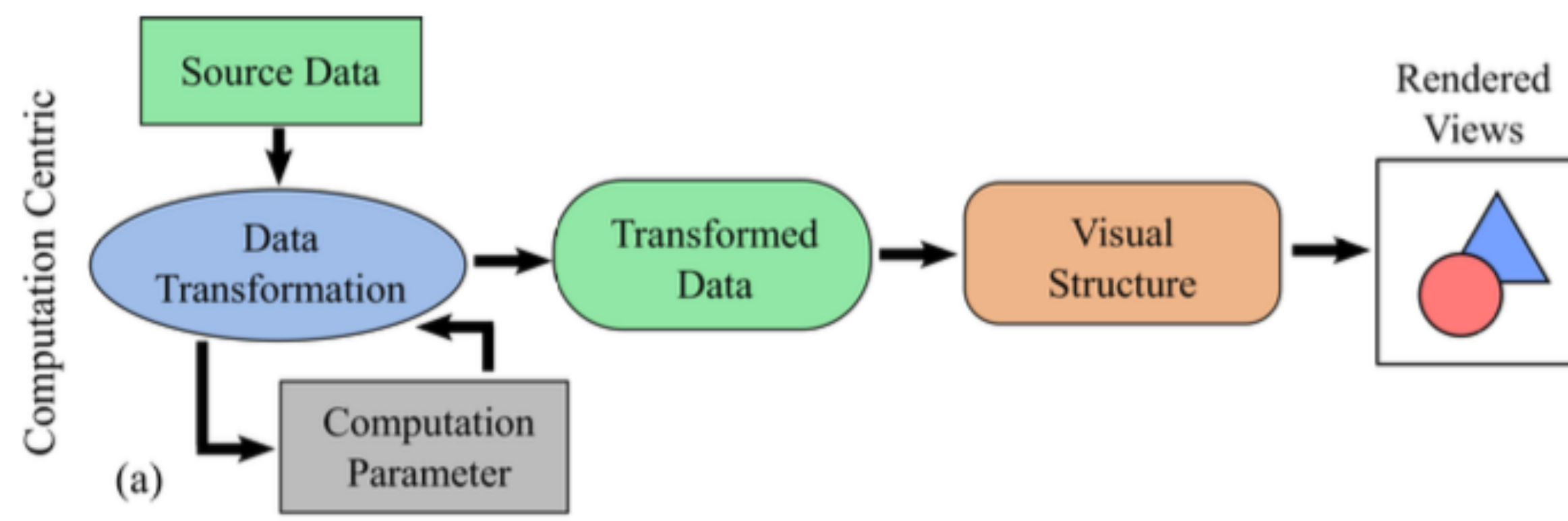


- Computation-centric approaches
- Interactive exploration
- Model manipulation

Fig. 7. The three types of user interaction paradigms with varying degrees of user involvement. Since each paradigm can interact with each processing stage in the visualization pipeline, the diagram highlights the most general patterns.

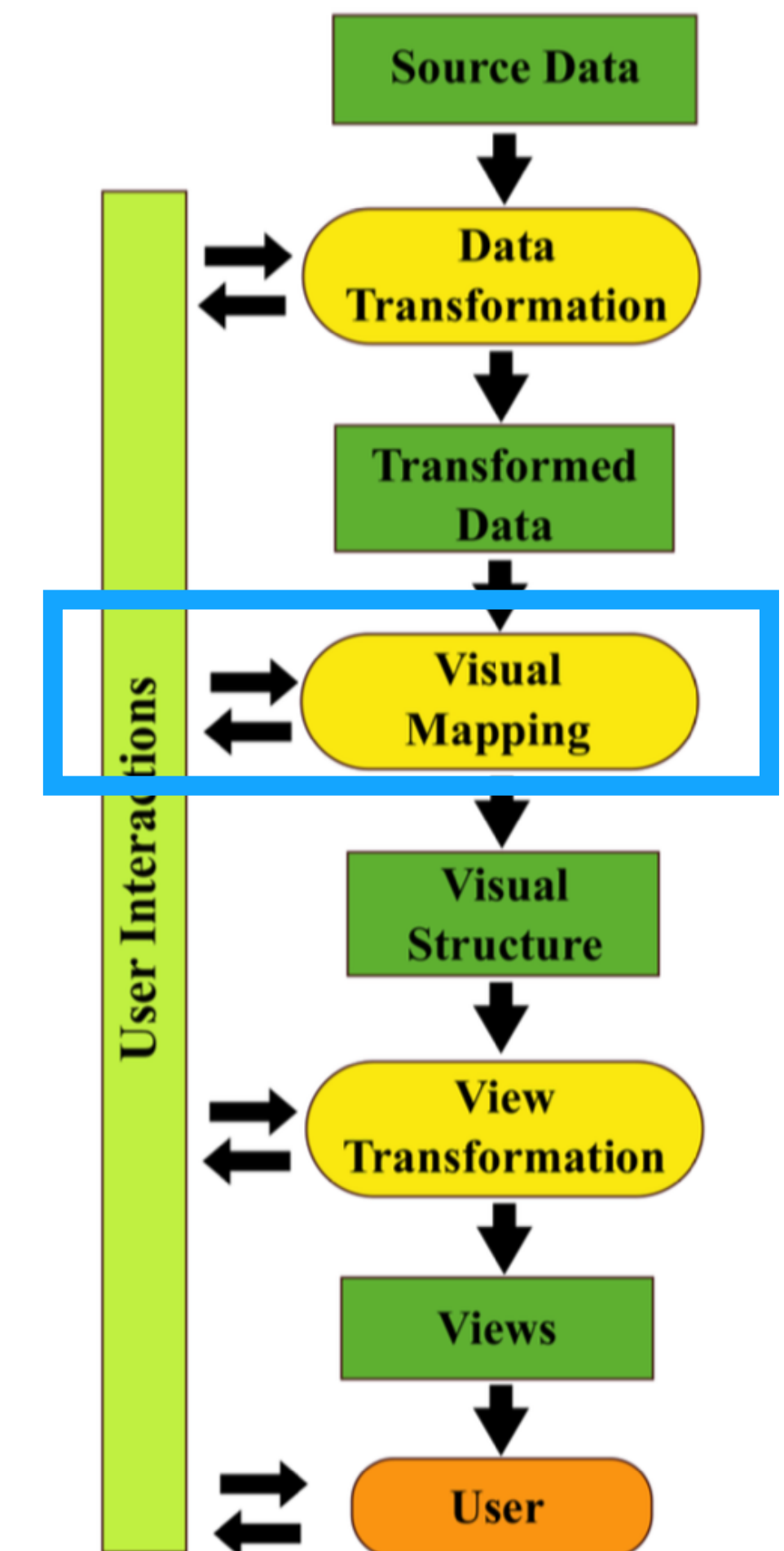
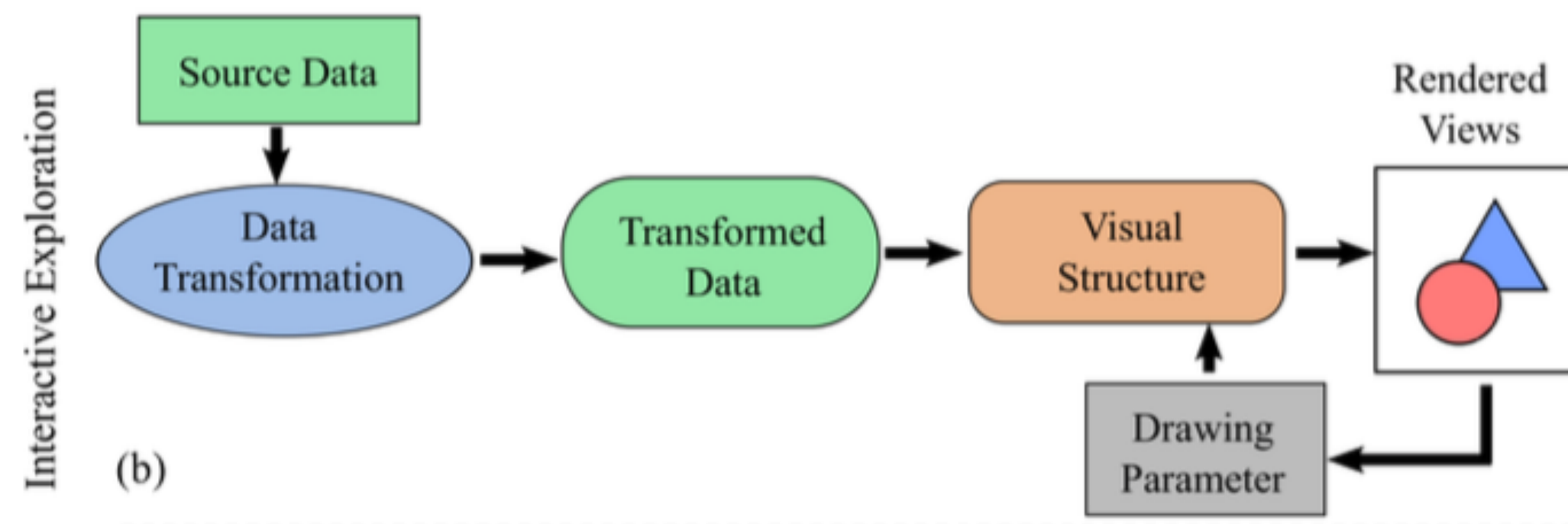
Computation-Centric approaches

- Require only limited user input: e.g., setting initial parameters
- Center around algorithms designed for well-defined computational problems:
 - DR, subspace clustering, regression, quality metric
- Concentrated in data transformation



Interactive Exploration

- Navigate, query, and filter the existing model interactively for more effective visual communication.
- Mostly in visual mapping stage: interactively modify visual structure
- User do not alter the underlying computation model in the interactive exploration.
- “Visual data mining”



Interactive Exploration: MDSteer

- Progressive layout of points
- Hierarchical binning
- Steering allows computational power to be focused where it is needed to support exploration in parallel with continuing the layout process.
- The algorithm alternates layout with binning computation:
 - At each layout step, we add $\sqrt{n/k}$ new points to the computation, find an initial position for each of these new points, and run MDS iterations.
 - Every k layout steps, we re-bin all of the points.
- Bins serve as a mechanism for the user to select a subset of the data as the target of the available computational resources

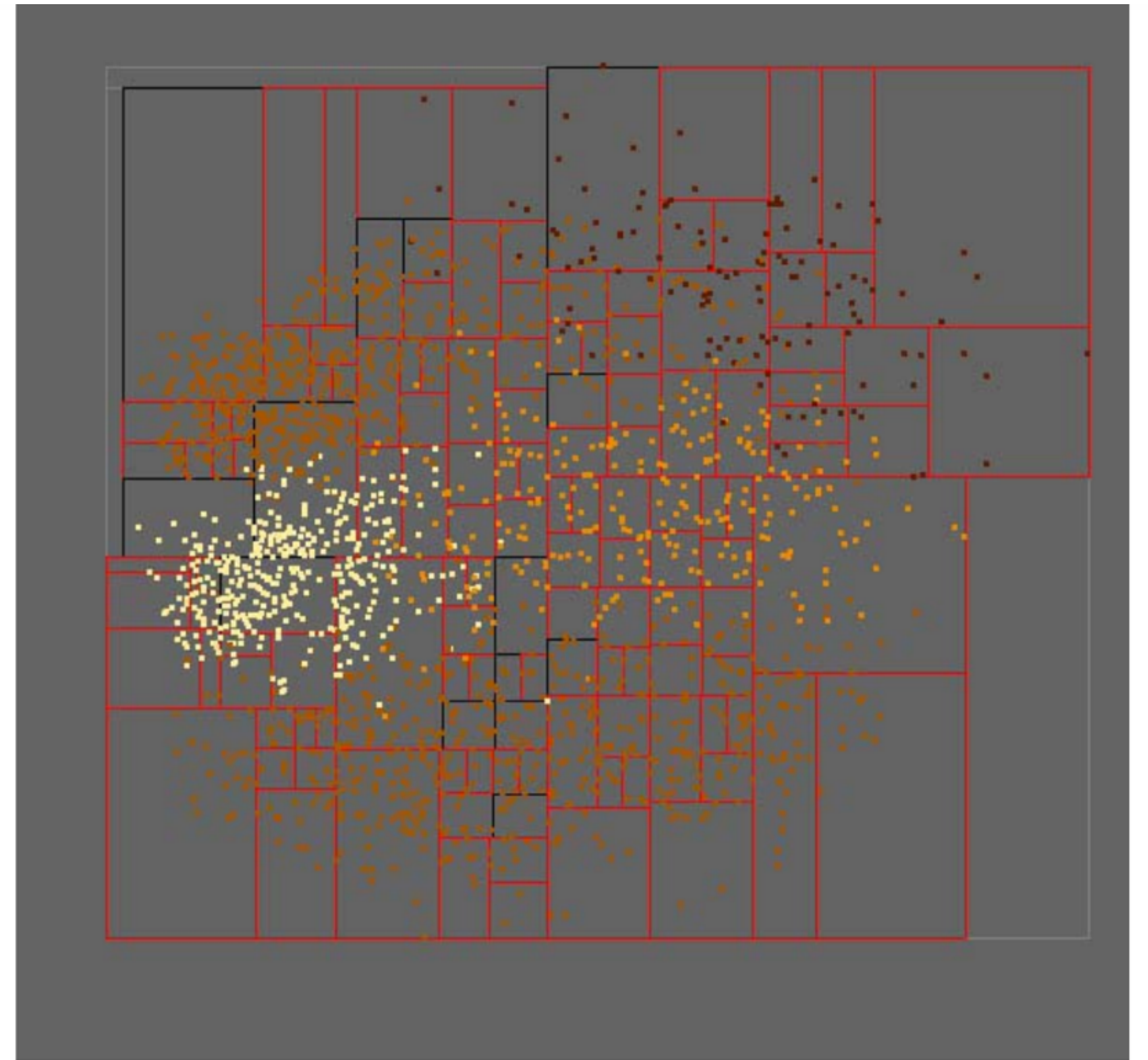
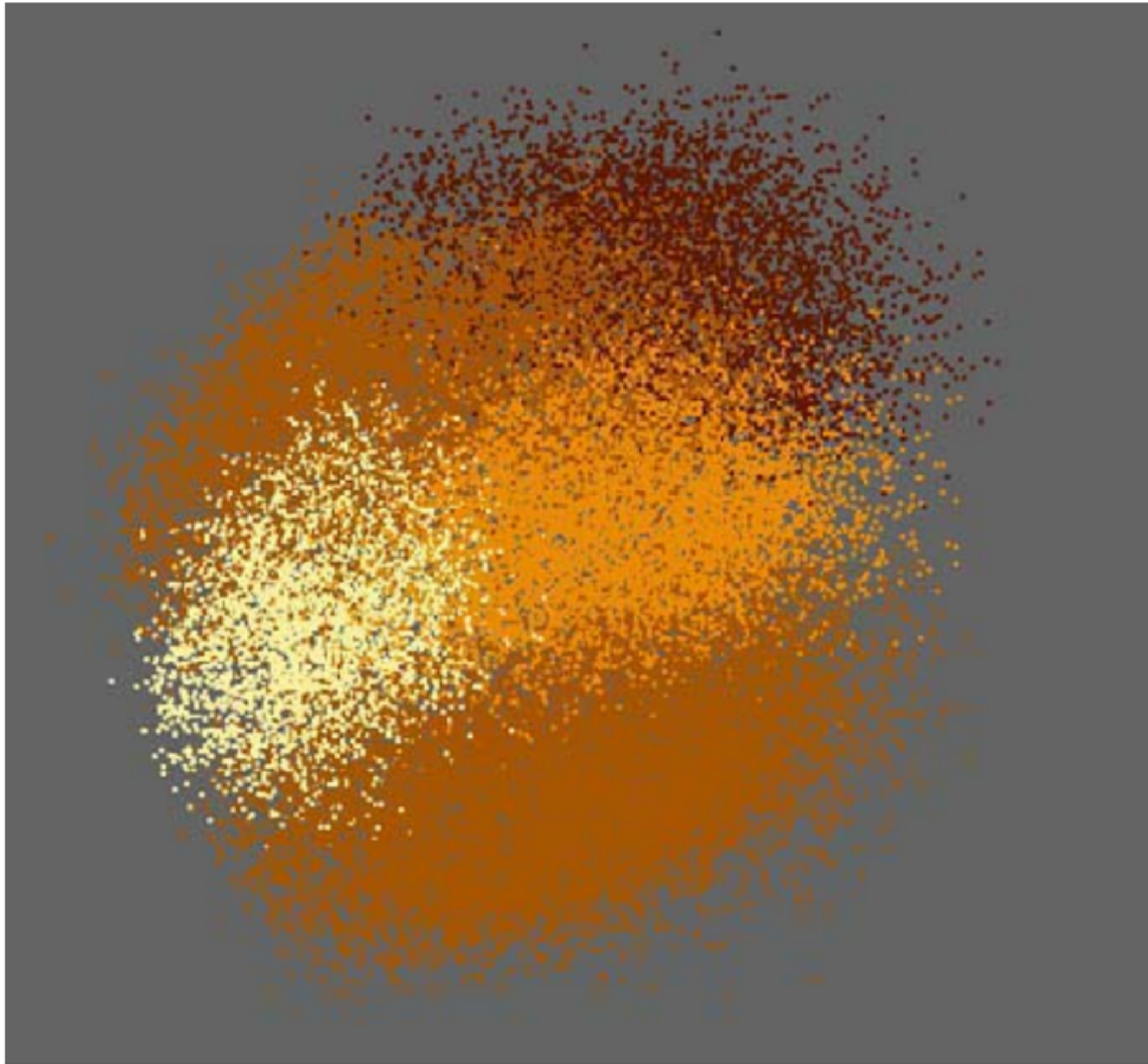
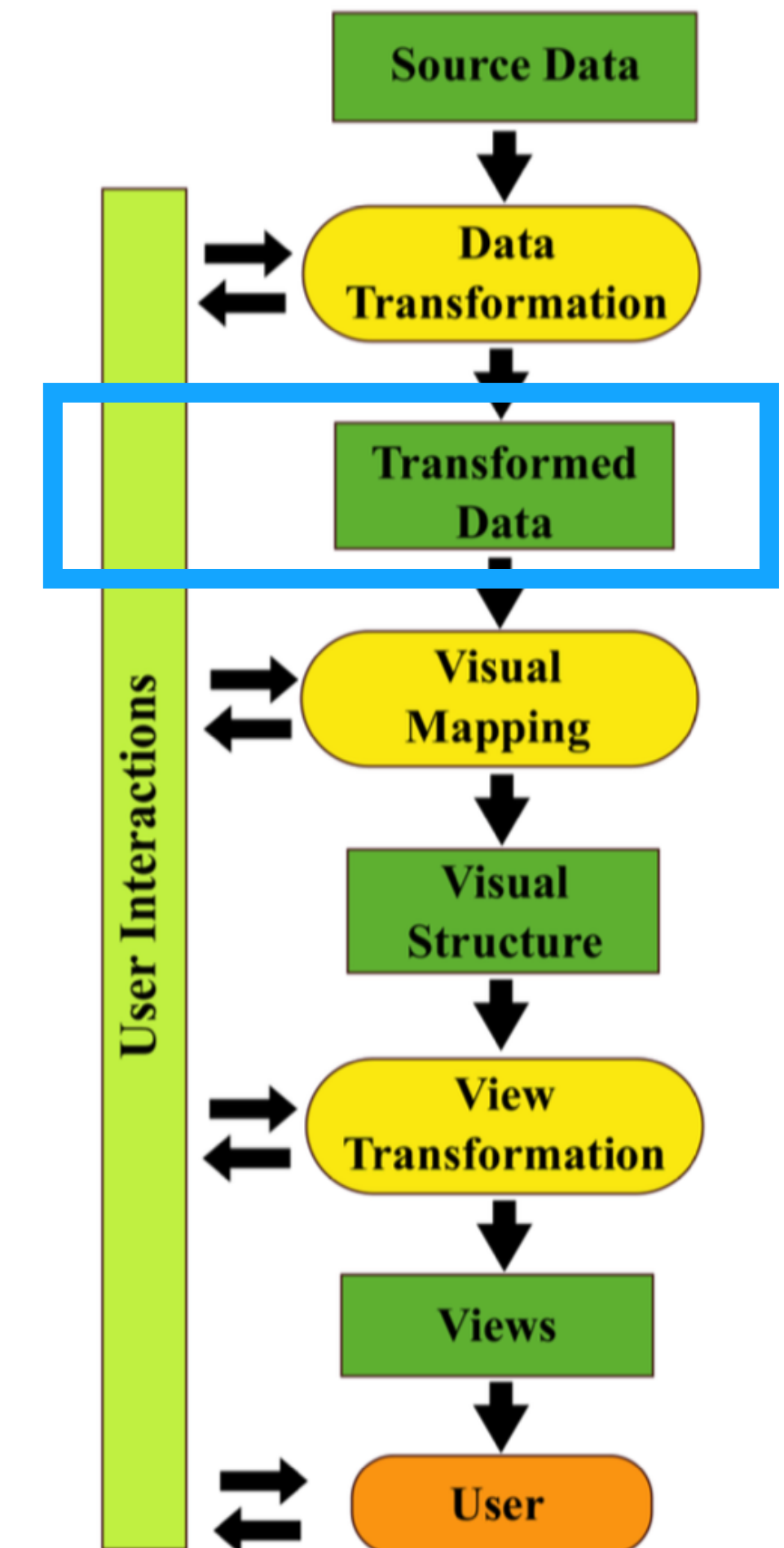
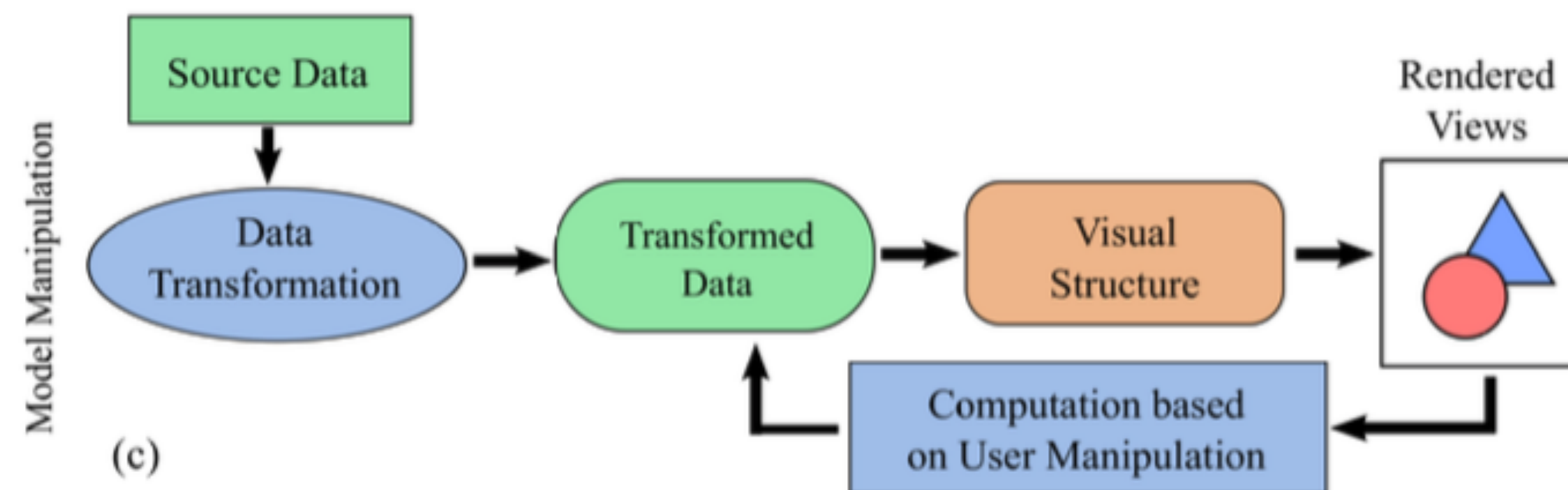


Figure 2: Visual Quality: Environmental Dataset. **Left:** We show the 40,000 point real environmental dataset laid out with the Morrison [12] algorithm, taken after a full layout computation that takes 16 minutes. **Right:** We show a partially placed version of the same environmental dataset after steering with MDSteer for roughly 2 minutes. Again, we see the same large-scale structure.

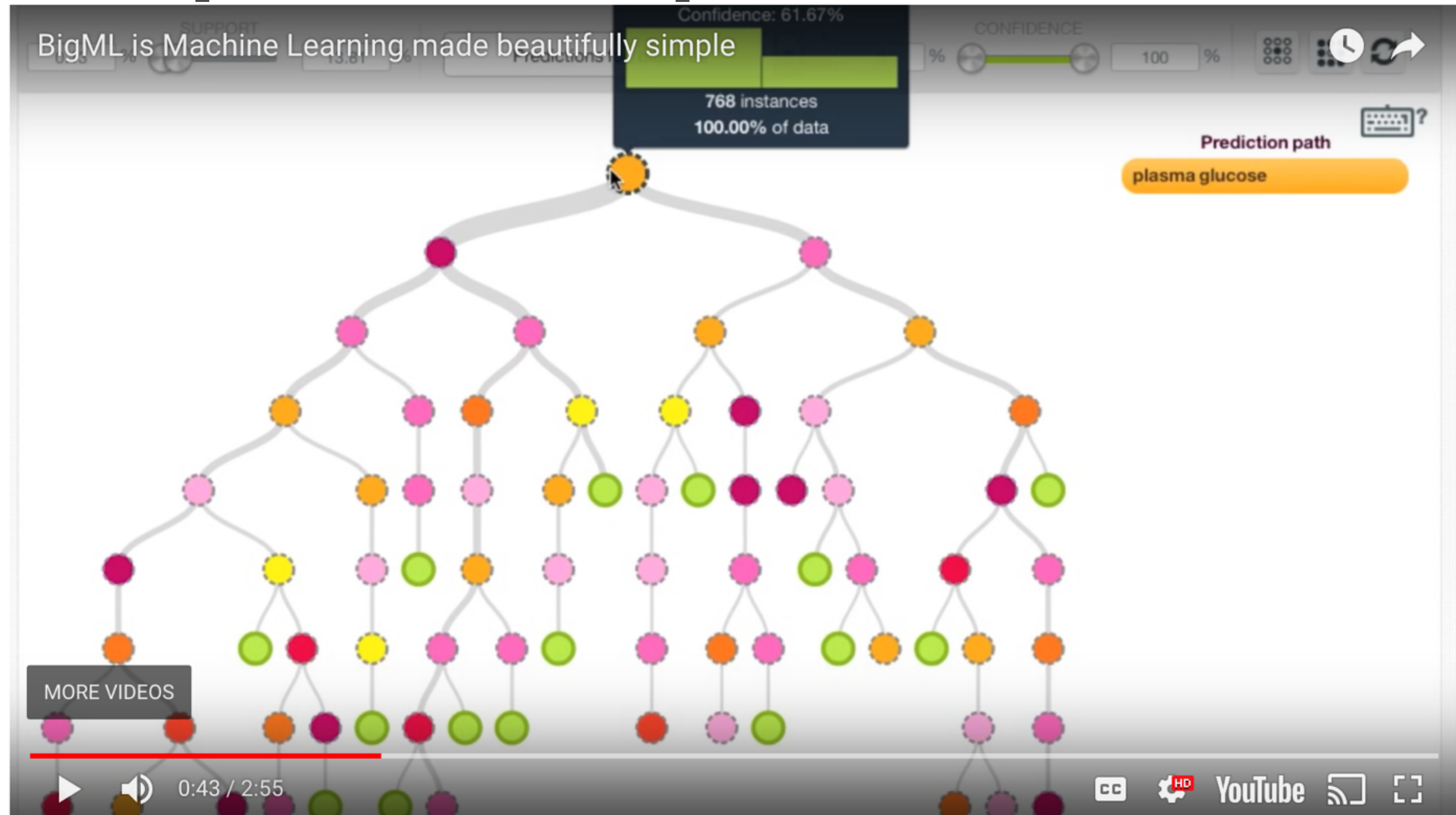
Model Manipulation

- Integrate user manipulation as part of the algorithm and update the underlying model to reflect the user input to obtain new insights.
- For example, control point based projection methods using user manipulation



Decision Tree and Vis

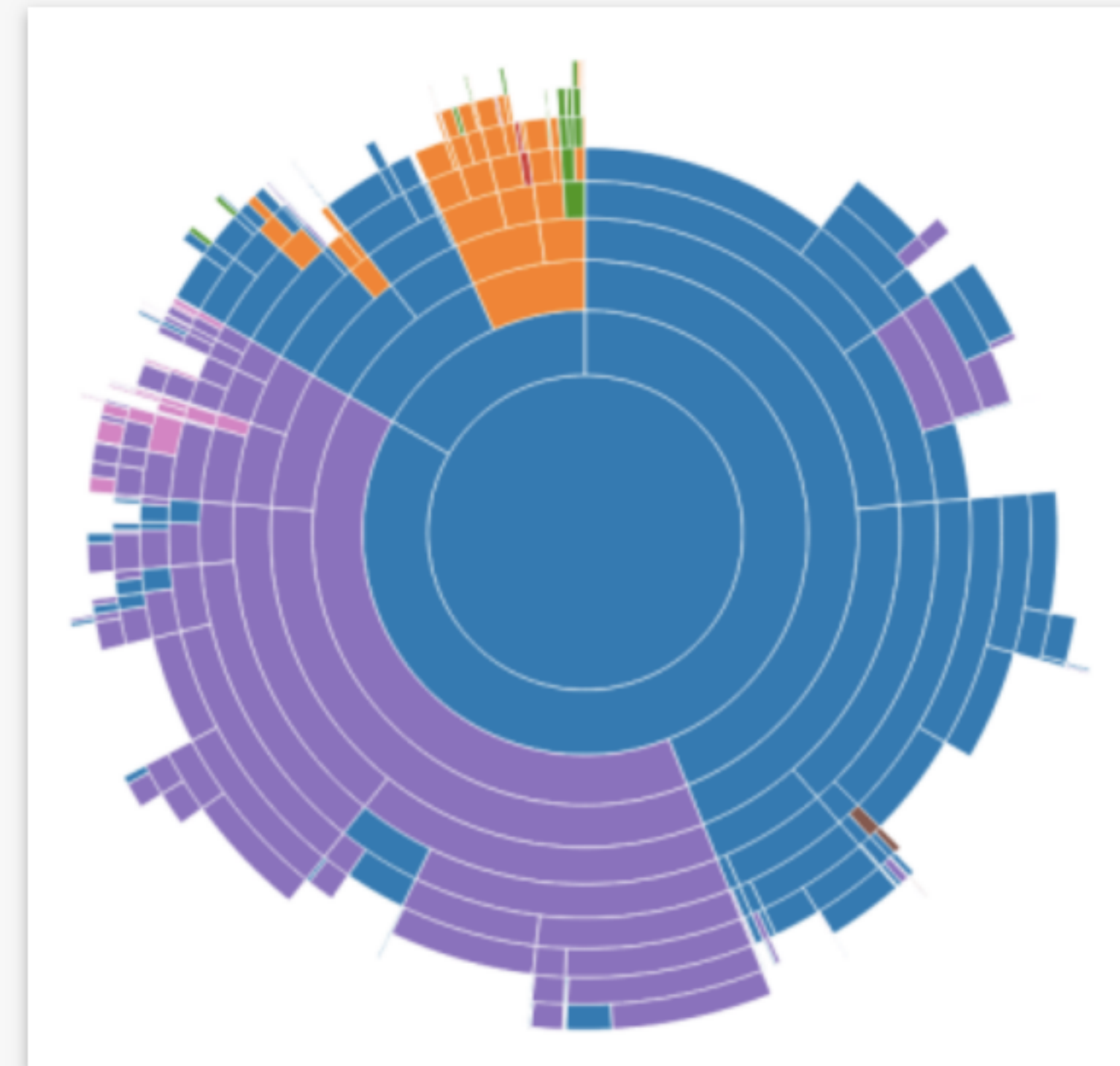
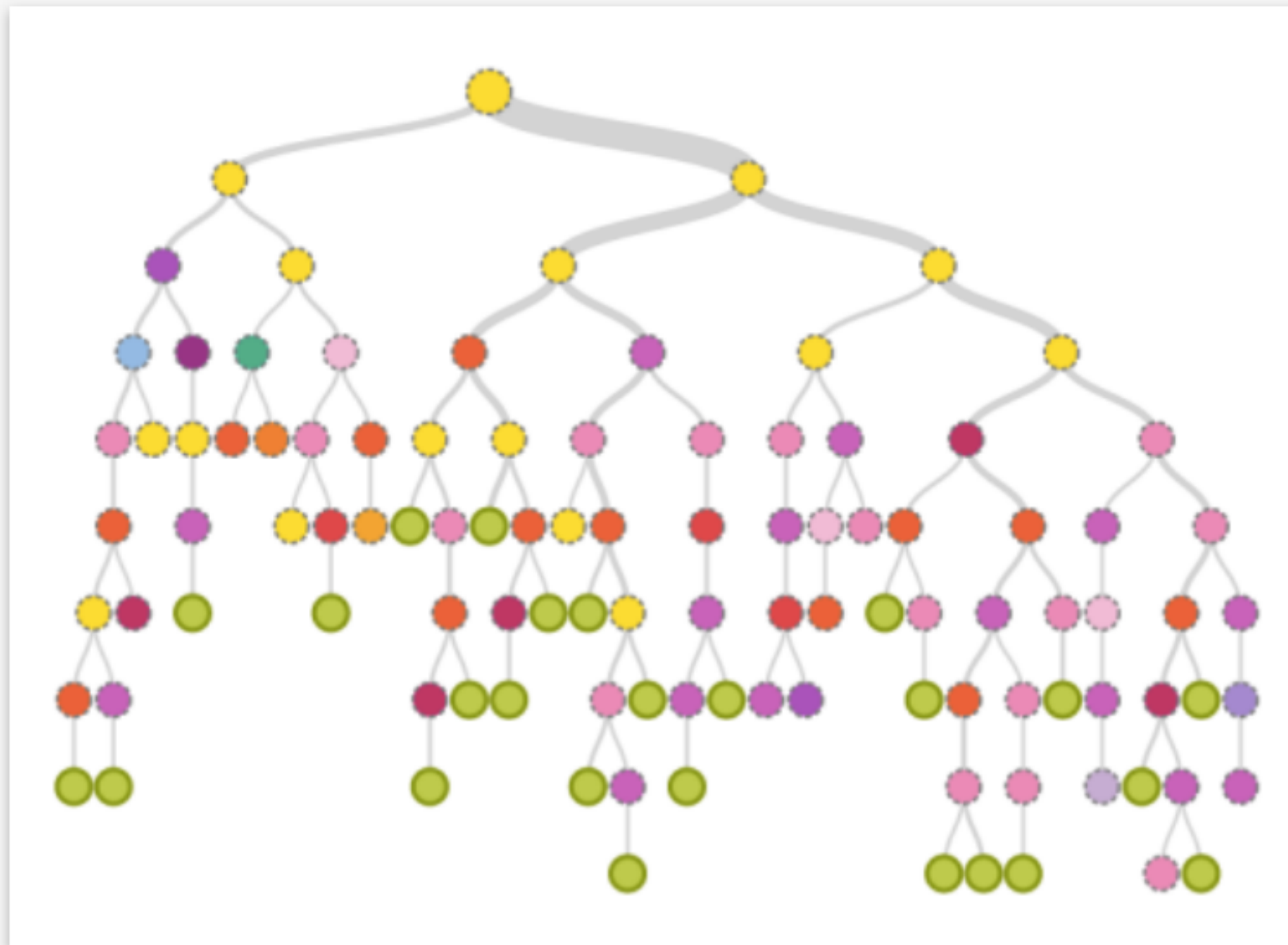
An simple example from industry



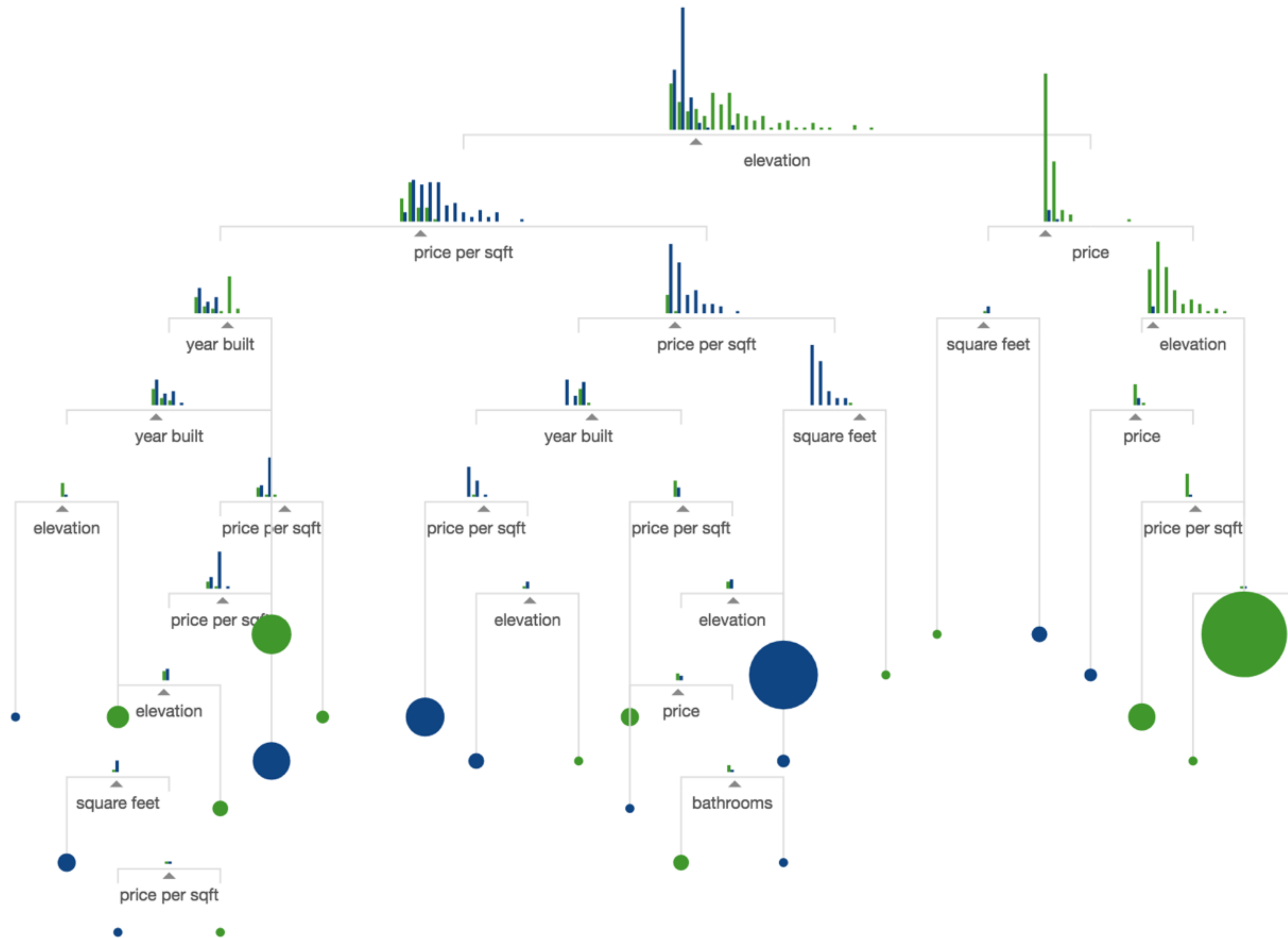
- BigML: <https://bigml.com/about>

Visualizing Decision Trees

- Typical, tree Visualization.
- Drawback: too wide to fit the display area.
- Collapsing, picking specific branches to drill down.
- Other visual design: e.g. SunBurst technique (for hierarchical structure, similar to nested pie charts) <https://www.cc.gatech.edu/gvu/ii/sunburst/>



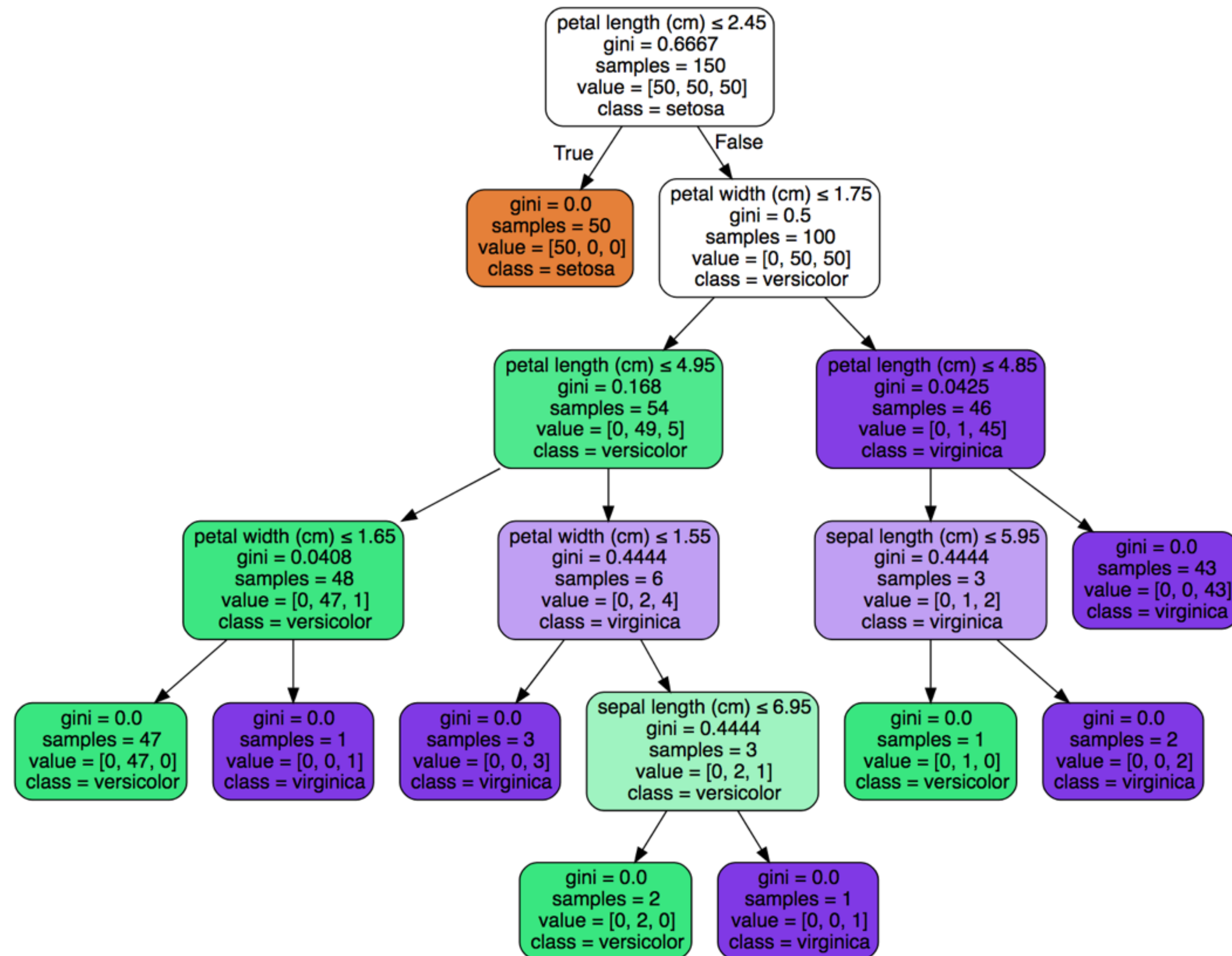
Decision Tree in a nutshell



By Tony Chu at noodle.io

<http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>

Scikit-learn's decision tree vis

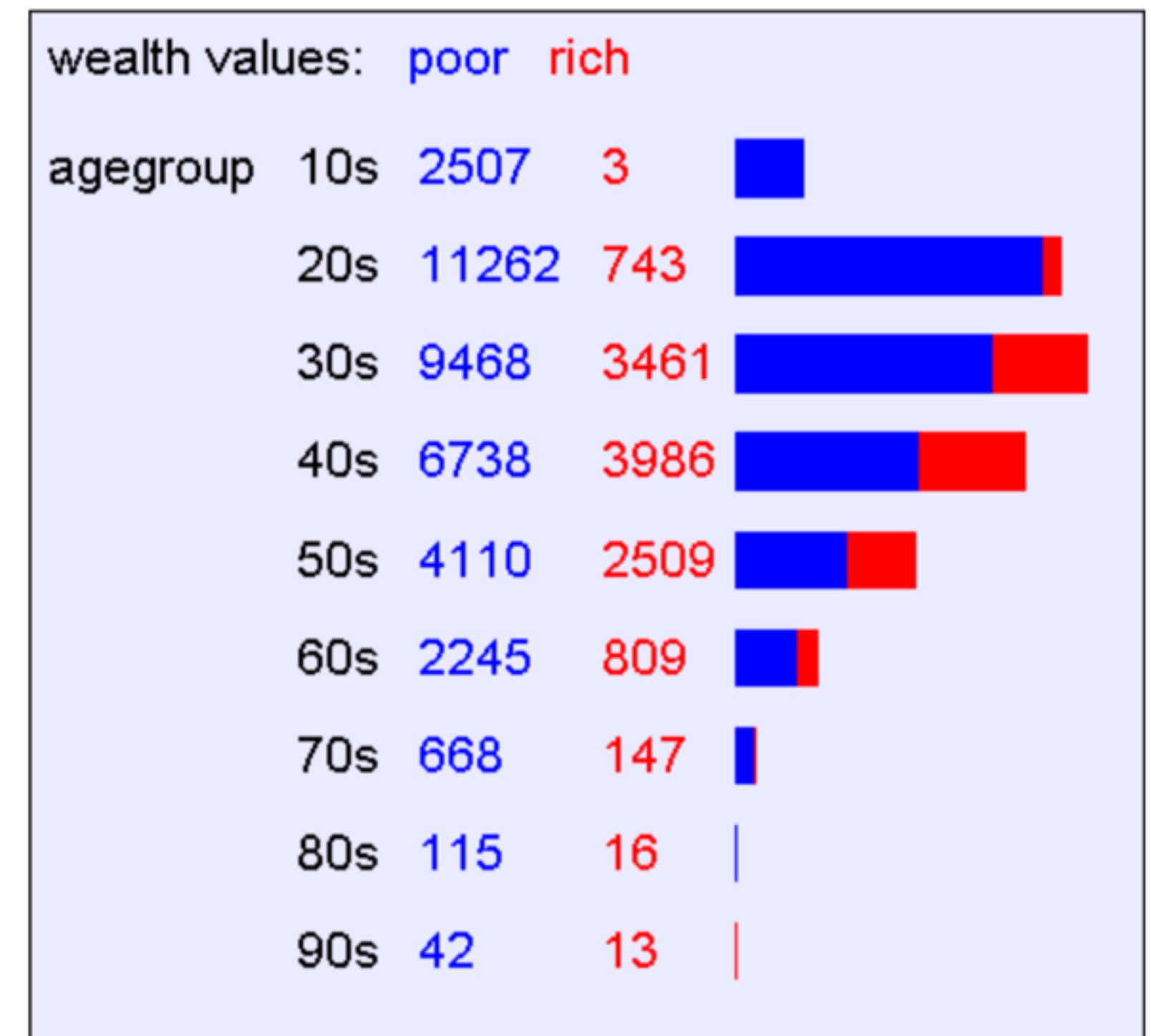


Decision tree on a high-level

- The notion of a **contingency table**: like 1D, 2D and 3D histograms

age	employe	education	edun	marital	...	job	relation	race	gender	hour	country	wealth
39	State_gov	Bachelors	13	Never_mar	...	Adm_cleric	Not_in_fam	White	Male	40	United_Stat	poor
51	Self_emp	Bachelors	13	Married	...	Exec_man	Husband	White	Male	13	United_Stat	poor
39	Private	HS_grad	9	Divorced	...	Handlers_c	Not_in_fam	White	Male	40	United_Stat	poor
54	Private	11th	7	Married	...	Handlers_c	Husband	Black	Male	40	United_Stat	poor
28	Private	Bachelors	13	Married	...	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	...	Exec_man	Wife	White	Female	40	United_Stat	poor
50	Private	9th	5	Married_sp	...	Other_serv	Not_in_fam	Black	Female	16	Jamaica	poor
52	Self_emp	HS_grad	9	Married	...	Exec_man	Husband	White	Male	45	United_Stat	rich
31	Private	Masters	14	Never_mar	...	Prof_speci	Not_in_fam	White	Female	50	United_Stat	rich
42	Private	Bachelors	13	Married	...	Exec_man	Husband	White	Male	40	United_Stat	rich
37	Private	Some_coll	10	Married	...	Exec_man	Husband	Black	Male	80	United_Stat	rich
30	State_gov	Bachelors	13	Married	...	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	...	Adm_cleric	Own_child	White	Female	30	United_Stat	poor
33	Private	Assoc_acc	12	Never_mar	...	Sales	Not_in_fam	Black	Male	50	United_Stat	poor
41	Private	Assoc_voc	11	Married	...	Craft_repa	Husband	Asian	Male	40	*MissingVar	rich
34	Private	7th_8th	4	Married	...	Transport	Husband	Amer_India	Male	45	Mexico	poor
26	Self_emp	HS_grad	9	Never_mar	...	Farming_fi	Own_child	White	Male	35	United_Stat	poor
33	Private	HS_grad	9	Never_mar	...	Machine_c	Unmarried	White	Male	40	United_Stat	poor
38	Private	11th	7	Married	...	Sales	Husband	White	Male	50	United_Stat	poor
44	Self_emp	Masters	14	Divorced	...	Exec_man	Unmarried	White	Female	45	United_Stat	rich
41	Private	Doctorate	16	Married	...	Prof_speci	Husband	White	Male	60	United_Stat	rich
:	:	:	:	:	:	:	:	:	:	:	:	:

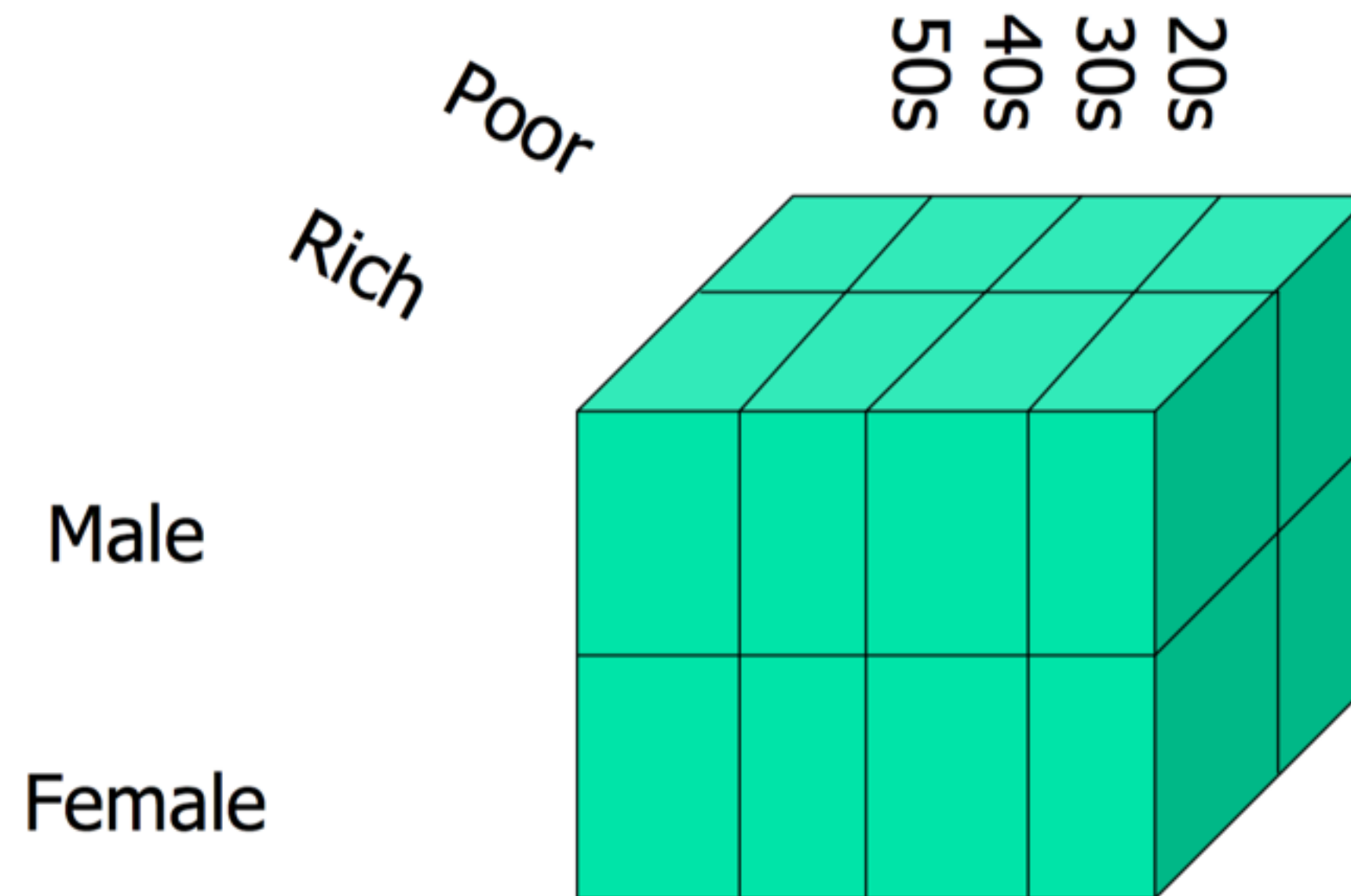
(agegroup,wealth)



2D contingency table

3D contingency table







- Goal: avoid manually looking at contingency tables
- For example, 100 variables, 161700 tables...
- Instead, using **information theory** to decide whether a pattern is interesting, such as **entropy** or **information gain**



Is a pattern interesting?

- Finding the attribute with the highest information gain

wealth values: **poor** **rich**

relation	Husband	10870	8846		$H(\text{wealth} \mid \text{relation} = \text{Husband}) = 0.992385$
	Not_in_family	11307	1276		$H(\text{wealth} \mid \text{relation} = \text{Not_in_family}) = 0.473439$
	Other_relative	1454	52		$H(\text{wealth} \mid \text{relation} = \text{Other_relative}) = 0.216617$
	Own_child	7470	111		$H(\text{wealth} \mid \text{relation} = \text{Own_child}) = 0.110192$
	Unmarried	4816	309		$H(\text{wealth} \mid \text{relation} = \text{Unmarried}) = 0.328606$
	Wife	1238	1093		$H(\text{wealth} \mid \text{relation} = \text{Wife}) = 0.997207$

$H(\text{wealth}) = 0.793844$ $H(\text{wealth} \mid \text{relation}) = 0.628421$
 $IG(\text{wealth} \mid \text{relation}) = 0.165423$

Information Gain

What is Information Gain used for?

Suppose you are trying to predict whether someone is going live past 80 years. From historical data you might find...

- $IG(\text{LongLife} \mid \text{HairColor}) = 0.01$
- $IG(\text{LongLife} \mid \text{Smoker}) = 0.2$
- $IG(\text{LongLife} \mid \text{Gender}) = 0.25$
- $IG(\text{LongLife} \mid \text{LastDigitOfSSN}) = 0.00001$

IG tells you how interesting a 2-d contingency table is going to be.

Entropy

General Case

Suppose X can have one of m values... V_1, V_2, \dots, V_m

$P(X=V_1) = p_1$	$P(X=V_2) = p_2$	$P(X=V_m) = p_m$
------------------	------------------	------	------------------

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X 's distribution? It's

$$\begin{aligned} H(X) &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_m \log_2 p_m \\ &= -\sum_{j=1}^m p_j \log_2 p_j \end{aligned}$$

$H(X)$ = The entropy of X

- "High Entropy" means X is from a uniform (boring) distribution
- "Low Entropy" means X is from varied (peaks and valleys) distribution

Conditional entropy

Specific Conditional Entropy $H(Y|X=v)$

X = College Major
Y = Likes "Gladiator"

Definition of Specific Conditional Entropy:

$H(Y|X=v)$ = The entropy of Y among only those records in which X has value v

Example:

- $H(Y|X=Math) = 1$
- $H(Y|X=History) = 0$
- $H(Y|X=CS) = 0$

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Conditional Entropy

X = College Major
Y = Likes "Gladiator"

Definition of Conditional Entropy:

$H(Y|X)$ = The average conditional entropy of Y

$$= \sum_j \text{Prob}(X=v_j) H(Y|X=v_j)$$

Example:

v_j	$\text{Prob}(X=v_j)$	$H(Y X=v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

$$H(Y|X) = 0.5 * 1 + 0.25 * 0 + 0.25 * 0 = 0.5$$

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Information Gain

Information Gain

X = College Major

Y = Likes "Gladiator"

Definition of Information Gain:

$IG(Y|X)$ = I must transmit Y .
How many bits on average
would it save me if both ends of
the line knew X ?

$$IG(Y|X) = H(Y) - H(Y|X)$$

Example:

- $H(Y) = 1$
- $H(Y|X) = 0.5$
- Thus $IG(Y|X) = 1 - 0.5 = 0.5$

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Learning a decision tree

- A Decision Tree is a tree-structured plan of a set of attributes to test in order to predict the output.
- To decide which attribute should be tested first, simply find the one with the highest information gain.
- Then recurse...

Decision tree on a high-level

- Tree structure
- Using the notion of **entropy** or **information gain** to choose which dimension to split
- Recurse

Learn more on decision tree

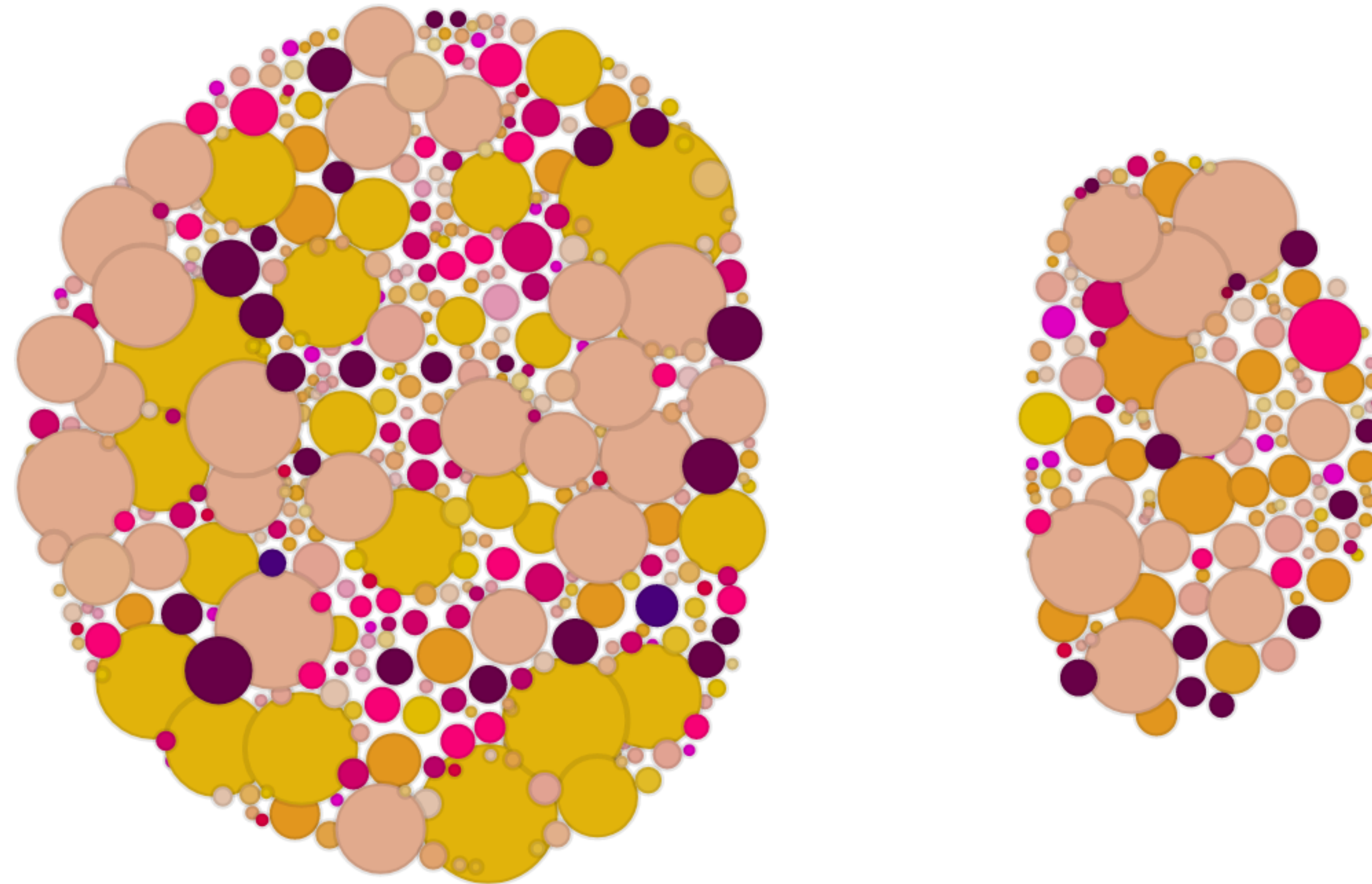
- Youtube, e.g. <https://www.youtube.com/watch?v=eKD5gxPPeY0>
- Decision tree tutorials
 - By Avinash Kak: <https://engineering.purdue.edu/kak/Tutorials/DecisionTreeClassifiers.pdf>
 - By Andrew Moore:
 - <http://www.cs.cmu.edu/~./awm/tutorials/dtree.html>
 - <http://www.cs.cmu.edu/~./awm/tutorials/infogain11.pdf>

Personal Vis Project...

Getting ready for Project 4

- Project 4 is due April 5th...I know it seems like a long time away
- However, project 4 has a flavor of personal visualization
- Start thinking about the following question:
 - What personal data of your own (or someone you know) would you like to visualize?
 - How do you get hold of the data?
 - Why is this important to you?

Example of personal vis



Visualize Gmail data

<http://luk3thomas.com/labs/gmail-archive-for-2013-20140224.html>



Thanks!

Any questions?

You can find me at: beiwang@sci.utah.edu

CREDITS

Special thanks to all people who made and share these awesome resources for free:

- ☐ Presentation template designed by [Slidesmash](#)
- ☐ Photographs by [unsplash.com](#) and [pexels.com](#)
- ☐ Vector Icons by [Matthew Skiles](#)

Presentation Design

This presentation uses the following typographies and colors:

Free Fonts used:

<http://www.1001fonts.com/oswald-font.html>

<https://www.fontsquirrel.com/fonts/open-sans>

Colors used

