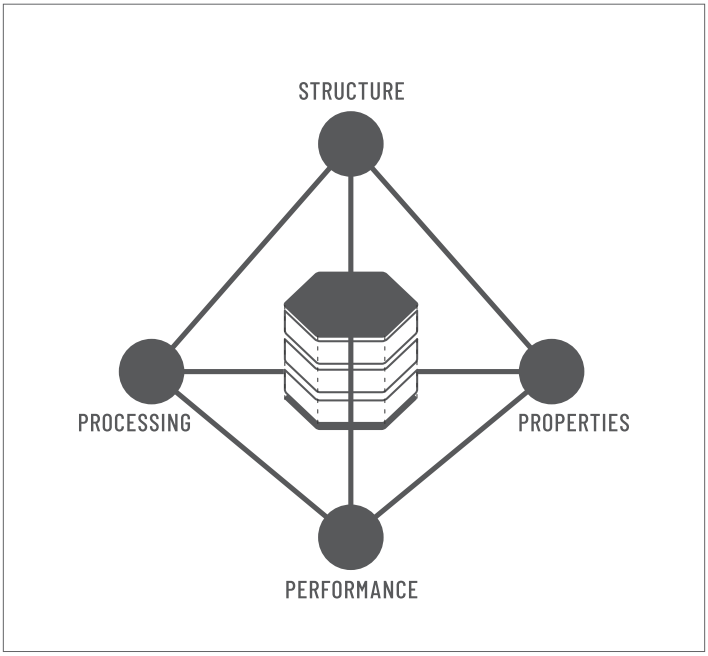
CITRINE INFORMATICS
citrine.io • info@citrine.io

UNLOCKING THE POWER OF DATA IN MATERIALS AND CHEMICAL DEVELOPMENT

THE BIG PICTURE

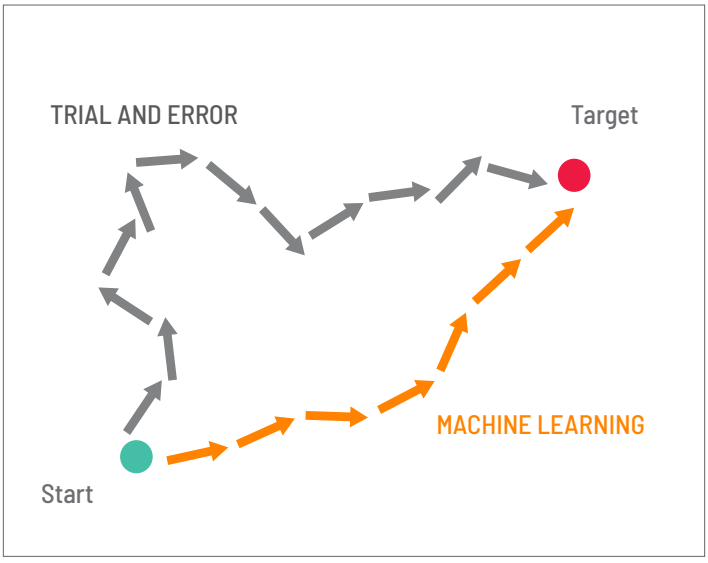
Digital transformation is occurring across industry, from the finance system to the PLM system, bringing such benefits as: enforcing best practices, better decision making, and faster processing as information is shared instantly. Materials and chemicals research and development can also benefit from digitalization and the use of data-driven tools such as machine learning algorithms. The US government's Materials Genome Initiative (MGI) highlighted the need for materials data infrastructure to accelerate innovation. A recent National Institute of Standards and Technology (NIST) MGI report attributes between $123B and $270B in economic value in the United States alone to next-generation materials innovation infrastructure[1].



HOW CAN DATA BE USED TO ACCELERATE MATERIALS INNOVATION?

OPTIMIZATION OF PROPERTIES

Good data management allows researchers to accurately analyze and understand their results as they try to optimize the properties of a new material. If simulation techniques or machine learning are being used to reduce the number of experiments needed, then data management becomes even more important, as algorithms need structured information to work.



CUSTOMER RESPONSIVENESS

When new customer requirements are expressed it is important to be able to quickly find out if a current product already matches the property profile requested and, if not, which material could be most easily and cheaply adapted to meet the demand. Materials Informatics enables companies to do this in minutes, but requires properly structured data to do so.

COST REDUCTION

Machine learning enables many more parameters to be systematically considered as materials are optimized. Rationalizing ingredients to enable bulk purchases, using cheaper ingredients to achieve the same result, or using process settings that require less energy or increase consistency of the product can all be efficiently considered using AI algorithms, so long as the algorithm can interpret the data.

MEETING REGULATIONS AND BECOMING MORE SUSTAINABLE

Whether it is meeting regulations around conflict minerals, restricted substances, or emissions and waste, data is key in identifying where certain substances and minerals are used. Data is also needed to power AI models that can optimize physical performance criteria at the same time as reducing dependence on critical materials, improving recyclability, or reducing the carbon footprint of production.

LEARNING

Both researchers and AI algorithms can learn from data. The more the better. By having a well structured, well curated, high-quality dataset focused on the materials or chemicals you produce, your team can understand how composition and processing conditions affect properties. By ensuring that your data can be reused in the future by the team that made it or other teams doing similar work at different sites, a powerful superset of data can be used to ensure accurate property predictions.

PRINCIPLES OF GOOD DATA MANAGEMENT FOR SCIENTIFIC DATA

Good data management requires:

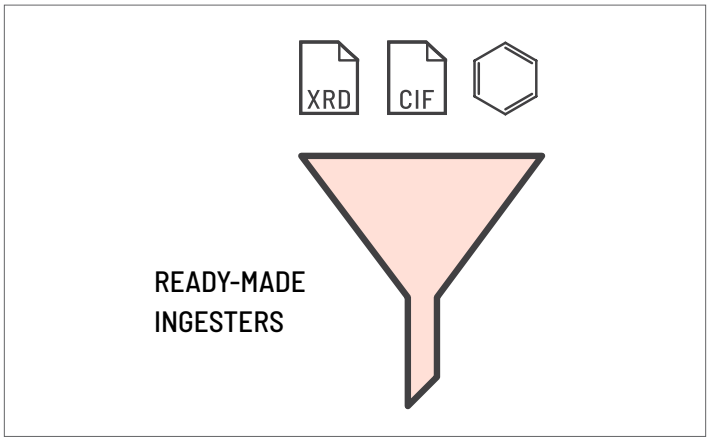
- A **scalable database** that can be quickly queried by multiple people around the globe.
- A **data model** that is flexible enough to work for different teams on different projects, but structured enough to ensure all data can be understood and used by all teams.
- A **workflow** that encourages best practice without getting in the researchers' way.
- An **intuitive user interface**, empowering researchers rather than just training them to do tasks.
- **Visualization and analysis tools** that researchers value and want to use.

Citrine Informatics has been working on this area for the last 5 years. Data is the lifeblood of both science and machine learning, and getting this right has been key to Citrine's success.

HIGHLIGHTS OF CITRINE'S DATA MANAGEMENT

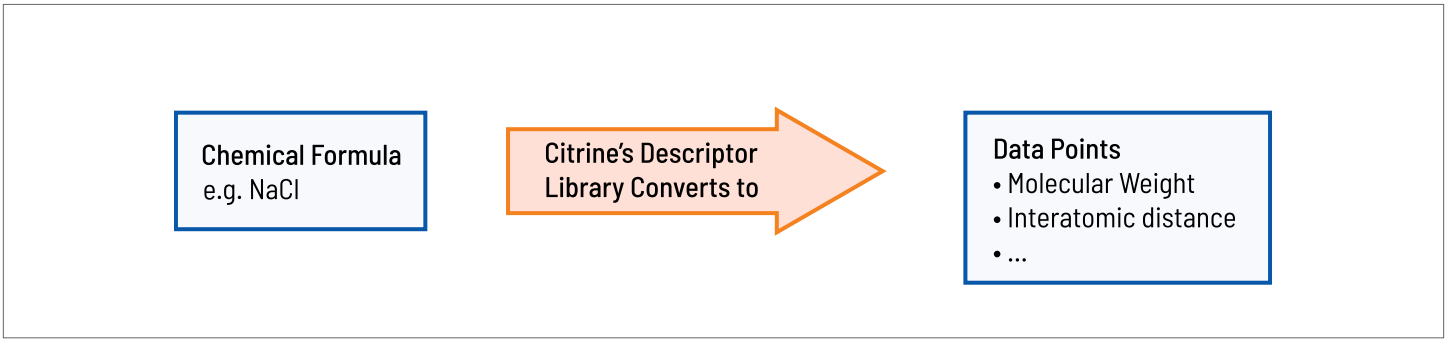
GETTING DATA IN

A library of "ingesters" is available to speed the importation of data from common instrument files such as those from x-ray diffraction equipment (.xrd) and crystal structure programs (.cif). Once in, the data is then much more valuable to researchers, as it can easily be analyzed, compared to other data, and used. Alternative systems such as ELN/LIMS merely warehouse the raw datafiles, requiring scientists to take these files out of the system and run them through proprietary software to use the data.



CONVERTING SPECIALIZED INFORMATION TO DATA

Descriptor libraries convert information such as chemical formulas to data points of the relevant features.



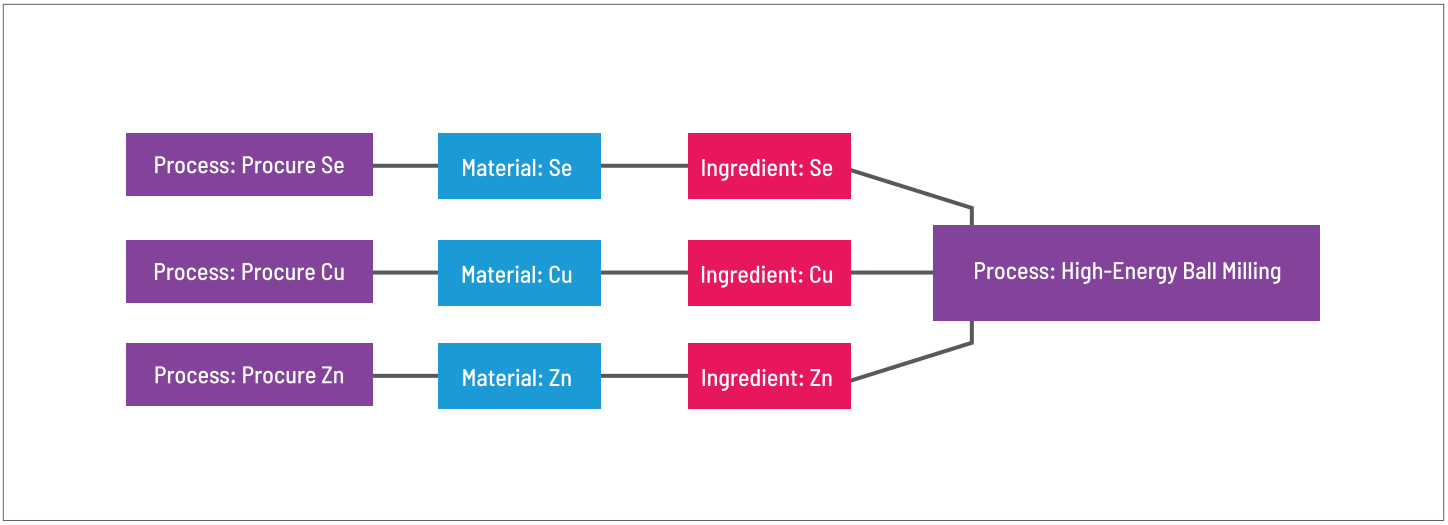
SECURE, SEARCHABLE, AND SCALABLE

Enabling quick access to the data, from anywhere in a global organization while maintaining the security of the data, is a software engineering challenge. Citrine's AI Platform is built from the ground up to be used enterprise-wide with data access controls. Teams can be self-sufficient but still benefit from network effects. Citrine's Platform is deployed privately on AWS in a way that scales dynamically as more teams get on board. It is built to the information security standard ISO27001.

DATA MODEL

Historically, database systems were designed for banks not for science! Relational databases have a rigid design, making adding/removing categories of information slow and difficult. The most fortuitous experiments are the unexpected results, so you need a system that doesn't get in the way of following those leads. In contrast, data management is often so unpleasant that for many companies the only structured data that is stored is metadata (experiment ID, data, sample ID, etc.) or core quality control measurements that are unlikely to change. This results in scientists using a sea of spreadsheets to store data from their ongoing work. What is needed is a data model that never turns away data. It needs to be structured but not rigid.

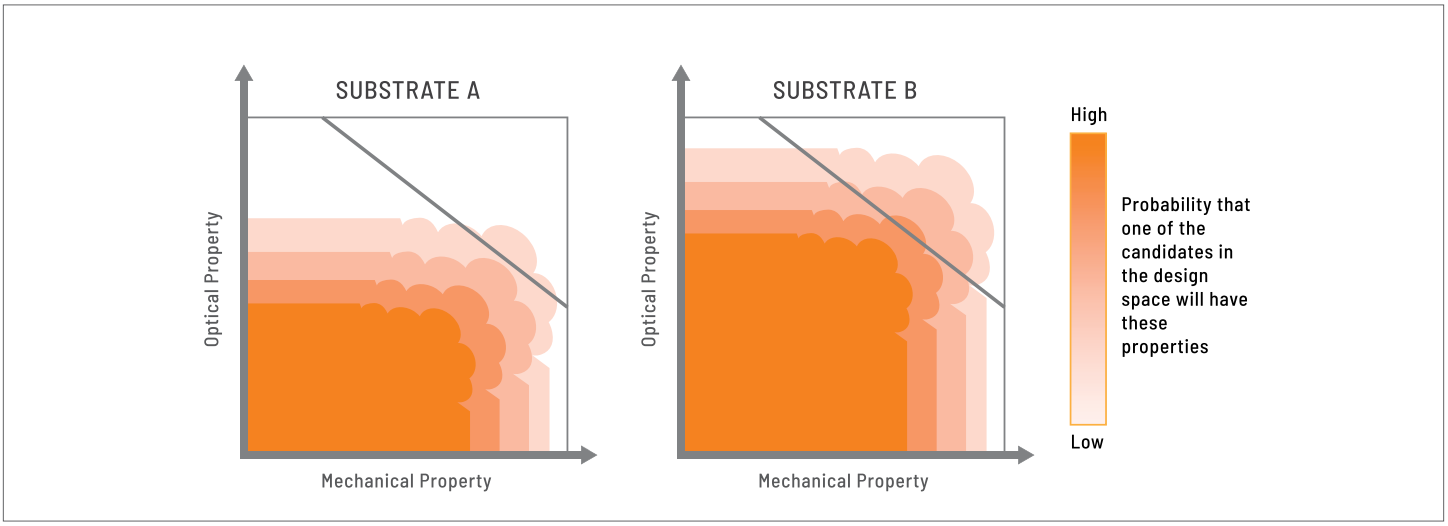
By partnering with and learning from customers across different materials and chemical classes, Citrine's team of data science and material science experts have developed a materials data model (Graphical Expression of Materials Data: GEMD). The key to the model is being able to systematically specify all the context surrounding the steps in a material processing history. All of the processing conditions and batch measurements are relevant to the final properties of the material. Each step from procurement to final product is visualized in the material history; clicking uncovers the details. At each stage, both the specified processing paramters (e.g. 200°F) and the actual measured parameter in that run e.g. (199°F) are recorded. The color-coded graphical user interface is easy to use, but those that are more comfortable with Python can use the Python client to add, review, and revise data.



DATA VISUALIZATION

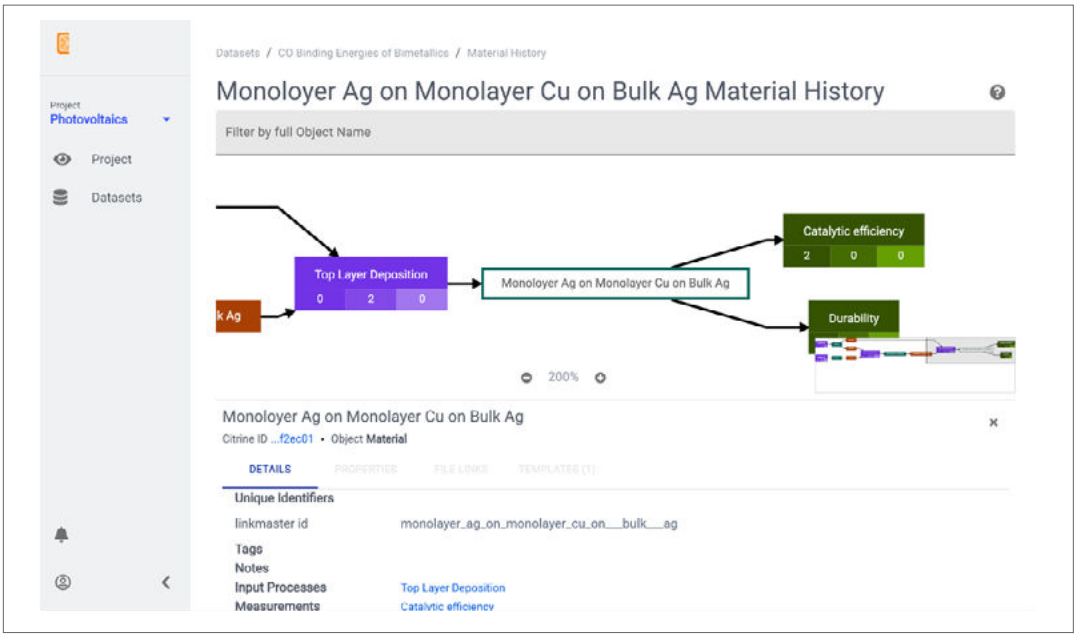
Making the data easily interpretable by researchers leads to more actionable insights. In turn, this usefulness incentivizes the team to consistently manage their data to high standards, which of course leads to a virtuous circle. Citrine's Platform can not only enable simple comparisons of the properties of materials, but also:

- Visually display the chronology of a material, fostering natural understanding of how it came to be
- Gather similar parts of very different processing histories, comparing apples to apples even when dealing with very different end products
- Depict the probability of achieving properties using different materials



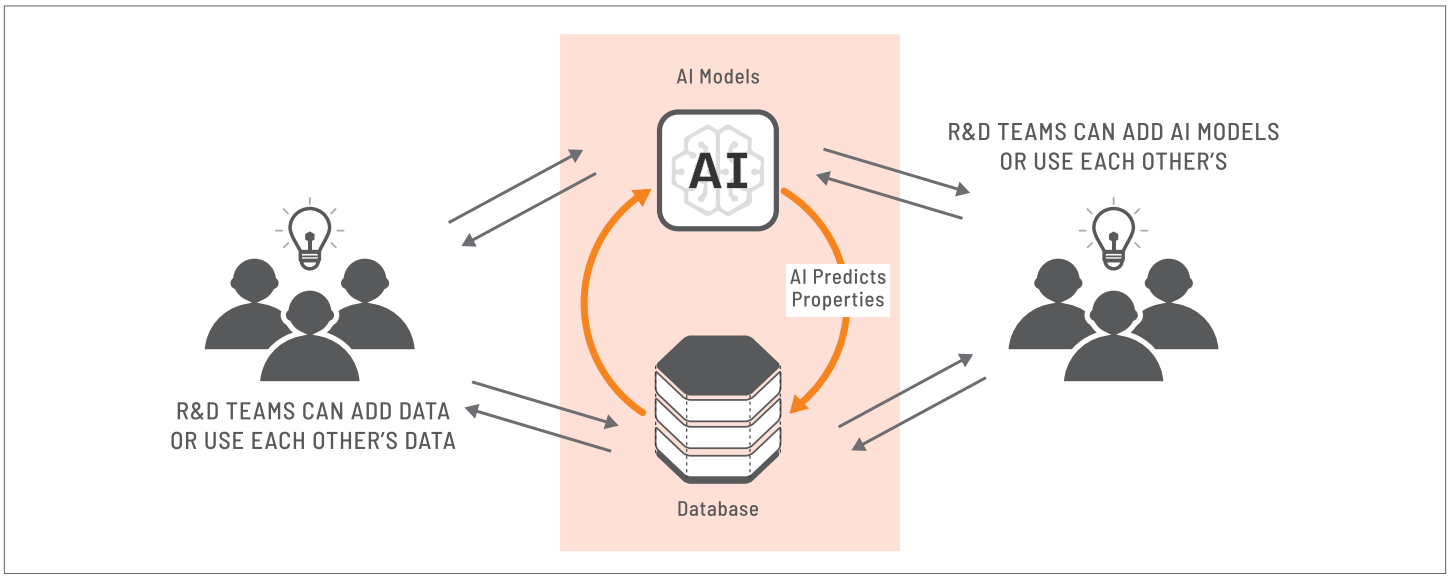
INTUITIVE USER INTERFACE AND TEMPLATING

Making data entry as easy as possible is important. This encourages researchers to add detailed context that is valuable for reuse of data across an organization without being burdened with constantly having to fill every little field. Citrine's Platform enables the quick creation of input forms and templates to ensure standardization.



USING THE LATEST DATA

By making sure that related data from different teams is linked, Citrine's Platform can help researchers to always have the latest information. Teams can use each other's data to improve their AI models.



SUMMARY

Smart and specialized data infrastructure is needed to support innovation in materials and chemicals. Data is the lifeblood of research and of modern AI tools used in research and development. The benefits of AI, such as optimized material properties, reduced costs, and increased customer responsiveness, are greater the more curated data you have available. Citrine has created a platform that works with teams to make data entry, curation, analysis, and visualization easy and powerful.

To find out more about Citrine's Platform, contact us.

REFERENCES

1: Hernandez, Paul. Materials Genome Initiative. NIST, 22 Oct. 2018, www.nist.gov/mgi
2: (Graphical Expression of Materials Data: GEMD, https://citrineinformatics.github.io/gemd-docs/

ABOUT CITRINE INFORMATICS

Citrine Informatics is the award-winning materials informatics platform for data-driven materials and chemicals development. It won the 2017 World Materials Forum Start-up Challenge, the 2018 AI Breakthrough award as the "Best AI-based Solution for Manufacturing", and 2020 Cleantech 100 honors. The Citrine Platform combines smart materials data infrastructure and Artificial Intelligence, which accelerates development of cutting-edge materials, facilitates product portfolio optimization, and codifies research IP; enabling its reuse and preventing its loss. Citrine's customers include Panasonic, LANXESS, and some of the biggest and most respected names in the materials and chemicals industry in Asia, North America, and Europe. For more information visit our website at Citrine.io, or contact us at +1 650-276-7318.