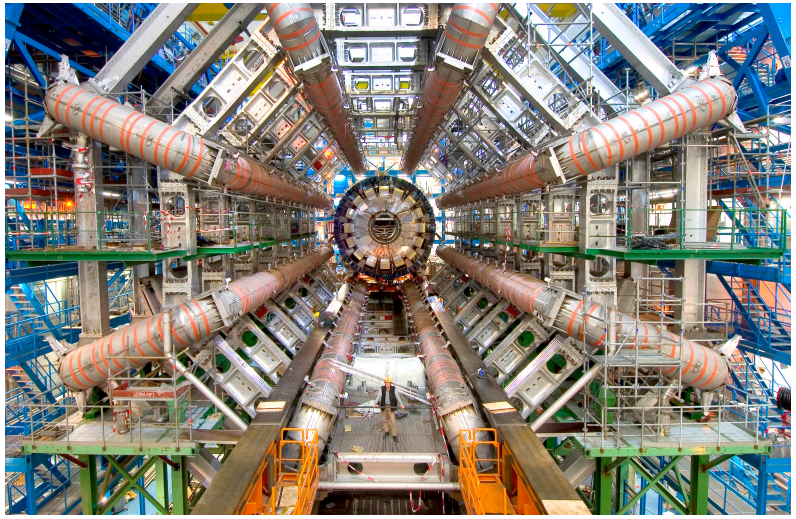
Virtual Experiments for Distributed Research Networks

Jennie Duggan

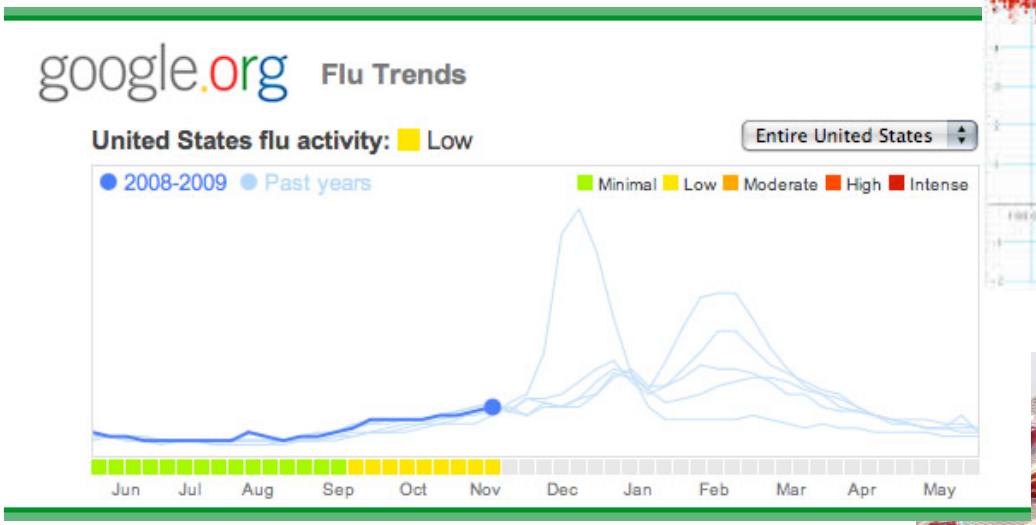
NORTHWESTERN
UNIVERSITY

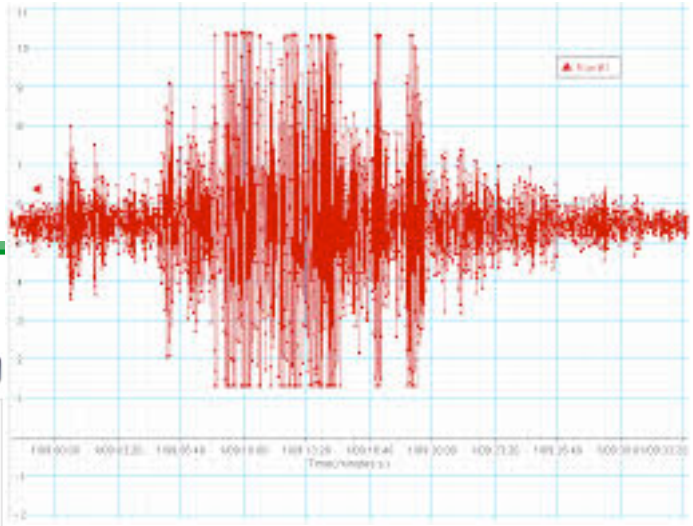
Science is Changing

- Research is increasingly data-intensive
- Data reuse is the new normal
- Reproducibility is low



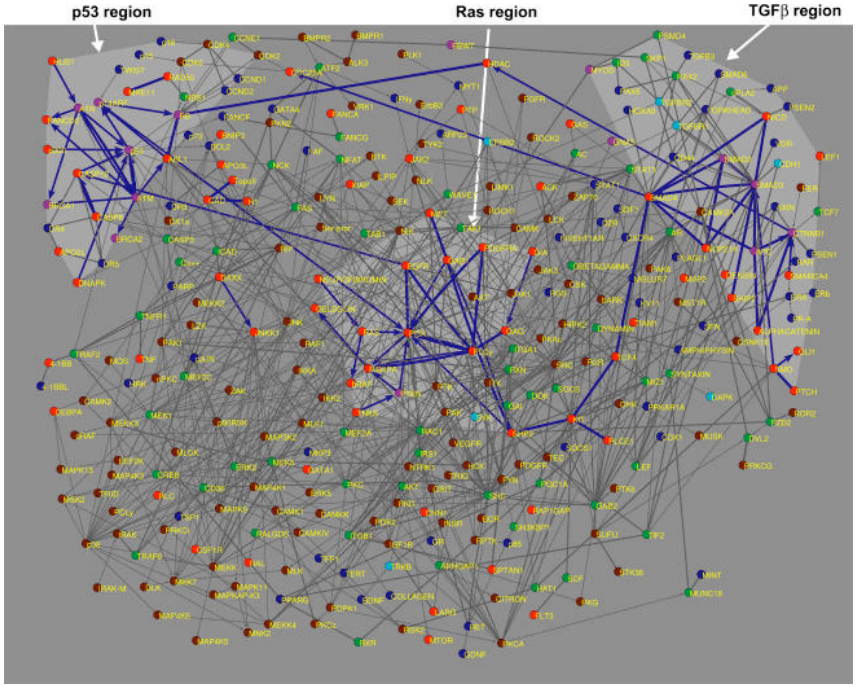
Trouble in Paradise





What changes with data reuse?

- Neither man nor machine is sufficient
- Experimental design is main artifact
- Explore a space of hypotheses
- Large remote data sources
- Discoveries can be continuously verified
- Community science



Cui, Qinghua, et al. "A map of human cancer signaling." *Molecular systems biology* 3.1 (2007).

Hephaestus

- Proposed data reuse platform
- Approach:
  - Virtual experiments
  - Probabilistic causal graphs
- Meta-system sits atop $\geq$ 1 data repos & computing resources

Virtual Experiments

- Queries to express experimental design
- Identify and statistically test hypotheses
- Creates “recipes” for discovery
- Meta-system translates VEs into queries on existing data sources

Skin Cancer Experiment

- Oncologist looking for root causes of disease
- Hypotheses: cancer result of sun exposure, fair skin, ?
- Controls for gender, age

Block: age < 25 and gender = female

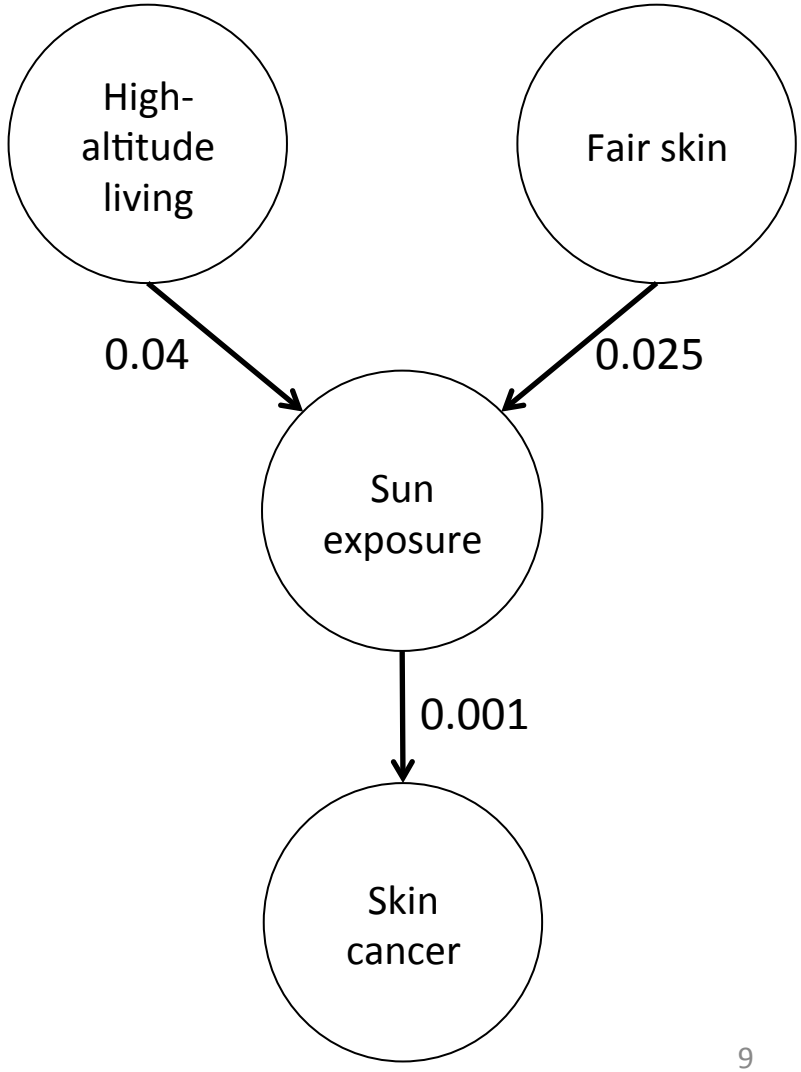
		Sun Exposure = 0	Sun Exposure = 1
Cancer	0	$n_{0,0}$	$n_{1,0}$
	1	$n_{0,1}$	$n_{1,1}$

Virtual Experiment Language

```
SELECT * LIMIT 10
FROM cancerSubjects
EFFECT 'skin cancer' as S
INTERVENTION sun exposure, skin tone, *
CONTROLLING FOR age, gender
ANALYSIS chisquare
SCORE BY pvalue(c) ASCENDING
WHERE pvalue(c) <= 0.05;
```

Probabilistic Causal Graphs

- Catalog discoveries
  - Edges are VEs
- Graphs are sharable for community science
- Relationships may be complex



Computing the Graph

- Some tasks better for machines
  - Check for inconsistencies
  - For rival frameworks find where they intersect and diverge
  - Re-weight edges w/ continuous verification

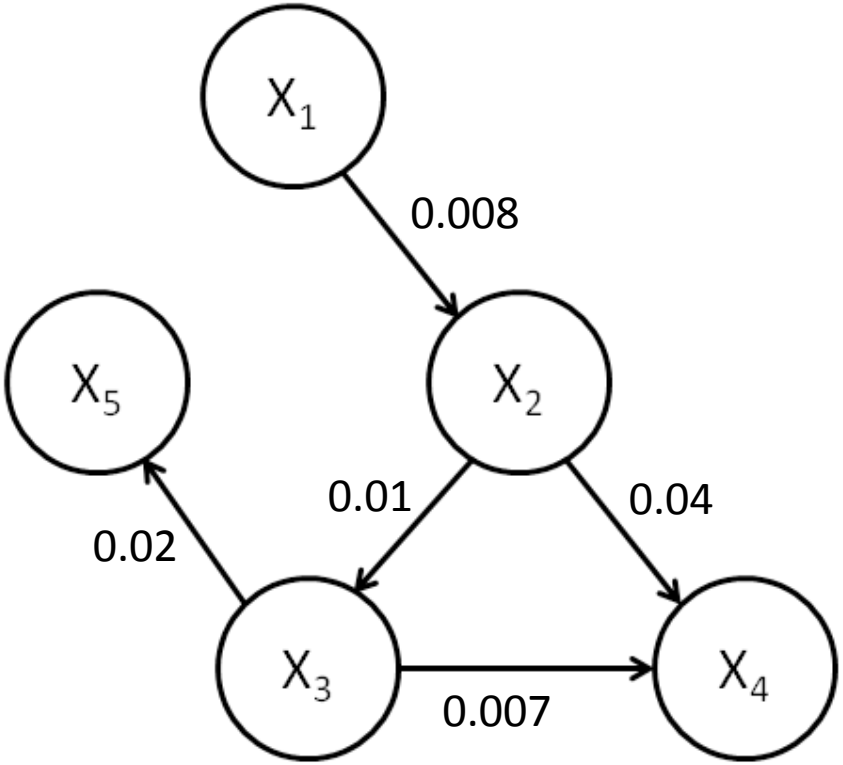


Image: http://www.michaelnielsen.org/ddi/if-correlation-doesnt-imply-causation-then-what-does/

Augmenting Human Researchers

- Viz proposed discoveries in context
- Vary VE parameters en masse
- Zoom in and out
- Sharing discoveries

Distributed Research Networks

- Many data providers
- Multiple data models
- Workloads rarely just relational-style queries
- Data often subject to governance – need to reuse in place

Use Case: Clinical Data Research Networks

pcornet

The National Patient-Centered Clinical Research Network

Why is this not just another federated db?

- Complex trust environment
- Queries to be executed in-situ
  - data ingest and cleaning
  - in-engine statistical analysis
- Sources come and go
- DBs have varying populations in data
- Heterogeneous compute power

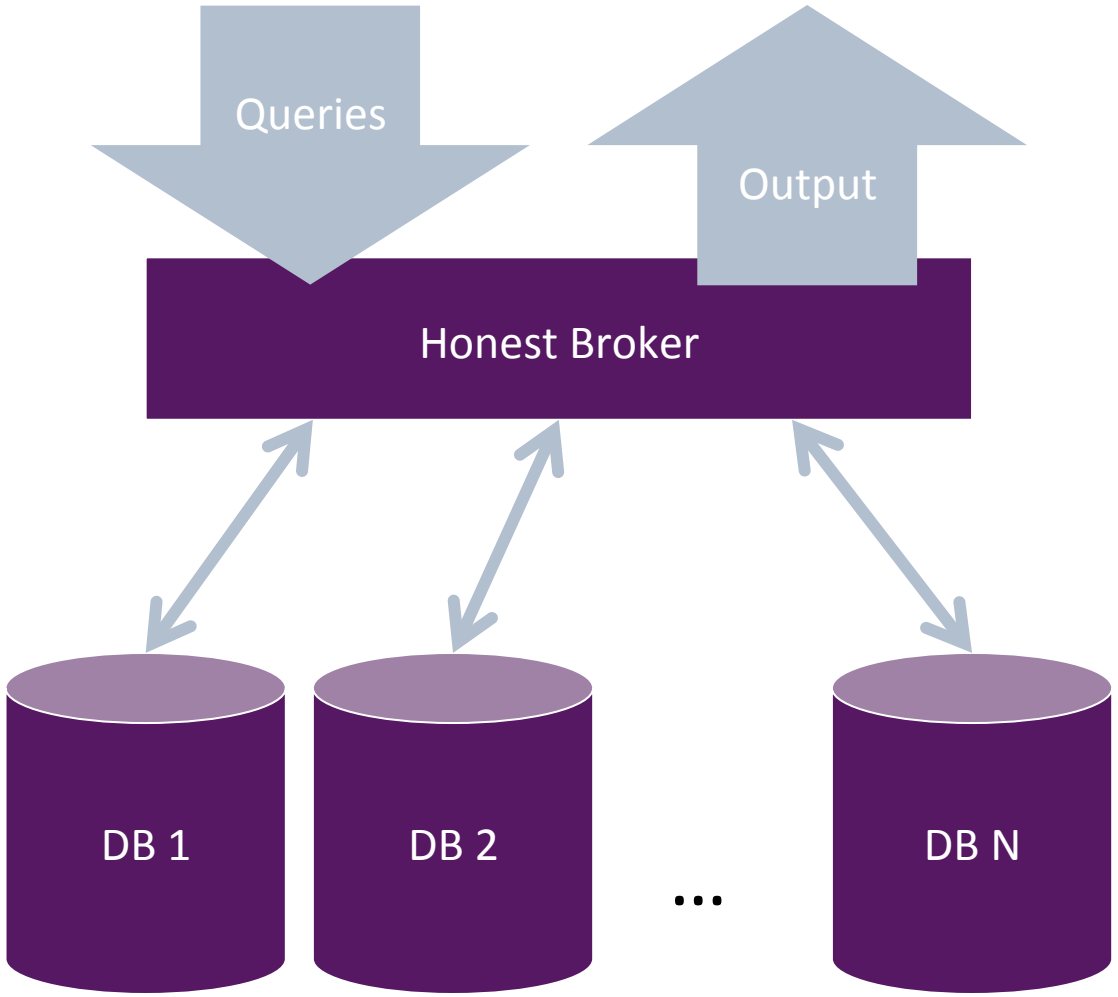
Example Clinical Data VE

- "Test for a change in mortality rates for lupus patients with fragmented care adjusting for race, gender, and insurance status."
  - Iteratively break down and verify confounders (8X chi-square)
  - Use logistic regression to get p-values for relationship (23X)
- Very slow
- Pull most of the data out of the db and use SAS for analysis
  - 640 lines of code!

Getting started…

- Security and privacy
- Simulating experiments
- Multi-data model support

DRN Query Execution Model



Simulating the Experiment

- Seek to imitate the conditions of a randomized controlled trial
- Two issues:
  - Statistical freshness
  - Data selection

Adaptive Data Reuse

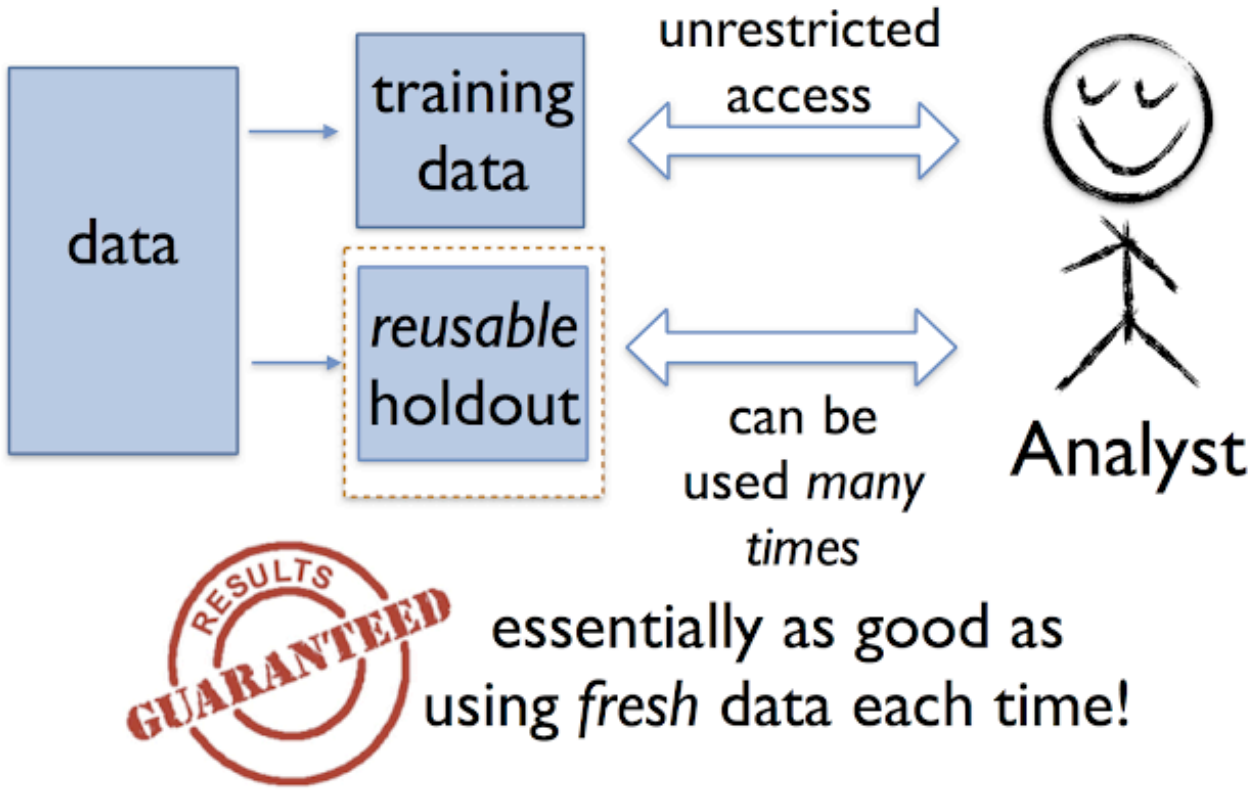
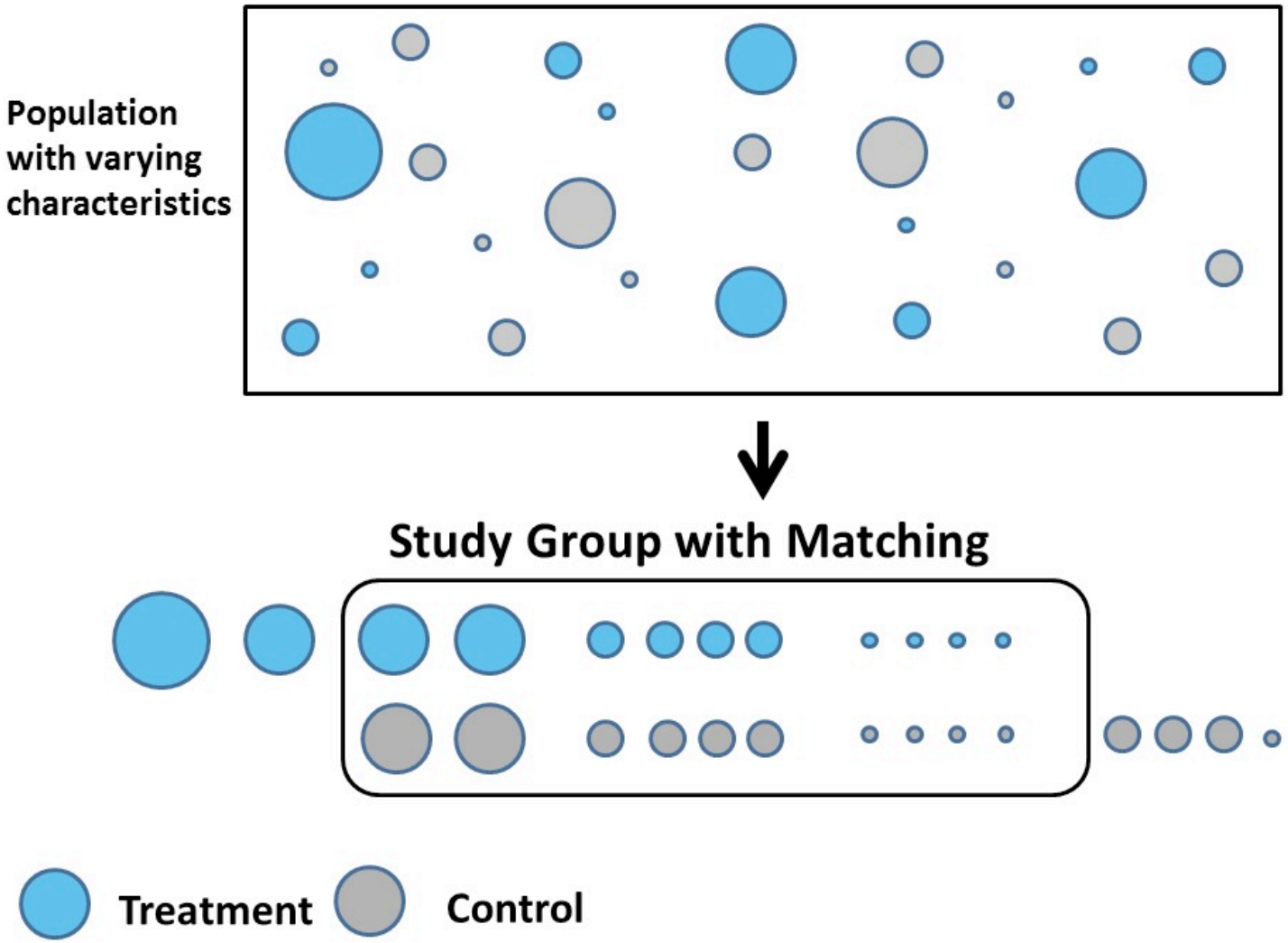


Image source: http://googleresearch.blogspot.com/2015/08/the-reusable-holdout-preserving.html

Propensity Score Matching



Source: http://www.summitllc.us/propensity-score-matching/

Querying Multiple Data Models

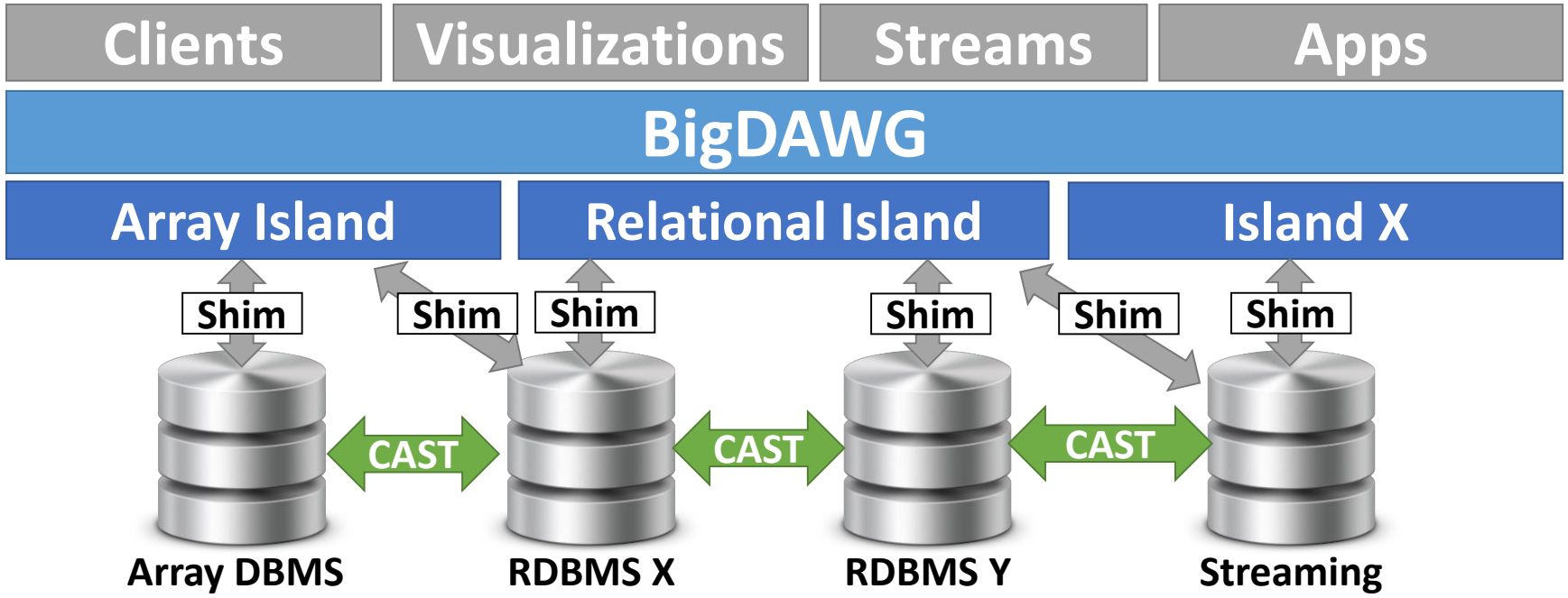
- In DRNs, users want to access a variety of data sources with many disparate models.
- Lots of data is not even initially in a db!
- Challenges:
  - Location transparency
  - Semantic completeness
  - N:N relationship between back ends and user semantics

BigDAWG Polystore Design

- Achieve data independence with **islands of information**. Each contains:
  - Data model, query language, and shims to supporting db(s)
- Users pose queries by invoking islands with `scopes` and by `casting` between disparate semantics
  - Example:

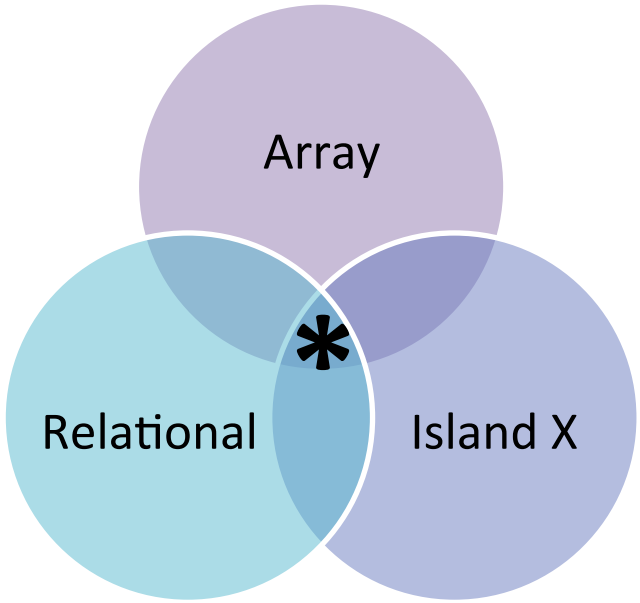
```
RELATIONAL(SELECT *
                FROM R, CAST(A, relation)
                WHERE A.v = R.v);
```

Polystore Architecture



Finding Overlapping Island Semantics

- Goal: identify semantics that span 2+ islands
- Define decision space for query optimization
- Approach:
  - Enumerate all or sample of n-gram plans for an island over all shims in canonical form
  - Diff output against same enumeration in other islands for overlapping semantics
- Identify transitive relationships between engines



Polystore Query Optimization

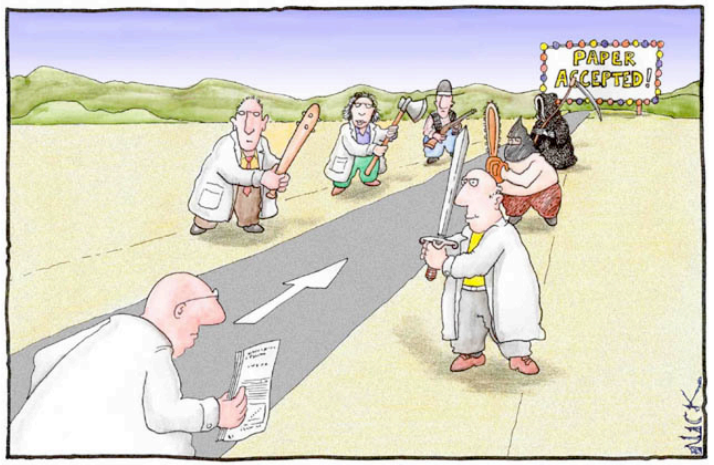
- How do we assign work to DBMSs?
- Classify island operators as *simple* or *complex*
- Simple ops have approximately same cost everywhere
- Complex ops have significant performance variance by engine
- Heuristics:
  - Copy avoidance: bring computation to the data when possible for simple ops
  - Reroute data to complex ops when profitable
- Iterative dynamic programming for stochastically optimizing partial plans

Long Term Research Challenges

- Modeling and pruning a hypothesis space
- Managing & visualizing uncertainty
- Data discovery
- Pay-as-you-go curation

Conclusions

- Scientists are decoupling data collection from analysis
- Reframe scientific method for data reuse (aka data science?)
- Many open questions on how to formulate and optimize queries in DRNs



Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'