

DIGITAL ACCELERATION IN INDUSTRIAL R&D

How to rapidly and safely deliver value from
data science and AI projects



CONTENTS

INTRODUCTION	3
1. PROVE VALUE BEFORE YOU COMMIT	4
2. IMMEDIATELY ACCESSING THE RIGHT DATA	7
3. THE RIGHT TYPE OF INTELLIGENCE	11
4. DEPLOYING MODELS AT SCALE	14
BRINGING IT ALL TOGETHER FOR RAPID RESULTS	16

INTRODUCTION

Digital R&D is advancing rapidly, creating lucrative opportunities for innovation, by speeding up the time to get new innovations to market and reducing wasted effort and cost.

Data science and AI are the jewels in the crown of Digital R&D, making it possible to spot unseen research opportunities, be more market driven, predict success or failure early, automate arduous processes, find new insights and evidence in small data sets, and optimize product development.

Getting this right is not just about what is technically possible. It is about being able to use these tools in ways that deliver tangible value to R&D, in timeframes that make it worthwhile.

Data science and AI are complex tools that must be carefully integrated across different areas of R&D, and carefully aligned to the context in which they operate. Thought must be given to the whole implementation process from data gathering, to model selection, to user experience.

As Digital R&D accelerates, data science and AI will have an ever-greater role in doing R&D at speed, but must not compromise accuracy. Meanwhile, expectations of these tools will become increasingly complex, as the low hanging

fruit of process automation gives way to more complex and nuanced uses of AI to predict product formulation outcomes. Data science and AI will be called upon to do ever more complex tasks, with ever less certain data.

Organizations will find themselves with a constantly evolving portfolio of data science projects, which move them towards Digital R&D maturity, hopefully with many successes on the way.

Those building a portfolio of data science R&D projects need people, processes, and technology to create robust models at speed and scale. They also need strategic approaches to assessing where AI can add value, which project to prioritize, and when to make changes or halt projects.

Success combines business strategy, project management, data engineering, data pipelining, building and validating models, software engineering, and user support, all of which must work together cohesively.

This whitepaper draws on a wide range of data and AI projects - across CPG, chemicals, agri-business and manufacturing industries to explore what factors deliver success.





1. PROVE VALUE BEFORE YOU COMMIT

How to decide which data projects to take forward


Delivering rapid value from Digital R&D will involve a portfolio approach, identifying a range of projects that can deliver overarching business goals, and pursuing them in parallel. Each stage of the R&D process may look at the range of places where data science can add value, from modeling product formulations to predicting demand.

The temptation may be to rush in with bold ideas and start building proof of concept. But before that, we should do a Proof of Value exercise.

Proof of Value looks at your planned data projects, or models in development, and asks: 'Is this possible with the existing data? And if we built it, would it be useful?'

This allows you to quickly identify which workstreams to progress now, which need more work to capture useful data, and which will cost more to deliver than the value they create. The process may also identify opportunities that were not considered, which can be added to the portfolio.

A proof of concept might say 'We want to use historic vessel location data to recommend which protective anti-fouling coatings offer the best performance and value. How can we design such a system?'. Proof of Value would first ask: 'Does our existing data contain the right information to make these predictions with the accuracy we (and our customers) need?'



THE RIGHT TECHNIQUE FOR THE JOB: MACHINE LEARNING TO IDENTIFY NEW PRODUCT FORMULATIONS WITH IMPROVED SHELF-LIFE

Capgemini Engineering developed a powerful predictive model for product shelf life on behalf of a major chemical firm.

We started using techniques such as principal component analysis and correlation testing to identify the predictive power of the data. This showed that 20 key ingredients had the most significant impact on shelf life and, more importantly, that the effect was highly non-linear. Based on this, a neural network was selected as the best approach.

After training the initial neural network, we observed systematic bias in the predictions. This was traced to the way that formulations scientists had traditionally approached the research: by generally trying safer combinations of ingredients in their testing. To counter this, we interviewed the domain experts to codify their understanding into a supplementary knowledge-based neural network. The adapted neural network had significantly greater predictive power, and as a supplementary benefit, the codification of the domain knowledge ensured better sharing and retention of expertise throughout the business.

The knowledge based neural network is now fully adopted in the business, reducing product waste by over 50% and accelerating formulation testing by over 20%.

How to Start a Proof of Value Initiative

To run a Proof of Value, we advocate starting with “**Art of the Possible**” workshops.

These look at a range of planned use cases that could deliver business value, and discuss the potential for new ones, including looking at successes in other organizations for inspiration. For each, they ask what the use case is trying to achieve and what data is available.

A portfolio of the most promising use cases should then be prioritized and investigated by data scientists, who use available data to explore possibilities against the intended business goals, and visualize insight. This allows a quick assessment of whether the proposed value can be delivered.

For example, in one project a client had lots of messy data they were struggling to get value from. We identified a sample data set to test for value, and set up one of our bioinformaticians to manually analyze the data and reach conclusions in the way a machine learning system would.

This allowed us to understand what insight could be gained from that data, and what its limitations were – including identifying some interesting associations that were not known to the client. This identified approaches we were confident would work and could be built out.

Valuable Data Science and AI Projects Start With Proof of Value

Proof of Value allows R&D to prioritize the most viable data and AI projects, and plan routes to deliver them before any serious investment is made. Such mapping not only derisks projects, but underpins clear, evidenced business cases that will secure organizational buy in for the project.

Once you have identified viable projects, you can move on to building the proof of concept for each. This will be covered in the following two sections, before we look at how to turn these into productionized usable products in Section 4.



“ Proof of Value allows a quick assessment of whether the proposed value can be delivered.”



2. IMMEDIATELY ACCESSING THE RIGHT DATA

How to ensure the right data is generated, prepared, controlled and accessible?

Good data is the foundation of any model. Any model – whether identifying candidate molecules, effects of temperature changes on food chemistry, or the impact of supply chain changes – needs to be trained on accurate, representative data that represents what is being modeled.

Even if a model is perfect, it will still produce a wrong result if the data going in is incorrect or incomplete. Garbage in, garbage out.

Accessing that data is a pain point for many modelers. Data is often in different formats, different locations or labeled according to different systems. Some will be subjectively captured and may reflect human biases. Such data may require significant work before it can be used for modeling.

If errors in data are missed, they will cause problems down the line, leading to sub-optimal or incorrect model outputs.

Getting Your Data in Order

Good data is FAIR (Findable, Accessible, Interoperable, Reusable). It is stored in a way that makes it easy to identify by anyone who searches for it. It is in formats that can be read by humans and machines. And it is clear about any limitations or rules about how it can be used.

The following four principles, inspired by Capgemini Engineering's Data Management Maturity Model, should be followed to ensure an organization's data can be effectively used for modeling.



Whether identifying candidate molecules, effects of temperature changes on food chemistry, or the impact of supply chain changes – needs to be trained on accurate, representative data that represents what is being modeled.

1. Data Must Be of Sufficient Quality for Modelers

Data must come from a trusted source. This may be simple for your own chemical analysis data, but will be more complicated for open source data, or third-party data from panel tests which may include bias or misreporting. It will be particularly challenging for public data.

A common problem is very flat data, ie data that has relatively few products but hundreds of different attributes. This leaves the data open to overfitting where it looks like the predictive power is increasing but, in fact, the model is becoming more and more fragile and will fail when applied to new data.

Data scientists need to ensure there is sufficient data, correct any missing or confounding elements, and work with domain experts to review and modify data, so that it accurately represents the things it is measuring in the real world.

2. Use Metadata to Make Data Searchable and Understandable

Metadata should be added to enhance understanding and usability. This will include descriptions of what the data represents – eg type of molecule, but also provenance, timestamps, etc. There must be a consistent taxonomy for naming things.

Good metadata allows different groups with different interests to find it easily in the system, and allow those reading it – including machines – to make sense of it and easily compare it to other data.

3. Consider Privacy and Security to Avoid Problems Down the Line

If models are trained on data that doesn't meet privacy rules, it could cause big problems down the line. Its provenance and allowable use should be made clear in the metadata. It must also have adequate security in place to protect it, where it is stored and used.

4. Make Data Consistent, Accessible, and Traceable

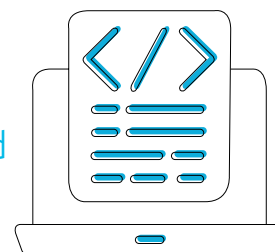
Data stores, lakes and warehouses need to be set up so that data is accessible to anyone who needs it, whilst restricted to those who don't. This also includes selecting tools and building integrators that would pipe data to data science teams.

Data must have a single source of truth. It must be linked together in the IT system, so that if one instance is changed, all others are updated.

Finally, all data must have a data steward, someone who makes decisions about how it is stored and managed, and someone who can be contacted by modelers who need further information.



Metadata should be added to enhance understanding and usability.”





CASE STUDY: THE VALUE OF DATA MANAGEMENT

We helped a large packaged goods company explore how R&D data could be used for in silico product design.

After speaking to the modelers, it became clear that they had the right data science skills, but the data was hard to find, hard to understand, laborious to use, and sometimes risky to draw conclusions from. Data lacked the right tagging and metadata to allow it to be linked across the R&D development process.

By improving their data management, they were able to improve their models and start to chain these models together. Getting the data right means better answers early on, and reduced risk of failure.

RAPIDE: A PROFESSIONAL GOVERNANCE FRAMEWORK FOR DATA SCIENCE PROJECTS

Capgemini Engineering's RAPIDE framework guides organizations through data science projects - from data selection to model development to productionization, with checks at key stages to ensure projects only progress when they are ready to do so.

i. Readiness Assessment

Assess what data you need and what is available. Understand the type of analytics problem: Is it classification/regression, supervised/unsupervised, predictive, root-cause analysis, statistical, physics-based? Understand how "dynamic" the problem is – ie will the nature of the incoming data change over time, necessitating periodic retraining? The Proof of Value exercise in Section 1 will help guide this first stage and confirm the project is worth taking forward.

ii-iii. Advanced Data Screening and Pinpointing Variables

Explore the data using a range of simple techniques to spot meaningful correlations between events of interest. For example, do product characteristics, such as tensile strength, correlate with a change in the extrusion process? Identify constraints in the data that might limit model choice; such as overly broad data that might obscure variables that dictate behavior. Early insights help direct your model to be most effective.

iv. Identify Candidate Algorithms

Based on outputs from the previous analysis, identify candidate modeling techniques (which could be empirical, physical, stochastic, or hybrid). Shortlist the most promising candidate algorithms and quickly assess the feasibility of each.

v. Develop Powerful Models

Decide on the most suitable model for the problem. Check the implementation requirements, such as user interface, required processing speed, architecture, etc to ensure it will be a usable solution before you commit. Gather validation data. Build it.

vi. Evolve and Embed

Embed the solution into the relevant business unit and refine using data gained from in-service use.

If these steps are carried out correctly, no model should fail after deployment.

3. THE RIGHT TYPE OF INTELLIGENCE

Selecting the most effective tools and techniques to get the answers you need

Once you're happy you have the right data, it's time to build models that work.

There is no rule for which approach is best for a particular problem. The nature and context of the problem, data quality and quantity, computing power needs, and intended use, all feed into model choice and design.

Techniques such as machine learning and neural networks can be very powerful where there is lots of well curated data. For example, developing accurate predictions of product performance, trained from decades of historical performance test data, can be used to develop and market high-value products by calculating the true total cost of ownership.

However, 'most powerful' is not the same as 'most suitable'.

If the problem is new and there is not much proven data available, this may limit your approach to well understood modeling techniques, such as cluster analysis, principle component analysis, or Bayesian uncertainty quantification.

Equally, there is no need to build complex powerful machine learning models, if the problem can be solved just as well with much simpler statistical approaches. A simple or naive model, based on a physical understanding of the product's properties may be able to demonstrate how product performance degrades over time. This may be perfectly good enough for some needs, such as redirecting research focus.

Building any particular model requires someone with the right skillset for that model. But the real challenge is knowing which model is best to use. Mistakes are often made when decisions are made based on what modeling skills are available, rather than what is best for the problem. The best decisions happen when organizations involve a range of data science experts, who can assess the best tools, based on extensive experience of similar problems. It is also important to be prepared to pivot your approach early in the process, to make sure you get the right final model for your data.



Mistakes are often made when decisions are made based on what modeling skills are available, rather than what is best for the problem.”



PROOF OF VALUE IN ACTION

Capgemini Engineering partnered with a major paint manufacturer to investigate whether their data could deliver new insights with real business value. Our 'Art of the Possible' workshop identified which business cases had the most value, which were more likely to succeed, and which could draw on a wide range of the available data sources.

One of the ideas chosen for investigation combined internal and market data sources to get insight into product positioning. Capgemini Engineering demonstrated that the data was good enough to show the expected macroscopic trends. The investigation also identified key data quality and reliability issues that should be resolved before productionizing the models into the client's decision making processes.

Before launching into a full proof of concept, we proved the value of the existing data, by using it to deliver working models that could generate business insights, along with detailed, concrete next steps.

BUILDING TRUST INTO AI

A trusted model is one that people are happy to use. It gives results that users understand and accept, are easy to use, and that don't raise privacy, legal or ethical issues. If it falls down here – even if the model is perfect - the user will not trust the results. This can be resolved through good practice in model development in five areas:

- 1. Assured:** Trusted AIs must use a well-designed model, and be trained and tested on data that is proven to be accurate, complete, from trusted sources, and free from bias.
- 2. Explainable:** A recommendation is much more useful if you understand how and why it was made. A good AI will have tools to analyze what data was used, its provenance, and how the model weighted different inputs, then report on that conclusion in clear language appropriate to the users' expertise.
- 3. Human:** An intuitive interface and easy-to-understand decisions help the user to trust AI over time. The complexity of the interface needs to be suited to the user's knowledge; a customer facing smartphone app will look very different from an automated safety assessment platform.
- 4. Legal and Ethical:** A trusted AI should reach decisions that are fair and impartial, meeting data protection regulations and giving privacy and ethical concerns equal weight to predictive power.
- 5. Performant:** A trusted AI continues to work after deployment. A performant AI considers future throughput of data, accuracy, robustness, and security.



4. DEPLOYING MODELS AT SCALE

Deliver applications that are robust and resilient enough to withstand real-world use

For a model to be successful, it must work and scale in the real world. The user is presented with a clear interface. They enter the relevant parameter – which may be the desired sensory characteristics, or optimal sustainability properties. The software runs, collects data from backend IT systems, executes the model, and presents the resulting insight to the user.

In most cases, this involves wrapping the model into a piece of software and integrating it into either a web or phone app, or a piece of technology such as a high throughput experimentation machine.

This is where a lot of data science projects fall down. Those building the models do not always appreciate the rules and complexities of enterprise IT or edge computing, where the model must operate. There is often a mismatch in expectations and language between the domain, modeling and IT functions. Software engineers who understand both sides need to be able to bridge this gap.

Integrating Models Into Real World Systems

Models built by data scientists often use languages not familiar to the enterprise, such as Python or R.

In some cases this can be overcome by requiring data science teams to build models in cloud environments, such as Azure and AWS, which are set up to reflect the enterprise's infrastructure and provide common toolkits which easily integrate.

However, complex models may need more sophisticated data science tools and programming languages, leaving them in a format which doesn't naturally integrate. The solution is usually 'containerization'; wrapping models in software ('containers') which translate incoming and outgoing data into a common format. The model then runs in isolation in the container, but slots into the wider IT ecosystem.

The growing field of

AI DEVOPS

can help ensure data science, IT and software teams can work together effectively towards shared goals



Models vary in power and compute demands. A model used to discover naturally occurring molecules may process petabytes of data from libraries once per month, whilst an insights system may process a continuous stream of big data from thousands of IoT devices. Compute needs to be allocated correctly, or it will slow down deployment and could alienate early users. Data security and regulatory compliance must also be considered. Those building the software which wraps around the models need to consider all these

issues. The growing field of AI DevOps can help ensure data science, IT and software teams work together effectively towards shared goals.

Slotting the software into the IT systems is not the end of the story. Models need ongoing retraining, maintenance and support to ensure they keep working and improving. This is often specific to the model, and thus requires support teams to be set up, which include people with expertise in that model or data.

BRINGING IT ALL TOGETHER FOR RAPID RESULTS

From business challenges, to data selection, to modeling, to scale up and use

Bringing this together requires a range of skills, deployed effectively across the organization.

Initial planning should bring together strategists, data scientists, IT teams, and domain experts. It should start by exploring whether proposed projects could deliver value, and whether the company has sufficient quality data to deliver that value.

Once a portfolio of projects has been identified and confirmed to be worth pursuing, a team should be established to guide projects and assess them at key stages to ensure they are delivering against business goals. This team should also align these projects to an agile governance framework, like RAPIDE.

For each project, data scientists need to work with the domain experts who will ultimately use the models, to understand what they need the models to do. The data scientists should assess the data available, feed back on what is possible, and adjust plans. Once agreed, they must work with data engineers to select and clean the data.

The project may need access to a pool of modelers who can bring different skills to bear on different problems, from statistical techniques to machine learning. Finally, software engineers will be needed to productionize the model.

Such complex, multi-skilled work benefits hugely from the involvement of 'translators', people who speak the language of the domain, data

science and the business units, who can communicate across the different teams and ensure the right skills are selected, and deployed at the right times - and that projects remain aligned to objectives.

With the Right Skills, Data and AI Projects Should Be Right First Time

Rapidly progressing valuable data science requires strategic planning and allocating the right skills at the right time.

Often, domain experts who understand data, rather than data experts, will be left to build models. They may be able to do it, but it is a poor allocation of resources, resulting in slower progress and high failure rates. Domain experts will often spend 80% of their time turning data into something useful, eg a process engineer producing a process engineering model, and just 20% of their time using that model to design processes for new products - where their true expertise lies.

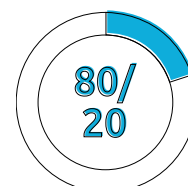
This is a manifestation of the Pareto principle, or 80/20 rule, which says 80% of the value comes from 20% of the work, and vice versa. Value can be realized much quicker if the data is taken off the hands of the 'domain experts who understand data', and given to 'data experts who understand the domain'. This speeds up the data work, and frees up domain experts to focus on what they do best.

Speed isn't about cutting corners, it is about doing things right as quickly as possible, so you get earlier results and don't need to repeat, correct, or abandon work.

That means efficient allocation of resources - the right skills for the right job. Getting data experts to handle the data, modelers to do the models, and software engineers to do the software, with someone in the middle managing it all. Critically, this means freeing up domain experts to focus on where their true expertise lies - understanding your customers, product characteristics and the development of the technology platform.

Getting all these moving parts to work effectively together is the quickest way to develop data science and AI projects that deliver value to the business, and which work first time.

Strategic planning and allocating the right skills at the right time is a manifestation of the Pareto principle, or 80/20 rule, which says 80% of the value comes from 20% of the work, and vice versa



ABOUT THE AUTHORS



Matt Jones
Head of Offer Development



Mark Knight
Head of Sales



David Hughes
Head of Technical Presales



James Downing
Head of Offer Presales



Sam Genway
Emerging Technologist Lead

Contact
hybridintelligence.coe.global@capgemini.com



John Godfree
Head of Consulting



James Hinchliffe
Senior Consultant

About Capgemini Engineering

World leader in engineering and R&D services, Capgemini Engineering combines its broad industry knowledge and cutting-edge technologies in digital and software to support the convergence of the physical and digital worlds. Coupled with the capabilities of the rest of the Group, it helps clients to accelerate their journey towards Intelligent Industry. Capgemini Engineering has more than 55,000 engineer and scientist team members in over 30 countries across sectors including Aeronautics, Space, Defense, Naval, Automotive, Rail, Infrastructure & Transportation, Energy, Utilities & Chemicals, Life Sciences, Communications, Semiconductor & Electronics, Industrial & Consumer, Software & Internet.

Capgemini Engineering is an integral part of the Capgemini Group, a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided every day by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of over 340,000 team members in more than 50 countries. With its strong 55-year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fueled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2022 global revenues of €22 billion.

For more information please visit:

www.capgemini.com

Contact us at:

engineering@capgemini.com