ClarifAI

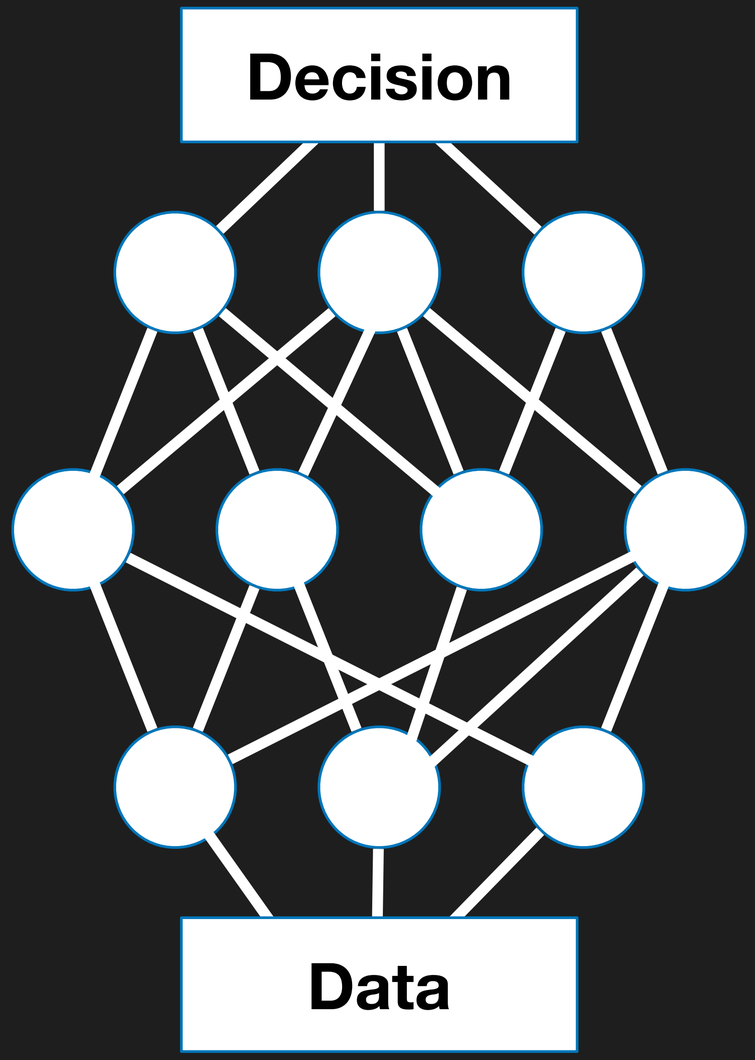
Unveiling the Black Box of AI with Precision and Clarity.

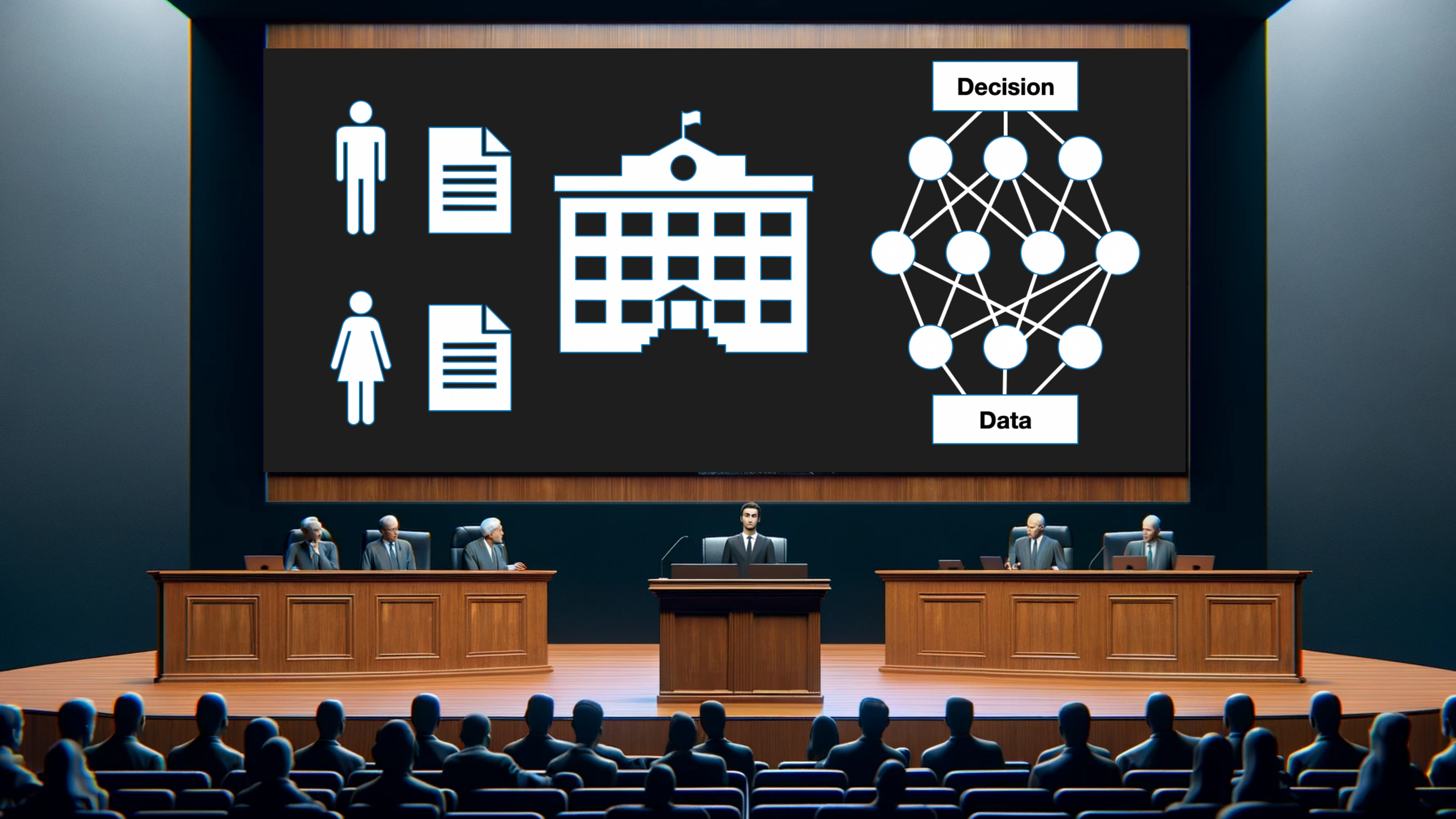
AI Safety Hackathon 11th-12th Nov, 2023.

Problem

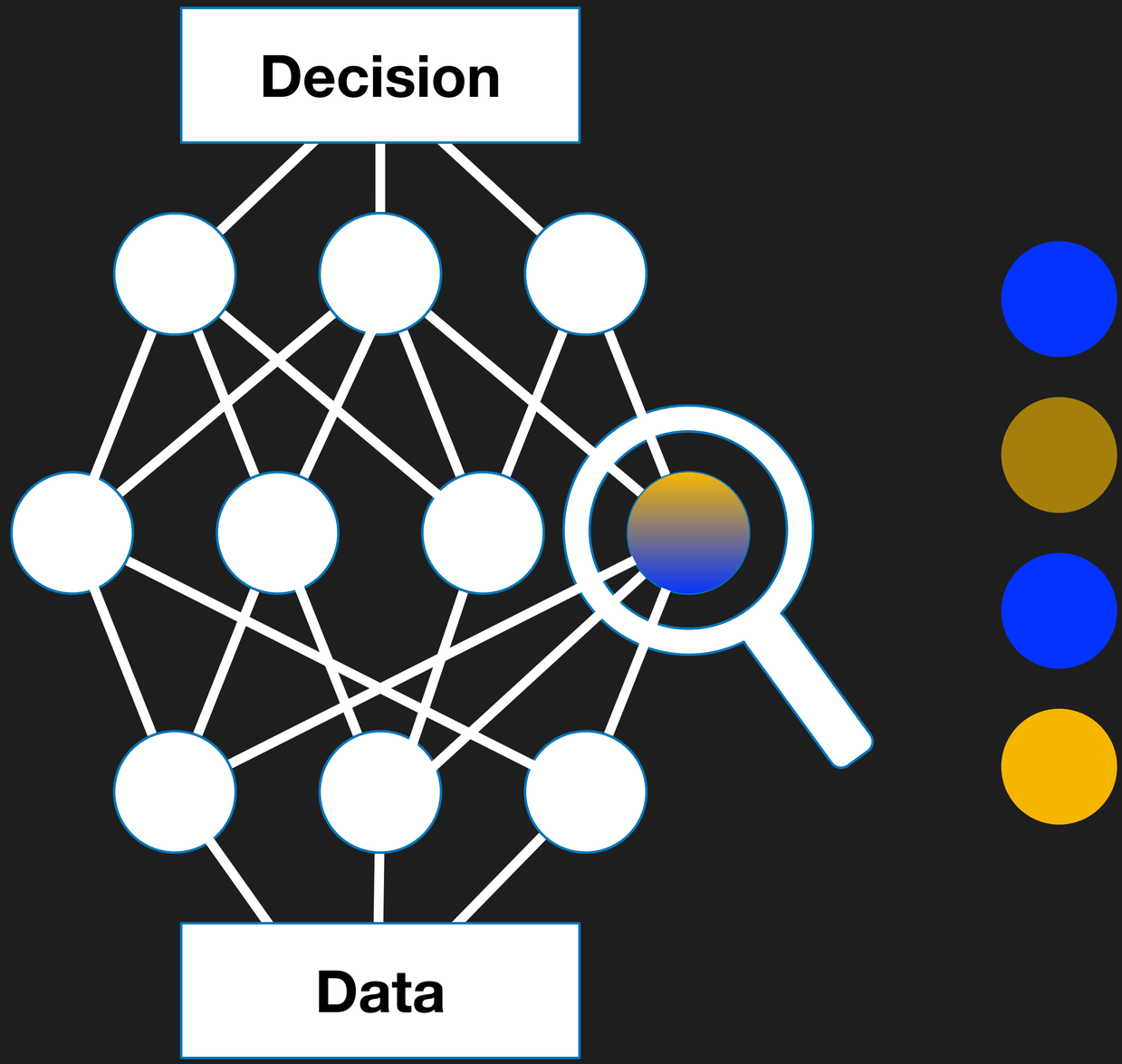
Understanding the 'Why' Behind AI Decisions

Legal clarity through feature attribution.



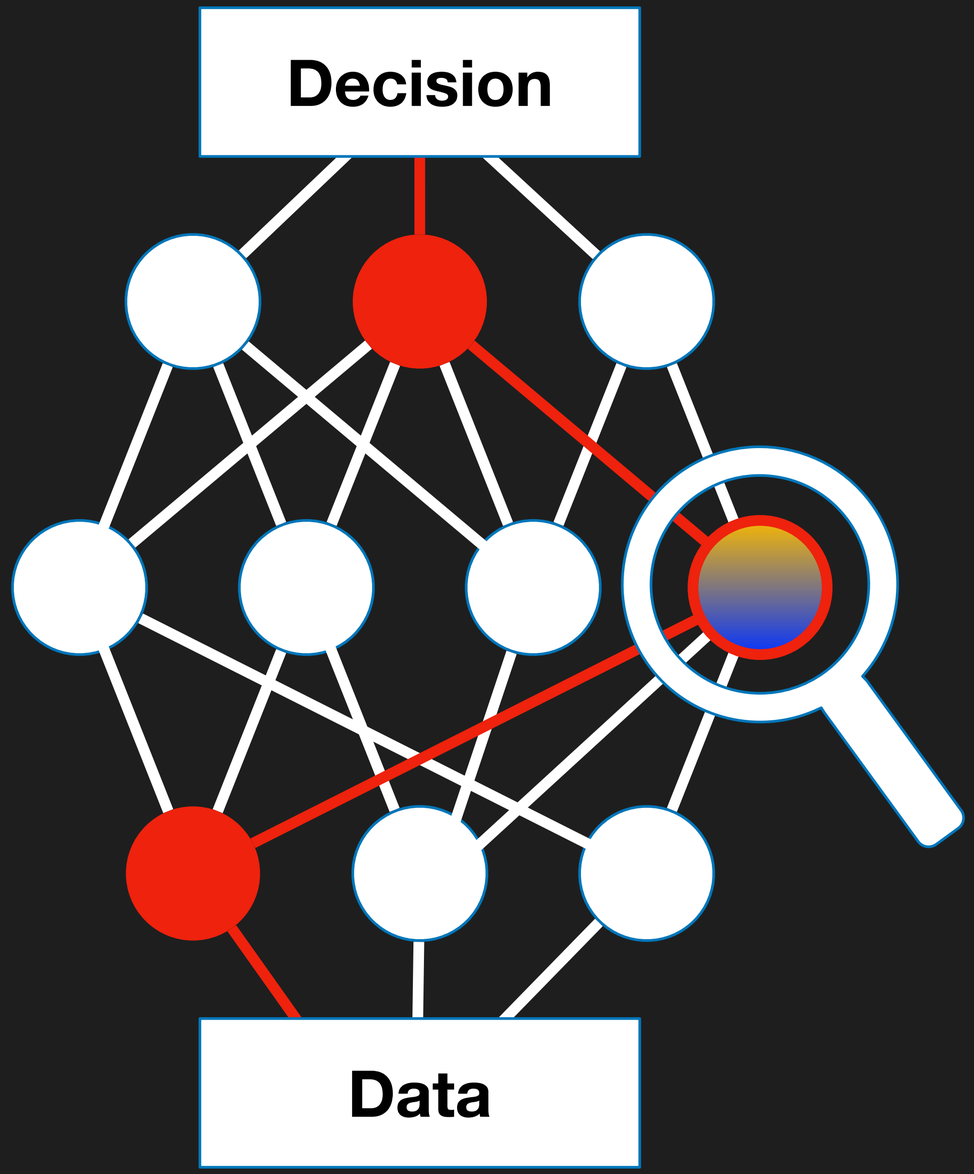

Decision
Data

Problem



Status Quo

Automated Circuit Discovery



Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S. & Garriga-Alonso, A. Towards Automated Circuit Discovery for Mechanistic Interpretability. *ArXiv* (2023).

What is Superposition?

Superposition Hypothesis:
Features >> Neurons.

- Features are represented as near-orthogonal directions.
- **Advantage**: Can represent more features: information compression outweighs the cost of *interference*.

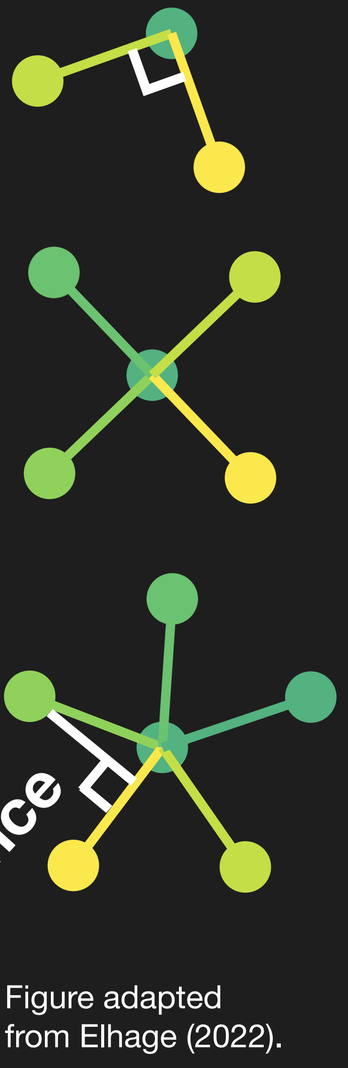
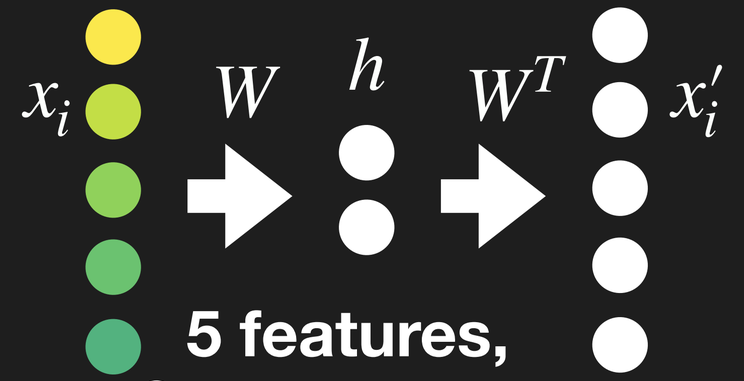


Figure adapted from Elhage (2022).

Importance I_i

- most
- medium
- least important

$$h = Wx$$
$$x' = \mathrm{ReLU}(W^T h + b)$$

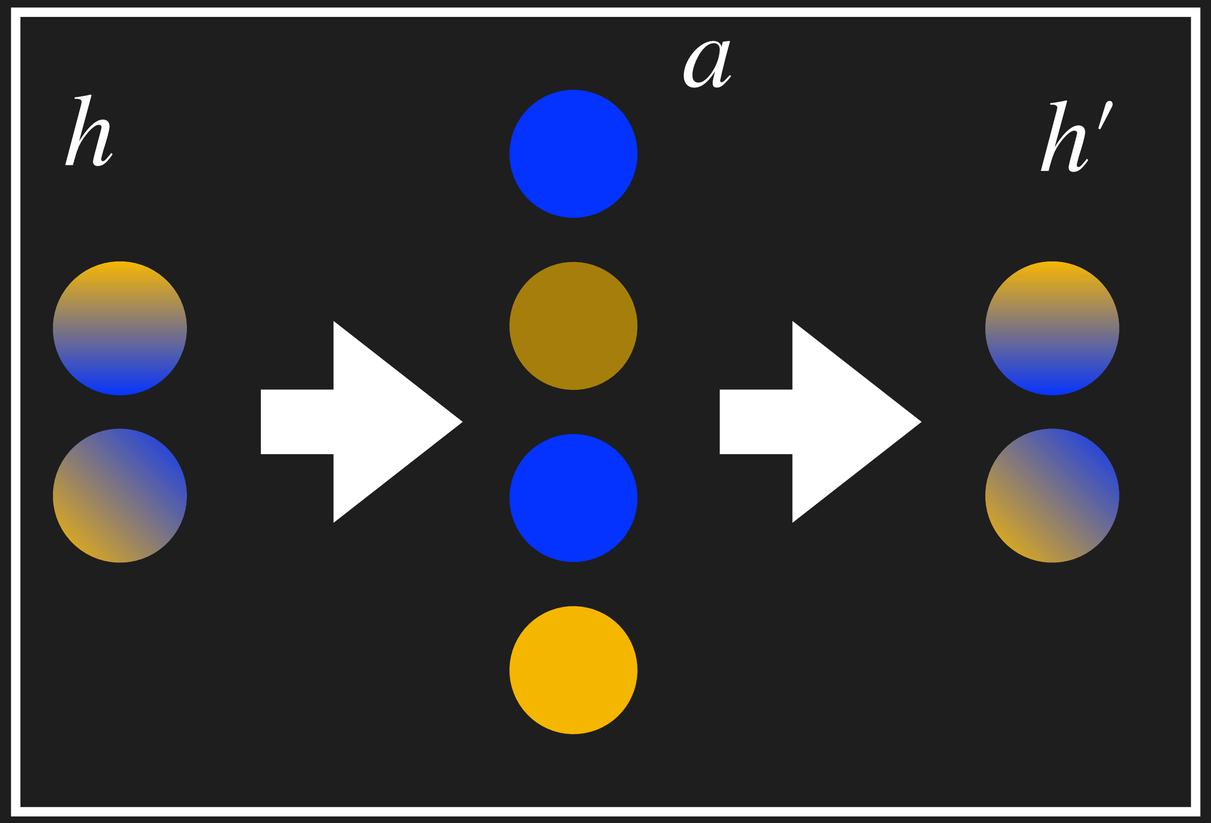


5 features, 2 neurons.

$$L = \sum_x \sum_i I_i (x_i - x_i')^2$$

Elhage, N. *et al.* Toy Models of Superposition. *Transformer Circuits Thread* (2022).

Sparse Autoencoders

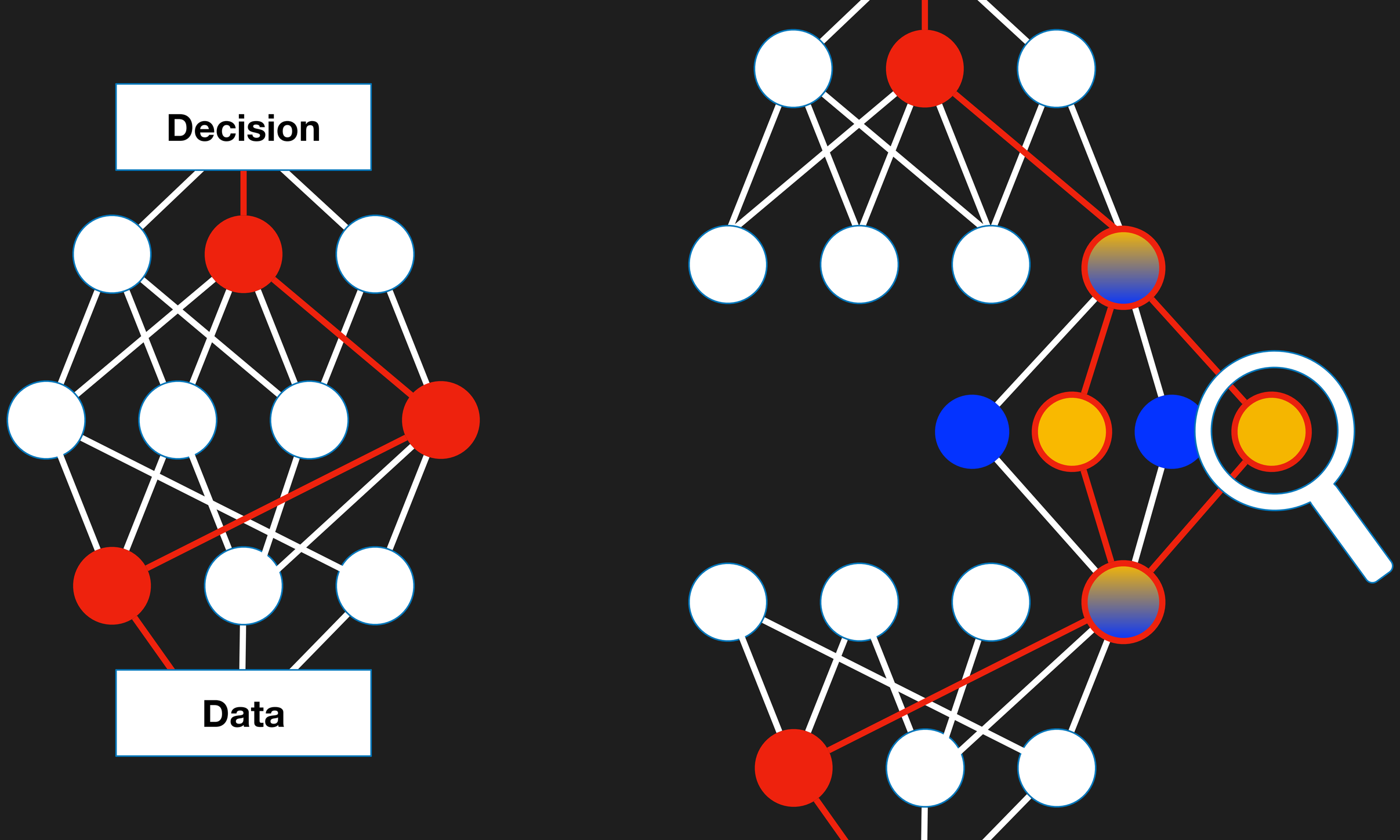


Sharkey, L. *et al.* Taking features out of superposition with sparse autoencoders. *alignmentforum* (2022).

Cunningham, H., e al. Sparse Autoencoders Find Highly Interpretable Features in Language Models. *ArXiv*, (2023).

Bricken, T. *et al.* Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread,* (2023).

Solution

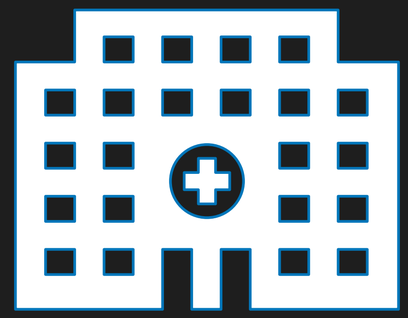

Decision
Data

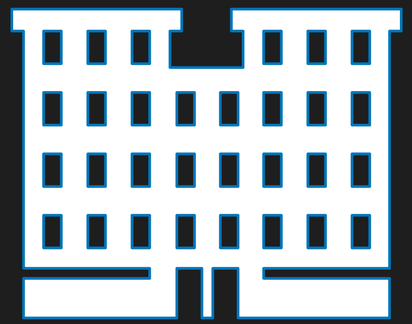
Business

Target customers

Small to Medium Enterprises

- Accessible AI interpretation

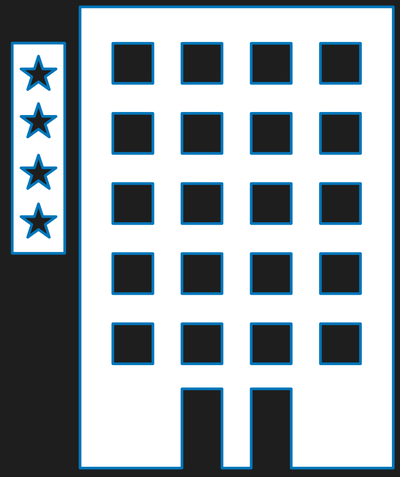




Target customers

Large Corporations

- Scalable, comprehensive AI interpretation
- Regulator compliance (e.g. GDPR), informed decisions



Revenue Streams

	Basic	Professional	Enterprise
Features	• Basic interpretability tools, • Manual reporting.	• Advanced interpretability tools, • Customizable automated reports, • Integration support.	• Full suite of interpretability tools, • On-premises deployment option, • Custom integrations.
Pricing	Pay-per-use	Subscription-based	Custom solution

Transparency for AI Safety

- Empirical alignment and model **analysis**
- **Predictive** understanding of AI scaling
- Enhanced auditing and **deception** detection

Thanks for your attention!