

ISSN 2186-7437

## NII Shonan Meeting Report

No. 183

# Understanding the “Why” of Data and Knowledge Models

George Fletcher  
Marie Katsurai  
Juan F. Sequeda  
Hsiang-Yun Wu

September 16–20, 2024



National Institute of Informatics  
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

# Understanding the “Why” of Data and Knowledge Models

Organizers:

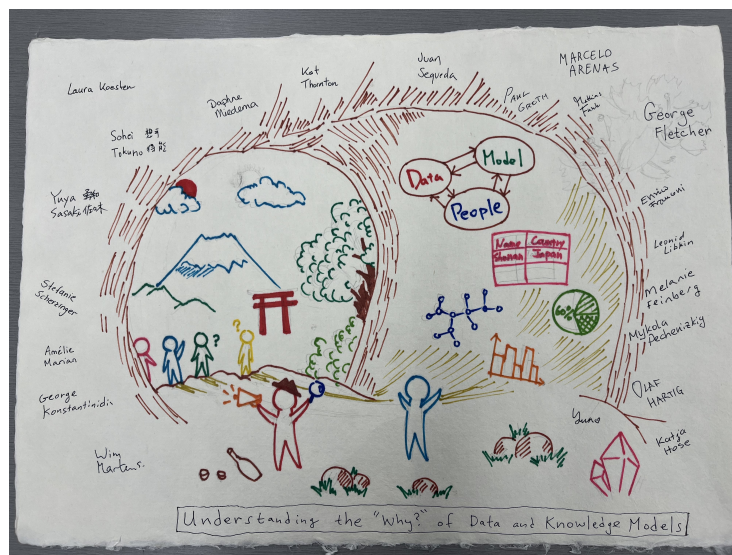
George Fletcher (TU Eindhoven, Netherlands)

Marie Katsurai (Doshisha University, Japan)

Juan F. Sequeda (data.world, USA)

Hsiang-Yun Wu (St. Pölten University of Applied Sciences, Austria)

September 16–20, 2024



## Background and introduction

The Wikidata Entry for the Seminar is at: <https://www.wikidata.org/wiki/Q130324785>

Data integration is the problem of 1) combining data residing in different sources, and 2) providing a unified view of this data. A modern manifestation of data integration is Knowledge Graphs which integrates not just data but also knowledge at scale in the form of a graph data model. This has gained wide popularity in academia and industry in areas ranging from search to mapping.

An argument to integrate data using graphs is that graphs bridge the conceptualization gap between how people think about data and how data is physically stored. One can see this for example when people end up drawing graphs on the

white board when describing data. Anecdotally, we often hear that a graph data model is a natural representation of data coming from heterogeneous sources.

However, if we look at the history of data models, we observe the following:

1. There has been numerous types of data models and corresponding query languages used for data integration. We can group these into three types of data models: tabular, graph, and tree each with many different flavors. The relational model continues to be dominant in practice, with additional light weight usage in the form of CSV widely used. Graph models such as RDF and Property graph data models have come to the fore. XML and JSON as tree data models are prevalent. Tailored query languages have been designed and engineered for each of these data models.
2. What goes around comes around. We have seen many data models come and go several times over the past 50 years. The first databases before the relational model were network (graph) and hierarchical (tree). During the 80s, Object Oriented Databases were common and foundational graph data models were developed [1].
3. We live increasingly in a data-driven world. Data and data analytics play a central role, not only in technical systems but also broadly in organizations and society, across government, academia, and industry and both in private and public spaces. A broad shift is underway, from data-centric analytics to human-centric analytics, where the emphasis is on understanding the role and impact of data in organizations and society, i.e., what do people do with data? and, what does data do to people?

Based on these observations and the current popularity of Knowledge Graphs for data integration, we believe it is the right time to reflect on the tripartite relationship between data models, their corresponding query languages, and the people both using and producing integrated data. Importantly, this reflection should drive us towards a more deeply empirical and social understanding of data models, not based in anecdotes but in data and methodologically rigorous investigation.

**Understanding the relationships between data models, query languages, and people** Thus, we argue that we need to understand how people perceive the way data is modeled and represented. In order to do so, we need to work with scientists and experts across communities to design methodologies, experiments and user studies. This requires bringing together data management expertise in theory, systems and semantics with communities who study people (e.g. human data interaction) and those who are actively using data.

Fundamentally, we need to embrace the role of human understanding in data modeling.

### Goals and outcomes of the meeting

- Build bridges between the data management, human computer interaction, knowledge engineering, and semantic web research communities and practitioners and users of data analytics technologies and the social sciences.

- Compile existing user studies and methodologies and take steps towards proposing new ones.
- Articulate a shared Vision Statement on open challenges, for peer-reviewed publication.
- Concrete action plans for collaborations in research and longer term international projects of broad ambition.



## List of Participants



Figure 1: Group Photo Shonan 183

- Marcelo Arenas, Pontifical Catholic University of Chile, Chile, and RelationalAI, USA
- Enrico Franconi, Free University of Bozen-Bolzano, Italy
- Daphne Miedema, University of Amsterdam, The Netherlands
- Stefanie Scherzinger, University of Passau, Germany
- Yuya Sasaki, Osaka University, Japan
- Mykola Pechenizkiy, Eindhoven University of Technology, The Netherlands
- Leonid Libkin, RelationalAI, USA, and University of Edinburgh, UK
- Olaf Hartig, Linköping University, Sweden
- Katherine Thornton, Yale University Library, USA
- Wim Martens, University of Bayreuth, Germany, and RelationalAI, USA
- Sohei Tokuno, Nara Institute of Science and Technology (NAIST), Japan
- Katja Hose, TU Wien, Austria
- George Konstantinidis, University of Southampton, UK
- Laura Koesten, University of Vienna, Austria
- Mathias Funk, Eindhoven University of Technology, The Netherlands

- Melanie Feinberg, University of North Carolina at Chapel Hill, USA
- Amelie Marian, Rutgers University, USA
- George Fletcher, Eindhoven University of Technology, The Netherlands
- Paul Groth, University of Amsterdam, The Netherlands
- Juan Sequeda, data.world, USA
- Hsiang-Yun Wu, St. Pölten University of Applied Sciences, Austria

## Meeting Schedule

### Check-in Day: September 15 (Sun)

- Welcome Banquet

### Day 1: September 16 (Mon)

- Introduction to the workshop
- Introductions by participants
- Collection of research questions

### Day 2: September 17 (Tue)

- Definition of Experiments and Studies
- Group Photo Shoot

### Day 3: September 18 (Wed)

- Reflection
- Excursion and Main Banquet

### Day 4: September 19 (Thu)

- Consolidating Work

### Day 5: September 20 (Fri)

- Next steps

## Summary of discussions

Driven by the need for users to make effective use of their data, we have identified three key areas in this seminar that require further research: exploratory queries, genres and genre conventions in data modeling, and the effective use of data. In this section, we highlight some research questions for each of these areas, emphasizing the need for collaboration between researchers focused on the technical aspects of data management and those working on the social aspects of data consumption. Specifically, in Section 1, we examine how the growing complexity of query languages and data architectures is transforming the way users design, debug, and maintain queries, particularly with the rise of generative AI. In Section 2, we introduce the concept of genre conventions in data models, suggesting that patterns in data modeling practices can be understood through a communicative framework. Finally, in Section 3, we reflect on the technical and social challenges that users face in effectively using data, highlighting the importance of revisiting data and knowledge work from both technical and social perspectives.

### 1 Exploratory Querying

The modern data landscape is filled with a wide range of data models and query languages, each becoming increasingly complex. The SQL standard alone is comprised of sixteen parts, of which only two—Framework and Foundations—are typically covered in university curricula. The total length of the standard has now grown to more than three thousand pages. The initial simplicity of declarative query languages has been overshadowed by the rising complexity and expressiveness, making some query languages Turing-complete, while adding numerous advanced features. Concurrently, data architectures have shifted from closed, controlled environments to open, cloud-based systems, where development and testing are much closer to production. This transition increases the cost and complexity of testing and development, as users no longer benefit from affordable and comprehensive access to data or a local, secure testing environment. The rise of diverse data models also introduces use cases where the schema is either highly intricate, undefined, or non-existent, such as in graph databases or RDF.

Moreover, the user base for databases has become increasingly varied. Differences in culture, socio-economic status, profession, and technical expertise can impact how users build queries, especially in interpreting intermediate results. Public interfaces like Wikidata, for example, are accessed by a growing number of users who have little or no formal training in query languages or domain-specific knowledge. As a result, the process of query construction must accommodate varying skill levels, especially given the shortage of highly-skilled data and software engineers in the labor market. Many users may gravitate toward specific data models or query languages based on their personal preferences or background, rather than selecting the most suitable tool for the task.

Recent developments in generative AI, backed by significant industry investment, are changing how people interact with data. These systems are being integrated into every stage of the data lifecycle, from formulating intent to crafting queries, acquiring domain knowledge, and even debugging. The tradi-

tional landscape of software-generated queries, previously driven by applications or web forms, is now undergoing a major transformation due to the arrival of these groundbreaking technologies.

As a result, query development is becoming a more exploratory process, where users engage in iterative cycles to refine their queries. We can identify three key activities within this exploratory query lifecycle: design, debugging, and maintenance. This lifecycle is framed around exploratory querying, acknowledging that much of data management research has focused on fixed queries rather than those generated through an iterative process.

As more users engage in exploratory queries, they could benefit from additional support tailored to this type of activity. To address this, we propose a research agenda that includes experimental methods and research directions based on the following questions:

- How do users create queries when they are familiar or unfamiliar with the schema or language?
- How do users approach designing, debugging, and maintaining queries?
- What kind of HCI support can enhance this process?
- How can query engines better support exploratory querying?
- What language support is needed?
- How do large language models (LLMs) influence exploratory queries?

Each of these questions can be applied to the design, debugging, and maintenance phases for different query languages. To advance this research, we aim to collaborate on a vision paper to be submitted to VLDB, calling for more research into the area of exploratory querying.

## 2 Genres and genre conventions in data modeling

Genre is a familiar concept from everyday life. For instance, many of us will immediately understand what is meant by the “detective noir” genre of fiction. In recognizing this genre, we will be able to identify various recurring patterns, or conventions, that occur in detective noir novels. One example is the figure of the lone investigator, such as Marlowe in Raymond Chandler’s *The Big Sleep* or Sam Spade in Dashiell Hammett’s *The Maltese Falcon*. Often, the lone investigator in a noir novel has a shadowy past that informs their cynical perspective on the present, but the details of that past remain hidden.

The conventions that shape a genre are not arbitrary stylistic flourishes; these conventions have a communicative purpose. In *detective noir*, the convention of the cynical, troubled lone investigator helps to create a dark (“noir”) mood, one that exposes the mayhem of base desires that lurks just beneath the apparent orderliness of human society. The figure of the lone investigator spurs the reader, likewise, to investigate: to look beyond the surface of everyday life, examining the chaotic undercurrents that may exist in all our realities.

An established tradition of genre studies exists within the field of composition and rhetoric (Bawarshi and Rieff, 2010). Genre, here, provides a mechanism to connect recurring regularities in expressive works (such as the lone investigator of noir fiction) to the communicative purposes that such regularities serve (such as persuading the reader that the orderliness of society is but a mirage). In an influential early article, for instance, Carolyn Miller (1984) defines genre as a formal response, implemented in a particular medium, to a specific and recurrent social situation. But scholars of genre studies are more likely to examine everyday sorts of texts rather than artistic works.

In our work, we are trying to understand the work of data in contemporary information systems (Feinberg 2022). Currently, there is no clear framework to make sense of the evolving relations between people, models and data. In particular, while there is a rich literature on many technical aspects of data models going back over the last half century and beyond, we do not have a good understanding of the social life of data modeling, an understanding necessary for building more people-centered futures together. We argue that looking at data modeling from the perspective of genre conventions might help us see pathways through this situation. Indeed, data models might seem like technical constructs, but they also serve communicative functions. Can the concept of genre, then, help us to understand the communicative work that data models perform? This is the motivating question that we hope to investigate here.

To establish whether genre provides a reasonable framework from which to examine data models and data modeling, one approach is to identify potential genre conventions, or recurring patterns (such as the lone investigator in noir novels), that seem to emerge in response to certain conditions in data modeling practice. If we can identify such examples, that will corroborate our intuition that examining data models through the lens of genre is a worthwhile enterprise.

## 2.1 Genre Conventions in Data Models

Anecdotal evidence suggests that genre conventions occur in data modeling practice. Here, we highlight four examples.

- Metadata in Excel workbooks. In many Excel workbooks, the column names are moved down, and a series of rows at the top describe metadata or information about the Excel workbook.
- English language labels in IDs: When publishing schemas on the Web, URLs serve as identifiers. In many schemas, the end part of the url (its slug) is an english language term. For example, the identifier for the type creative work within the schema.org schema is <https://schema.org/CreativeWork>.
- Inserting JSON into columns in relational structures.
- Log data. When logging system outputs, a practice well-known to programmers and software engineers from beginner to expert, certain conventions tend to emerge, such as including a timestamp in a standard format, using delimiters, a recognizable order of elements, and placing each piece of data on its own line

These examples illustrate the emergence of bottom-up conventions in data modeling practice, shaped by practical needs and common use cases. Each of

these conventions serves specific purposes, whether it's organizing metadata, creating meaningful identifiers, storing complex data formats like JSON, or managing system logs. The examples also show that, in our first conceptualization of genres in data modeling, recurring patterns emerge at various levels of abstraction and scope: from small-scale IDs to an entire category of log data. Are these individual patterns initial evidence of potential data modeling genres? This is a question for future research.

## 2.2 Research Opportunities

The genre perspective on data, conceptual, and knowledge modeling raises many new questions. One area to examine is the concept of genre conventions in modeling. For example, do genre conventions exist in data modeling? Can we identify distinct genres? Additionally, there are broad research questions surrounding the perception and understanding of genres in individual and group work. For instance, how do individuals comprehend specific data and conceptual modeling genres and their formation and evolution? In addition to understanding perception and prevalence, the impact of genre conventions on practice deserves attention. For example, how does understanding data models create common practices?

Building on the anecdotal evidence that genre conventions provide a productive lens for investigating data and conceptual modeling, we now consider research programs that could help us gain deeper insights and evidence into these conventions.

**Model Corpora** One major study is to build corpora that consist of instances or manifestations of a genre. For example, one could examine corporate data models to explore genres by collecting examples of data models in the wild. Another potential corpus for study is Wikidata schemas, particularly focusing on SHACL (ShEx) shapes used in the Wikidata community

**Ethnographic studies of modeling** Another complementary approach we could take is applying research designs from the social sciences to study modeling in practice. Conceptual modeling can be viewed as an unrepeatability activity, where at each instant, the model is being shaped by a specific group of people in a particular context. Using an ethnographic lens, borrowed from anthropology, would allow us to capture the embedded, contextual, and evolving nature of genre conventions as responses to recurring social situations.

**Lab-based studies of modeling** In addition to ethnographic studies, we also propose to study genre conventions in lab-based studies, to see whether they surface in the way we create conceptual models of data. The aim is to gain a better understanding of to what extent these conventions shape the way people think about data modeling and how they impact the way that existing data modeling methodologies and practices are understood.

## 2.3 Implications

We believe that examining data models and data modeling through the lens of genre will promote a richer understanding of the continually evolving, complex

relationships between people, models, and data.

### 3 A Reflection on the Effective Use of Data

Data drives modern society, influencing decisions in almost every sector from academia to industry. Developing infrastructure for the effective use of data is a collective objective shared by a range of communities, including data management, information science, data science, semantic web, machine learning, business intelligence, and digital humanities, among others. However, despite decades of investment in research and development and considerable progress, users continue to face significant challenges in harnessing data to its fullest potential.

The purpose of this paper is to reflect on the challenges that remain in achieving this objective, with a particular focus on data and knowledge work as a socio-technical endeavor.

#### 3.1 The Objective: Effective Use of Data

At its core, the effective use of data refers to enabling users to extract meaningful value from data for their tasks. However, the term “effective” varies depending on the context and user needs. It is clear that a one-size-fits-all approach does not work—effectiveness is highly contextual, and what may be effective in one scenario might not be in another.

In the last half century, within modern computation, we have invested heavily in developing an infrastructure for the effective use of data. A non exhaustive list of investments have been:

- Technical Infrastructure
  - Scalable data architecture
  - Cloud Computing
  - Data integration and interoperability
  - Data quality and cleaning
  - Real-Time Data and Streaming
- AI/Automation
  - Machine Learning (ML)
  - Knowledge Representation
  - Semantic Technologies
- Usability
  - Human-Computer Interaction (HCI)
  - Data Visualization
- Data Standardization and Metadata Management
- Data Governance, Privacy, and Security

- Data literacy, culture and training programs

While there has been substantial progress, the fact that users still struggle to effectively use data underscores the need for further exploration. We ask ourselves: to what extent have we actually accomplished our objective? Are we on the right path but simply in need of more time and resources? Or is there something fundamentally new that we have yet to realize?

### 3.2 The Socio-Technical Divide

We postulate that this disconnect is often the result of examining problems through purely technical or purely social lenses, when in reality, a combination of the two is necessary.

On the technical side, we focus on tasks like data cleaning, transformation, standardization, and mapping. These tasks are often approached from an objective standpoint, with clearly defined metrics for success, such as processing time or storage efficiency. The improvements in technical infrastructure and algorithms have led to impressive gains in these areas.

On the social side, the focus shifts to human interaction with data. Here, the challenges are more subjective. For instance, we often discuss “better user experience” without fully defining what “better” means for different users. Engaging users in exploratory queries, gathering requirements from stakeholders, or fostering collaboration between technical and non-technical teams are vital tasks, but they are often undervalued because they resist quantification.

Our position is that these two streams should not be viewed in isolation. We believe that the tasks that fall into the middle ground, which we call *data and knowledge work*, require a socio-technical perspective that takes into account both the tools and the people involved. A key challenge to enable the effective use of data is bridging this gap.

### 3.3 Data and Knowledge Work

We consider data and knowledge work to be tasks that fall into two categories: technical tasks which are those that rely on tools, algorithms, and technical expertise, and social tasks which are driven by human interaction and non-technical expertise. Based on further reflection, we present a non-exhaustive list of these tasks.

- **Technical Tasks:**

- *Data Preparation and Cleaning*: Fixing errors, handling missing values, etc.
- *Data Transformation*: Reformatting or restructuring data to make it usable in different systems or for analysis.
- *Data Quality*: Ensuring and checking accuracy, completeness, and compliance against rules and standards.
- *Data Standardization*: Ensuring data follows consistent formats and conventions across systems.
- *Semantic Mapping*: Aligning schemas, vocabularies, and taxonomies to ensure shared understanding across systems.

- *Information Extraction*: Identifying and pulling structured information from unstructured or semi-structured data sources.
- *Conceptual/Ontology Modeling*: Creating abstract representations of how knowledge fits together.

- **Social Tasks:**

- *Collaborating with Subject Matter Experts*: Gathering domain-specific knowledge to understand the data.
- *Data Governance*: Ensuring data meets organizational policies and standards (e.g., access, quality, ethics, privacy).
- *Data Stewardship*: Assigning responsibility for managing policies within an organization.
- *Reaching Agreements*: Aligning terminology, models, and rules between stakeholders to resolve ambiguities and conflicts.

As shown in the previous list, what characterizes these tasks is their reliance on context, collaboration, and negotiation, making it difficult to apply purely technical solutions without considering social factors:

- Contextual awareness: understanding and applying domain-specific knowledge so data can be interpreted effectively
- Collaboration and communication: working with people to define, refine, align, diagnose, resolve
- Conflict resolution: Resolving ambiguities and inconsistencies which requires negotiation and alignment
- Balancing automation and human input: Many tasks can be automated, but human judgment is still critical for interpreting, validating, or refining results.
- Emphasis on organizing/classifying: defining structure through relationships between entities, attributes, and concepts
- Integration across systems: integrating disparate data sources and ensuring interoperability between systems and domains

### 3.4 A Call to Revisit the Socio-Technical Approach

Our reflection leads us to advocate for revisiting both technical and social tasks from a socio-technical perspective. Specifically:

- **Revisit Technical Tasks with a Social Perspective:** For example, when designing conceptual models, we should consider how these tasks impact and are understood by users. Are users able to fully leverage the tools we provide, and do they understand the implications of these technical decisions?
- **Revisit Social Tasks with a Technical Perspective:** Likewise, social tasks such as data governance or policy development should be informed by technical insights. Can technical tools be leveraged to automate or support these processes, and how can they be adapted to fit social needs?

### **3.5 Call for Action**

We have outlined the need for a socio-technical approach to data and knowledge work in order to achieve our goal of developing infrastructure for the effective use of data by users. While much progress has been made in both technical and social dimensions, the gap between the two remains a significant barrier to achieving truly effective data use. By revisiting technical tasks through a social lens and vice versa, we believe we can make meaningful strides toward empowering users and enabling the effective use of data.

Our call to action is clear: we must engage in interdisciplinary research, develop new frameworks that integrate social and technical perspectives, and design tools that acknowledge the complexities of data and knowledge work. Only through these efforts can we close the gap and fully realize the potential of data for all users.

## Summary of new findings and next steps

During the seminar, the foundations for at least three papers were produced. Additionally, over 130 research questions were identified as listed in the appendix. Overall, this provides input to the next steps after the seminar. The main directions include

- finalizing the articles started during the week;
- continued collaborations on the studies identified; and,
- community building activities.

## Acknowledgement

We would like to acknowledge Professor Paul Groth (University of Amsterdam, Netherlands) for his valuable organizational support and assistance before and throughout the seminar.

## References

- [1] R. Angles and C. Gutierrez. Survey of graph database models. *ACM Computing Surveys*, 40(1), feb 2008. doi: 10.1145/1322432.1322433

## A Collected Research Questions

- Why isn't knowledge work done properly?
  - Knowledge work involves tasks like conceptual design and interviews with subject matter experts.
  - Why aren't organizations effectively modeling things?
  - Do people understand the consequences of improper conceptual modeling?
  - How can we help them do things correctly, efficiently, and quickly while meeting deadlines?
  - Does caring about conceptual modeling make a difference?
- Why don't people care?
  - They design, build software or data, and it evolves quickly, becoming disconnected.
  - Doing this properly may be seen as dogma, and there is no clear evidence that it's beneficial—at best, the evidence is anecdotal.
- Is there a correlation between proper knowledge work and mission-critical tasks?
  - Is it only worth doing proper knowledge work for critical tasks?
  - What is the value of knowledge work?
  - Do organizations find that investing in knowledge work provides enough value to justify the cost?
  - If designers or builders don't use proper conceptual models, does that mean they don't find value in them?
  - We teach these concepts—so what is happening?
  - What is the value of a conceptual model?
  - How do you measure this value?
  - How do we communicate value that might not be realized until years later?
  - How can we convince or incentivize people to do knowledge work?

- Why do we fail to convince people to do knowledge work?
- How can we convince people to do knowledge work?
- How can we weigh the positives and negatives of a solution?
- Why do we fail to convince people to adopt a new model?
- How do we explain the pros, cons, and implications of choosing different models?
- What are the situations or tipping points that trigger people to realize the need to invest in knowledge work?
- What is the minimal value required for people to invest in knowledge work?
- How is knowledge work measured?
  - How do you measure the value of a conceptual model?
  - How do you measure the naturalness of a (data) model? What about a query language?
  - What are the key things to measure?
  - How do you assess the effectiveness of knowledge work? Is it useful beyond just financial measures?
  - How do we know if a conceptual model or ontology is improving or deteriorating?
  - What are the criteria for better ontology? Does it depend on the user or the query?
- Do we have the correct educational programs for knowledge work?
  - How well are we teaching knowledge work?
  - If designers or builders aren't using proper conceptual models, does that indicate they don't see the value in it?
  - What is the ideal skill set for knowledge work?
  - Where should this training come from—Computer Science or Information School?
  - If leadership doesn't advocate for investment in knowledge work, it won't happen. So, where are we educating leadership?
- What are knowledge work methodologies?
  - What are the established methodologies? Do they require a long chain of elements?
  - How do we reverse engineer, from the mess that exists?
  - How do we explain the pros and cons and the implications of selecting different models?
  - What are the friction points in the process to build data systems?
  - Is this a disaster due to agile? Do something quickly, check.
  - How much is agile affecting the proper design?

- What are the tasks that need to be accomplished? What do users need for the task?
- When should you push the knowledge work to the left, where things start, or live with that and move it after, to the right?
- How can we make this scale? Is it even possible?
- What if the right person is not in the room?
- Do people have a shared agreement? Do they know? What are they assuming?
- How do you confirm that there is a shared agreement? How do you realize they don't? And how do you make a decision on which one? Who makes these decisions?
- How do you manage disagreements?
- What is the relationship between query and the conceptual design?
  - Queries are a mess/complicated due to poor conceptual design.
  - Once you have a conceptual model, how do you know which data model to use? What is the systematic way to decide? Does it depend on the queries/questions?
  - Some query languages are more interesting to be used to users?
  - Is the reason to use a data model because of the query language? Does a user like it? Are they able to express things more easily?
  - What is naturalness? “This is one that I already know, that I feel comfortable with.”
  - Naturalness depends on how the data model is used.
  - Is the naturalness applied more to vocabulary than the data model?
- In which cases do knowledge graphs catch on?
  - Cost/reward: What are the costs and rewards associated with using a knowledge graph?
  - Completeness/representation of reality: Is there a continuous source of data such that the graph must be amended?
  - Understanding, mental models, semantic layer: What are appropriate mental models (notional machines) for understanding data models / knowledge graphs?
  - What forms the minimal semantic layer to facilitate laypersons' understanding of knowledge graphs? How can we design for supporting appropriate mental models of a KG?
  - (Dis)trust: What properties make users trust/distrust knowledge graphs?
  - Support: How can we facilitate usage of knowledge graphs? How can we show how much the data model / the KG reflects reality?
- Questioning Database 101 assumptions:
  - How has a representationalist dogma constrained our work?

- How does a representational view inform or constrain our understanding of data quality (clean vs dirty)? Of structured data vs unstructured data?
- What would data models look like if designed using a pluralist view of the world? E.g., using a relational view.
- How can methodologies of data critique and interpretation complement and coexist with data engineering?
- What are the running competitors of the classical representational view of data modeling?
- The evolving relations between people, models, and data.
- How do people (re)construct conceptual models when using a database instance, even when given the “correct” conceptual model?
  - Experiment: Give the people a database and ask them to say aloud.
  - How have we (not yet) addressed the temporal and plural instability of the connections/mappings/interactions between the unstable “layers” of an information system?
    - \* People using and impacted by the information system: We inhabit multiple “roles” at any given point in time. We have different roles interacting with the system over time. We have different “publics” /”stakeholders” at any given point in time and at different points in time.
    - \* Conceptual model of the information system: Conceptual instability, conceptual plurality/heterogeneity.
    - \* Database instances of the information system: E.g., in the presence of (de)normalized schemas, hacked schemas.
    - \* Interaction/interplay of conceptual models and interaction paradigms: Visual, aural, olfactory, tactile. What is the smell of an ER diagram? What does an ER diagram sound like?
- The impact of an influential model on subsequent models:
  - E.g., “wheels” to schematize the taste and smell of wine, and the subsequent transposition of this structure to other domains (coffee, chocolate).
  - What are the affordances and effectiveness of different notations and syntaxes?
  - Can we determine the flexibility of data models? What do we mean by flexibility?
- Understanding practice:
  - What are the uses of the word “user” in literature and practice?
  - What are the uses of the phrase “real world” in literature?
  - How do users “actually” use data systems?
  - Can we create a database and understand the characteristics of “homebrew conceptual models”? E.g., conceptual models done on whiteboards.

- Is data reuse a net good? How can we measure this?
- How do we scale empirical studies of data in the wild?
- To what extent can we provide evidence for a generic set of tasks in data work?
- What is the relationship between data models and representation learning?
- Situatedness of data and models: How do community data/knowledge resources propagate in the world? E.g., wordnet, wikipedia categories. Do such resources constrain or facilitate futures?
- To what extent is the design of data models influenced by external incentives? For example, are ML annotation schemes influenced by the motivation to get higher F1 scores?
- Side effects of choices in data modeling:
  - What is unstructured vs structured data?
  - What makes data “unstructured”?
  - What is the process of structuring?
  - What is clean vs. dirty data? What makes data “clean”?
  - How can we maintain and communicate context and uncertainty when datafying (also visualizing)?
- What is the gap between what is being taught about knowledge work and what industry requires?
  - How do universities teach conceptual modeling?
  - How are modeling skills used in capstone projects?
  - How do bootcamp-like trainings teach conceptual modeling?
  - What knowledge work-related skills do job openings show?
  - What gaps do industry professionals observe with respect to conceptual modeling skills?
- Exploring Exploratory Queries, Core Research Questions:
  - Research Questions on Queries in Today’s Systems in General:
    - \* Traditionally, a large share of queries is generated programmatically. Nowadays, what is the share of queries that are generated programmatically (even by LLMs) and that is hand-crafted by humans?
    - \* How can we distinguish generated and crafted queries?
    - \* Is there a difference, and if so, should we cater to both kinds of queries?
    - \* How are LLM generated queries different from human or software-generated queries?

- \* Given a query, can we formalize its intent? Intent will differ along a query's progress through the stack, from describing a high-level task to a low-level query. Do we need a language for capturing query intent? Might this be logic-enhanced NL? Given generated/hand-crafted queries, can we extract their intent?
  - \* Do developers have different ways of thinking about writing queries, such as a declarative versus a procedural "brain"? Can we classify query authors by their style of thinking? Do we need to custom-design query languages that cater to these styles of thinking, such as PRQL versus SQL?
- Research Questions on Exploratory Querying:
- \* How do users query data when they know/don't know schema/language?
  - \* What HCI support is needed?
  - \* How can a query engine support exploratory queries?
  - \* What language support is needed?
  - \* How do LLMs impact exploratory queries?