How Data Science Entered the Corporate World

A brief overview of how an academic branch of analysis came to dominate on Wall Street and Main Street

by Jon Shepherd

ANACONDA. EXPLORIUM Julia computing

vectice

Jon Shepherd

Sales Expert, Advisor to Vectice CEO

LinkedIn

Jon Shepherd brings valuable knowledge and experience in machine learning/AI, RDBMS, enterprise application software, cloud computing, and much more to Vectice. Jon is a product strategist with a proven track record of bringing data products to market. He is a skilled coordinator between sales, product management, development, finance, marketing, and other executive functions. Previously, he has worked at Explorium, Anaconda, Julia Computing and Netezza in various sales-related roles.

Jon has 30 years of professional experience and witnessed the explosive growth of data science first-hand. He decided to write a paper for Vectice which chronicles the major turning points in the history of data science.

Introduction

On October 2, 2006, Netflix launched a million-dollar competition that would change the world of data science. The competition, dubbed “The Netflix Prize”, challenged anyone to improve their Cinematch movie recommendation system. Just 6 days after the announcement, a team called “WXYZConsulting” had already beaten Cinematch by a small margin.

However, to be awarded the grand prize of $1 million, an improvement on the margin of error of more than 10% was needed. While gradual improvements were made, this benchmark was not reached until September 18, 2009, when team "BellKor's Pragmatic Chaos" achieved an improvement of 10.06%, and was awarded the prize money in a ceremony.

> *“The Netflix Prize” sparked public interest in the data science field and how this academic tool could be used to solve real-world business problems.*

For many participants, the Netflix competition was their first exposure to big data and predictive analytics. While the internet was already growing strong and companies were collecting large amounts of data, the tools to analyze data and make predictions were still in their infancy. Let’s look at some historical trends that enabled the emergence of machine learning as an approach to help organizations make predictions and decisions based on insights from data.

The Early Waves

One of the earliest appearances of quantitative analysis was in New York, when the rise of electronic trading systems brought math-minded academics to Wall Street to try and beat the markets. Their strength was building complex models that crunched insane amounts of data and attempted to generate higher returns by outsmarting the smartest traders. They were dubbed “quants”, and used specialized software to build proprietary algorithms to identify profitable trading opportunities.

In contrast to the qualitative approach of stock picking (investigating their product, studying the founders’ team, understanding the market, rating the balance sheet), quants only care about technical analysis of stock indicators. They use statistics and mathematics to outperform traditional stock pickers by looking for predictive clues in company and trading data.

By the early 2000s, tech-savvy investors and the insatiable demand for algorithmic trading had come to dominate Wall Street. Trading floors had mostly been replaced by computer programs running on specialized servers. Data science had evolved from small academic circles to the larger business world. The value of capturing and analyzing huge amounts of data to discover emerging trends would soon transpire to other industries.

Growing Demand For Tools

By the early 2000's, BI tools like MicroStrategy and Tableau provided dashboards and reporting tools for an organization's data warehouse. Soon, having a data infrastructure became an essential part of running an organization. This revolution was comparable to oil refinement techniques in the 19th century. While raw oil was already being extracted, it was the value of derivatives like kerosene and rubber that fueled the demand for better processing techniques.

Likewise, many companies had the ability to capture and store data, but quickly realized how much value they were missing out on because they didn't have the tools to investigate their data. Appetite for analytics grew, and the market responded by building supercomputers to handle the growing volume of data, and developing new tools to find insights in raw data.

Two toolboxes for data science that experienced a resurgence are Python (developed in 1991) and R (developed in 1993). While these programming languages had been around for years, it was only until the growth of data science in the early 2000s that their popularity exploded. New open-source packages and libraries were developed and improved by developer communities, and Python was used as the foundation of TensorFlow (Google's own machine learning library).

Some of the most popular tools in the early days of data science.

Data Science Becomes Mainstream

In 2008, the leaders of data and analytics at LinkedIn and Facebook, DJ Patil and Jeff Hammerbacher, coined the title “data scientist.” The profession blends principles from statistics, actuarial science, economics, mathematics and business into one role. Job applications flooded the job market and companies started competing to hire the brightest analytical minds, a trend that continues to this day.

By 2009, Linkedin employees took data analytics one step further. They noticed that their data revealed economic trends that they could use to predict skill gaps in the job market. Data scientists started predicting the future, not just reporting what had happened in the past. This led to a revolution in machine learning and artificial intelligence tools, which were able to learn from past data, predict the future trends and make autonomous decisions.

Cloud Providers And ML Tools

The convergence of ML/AI, big data, and cloud native has largely been built upon the back of Amazon Web Services, Microsoft Azure and Google Cloud. These cloud vendors provided flexible compute and storage and the ability to manage their infrastructure remotely or in a hybrid of cloud and on-premise data centers.

Each vendor made their own machine learning and analytics capabilities to differentiate from competitors and make ML easier on their platforms. Amazon Sagemaker offers Amazon customers capabilities to prepare, build , train and deploy models. Microsoft developed Azure Machine Learning and Google developed TensorFlow, open-source libraries for machine learning and artificial intelligence. This gave rise to “ML as a Service”, allowing developers to bring production models to market faster and cheaper than ever.

Cloud Native Companies

For the first time in history, companies with AI/ML at the core of their business plan emerged. In the Fintech space, companies like Stripe, Klarna, Quadpay and Chime challenged traditional brick and mortar banks. eCommerce giants like Amazon, FarFetch, and Warby Parker operated on such economies of scale that they nearly put physical stores out of business. Being able to hire specialists in the cloud and analytics can create different economics for ecommerce and can be managed differently than brick and mortar stores. All of this activity has put pressure on every organization to become more efficient in the use of analytics and machine learning.

The Future Of Data Science

Demand for data science talent and tools continues to outpace supply. Current trends point towards more industry adoption and the development of “no-code solutions” that allow anyone without technical skills to enter the market. While the technology is currently reserved for highly trained professionals, no-code/low-code applications will allow anyone to leverage the power of data science and machine learning. Much like the internet democratized knowledge, low-code ML and cloud computing will allow anyone with creative solutions to join the data revolution.

What is Vectice?

Vectice offers the most automated data science documentation solution, designed for instant adoption by the team with zero learning curve. Enable trustworthy AI models to be delivered 25% faster in production, no matter if they drive revenue or fall under regulation.

Vectice automates the documentation of models and data analysis by capturing metadata, lineage, and context with a single line of code. On top of the Python and R libraries, Vectice offers a web app providing real-time visibility for all your AI projects and their details.

With Vectice, you can scale AI/ML best practices, foster collaboration among various audiences from technical, legal, model risk management, and business teams, and ensure robust end-to-end documentation governance.

Try Vectice today

contact@vectice.com