Open and Sustainable Innovation Systems (OASIS) Lab

Knowledge synthesis: A conceptual model and practical guide

Joel Chan

Published on: Dec 19, 2020

> **NOTE (March 10th, 2022): The system in this guide has remained stable and useful since the last release, with some minimal implementation-level tweaks. If it helps, you can view some "game film" here. I have also developed an experimental software extension in RoamResearch that smooths this process considerably. This extension shows up in the more recent (from late 2021) videos in the game film playlist.**
>
> **This article will be updated soon to reflect these changes. Sooner if there are more requests!**

Motivation

This document shares a conceptual model and practical approach for knowledge synthesis that I have developed for myself.

It is forged in practice, and adapted from other practices, but also with an eye to being as theoretically grounded as possible. Creative knowledge work is my area of research, after all, and knowledge synthesis is an inherently creative act.

My initial audience for this document is researchers who struggle with **knowledge synthesis** (aka a "real" literature review"), which is the nebulous "black box" in between

> "I have found a bunch of papers to read"

and

> "I now have synthesized the literature and have a set of promising angles of attack on my research problem"

This used to include me! I'm still learning, but this approach has really really helped me gain traction on this problem, and learn how to be able to develop this skill further, and pass it on to my students.

My hope is that you, the reader, will be equipped by this document to:

1. Replicate my system for knowledge synthesis for your own work, should you choose to do so
2. Understand the rationale and theoretical grounding of the system enough to adapt it to the particulars of your context and/or be inspired to develop your own system

I suspect this document will be most successful at achieving these goals if you resonate with the challenges I describe below for a synthesis system (ideally by having experienced them!), and have some familiarity with RoamResearch (or related networked notebooks like Obdisian.md, Remnote, TiddlyWiki, or Tinderbox).

What is synthesis?

Before we go further, let's define our target: what is **synthesis?**

A good place to start is to compare extreme examples:

Here is a "lit review" of observations with no synthesis:

> Species vary: some variations are bad, and some help with survival. Species struggle to survive. Some, but not all, organisms pass on new offspring.

And here is the same set of observations, synthesized:

> Species struggle to survive. Species also vary, and some variations are good and some are bad for survival. Therefore, one precondition for species to survive and pass on offspring is by having or inheriting beneficial variations. This variation and selection process explains how we get the diversity of species we see today.

You may recognize the topic here: it's Darwin's theory of evolution by natural selection!

Notice how the second example ***creates something new***, greater than the sum of its parts, namely the explanation for the origin of species. This is a core aspect of synthesis: the ***construction*** of a new point of view from a set of observations, that directly advances knowledge and/or opens up a path to advancing knowledge.

In this case, the synthesis yielded a ***theory***, which to me is a paradigmatic example of synthesis. But a good synthesis can also take other forms, such as a critical literature review that leads to a set of powerful new research questions, or a design argument or problem frame.

I won't belabor this point further, but I do recommend these sources for understanding what synthesis is and what it looks like (and what it does *not* look like):

- Strike & Posner (1983). Types of synthesis and their criteria.
- Levy & J. Ellis (2006). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. Informing Science: The

International Journal of an Emerging Transdiscipline
- Boote & Beile (2005). Scholars Before Researchers: On the Centrality of the Dissertation Literature Review in Research Preparation. Educational Researcher
- Blake & Pratt (2006). Collaborative information synthesis I: A model of information behaviors of scientists in medicine and public health. Journal of the American Society for Information Science and Technology
- Holbrook (2008). Levels of success in the use of the literature in a doctorate. South African Journal of Higher Education

Suffice it to say that this kind of intellectual product is what I'm optimizing my system for, to help me do this in a sustainable way. It is my job, after all!

Challenges and desiderata for a synthesis system

Here are some common failure modes for a synthesis system and process that I have experienced and observed in others (not mutually exclusive!):

1. **Too much detail** (too low-level, missing forest for trees). This manifests as a lack of higher-level synthesis of what a collection of results means. A common manifestation is the "x said this, y said this, z said this" form of literature review.
2. **Too little detail** (too high-level, missing the devil/diamonds in the details). This manifests as overgeneralization of claims, or glossing over critical inconsistencies or contradictions. A good example of this is debates about the role of "children" in COVID-19 transmission that ignore the details of differences between young children (under 10).
3. **Insufficient context**. This is related to the lack of details, but separate in that context can also come from connection to other claims: if this is missing, even observation notes can be lost because their significance isn't recognized.
4. **Information silos.** This manifests in part also due to inordinate detail-orientedness, where important connections across disciplines or topics are ignored. This can also come from too little detail! If results are described at too high a level, we might miss important connections at the subproblem level between problems and results.
5. **Information overload**. There are often too many papers to read and process in a rigorous and iterative way, which leads to / exacerbates the preceding set of problems!

I won't spend too much more space here to give fully fleshed out concrete examples of this for this release, but if you recognize/resonate with these, then this document is for you!

This is a future that I want: a research group can confidently aim their sights at a complex, interdisciplinary problem area, and construct an effective synthesis together with minimal "busywork overhead": they can just focus on the core task of synthesis, instead of fighting to extract the "trapped data" (Knight, Wilson, Brailsford, & Milic-Frayling, 2019) in PDFs and long documents! The results *and* intermediate products of their synthesis work also provide a stronger foundation for themselves and others in the future to build on.

For this to be true, we need a system that helps us **achieve a generative dialectic** between compression/divergence/abstraction/theory and context/convergence/particulars/data. We also need the system to enable us to **accrete insight over boundaries of time and projects/disciplines**. We don't always have the luxury of being able to devote (funded) time and attention at an intense level for a given project. We often have multiple irons in the fire (good for creativity), and we often want to reuse and remix ideas from the past (Blake & Pratt, 2006). Finally, we need the system to enable us to **distribute work across multiple people**. There are just too many papers for any one person to absorb by themselves!

In the next section, I'll start to describe a conceptual model I've found helpful for thinking about and implementing a synthesis system that meets these requirements. I also describe in detail how I implement this in my own work, with hopefully enough detail that you, the reader, can try it out. In later releases, I will also add some discussions of open problems and ideas for improving the system.

The conceptual model

"Data model"

In this model, we create and update four basic kinds of entities in the synthesis process:

1. **Question** notes, which express an open research question,
2. **Synthesis** notes, which express a single, generalized idea, such as a claim,
3. **Observation** notes, which express a single, highly contextualized and specific observation that, together with other observation notes, can form the basis of a synthesis note, and
4. **Context snippet** notes, which help to ground and contextualize observation notes.

Let's consider each kind of entity in a bit more detail.

QUESTION notes

Question notes express an open question (e.g., *"What is the effect of analogical distance of inspirations on creative output?"*). They can be readily mapped to research questions in research projects.

SYNTHESIS notes

Synthesis notes articulate a single, generalized idea, such as a claim (e.g., *"Inspirations that are of intermediate distance from the problem domain strike the best balance between benefits for novelty and quality of ideas"*). In some cases, a synthesis note can encapsulate a more complex single idea, such as a theory (e.g., *"Darwinian theory of evolution by natural selection"*), high level argument (*"Scientific observations are theory-laden"*) or problem (e.g., *"The demarcation problem in philosophy of science"*)

By generalized, we mean that synthesis notes should aim at something that is true of an *equivalence class* of instances instead of expressing a bounded statement about a single instance.

Synthesis notes can be mapped to citation statements in academic publications, which are typically generalized and drawing on more than a single source. They are also similar in flavor to "claims" in the micropublications model (Clark, Ciccarese, & Goble, 2014) (although it can encompass more complex things than a simple assertion), and "permanent notes" in the Zettelkasten method.

OBSERVATION notes

Observation notes articulate a single, highly contextualized observation (e.g., *"the finches on the island had different colored beaks after two generations"*).

By contextualized, we mean that observation notes should tend towards being bounded in the particulars of time, authorship, and setting, as opposed to trying to describe a generalized claim that holds over an equivalence class (that is the function of synthesis notes). The intuition is that observation notes should be as close to "the data" as possible. They should be similar to how results are described in results sections of academic publications.

By convention, we write them in the past tense (to ground them in *time*), bind them to an assertor where possible (to ground them in the standpoint of the author), and tend towards lower levels of abstraction (to ground them in relevant particulars).

Observation notes can be mapped to “literature notes” in the Zettelkasten method, or “lines of evidence” in models of scholarly argumentation like the SEPIO model (Brush, Shefchek, & Haendel, 2016)

CONTEXT SNIPPET notes

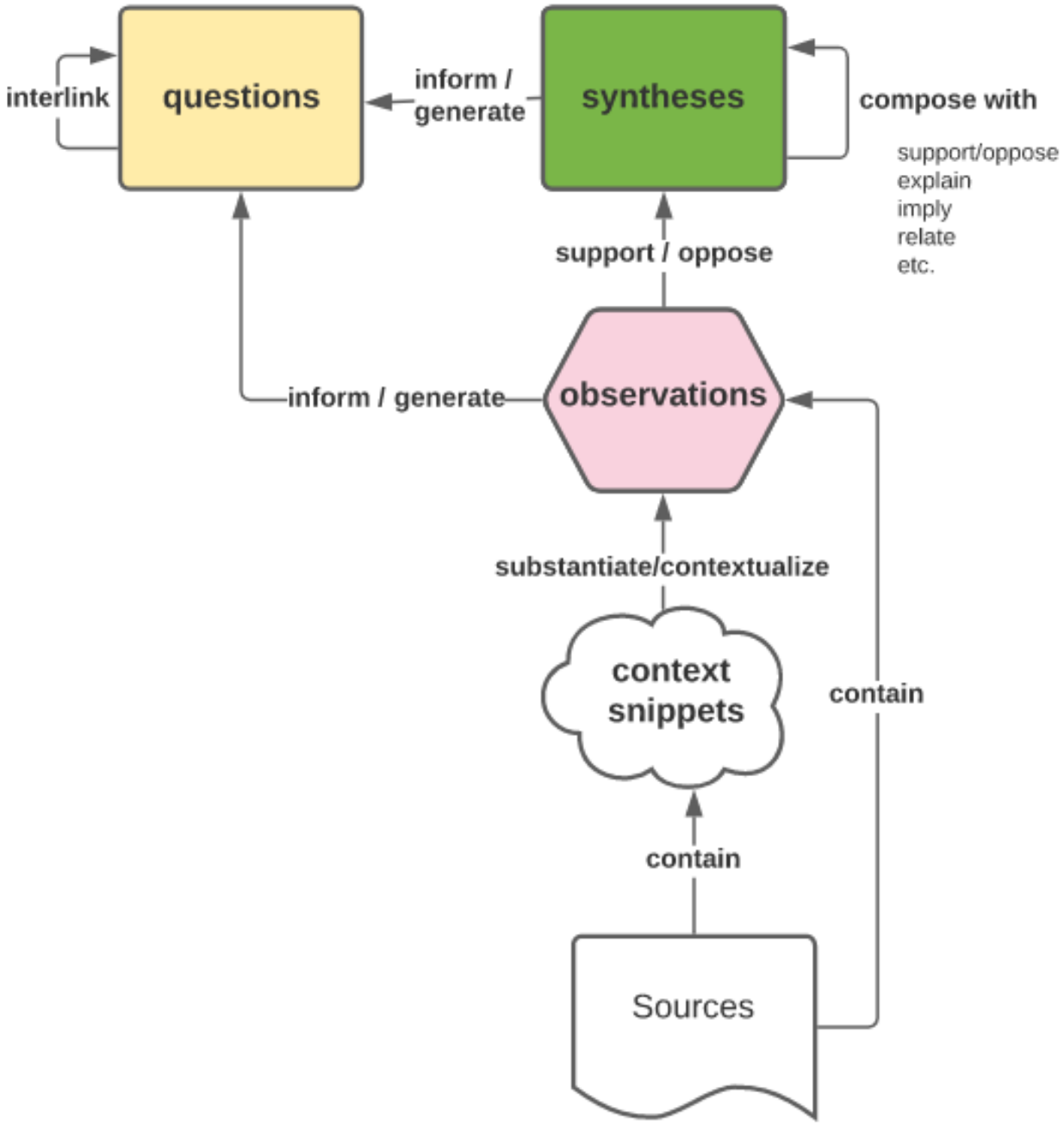
Context snippet notes capture (and optionally describe) contextual details that ground the observation notes.

Contextual details is a broad term, but generally includes things like specific figures, data items, tables, or quotes that are the basis for an observation, as well as metadata (e.g., authors, year, publication) and methodological details that are important for understanding and evaluating an observation note.

As a practical matter, I’ve found it more useful to use screenshots as context snippets, rather than plain text grabs. I find that this gives me the freedom to be a bit more sloppy and inclusive in the context of the quote (vs. very precisely specifying something), easier handling of images/figures/tables, and forces me to redescribe the context snippet, which enhances comprehension and recall. I also don’t need to waste time fixing up text (OCR mistakes, formatting, etc.)[1]!

Relationship between entities

Here is a visual diagram of the entities and how they relate to each other to form a system for synthesis.



There is a hierarchical relationship between the artifacts: **question** and **synthesis** notes (at the top "layer") are supported/opposed/informed by one or more **observation** notes (at the middle "layer"), which are substantiated/contextualized by one or more **context snippet** notes (at the bottom "layer").

Synthesis notes can also be composed into more complex structures (such as arguments or theories or models) through relations with other synthesis notes that

vary in complexity from simple "relates to", to implication/explanation and support/opposition.

From a practical standpoint, it’s probably most important to implement the typed distinction between entities (synthesis vs. observation vs. context snippet); typed distinctions between relations could significantly enhance the system's ability to augment human synthesis, but significant boosts in synthesis will likely accrue with implementation of only the three distinct artifacts (without explicit typed distinctions between relations).

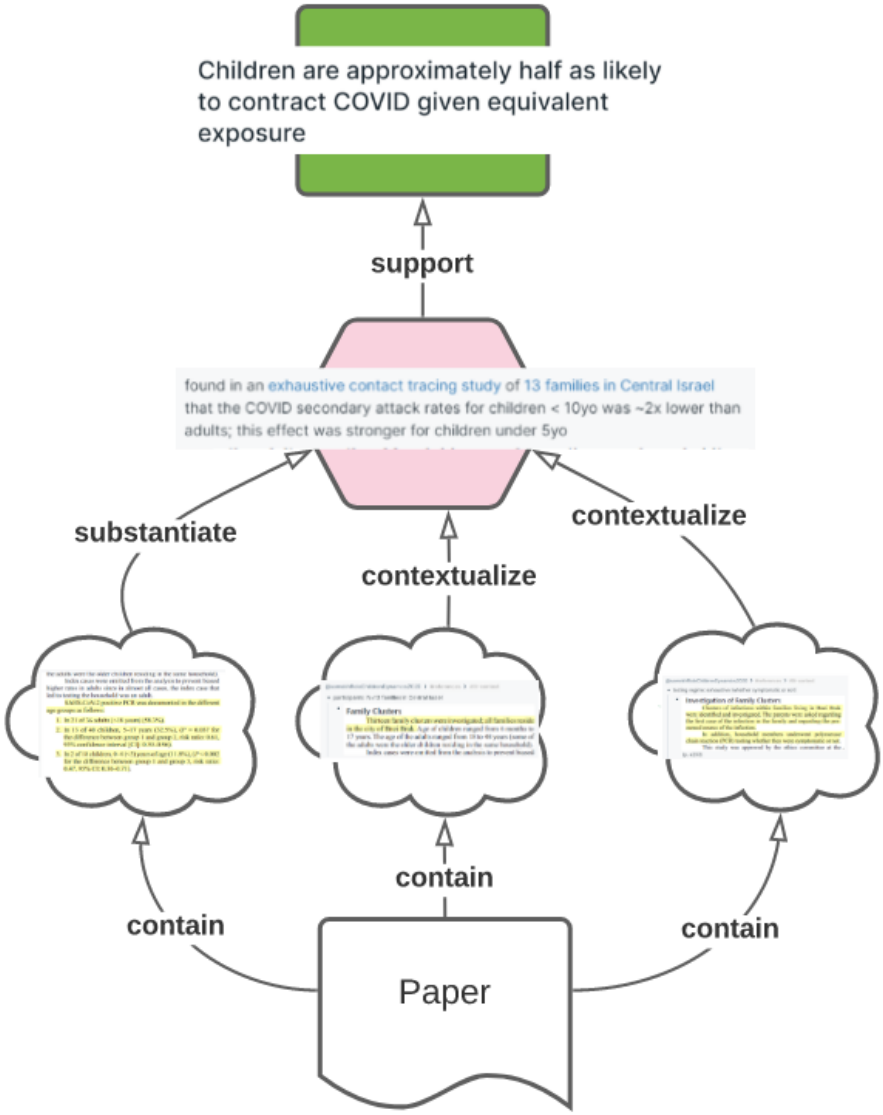
Thus, a minimal implementation of our model will include distinctions between the three types of entities, and explicit (but optionally typed) links between them.

Some examples

This discussion has been quite abstract. Let’s see how this conceptual model plays out with some examples. These are concrete, real examples from my own work (one personal, trying to figure out some decisions regarding COVID, which requires synthesis, and two from my own research projects).

Note: while synthesis notes should, whenever possible, be supported/opposed by multiple observation notes, for simplicity here I will only show a single thread through from context snippet to observation to question/synthesis.

Understanding COVID-19 transmission risks with children

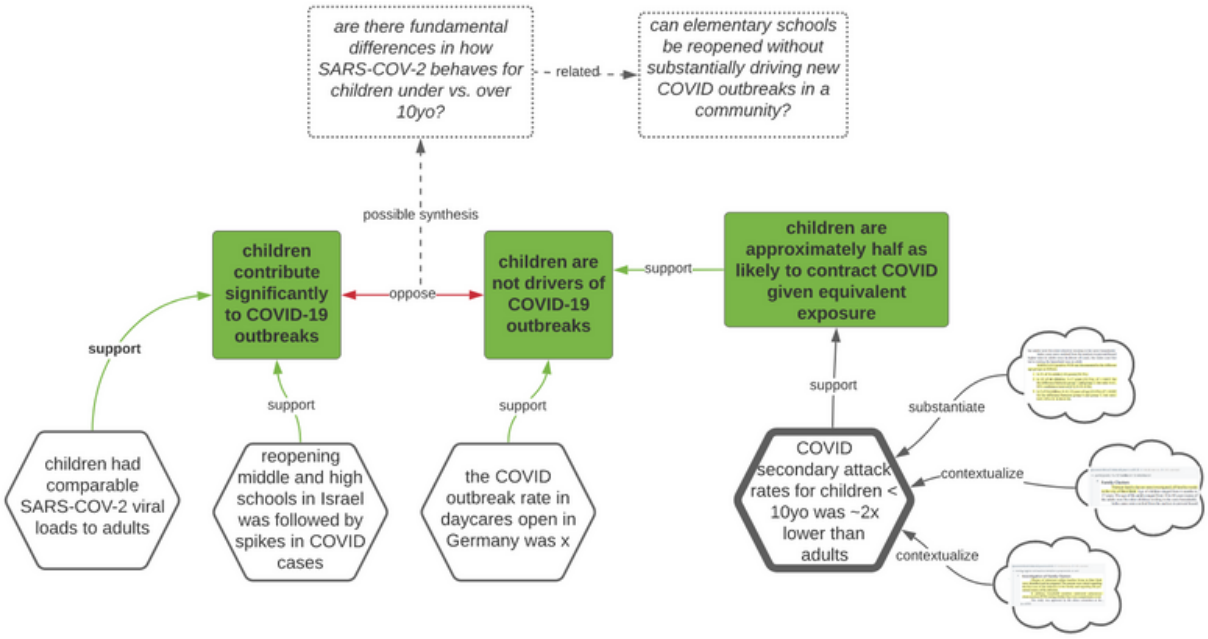


In this example, we have:

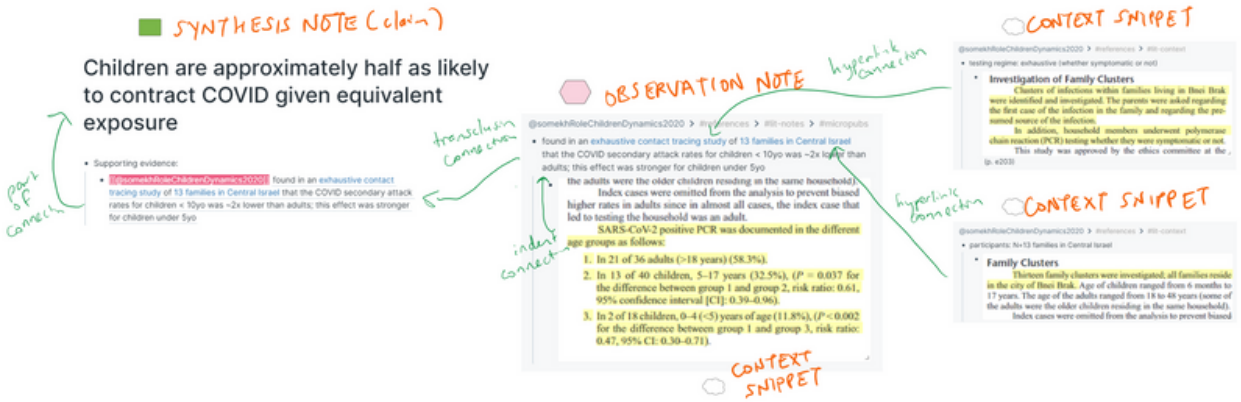
1. A **synthesis** note that *"Children are approximately half as likely to contract COVID given equivalent exposure"*
2. An **observation note** from a paper (Somekh et al., 2020) that *"in an exhaustive contact tracing study of 13 families in Central Israel, the COVID secondary attack rates for children < 10yo was ~2x lower than adults"*. This observation note supports the synthesis note.
3. Three **context snippets**, including a screenshot of the raw descriptive results and test statistics for the differences between age groups, one context snippet about the

exhaustive testing regime (i.e., regardless of symptoms), and the number of participants and setting. These context snippets ground the observation note, and are extracted from the paper's PDF.

This bundle of synthesis, observation, and context snippet notes is embedded in a larger network of synthesis/observation/context notes that is focused on understanding the risk of reopening schools for elementary school-aged children (a personal concern of mine!). The following figure demonstrates this. The focal synthesis note we just discussed is highlighted in bold.



Finally, let me illustrate how this is instantiated in RoamResearch, which provides many rich affordances for linking granular information items.



Here, the focal synthesis note is instantiated as a page.

The observation note about secondary attack rates is *transcluded* into (via block reference, and therefore a *part-of*) the synthesis page.

The context snippets are linked to the observation note via indentation in the outline as well as via hyperlinks to their block references, all of which create explicit bi-directional links in the underlying database between the items. Note here also how one of the context snippets includes a crucial detail that the testing regime was exhaustive (regardless of symptoms), which lends additional strength to the observation and how it might support a more informed synthesis claim about COVID transmission risks for children.

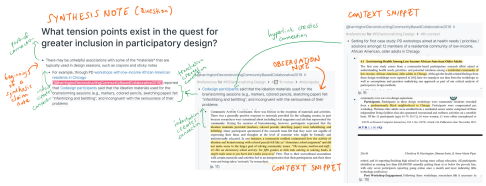
Both observation notes and context snippets are also *part-of* a page dedicated to the particular paper from which they came. In this way, other metadata such as the authors, institutions, year of publication, as well as high-level observations about the paper's context (e.g., number of citations, status as preprint or peer-reviewed) are also explicitly available as context for the observation note.

Understanding whether/how deep learning models of language might enable us to model analogical similarity

In this example, we have:

1. A **question** note, asking *"Can deep learning really discover analogical representations?".*
2. A **synthesis** note, that "*Vector-space models of language struggle with relational similarity",* which informs the question note.
3. An **observation note** from a paper (Schwartz, Reichart, & Rappoport, 2016) that a *"skip-gram model with vanilla BOW contexts performed ~50% worse on verbs compared to nouns and adjectives in terms of predicting human similarity judgment son SimLex999".* This observation note supports the synthesis note.
4. Three **context snippets**, including a screenshot of the raw descriptive results, one context snippet with details about the model, and one context snippet with details about the similarity judgments task. These context snippets ground the observation note, and are extracted from the paper's PDF. Note that the last observation is crucial for interpreting this result, since in technical terms, association and similarity are quite different (e.g., sit and stand are associated, but not similar), and tests of association may not reveal quite as stark of a difference in performance.

Understanding how the setup details of a participatory design influence how inclusive it is



In this example, we have:

1. A **question** note, asking *"What tension points exist in the quest for greater inclusion in participatory design?"*. Note that in the page here, there is the beginning of a synthesis note, to the effect of *"There may be unhelpful associations with some of the *materials* that are typically used in design sessions, such as crayons and sticky notes"*. However, I have refrained from making this a synthesis note proper until I see it show up in at least one other observation note, so I can write a sharper note. This choice also reflects the relatively early stage of this inquiry. More on this later when I talk about the process.
2. An **observation note** from a paper (Harrington, Erete, & Piper, 2019) that *"Codesign participants said that the ideation materials used for the brainstorming sessions (e.g., markers, colored pencils, sketching paper) felt "infantilizing and belittling" and incongruent with the seriousness of their problems"*. This observation note supports the draft synthesis note, but also stands by itself as a possible answer to the question note.
3. Two **context snippets**, including a screenshot of the quotes from the participants, and one context snippet with details about the co-design setting and participants. These context snippets ground the observation note, and are extracted from the paper's PDF. Here, both context snippets are especially crucial for me, since I am relatively new to this topic, and the core observation here depends on a *lot* of rich qualitative details that I would be foolish to lose.

This example illustrates how this approach can start to generalize from more quantitative empirical work. In later releases, I will discuss how other genres of research, such as formal modeling, simulations, philosophical arguments, and case studies, might also be instantiated in this model. My short answer for now is that they can.

Process

How does the model play out in my ***process*** of synthesis?

At a high level, I begin with

> a set of question notes and papers

and end with

> a network of synthesis notes grounded in observation notes that are themselves grounded in context snippet notes, and (usually) one or more new compelling question or synthesis notes that are not as well supported by observation notes and may be contradictory in some interesting ways. These indicate promising next steps for research.

There is forward progression in this model, but the process in between is iterative and nonlinear. It goes roughly as follows.

Phase 1: Articulate question notes.

Every project is aimed at one or more high level research questions. These questions are expressed as question notes.

These questions frames how I collect and process papers: every paper is considered or read with these key questions in mind[2].

Phase 2: Create observation notes from papers.

Next, I select and read sources (e.g., papers, books, early reports of data from colleagues) that have the potential to inform one or more questions of interest.

Reading will produce a variety of scratch notes and annotations, but should culminate in one or more observation notes that inform question notes.

These observation notes are grounded in at least one context snippet note, and explicitly linked to relevant question notes[3].

Phase 3: Develop synthesis notes.

As I begin to accumulate observation notes[4], I can then begin to articulate synthesis notes: what does the literature have to say about my questions of interest? Synthesis notes get explicitly linked to the relevant question note.

The process of developing synthesis notes is iterative with the previous phase. As interesting claims surface, I sometimes return to previous papers, or collect new papers, to stress test the ideas and find points of uncertainty. Revisiting papers and

collecting new ones in turn often spawns new observation notes. These new observation notes then may also in turn spawn yet more question or synthesis notes. The process of refining and juxtaposing synthesis notes may also spur refinement of observation notes (e.g., sharpening a description, adding context snippets that turn out to be important), or new question notes.

Phase 4: Compose synthesis notes into arguments or theories

The process culminates as I compose these synthesis notes into arguments or theories for my high-level question(s) of interest.

If done well, this process reveals further, sharpened question notes that lack satisfactory answers from the literature. Operationally, in this model, these question notes would be ones that cannot be traced to a sufficient critical mass of observation notes, or ones where there are multiple competing synthesis notes that require additional data to resolve. These would signal high-value directions to explore next to maximize knowledge gain.

Note also that this phase is likely to be iterative with the previous two steps: as a higher-level argument or theory begins to emerge, I will discover, or seek to discover, points of weakness or uncertainty, and dive back down to reconsider and refine synthesis and observation notes to further develop the argument or theory.

These compositions of synthesis notes can be encapsulated into complex synthesis notes (if I think I might want to reuse the whole package), or simply collated together in the body of a question note.

Frequently, these compositions will also inform the drafting of a research argument, components of a research proposal, or other shareable scholarly artifact, expressed externally (e.g., in a pubpub draft, LaTeX document, or otherwise). Alternatively, the composition could also form the basis of a contribution its own right, as a published review/synthesis paper, or theoretical paper.

Practical guide

I believe this model and process can be (at least in principle) implemented with three things:

1. A PDF reader
2. A reference manager (to make it easier to capture and manage metadata)

3. A networked notebook, such as RoamResearch, Obsidian, Tiddlywiki, RemNote, Notion, or any tool that implements bi-directional linking (some other pointers here).

As I note above, I am implementing this conceptual model and process in the software RoamResearch, with some support from Zotero, and regular PDF reader. In upcoming releases, I may share here how I do this in more detail[5].

But for now, here is a video of a synthesis session to get started, where I write observation notes from a paper, and use the resulting ideas to develop a synthesis note (in this case a question that was described above):

Visit the web version of this article to view interactive content.

Knowledge synthesis session in RoamResearch | 2020-11-28

What does this model buy us?

I want to test these claims more rigorously (that is in part why I'm sharing this document!), but here are some benefits that theory predicts and/or I've experienced personally. In later releases I will flesh out the conceptual and theoretical basis for this model more fully.

Effective synthesis

This model allows for rich layers of context to aid synthesis.

Distinguishing between observation notes and synthesis notes helps prevent me from rushing too quickly to generalizations, and allows for careful, nuanced questioning of past claims (e.g., does X really not work?), and consideration of possible syntheses between opposing claims. Directly including context snippets also allows me to have crucial details "on hand" that are necessary for this nuanced questioning.

In this way, the conceptual and process model helps mitigate the core challenge of lossy compression or premature ossification. Writing a synthesis note involves abstraction, which is a form of compression: removing details to generalize. If this is done in a way that breaks connections with the details (e.g., by writing a note without referencing even a page number, or even functionally breaking the reference by simply noting the bibliographic source), this compression is lossy. I believe compression that is more lossy or descriptions that are more reified are ok at much later stages of

knowledge production, where there is sufficiently high confidence in the articulation and certainty of those ideas. But I suspect this is rare when working on hard, creative, open-ended knowledge problems like in research!

So my belief is that knowledge synthesis is severely hampered by lossy compression. This relates to Strike and Posner's (1983) observation that an effective synthesis clarifies and resolves, rather than obscures, inconsistencies and tensions between material synthesized.

Reusability of ideas across barriers of time, people, projects, and disciplines

I believe this flexible compression not only helps synthesis right now, but also enables me to earn compound interest on the notes over time. One mechanism by which this happens is that the overhead for regaining context for my notes is reduced for my future self, and possibly for others as well, since the details are much more directly accessible through the three-part model. This is important, because the devil/diamond is in the details, and details fade over time from memory. I suspect that synthesis notes and systems that omit details (or at least make it hard to access details later), will have a much shorter half-life.

A less obvious benefit of retaining context is an increased capacity to notice points for intellectual progress, since anomalies and inconsistent results can often be a pointer to where a conceptual breakthrough is most needed (see, e.g., Kuhn's ideas about scientific revolutions). I can also question results more readily, and/or remix ideas for different settings.

I view the time I take to write notes in this structure in this way: instead of making one-off purchases, I am trying to amortize the cost of sensemaking by making investments that can pay off over time.

Note: For shorter-term or one-off cases, a lightweight version of what I describe here, like a synthesis matrix, is probably ok.

The ability to distribute the synthesis process

If I'm right that these sorts of notes are more shareable, then I should be able to distribute the process across a team of people. Hopefully this also means we get to substantially reduce the time needed to do effective synthesis. I am testing this hypothesis right now with my lab, and hope to get others to join me. Stay tuned on this!

Footnotes

1. Thank you to Darin Flynn for pointing this out! https://www.youtube.com/watch?v=uUF0XWk0bns&feature=youtu.be ↩

2. I might return to them later with different questions in mind, and get different things out of them! In this way, I never consider papers to be "processed" or "read" ↩

3. In a hypertext authoring environment like RoamResearch or TiddlyWiki or Obsidian, these links can be in-line in prose, rather than encapsulated in a section of links. I suspect the former is preferable to the latter, because it provides more context for links between notes. This is true also of links between claim_synthesis notes (see next phase) and question_synthesis notes. ↩

4. Where possible, I may also draw from past/other synthesis/observation note networks that could inform our questions. The intent is that over time, this Phases 2 and 3 get shortened as I accumulate critical mass of reusable observation and synthesis notes. ↩

5.
A video walkthrough + template might be more helpful for this!

If you are trying to implement this and would like more feedback or examples, I would be open to showing you around my Roam database and/or talking with you about issues you encounter. If this sounds useful to you, please email me at joelchan@umd.edu ↩

Citations

1. Knight, I. A., Wilson, M. L., Brailsford, D. F., & Milic-Frayling, N. (2019). Enslaved to the Trapped Data. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM. https://doi.org/10.1145/3295750.3298937 ↩
2. Blake, C., & Pratt, W. (2006). Collaborative information synthesis II: Recommendations for information systems to support synthesis activities. *Journal of the American Society for Information Science and Technology*, *57*(14), 1888–1895. https://doi.org/10.1002/asi.20486 ↩
3. Clark, T., Ciccarese, P. N., & Goble, C. A. (2014). Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical

communications. *Journal of Biomedical Semantics*, *5*(1), 28. https://doi.org/10.1186/2041-1480-5-28↩

4. Brush, M. H., Shefchek, K., & Haendel, M. (2016). SEPIO: A Semantic Model for the Integration and Analysis of Scientific Evidence. In *CEUR Workshop Proceedings* (p. 6). ↩

5. Somekh, E., Gleyzer, A., Heller, E., Lopian, M., Kashani-Ligumski, L., Czeiger, S., … Stein, M. (2020). The Role of Children in the Dynamics of Intra Family Coronavirus 2019 Spread in Densely Populated Area. *The Pediatric Infectious Disease Journal*, *39*(8), e202. https://doi.org/10.1097/INF.0000000000002783 ↩

6. Schwartz, R., Reichart, R., & Rappoport, A. (2016). Symmetric Patterns and Coordinations: Fast and Enhanced Representations of Verbs and Adjectives. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 499–505). San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1060 ↩

7. Harrington, C., Erete, S., & Piper, A. M. (2019). Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 216:1-216:25. https://doi.org/10.1145/3359318 ↩

8. Strike, K., & Posner, G. (1983). Types of synthesis and their criteria. ↩