Visualizing Ambiguity: A Grammar of Graphics Approach to Resolving Numerical Ties in Parallel Coordinate Plots

Denise Bradford

2026-01-01

Table of contents

1 Introduction

This proposal outlines a systematic approach to visually distinguish tied numerical values in multidimensional datasets by employing parallel coordinate plots (PCPs). Parallel coordinates, first popularized by Alfred Inselberg, are a powerful technique for investigating patterns across multiple attributes simultaneously (Inselberg 2009). However, when datasets contain exact numerical ties, the resulting overlapping lines in PCPs can obscure critical distinctions.

To address this, we propose a uniform method for collected spacing to tied values. The method will be integrated into the `ggpcp` package in R, ensuring a streamlined workflow for users seeking enhanced clarity in their parallel coordinate visualizations.

Importantly, our approach complements recent work on generalized parallel coordinate plots (GPCPs), an extension of PCPs that supports categorical variables (VanderPlas et al. 2023). The `ggpcp` package for R implements these GPCPs using a grammar of graphics framework, which seamlessly incorporates both continuous and categorical variables in a single parallel coordinate plot. One of the key contributions of that work is a robust tie-breaking mechanism for categorical variables, implemented through the `pcp_arrange()` function with methods including "from-left" and "from-right" hierarchical sorting. This ensures that individual observations can be traced across multiple dimensions, even when categories induce identical or "tied" values.

By adding multiple numerical tie-breaking techniques for continuous data—including our three deterministic approaches—we further refine GPCPs' capacity to handle the visualization of real-world datasets exhibiting many types of ties.

2 Background and Motivation

2.1 Parallel Coordinate Plots

Parallel coordinate plots assign each dimension of an n-dimensional dataset to a vertical axis arranged in parallel (Wegman 1990). Each observation is drawn as a polyline connecting its values on these axes, providing a visual representation that can illuminate underlying data structures.

2.2 Numerical Ties and Visual Overlap

When multiple observations share the same value in a given dimension, their polylines perfectly overlap, creating "visual collisions." This masks information about distribution, density, or potential outliers. The treatment of ties is an aspect not generally addressed in the original parallel coordinate plots of Inselberg (1985) and Wegman (1990). However, the `ggpcp` implementation has demonstrated that careful tie-handling is essential for both continuous and categorical variables.

Introducing a small offset ("jitter") to these tied values can mitigate overlap without distorting the overall relationships in the data (Peng, Ward, and Rundensteiner 2004). In the context of generalized parallel coordinate plots, the `ggpcp` package separates data management from visual rendering into three distinct components: variable selection and reshaping, scaling of axes, and treatment of ties in categorical axes (VanderPlas et al. 2023).

2.3 Existing Solutions for Categorical Ties in ggpcp

The `ggpcp` package currently addresses categorical ties through sophisticated tie-breaking algorithms. The package implements hierarchical sorting through the `pcp_arrange(data, method, space)` function, with two primary methods: "from-left" and "from-right", meaning that tie breaks are determined hierarchically by variables' values from the specified direction. The parameter `space` specifies the amount of the y-axis to use for spacing between levels of categorical variables, with a default of 5% of the axis used for spacing.

This hierarchical sorting approach serves as "external cognition," the additional computational processing reduces the cognitive load required to untangle overlapping lines in the parallel coordinate plot. The categorical tie-breaking creates equispaced tie-breaking that reduces line crossings and allows users to follow individual observations from left to right through the plot even for categorical variables.

2.4 The Challenge of Visualizing Mixed-Type Data

Parallel coordinate plots (PCPs) have been established as valuable tools for exploratory data analysis of high-dimensional numerical data since their introduction (Inselberg 1985; Wegman 1990). However, the use of PCPs is fundamentally limited when working with categorical variables or mixed categorical-continuous data. As VanderPlas et al. (2023) note in their introduction to generalized parallel coordinate plots (GPCPs), existing solutions for categorical values become insufficient when attempting to maintain visual continuity across both data types.

The treatment of ties, multiple observations sharing the same value, represents a critical design decision that affects both perceptual effectiveness and analytical utility. As VanderPlas et al. (2023) observe, "The treatment of ties is an aspect not generally addressed in the original parallel coordinate plots of Inselberg (1985) and Wegman (1990). We have found a need to deal with ties" (p. 6). This observation extends naturally from categorical to numerical variables.

The ability to follow individual observations is central to the analytical power of PCPs, enabling users to identify patterns, outliers, and relationships that span multiple variables simultaneously.

2.4.1 Step 1: Identify the Parallel Axes

Begin by identifying each vertical axis in the plot. Each axis represents one variable from the dataset. The axes are typically arranged from left to right, and the order may be determined by the data analyst to highlight specific relationships or minimize visual clutter.

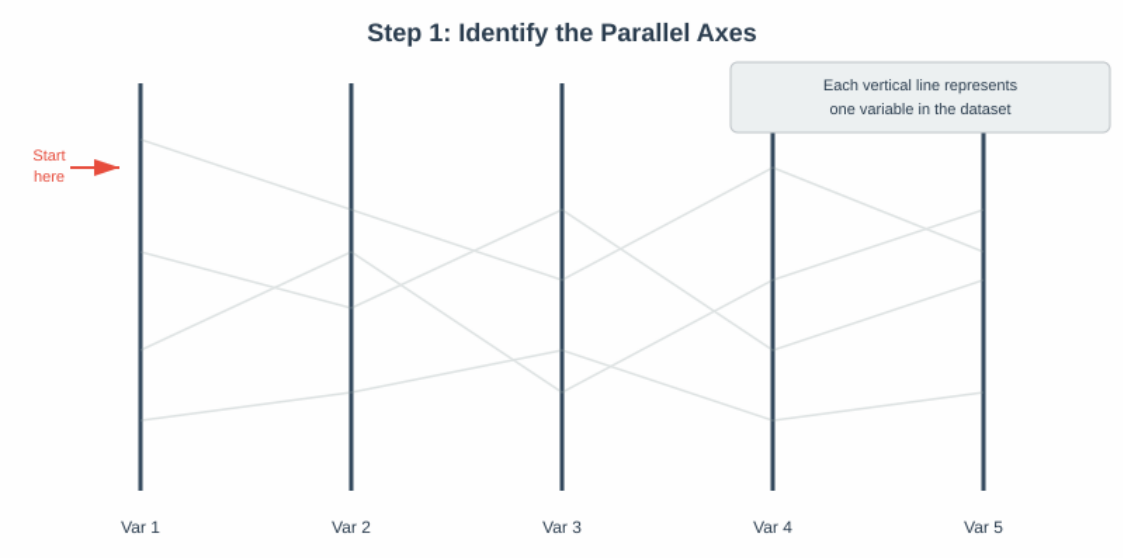

Figure 1: The parallel axes form the structural framework of the visualization.

2.4.2 Step 2: Locate the Starting Point

Find the observation of interest on the leftmost axis. The vertical position indicates the scaled value of that observation for the first variable. If you are examining a highlighted or color-coded observation, look for its distinctive marker at this starting position.

2.4.3 Step 3: Follow the Line Segment

Trace the line segment from the starting point to its intersection with the next axis. The human visual system naturally follows smooth, continuous paths due to the Gestalt principle of good continuation. This principle allows viewers to perceive connected lines as unified objects, making it easier to track observations across multiple variables.

2.4.4 Step 4: Read Values at Intersections

At each axis intersection, the vertical position of the line indicates the observation's value for that variable. Read these values to understand how the observation changes across different dimensions of the data. The slope of line segments between axes provides information about the relationship between consecutive variables for that specific observation.

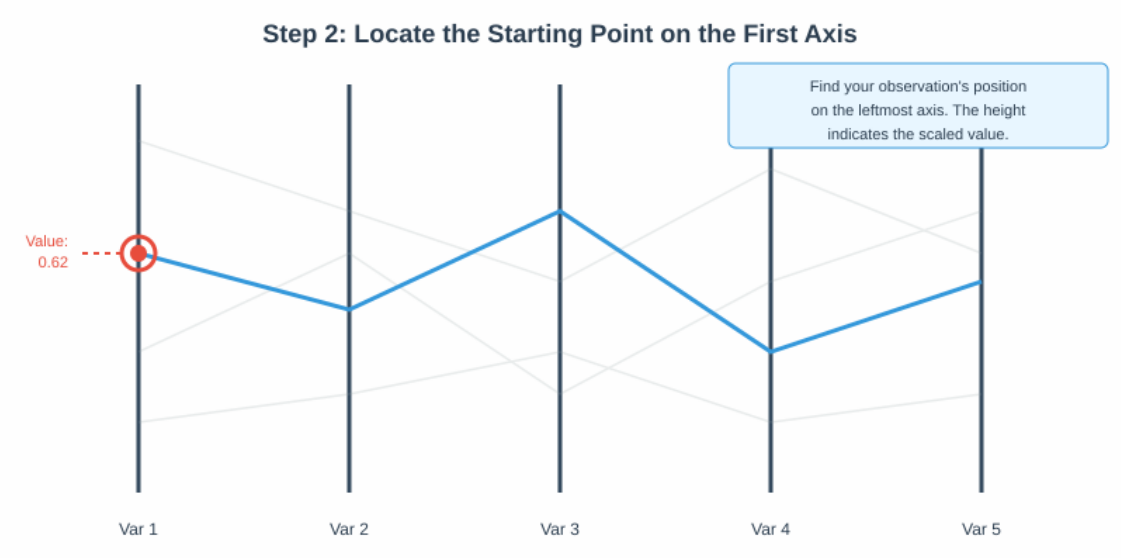

Figure 2: The starting point is identified on the first axis with its corresponding value.

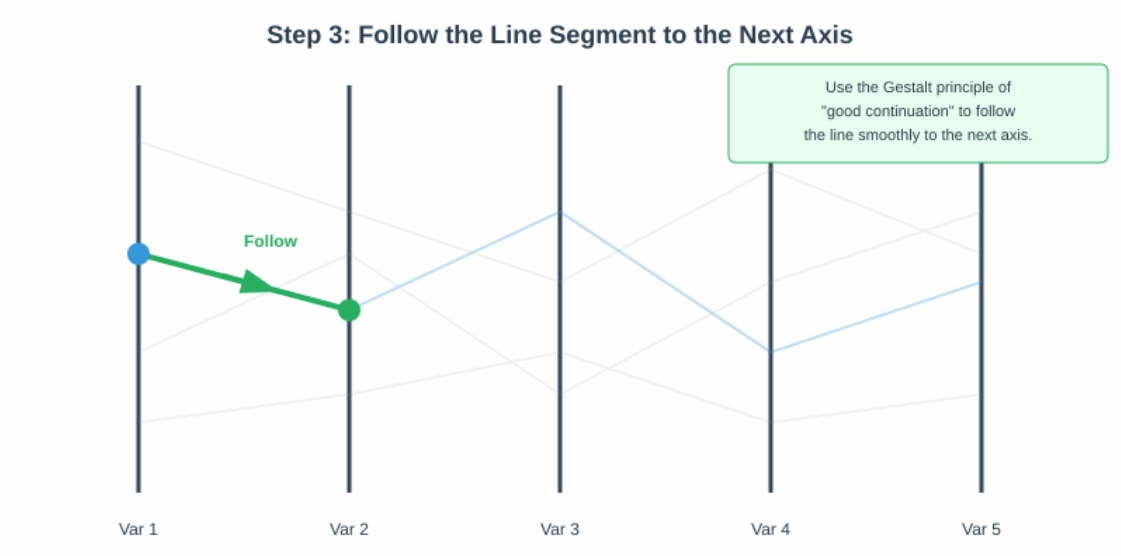

Figure 3: Following the line segment uses the Gestalt principle of good continuation.

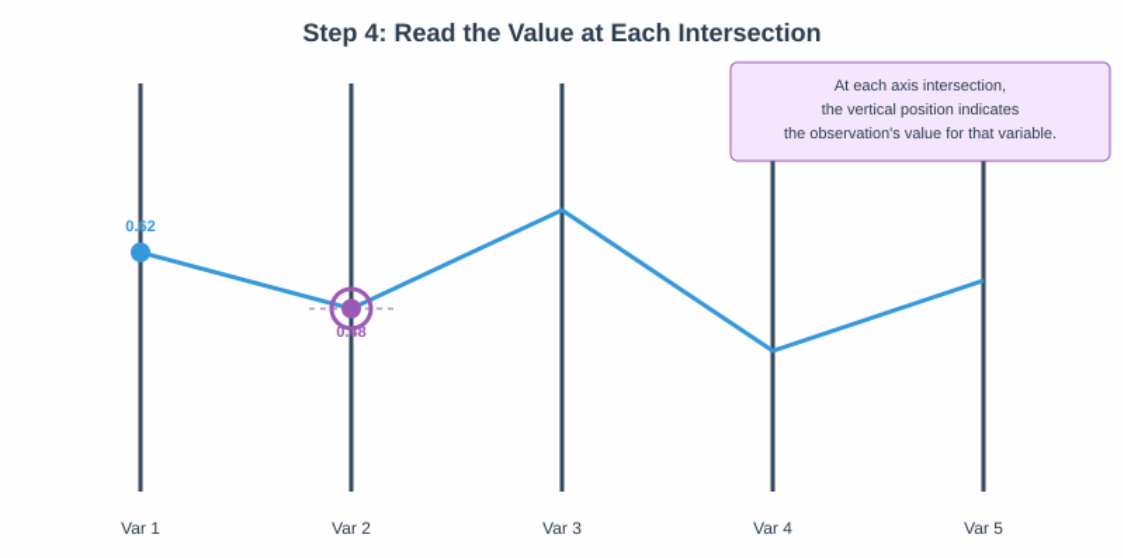

Figure 4: Values are read as approximate scaled value relative to the total range.

2.4.5 Step 5: Continue Across All Axes

Repeat the tracing process for each consecutive pair of axes until reaching the rightmost axis. By following the complete path, you obtain a comprehensive view of how that particular observation behaves across all measured variables. The vertical positions along each axis represent scaled values that can be interpreted as quantiles when the data are appropriately transformed. This enables identification of unique characteristics, cluster membership, or outlier status.

Figure 5: The complete traced path reveals the observation's values across all variables.

2.4.6 Current State: ggpcp's Categorical Tie-Breaking

The ggpcp package implements a sophisticated tie-breaking algorithm for categorical variables that maintains individual observation traceability. The approach spaces observations evenly within categorical levels:

> "All observations are spaced out evenly. This results in a natural visualization of the marginal frequencies along each axis (additionally enhanced by the light gray boxes grouping observations in the same category) that is not as prominent in the previous three panels. The ordering of the observations within the level is such that a minimal number of line crossings occurs between the axes." (p. 11)

The algorithm achieves this through hierarchical sorting implemented in `pcp_arrange(data, method, space)`, where the `space` parameter specifies the proportion of the y-axis used for spacing between categorical levels (default 5%). This optimization can be formalized as:

$$d_i = \frac{S_i - S_i^- - S_i^+}{n_i - 1}$$

where:

- S_i is the total space allocated to category i
- S_i^- is the spacing below category i
- S_i^+ is the spacing above category i
- n_i is the number of observations in category i
- d_i is the optimal spacing distance between consecutive observations

2.4.7 Alternative Approach: Hammock Plots

Hammock plots, introduced by Schonlau Schonlau and Yang (2024), take a fundamentally different approach to handling both categorical and numerical variables. Rather than using individual lines, hammock plots employ two-dimensional boxes to connect adjacent axes, with box width proportional to the number of observations.

As Schonlau and Yang (2024) describes:

> "Like a parallel coordinate plot, the axes are aligned parallel to one another. Categories of adjacent variables are connected by boxes. (The boxes shown are parallelograms; I use the word boxes for simplicity). The width of boxes is proportional to the number of observations." (p. 3)

For numerical variables specifically, hammock plots maintain constant-width boxes throughout the visualization. Schonlau and Yang (2024) explains the spatial constraint this imposes:

> "When treating this variable as numerical, the range from 0 to 20 leaves 1/21th of the space for each unit length. Consequently, the widths of the boxes have to be more frugal." (p. 21)

This creates a fundamental trade-off: hammock plots explicitly encode frequency through box width but sacrifice individual observation traceability. As Schonlau and Yang (2024) notes, "For small data sets, GPCP plots beautifully show all individual observations whereas hammock plots require highlighting to feature individual observations" (p. 19).

2.4.8 Visual Comparison: Triangles vs. Boxes

A key visual difference emerges when connecting categorical to numerical variables. VanderPlas et al. (2023) observe:

> "When many observations have the same value for a categorical and an adjacent numerical variable, the corresponding area looks like a triangle… Notice the lines/boxes between the variables hospitalizations and comorbidities in the GPCP (Figure 13) and hammock plots (Figure 2). Most of the observations are in the boxes leading from hospitalizations=0 to either comorbidities=0 or comorbidities=1. This is far more obvious in the hammock plot than in the GPCP plot." (p. 19)

This observation suggests that each approach has perceptual advantages in different contexts, motivating rigorous comparative evaluation grounded in perceptual science.

Understanding the perceptual and practical challenges of parallel coordinate plots provides essential context for our tie-resolution methods. This section synthesizes empirical evidence from the visualization literature that directly informs our design decisions and evaluation strategy.

2.4.9 The Line Width Illusion and Perceptual Constraints

Both approaches must contend with perceptual illusions that affect visual interpretation. Schonlau and Yang (2024) discusses the line width illusion:

> "The distance between two parallel lines is perceived at a right angle rather than as the vertical distance between the lines (Wallgren et al., 1996; Tufte, 2001)… The line width illusion is part of the family of Müller-Lyer illusions where two lines of same length appear to be of different lengths." (p. 8)

Hofmann and Vendettuoli (2013) further identify the "reverse line width illusion," where centering of lines creates a contextual cue encouraging evaluation of line widths using vertical rather than orthogonal measures. As they demonstrate empirically, this illusion can lead to systematic biases in frequency estimation.

The equispaced line approach proposed here may inherently avoid these illusions by maintaining parallel lines at consistent vertical spacing rather than using area-based encoding, though this requires empirical validation.

3 Handling Numerical Ties

A significant challenge arises when multiple observations share identical values on an axis. In traditional PCPs, these observations converge to a single point, creating overlapping lines that make individual tracking impossible. The ggpcp package addresses this through the tie_spread algorithm.

3.1 The Problem: Overlapping Lines

When multiple observations have the same value for a variable, their lines converge to a single point on that axis. This convergence creates visual clutter and breaks the continuity needed for individual observation tracking.

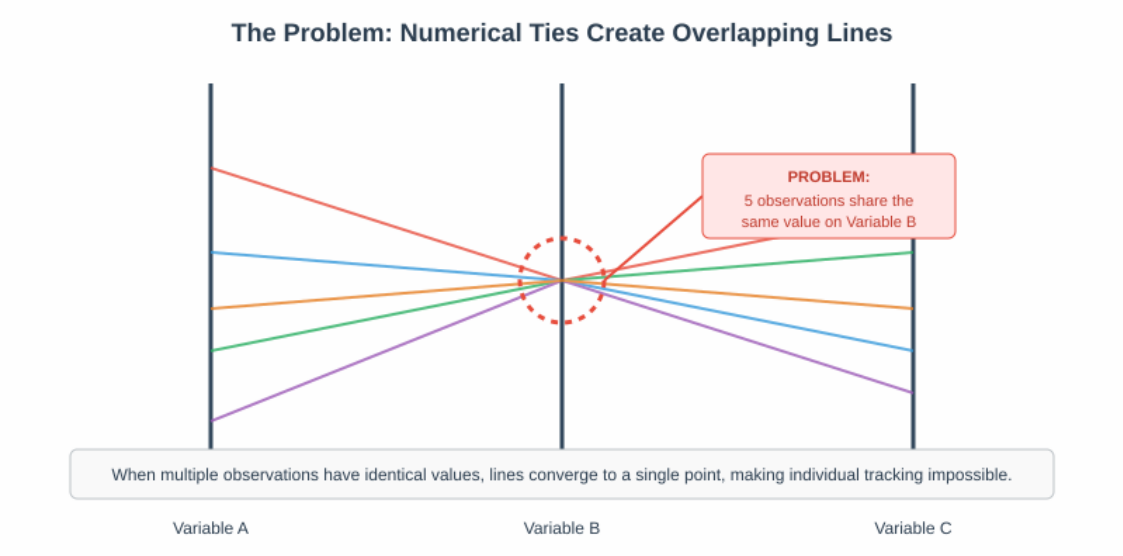

Figure 6: Numerical ties cause lines to converge at a single point.

3.2 The Solution: Tie Spreading

The `tie_spread` algorithm separates tied observations by distributing them evenly within a small range centered on their original value. By default, this range is set to maximize the space available for tie breaking while preserving the approximate position of values and maintaining visual separation.

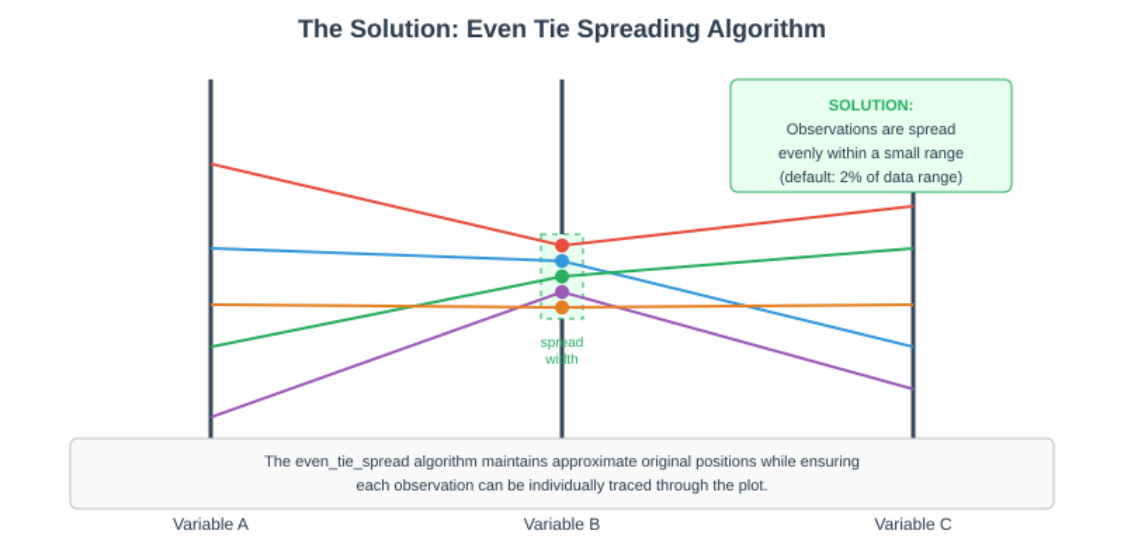

Figure 7: Tie spreading maintains visual separation while preserving approximate positions.

3.3 Hierarchical Sorting for Minimal Crossings

Beyond simply spreading tied values, the ggpcp package implements hierarchical sorting to minimize line crossings as in categorical tie breaking.

By ordering observations within a tie group based on their values on adjacent axes, the package leverages the Gestalt principle of common fate. This strategy ensures that observations with similar trajectories are positioned in close proximity, creating cohesive visual bands. These bands facilitate the perception of distinct groups as they move together through high-dimensional space, reducing visual noise and highlighting underlying patterns.

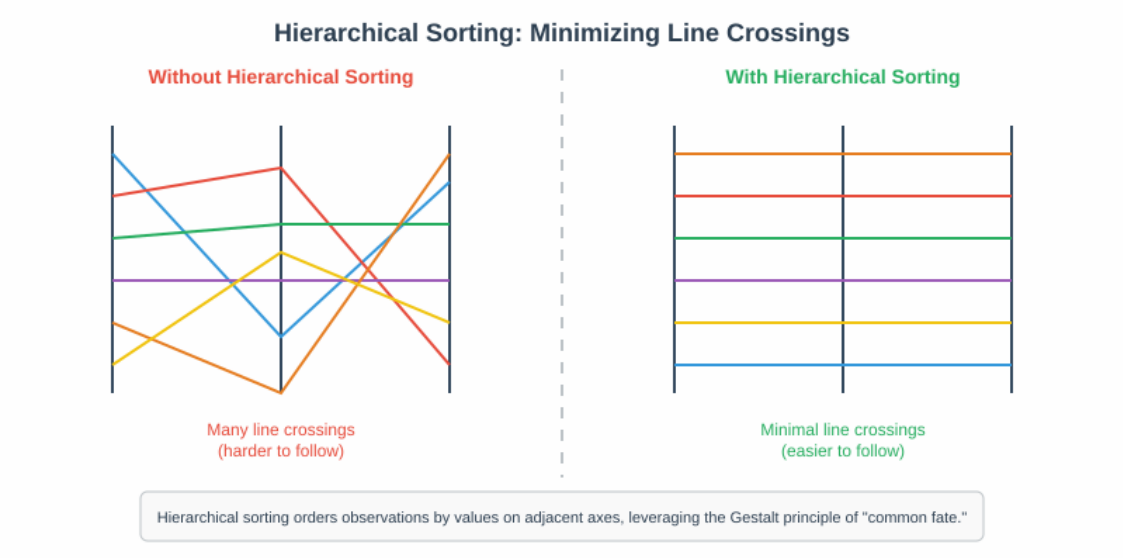

Figure 8: Hierarchical sorting reduces line crossings, making patterns easier to perceive.

3.4 Theoretical Framework: Perception-Driven Design

Colin Ware's "Information Visualization: Perception for Design" provides a comprehensive framework for understanding how humans process visual information (Ware 2012, 3rd ed., pp. 20-21, Figure 1.11). Ware describes visual perception as occurring in three distinct stages, each with specific implications for visualization design.

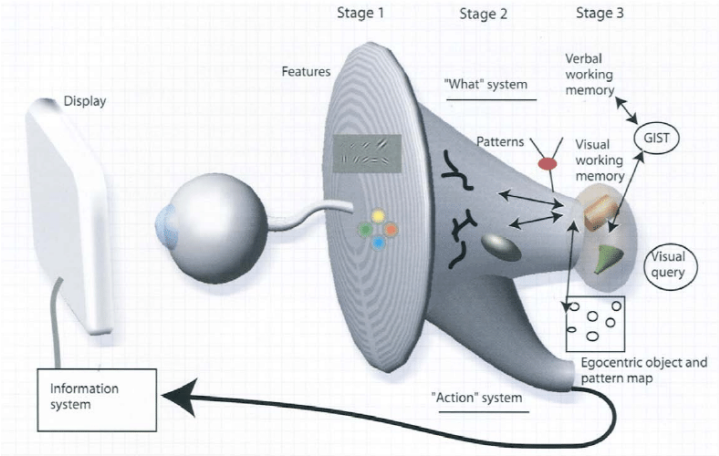

Figure 9: Colin Ware's three-stage model of visual information processing, illustrating the flow from low-level feature extraction to goal-directed processing (adapted from Ware, 2012).

3.4.0.1 Stage 1: Parallel Processing of Low-Level Properties

The first stage involves rapid, parallel extraction of basic visual features across the entire visual field. As Ware describes, this preattentive processing occurs in under 500 milliseconds and requires no conscious effort. Four fundamental categories of preattentive visual properties exist:

Form is one of the most basic. It includes simple geometric cues—orientation, length, width, and size—as well as more structured cues like curvature and the way marks are grouped. Form also depends on how clearly something is rendered (sharp vs. blurred) and whether boundaries are emphasized through added marks or enclosing shapes.

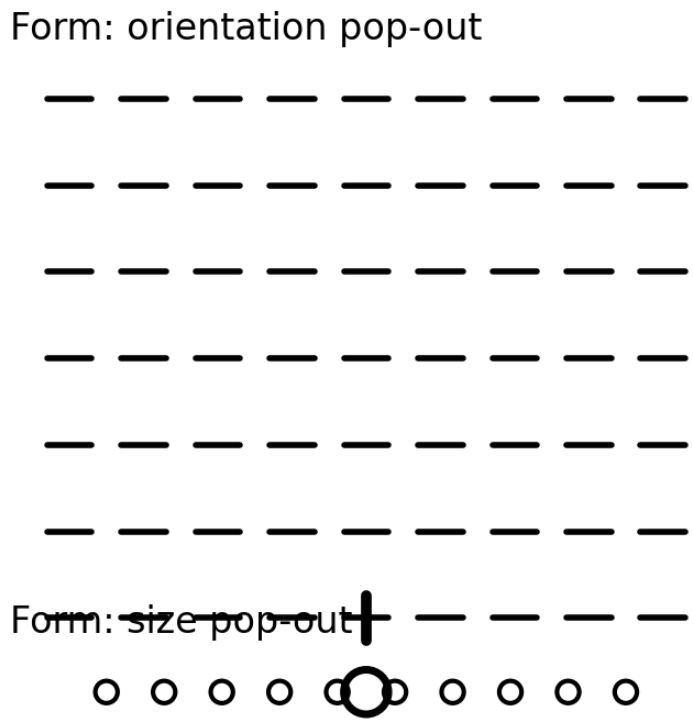

Figure 10: Form-based pop-out: orientation and size.

Color provides another immediate signal. Differences in hue, brightness (intensity), and saturation let viewers separate categories or perceive gradients quickly, typically without much mental work.

Spatial position underlies nearly every visualization. At its simplest, it's where something lands on a 2D plane via (x, y) coordinates. Position can also suggest depth or distance through stereoscopic cues or shading that makes shapes appear concave or convex.

Finally, **motion** is especially effective at grabbing attention. Flicker or directional movement stands out strongly and can quickly indicate changes or relationships in data that unfold over time.

These preattentive properties are critical for visualization design because they allow information to be perceived "at a glance" without requiring serial visual search. As Ware notes, it is easy to spot a single hawk in a sky full of pigeons, but if the sky contains many types of birds, the hawk becomes much harder to see—the distinctiveness of preattentive properties diminishes as visual variety increases.

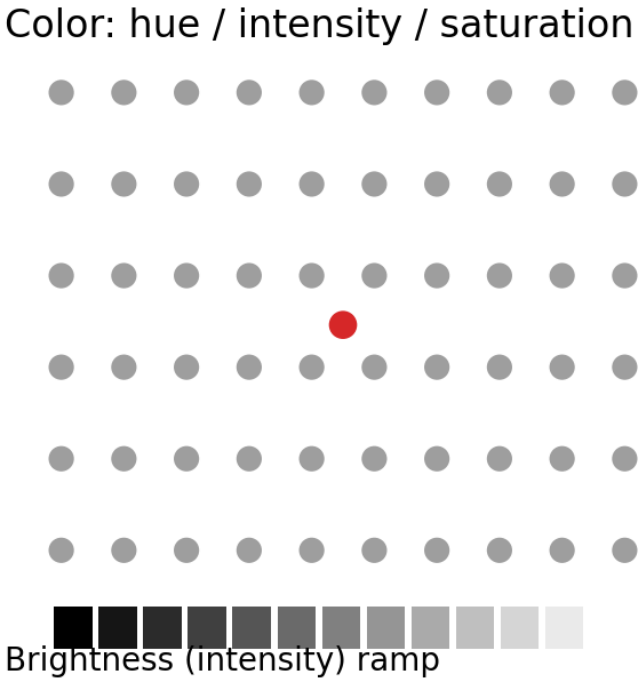

Figure 11: Color-based pop-out: hue; plus an intensity ramp.

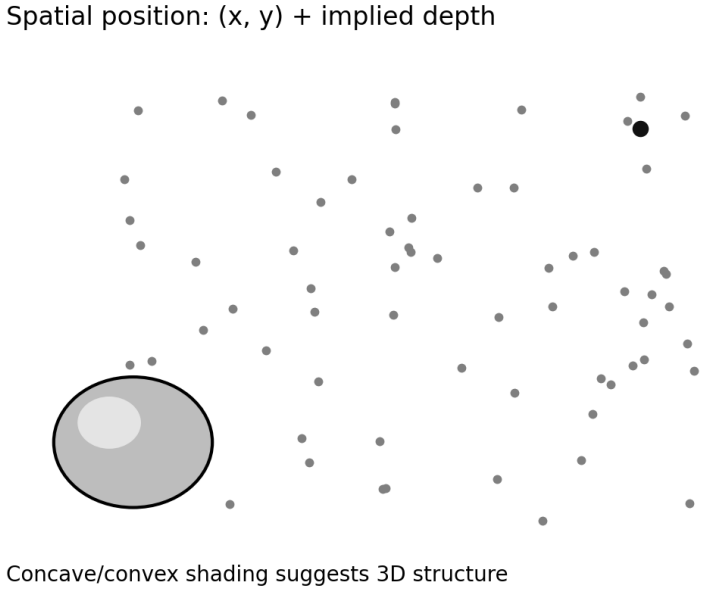

Figure 12: Spatial position (x,y) and implied depth via shading.

While Colin Ware's model provides a foundational understanding of the physiological stages of vision, researchers like Lace Padilla have expanded this framework to include the complex cognitive processes involved in decision-making. Padilla et al. (2018) integrates Dual-Process Theory, which suggests that users oscillate between fast, intuitive "System 1" thinking and slow, deliberative "System 2" reasoning when interpreting data.

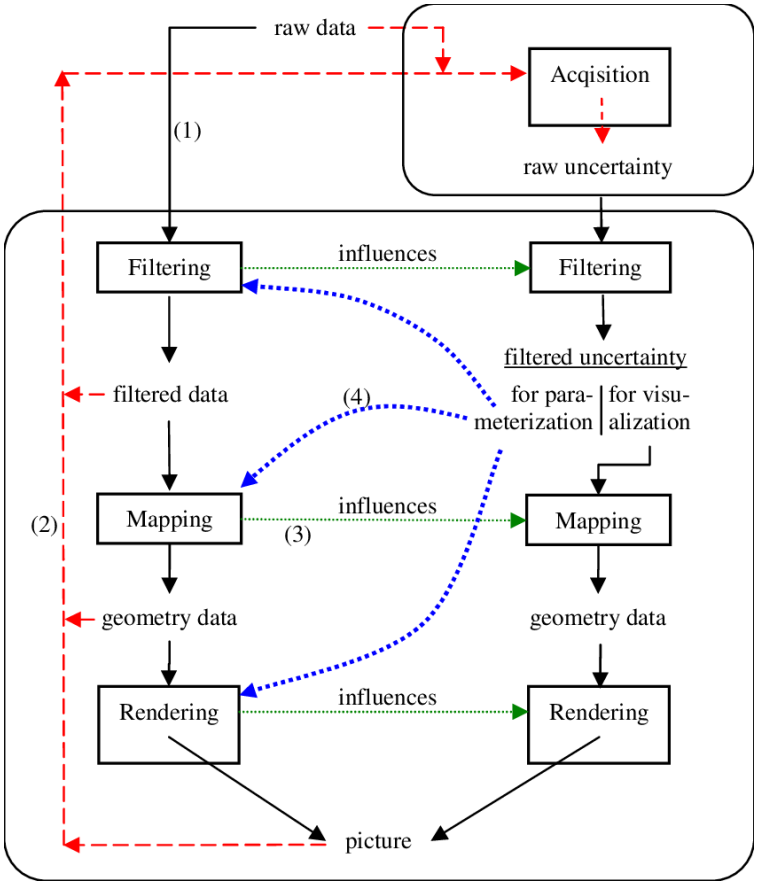

Figure 13: Lace Padilla's cognitive model (Padilla et al., 2018).

Ware's framework is primarily a bottom-up, perceptual model that describes how the physical properties of a graphic (like color and orientation) are mechanically processed by the eye and brain. In contrast, Padilla's model is a cognitive-integrative model that places greater emphasis on the user's prior knowledge and the dual nature of reasoning. While Ware's "Stage 3" (Sequential Goal-Directed Processing) touches on cognition, Padilla delves deeper into how a user's working memory and long-term expertise can override or bias the initial perceptual signals. For a package like ggpcp, Ware's model explains why categorical tie-breaking helps the eye follow a line (low-level tracking), whereas Padilla's model explains how those clear paths allow the user to transition from an intuitive "gut feeling" about a trend to a rigorous, analytical conclusion about the high-dimensional data.

3.4.0.2 Stage 2: Pattern Perception and Gestalt Principles

The second stage involves active pattern perception processes that segment the visual scene into coherent regions. The Gestalt laws of perceptual organization, articulated by German psychologists in 1912, remain fundamental principles for visualization design:

- **Proximity:** Objects near each other are perceived as belonging together
- **Similarity:** Objects sharing visual properties (color, shape, size) are grouped perceptually
- **Connectedness:** Objects connected by lines are seen as related

- **Continuity:** The visual system prefers interpretations with smooth, continuous paths
- **Symmetry:** Symmetric patterns are more easily perceived and remembered
- **Closure:** Incomplete shapes are perceptually completed
- **Common Fate:** Objects moving together are perceived as a group
- **Relative Size:** Smaller elements are perceived as figures against larger backgrounds

VanderPlas et al. (2023) explicitly invoke these principles in their design rationale for ggpcp:

> "By reducing the number of line crossings at non-axis points simplifies the plot, reducing the overall cognitive load required to 'untangle' (literally and metaphorically) the individual observations and leveraging the Gestalt principle of common fate." (p. 4)

The principle of **good continuation** is particularly relevant to parallel coordinate plots: can viewers smoothly follow individual lines through numerical ties, or do visual discontinuities disrupt the perceptual flow?

3.4.0.3 Stage 3: Visual Working Memory and Attention

The third stage involves conscious attention and visual working memory, which holds only a limited number of objects (typically 3-4) for active processing. As Ware emphasizes, the severe capacity limitations of visual working memory mean that effective visualizations must reduce cognitive load by leveraging the powerful parallel processing of stages 1 and 2.

This three-stage model has direct implications for evaluating tie-breaking strategies:

Equispaced Lines:

- Leverage preattentive orientation and position
- Support continuous line following (good continuation)
- Implicitly encode frequency through density (requires stage 2 processing)

Constant-Width Boxes:

- Leverage preattentive size and area
- Explicitly encode frequency (reduced cognitive load)
- May disrupt line continuity (challenges good continuation)

3.4.1 Preattentive Processing and Visual Search

A key distinction between the two approaches concerns the nature of visual search required for different analytical tasks. Ware distinguishes between:

- **Parallel Search (Preattentive):** Target pops out immediately, search time independent of number of distractors
- **Serial Search (Attentive):** Search time increases linearly with number of items, requiring conscious attention

For tracing individual observations through equispaced lines, the task may benefit from preattentive processing of orientation and position. Each observation maintains a consistent visual identity as

a continuous line with specific orientation. In contrast, tracing through constant-width boxes requires conscious attention to reconstruct the path, as the area-based encoding does not support preattentive line following.

However, for frequency estimation tasks, constant-width boxes may have an advantage. Ware notes that area is preattentively processed, allowing immediate perception of relative quantities. Equispaced lines require inferring frequency from density, which is a second-order visual property requiring pattern perception (stage 2).

3.4.2 External Cognition and Computational Offloading

Scaife and Rogers (1996) concept of "external cognition":

> "While hierarchical sorting requires additional computations relative to the jittering or equally spaced solutions in Figure 5, this extra processing serves as 'external cognition' [Scaife and Rogers, 1996] - the additional computer time reduces the cognitive load required to untangle the lines" (p. 11)

This principle aligns with Ware's emphasis on reducing working memory load through careful visual design. By computing optimal line positions algorithmically, the equispaced approach offloads cognitive work from the viewer. Rather than the viewer mentally sorting and organizing lines, the visualization presents pre-organized information that leverages natural perceptual grouping.

Hammock plots achieve a different form of cognitive offloading by aggregating observations into boxes, explicitly computing and displaying frequency information. The trade-off is between individual-level detail (equispaced lines) and aggregate-level clarity (boxes).

3.4.3 Gestalt Principles and Line Continuity

The principle of **good continuation** states that the human visual system preferentially perceives smooth, continuous contours over interpretations requiring abrupt changes in direction. This principle directly impacts the effectiveness of parallel coordinate plots for tracing individual observations.

Equispaced lines maintain perfect continuity—each observation is represented by a continuous polyline from the leftmost to rightmost axis. When observations are tied at a numerical value, the spacing ensures that lines remain visually distinct without requiring discontinuities.

Constant-width boxes, in contrast, represent observations as area segments rather than lines. While this encoding effectively communicates aggregate quantities, it disrupts the continuity principle. To trace an individual observation through a box, viewers must mentally reconstruct the implied path, engaging conscious attention and working memory.

This distinction becomes critical as dataset size increases. With many observations, equispaced lines may appear as ribbon-like bands, but the continuity principle still applies—viewers can perceive flow patterns even when individual lines are indistinguishable. Boxes maintain their aggregate interpretation but lose any individual-level information.

3.4.4 The Challenge of Area Perception

While area is preattentively processed, humans are notoriously poor at accurately comparing areas, especially when aspect ratios vary. This limitation is well-documented in Cleveland and McGill

(1984) hierarchy of elementary perceptual tasks, which ranks position along a common scale as most accurate, followed by length, angle, slope, area, and volume.

Hammock plots rely on area perception for frequency encoding. Observers must compare box widths (area divided by length) to estimate relative frequencies. The line width illusion and reverse line width illusion Hofmann and Vendettuoli (2013) further complicate this judgment by creating systematic biases.

Equispaced lines instead rely on density estimation—counting or estimating the number of lines within a region. While density estimation is also challenging, it may be less susceptible to geometric illusions. Additionally, density provides a more continuous encoding: frequencies can be estimated at any scale, from individual line counts to overall patterns.

3.4.5 Visual Clutter and Information Density

Ware discusses the concept of "visual clutter" as a fundamental limitation on visualization effectiveness. Clutter occurs when too many visual elements compete for attention, overwhelming the viewer's ability to extract meaningful patterns.

The two approaches differ in how they manage clutter:

Equispaced Lines:

- Clutter increases with number of observations
- Hierarchical sorting minimizes line crossings (reduces clutter)
- Graceful degradation: lines → ribbons → filled areas as density increases
- Individual observations remain theoretically accessible

Constant-Width Boxes:

- Clutter depends on number of unique value combinations
- Aggregation inherently reduces clutter
- Information loss: individual observations not accessible
- Clear representation of bivariate contingency tables

Dennig et al. (2021) formalize clutter metrics for parallel sets visualizations, measuring ribbon overlap, crossing angles, and ribbon width variance. These metrics can be adapted to compare the two approaches quantitatively.

3.5 Problem Statement and Research Questions

3.5.1 Gap in Research

Present implementations show a split:

1. **ggpcp**: This method uses Gestalt principles of continuity and common fate to break ties between categorical variables in a smart way. However, when numbers are tied, the lines overlap.
2. **Hammock plots**: All variable types have boxes with the same width, which makes frequency encoding clear, but you can't follow individual observations without highlighting them.

There is no current research that thoroughly evaluates these methodologies from the perspective of perceptual science or adapts ggpcp's optimization framework to numerical variables.

This gap is important because:

- Mixed-type datasets are common in modern data analysis.
- Different ways of visualizing data may work best for different types of analysis.
- Perceptual trade-offs have not been quantified or empirically validated.
- Design choices are not based on established principles of human visual perception.

3.5.2 Suggested Fix

We suggest that ggpcp's categorical tie-breaking algorithm be expanded to include numerical variables, using the same optimization framework that keeps things evenly spaced.

For a number v with n tied observations:

1. Find out how much space is available by looking at the values and data density of nearby items.
2. Use spacing optimization: $spacing = \frac{available\_space}{n-1}$
3. Organize observations in a hierarchical way and spread them out so they are perpendicular to the axis direction.
4. Keep the rules of visual continuity in mind by making sure that lines flow smoothly from one to the next.

This creates a single framework where the same algorithm deals with both categorical and numerical ties. This could make things easier on the brain by keeping the visual grammar consistent and keeping the visual continuity across mixed-type data.

3.5.3 Main Research Question

Is the equispaced line method better for numerical ties than the constant-width box method in hammock plots, based on how people see things?

This question breaks down into smaller, more specific questions based on perceptual theory:

Q1 (Preattentive Processing): Do equispaced lines facilitate preattentive detection and visual search more effectively than constant-width boxes for individual observation tracking?

Q2 (Pattern Perception): How do the two methods use Gestalt principles in different ways, and which one works better for seeing relationships in multivariate data?

Q3 (Memory and Cognition): Which method is better at lowering the cognitive load and visual working memory needs for everyday analytical tasks?

Q4 (How Often You Think): How well can viewers guess frequencies when using implicit density (equispaced lines) instead of explicit area encoding (boxes)?

Q5 (Scalability): How do the pros and cons of perception change when the size of the dataset and the distribution of ties change?

3.6 Hypothesis Grounded in Perceptual Theory

3.6.1 Hypothesis: Categorical-Numerical Consistency

Participants using consistent visual encoding (same representation for categorical and numerical variables) will show significantly lower cognitive load, defined as mental effort required to complete

visualization tasks, than those using mixed encoding strategies, tested through task completion time, error rates, and self-reported mental demand (Hart and Staveland 1988).

Theoretical Framework:

- Consistent visual grammar reduces extraneous cognitive load by eliminating encoding strategy switches
- Single mental model reduces working memory demands versus maintaining multiple schemas
- Computational optimization at the visualization level reduces processing demands at the viewer level

Operationalization:

- Independent variable: Encoding consistency (3 levels)
  - Condition A: Consistent equispaced lines (ggpcp)
  - Condition B: Mixed encoding (categorical boxes, numerical lines)
  - Condition C: Consistent boxes (Hammock)
- Dependent variables: Response time (ms), error rate (proportion), NASA-TLX cognitive load score
- Task: 30 trials alternating between pathway tracing and frequency estimation

Experimental Design:

- Between-subjects design with three conditions
- Sample: N = 60 (n = 20 per condition)
- Procedure: Training (10 min), Practice (5 min), Test phase (15 min)
- Analysis: One-way ANOVA for group differences, hierarchical linear modeling for learning effects, pairwise comparisons with Bonferroni correction

3.7 Mathematical Formalization

3.7.1 Optimization Problem

For a numerical axis with values $V = \{v_1, v_2, ..., v_m\}$ ordered such that $v_1 < v_2 < ... < v_m$, let n_i denote the number of observations with value v_i.

Objective: Assign positions $\{y_{i,j}\}$ for $j = 1, ..., n_i$ such that:

1. **Non-intersection constraint**: $y_{i,j} \neq y_{i,k}$ for $j \neq k$ (maintain visual distinctiveness)
2. **Monotonicity constraint**: Within value v_i, hierarchical ordering is preserved (support common fate Gestalt)
3. **Continuity constraint**: Lines maintain smooth transitions between axes (leverage good continuation)
4. **Space efficiency**: Utilize available space proportionally to frequency (balanced visual density)

Space allocation:

For each value v_i, allocate proportional space:

$S_i = \frac{n_i}{\sum_{k=1}^{m} n_k} \cdot (1 - space \cdot (m - 1))$

where $space \in [0, 0.5]$ is the spacing parameter controlling separation between value regions.

Position assignment: (Check this please!!!), the buffer should not be in the equal anymore

For observation j within value v_i (after hierarchical sorting):

$y_{i,j} = start_i + \frac{S_i - 2 \cdot buffer}{n_i - 1} \cdot (j-1) + buffer$

where: - $start_i = \sum_{k=1}^{i-1}(S_k + space)$ is the starting position for value v_i - $buffer$ provides margin space at boundaries (typically S_i) - The formula ensures even distribution across available space

3.7.2 Properties

Theorem 1 (Non-intersection): For any two observations $j \neq k$ with the same value v_i:

$$y_{i,j} \neq y_{i,k}$$

Proof: By construction, $y_{i,j} = start_i + d_i \cdot (j-1) + buffer$ where $d_i = \frac{S_i - 2 \cdot buffer}{n_i - 1}$. For $j \neq k$, we have $(j-1) \neq (k-1)$. When $n_i > 1$, $d_i > 0$, thus $y_{i,j} \neq y_{i,k}$. When $n_i = 1$, the property holds trivially.□

Theorem 2 (Perceptual Separability): Any two observations with different values occupy non-overlapping vertical regions:

$$\max_j(y_{i,j}) + \frac{space}{2} < \min_j(y_{i+1,j}) - \frac{space}{2}$$

Proof: The maximum position for value v_i is $start_i + S_i$. The minimum position for value v_{i+1} is $start_{i+1} = start_i + S_i + space$. The spacing parameter $space$ creates explicit separation. □

This property ensures preattentive separability: observations with different values can be distinguished without serial search.

Theorem 3 (Space Utilization): The algorithm achieves optimal space utilization:

$$\sum_{i=1}^{m} S_i = 1 - space \cdot (m-1)$$

Proof: By construction, $S_i = \frac{n_i}{\sum_k n_k} \cdot (1 - space \cdot (m-1))$. Summing over all values:

$$\sum_{i=1}^{m} S_i = \sum_{i=1}^{m} \frac{n_i}{\sum_k n_k} \cdot (1 - space \cdot (m-1)) = \frac{\sum_i n_i}{\sum_k n_k} \cdot (1 - space \cdot (m-1)) = 1 - space \cdot (m-1)$$

□

Corollary (Visual Density Proportionality): Visual density within each region is proportional to observation count, supporting density-based frequency inference.

3.7.3 Hierarchical Sorting for Line Crossing Minimization

To minimize line crossings and leverage the Gestalt principle of common fate, observations within each tie group are sorted hierarchically based on values on adjacent axes.

Algorithm (Simplified):

```
For each tied group on axis i:
  Sort observations by (value on axis i+1, value on axis i-1)
  Assign positions y_{i,j} in sorted order
```

This ensures that observations with similar trajectories are positioned near each other, creating visual bands that support the perception of common fate.

3.8 Implementation and Evaluation Plan

3.8.1 Phase 1: Algorithm Development (Weeks 1-4)

Deliverable: Functional R package extension with comprehensive documentation

Tasks:

1. Extend `pcp_arrange()` to detect and handle numerical ties
2. Implement `optimize_spacing_numerical()` or `tie_spacing()` function with perceptually-motivated parameters
3. Handle edge cases while preserving perceptual properties:
   - Single unique value (centered position)
   - Extreme skew (adaptive buffer sizing)
   - Missing values (explicit separation region)
4. Integration testing with existing ggpcp workflow

Code Structure (Simplified):

3.9 Phase 2: Perceptual Validation Studies (Weeks 5–7)

Study 1: Frequency Perception

- **Duration:** Weeks 5–7
- **Participants:** n = 60 (30 per condition)
- **Design:** Between-subjects

Procedure:

1. Magnitude estimation (12 trials): "What percentage have value X?"
2. Ordinal comparison (12 trials): "Which value is more frequent?"
3. Ratio judgment (12 trials): "A is how many times B?"
4. Confidence ratings (7-point scale) for each response

Analysis:

- Absolute percentage error for magnitude estimation
- Accuracy rates for ordinal and ratio tasks
- Bias analysis: systematic over/under-estimation
- Confidence calibration: accuracy vs. subjective confidence

3.10 Phase 3: Comparative Benchmarking (Weeks 6–10)

Computational Metrics:

1. Rendering performance (time, memory, scalability n = 2 to 100)

Table 1: Research Timeline (12 Weeks)

Week	Milestone	Deliverable	Perceptual Focus
1–2	Literature synthesis	Annotated bibliography with Ware framework integration	Ware framework integration
3–4	Core algorithm development	Working R code with perceptual properties verified	Perceptual properties implemented
4	IRB submission	Approved protocol for human subjects research	Human subjects clearance
5–7	Study 1: Frequency perception	Raw data and statistical analysis	Magnitude estimation validation
8–10	Benchmarking	Performance report with computational metrics	Computational metrics
10–12	Integration	Draft manuscript and R package release	Synthesis and dissemination

2. Visual quality metrics (Dennig et al. 2021): line crossing count, ribbon overlap, visual clutter index
3. Perceptual quality estimates: modeled eye movements, predicted visual search time

Case Studies:

- Palmer Penguins: Mixed categorical-numerical data with known correlation structure
- Iris Dataset: Multiple numerical variables with natural ties
- Asthma Data (Schonlau and Yang 2024): Direct comparison with published hammock plot

3.11 Phase 4: Integration and Dissemination (Weeks 8–12)

Deliverables:

1. **R Package Release (ggpcp v2.0):**
   - CRAN submission with full documentation
   - Vignette: "Handling Numerical Ties in Parallel Coordinate Plots"
   - Unit tests achieving >95% coverage
2. **Academic Paper:**
   - Target: IEEE Transactions on Visualization and Computer Graphics
   - Submission deadline: March 2026 (IEEE VIS)
3. **Supplementary Materials:**
   - Open Science Framework repository
   - All experimental stimuli and data
   - Reproducibility package

4 Timeline and Milestones

5 Expected Contributions

5.1 Theoretical Contributions

1. **Unified Tie-Breaking Theory:** Formalization of equispaced optimization for both categorical and numerical variables, with mathematical proofs of key perceptual properties (separability, continuity, space efficiency)
2. **Empirical Characterization:** Rigorous experimental evidence quantifying perceptual trade-offs between individual-level and aggregate-level encoding strategies across multiple task types

5.2 Practical Contributions

1. **Production Software:** Open-source R package extension immediately usable by practicing data scientists and researchers
2. **Evidence-Based Guidelines:** Design recommendations grounded in empirical evidence and perceptual theory:
   - When to use equispaced lines vs. constant-width boxes
   - How to set spacing parameters for optimal perceptual clarity
   - Task-specific visualization selection criteria
3. **Benchmark Suite:** Reusable experimental framework and stimuli for evaluating future parallel coordinate plot innovations

5.3 Methodological Contributions

1. **Theory-Driven Evaluation:** Demonstrates how established perceptual theory (Ware, Gestalt principles, Cleveland & McGill) can generate specific, testable hypotheses about visualization effectiveness
2. **Multi-Method Approach:** Integrates behavioral experiments with computational benchmarking to triangulate findings
3. **Reproducible Research:** Complete open science package enabling replication and extension by other researchers

6 Broader Impact

This research addresses visualization challenges across numerous domains where mixed categorical-numerical data is common:

By providing both theoretical understanding and practical tools, this work enables more effective exploratory data analysis across these domains, ultimately supporting better data-driven decision making.

Table 3: Application Domains for Research Impact

Domain	Application	Data Characteristics
Healthcare	Patient trajectory visualization	Categorical diagnoses + numerical measurements
Manufacturing	Quality control monitoring	Defect categories + continuous sensor readings
Social Science	Survey analysis	Demographic categories + Likert scales
Finance	Portfolio analysis	Categorical sectors + numerical performance metrics
Education	Learning analytics	Course completion + assessment scores

7 Limitations and Future Directions

7.1 Study Limitations

- **Sample Characteristics:** University student population may not generalize to domain experts
- **Experimental Constraints:** Laboratory tasks may lack ecological validity
- **Scope:** Limited to static visualizations; interactive features not evaluated

7.2 Future Extensions

- Longitudinal study with domain experts in real analytical workflows
- Investigation of interaction effects between individual differences and visualization method
- Hybrid approaches: smooth interpolation between lines and boxes based on dataset properties
- Integration with animation and interactive highlighting techniques

8 Conclusion

This comprehensive examination proposal addresses a fundamental question in information visualization by grounding design decisions in Colin Ware's three-stage model of visual information processing. The proposed equispaced line approach extends ggpcp's proven categorical tie-breaking algorithm to numerical variables, creating a unified framework that maintains visual continuity, leverages preattentive processing, and supports Gestalt principles of good continuation and common fate.

Compared to hammock plots' constant-width boxes, equispaced lines offer different perceptual trade-offs: enhanced individual observation traceability and smoother visual flow at the potential cost of explicit frequency encoding. This research contributes to theory by demonstrating how established perceptual principles generate testable hypotheses about visualization effectiveness, to practice by delivering open-source software and evidence-based design guidelines, and to methodology by exemplifying theory-driven, multi-method evaluation.

As Ware (2012) emphasizes, effective visualization must be grounded in understanding of human perception. This research takes that principle seriously, asking not just "which visualization looks better?" but rather "which visualization better aligns with the architecture of human visual pro-

cessing for specific analytical tasks?" The answer, as with most visualization questions, is nuanced and context-dependent—but now it will be grounded in empirical evidence and perceptual theory.

References

Cleveland, William S, and Robert McGill. 1984. "Graphical Methods for Data Presentation: Full Scale Breaks, Dot Charts, and Multibased Logging." *The American Statistician* 38 (4): 270–80. https://doi.org/10.1080/00031305.1984.10483223.

Dennig, Frederik L., Maximilian T. Fischer, Michael Blumenschein, Johannes Fuchs, Daniel A. Keim, and Evanthia Dimara. 2021. "ParSetgnostics: Quality Metrics for Parallel Sets." *Computer Graphics Forum* 40 (3): 375–86. https://doi.org/10.1111/cgf.14314.

Hart, Sandra G., and Lowell E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Human Mental Workload*, edited by Peter A. Hancock and Najmedin Meshkati, 139–83. Amsterdam: North-Holland.

Hofmann, Heike, and Marie Vendettuoli. 2013. "Common Angle Plots as Perception-True Visualizations of Categorical Associations." *IEEE Transactions on Visualization and Computer Graphics* 19 (12): 2297–2305. https://doi.org/10.1109/TVCG.2013.140.

Inselberg, Alfred. 1985. "The Plane with Parallel Coordinates." *The Visual Computer* 1 (2): 69–91. https://doi.org/10.1007/BF01898350.

———. 2009. "Parallel Coordinates: Visual Multidimensional Geometry and Its Applications." *Springer Science & Business Media*.

Padilla, Lace M., Sarah H. Creem-Regehr, Mary Hegarty, and Jeanine K. Stefanucci. 2018. "Towards an Optimistic View of Visual Decision-Making." *Cognitive Research: Principles and Implications* 3 (1): 1–25. https://doi.org/10.1186/s41235-018-0120-9.

Peng, Wei, Matthew O Ward, and Elke A Rundensteiner. 2004. "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering." In *IEEE Symposium on Information Visualization*, 89–96. IEEE.

Scaife, M., and Y. Rogers. 1996. "External Cognition: How Do Graphical Representations Work?" *International Journal of Human-Computer Studies* 45 (2): 185–213.

Schonlau, Matthias. 2003. "The Hammock Plot: Visualizing Mixed Categorical and Numerical Data." In *Proceedings of the American Statistical Association*.

Schonlau, Matthias, and Rosie Yuyan Yang. 2024. "Hammock Plots: Visualizing Categorical Data Beyond Parallel Coordinates." *Journal of Computational and Graphical Statistics*.

VanderPlas, Susan, Yawei Ge, Antony Unwin, and Heike Hofmann. 2023. "Penguins Go Parallel: A Grammar of Graphics Framework for Generalized Parallel Coordinate Plots." *Journal of Computational and Graphical Statistics* 32 (4): 1405–20. https://doi.org/10.1080/10618600.2023.2181762.

Ware, Colin. 2012. *Information Visualization: Perception for Design*. 3rd ed. Waltham, MA: Morgan Kaufmann.

Wegman, Edward J. 1990. "Hyperdimensional Data Analysis Using Parallel Coordinates." *Journal of the American Statistical Association* 85: 664–75.