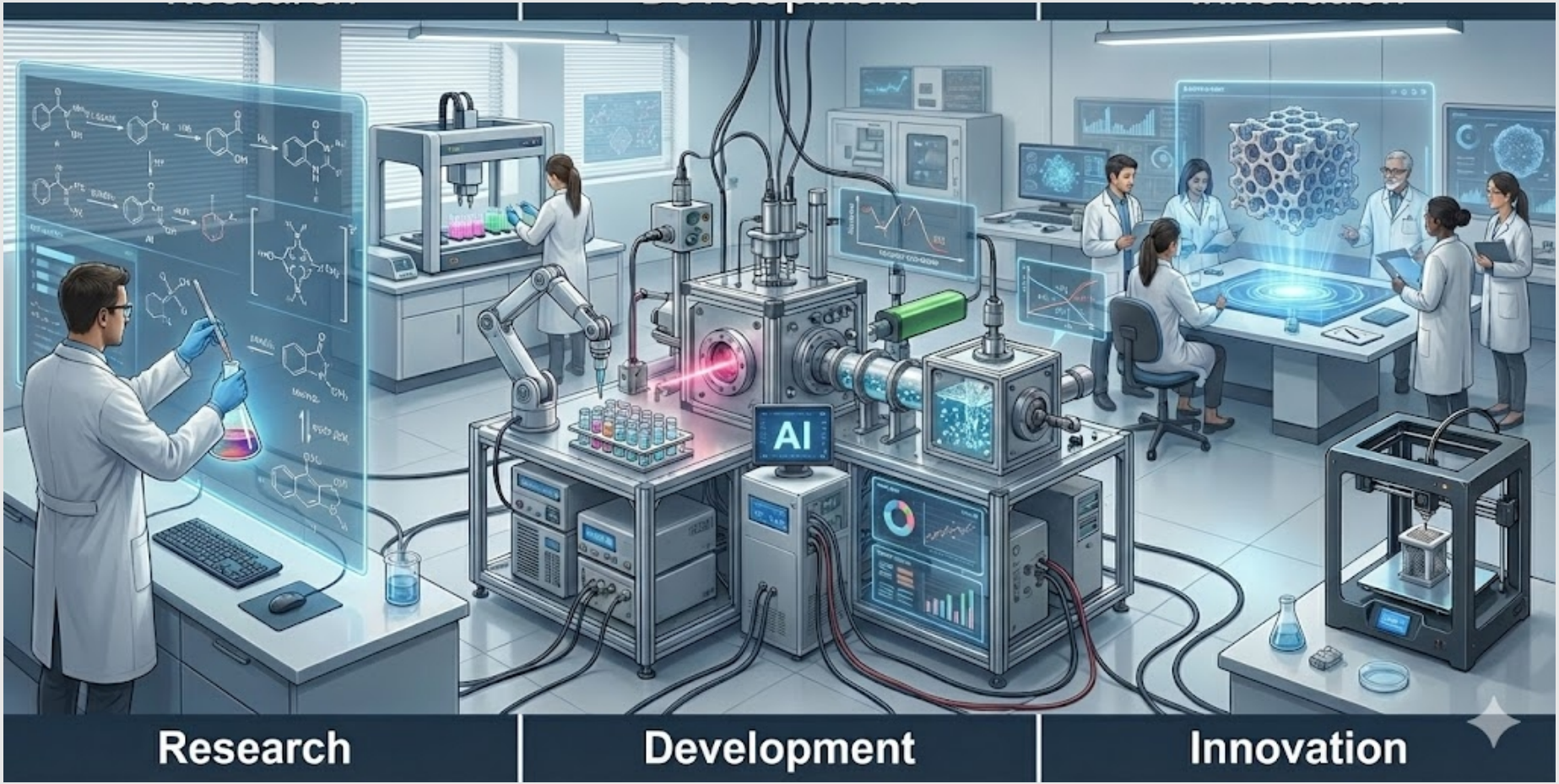
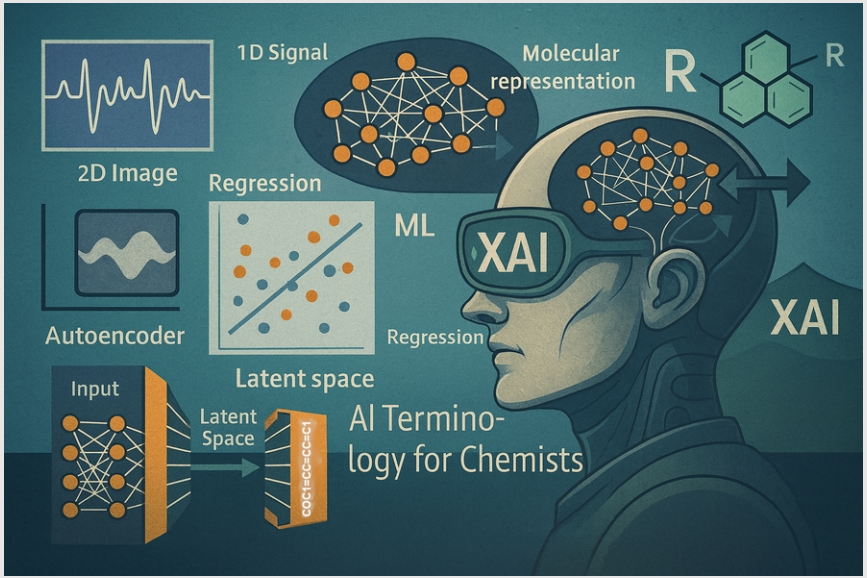

AI
Research
Development
Innovation

Typical workflow – with a bit of xAI



Strategy

Which data ?

How much?

Collect data

Analyze, curate and format data

Hidden patterns?

Train and optimize models

Which algorithm?

Validate model performance on new data

What did the model learn?

Explainable AI tools to see through the *black box* models

How many data?

There is no single magic number

Data requirements vary by several orders of magnitude depending on whether you are building a generalist "chemical brain" or a specialized expert tool

Typically

- **Generalist AI: 10^9. Example = LLMs that can understanding the language and concepts of chemistry**
- **Specialist AI: 10^6. Example = Proprietary databases of chemical reactions**
- **Expert AI: 10^2 - 10^3. Example = Structure-Activity Relationships (SAR)**

Don't forget: quality is better than quantity