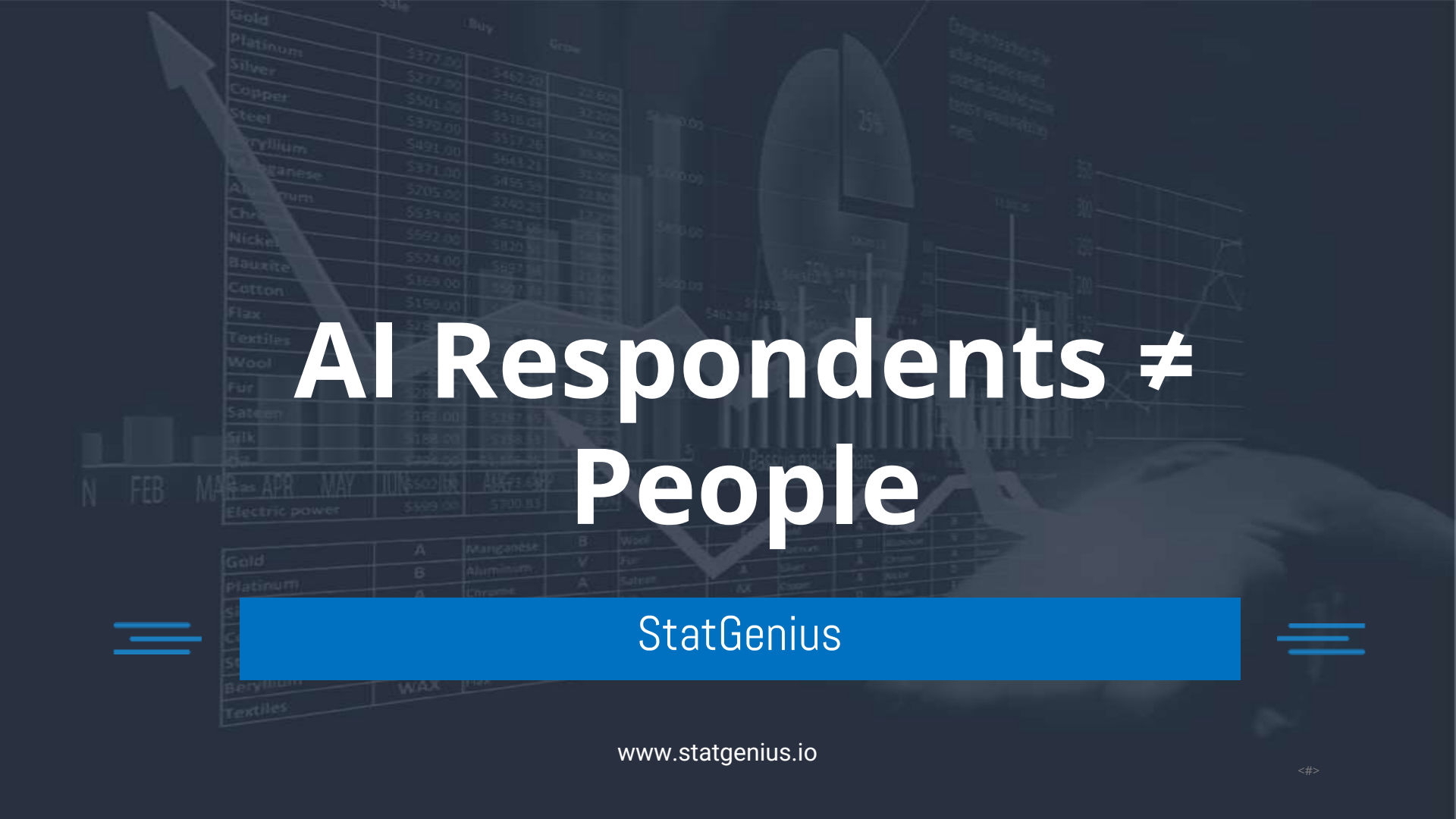


AI Respondents ≠ People

StatGenius

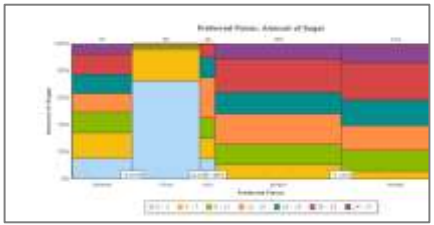
www.statgenius.io

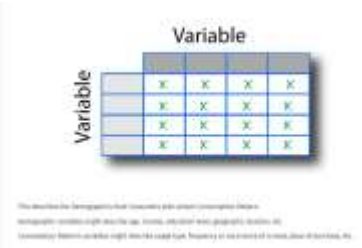
What is StatGenius ?

Co-Pilot for **Quant Insight Generation**

- Has the intelligence to mathematically analyze data *(just like a PhD)*
- Generates scientific and research discoveries

The Only AI trusted for quant research





What are Synthetic Panels?

How Synthetic Respondents Work
Technical Review

"Right now, synthetic sample... lacks variation and nuance in both qual and quant analysis. On its own, as it stands, it's just not good enough to use as a supplement for human sample."

Jane Ostler & Ashok Kalidas (EVP & Chief AI Scientist, Kantar)

In Practice - How Panels Work

- Target audience (persona) defined, prompt written
- Prompts are fed into AI model (along with survey)
- Survey is filled out using responses from AI

Target Profile

Age: 25-40, Consume Energy Drinks 2x a Week

Example Prompt:

"You are a 32-year-old woman in Chicago who drinks Monster twice a week. Answer the survey as her."

Tech Review - How Panels Work

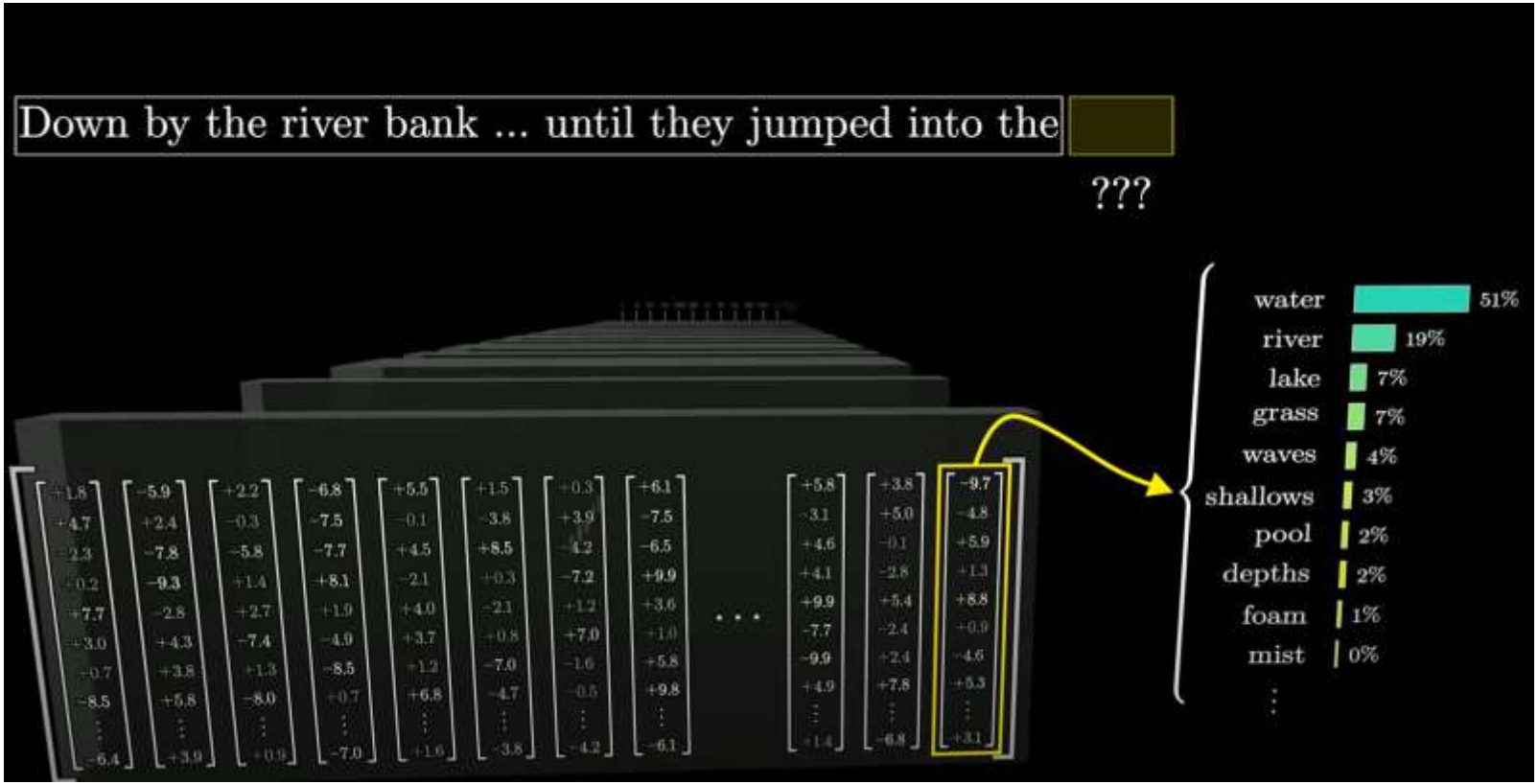
How it Works:

- Large Language Models (LLMs)
- Repeated prompting to simulate responses (randomized by sampling)
- Latent Space, Public Data Sources, Private Data (RAG)

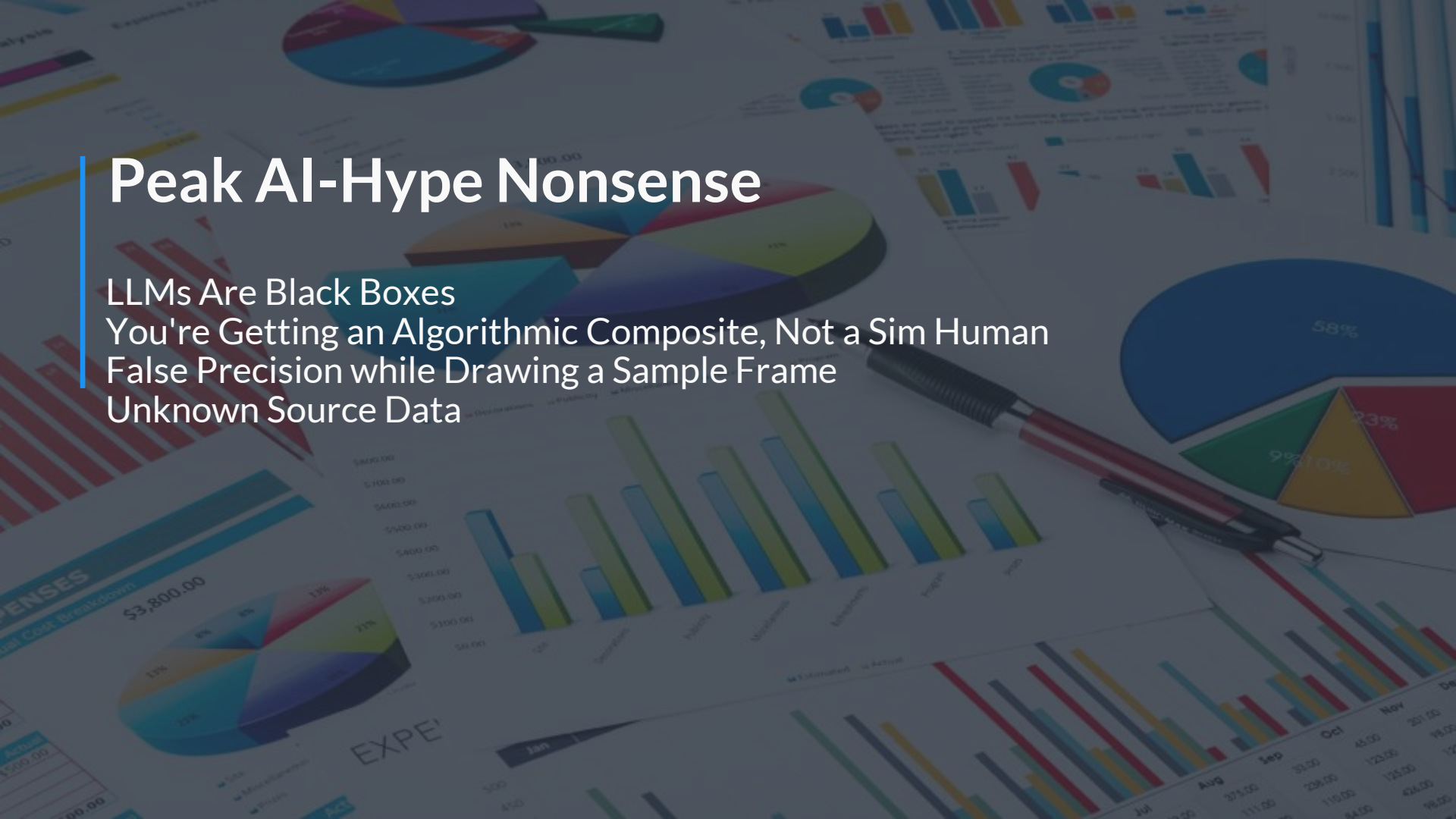
Data Sources

- Common sources: Common Crawl, Wikipedia, Reddit, news archives, scientific papers, etc.
- Documents are stored to define a persona: background, demographics, attitudes, behaviors

Sampling



https://www.youtube.com/watch?v=LPZh9BOjkQs&t=53s



Peak AI-Hype Nonsense

LLMs Are Black Boxes
You're Getting an Algorithmic Composite, Not a Sim Human
False Precision while Drawing a Sample Frame
Unknown Source Data

LLMs are Black Boxes

"**We don't really understand how these models work**. The behavior of large neural networks is very hard to explain."

— *Ilya Sutskever, Co-founder and Chief Scientist at OpenAI*
(Source: *Wired* article, 2021)

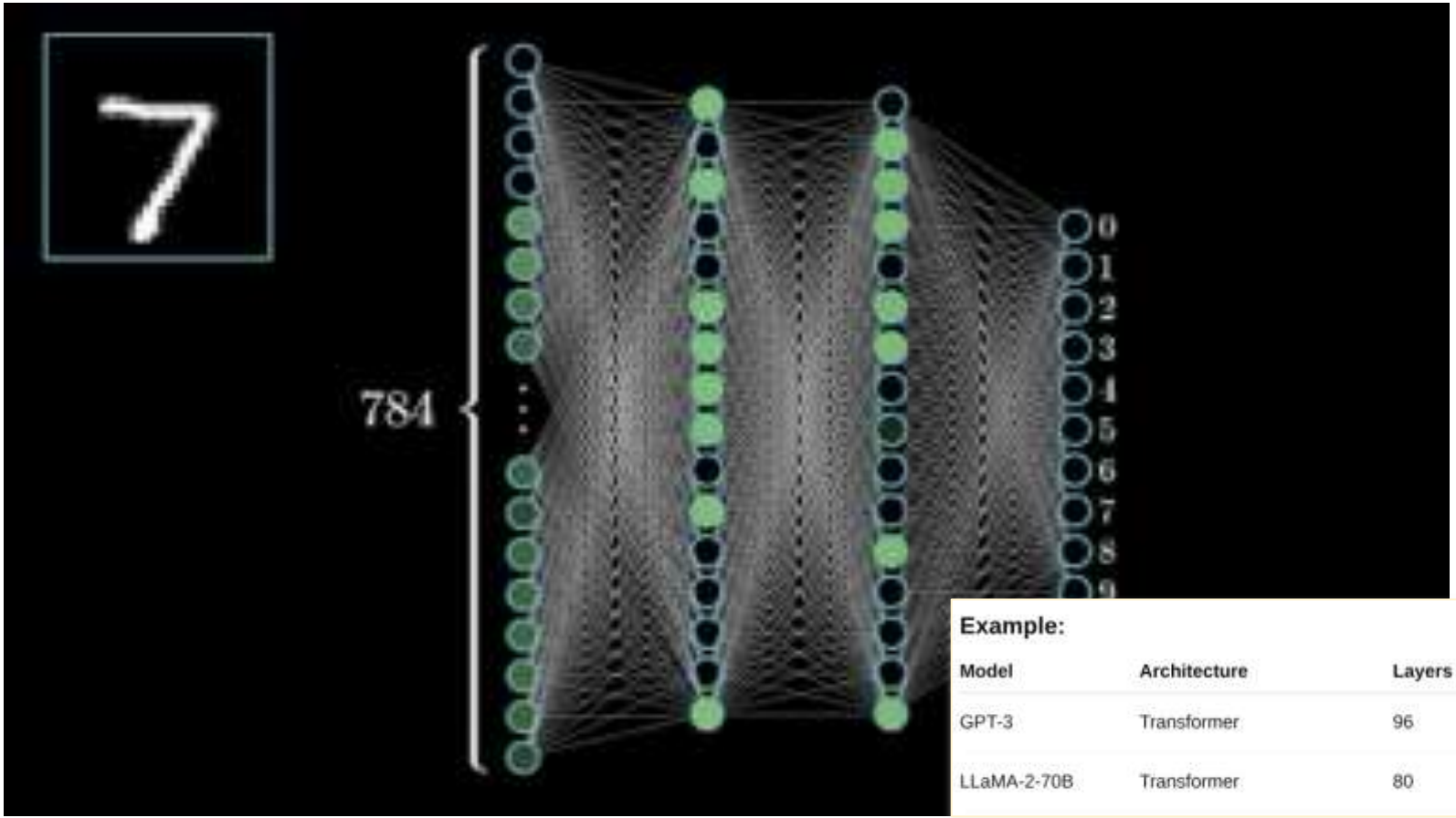
"I **don't think anyone can explain** how a deep neural network works… It's not that we don't know how to train them; we don't fully know what's going on under the hood."

— *Yann LeCun, Chief AI Scientist at Meta*
(Source: *Wired*)

"The fact that we don't know exactly why some parts of these models work is a huge problem. And **we really don't know what could go wrong** when these systems are used in different contexts."

— Chris Olah, *Co-founder of Anthropic* (Source: Blog post at Anthropic)

LLMs are Black Boxes



Example:

Model	Architecture	Layers	Parameters
GPT-3	Transformer	96	175 billion
LLaMA-2-70B	Transformer	80	70 billion

Synthetic Respondents are not Individual Minds

You're Getting an Algorithmic Composite, Not a Sim Human

Unknown Source Data

AI Doesn't Store Data—It Compresses Patterns into Latent Space

No Visibility Into Data Sources

- Sources are purposely confuscated as trade secrets
- Full transparency would expose them to legal risk (e.g., copyright infringement)

If you wanted to pull response data from online sources, then why not just use social listening?

False Precision while Drawing a Sample

Samples Pulled from Unreliable Prompt Engineering

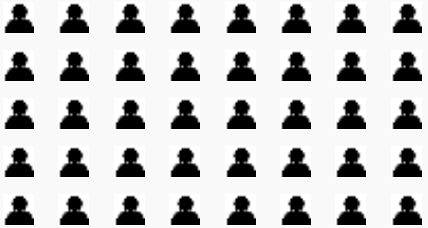
Prompts designed from linguistics, not demographic or behavioral data

Alternatives to Panels

Reusing Older (but still relevant) Data Scaling Analysis across the Enterprise

Mining Insights Across the Enterprise

Unlocking Dormant Data



Virtual Scientists at Scale

From Weeks to Hours

User Asks via Prompt:

"Help me…

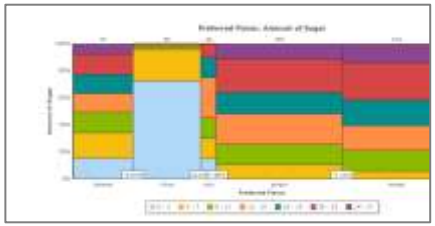
…determine the best segments to market my service"

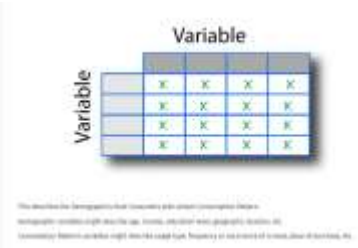
…explore satisfaction scores and tell me where to improve"

…identify customers most likely to cross-sell or upsell"

…tell me how NPS scores have changed over time"

AI Answers with Actionable Insights:





Solving Problems and Insights like a Researcher

Technical Overview for Quant Analysis

Data Mining & Transformation

Classification & data transformation for research

Mathematical Computation

Rule-based systems to ensure reliability, and output translated for non-mathematicians

Research Cases

Cases built by enterprise or provided out-the-box, based on insights & research best practices

Analysis, Insights & Discovery

Crawlers apply Research Case and Rule-Based System for data discoveries

- Exploratory Analysis
- Segment Comparison
- Descriptive Analysis
- Multivariate Regression
- Cluster Analysis

Enterprises Leveraging AI Discovery

- **Enterprise**-Wide Data **Mining** (in near real-time)
- **Unlimited** Statistical Power
- Accelerated **Innovation** & Strategy
- Subscription Service for Decision-Makers
- **Reuse** and **Amplify** Past Research

Agile Turnaround & Workflow

-60% Time spent on analysis & coordination

Scalability & Cost

+70% Savings on labor cost to complete analysis

Improved Insights Quality

120% Number of additionally identified trends

Real-World Example: Making Insights Actionable

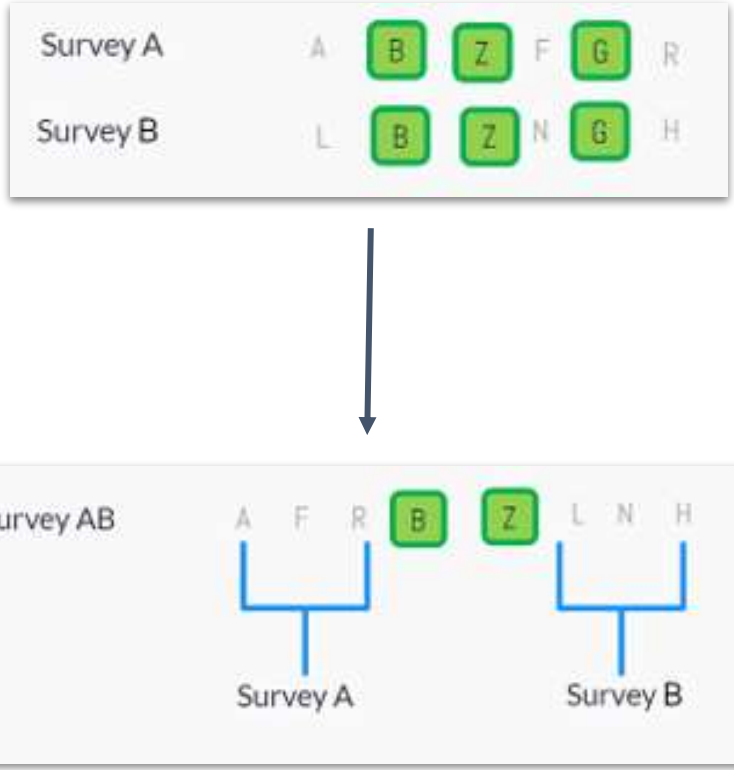
AI & Quant Research: Case Example

Synthesis - Without the AI

Statistically merging data sets to draw insights across surveys

How the Algorithm Works

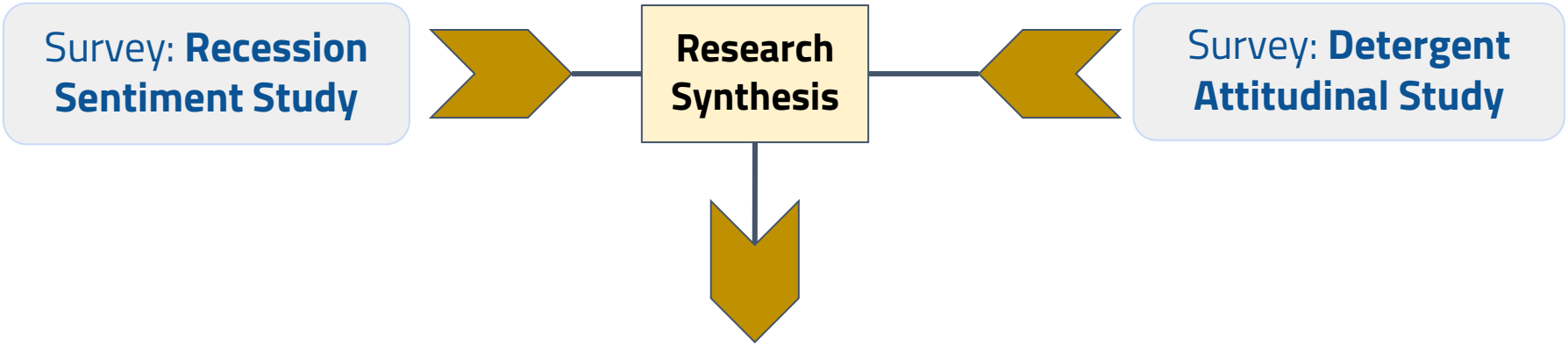
- Matching Common Variables
- Establish Segment Schema
- Partitioning Records
- Randomly Impute Values by Micro-Segment into Joined Study



Patent No. 12,019,593 - SYSTEM AND METHOD OF JOINING RESEARCH STUDIES TO EXTRACT ANALYTICAL INSIGHTS FOR ENABLING CROSS-STUDY ANALYSIS

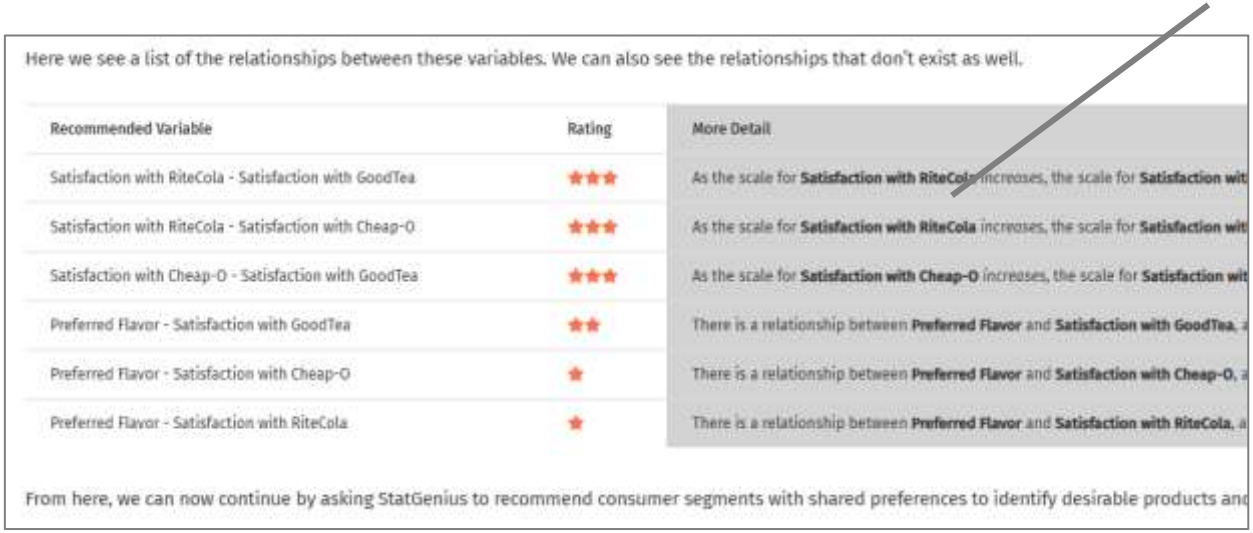
Case Example: Top 50 CPG Firm

Multinational CPG Enterprise - Seeking Insights from Past Research across Divisions



Detergent Preferences in a Recessionary Environment

Statistical relationships explained in plain English

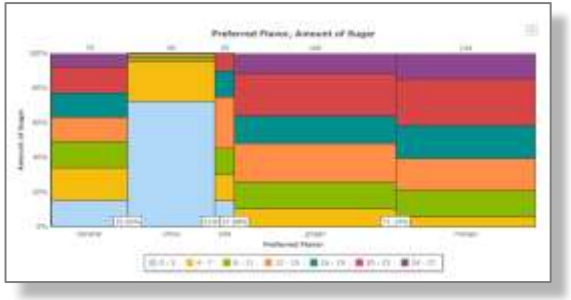


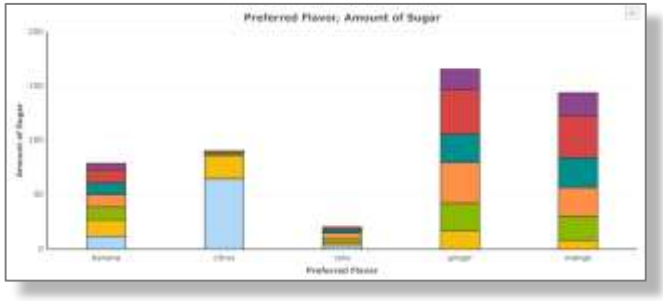
Here we see a list of the relationships between these variables. We can also see the relationships that don't exist as well.

Recommended Variable	Rating	More Detail
Satisfaction with RiteCola - Satisfaction with GoodTea	★★★	As the scale for **Satisfaction with RiteCola** increases, the scale for **Satisfaction wit**
Satisfaction with RiteCola - Satisfaction with Cheap-O	★★★	As the scale for **Satisfaction with RiteCola** increases, the scale for **Satisfaction wit**
Satisfaction with Cheap-O - Satisfaction with GoodTea	★★★	As the scale for **Satisfaction with Cheap-O** increases, the scale for **Satisfaction wit**
Preferred Flavor - Satisfaction with GoodTea	★★	There is a relationship between **Preferred Flavor** and **Satisfaction with GoodTea**, a
Preferred Flavor - Satisfaction with Cheap-O	★	There is a relationship between **Preferred Flavor** and **Satisfaction with Cheap-O**, a
Preferred Flavor - Satisfaction with RiteCola	★	There is a relationship between **Preferred Flavor** and **Satisfaction with RiteCola**, a

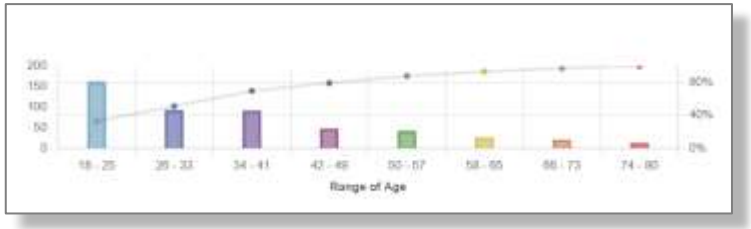
From here, we can now continue by asking StatGenius to recommend consumer segments with shared preferences to identify desirable products and

Discover Hidden Trends & Relationships





Preferred Flavor / Amount of Sugar	banana	citrus	cola	ginger	mango
8 - 3	12	65	3	0	0
4 - 7	13	21	3	17	8
8 - 11	12	2	3	25	22
12 - 15	11	2	6	38	27
16 - 19	11	0	3	26	27
20 - 23	12	0	2	47	28
24 - 27	6	0	0	19	21



Banners & Crosstabs, Charts, and Comparatives

Thank You

Invite us to speak to your team!

Review our Page: Stop Synthetic!

Upcoming Events:

- Automating Statistical Analysis w/ AI on an Enterprise Scale
- Synthetic Respondents Are AI Hype

Josh Speyer

js@statgenius.co

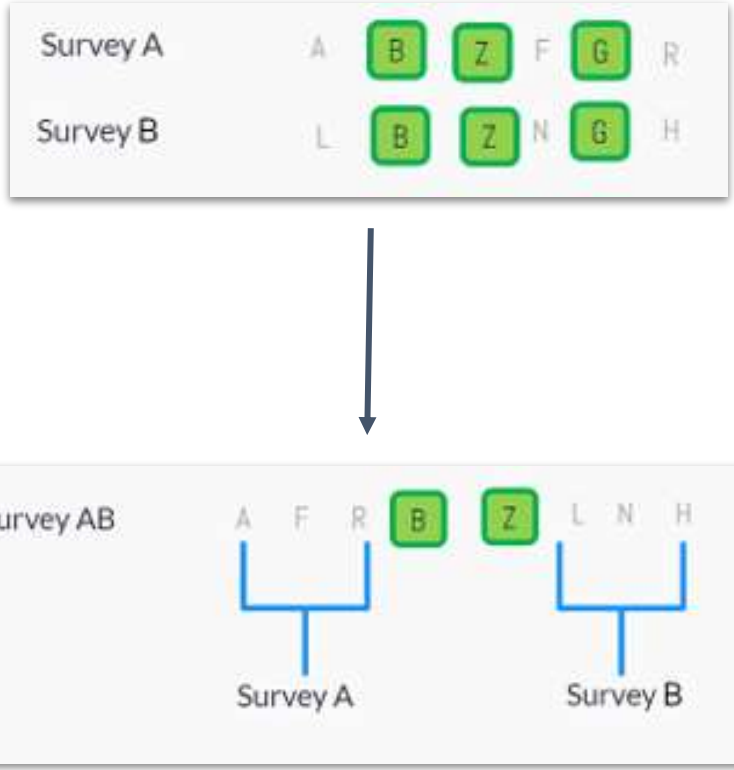

Joining Data Sets

Synthesis - Without the AI

Statistically merging data sets to draw insights across surveys

How the Algorithm Works

- Matching Common Variables
- Establish Segment Schema
- Partitioning Records
- Randomly Impute Values by Micro-Segment into Joined Study



Patent No. 12,019,593 - SYSTEM AND METHOD OF JOINING RESEARCH STUDIES TO EXTRACT ANALYTICAL INSIGHTS FOR ENABLING CROSS-STUDY ANALYSIS

Step 1: Matching Common Variables

Shared variables (and values) are normalized, for inclusion in segment schema selection

Harmonize matched variables found in both studies

- Identify variables common in both data sets
- Harmonize values in ways that individual responses can be matched between both data sets

Matching Common Variables in Both Data Sets

Example Fields or Dimensions

- *Demographic data (age, gender, marital status)*
- *Psychographic Data (Lifestyle choices, hobbies, interests, values, personality traits)*
- *Geographical Data (country, city, postal code, region)*

Step 2: Establish Segment Schema

Potential Segment Schemas are built as combinations of common variables that will be used to form segments

Finding segments as a combination of matched vars

- Create Combinations of Common variables

Combination Testing Using Confidence Interval

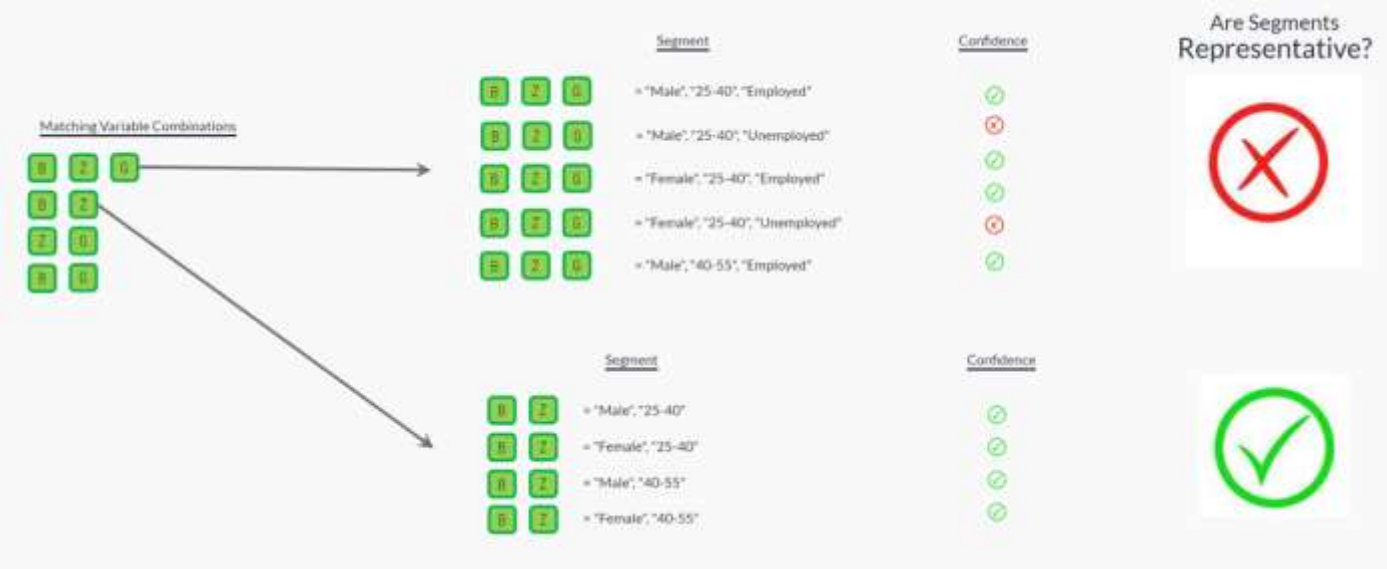
- Count minimum n responses for each Common Variable Combination
- Use confidence interval testing to determine if generated segments have enough responses for inferential analysis

Determine Optimum Segment Schema

- Find the "deepest" combination of variables that present statistically significant samples

Step 2: Establish Segment Schema

Potential Segment Schemas are built as combinations of common variables that will be used to form segments



Each variable combination is evaluated, to find a balance between statistically representative populations that are still inferential

Step 3: Partitioning Records

Prepares response data for appending / imputed by partitioning via schema segmentation

Using the chosen Segment Schema, partition data set by micro-segments

- Segments engineered to include every possible combination of variable / values
- Response data is partitioned according to micro-segment assignment

Step 4: Allocate Values into Joined Study

Randomly Impute Values by Micro-Segment into Joined Study

Data Joined using Micro-Segments as the "Primary Key"

- Results combined by matching responses in both studies within each micro-segment
- Values are imputed using unaltered data, by respondent