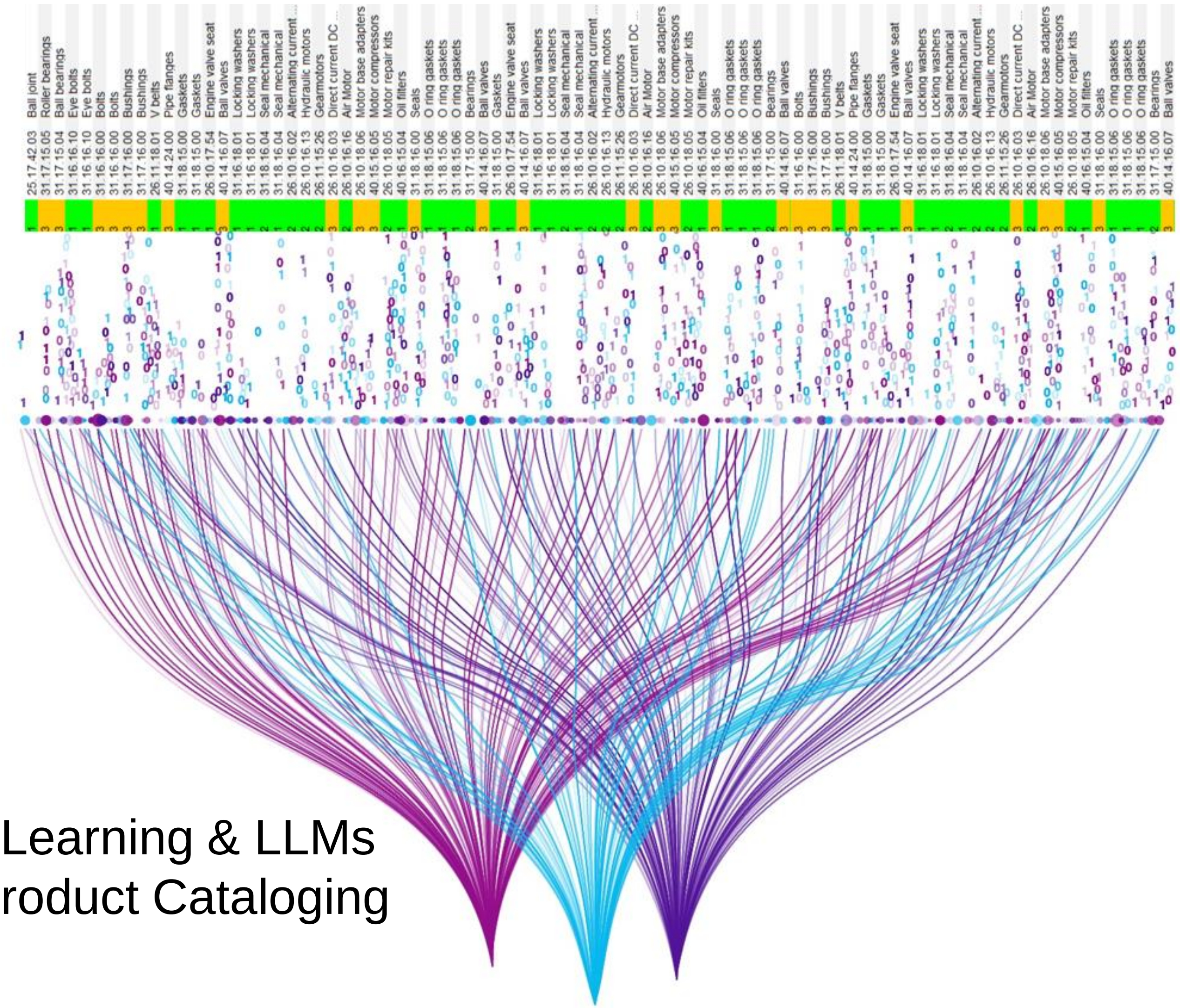
Refresh™

Machine Learning
AI Data Standardization
Software for the Enterprise



10 years of AI, Machine Learning & LLMs
for Massive Multiclass Product Cataloging

Refresh™
YOUR DATA MASTERED

A.I. WARS

Last year, one of the worlds largest companies went out to tender for specialist software to help them classify millions of products...

This wouldn't be an interesting story except that 30 of the world's best enterprise A.I. companies were invited. Some had expected Watson to do it...   but a small company from Switzerland won.

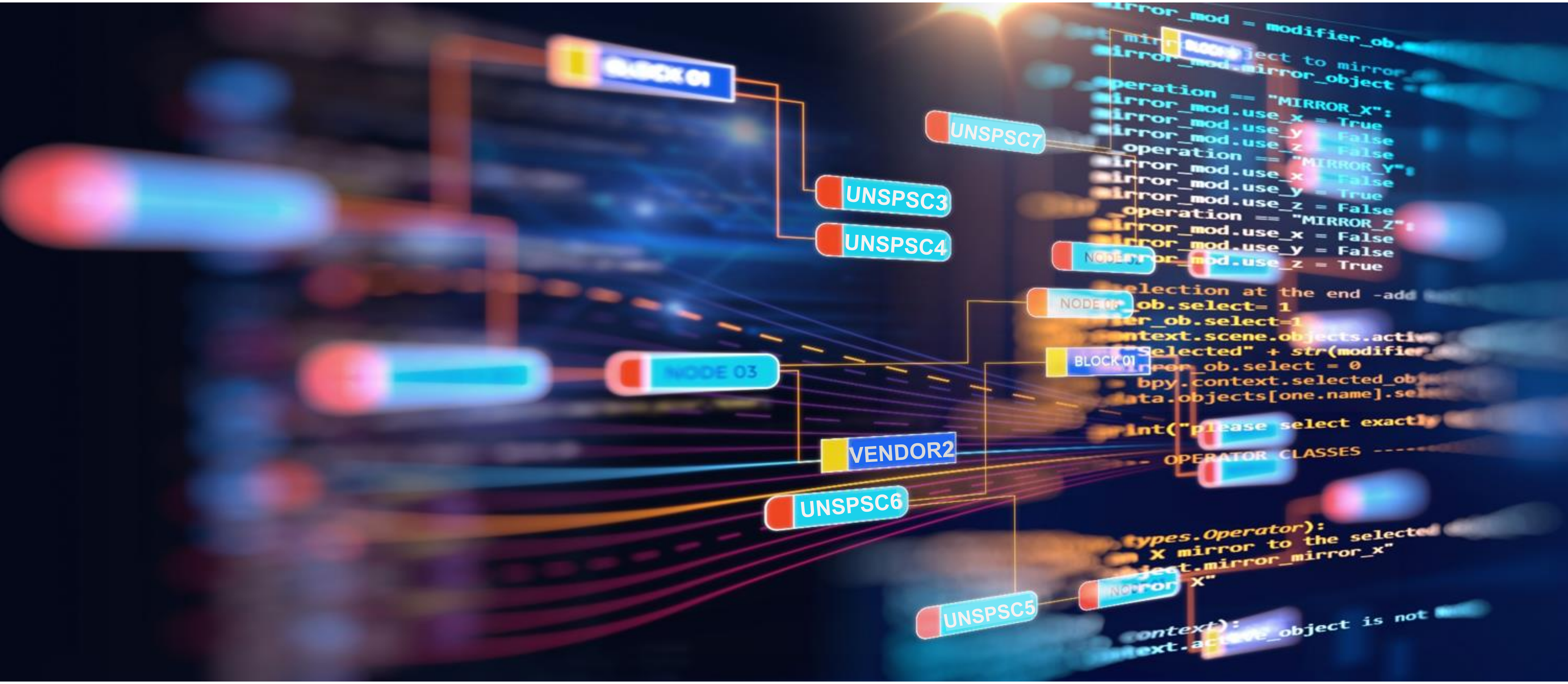
The goal seemed simple enough – take the millions of line items that this company buys each year, and classify each one to the correct spend class. While you're at it, mine out key attributes and re-describe the items correctly where the descriptions are poor.

Live demos on real data by WebEx meant that after a few months only 5 software companies were left shortlisted.

The final showdown was a live demo week where each finalist was given a blind sample of thousands of items to classify in real-time at the customer HQ and recorded in detail for clarity and fairness.

We emerged the winner.
This paper aims to explain why.

1. It is NOT good enough to ‘just use the latest’ AI Model



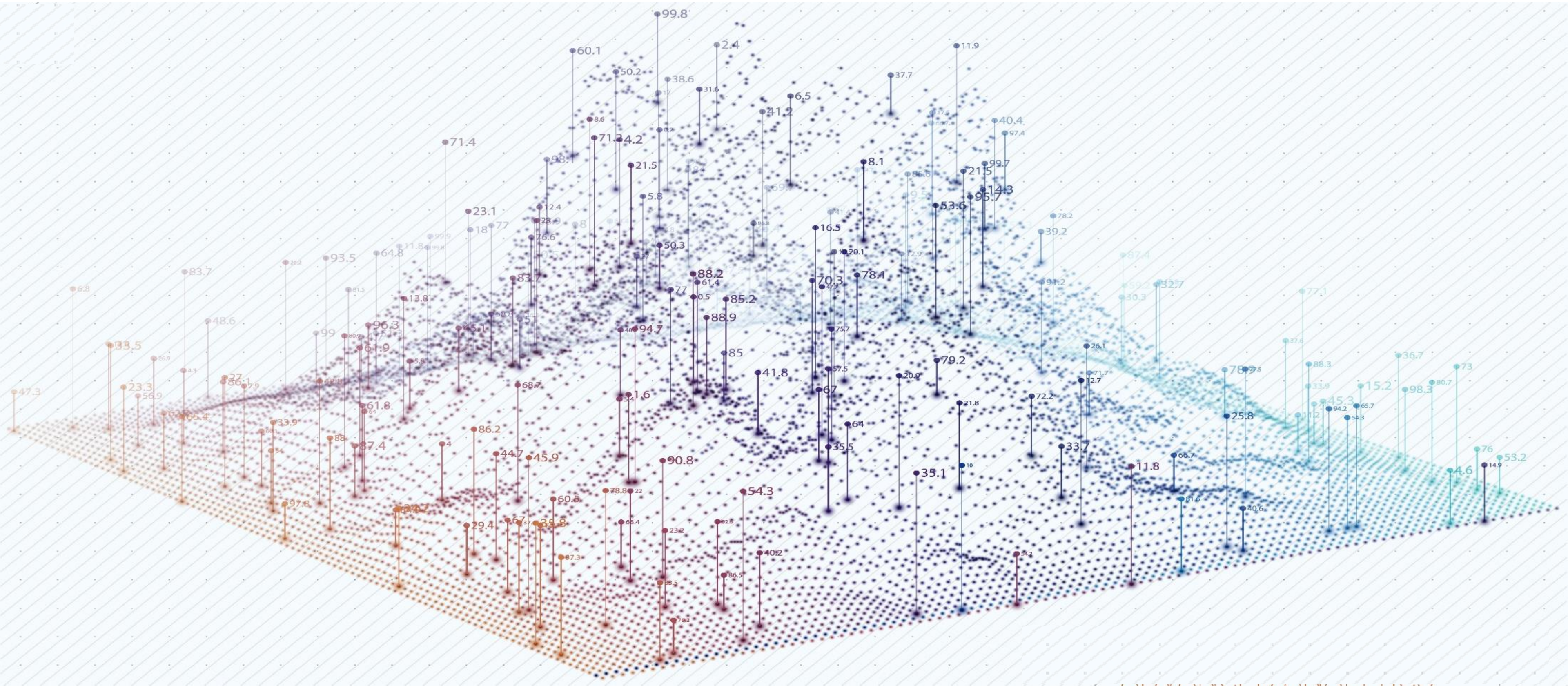
One of our founders here at Fresh likes to remind us that while he was doing his post-grad in computer science in California, Neural Networks were all the rage. That was 20+ years ago. They are back in fashion again today… and still not very good *on their own* at million-line product & service classification. A modern LLM vocabulary of 50k+ tokens in 10k+ dimensional space, however, can be very useful; especially for enrichment workflow, but billions of clean feature embeddings do need practical curation for the application layer. LLMs alone are not currently out-of-the-box optimized for few texts and many classes (whether spend, tariff or technical), and massive Multiclass classification is non-trivial, exhibiting real heterogeneity of decision boundaries. The state of the art is instead now a bespoke hybrid model with pre-tested workflows, some RL (reinforcement learning) & a little dash of HITL (you).

2. You need lots of CLEAN training data to train an AI model

Machine Learning algorithms, especially Large Language Models, require lots of training data to accurately predict the next item's correct class. If you want to stand out from the crowd and beat your competitors, you'll need even more truly clean, relevant training data. But to actually get this real-world data takes many years and many successful cleansing projects. Data should also cover all technical noun-modifier pairs or e.g. 80 000 diverse UNSPSC or HTS's, with each class well-represented to prevent overfitting. This means tens of millions of CLEAN, relevant & preferably unpublished data lines – hard to get without a head start.

Refresh™ is the world's leading product auto-classifier and we've collected only CLEAN data for 10 years.

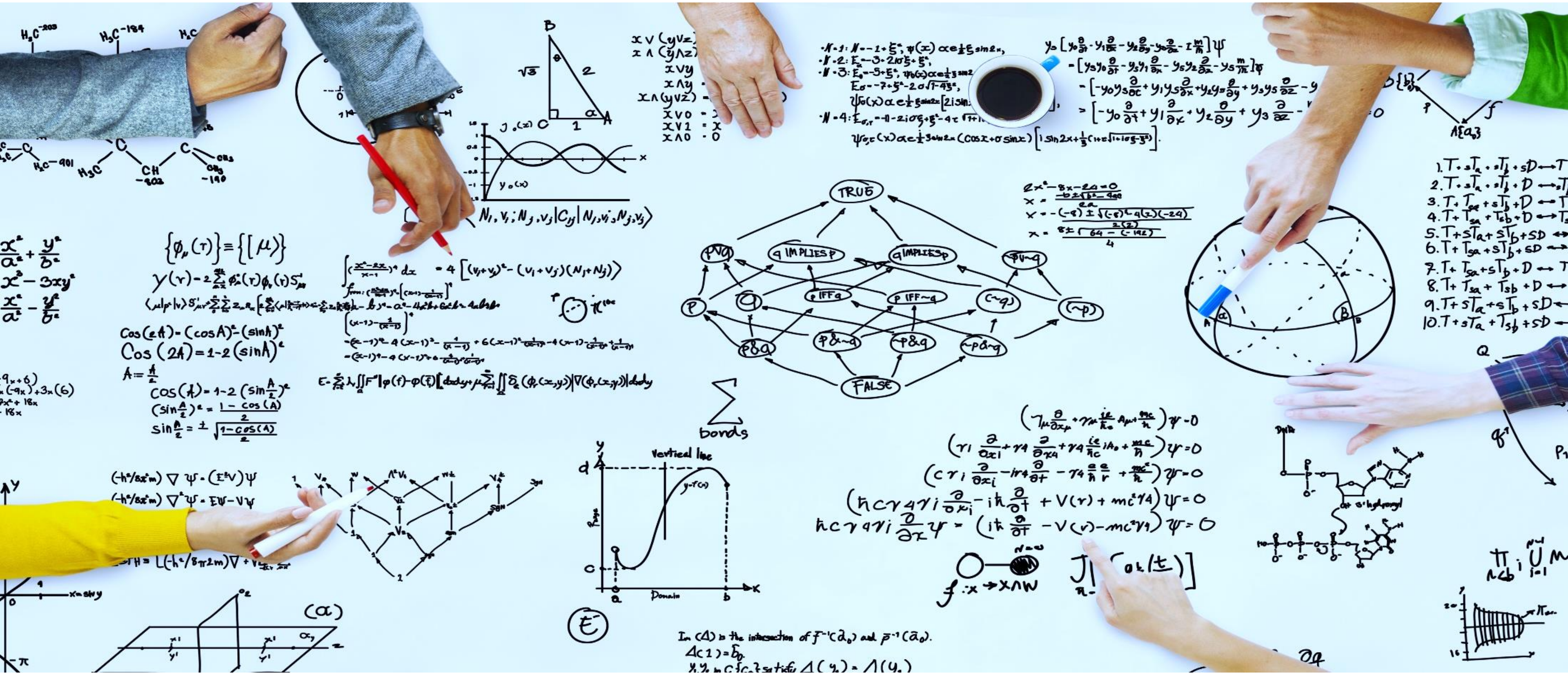
3. Building a bespoke machine learning algorithm takes time



The success of any enterprise AI application is actually heavily skewed towards the ingredients and not the methods of combining them in a recipe: knowing the problem, asking the right questions, thoroughly understanding the data, and combining results from your test algorithms with expert knowledge are all vital. Finding better features in preprocessing also brings massive gains, as does improving the training set. All of these things take time, trial and error feedback, updates, deep testing, and knowledge transfer.

We’ve been preparing our LLM-enabled algorithms for 10 years now & we still see fine-tuning gains today.

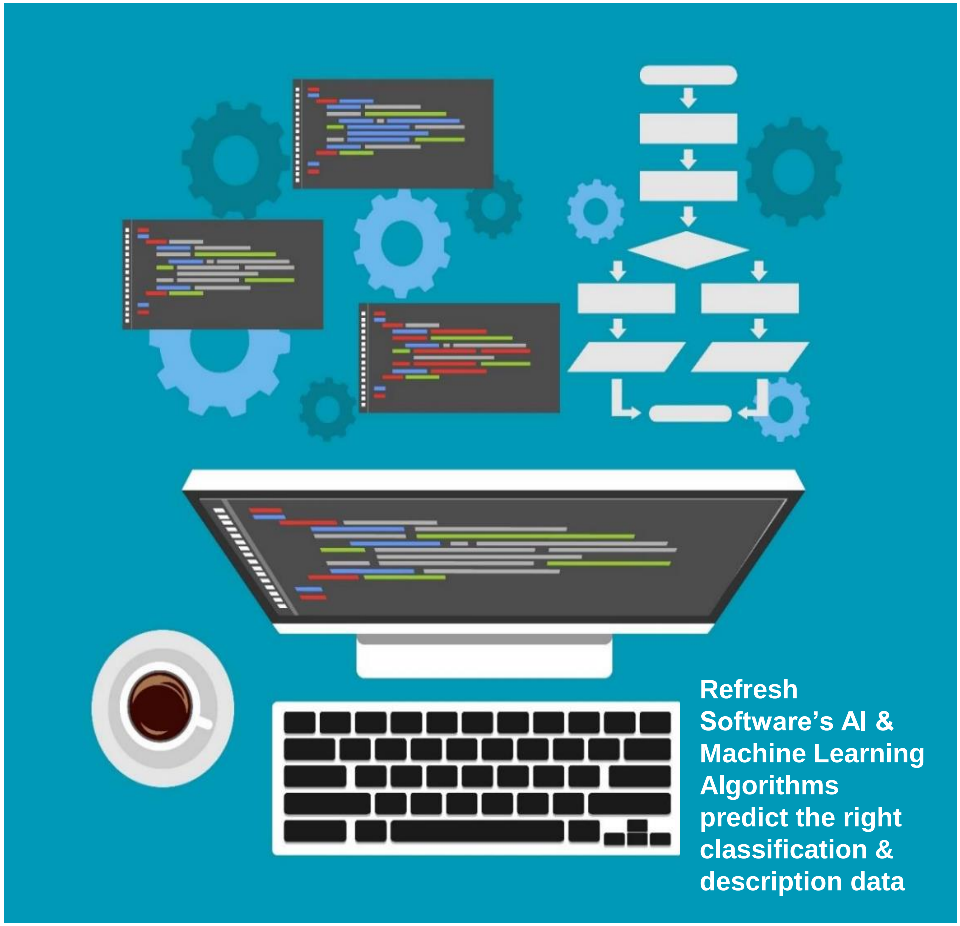
Building vs. Buying your AI System



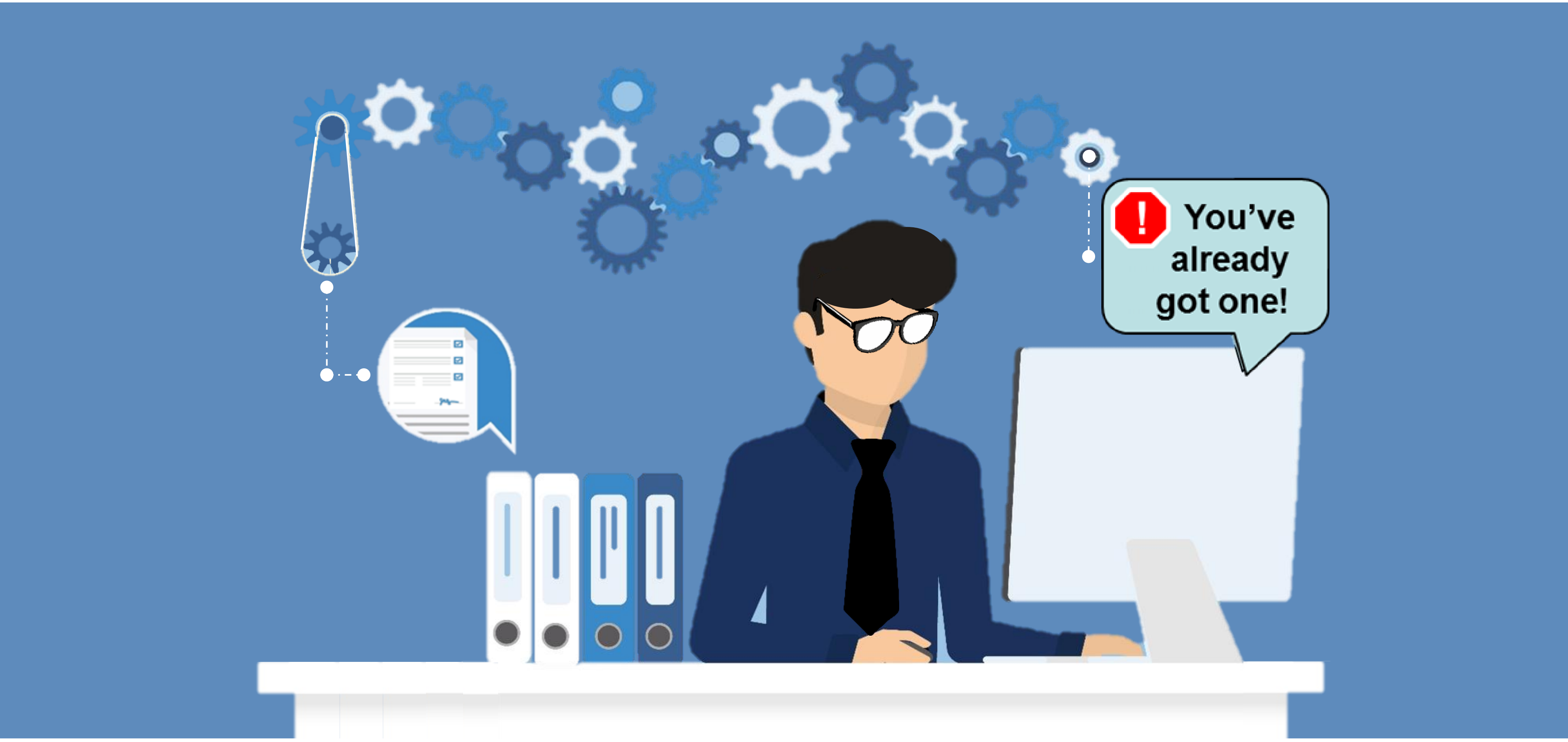
There are two massive challenges companies face in DIY Machine Learning: data and talent. To build good AI, companies first need lots of their own relevant and clean training data, but maybe you haven't had a data culture for very long, maybe you have grown by acquisition, or maybe you just inherited a little mess. Even if you did have lots of perfect training data, you would still need talented data scientists, deep domain expertise from the business users, and then some years of development, testing, workflow and integration. All of this is key if you are to avoid the trap of building just another underwhelming LLM prompt wrapper.

So using a tool like our Refresh Desktop classifier would effectively help get you to your goals faster…

Just like Refresh™ helps these customers today:



Our algorithms, built over a decade, have learned from classifying tens of millions of real customer items



This automation gives you real-world savings from day 1

“an overnight success is ten years in the making”

> 10 years ago

Joined forces and founded Fresh as a research org, submitting our 1st patent titled: “Automatic Material Data Standardization Tool”

Founder 1: Physics grad with computer science graduate studies at CalPoly on NNets

Founder 2: Engineering grad with final year thesis on fuzzy logic software for ‘self-driving’ car

2015

OG Refresh Application is 80% traditional AI. 1st Million of our paying customer’s materials cleansed

2018

1st Acceleration from Machine Learning. >10 models proto. in Weka incl. kNN, SVM & ffNN. Re-coded something like J48 / C4.5 in Python & Java to fit our use-case

Many customers, lots of new data

S-curve step change

2019

Large gains in pre-processing

2020

1st customer using the “ML only” version, Canada

2021 we needed to integrate all the clean new customer data. Looked at Hoeffding trees & ffNNs but we all faced out of memory errors & long runtimes. A costly NVIDIA CUDA migration wasn’t an option

2022

2nd Acceleration from Machine Learning. LLMs at large scale are now available. This is a game changer for us

2023

By end 2023 LLMs are on-tap / a utility. This massively complimented our approach, migrating the compute bottleneck to the cloud for us, freeing up Refresh to do what it was built for (we were lucky we had this architecture). LLMs on their own needed a lot of checking / templating. But by now we had curated thousands of templates

2024

Many customers, lots of new data

Hybrid Java AI with LLM cherry pick. Thousands of workflow steps, many thousands of templates live

2025

3rd Acceleration from Machine Learning. Multi / Hybrid Model: Proprietary, LLM API and HITL. Horses for courses: LLMs on our own workflows for attribution / translate but many new break-throughs are at the human level again. Refresh HITL (GUI) is catalytic for high-spec, hard projects that previously were un-trustworthy / un-doable. Best of breed accuracy, acceptable speed, means more clean data & more delighted customers

Timeline of how we built the Refresh AI Product

Available today. Installed in an hour. 1st Results in a week.

Campbell's | GLENCORE | Coca-Cola | Government of Canada | DOW | Continental | RioTinto | University of Oxford | ExxonMobil

@materialcleanse
www.freshinternational.com/contact
accounts<at>freshinternational.com

Contact us for a free live demo.