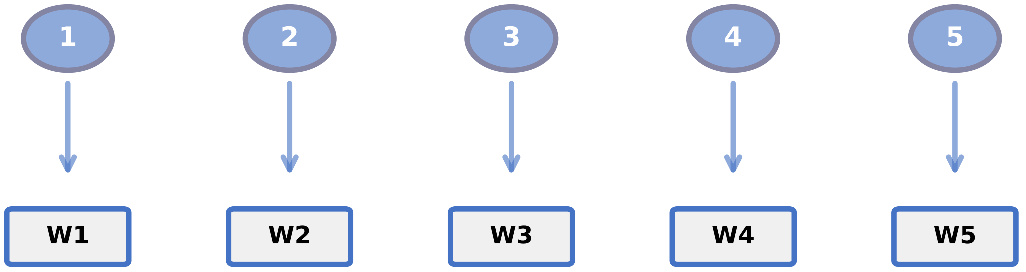
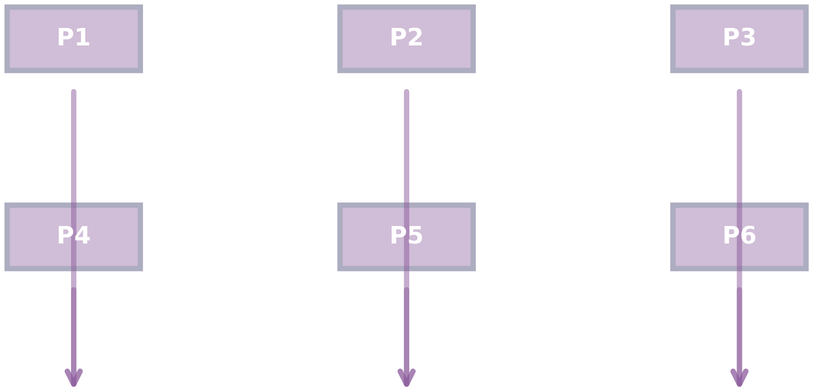
Four Key Principles from Transformers

1. Sequential Processing Not Always Necessary



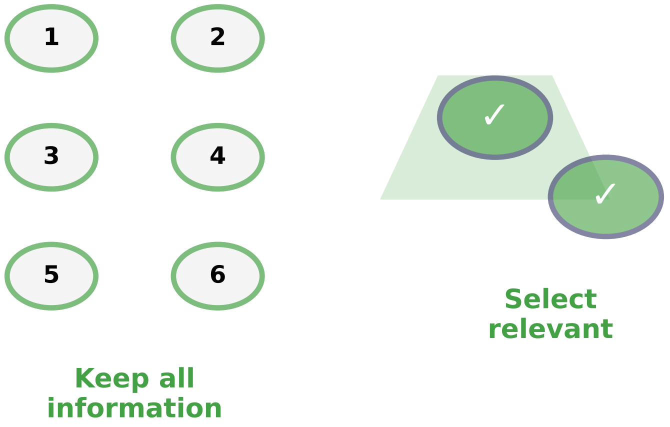
Order can be encoded,
not just processed sequentially

2. Parallelization Through Independence



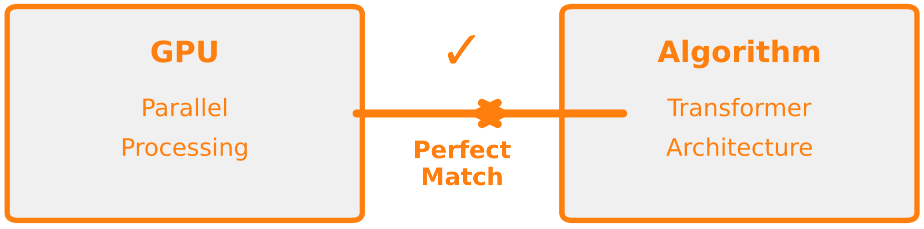
Trade more compute operations
for less wall-clock time

3. Selective Attention vs Compression



Keep information, let model
decide what's relevant

4. Hardware-Algorithm Co-Design



Match architecture to hardware capabilities

Transformer perfectly utilizes
GPU parallel processing power