Human-Centred Data Management

Alan Fekete

University of Sydney

SSRG Summer School 2016

Acknowledge: use of images from cited papers and related talk presentations

DB: A Prominent and Active Research Community

- Major conferences, always considered A*, top-tier, etc
  - SIGMOD since 1975 (and prior workshops), VLDB since 1975, ICDE since 1984
- Strong on many metrics
  - many active researchers
    - careers helped by lots of opportunities to publish well
  - strong agreement on what is “good work”
  - plenty of funds from industry (vendors and users) as well as govt
  - many already recognized (helps as referees etc)
    - Turing Awards: Bachmann (73), Codd (81), Gray (98), Stonebraker (14)
    - so many ACM Fellows

Importance of DB field

- The monetary value of data (especially, of rapid action based on data)
  - Billion-dollar industries
- The scale of data
  - especially now, born-digital
  - especially with history included
- Some major vendors started with DB, others created a division to provide this essential component

Typical DB research

- Systems-engineering-style
  - Improve the performance of some part of the system
  - Lots of sophisticated interactions with hardware characteristics and processing algorithms
  - Evaluate by metrics (esp throughput) under varying workloads
- Theory-style
  - Mathematical (usually log-based) model of some aspect, with theorems, complexity bounds, etc
  - Has often led to innovative systems designs
  - Most famous example: relational model and the theory of query equivalence, led to SQL (now almost universal as query interface)
  - “Evaluate” by proof, and also by judgment about practical implications
  - Also: transaction management

What's missing in DB research?

- People!
- For dbms to deliver value, it must interact with people
  - people enter data without error
  - people compose queries that capture intent
  - people interpret output sensibly
  - people configure system parameters to get good performance
- Few papers in the db literature consider people and their characteristics
  - how people's bodies/senses work
  - how people's minds work

Three examples of human-centred research

- Query interface for novel hardware
- Hybrid computation platform
- Empirical study of problems with application coding

THE UNIVERSITY OF SYDNEY

User interface

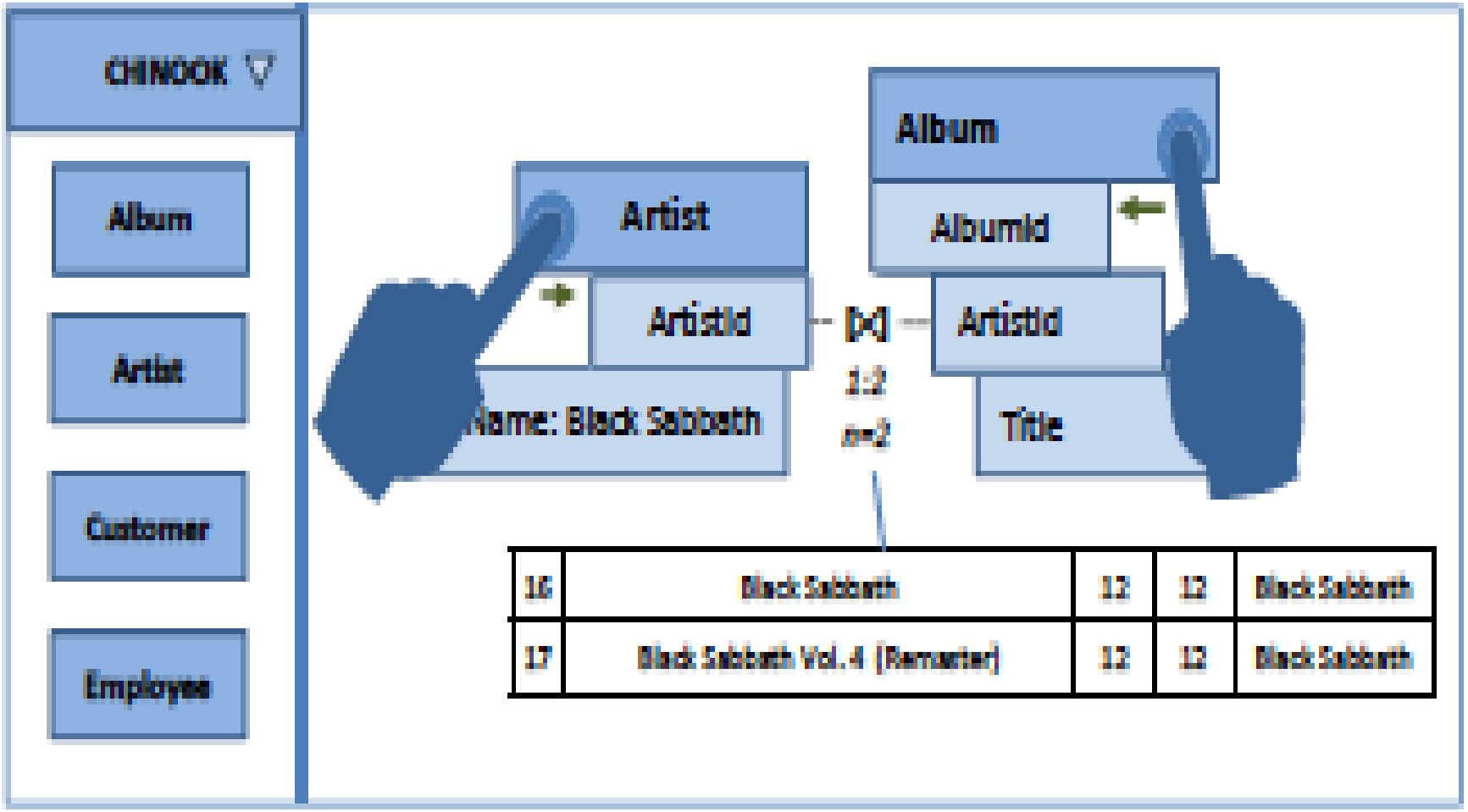
- Gestural Query Specification
- A. Nandi, L. Jiang, M. Mandel (Ohio State U)
- VLDB 2014

Summary

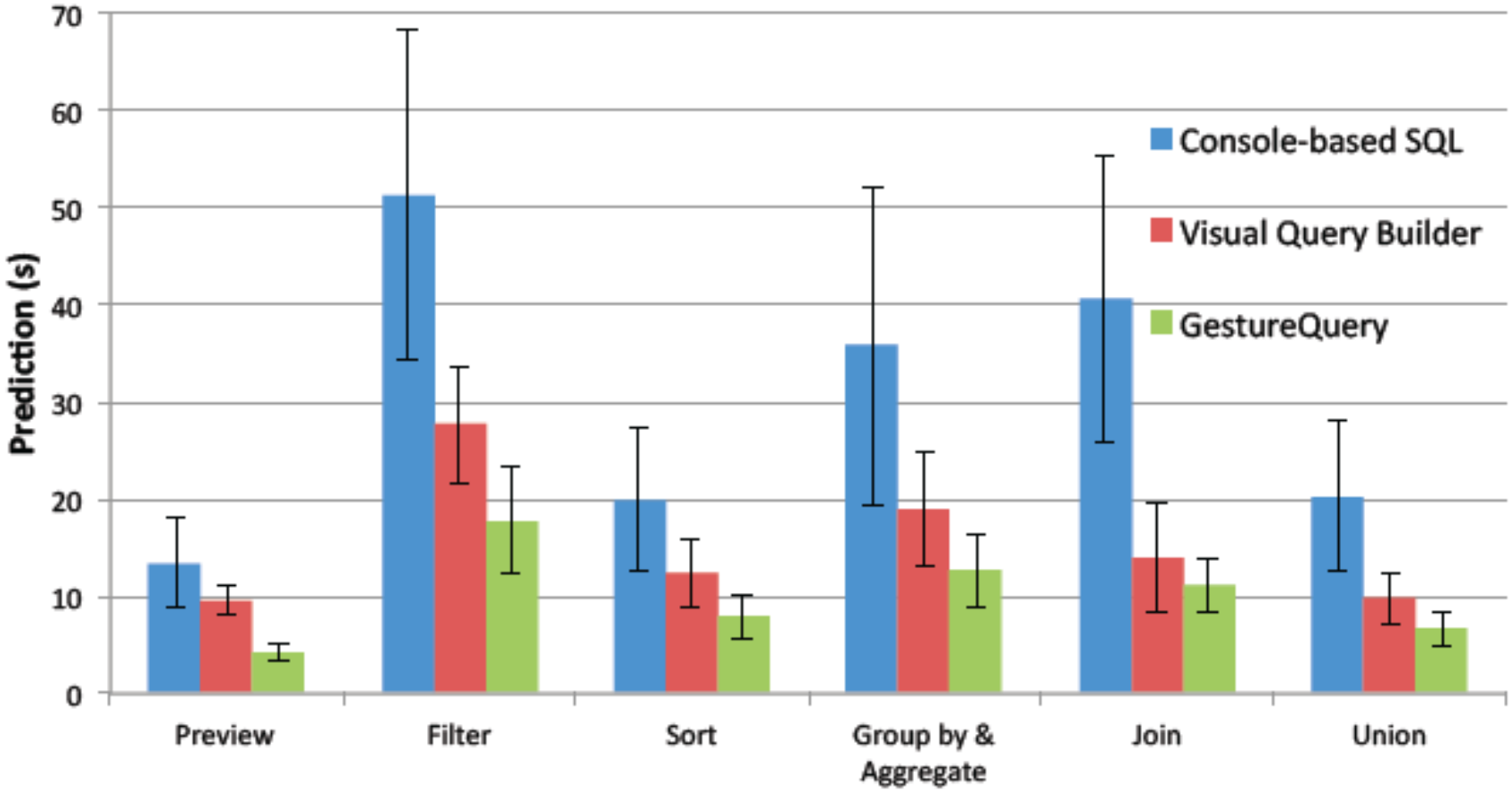
- Driver: new hardware for input with multitouch but without easy keyboard (eg iPad)
- Contribution: Proposes a new way to compose queries through gestures on a multitouch tablet
- Evaluated by user study (time to complete tasks) and performance

- See short video at https://www.youtube.com/watch?v=R7V7Om9gZHI&feature=youtu.be

Gestural join with preview



Gestural Querying is Faster



People-in-the-loop calculation

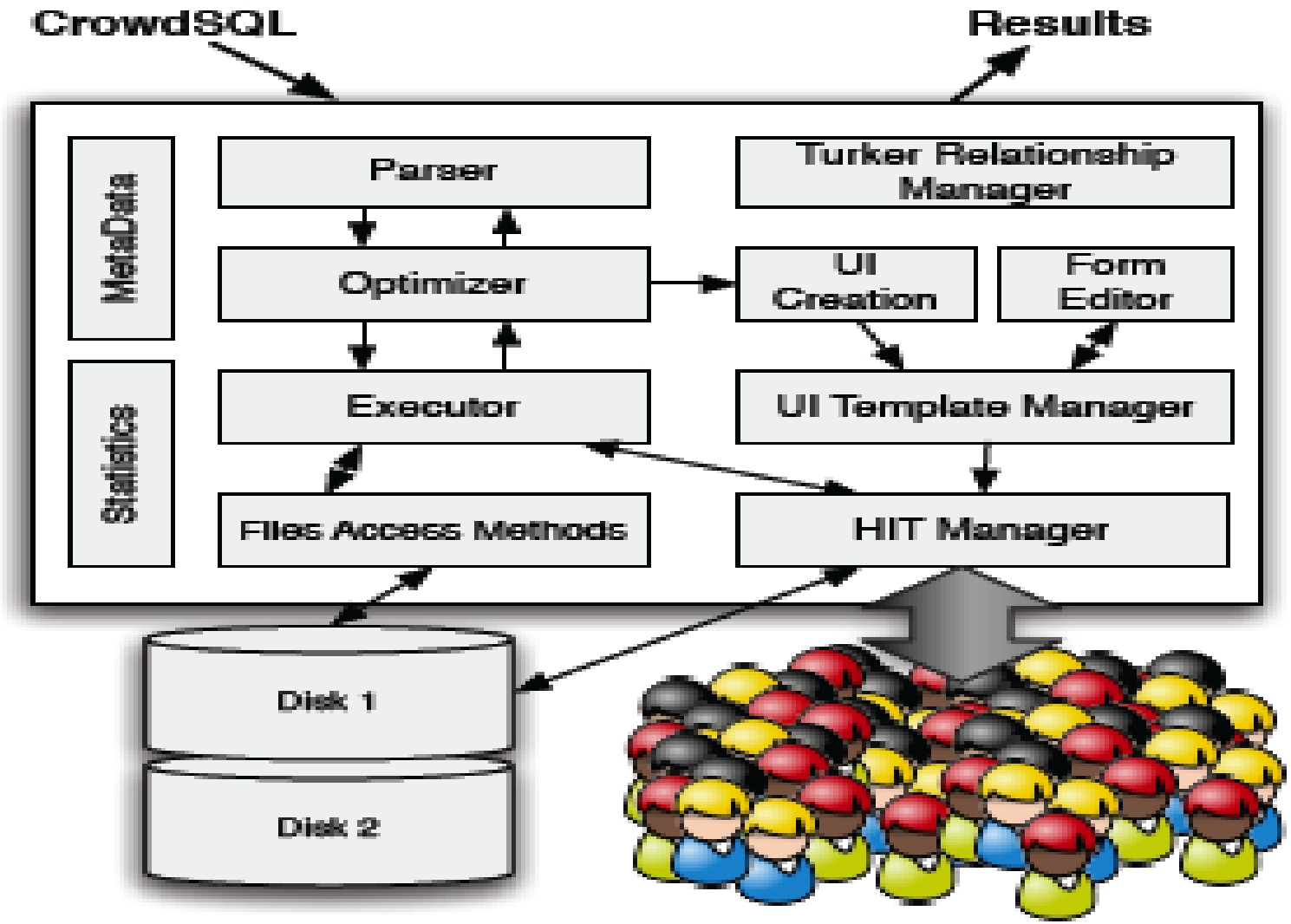
- CrowdDB: Answering Queries with Crowdsourcing
- M. Franklin, D. Kossman, T. Kraska, S. Ramesh, R. Xin (UC Berkeley and ETH)
- SIGMOD’11

THE UNIVERSITY OF SYDNEY

Summary

- Driver: using people to compute steps such as matching, ranking, information extraction that are not algorithmic
- Contribution: system design that uses crowdsourcing to determine fuzzy information that can be incorporated in queries; SQL extensions; integration of crowd operations into query processing; interface design for presenting the tasks to crowdworkers
- Evaluated by measuring response time, and by exploring variation in performance with aspects of the crowdsourcing approach

CrowdDB architecture



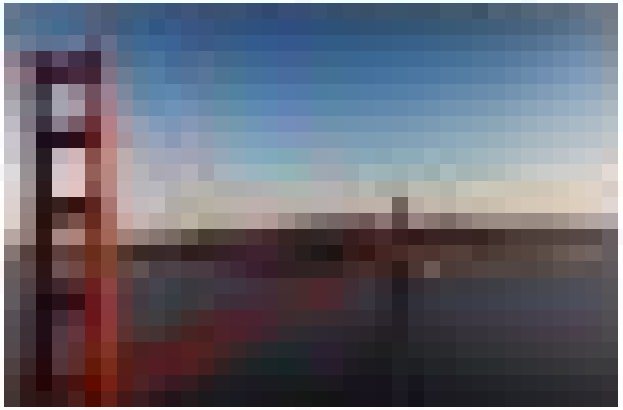
THE UNIVERSITY OF SYDNEY

Example query

```
CREATE TABLE picture (
  p IMAGE,
  subject STRING
);
SELECT p FROM picture
WHERE subject = "Golden Gate Bridge"
ORDER BY CROWDORDER(p,
"Which picture visualizes better %subject");
```

Generate HIT

Which picture looks better "Golden Gate Bridge"



Submit

Empirical study of developers

- Feral Concurrency Control: An Empirical Investigation of Modern Application Integrity
- P. Bails, A. Fekete, M. Franklin, A. Ghodsi, J. Hellerstein, I. Stoica (UC Berkeley and U Sydney)
- SIGMOD’15

THE UNIVERSITY OF SYDNEY

Summary

- Driver: ORM-based application development (contrast to traditional applications that are more tightly linked to DB)
- Contribution: study how integrity is managed in 67 substantial, active Ruby on Rails projects from Github
  - what mechanisms are used
  - how well those mechanisms preserve integrity
- Evaluated by: contributions are not really evaluated, but discussion of potential for generalization to other ORM frameworks

Rails Integrity Mechanisms

1.) Transactions: in business logic

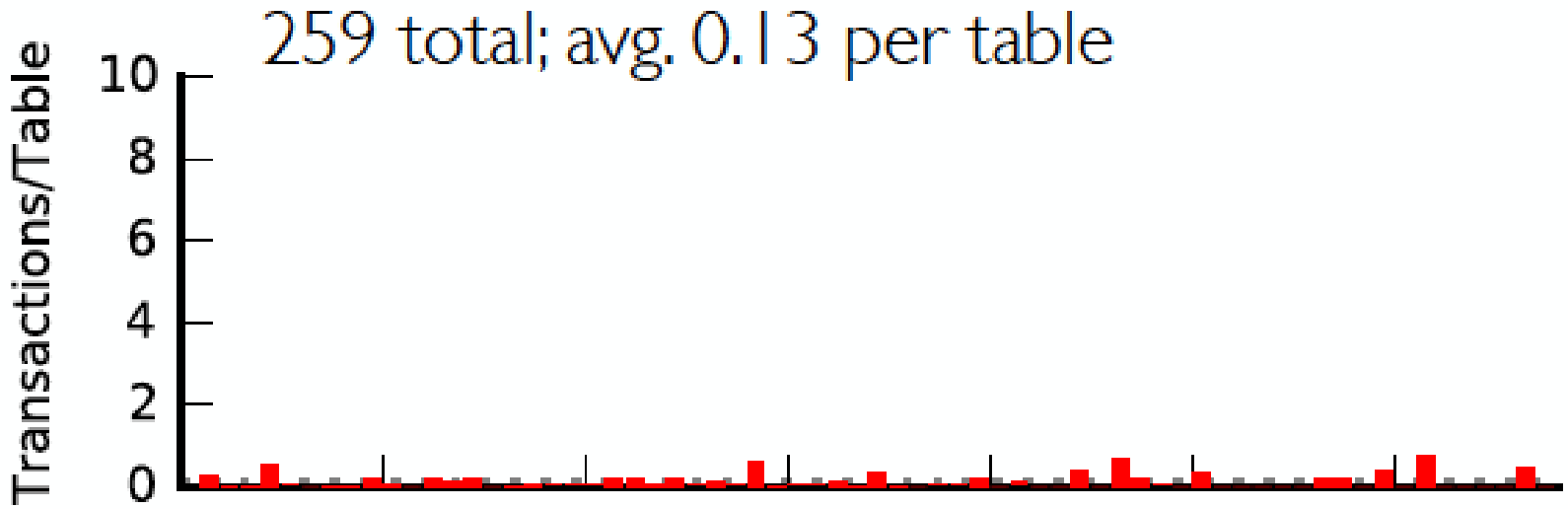
2.) Validations: invariants on schema

```
class Person < ActiveRecord::Base
  validates :name, presence: true, uniqueness: true
end
```

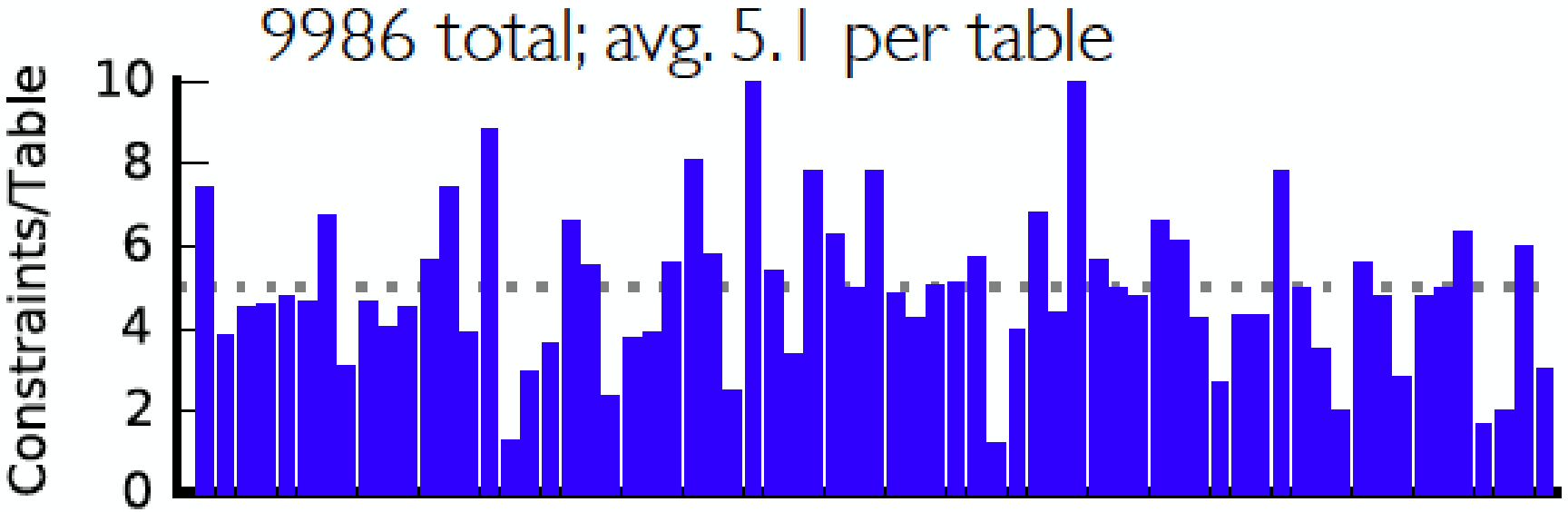
3.) Associations: relational integrity

```
class Order < ActiveRecord::Base
  belongs_to :customer
end
```

Transactions in the App?



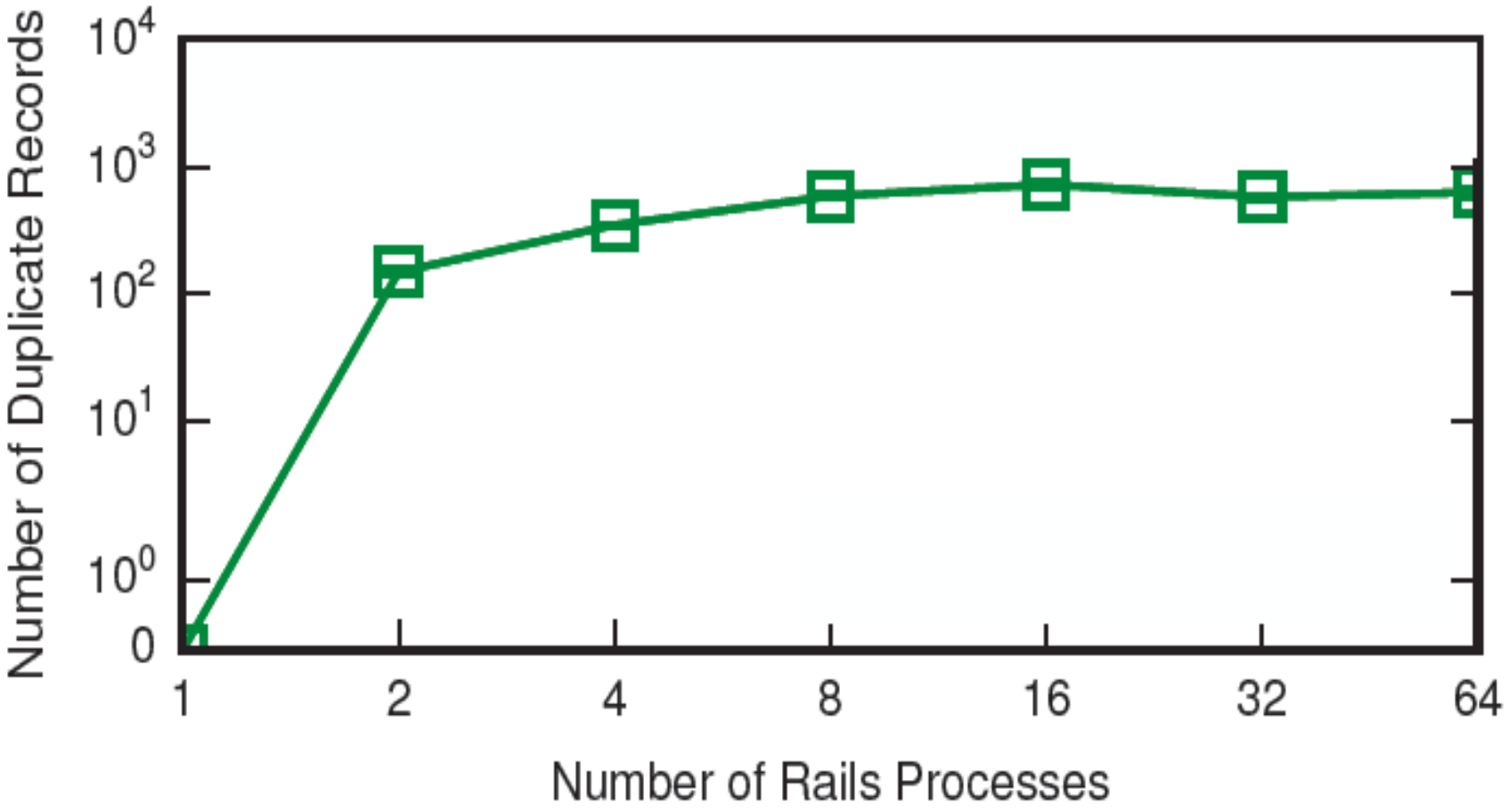
Assertions and Associations



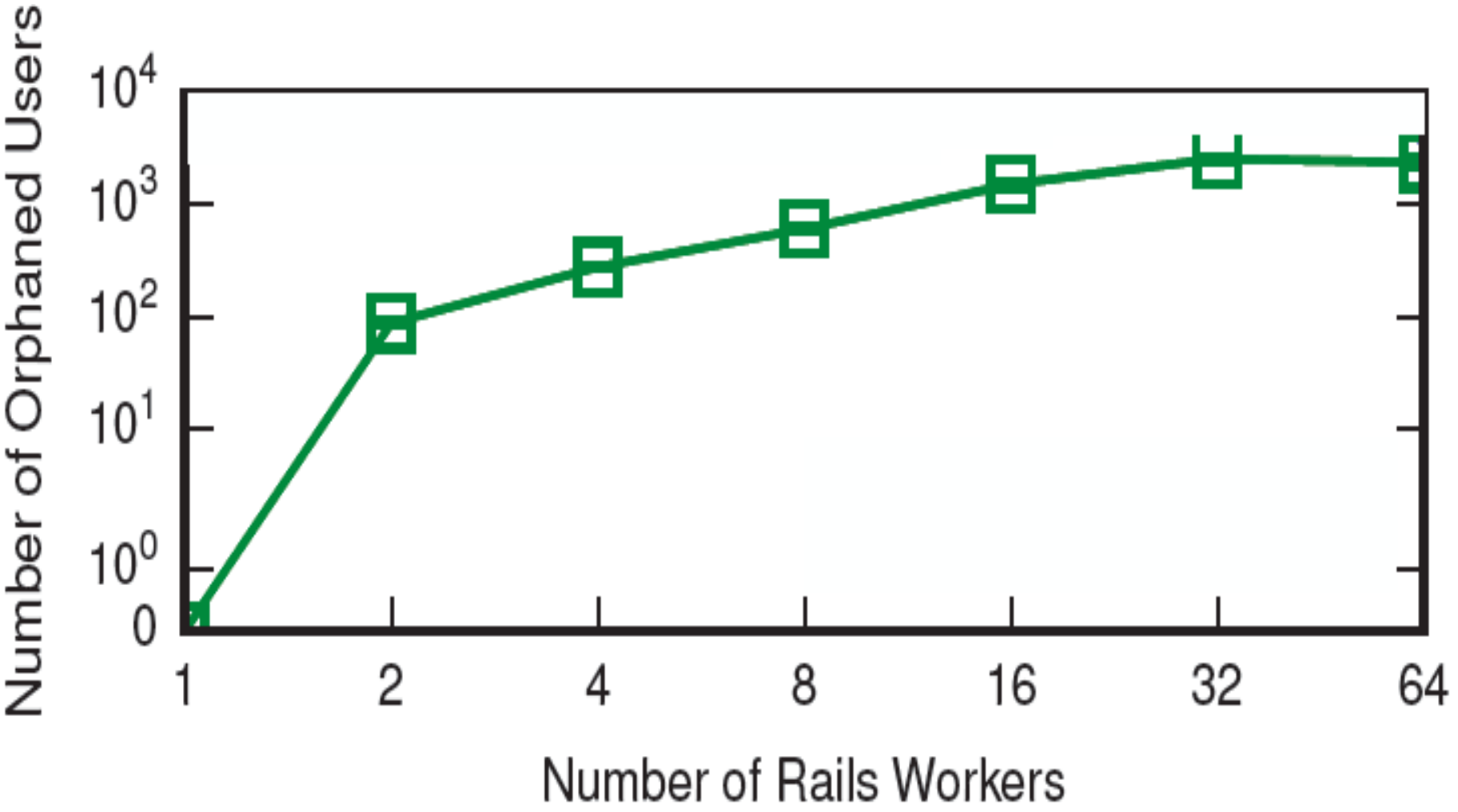
Feral Mechanisms

- In Rails, Assertions and Associations are opaque to DBMS
  - checked in framework, not passed throug hto schema as relational constraints
- Do they work correctly?
- Theory: Invariant Confluence Test (from Bailis et al, VLDB'15)
- In the studied repository, 86.9% of constraints pass ICT (safe!), but 13.1% do not pass ICT (potentially unsafe)

Run an example of Uniqueness Validation



Run an example of Association Validation



THE UNIVERSITY OF SYDNEY

Where to next?

- HILDA (Human In the Loop Data management) workshop at SIGMOD'16
  - http://hilda.io
- Please submit if you have relevant work!