IMEN891K / IDSC501 / AXGS-501 · DATA ENGINEERING · FALL 2026

3.

Data Exploration

Sangwoo Mo · POSTECH

Industrial & Management Engineering

Where We Are

Part I (Lec 2–5): collection, exploration, and preprocessing



You cannot clean what you have not looked at

Outline: Data Exploration

01

Descriptive Analysis

Reduce the data to a few numbers

02

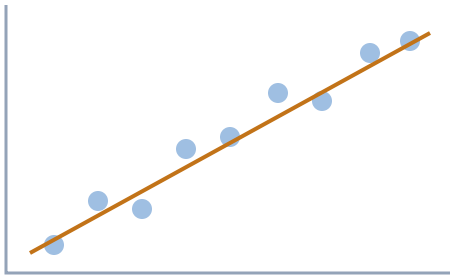
Visual Exploration



Find patterns by plotting the data

03

Statistical Exploration



Read relationships between variables

Three Ways to Read One Table

Numbers, pictures, relations: what each of today's three parts produces

$\bar{x}$ s Σ **Numbers**	Summarize location, spread and distribution shape	PART 1 **Descriptive Analysis**
Pictures	Choose plots that reveal structure and time patterns	PART 2 **Visual Exploration**
Relations	What ties two columns, and whether it is a cause	PART 3 **Statistical Exploration**

Descriptive Analysis

01

Descriptive Analysis

Reduce the data to a few numbers

02

Visual Exploration

Find patterns by plotting the data

03

Statistical Exploration

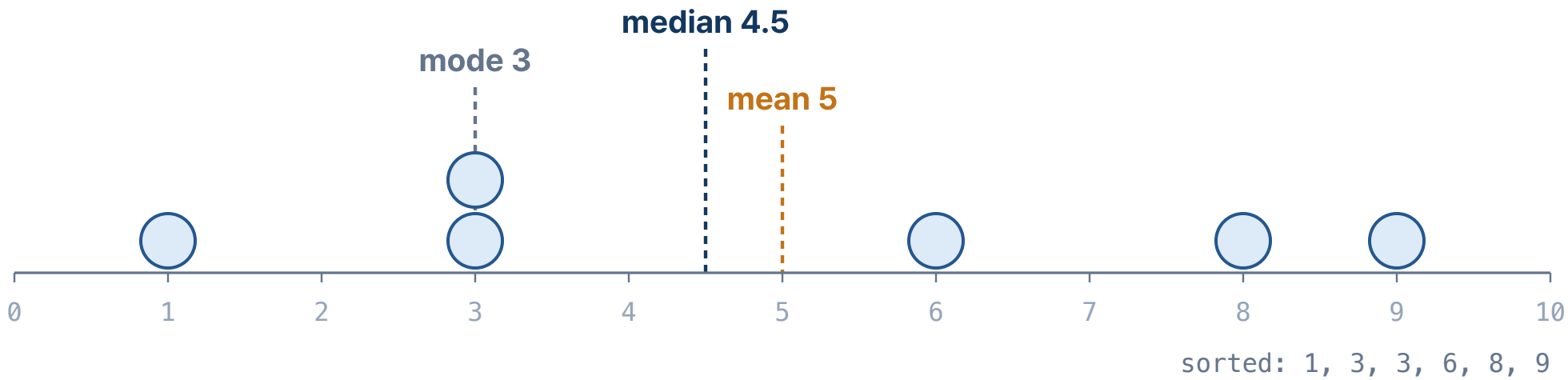
Read relationships between variables

Central Tendency



Recap: Mean, Median, Mode

Three answers to “the typical value”

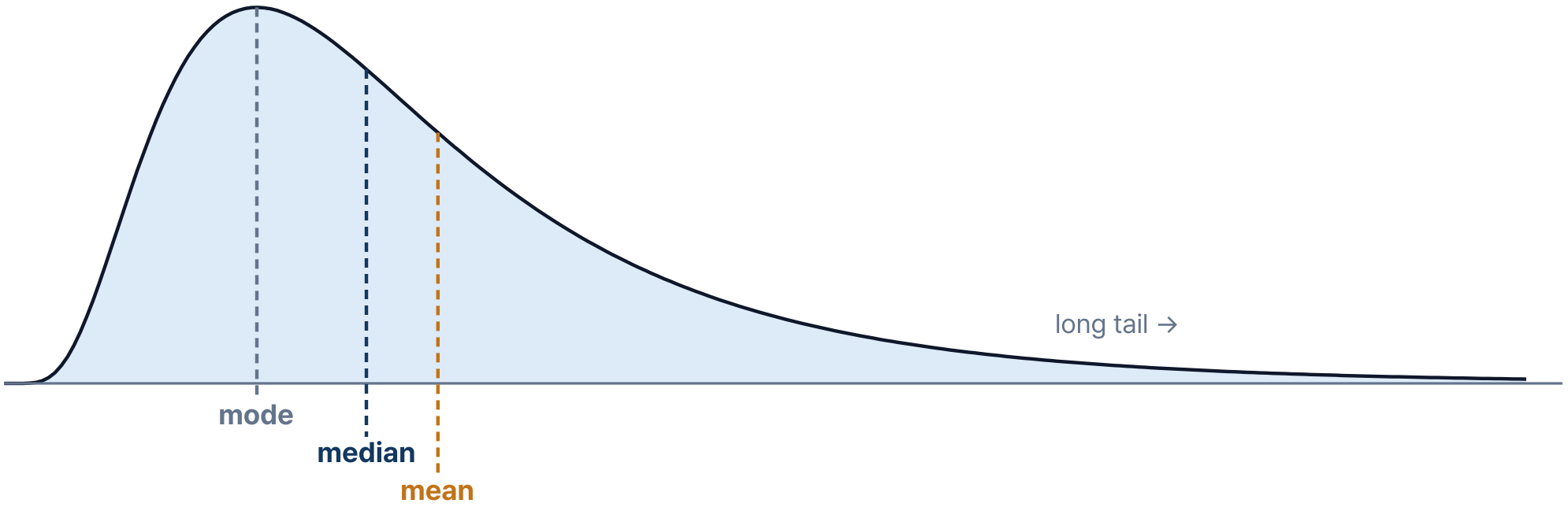


MEAN $\bar{x} = (1/n) \sum_i x_i$ | MEDIAN $(x_{(n/2)} + x_{(n/2+1)}) / 2$ | MODE $\text{argmax}_x \text{count}(x)$

From 9, 3, 1, 8, 3, 6: mean **5**, median **4.5**, mode **3**

Skew and the Three Centers

In this right-skewed example, the mean lies farther into the tail



`nonparametric skew = (x̄ − median) / s` a mean–median comparison, not a complete description of shape

Here, mode < median < mean; check the distribution before generalizing

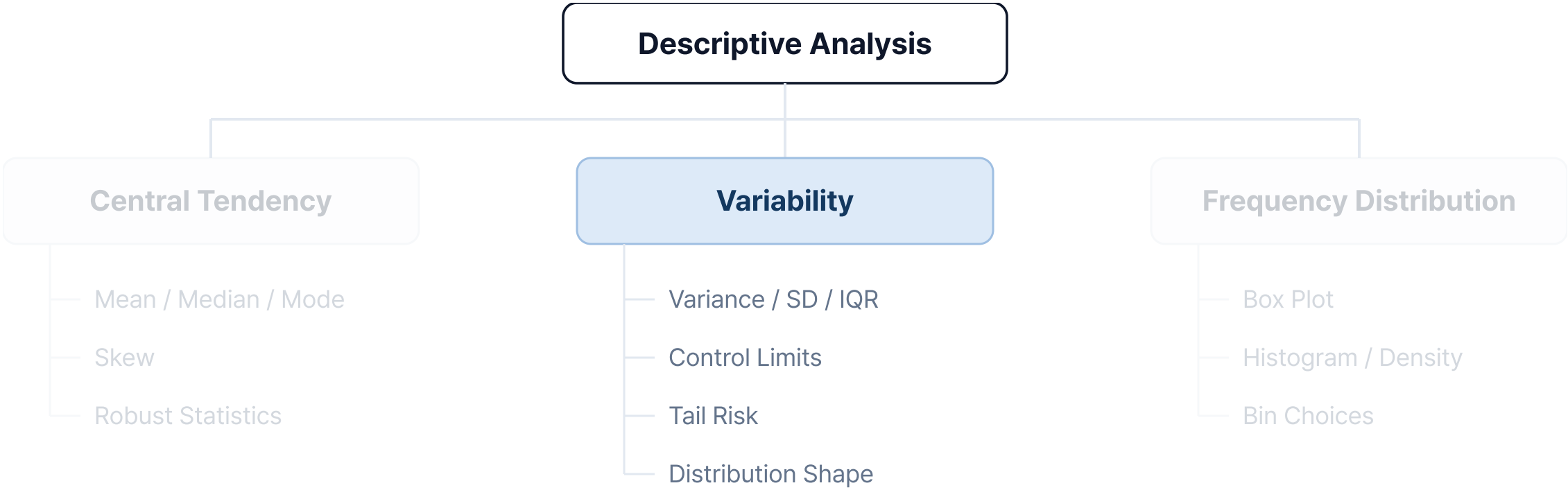
Robust Statistics

Every summary comes in a version built to ignore the far end

WHAT YOU SUMMARIZE	READS EVERY VALUE	READS ONLY THE MIDDLE
Center	**Mean** one far row moves it	**Median** trimmed mean sits between them
Spread	**Standard deviation** squares every distance first	**IQR, MAD** interquartile range, median absolute deviation

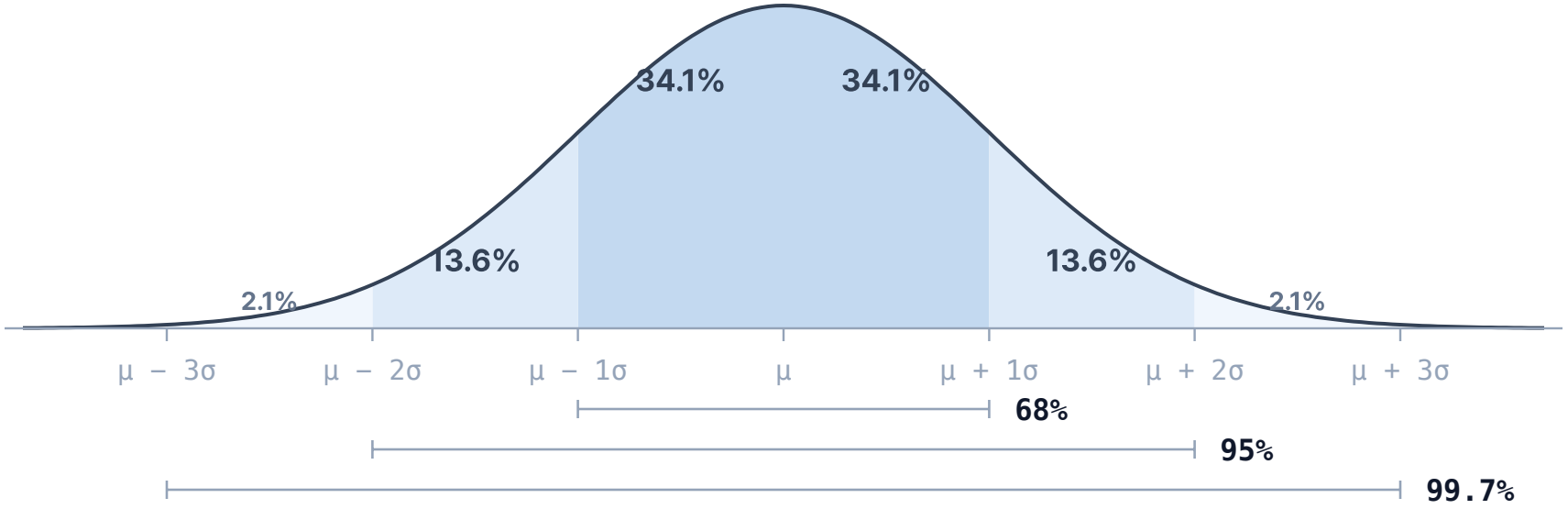
The split is not about the mean: **every summary number has both versions**

Variability



Recap: Standard Deviation

The 68–95–99.7 rule

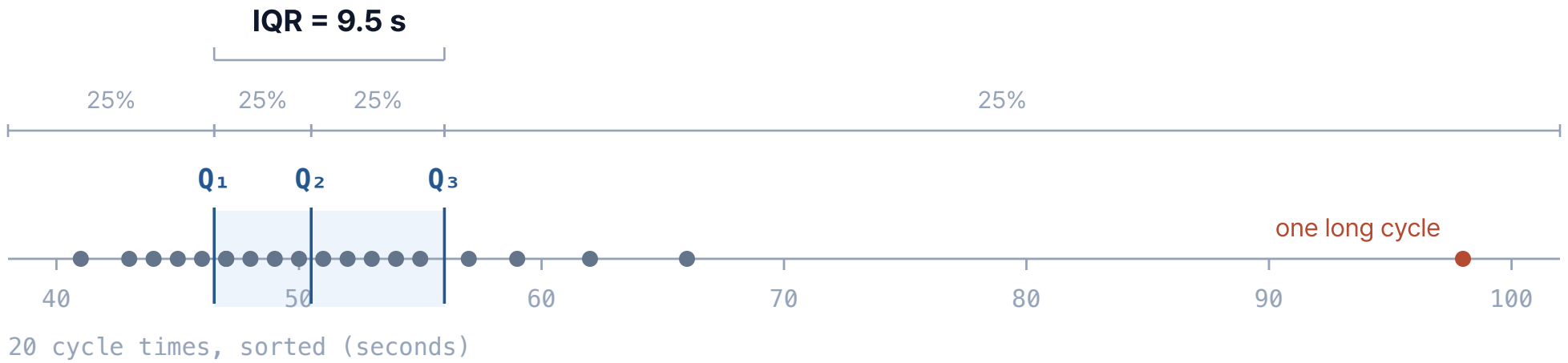


$$s = \sqrt{\ (1/(n-1))\ \sum_i\ (x_i - \bar{x})^2\ }$$ the square is what lets one far-out row dominate

These percentages hold for a normal distribution **and only then**

Interquartile Range (IQR)

Q3 − Q1, the width of the middle half: it does not care how far the outlier sits

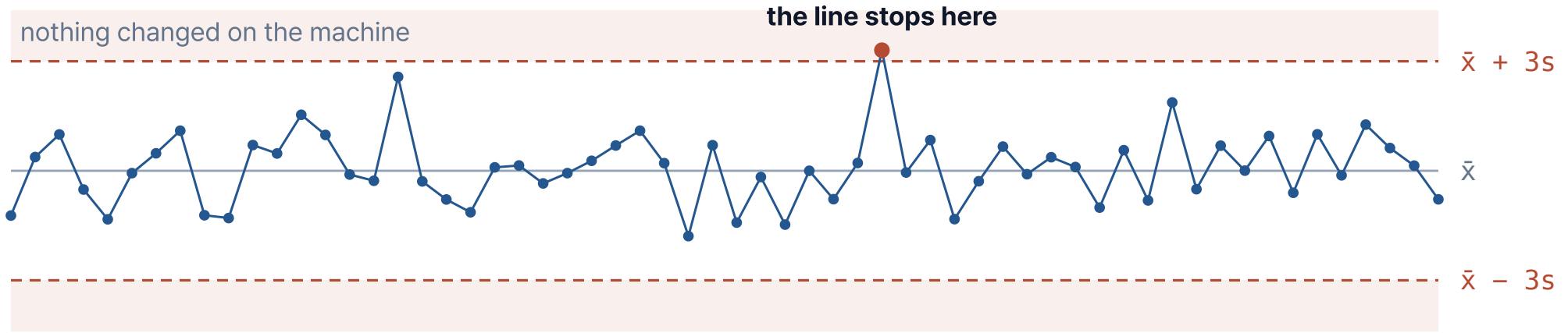


$$IQR = Q_3 - Q_1$$ Q_1, Q_3 are the 25th and 75th percentiles

In this example, moving the largest value farther out leaves the quartiles unchanged

Control Limits and the Normal Curve

A normal curve puts about 0.27% of its mass outside μ ± 3σ

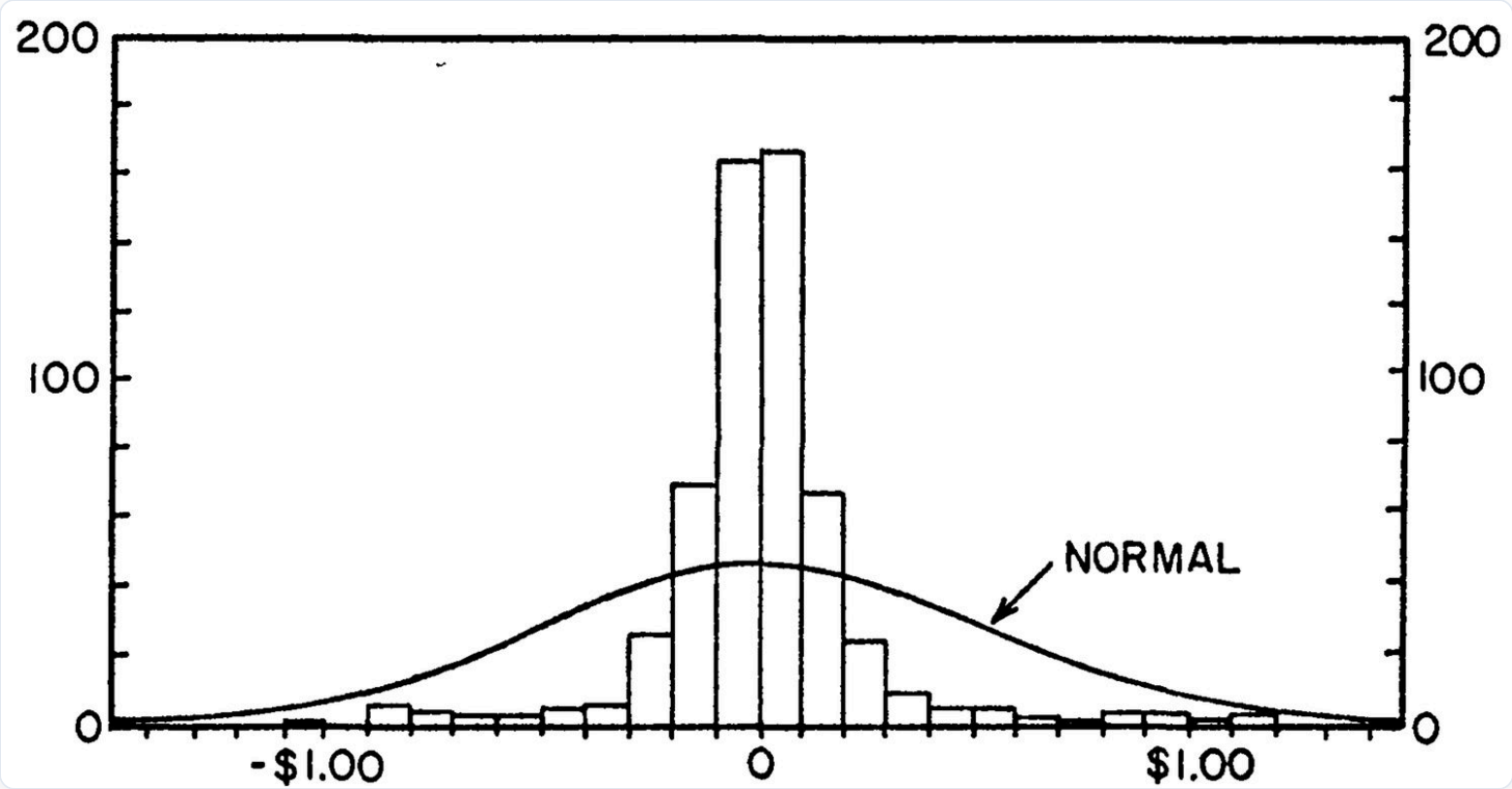


bearing temperature, one reading per hour

0.27% assumes a normal distribution with known parameters; estimated limits can behave differently

Case study: Heavy Tails in Price Changes

Mandelbrot, 1963: the risk models of the day assumed Gaussian moves



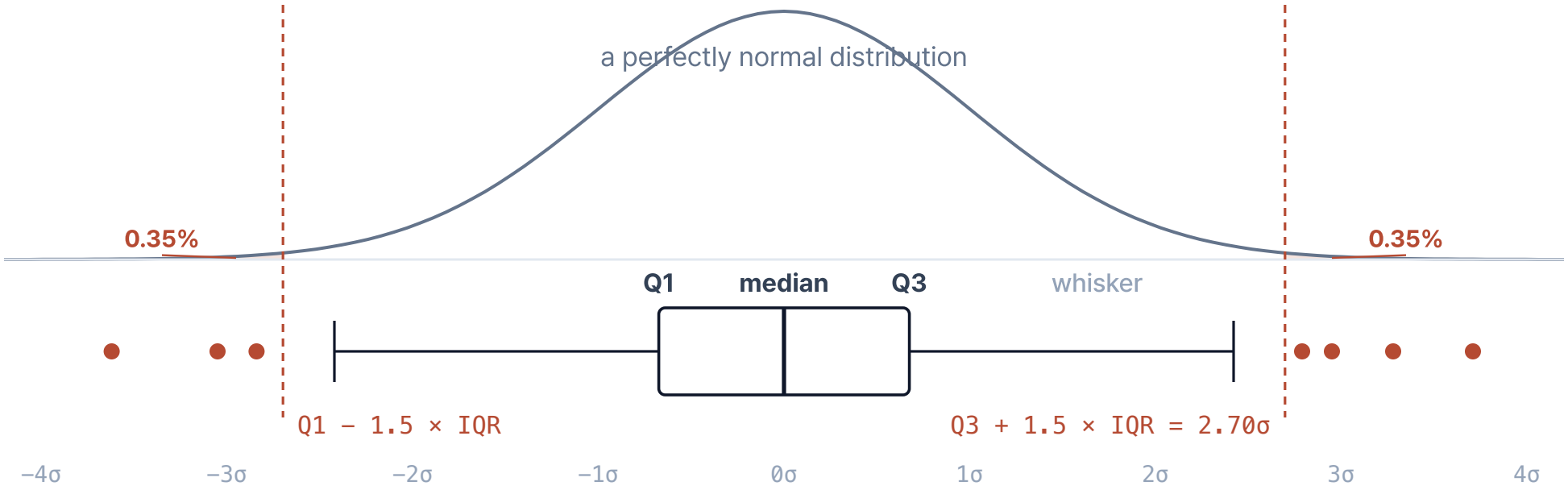
Quieter than Gaussian on most days, and **far wilder on the rare ones**

Frequency Distribution



Box Plot

Fences flag extreme values; whiskers end at observed values inside them

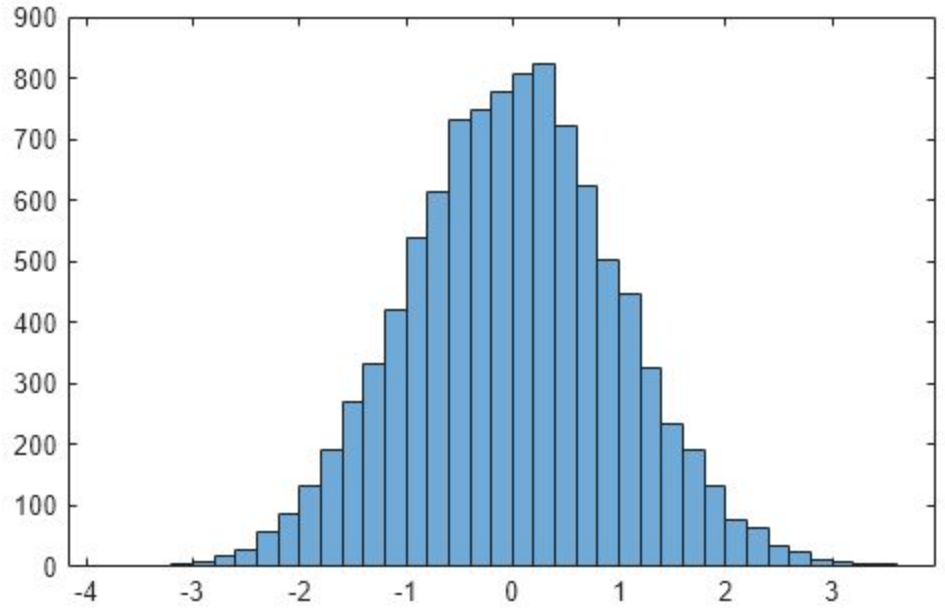


$$\text{IQR} = Q_3 - Q_1 \quad | \quad \text{fences at } Q_1 - 1.5 \cdot \text{IQR} \text{ and } Q_3 + 1.5 \cdot \text{IQR}$$

Fences are thresholds; whiskers end at data points, and flagged values still need investigation

Histogram as Empirical Distribution

Bins summarize how many observations fall in each interval



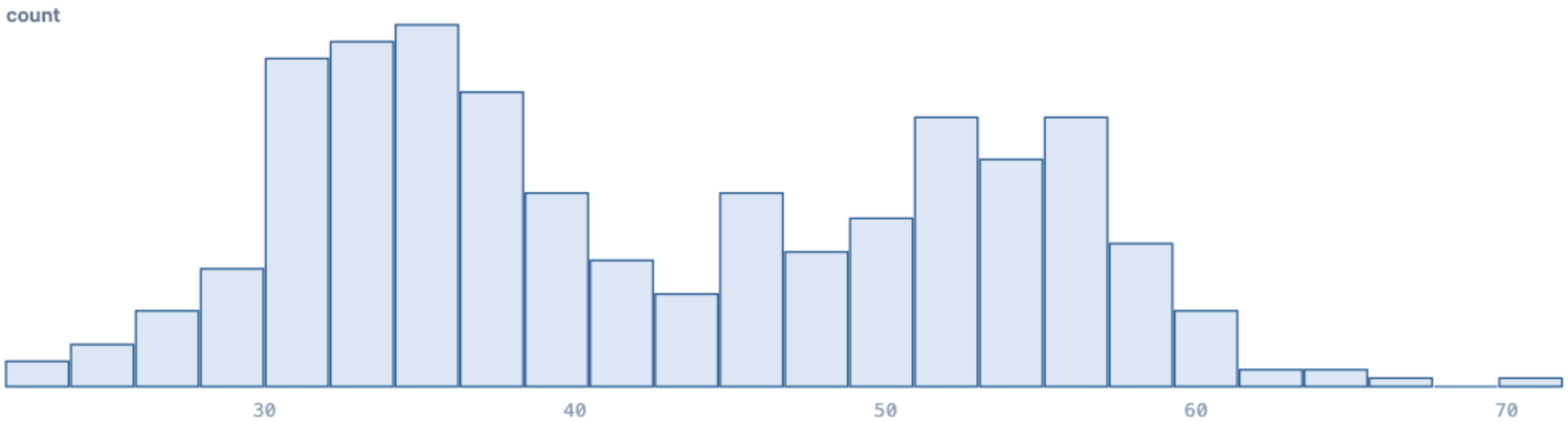
Bin counts show distribution shape while hiding within-bin detail

Demo: Histogram Bin Count

Same 420 numbers, bin count is the only thing changing

1. Compare 4, 24 and 70 bins. 2. Which pattern survives your choices?

Bins 24 Reset



bin width **2.09** · tallest bin sits at **35.3**

Answer: Histogram Choices

Read what changed before choosing a setting

Setting	Interpretation
Few bins	Distinct peaks can merge
Many bins	Sparse bars can mimic structure
Several settings	Report stable distribution features

The setting changes the picture; the underlying data stay fixed

Visual Exploration

01

Descriptive Analysis

Reduce the data to a few numbers

02

Visual Exploration



Find patterns by plotting the data

03

Statistical Exploration

Read relationships between variables

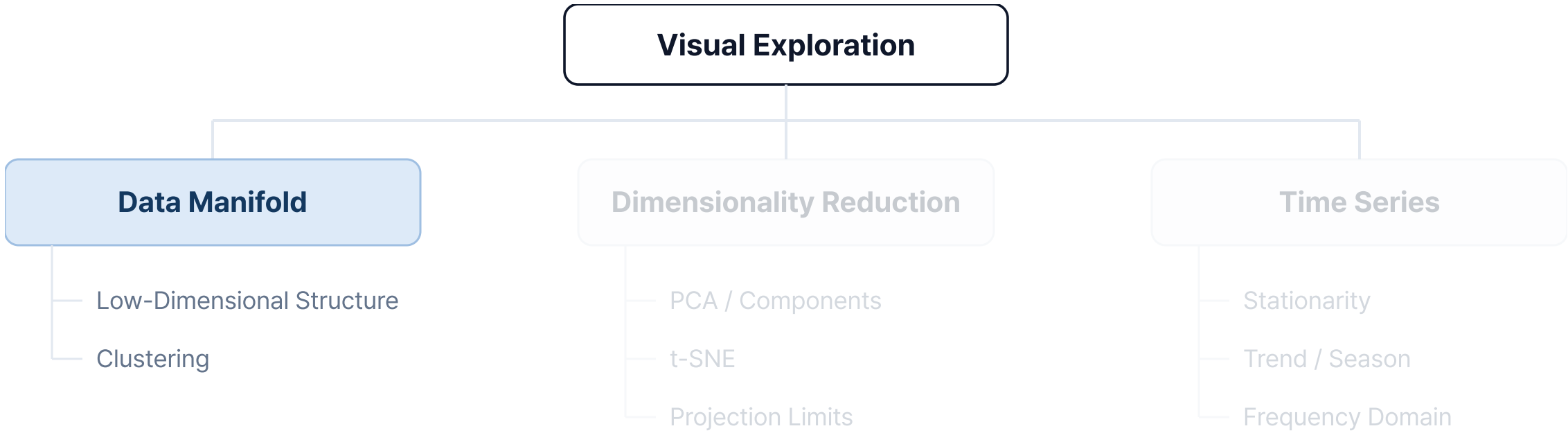
Recap: Five Standard Charts

The five you already know

	Which category is bigger?	BAR CHART
	How does the whole split up?	PIE OR STACKED BAR
	Which values are common?	HISTOGRAM, BOX PLOT
	Do these two move together?	SCATTER PLOT
	What changed over time?	LINE CHART

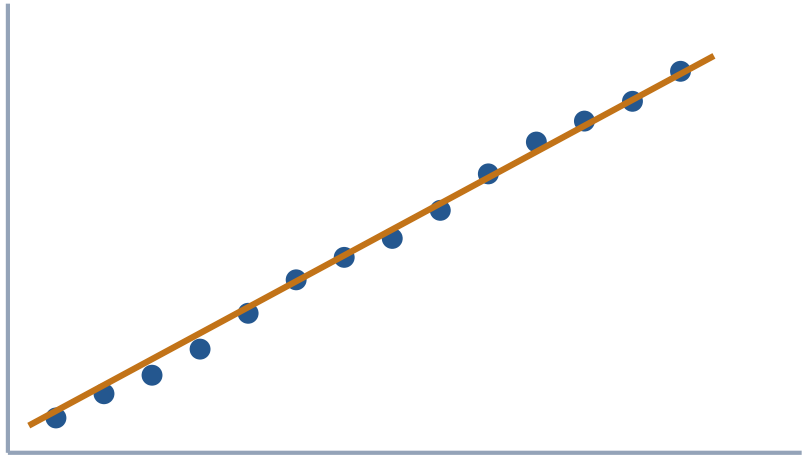
Every one of these needs you to **already know which columns matter**

Data Manifold



Data Manifold

A lower-dimensional structure can explain variation in many measured columns



two measured columns

One direction explains most variation

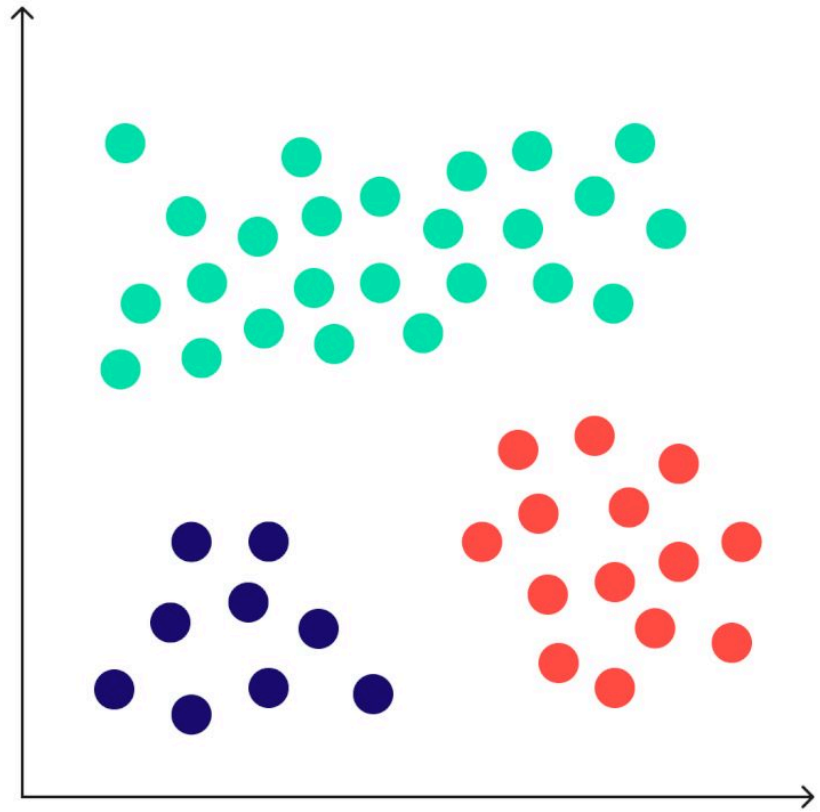
The points occupy a narrow band

A 1D representation can approximate it

These two columns vary mostly along one direction

Clustering with k-means

It always returns groups: the same points after asking k-means for three



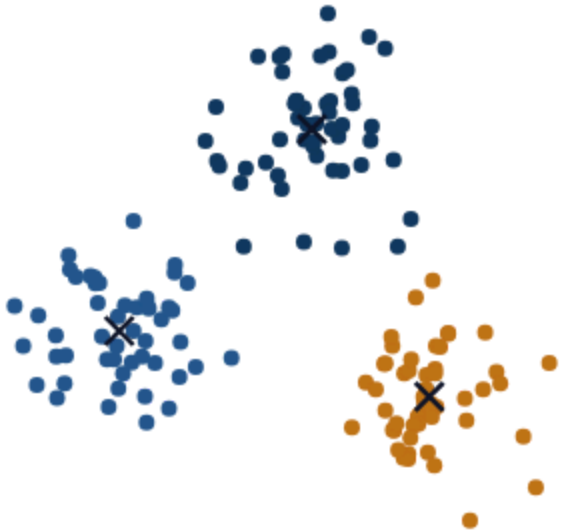
Choosing k asks for that many clusters; it does not establish the true group count

Demo: k-means Clustering

Same 150 points every time, k is the only input k-means gets

1. Compare k = 2, 3 and 6. 2. Can the smallest WCSS select k?

k 3 Reset



k = **3** · this matches the **three groups that are actually there** · within-cluster sum of squares **120**, which only ever falls as k grows, so it cannot pick k for you.

Answer: Cluster Count

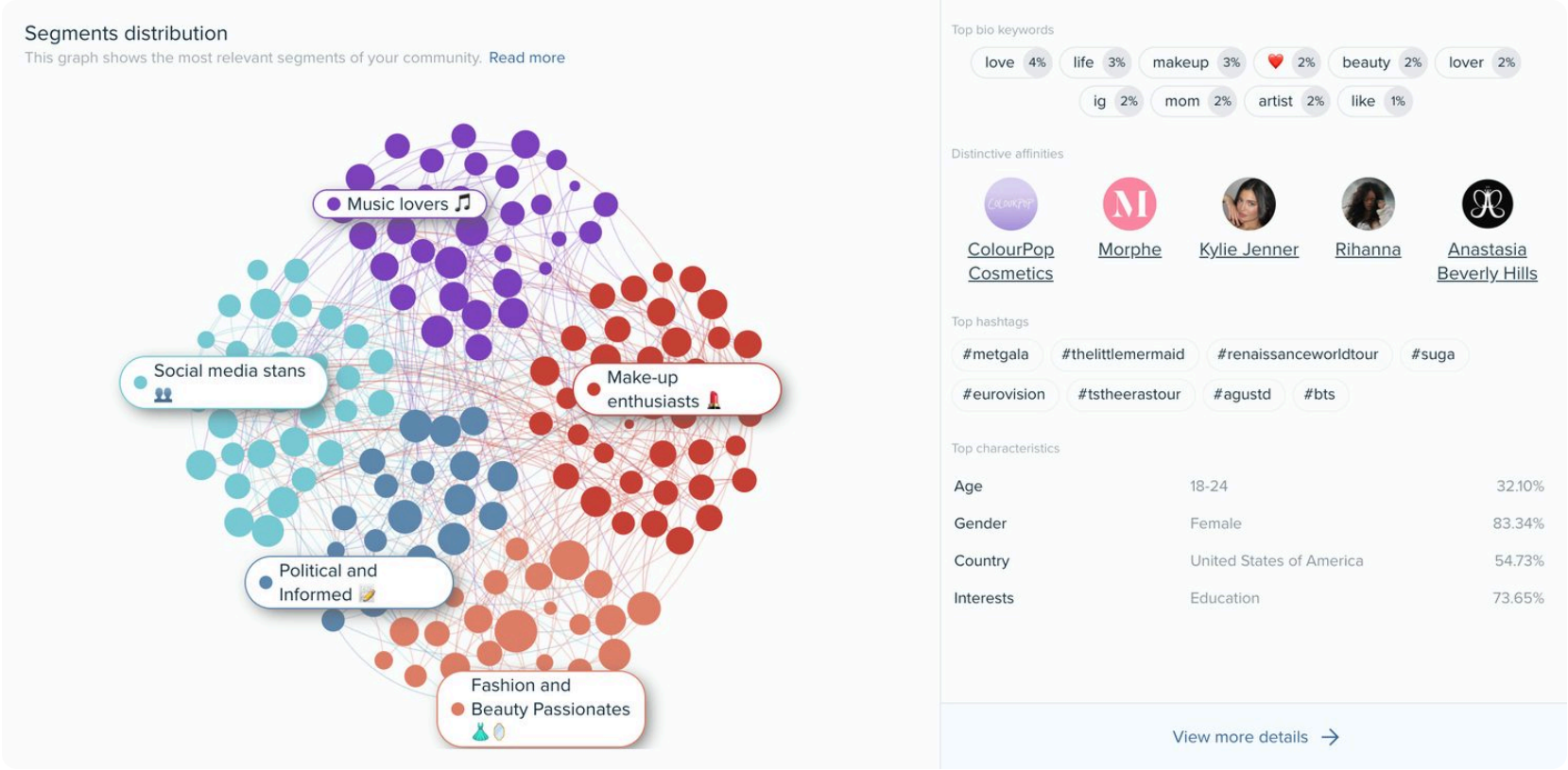
Read what changed before choosing a setting

Setting	Interpretation
`k = 2`	`Groups merge`
`k = 3`	`Three groups in this synthetic example`
`k = 6`	`Groups split; WCSS can still fall`

The setting changes the picture; the underlying data stay fixed

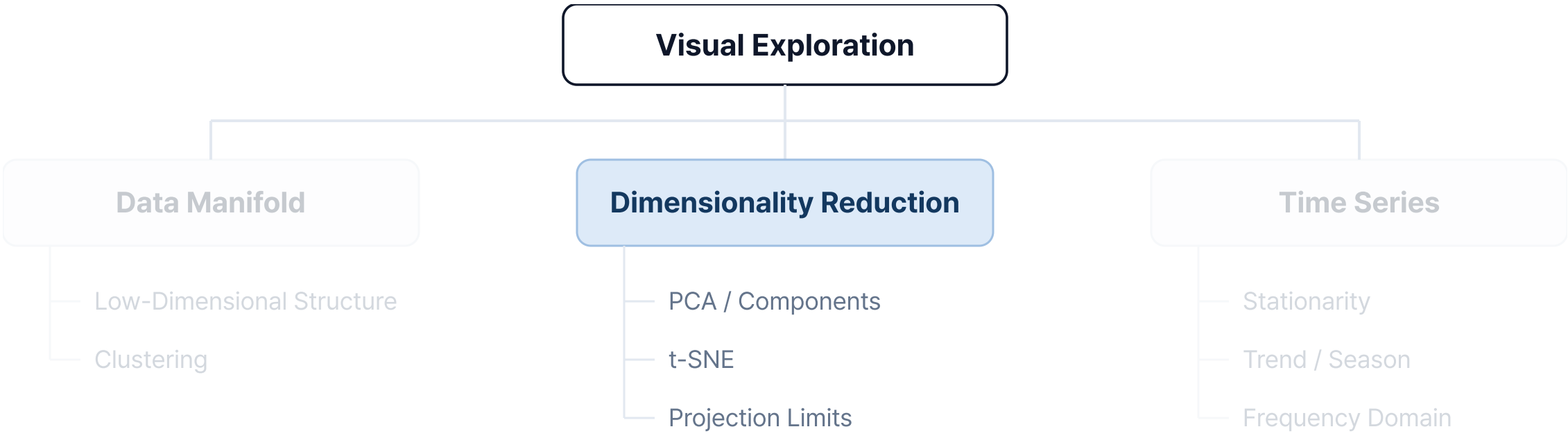
Case study: Marketing Segmentation

A cluster needs a name to be useful: from a cluster number to a readable persona



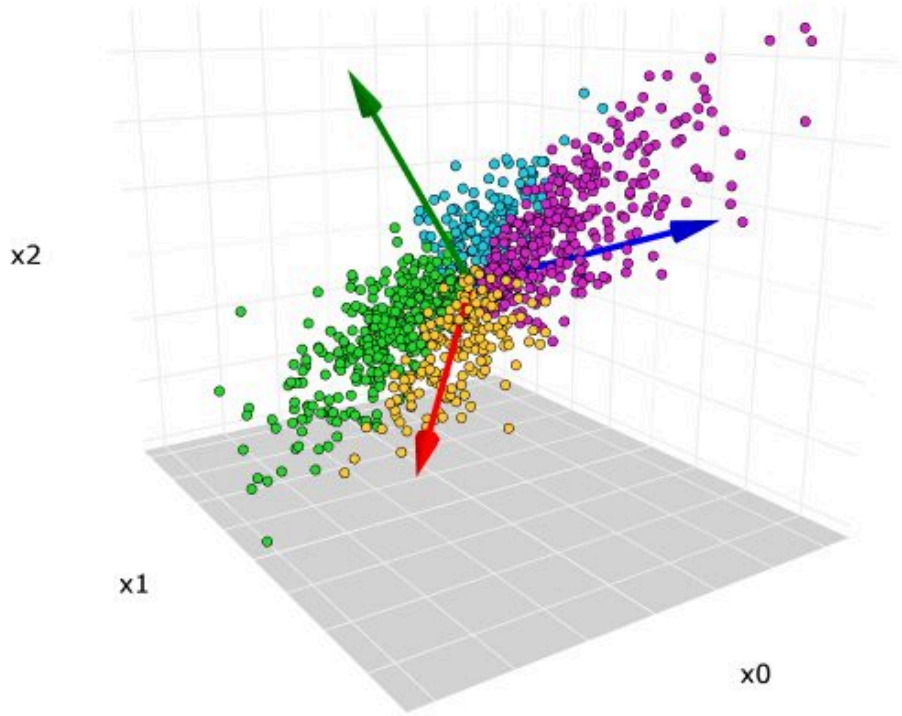
A cluster number is useless until someone can say **what makes that group different**

Dimensionality Reduction



Principal Component Analysis (PCA)

Rotate the axes so the spread lands on the first few



The three arrows are the new axes. **The longest one points along the widest spread**

Projection onto a Component

PCA turns each original row into coordinates along the fitted directions

Two measured columns

```
temperature ·
vibration

2 numbers per
observation
```

→

Fitted PCA direction

```
center on training
mean
project onto the
learned axis
```

→

One component

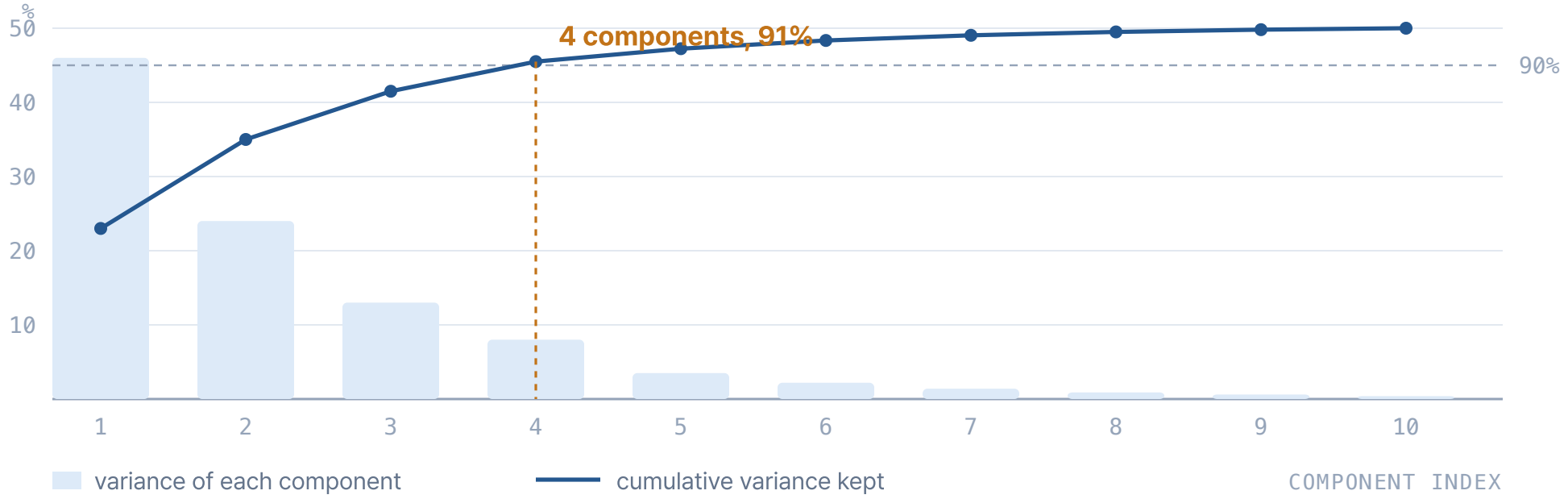
```
z = (x − μ) · v

1 number per
observation
```

Keeping fewer coordinates discards variation along the omitted directions

Choosing the Number of Components

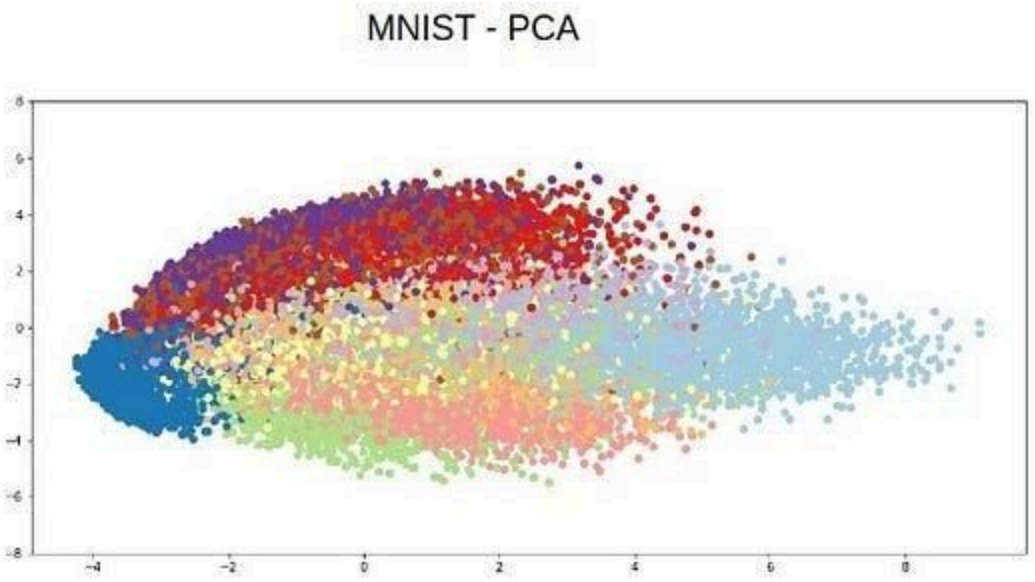
Retained variance measures preserved spread; useful information depends on the task

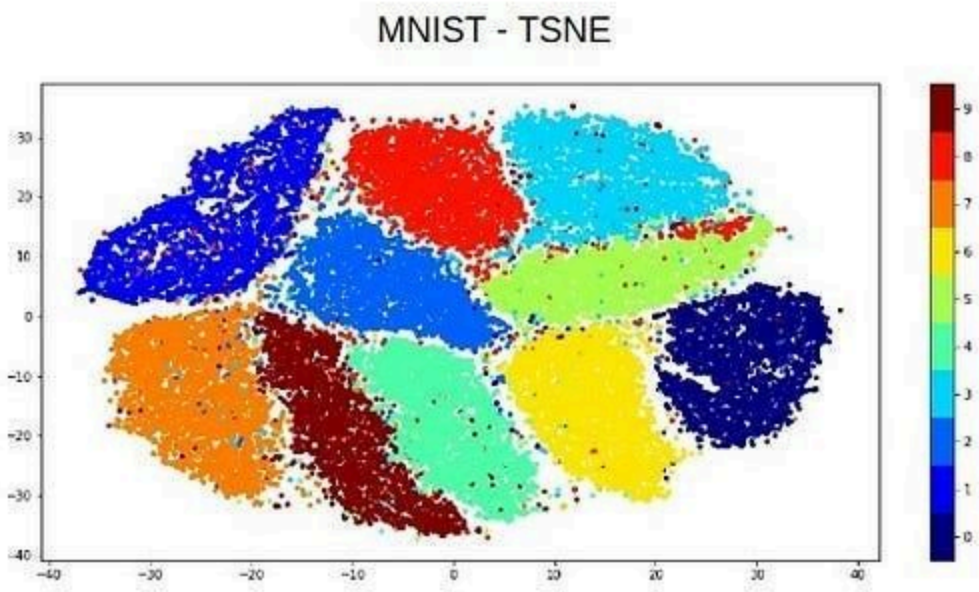


Four components keep 91% of variance; task-relevant information still needs checking

t-SNE vs PCA

t-distributed stochastic neighbor embedding separates what PCA leaves overlapping





The same digits, twice: **the classes overlap on the left and separate on the right**

Demo: t-SNE Perplexity

Same two datasets every time, only the perplexity moves

1. Compare perplexity 2 and 30. 2. Can a visible island prove a real cluster?

Perplexity `30 default` **Reset**

three real groups



no groups at all, pure noise



perplexity **30** · the **default in scikit-learn and openTSNE**. The effective neighborhood size is about 30: the left settles into **three**, the right into **one cloud**.

Perplexity is roughly how many neighbors each point is asked to preserve. Try several values and seeds; stable patterns still need checks in the original data.

Answer: Reading t-SNE

Read what changed before choosing a setting

Setting	Interpretation
Small perplexity	Local islands appear, even in random data
Larger perplexity	Some islands disappear
Original data	Check neighbors, stability and domain meaning

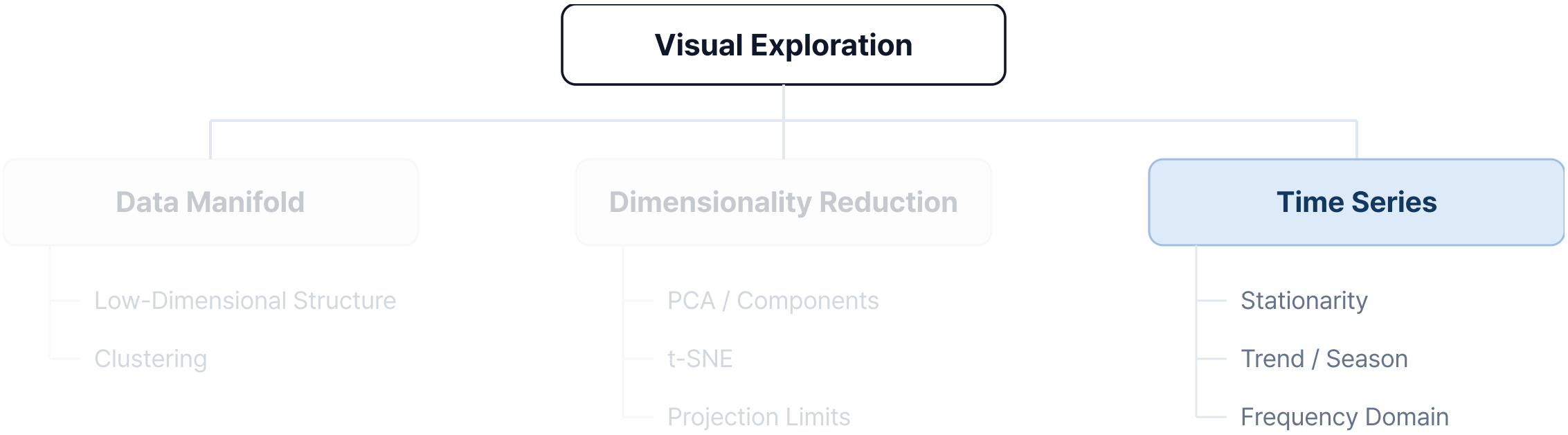
The setting changes the picture; the underlying data stay fixed

Reading a t-SNE Plot

Check stability; axes, blob areas, and gaps do not directly measure the original geometry

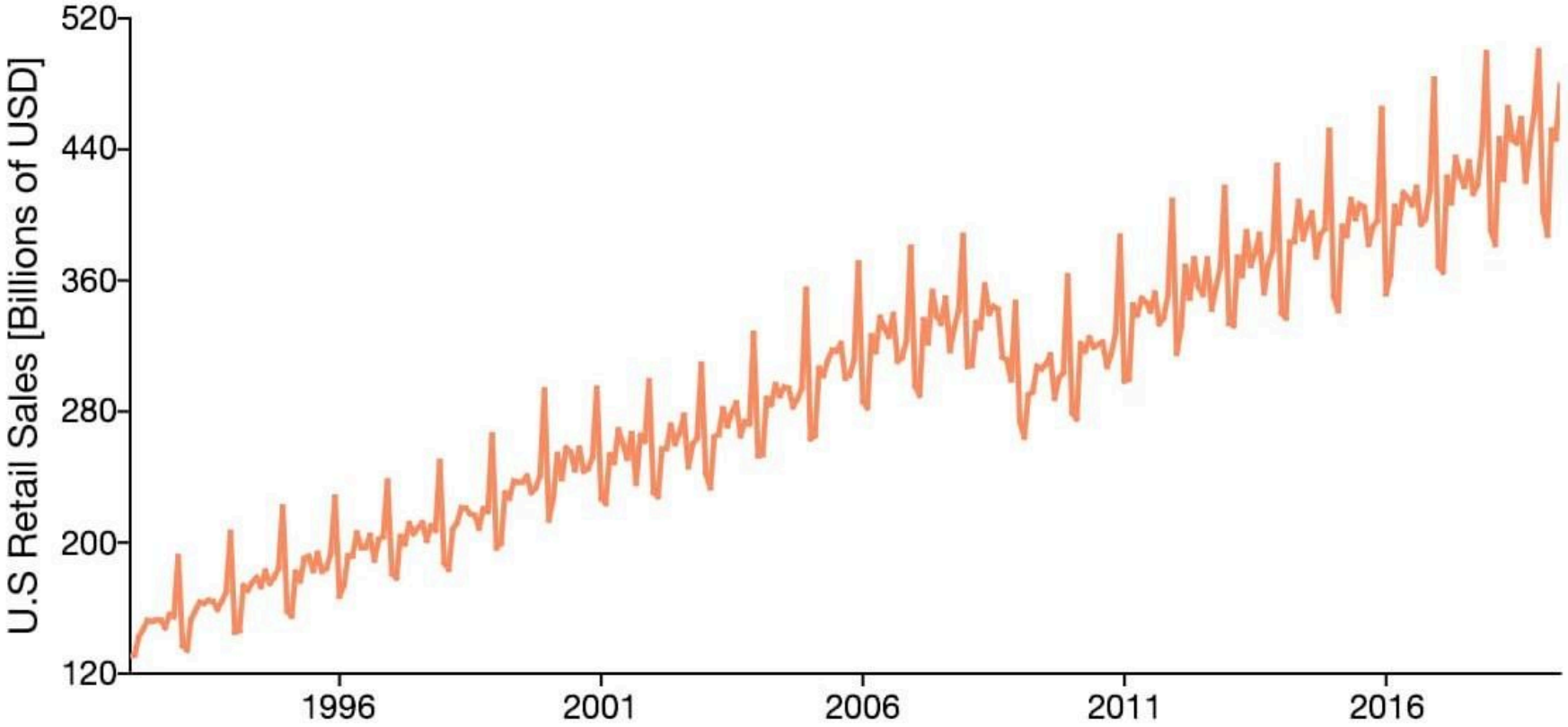
- **The axes carry no meaning.** PCA rotates, so its axes can be read; these two cannot
- **Blob area does not directly measure density.** Dense regions are inflated and sparse ones are shrunk
- **The distance between two clusters is not their dissimilarity.** Local neighborhoods are an approximate optimization target

Time Series


Visual Exploration
Data Manifold
Low-Dimensional Structure
Clustering
Dimensionality Reduction
PCA / Components
t-SNE
Projection Limits
Time Series
Stationarity
Trend / Season
Frequency Domain

Time Series and Stationarity

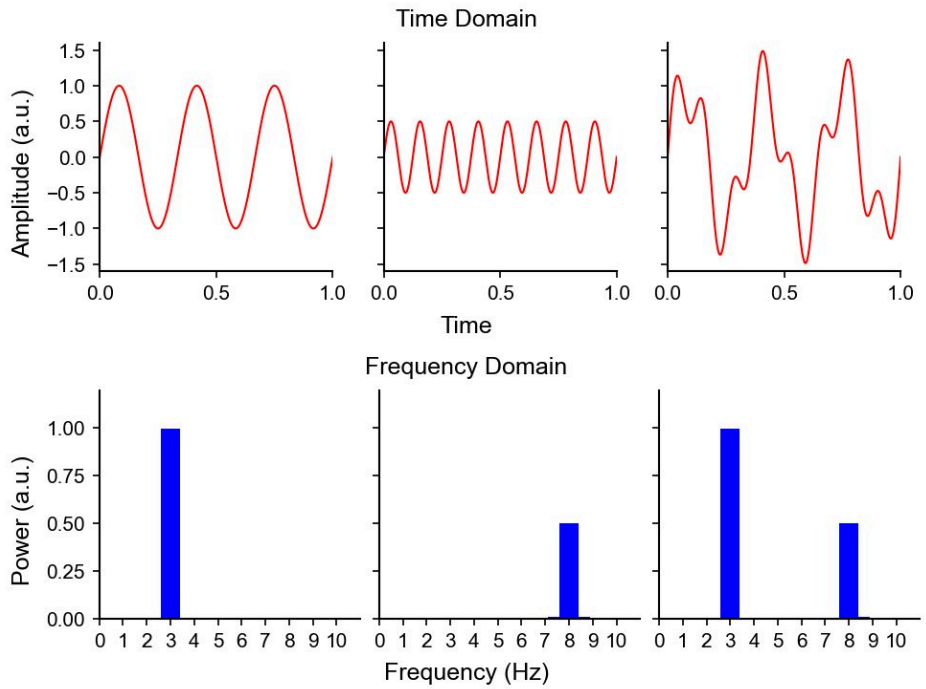
Order, trend, and seasonality contain information that one mean cannot capture



A changing mean violates weak stationarity; a single average hides the trend

Time Domain vs Frequency Domain

The same signal: a tangled wave in time becomes two bars in frequency



$$X(f) = \int x(t)\, e^{-2\pi i f t}\, dt$$ the Fourier transform; the fast Fourier transform (FFT) computes it quickly

What looks tangled in time becomes **two bars in frequency**

Statistical Exploration

01

Descriptive Analysis

Reduce the data to a few numbers

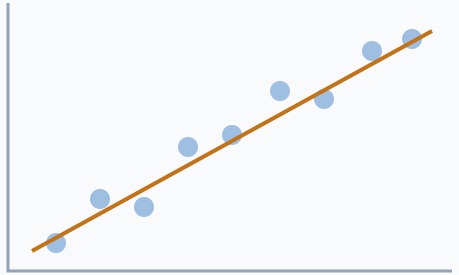
02

Visual Exploration

Find patterns by plotting the data

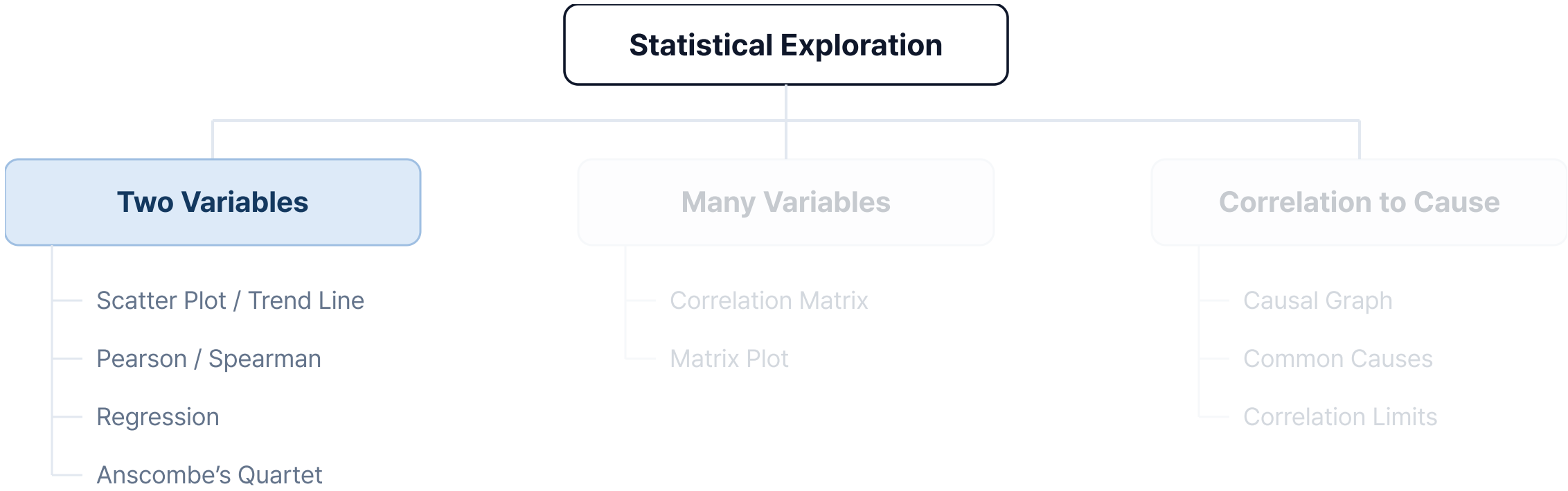
03

Statistical Exploration



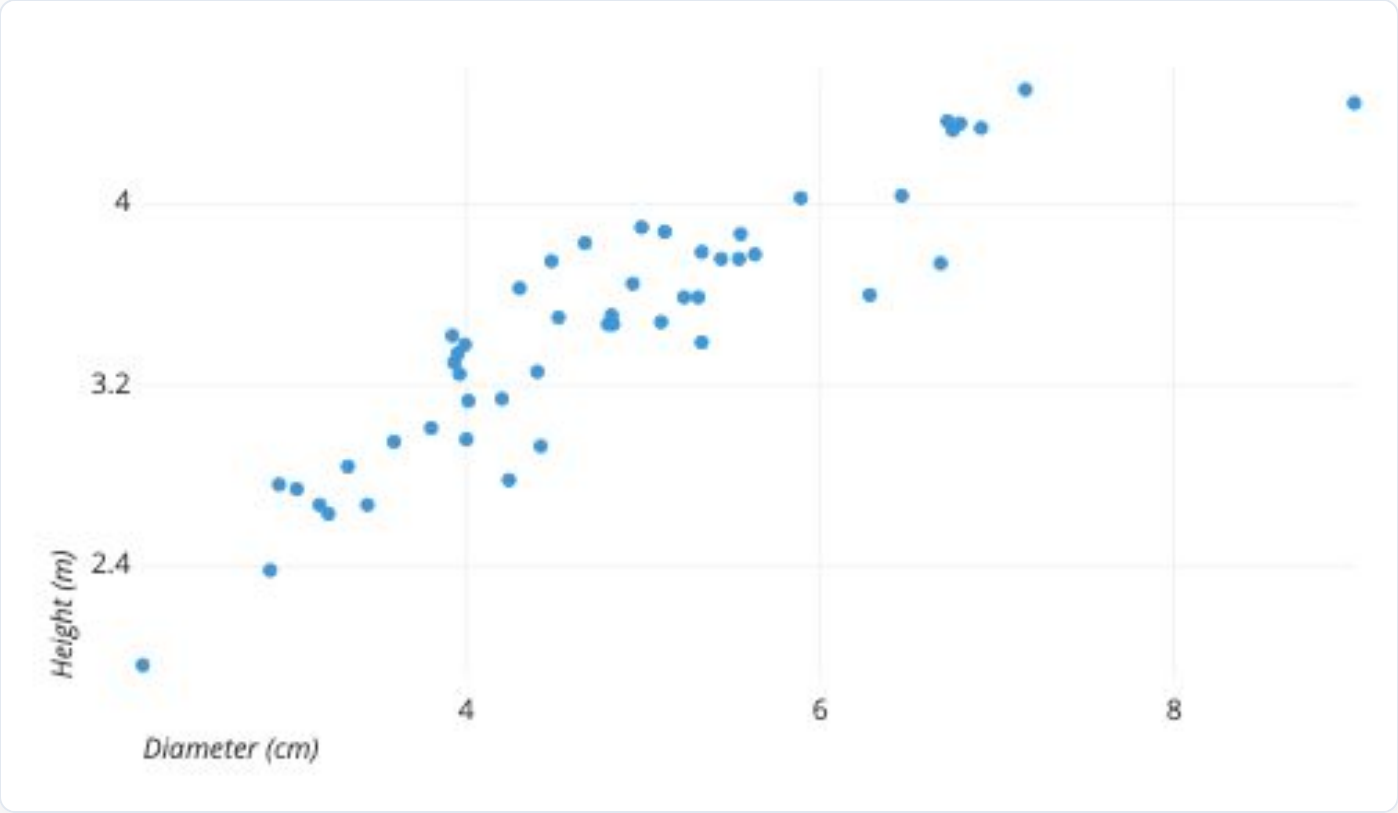
Read relationships between variables

Two Variables



Scatter Plot

One point per observation shows the relationship before a fitted model



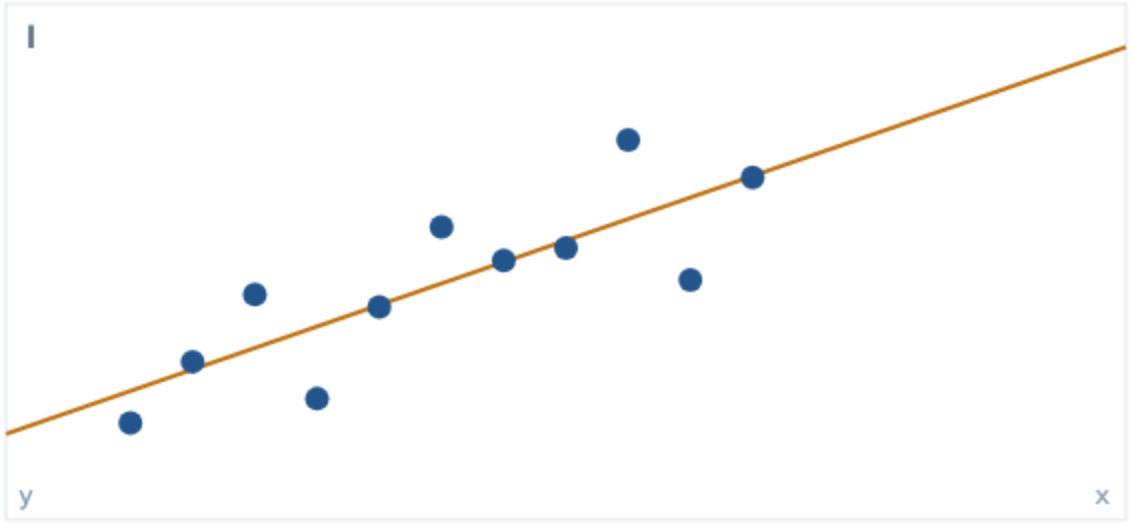
Direction, shape, spread, and the rows that sit apart, **all before anything is computed**

Demo: Anscombe's Quartet

Same mean, same variance, same correlation, same fitted line

1. Do equal summaries imply equal shapes? 2. Which plot makes the fitted line misleading?

I | II | III | IV | All four



mean x **9.00** · sd x **3.32** · mean y **7.50** · sd y **2.03** · r **0.82** · fit **y = 3.00 + 0.50x**

Answer: Anscombe's Quartet

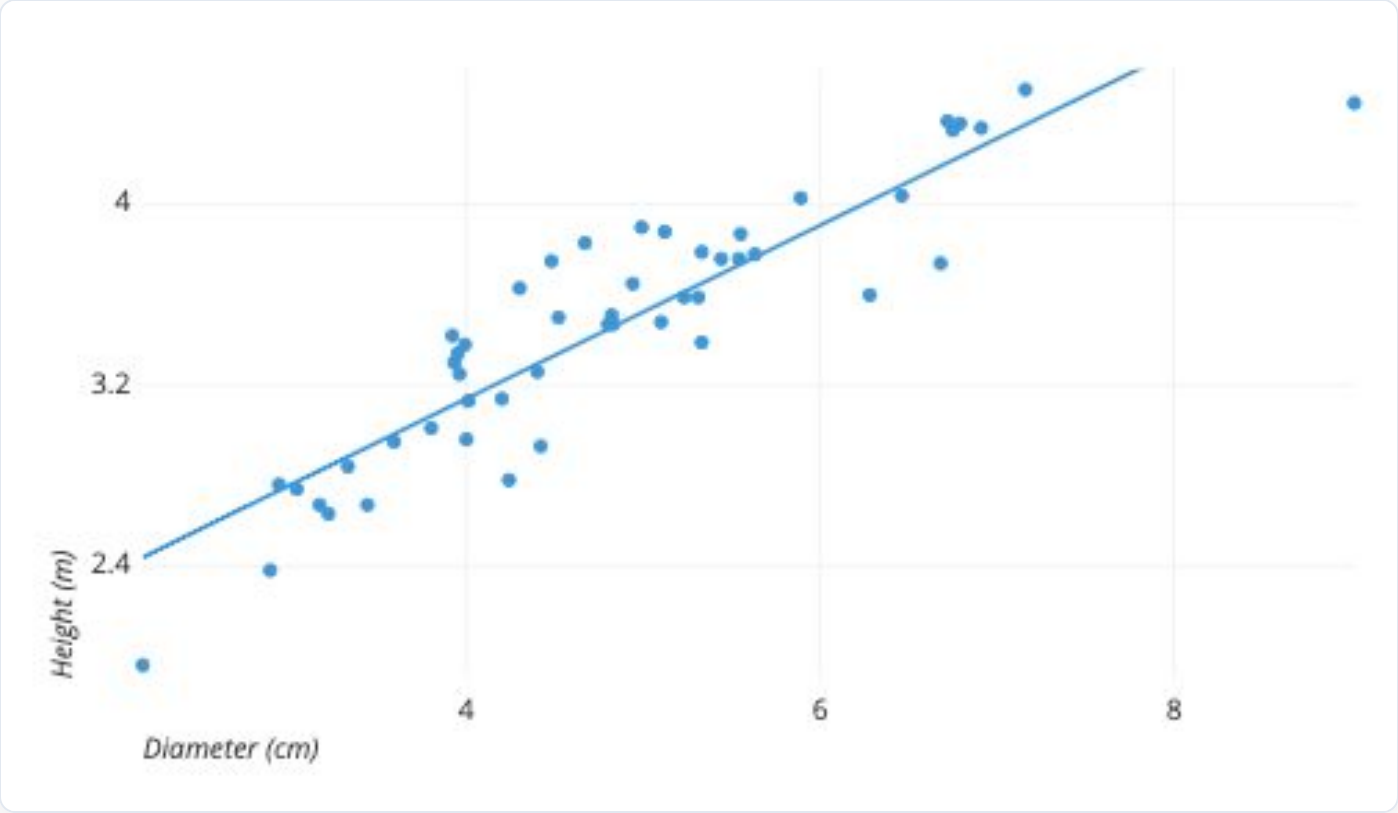
The same summary statistics describe different relationships

What stays the same	What the plots reveal
Mean and spread	Different geometry
Correlation ≈ 0.82	Curve, outlier and leverage
Line y ≈ 3 + 0.50x	A line is not equally informative in every plot

Plot the observations before interpreting a fitted relationship

Fitting a Line

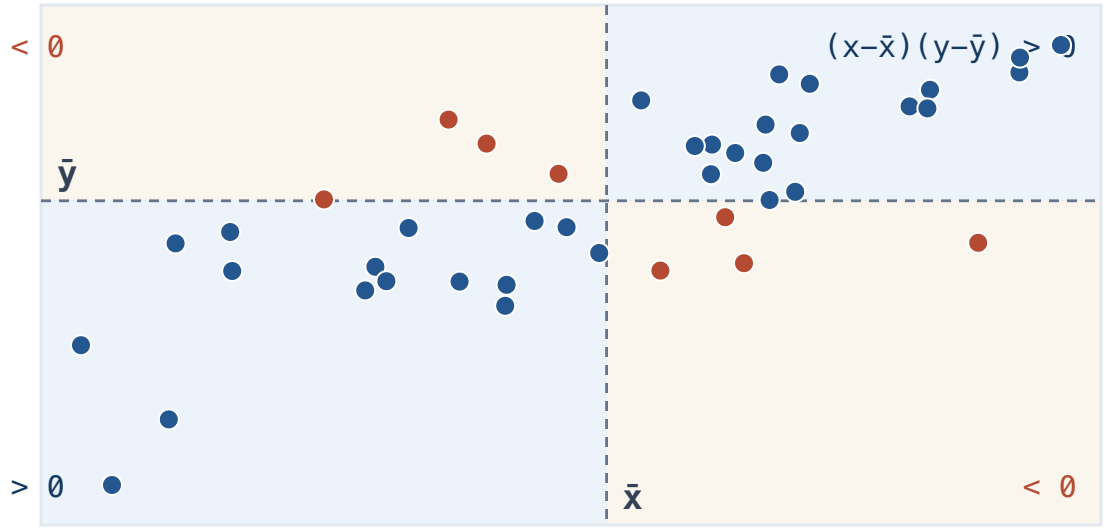
A fitted line summarizes a linear trend; residuals show departures



A fitted line defines residuals; a distant point warrants investigation

Covariance of Two Columns

Two columns can spread out together, measured in the units of both



one column

$s^2 = \mathrm{cov}(x, x)$

p columns

$\Sigma_{ij} = \mathrm{cov}(x_i, x_j)$

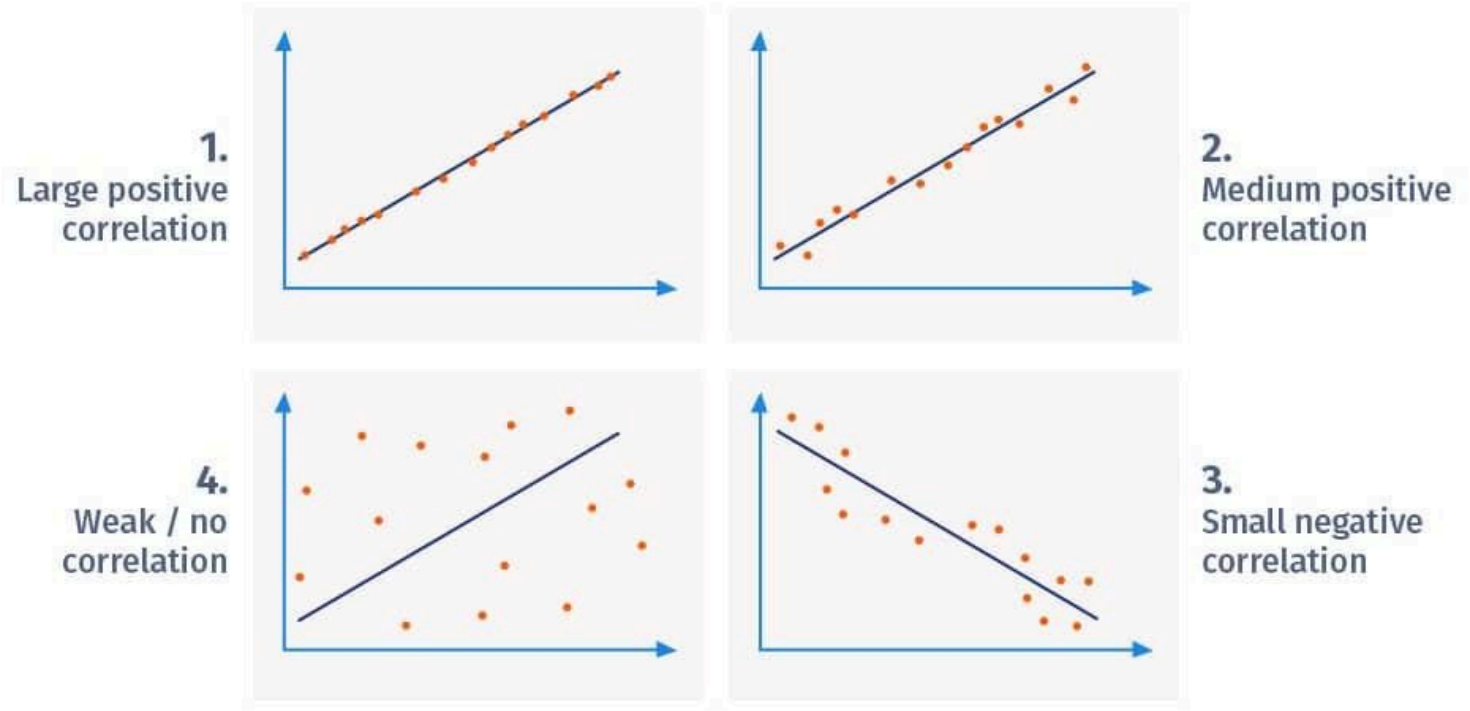
a p × p matrix, s^2 on the diagonal

$\mathrm{cov}(x, y) = (1/(n-1)) \sum_i (x_i - \bar{x})(y_i - \bar{y})$ the SD formula with the second factor taken from another column

Positive products add, negative products subtract; their weighted sum determines the covariance

Pearson Correlation

r measures a straight line, not any relation: a scatter plot reduced to one number

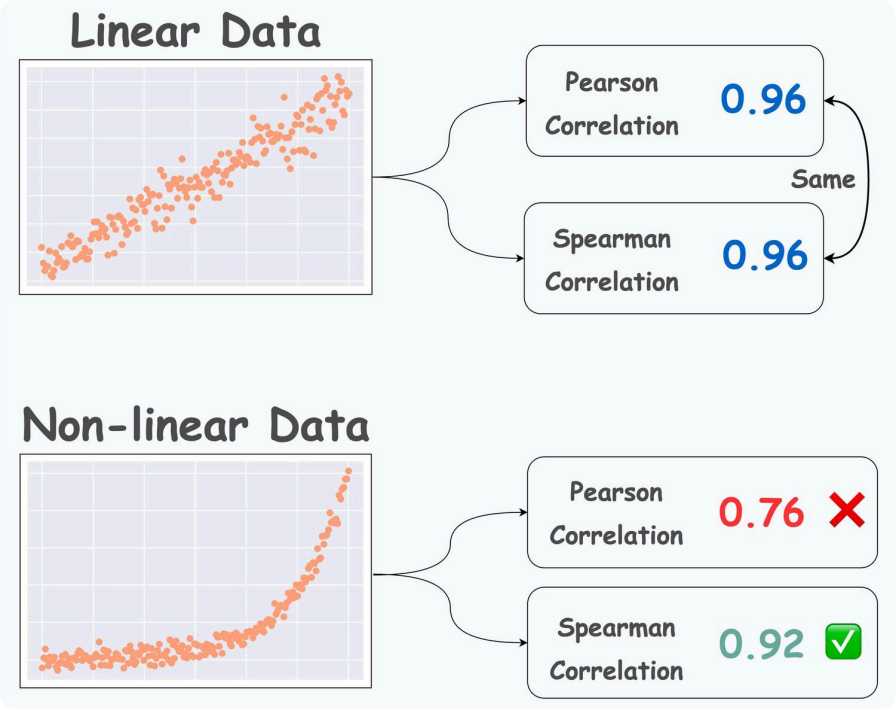


$$r = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \,/\, \sqrt{\left(\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2\right)}$$ the same thing as $\mathrm{cov}(x, y) / \sigma_x \sigma_y$

Near-zero r rules out strong linear association; nonlinear relationships may remain

Spearman's Rank Correlation

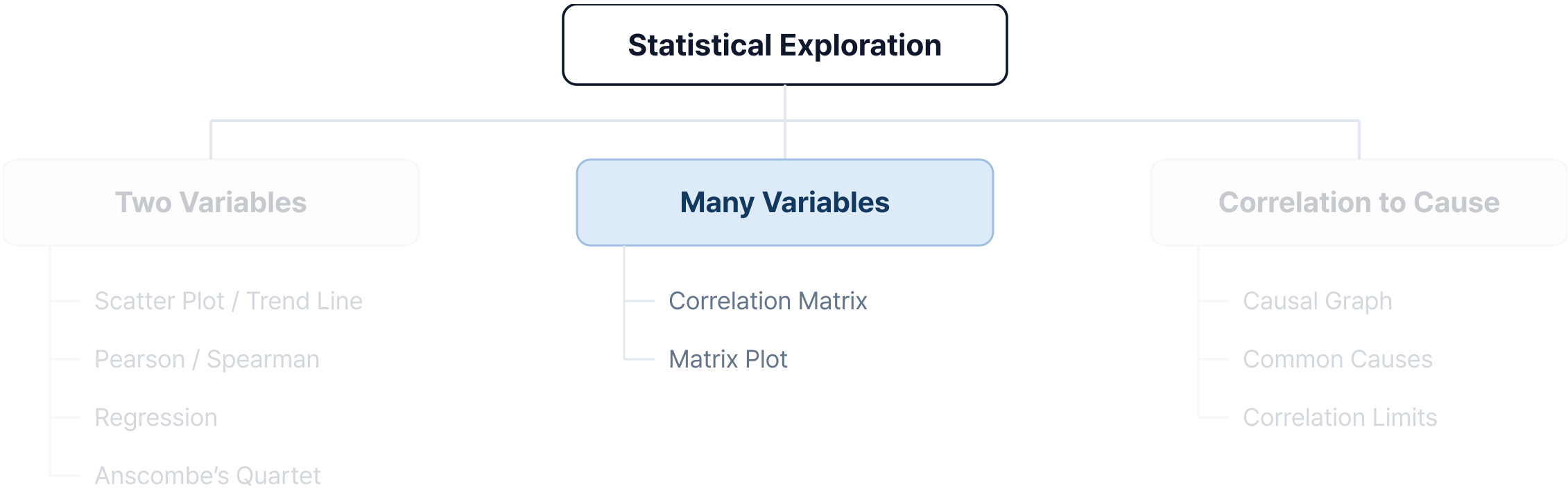
Ranks catch the curve that Pearson underrates: does it keep going one way?



$$\rho_s = \text{corr}(\text{rank}(x), \text{rank}(y))$$

Rank correlation captures the monotone curve: Pearson 0.76, Spearman 0.92

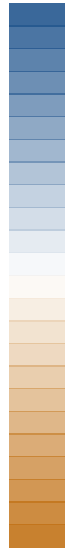
Many Variables



Correlation Matrix

Pairwise correlations reveal related groups and a weakly related humidity column

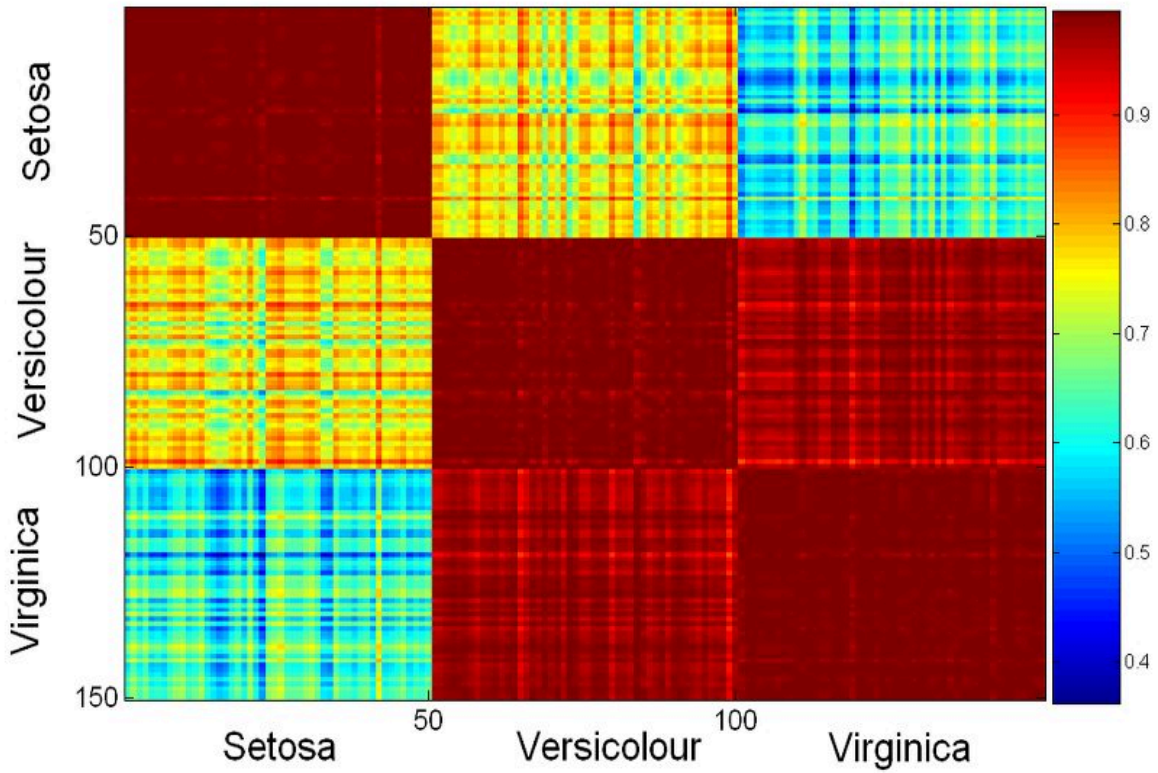
	Speed	Vibration	Temp	Defect	Rework	Humidity
Speed	1.00	0.82	0.81	0.47	0.42	0.08
Vibration	0.82	1.00	0.81	0.40	0.39	0.10
Temp	0.81	0.81	1.00	0.42	0.40	0.12
Defect	0.47	0.40	0.42	1.00	0.76	−0.03
Rework	0.42	0.39	0.40	0.76	1.00	−0.02
Humidity	0.08	0.10	0.12	−0.03	−0.02	1.00



Speed, vibration and temperature are strongly related; humidity stays weakly related

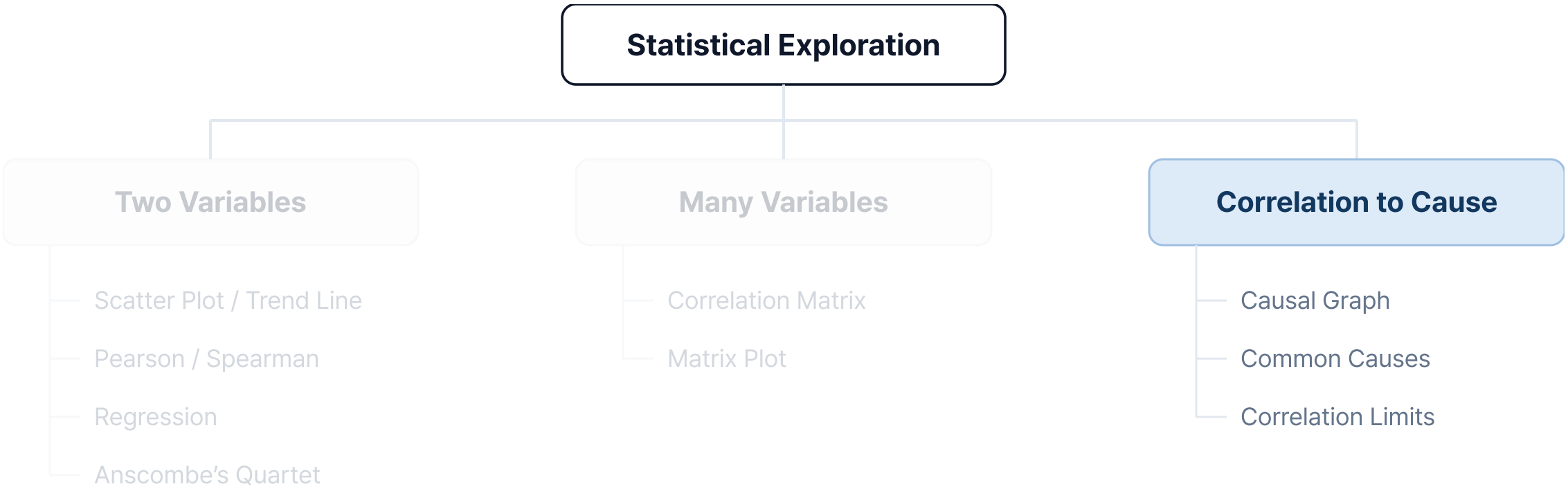
Matrix Plot

Reorder both axes by species and similar samples form blocks along the diagonal



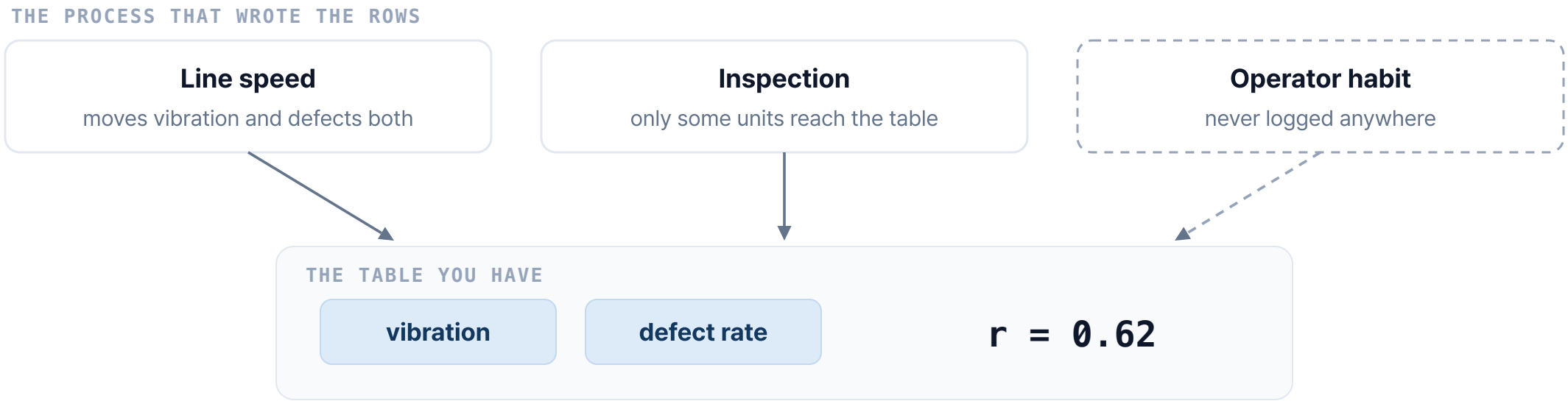
Reordering both axes reveals **similarity blocks** that a random order hides

Correlation to Cause



What wrote this table?

Exploration reads the rows; what the rows mean is decided one level up



The correlation is observed in the table; **its meaning depends on the process above it**

Causal Graphs

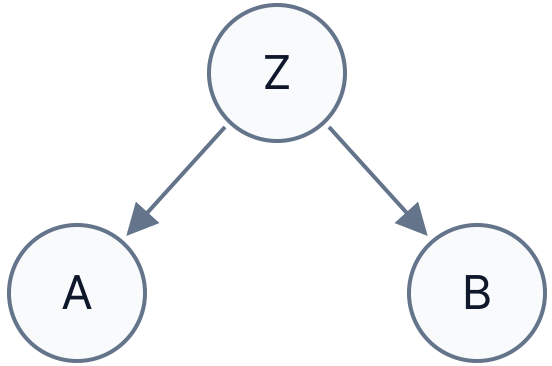
The same correlation can reflect either causal direction or a shared cause



A causes B



B causes A

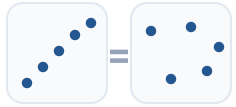


Z causes both

Three structures give one correlation; **the correlation alone cannot tell them apart**

Takeaway: Look before you summarize

Exploration reveals patterns whose meaning depends on measurement and modeling choices



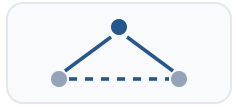
A summary is a lossy compression

Anscombe’s datasets share several summary statistics but have different shapes



Every tool here hides a choice

Bin width, perplexity, k, and the distribution you assumed



A correlation is a finding, not a causal conclusion

What it means depends on what wrote the table



Explore before you clean

You can only decide what to fix once you know what is strange