Architectur (al innovation) is (mostly) all you need

Saqib Choudhary saqibm@gatech.edu
Lee Fleming lfleming@berkeley.edu

October 2025

Recombinant perspectives of innovation imply a definition of novelty as the introduction of a new component, the selection of a new set of components, or a change in the relationships between components. Empirical work has mainly focused on the first two definitions, using technology classifications or word tokens as proxies for components. Here we use NLP and AI to parse patent claims to identify novel subjects, verbs and objects, which we then categorize as Components, Sets of components, and Relationships (CSR). We find strong correlations of our Component and Sets of components measures with traditional measures of novelty and yet also find that Relationship novelty accounts for the bulk of novelty as well as the most highly cited novelty. We document a more nuanced technology life cycle; while PQRs (Pasteur Quadrant Researchers) are more likely to introduce novel components into technology, they are less likely to then search for new Relationships. We are currently developing a shift share instrument in order to infer the causal impact of PQRs on the success of each type of novelty.

Methods

We analyze claims from patents filed between 1975 and 2020 across three technology classes (A61, C07, C12). Our approach converts each claim into subject–verb–object (SVO) structures that capture the components (nouns) and relations (verbs) at the heart of the invention. Because patent claims are long and syntactically dense, we first split them into sub-clauses, then use a two-pass NLP pipeline: (i) dependency parsing and named entity recognition (NER) to extract candidate components appearing in subject or object positions, and (ii) a second parse to identify the governing verb linking these components. In cases where a verb governs multiple objects, all implied subject–object links are retained. This yields a structured list of SVO triples for every claim.

To prevent spurious novelty signals, we lemmatize nouns and verbs, merge common synonyms and abbreviations (e.g., "global positioning system" and "GPS"), and filter out high-frequency, generic triples (e.g., *display–show–digits*). We further standardize the action vocabulary by grouping verbs both morphologically and semantically: verbs are lemmatized, embedded using a patent-specific sentence-transformer (PatentSBERTa), and clustered into groups via cosine similarity, with each group labeled by its canonical lemma. A parallel procedure is applied to entities (components), using a higher similarity threshold to ensure that only very close semantic matches are merged. These steps produce a compact inventory of verbs and entities that collapses inflectional and near-synonym variation, improving the precision of our novelty classification.

Novelty is then defined independently across three dimensions (CSR) for each patent claim relative to the historical record within each technology class:

1. **New Component** – a component not observed in any prior claims,
2. **New Set of components** – a new selection of previously known components,
3. **New Relationship** – a novel verb linking an already observed component pair,
4. **Not novel** – if none of the above apply.

Claims are classified at the clause level. For each patent we aggregate the number of new components, new combinations, and new relationships introduced across its claims. This framework allows us to separate innovations that arise from introducing new building blocks, from those that result from recombining or reconfiguring known ones.

<u>Summary Statistics:</u>

Data: Patents in three classes from 1975-2020
Sample contains 936,536 unique patents

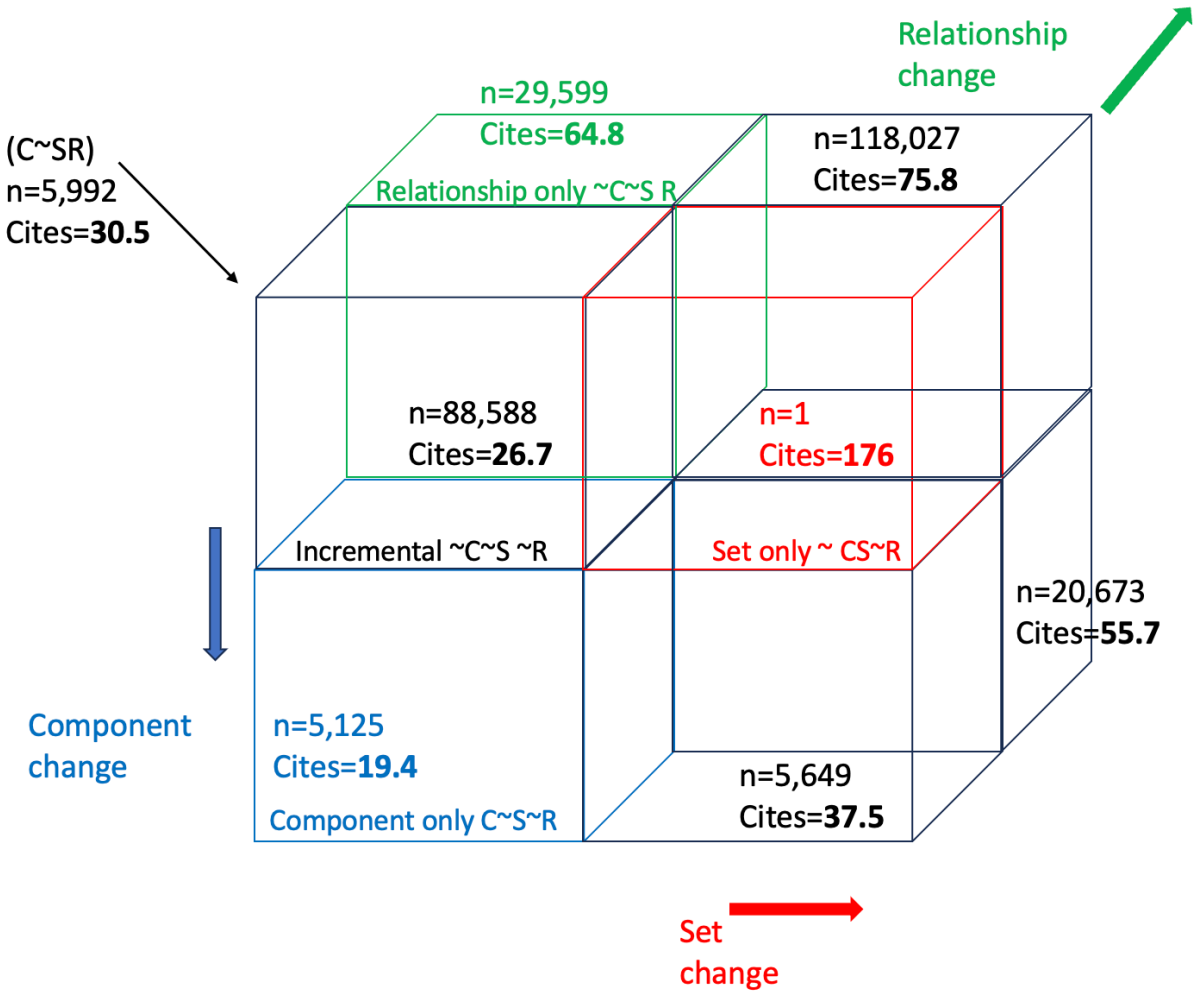
Total relationships analyzed: 14,659,176
New concepts introduced: 79,964 (0.55%)
New subject–object combinations introduced: 923,045 (6.30%)

New relationships introduced: 1,858,424 (12.68%)
Unique subjects seen: 63,494
Unique objects seen: 45,369
Unique subject–object combinations seen: 977,090
Unique verbs seen: 23,397

The Figure illustrates citation outcomes across the octants implied by the definition (these descriptive relationships strengthen in count regressions – the numbers are from an earlier sample and may not correspond to the descriptive statistics listed above).

The following points emerge from the Figure:

- Most novelty occurs with a new Relationship (the green octants in the back)
- Approximately a third of the patent sample indicates no novelty
- Patents which introduce new components are relatively rare and not highly cited



Limitations and Next Step

While NLP-based methods are fast, transparent, and benefit from well-defined rules, they can also introduce errors when applied to complicated patent text. Claims are unusually complex, with long sentences, layers of modifiers, and legalistic phrasing that make precise parsing difficult. Our current approach, which spreads the subject of a clause across all objects governed by a verb, helps capture possible relationships but also adds noise by generating links that may not be explicitly intended. Large language models (LLMs) provide an alternative: their ability to process longer contexts allows them to capture relationships more faithfully across complex claims. In benchmarking, state-of-the-art models (e.g., GPT-5) approach human-level accuracy in extracting subject–verb–object relations from patent claims, though their cost makes them impractical at corpus scale. Smaller models are less reliable on SVO extraction task, but all models perform comparably well in summarizing claims to strip away modifiers and legal boilerplate. We are therefore exploring a hybrid pipeline that first uses an LLM to produce concise summaries of claim text, and then applies NER or related extraction tools to recover structured SVO triples.