LIFE SCIENCES - HOW TO DELIVER RAPID VALUE FROM YOUR AI PROJECTS

CONTENTS

Digital R&D is advancing rapidly, creating lucrative opportunities for therapeutics and diagnostics innovation, with shortened R&D timeframes. During the COVID-19 pandemic we saw that quality R&D at speed is possible, thanks in large part to data and digital technology.

Data science and AI are the jewels in the crown of Digital R&D, bringing possibilities to spot unseen research opportunities, predict success or failure early, automate arduous processes, find new insights in small data sets, and optimise clinical trials.

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 which 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.

> 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 quality R&D at speed. 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 chemical, biological, and epidemiological outcomes. They 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.

> THEY ALSO NEED STRATEGIC APPROACHES TO ASSESSING WHERE AI CAN ADD VALUE

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 prioritise, 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 in life sciences to explore what factors deliver success.

About Tessella

At Tessella, we have 40 years' experience helping the world's leading life sciences organisations make sense of complex data. We have a strong understanding of the data techniques themselves, and how to apply them to health and scientific data.

Our data science and AI experience includes:

- Validating a rapid diagnostic technique by showing what PCR amplification curves say about prevalence of pathogens
- Working with a pharma company to develop an AI that identifies candidate drug molecules
- Helping a pharma company identify and enrol targeted groups for clinical trials
- Developing text analysis tools that can use social media and online reports to improve disease spread models

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 that data science can add value, from predicting success rates to optimising clinical trials.

EACH STAGE OF THE R&D PROCESS MAY LOOK AT THE RANGE OF PLACES THAT DATA SCIENCE CAN ADD VALUE

The temptation may be to rush in with bold ideas and start building proof of concepts. 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 our past clinical trial data to predict late stage failures, how can we design such a system?' Proof of Value would first ask: 'Does our trial data contain the right information to predict future failures with the accuracy we need?'

Proof of Value in Action: Predicting Smokers' Habits

Proof of Value is a business and data science led initiative which involves working with a simple extract of existing data, to rapidly explore possibilities.

For example, a pharma company wanted to understand craving levels of smokers, to develop products to help them quit.

In this case, they already had a machine learning model they wished to validate. Instead of just checking the model, we instead carried out a Proof of Value exercise. We found the data itself was adequate, but by looking at it in different ways, we showed that the end goal could be achieved just as effectively with much simpler classical statistics methods. This allowed them to focus on a cheaper and more transparent approach, and saved them investing in a product that would not have delivered maximum value.

The Proof of Value allowed us to identify whether the project was worth pursuing and explore options that delivered greater cost-benefit. Had we just said, 'let's make the existing model as accurate as possible', we would have missed an opportunity to find a better way of doing things, and leaped into an unnecessarily complicated and costly project.

Another very common outcomes of a Proof of Value is to identify holes in data sets that will prevent value being realised. At the start of this process, the client often assures us there are several data sets we could use as the basis for a Proof of Value exercise. But once they start gathering them for us, many either can't be found, or have fundamental issues like missing pages. Flaws in data like this are probably the most common reason that data projects become derailed. Spotting this early through a Proof of Value means the company can understand how to get data ready before entering full-on project territory with expectations and deadlines.

FLAWS IN DATA LIKE THIS ARE PROBABLY THE MOST COMMON REASON THAT DATA PROJECTS BECOME DERAILED

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 organisations for inspiration. For each they ask what they are trying to achieve and what data is available.

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

THIS ALLOWS QUICK ASSESSMENT OF WHETHER 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 setup one of our bioinformaticians to manually analyse 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 organisational buy in for the project.

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

ONCE YOU HAVE IDENTIFIED VIABLE PROJECTS, YOU CAN MOVE ON TO BUILDING THE PROOF OF CONCEPT FOR EACH

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 predicting chemical properties of a new compound, predicting the right first human dose level, or classifying features of a disease - need to be trained on data that truly represents what is being modelled and is accurate and representative.

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 modellers. Data is often in different formats, different locations or labelled according to different systems. Some will be subjectively captured and may reflect human biases – eg doctors notes. Such data may require significant work before it can be used for modelling.

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

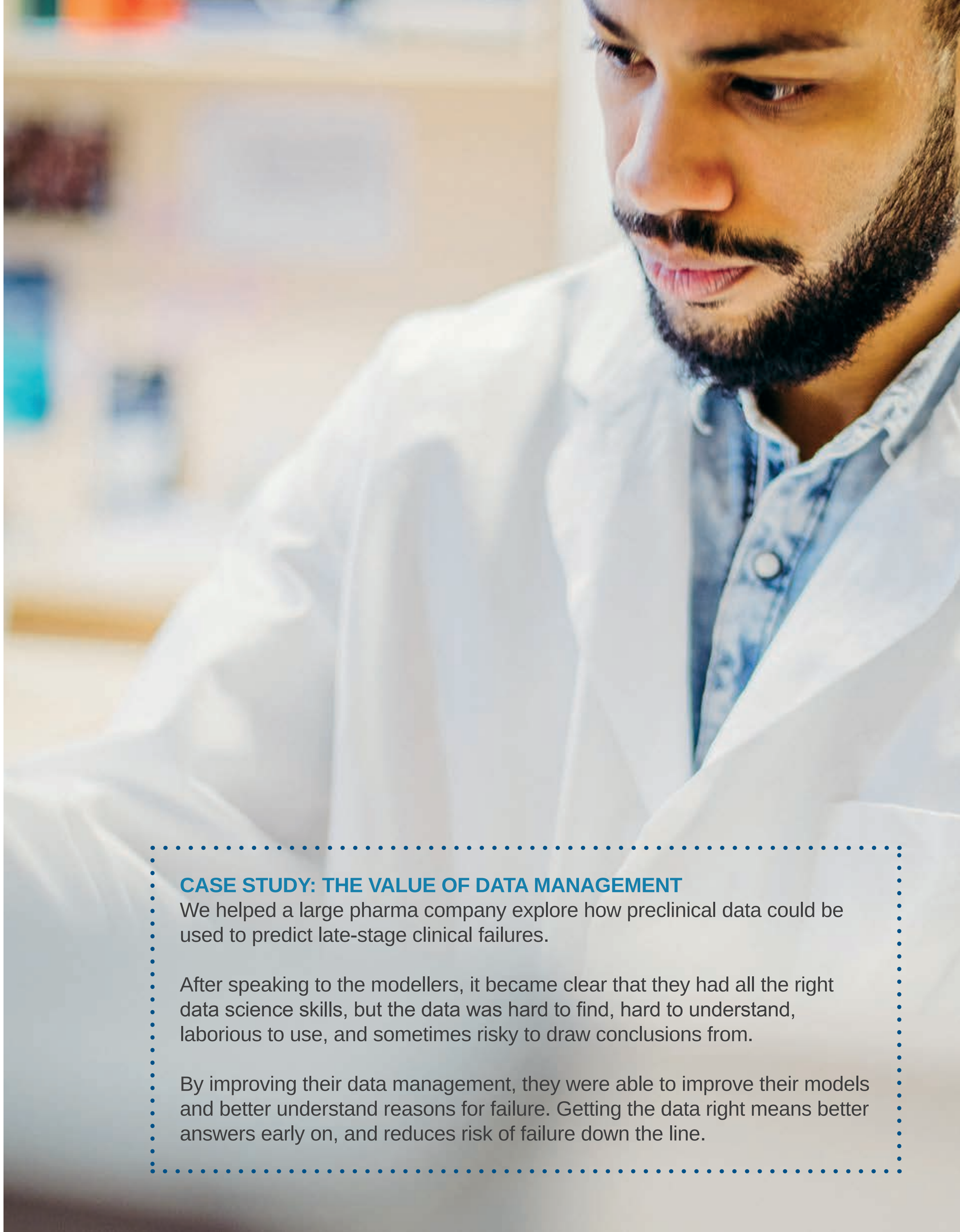
This is particularly challenging when dealing with new data – such as that on a new virus or treatment where data is coming in from poorly understood environments, or lacks standard approaches to data capture.

IN FORMATS THAT CAN BE READ BY HUMANS AND MACHINES

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 the Tessella Data Management Maturity Model should be followed to ensure an organization's data can be effectively used for modelling.



CASE STUDY: THE VALUE OF DATA MANAGEMENT

We helped a large pharma company explore how preclinical data could be used to predict late-stage clinical failures.

After speaking to the modellers, it became clear that they had all 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.

By improving their data management, they were able to improve their models and better understand reasons for failure. Getting the data right means better answers early on, and reduces risk of failure down the line.

1. Data Must Be of Sufficient Quality for Modellers

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 data from clinical trials or hospitals which may include bias or misreporting, and particularly challenging for public data.

An all-too-common problem is diagnostic AIs comes from training them with data which contains labels added by the diagnosing physician. The AI learns to spot the label, not the disease indicator.

Data scientists need to correct 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 or toxicology, but also provenance, timestamps, usage licences, 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 which 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 setup 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 the data science teams.

Data must have a single point 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 modellers who need further information.

RAPIDE: A PROFESSIONAL GOVERNANCE FRAMEWORK FOR DATA SCIENCE PROJECTS

Tessella's **RAPIDE framework** guides organisations through data science projects from data selection to model development to productionisation, 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 underlying health conditions such as obesity correlate with a reduction in efficacy of a new asthma drug. Identify constraints in the data that might limit model choice; such as overly broad data that might obscure variables that dictate behaviour. Early insights help direct your model to be most effective.

iv. Identify Candidate Algorithms

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

v. Develop Powerful Models

Decide on the most suitable model for the problem. Check 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 diagnostic tools for well understood conditions with years of historical diagnosis data which they can be trained upon.

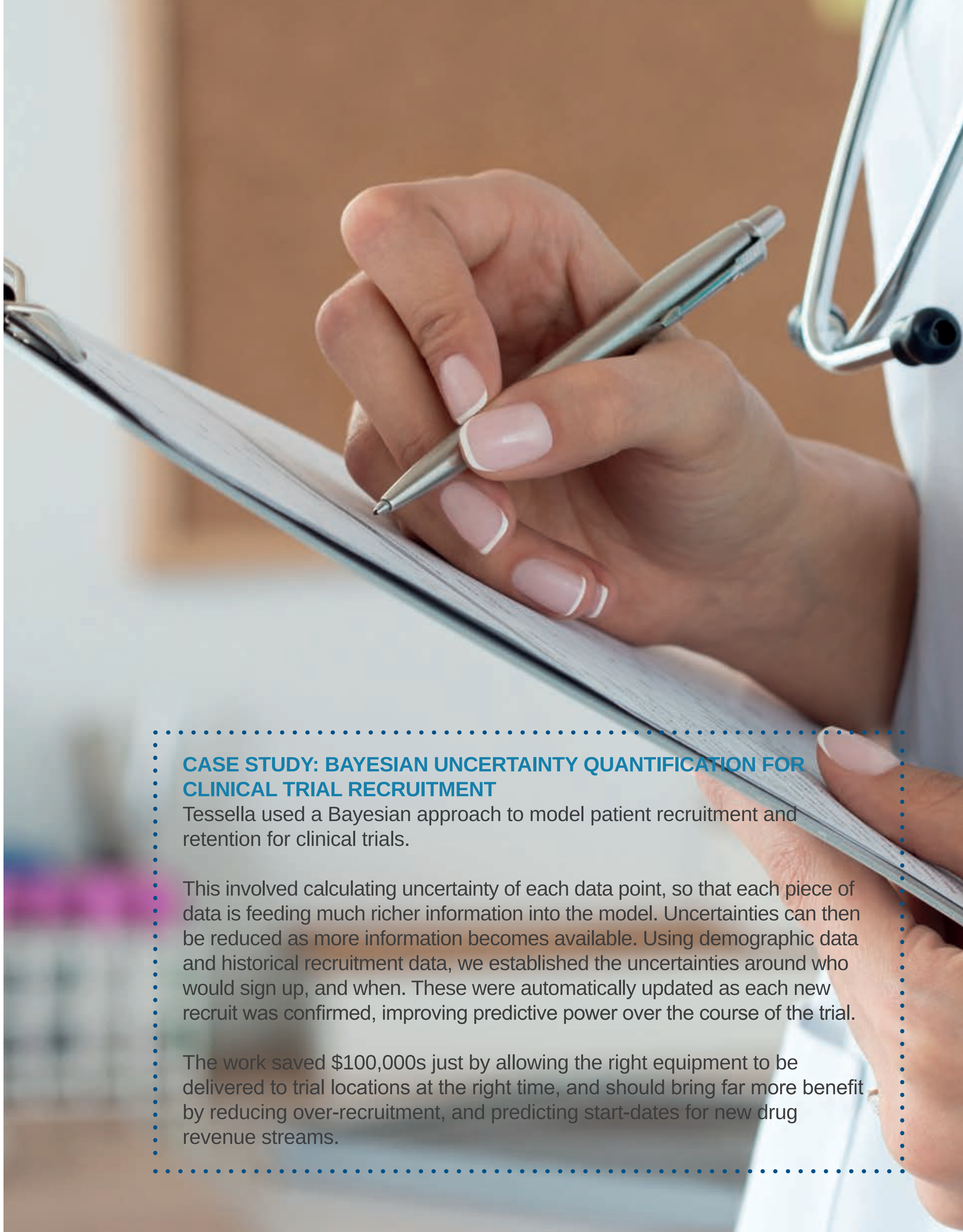
However, these will not be suitable for challenges with limited, uncertain, or changeable data. 'Most powerful' is not the same as 'must suitable'.

Real time disease spread monitoring, for example, involves constantly updated information from apps, hospitals, and potentially less curated sources, such as social media and doctors notes. Trials of new treatments involve collecting data quickly on patients responses, some of it subjective.

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

BUILDING ANY PARTICULAR MODEL REQUIRES SOMEONE WITH THE RIGHT SKILLSET FOR THAT MODEL

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 based on what modelling skills are available, rather than what is best for the problem. The best decisions happen when organisations involve a range of data science experts, who can assess the best tools based on extensive experience of similar problems.



CASE STUDY: BAYESIAN UNCERTAINTY QUANTIFICATION FOR CLINICAL TRIAL RECRUITMENT

Tessella used a Bayesian approach to model patient recruitment and retention for clinical trials.

This involved calculating uncertainty of each data point, so that each piece of data is feeding much richer information into the model. Uncertainties can then be reduced as more information becomes available. Using demographic data and historical recruitment data, we established the uncertainties around who would sign up, and when. These were automatically updated as each new recruit was confirmed, improving predictive power over the course of the trial.

The work saved $100,000s just by allowing the right equipment to be delivered to trial locations at the right time, and should bring far more benefit by reducing over-recruitment, and predicting start-dates for new drug revenue streams.

BUILDING TRUST INTO AI

A trusted model is one that people are happy to use. It gives results that users understand and accept, is easy to use, and does not 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 analyse 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 trust AI over time. The complexity of the interface needs to be suited to the user's knowledge; a smartphone diagnostics app will look very different from a drug discovery 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.

Our whitepaper on Trusted AI explains the challenges around AI and Trust and offers a detailed look at this framework.

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 should be presented with a clear interface. They enter the relevant parameter – which may be symptoms in a diagnostics app; or desired molecule properties in a drug discovery platform. 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 diagnostics 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, modelling 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 setup to reflect the enterprise's infrastructure and provide common toolkits which easily integrate.

COMPLEX MODELS MAY NEED MORE SOPHISTICATED DATA SCIENCE TOOLS AND PROGRAMMING LANGUAGES

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 translates incoming and outgoing data into a common format. The model then runs in isolation in the container but slots into the wider IT ecosystem.

Models vary in power and compute demands. A drug discovery model may process petabytes of data from libraries once per month, whilst a track and trace app may process a continuous stream of big data from millions of devices. These need to be allocated correctly or it will slow down deployment and could alienate early users. Data security and regulatory compliance around must also be considered. Those building the software which wraps around the models need to consider all these issues.

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 so needs support teams to be setup which include people with expertise in that model or data.

MODELS NEED ONGOING RETRAINING, MAINTENANCE AND SUPPORT TO ENSURE THEY KEEP WORKING AND IMPROVING

BRINGING IT ALL TOGETHER FOR RAPID RESULTS: FROM BUSINESS CHALLENGE, TO DATA SELECTION, TO MODELLING, TO SCALE UP AND USE

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

Initial planning should bring together strategists, data scientists, IT teams and domain experts to effectively plan. 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, and align them to an agile governance framework such as 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 available data and feedback 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 to access to a pool of modellers who can bring different skills to bear on different problems, from statistical technique to machine learning. Finally, software engineers will be needed to productionise the model.

SOFTWARE ENGINEERS WILL BE NEEDED TO PRODUCTIONISE 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 time and 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 drug prediction model, and just 20% of their time using that model to design drugs – 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 realised 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, modellers 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 the disease, developing drugs and diagnostics, or running clinical trials.

> THIS MEANS FREEING UP DOMAIN EXPERTS TO FOCUS ON WHERE THEIR TRUE EXPERTISE LIES

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

AUTHORS

Andrew Alderman
Sector Director - Life Science

Matt Jones
Lead AI and Data Science Strategist

David Hughes
Analytics Solutions Lead

Sam Genway
Senior AI Consultant

John Godfree
Head of Consulting

David Dungate
Senior Consultant

James Hinchliffe
Senior Consultant

Contact:
info@tessella.com

Capgemini