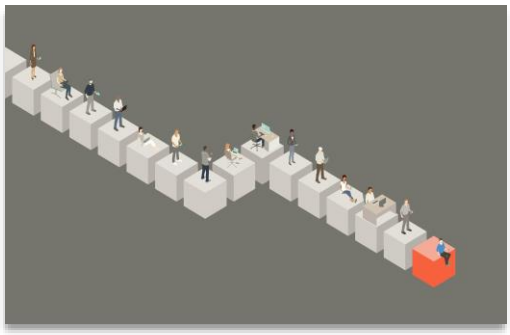
Shifting Focus to Small and Wide Data

May 2021
Coleman Harris

"By 2025, 70% of organizations will be compelled to shift their focus from big to small and wide data, providing more context for analytics and making AI less data hungry", according to Gartner's report on the Top Trends in Data and Analytics for 2021.

While machine learning methods have grown more complicated, the "big" historical data typically used to train them can become outdated quickly. This is especially true after the COVID-19 pandemic, where customer behaviors and expectations dramatically shifted and new modeling efforts were needed. This suggests a need for smaller data sources and models that are easier to implement, even as machine performance is unaffected by these shifts. These new techniques that can take advantage of small and wide data, implementing a more nimble, dynamic approach, are emerging as the immediate future of data science methods.

So, what makes data small and wide? Ross Dawson, noted strategic advisor, defines it as "diverse inputs, taking disparate sources and learning from them and their correlations without necessarily requiring the brute force of size." This differs from typical "big data" approaches like neural networks and artificial intelligence, that require copious amounts of data to reveal hidden trends. Companies that can leverage new techniques that are flexible to rapid changes in the data will provide more accurate, more adaptable business decisions in the coming future.



Data Lake and Agile Principles

One key demonstration of the power of small and wide data is the data lake. Data lakes are a type of data storage that hold large amounts of raw data from many different sources. As "big data" exploded into the technology scene, disparate data sources often were collected into a data lake or other types of infrastructure.

However, aggregating massive amounts of raw data yielded poor returns on investment - without use cases or a narrative driving the additional new sources to the data lake, what often resulted was a less useful data resource. This motivates the need for a reason, a use case, to collect new data sources. The ultimate goal is to implement a piece of data, draw value out of it, and test whether that value is relevant, before investing any costs in data infrastructure.

Shifting Focus to Small and Wide Data

Employing agile principles relies on fast iteration of new data and resources. In this scenario, teams cannot wait for large data sources to be stitched together and connected to begin work. To succeed in an agile framework, new data sources are identified and tested before the data is cleaned, transformed, and joined into the master data – saving on time and money.

Our Perspective: Rapid Data Development

The Kaizen Analytix team discovered one of our clients needed to acquire daily natural gas prices to better tune company projections. In a big data approach, an analytics team may compile natural gas, oil, and propane prices at varying levels of detail into a large database to get results. But that return on investment may be minimal – at Kaizen, we focused on acquiring the natural gas data that added direct value and then tested our hypothesis, before adding it into the data infrastructure. This acquisition was motivated by a specific use case, with immediate value - there is little reason to invest the time and cost of acquiring more data, if there is little evidence to support its utility.

Implementation

With this pivot to more valuable, more direct data sources, data scientists are turning away from black box machine learning methods to more statistically rigorous ones. These include techniques like Bayesian modeling and ridge regression, which are more powerful with less data, and can be tuned to not rely heavily on historical data.

This change in direction highlights a key difference in how data science is implemented - strategic planning vs. tactical planning. Strategic planning is a higher-level approach to hit downstream targets, typically years down the line, where tactical planning is the month-to-month shifts necessary to meet that strategic goal. As scenarios like the COVID-19 pandemic show, often the strategic goals that companies set must be changed and updated along the way.

This again requires data science modeling that is smaller and more nimble - models trained on the past 10 years of data will be overfit, or tuned to the wrong channel, for historical scenarios. Techniques that rely more heavily on statistical assumptions can be changed or updated with

major shifts, allowing for more accurate results alongside these rapid changes. This is not to say that "big data" and powerful machine learning methods have no place - these perform exceedingly well at understanding deep underlying relationships and long-term trends. But, to make tactical decision with the ability to update those forecasts nimbly requires small and wide data, and the methods that best utilize it.

The Value of Small and Wide Data

Small and wide data can be incredibly powerful, especially alongside new developments in methods that use rigorous assumptions to best handle rapidly changing data. We focus on this by motivating our analytics solutions with use cases that drive immediate value, and modeling efforts that are prescriptive. Ultimately, as we shift the focus from powerful black box solutions to defined addition of value, Kaizen sits at the forefront of implementing small and wide data in the analytics space.

About the Authors: Coleman Harris is a Content Writer for Kaizen Analytix.

Want to read more? Check out more papers here!