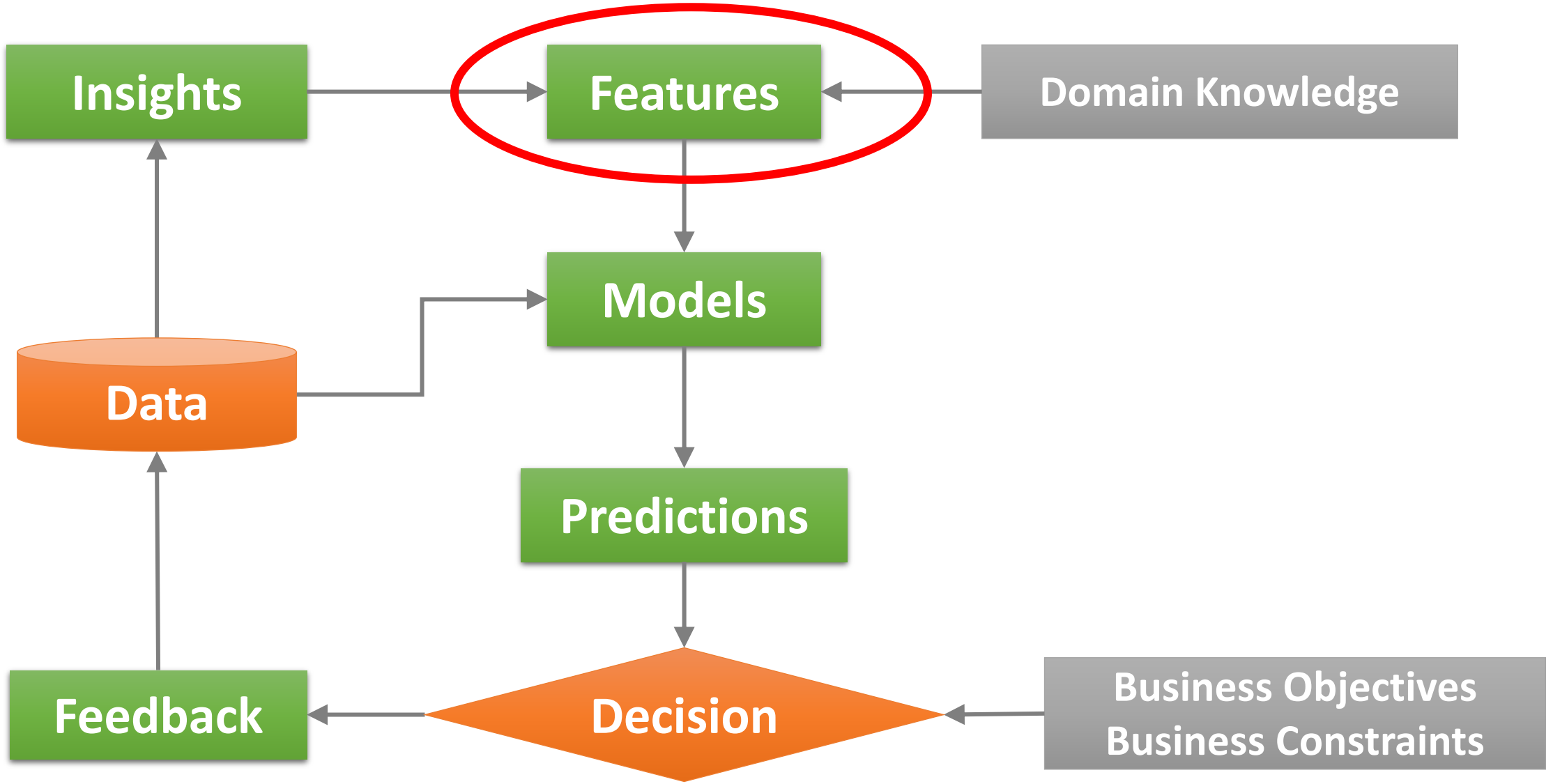
Machine Learning

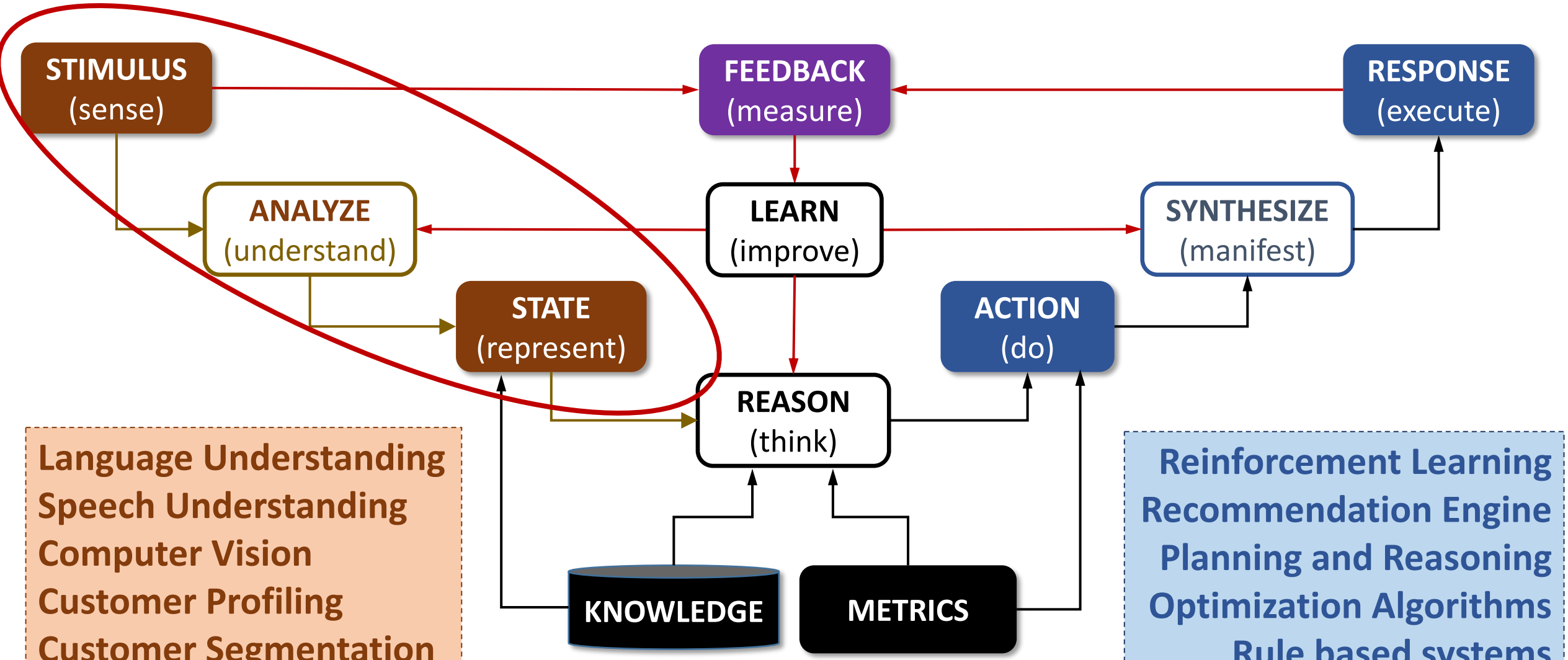
FEATURE ENGINEERING

Shailesh Kumar

Feature Engineering



Feature Engineering



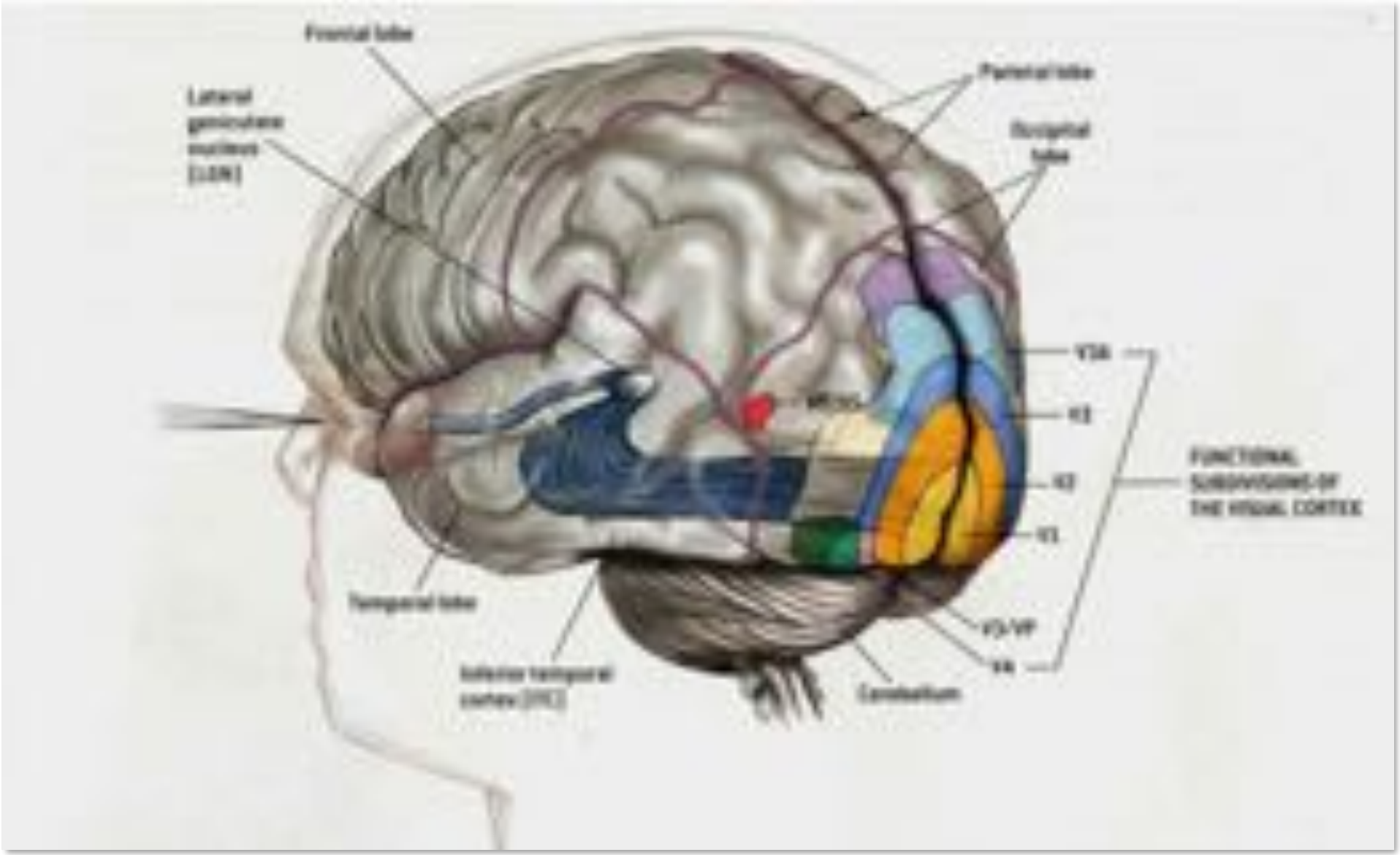
Language Understanding
Speech Understanding
Computer Vision
Customer Profiling
Customer Segmentation
Demand Forecasting

Reinforcement Learning
Recommendation Engine
Planning and Reasoning
Optimization Algorithms
Rule based systems
Business Logic

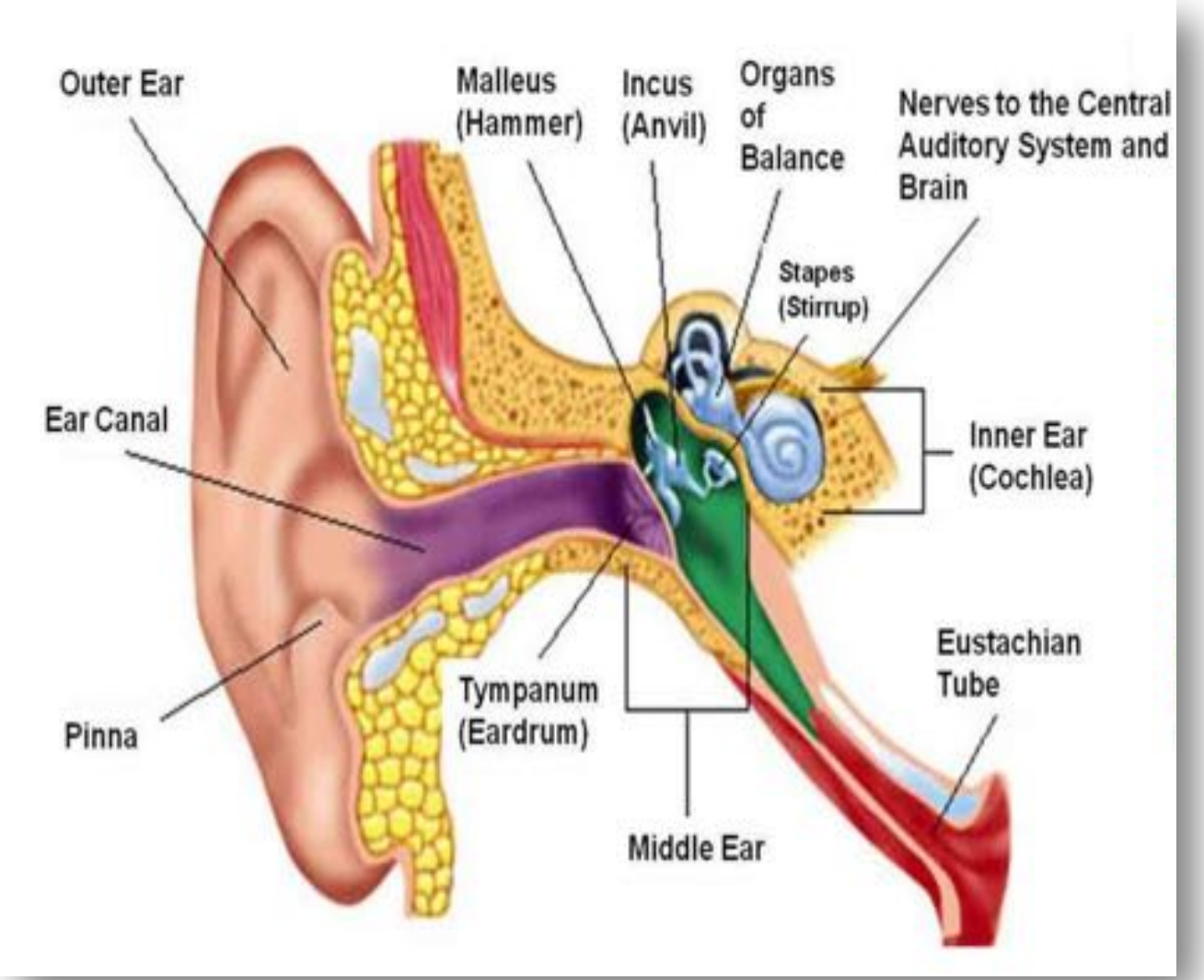
Visual Cortex | The Most Complex Feature Engineering

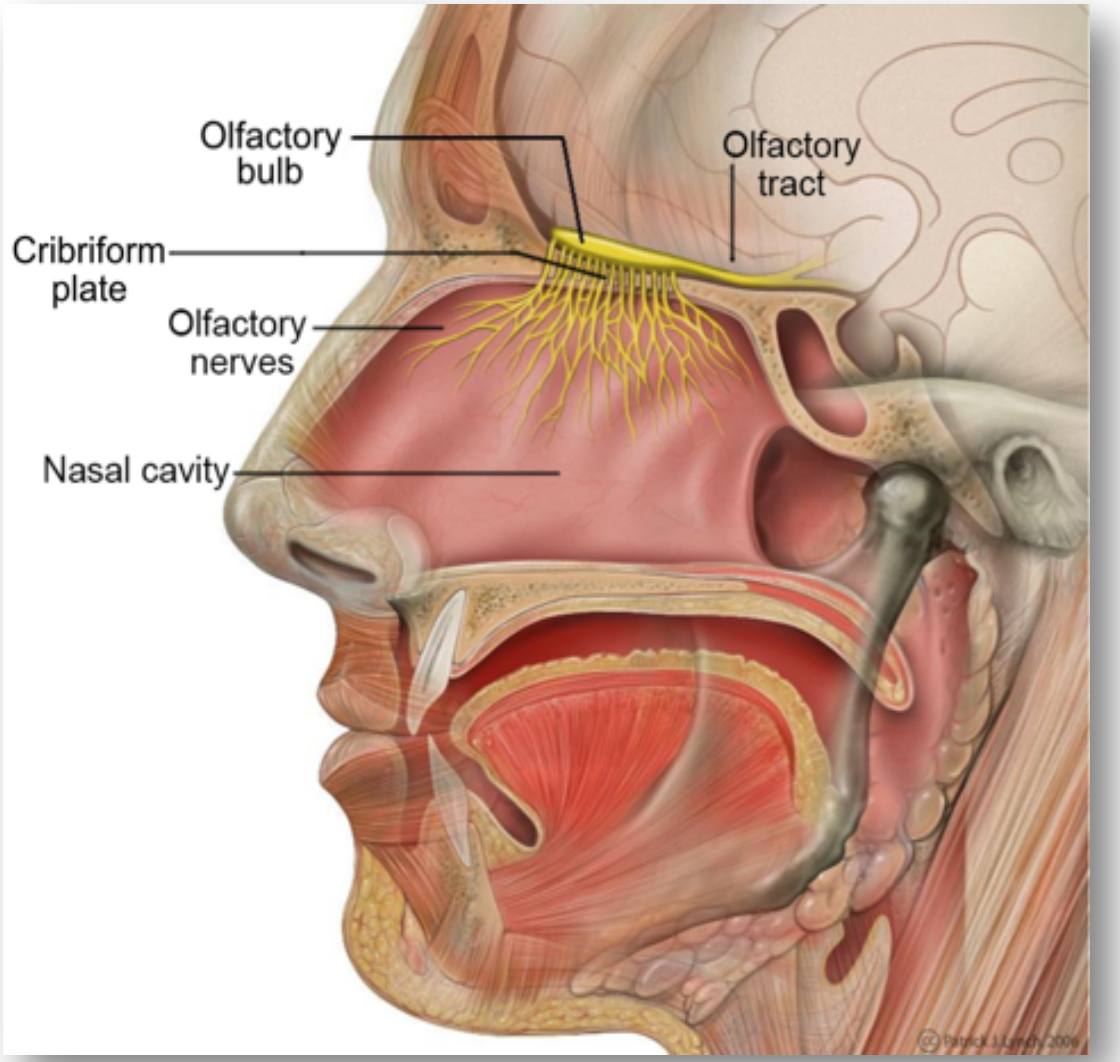
40% of brain activity is focused on visual processing alone!

Modern Deep Learning for Vision is inspired in part by the Visual Cortex

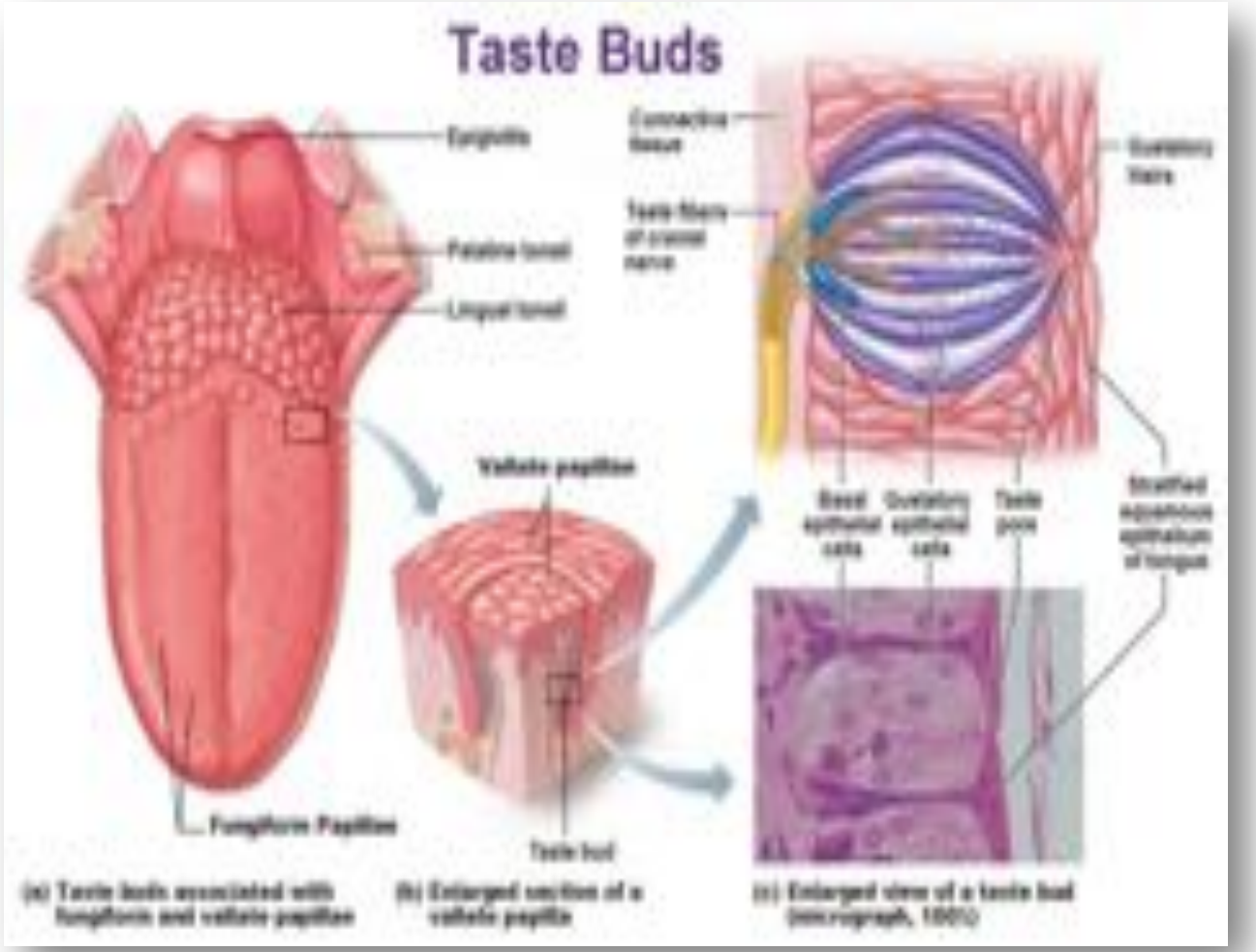


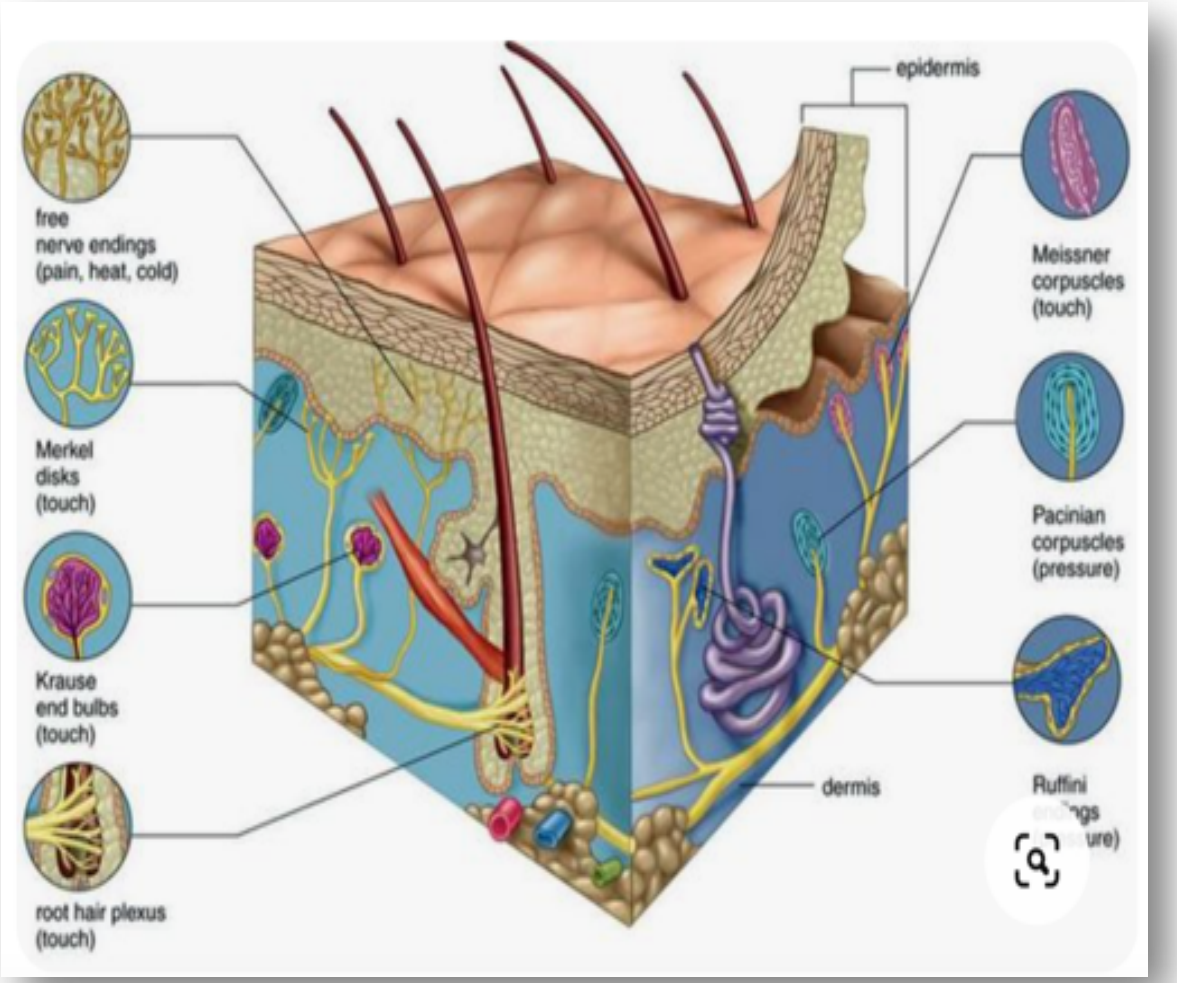
Our Sensory organs have in-built feature engineering





Our Sensory organs have in-built feature engineering

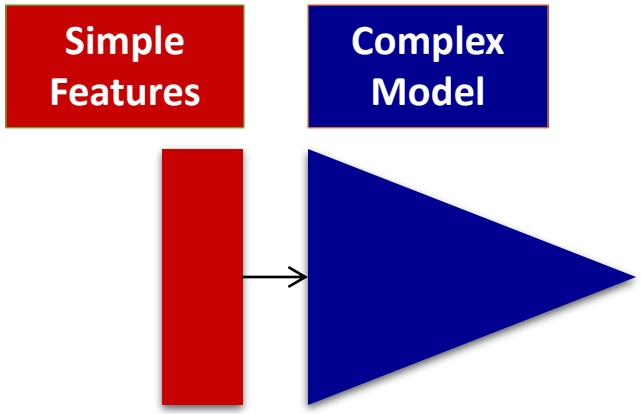




Two Mindsets to Modelling

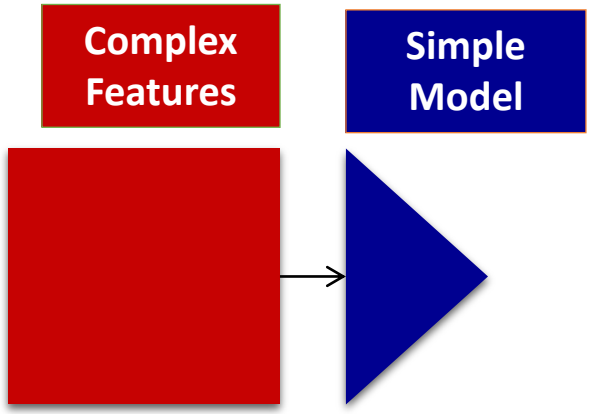
Model-Centric

- Throw all features in!
- Have enough data
- Build Complex models



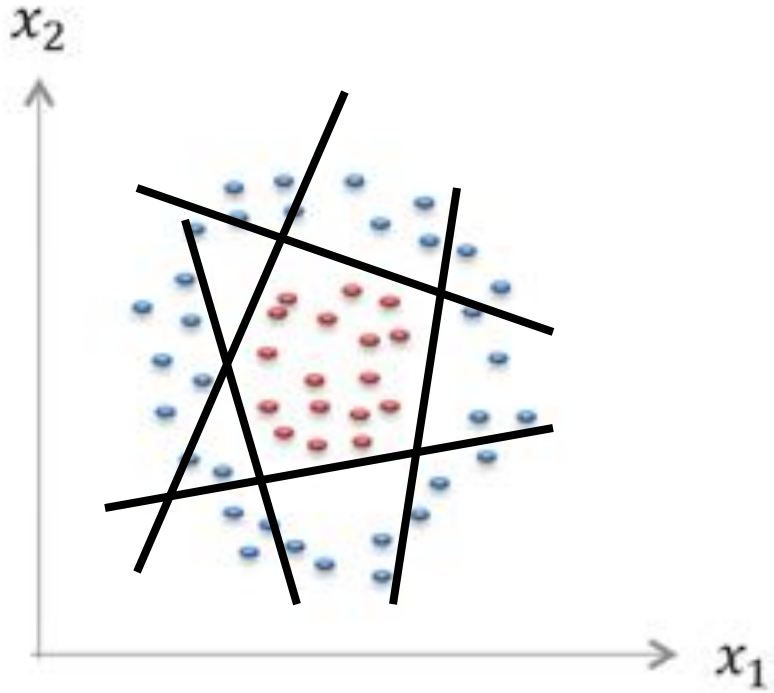
Feature-centric

- Carefully craft features
- Use Domain Knowledge
- Build Simpler Models

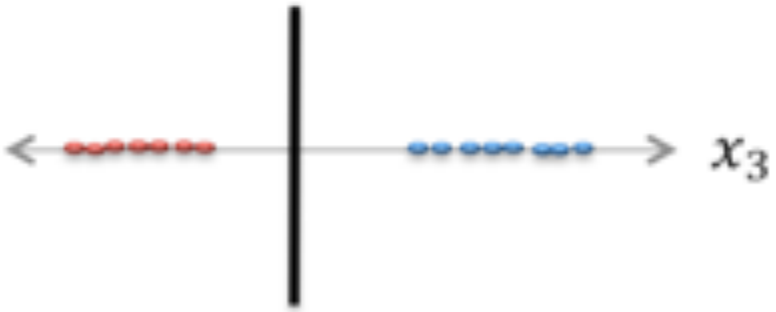


The Law of Conservation of Complexity

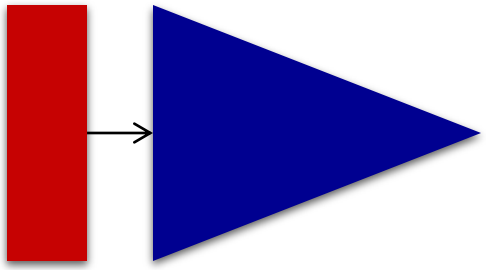
Two Mindsets of Modelling



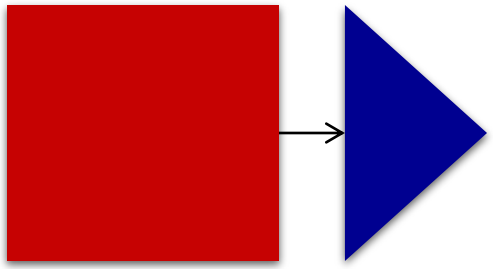
$$x_3 = (x_1 - a_1)^2 + (x_2 - a_2)^2$$



Simple Features → **Complex Model**



Complex Features → **Simple Model**



Feature Engineering using Domain Knowledge

Raw Input

- **Time** of **current** trans.
- **Place** of **current** trans.
- **Time** of **prev.** trans.
- **Place** of **prev.** trans.

Derived Feature - 1

- **Distance**(Prev→Current)
- **TimeLag**(Prev→Current)

Derived Feature - 2

- **Velocity**(Prev→Current)

$$\text{Velocity}(\text{Prev} \rightarrow \text{Current}) = \frac{\text{Distance}(\text{Prev} \rightarrow \text{Current})}{\text{TimeLag}(\text{Prev} \rightarrow \text{Current})}$$

Model the Variance that Matters

- *Match*(**field** | **query**)= how relevant is a field to a query

$$Match(\mathbf{field} \mid \mathbf{query}) = \sum_{token \in \mathbf{query}} Weight(token) \times Match(\mathbf{field} \mid token)$$

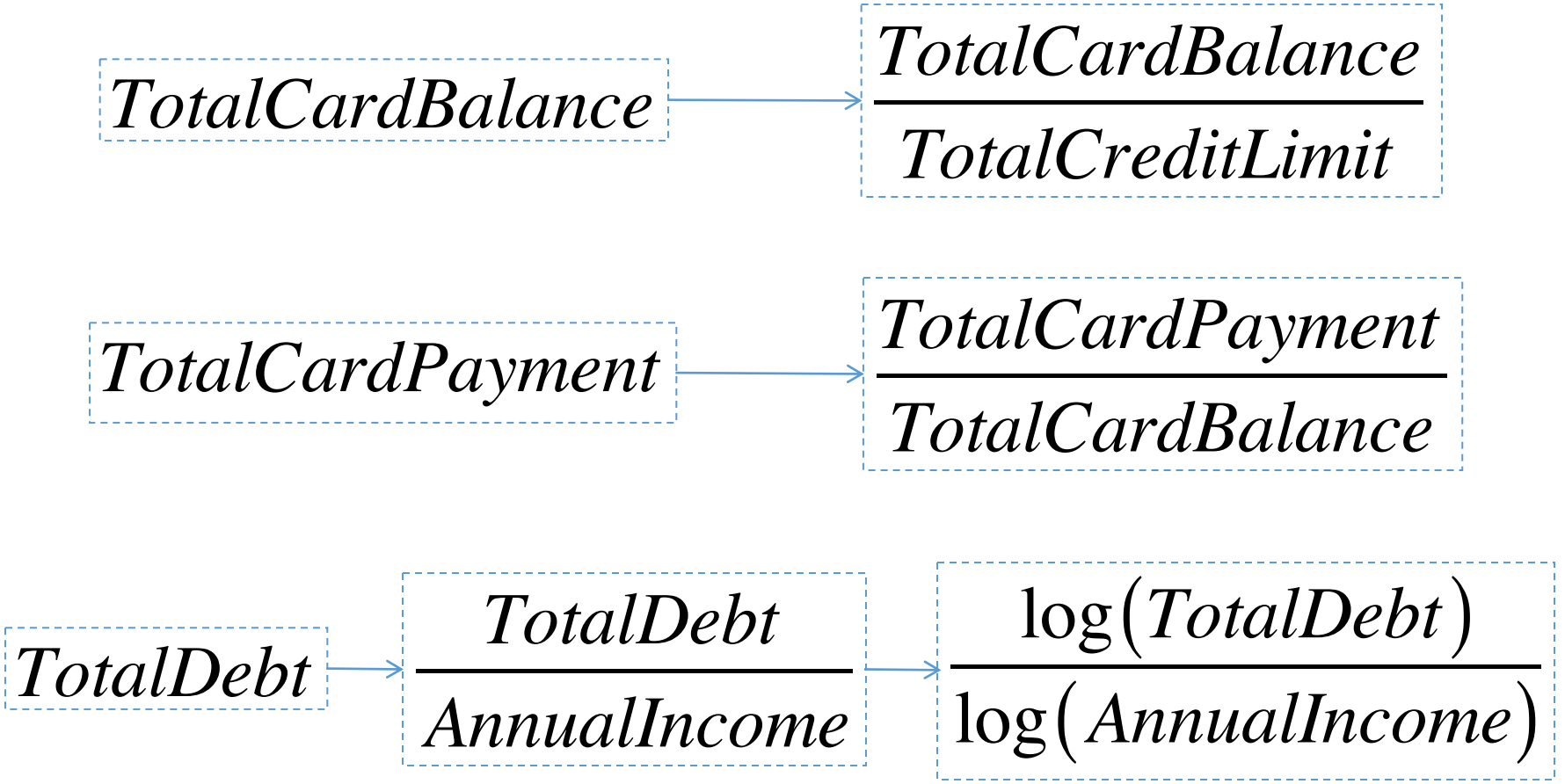
- Does **query length** matter to overall relevance?

$$Match(\mathbf{field} \mid \mathbf{query}) = \frac{\sum_{token \in \mathbf{query}} Weight(token) \times Match(\mathbf{field} \mid token)}{\sum_{token \in \mathbf{query}} Weight(token)}$$

- What about **field length**?

Model the Variance that Matters

Raw Feature **Normalized Feature**



Model Deviations from Expected

$$TotalSales(Context) \rightarrow \log\left(\frac{TotalSales(Context)}{ExpectedSales(Context)}\right)$$

$$CTR(query, url, position) \rightarrow \log\left(\frac{CTR(query, url, position)}{ExpectedCTR(position)}\right)$$

“Bugs” in Feature Engineering

- Observation: Model is **Unexpectedly Complex**
- Hypothesis: It is **Compensating for some “bug”**

$$\mathbf{field} = \textit{the quick brown fox jumped over a lazy dog}$$

$$TermFrequency(query|\mathbf{field})$$

$$TermFrequency(quick|\mathbf{field}) = 1$$
$$TermFrequency(brown|\mathbf{field}) = 1$$
$$TermFrequency(dog|\mathbf{field}) = 1$$
$$TermFrequency(cat|\mathbf{field}) = 0$$

$$FirstOccurence(query|\mathbf{field})$$

$$FirstOccurrence(quick|\mathbf{field}) = 1$$
$$FirstOccurrence(brown|\mathbf{field}) = 2$$
$$FirstOccurrence(dog|\mathbf{field}) = 8$$
$$FirstOccurrence(cat|\mathbf{field}) = 0$$

Is there anything wrong here?

Careful with those “Defaults”

$$\mathbf{field} = the\ quick\ brown\ fox\ jumped\ over\ a\ lazy\ dog$$

$$FirstOccurrence(quick|\mathbf{field}) = 1$$
$$FirstOccurrence(brown|\mathbf{field}) = 2$$
$$FirstOccurrence(dog|\mathbf{field}) = 8$$
$$FirstOccurrence(cat|\mathbf{field}) = 0$$

- If query term **present** in field → LOWER is BETTER
- If query term **absent** in field → **No Match = Best Match**
- What is the correct DEFAULT? How about **-1**?
- If query term **absent** in field → **field_length + K**