A Knowledge Graph Foundation for Real-World Data Sourcing, Lineage, and Evidence Traceability

Johnson&Johnson
Innovative Medicine

Real World Data Autumn Event 2026

30 September 2026

Zhimin (Jimmy) Liu
R&D Data, Data Science and AI
Johnson & Johnson

Agenda

1. **Why RWD sourcing needs a different foundation**
   - The decisions we make, and what is lost when we make them
2. **What we built**
   - An AI-assisted landscaping pipeline feeding a knowledge graph
3. **The data model and design principles**
   - Entities, relationships, provenance, versioning
4. **Traceability in practice**
   - From research question to dataset to published evidence
5. **Building on the foundation**
   - Where a graph takes RWE next, and what we learned doing it

Why RWD sourcing needs a different foundation？

Selecting fit-for-purpose data is still a bespoke exercise

Every study team re-solves the same problem, largely from scratch

How it runs today

Ad-hoc scan of familiar datasets, plus colleague memory

Real evidence sits in papers' methods sections, not a catalog

Landscaping is outsourced or hand-built, and takes months

Assessments land in a deck, then stop being findable

Next team, next question: start again

What it costs us

Better-fit data doesn’t surface on time

Feasibility is assessed after the protocol is written

The same dataset is re-assessed for each project

Provenance and selection rationale are hard to reconstruct

Transparency is met by narrative, not by record

Catalogs capture metadata. They do not capture relationships

Where the information we most need for reuse is lost

A catalog: rows of datasets and attributes

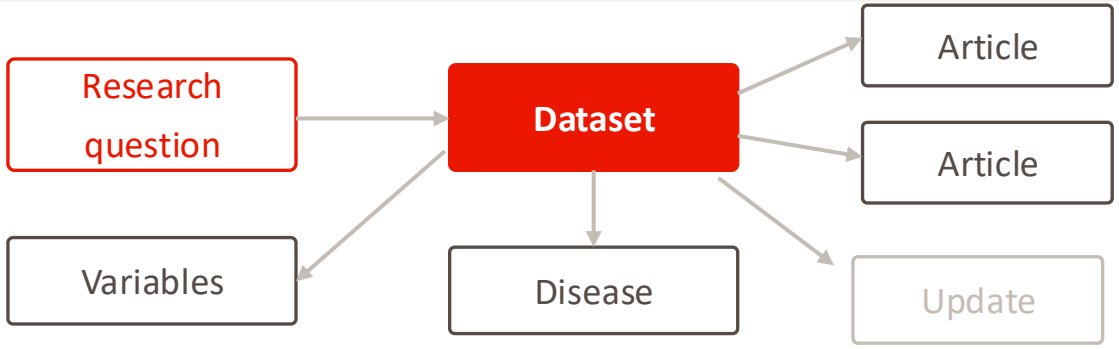
Dataset A	EHR · region · n = … · variables
Dataset B	Claims · region · n = … · variables
Dataset C	Registry · region · n = … · variables

Answers what exists, and what is in it.

Cannot answer:

- Which evidence came out of this dataset?
- Which question was it selected for, and why?
- What changed since we last looked?

A graph: datasets, evidence and context, connected



Answers the same questions - plus the ones that matter for reuse: this dataset, in this disease context, produced this evidence, was assessed for this question, and was last refreshed here. Relationships are the asset - and exactly what a flat catalog throws away.

Where sourcing sits in the evidence generation cycle

Decisions are made at every step - and rarely documented



The traceability gap

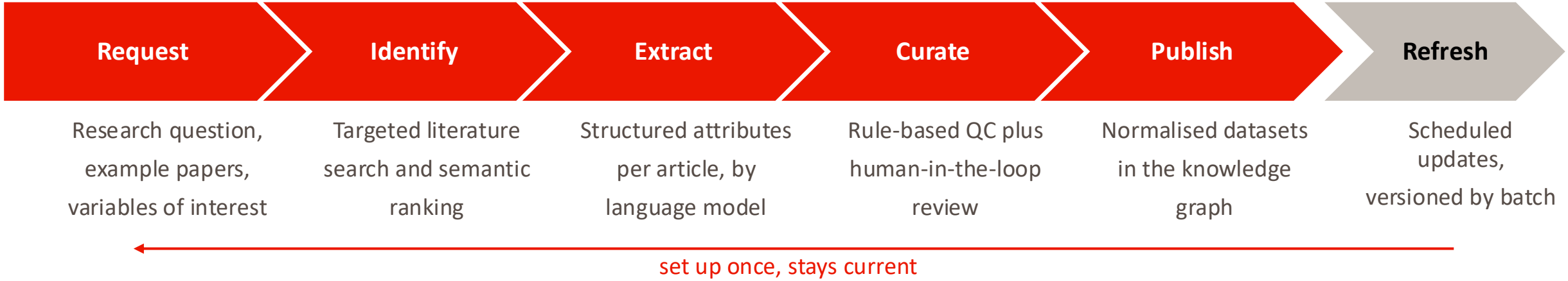
Every cycle generates real intelligence - sources considered, sources rejected and why, what feasibility showed, what evidence came out. Today it lives in reports, inboxes and memory.

Captured as connected data, the next study starts where the last finished, and the audit trail is a by-product.

What we built？

From a research question to a living data landscape

One request sets up a pipeline that keeps running



A rapid landscape scan

The candidate universe for a disease area, built from published literature in weeks, not months.

A question-specific shortlist

Sources ranked against the actual question, with supporting publications attached.

Documentation as a by-product

Request, search, extraction, curation and output in one reproducible structure.

Step 1 - Identify the universe of data sources from literature

Published studies are the most complete record of which datasets are usable in practice

Design choices

Two retrieval routes

- Keyword query for recall; semantic similarity to example papers for precision

Rank, then cut at a threshold

- Extraction runs only on candidates above a relevance threshold - reading every paper with a language model is very expensive at this scale

The article is the unit of retrieval

- A dataset earns its place because someone published research with it

Designed for repetition

- The same configuration re-runs on a schedule, processing only what is new

Why literature first

A published dataset has been through peer review of its fitness for a purpose - the methods section gives population, variables, follow-up and caveats.

Vendor material says what a dataset contains. Publications say what it has actually supported, how often and how recently.

Literature is open, refreshable and citable, so the landscape is defensible externally.

Known limitation

Literature under-represents new, proprietary or unpublished assets. The graph takes other sources of truth alongside it.

Step 2 - Turn articles into normalised, reviewable dataset records

Two levels of structure: what one paper says, and what the evidence says in aggregate

Article level

One record per publication
Attributes extracted by a language model:
source name, population, geography, study type, variables, follow-up, treatment, analysis type
Every value stays linked to its source article

Dataset level - aggregated from article records

Article records aggregated on a normalised name, with aliases resolved
Publication count and recency, population and variable coverage, countries, study types
The dataset becomes an entity with evidence

Normalization is the hard part

One dataset appears under many names and acronyms. Resolving them is what makes counting - and ranking - meaningful.

Rule-based QC

Automated checks on completeness, format, implausible values and duplicates before review.

Human-in-the-loop curation

Curators confirm names, resolve ambiguity and correct errors. Decisions are recorded, not just applied.

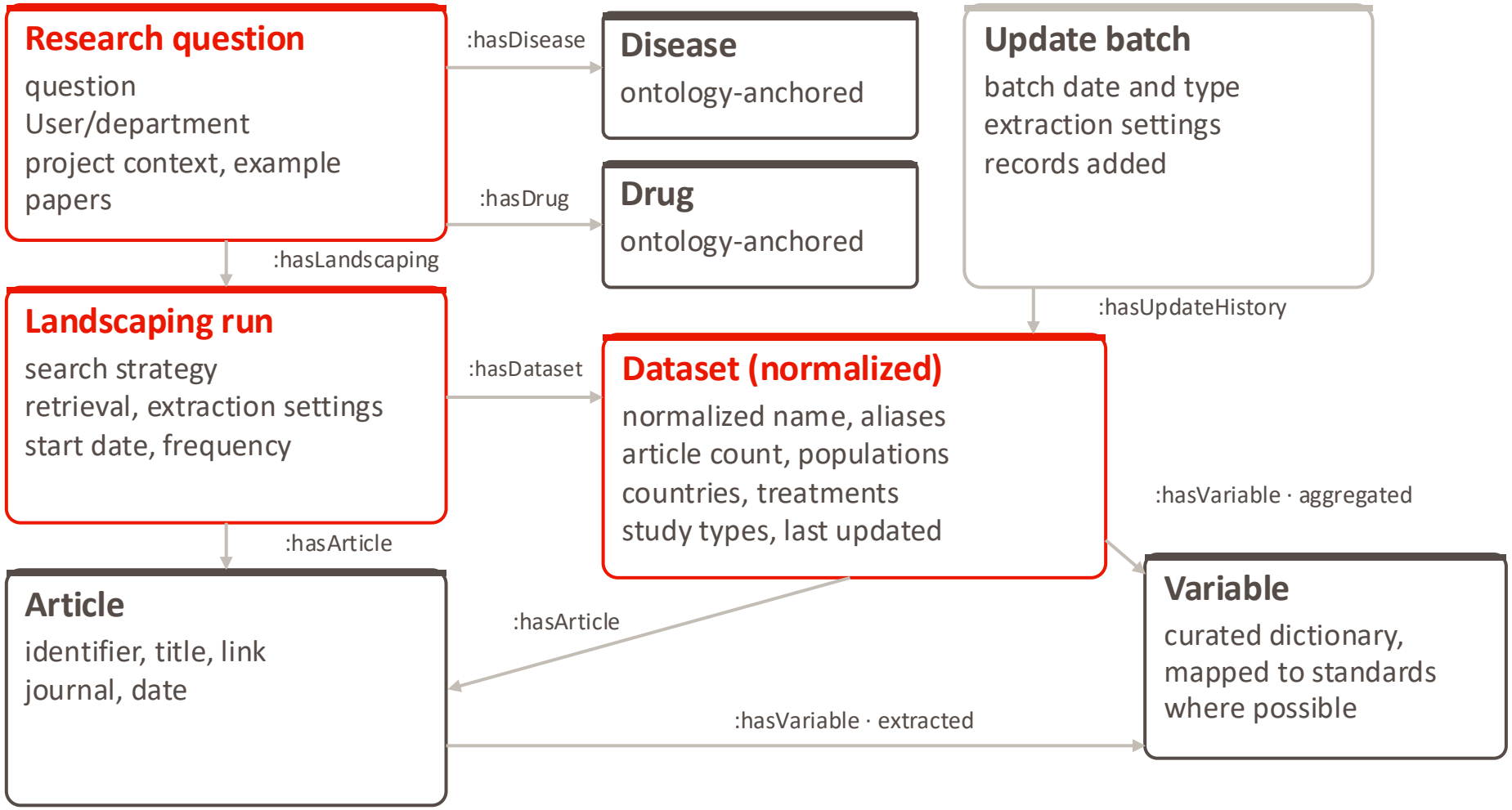
Versioned updates

Every run is stamped as a batch, so any figure traces to the run that produced it.

The data model and design principles

The data model

Entities, relationships and the attributes that make them useful



Reading the model

A question is scoped to ontology-anchored disease and treatment nodes, then triggers a run; the run collects articles; articles name datasets.

The same variable appears at both levels with different values: extracted per article, aggregated across articles at dataset level.

Provenance runs both ways: dataset back to its papers, question forward to the evidence it produced.

Update batches make the graph a history, not just a current state.

Design principles

What we would keep if we rebuilt it tomorrow

Normalize, then count

Aggregate metrics are only meaningful once entity names are resolved.

Alias resolution is a permanent, curated function - not a one-off cleaning step.

Never break the link to the source

Every extracted value keeps a pointer to the article it came from.

A claim you cannot trace is a claim you cannot defend in a submission.

Make time a first-class dimension

Runs are versioned; the graph records what was known when.

This is what makes a refresh incremental and an audit possible.

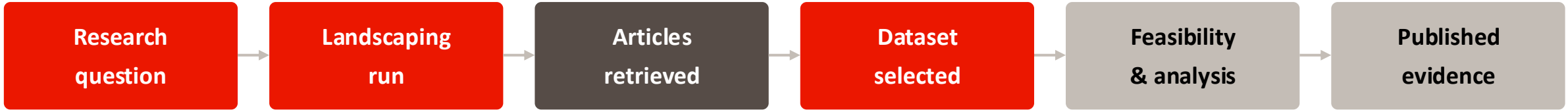
Anchor to standards at the edges

Disease, treatment and variable nodes align to public ontologies (MeSH, MONDO, OMOP-aligned vocabularies).

Interoperability is designed in, not retrofitted.

Traceability in practice

One path through the graph, and what it makes possible

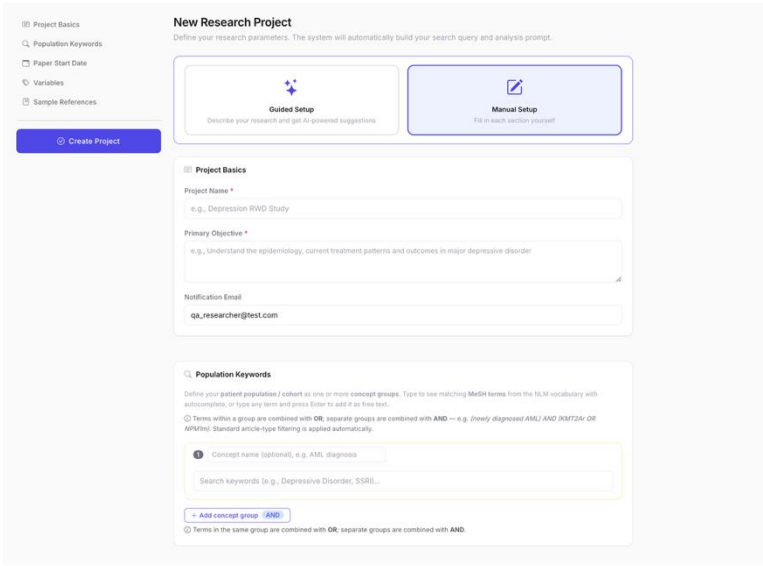


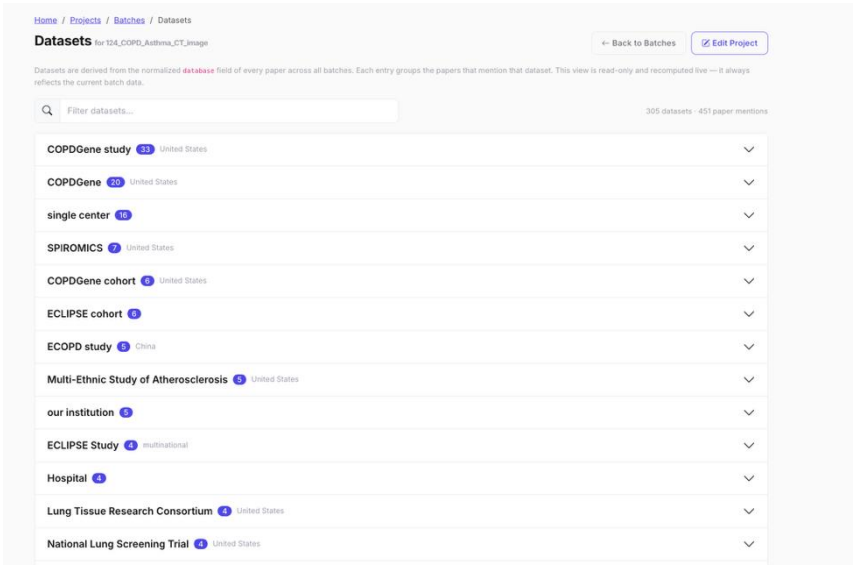
every step is a node or an edge, recorded when it happens - not reconstructed afterwards

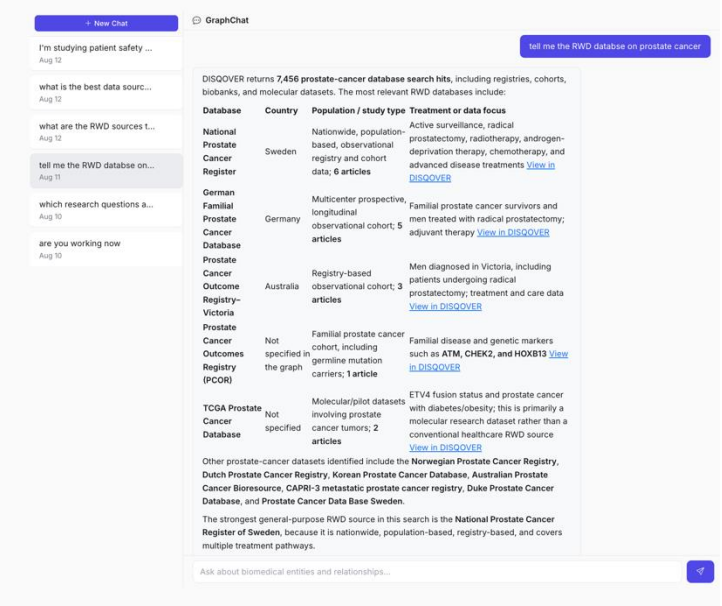
Reconstruct a decision	Why was this source chosen for that question? The candidates, their evidence base and the curation decisions are still in the graph.
Reuse prior work	A new question in an adjacent area starts from an existing landscape instead of a blank page.
Refresh without re-doing	An incremental run adds what is new and flags what changed, against a recorded baseline.
Support transparency expectations	Provenance and source-selection rationale export as a record - what reviewers increasingly ask to see.

What users actually work with - Demo

A web application over the graph: set up a pipeline, browse the evidence, ask questions







Pipeline set-up

Define the research question, the search strategy and the variables of interest; schedule the refresh.

Dataset and article browser

Ranked datasets with their attributes, each opening into the publications behind it.

Graph-grounded assistant

Ask in natural language; answers are composed from graph entities and cite the underlying records.

Why the assistant sits on the graph: retrieval over free text gives a fluent answer with no traceable basis. Answers composed from graph entities carry their provenance.

Questions the foundation can answer

Same graph, different traversals

Today

Which sources could address this question, and what evidence supports that?

What has this dataset been used to study - populations, variables, designs?

How active is a source: how many studies, how recent, in which settings?

Which sources did we consider for a past question, and what did we conclude?

What is new since the last refresh?

Enabled by the graph structure

Which sources recur across disease areas - candidates for long-term relationships?

Which carry the strongest published track record for the evidence we need?

Where are the white spaces - questions with no adequately evidenced source?

Given our prior experience with a source, what should we expect next time?

From a published result, what is the path back to the sourcing decision?

Building on the foundation

What the graph makes possible next

Near term

Deeper extraction: full text as well as abstracts, for the methods detail that determines fitness.

Broader ontology alignment, so the graph interoperates with other platforms.

More automated QC, with curators focused on genuine ambiguity.

Next

Composite indicators of a source's evidence maturity, derived from the published record.

An operational layer: what feasibility work and prior collaborations taught us.

Cross-disease views on where sustained data relationships are worth building.

The direction

Closing the loop: study outcomes flow back, so each project improves the next recommendation.

Sourcing intelligence becomes an asset with a history - not a document that ages.

Method transparency: publishing the approach so it can be critiqued and reused.

Takeaways

What we would ask you to take away

Sourcing intelligence is relational

A catalog answers what exists. A graph answers what a dataset has supported, for which question, with what evidence - and that makes prior work reusable.

AI industrializes discovery, curation earns the trust

Automated retrieval and extraction make landscaping continuous; QC and curation make it defensible.

Lineage has to be designed in

Provenance to the source record and versioned updates are architectural choices, not additions to a finished catalog.

A foundation, not a tool

Once datasets, evidence and study context are connected, recommendation, scoring, cross-disease analysis and traceable reporting are traversals of one graph.

The relationships between datasets and the evidence they produce are the part of RWD sourcing we discard - and the part that compounds.

Thank you

Epidemiology
Valerie van Baalen

EMEA HEMAR
Christina Wennerström

Data, Data Science and AI
Nick Baro
Garrett Jenkinson
Yigit Kucuk
Alice Chang
Timothy Schultz
Alan Andryc
Julia Brodsky