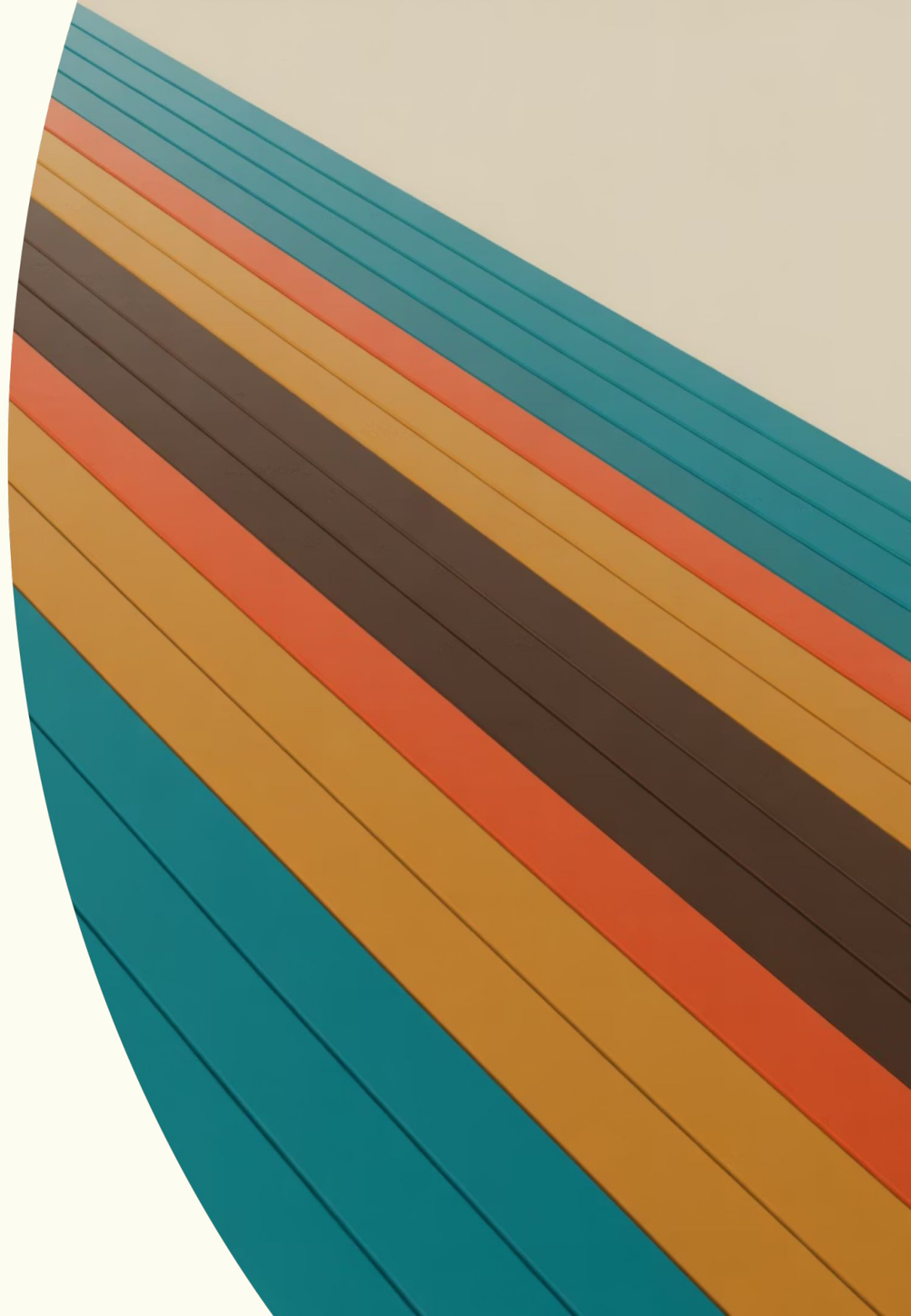
Module 10.4: Assumptions & Limitations of Linear Regression

Understanding when linear regression works and when it fails



Why Assumptions Matter

> Linear regression is powerful, but it only works reliably when certain conditions are met. When these assumptions are violated, you risk:
>
> **Unreliable predictions** that don't match reality
> **Misleading coefficients** that distort relationships
> **Increased errors** that reduce model accuracy

Think of assumptions as the foundation of a building — without them, everything crumbles.

Example: Predicting house prices based on square footage requires meeting key assumptions for accuracy.

Assumption 1: Linearity

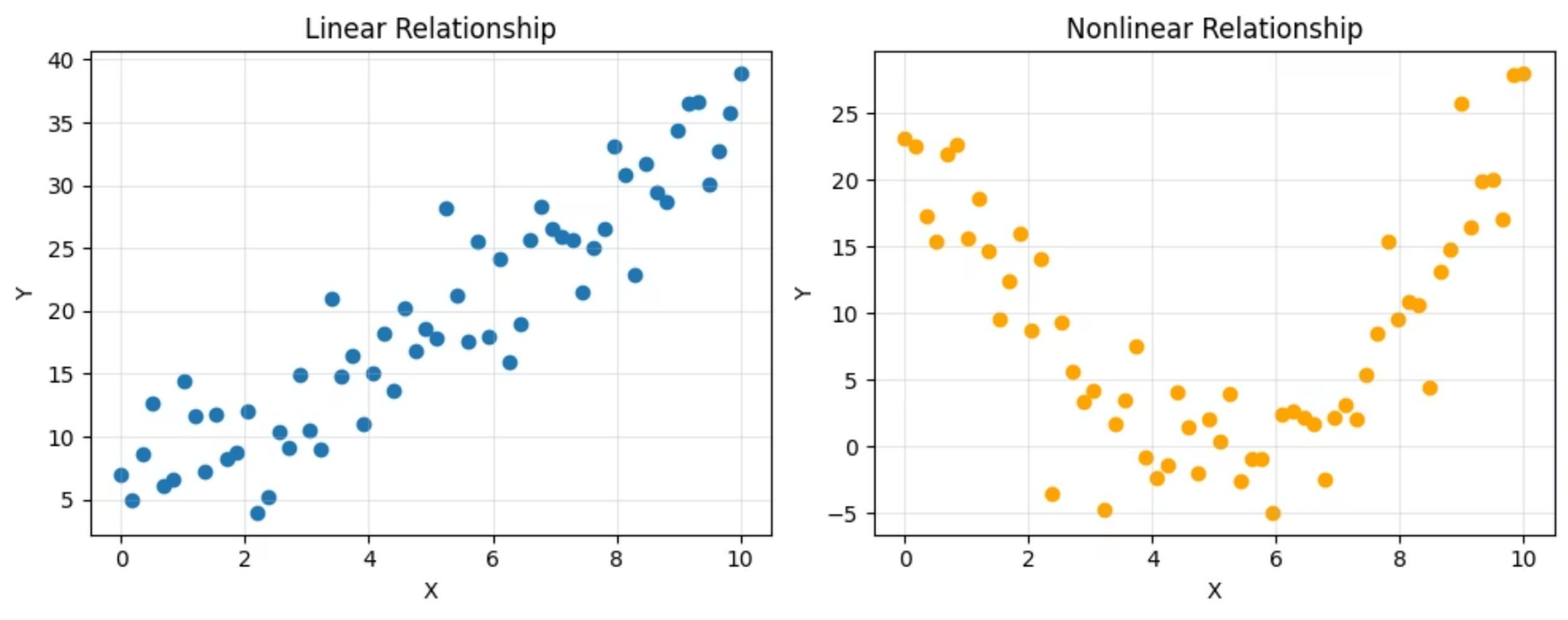
The relationship between your input and output variables must follow a straight line. If the data curves, linear regression will miss the pattern entirely.

✓ Linear Relationship

Good example: Hours studied vs. test marks shows a clear straight-line pattern.

✗ Non-Linear Relationship

Bad example: Age vs. income follows a curve — early growth, then plateau.



Assumption 2: Independence of Errors

Each prediction error (residual) should be independent — one error shouldn't influence the next. This is especially important in time series data.

Independent Errors ✓

Errors appear random with no detectable pattern — exactly what we want.

Dependent Errors ✗

Wave-like patterns suggest errors are related — common in temperature predictions over time.

Real-Life Example: Imagine predicting daily ice cream sales.

Independent errors ✓: Sales vary randomly day-to-day due to unpredictable factors (someone's birthday party, a food truck nearby). Each day's error is unrelated to the previous day.
Dependent errors ✗: If Monday is hot and sales are high, Tuesday (also hot) will likely have high sales too. The errors are connected because temperature patterns persist across consecutive days, creating a pattern in our prediction mistakes.

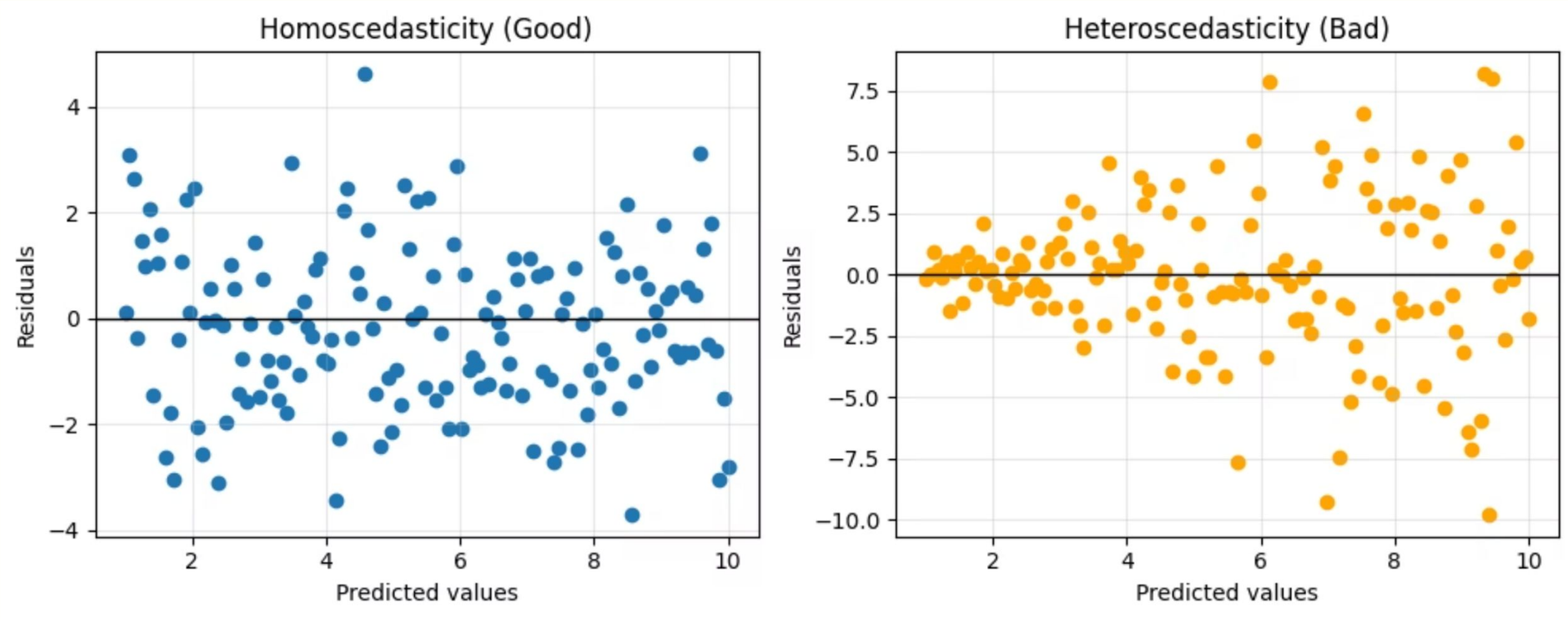
Assumption 3: Homoscedasticity

The variance of residuals should stay constant across all predicted values. When variance increases or decreases with predictions, we have *heteroscedasticity* — a fancy word for unequal spread.

Example: Salary vs. Experience

Early-career salaries cluster tightly, but senior salaries spread widely. This creates a funnel shape in residuals — a red flag for linear regression.

The model becomes less reliable as experience increases.



Assumption 4: Normality of Residuals

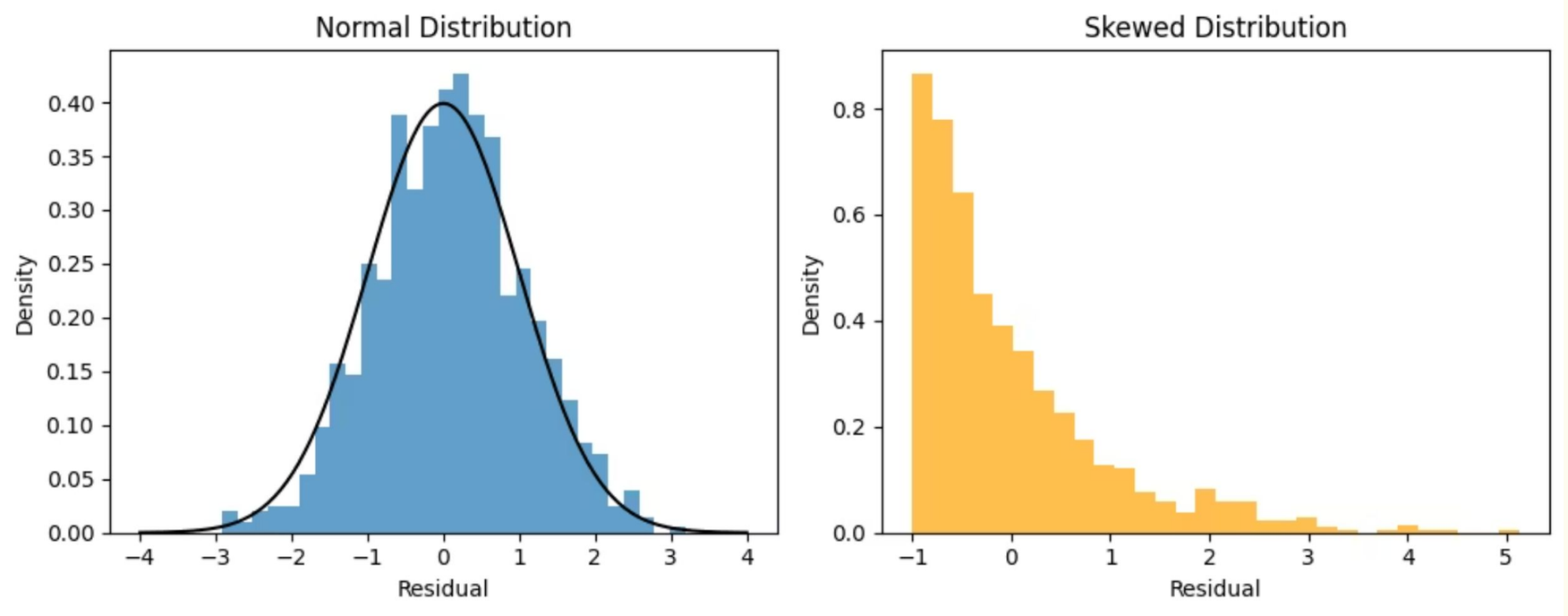
Residuals should follow a bell curve (normal distribution). When residuals are skewed or have heavy tails, confidence intervals and hypothesis tests become unreliable.

Normal Distribution

✓ Errors are symmetrically distributed around zero — model assumptions are met.

Skewed Distribution

✗ Errors lean heavily to one side — predictions and confidence intervals will be off.



Assumption 5: No Multicollinearity

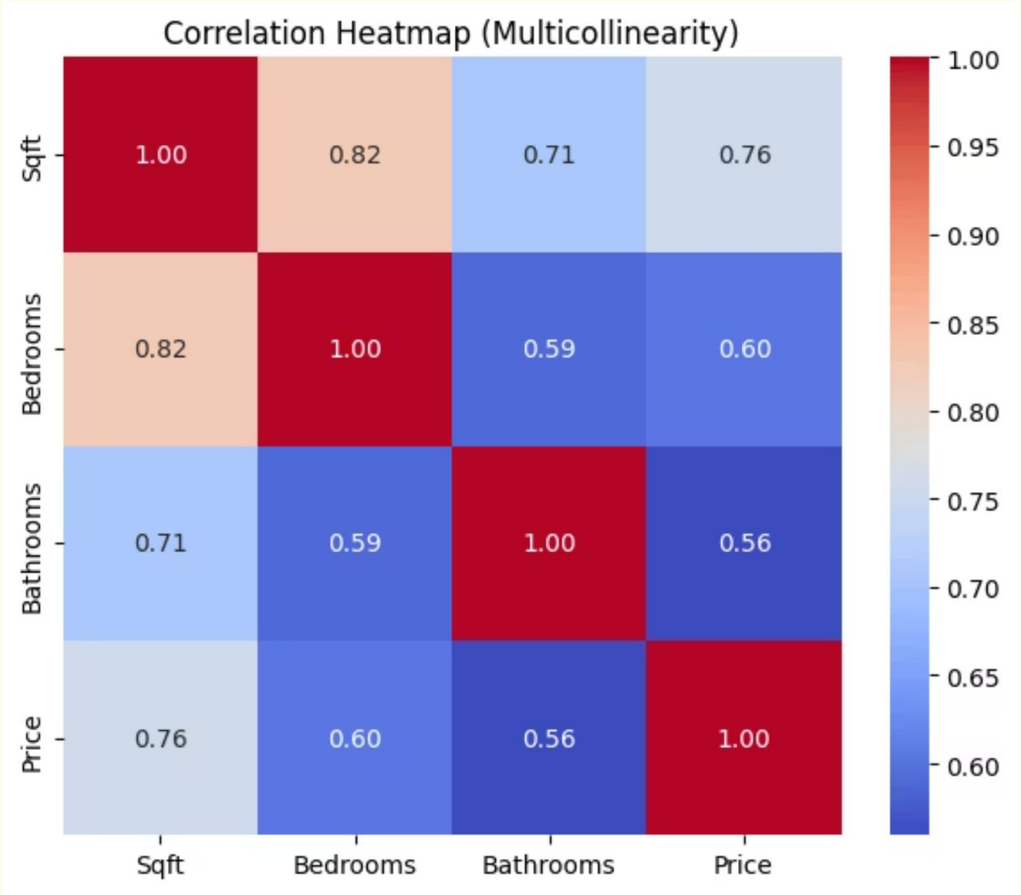
When predictor variables are highly correlated with each other, it becomes impossible to isolate individual effects. This inflates coefficient standard errors and makes interpretation unreliable.

Classic Example: House Features

Variables like **bedrooms**, **bathrooms**, and **square footage** are naturally correlated. Bigger homes tend to have more of everything.

When features move together, the model can't distinguish which one truly drives price changes.

Darker squares reveal dangerous correlations between predictors.



Real-Life Assumption Violations

Even carefully designed models can violate assumptions. Here are three common scenarios data scientists encounter in practice:

Stock Price Prediction

Violation: Autocorrelation

Yesterday's price strongly influences today's price, creating dependent errors that violate independence assumptions.

Medical Cost Estimation

Violation: Heavy Outliers

Rare catastrophic illnesses create extreme values that distort the entire model and skew predictions.

Social Media Engagement

Violation: Heteroscedasticity

Viral posts create huge variance in likes, while most content has predictable, low engagement — unequal variance throughout.

Fundamental Limitations of Linear Regression

Beyond assumption violations, linear regression has inherent boundaries that no amount of data cleaning can overcome:

Cannot Handle Non-Linear Patterns

Complex curves, exponential growth, and polynomial relationships require different modeling approaches.

Highly Sensitive to Outliers

A single extreme value can dramatically shift the regression line and distort all predictions.

Cannot Classify Categories

Predicting yes/no, win/lose, or multiple classes requires logistic regression or other classification methods.

Misses Complex Interactions

When variables combine in sophisticated ways, linear regression simple additive approach falls short.

When NOT to Use Linear Regression

Recognizing inappropriate use cases is just as important as knowing when to apply linear regression. Watch for these warning signs:

Non-Linear Relationships

When scatter plots reveal curves, switches to polynomial regression or other flexible models.

Strong Seasonality

Repeating patterns over time require time series models that capture cyclical behavior.

Extreme Outliers Present

Consider robust regression techniques or remove outliers before fitting standard linear models.

High Multicollinearity

Use regularization methods like Ridge or Lasso regression to handle correlated predictors.

Classification Problems

Predicting categories requires logistic regression, decision trees, or neural networks instead.

Key Takeaways

The Five Critical Assumptions

Linearity — Straight-line relationships
Independence — Errors don't influence each other
Homoscedasticity — Constant variance
Normality — Bell-curved residuals

No Multicollinearity — Uncorrelated predictors

Remember the Limitations

- Can't handle curves or complex patterns
- Sensitive to outliers and extreme values
- Not suitable for classification tasks
- Struggles with feature interactions

Always check diagnostics: Plot residuals, test assumptions, and validate predictions before trusting your linear regression model. When assumptions fail, explore alternative modeling approaches.

Build Better Models

Linear regression is a powerful tool when used correctly — but knowing its limits makes you a better data scientist.

> "The key to successful modeling isn't just knowing what to use — it's knowing when *not* to use it."