The Henry and Marilyn Taub
Faculty of Computer Science

Measuring the Importance of Database Elements

Benny Kimelfeld

Deutsche Forschungsgemeinschaft

Importance of Database Tuples

Ester
Livshits

Leopoldo
Bertossi

Mikaël
Monet

Daniel
Deutch

Nave
Frost

From Explanation to Responsibility Attribution

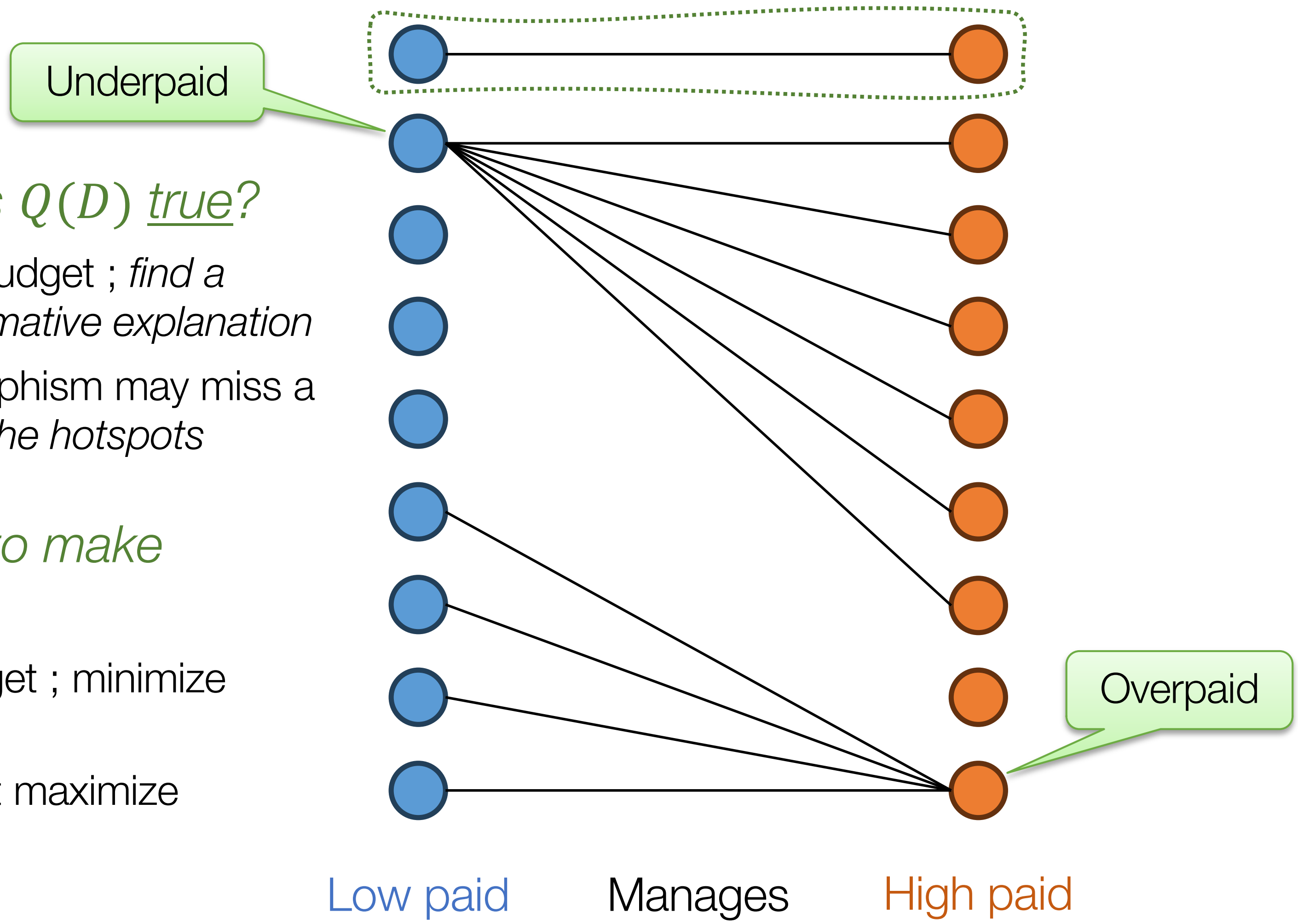
$$\exists x, y \; [\, \text{Salary}(x, \text{low}) \wedge \text{Salary}(y, \text{high}) \wedge \text{Manages}(x, y) \,]$$

Descriptive: *Why is $Q(D)$ true?*

- Limited attention budget ; *find a compact and informative explanation*
- Arbitrary homomorphism may miss a bigger story ; *find the hotspots*

Prescriptive: *How to make $Q(D)$ false?*

- Limited repair budget ; minimize the *extent* of $Q(D)$
- Find the tuples that maximize the benefit of fixing



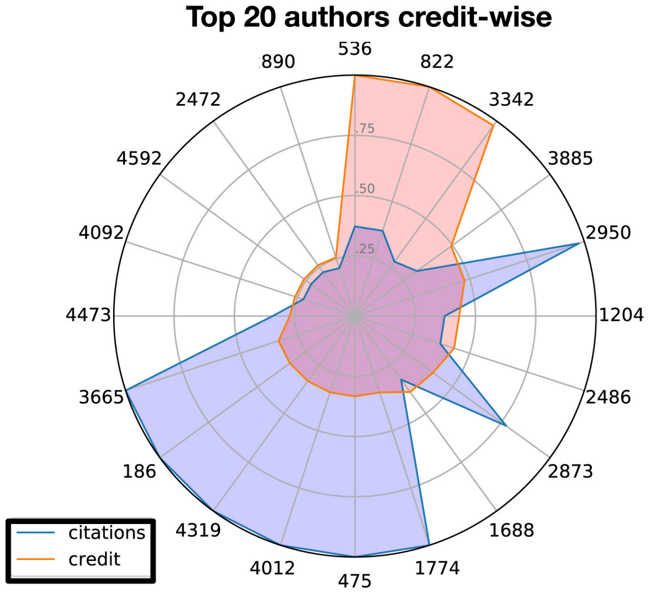
Beyond simple degrees, e.g., $\text{Manages}(x, y) \wedge \text{Manages}(y, z) \wedge \text{Family}(x, z)$

Another Example (Data Credit Distribution)

[Dosso-Davidson-Silvello22]



Credit to tuples of curated data based on references from *British Journal of Pharma.*



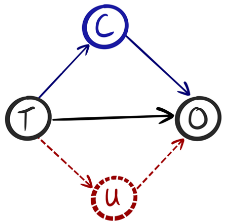
Credit to data curators (vs. their citation scores)

* Dennis Dosso, Susan B. Davidson, Gianmaria Silvello: *Credit distribution in relational scientific databases, Information Systems*, Volume 109, 2022.

Annotation vs. Contribution

- Opposite flows:
  - Annotated DBs: annotate output tuples according to the annotation of input tuples
  - Tuple contribution: annotate input tuples according to their impact on output tuples
- Abstractly – *how does each input annotation contribute to the output annotation?*
  - Useful abstraction for aggregate queries (e.g., sum)
- Relationship between the two… to be explored

Approaches to Contribution Measurement



- **Causality** (level of responsibility)
  - Idea: *Query answers depend causally on tuples ; to what degree?*
  - *Counterfactual dependence* under contingency [Meliou-Gatterbauer-Moore-Suciu10] [Meliou-Roy-Suciu15]
    - Based on [Chockler-Halpern04]: "*… minimal number of changes […] to obtain a contingency where B counterfactually depends on A*"
    - Here, min #tuples to delete so the answer depends on the tuple's existence
  - *Causal effect* [Salimi-Bertossi-Suciu-VanDenBroeck16]
    - Based on Pearl's degree of responsibility [Pearl09]
    - $\mathbb{E}[Q|\ \text{tuple}] - \mathbb{E}[Q|\ \neg\text{tuple}]$ when the DB is considered a probabilistic DB
    - Similar to earlier ideas [Kanagal-Li-Deshpande11]

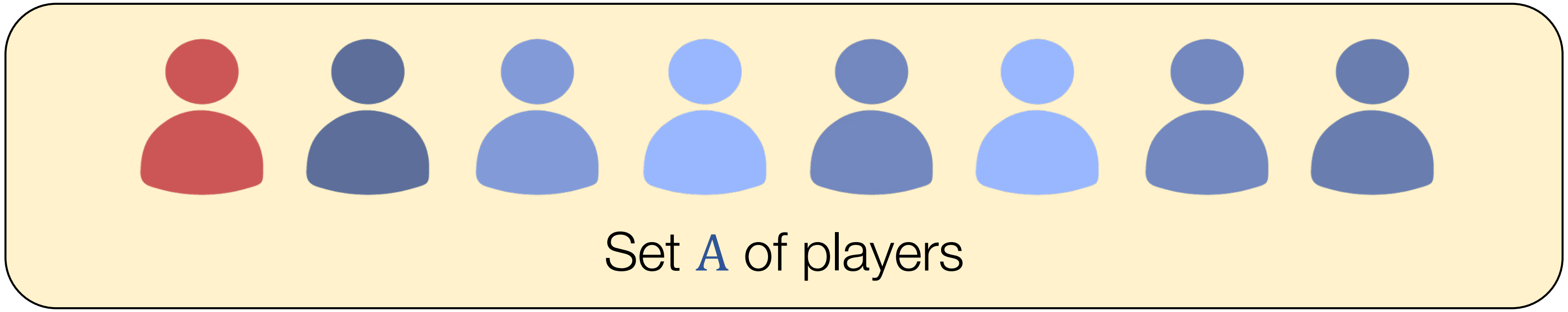
- **Cooperative Games** (profit sharing)
  - Idea: *tuples cooperate towards the answer ; what is their "share"?*
  - The Shapley value [Livshits-Bertossi-K-Sebag20] (next…)
  - The Banzhaf Power Index (= causal effect) [Abramovich-Deutch-Frost-Kara-Olteanu23]

The Shapley Value

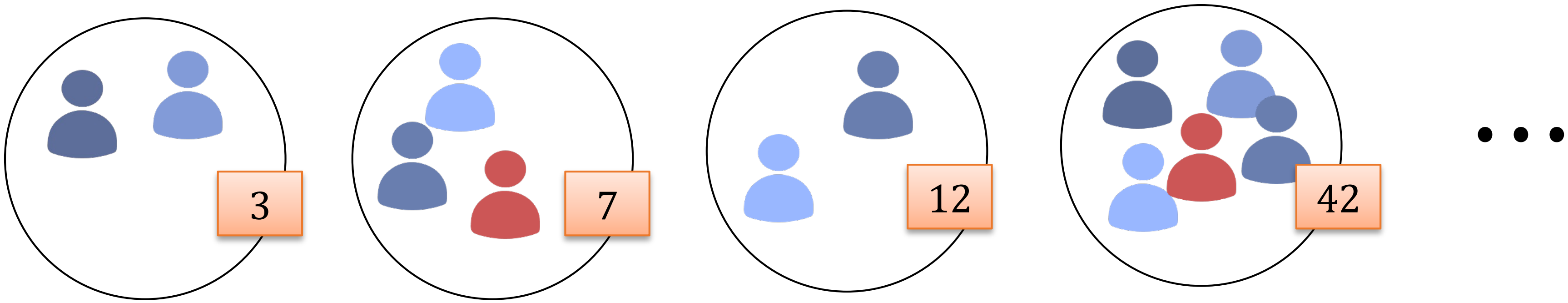
Lloyd Shapley
[1923-2016]

- Widely known profit-sharing formula in cooperative game theory by Shapley
  - [L.S. Shapley: *A value for n-person games*, 1953]
- Theoretical justification: **unique under axioms of rationality** (symmetry, linearity, efficiency, null player)
- Many application areas
  - Pollution responsibility in environmental management
  - Influence measurement in social networks
  - Identifying candidate autism genes
  - Bargaining foundations in economics
  - Takeover corporate rights in law
  - Explanations (local) in machine learning
  - Explanations in databases
  - …

Shapley Definition



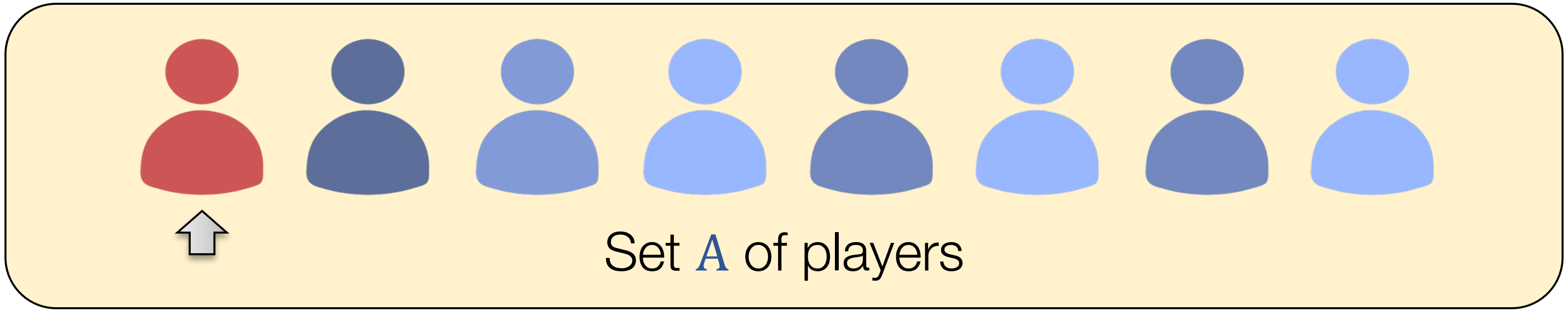
Wealth function $v: \mathcal{P}(A) \longrightarrow \mathbb{R}$ maps each coalition to a utility

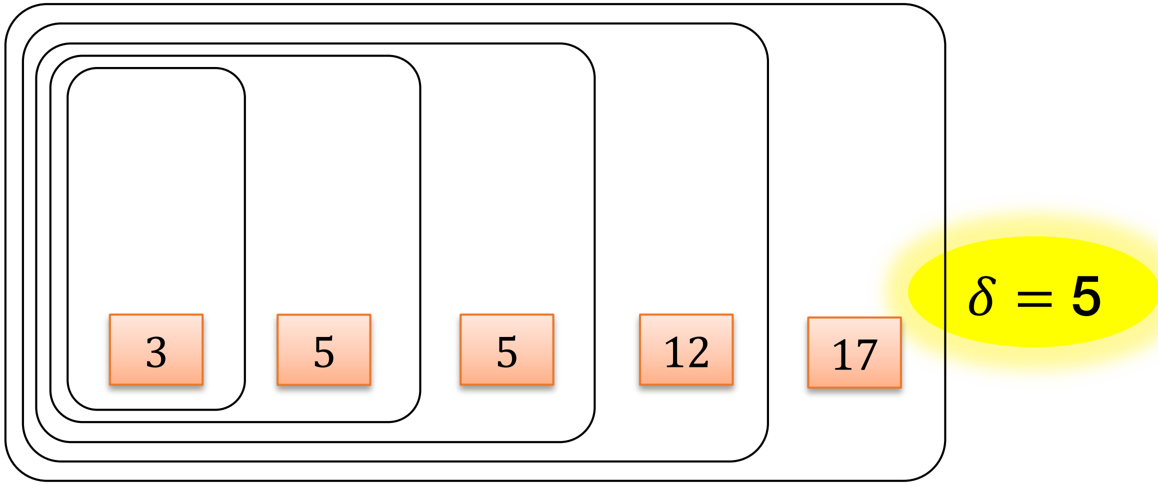


How to share the wealth among the players?

$$\text{Shapley}(A, v, a) = \sum_{B \subseteq A \setminus \{a\}} \frac{|B|!\,(|A| - |B| - 1)!}{|A|!} \big(v(B \cup \{a\}) - v(B)\big)$$

Shapley Explained

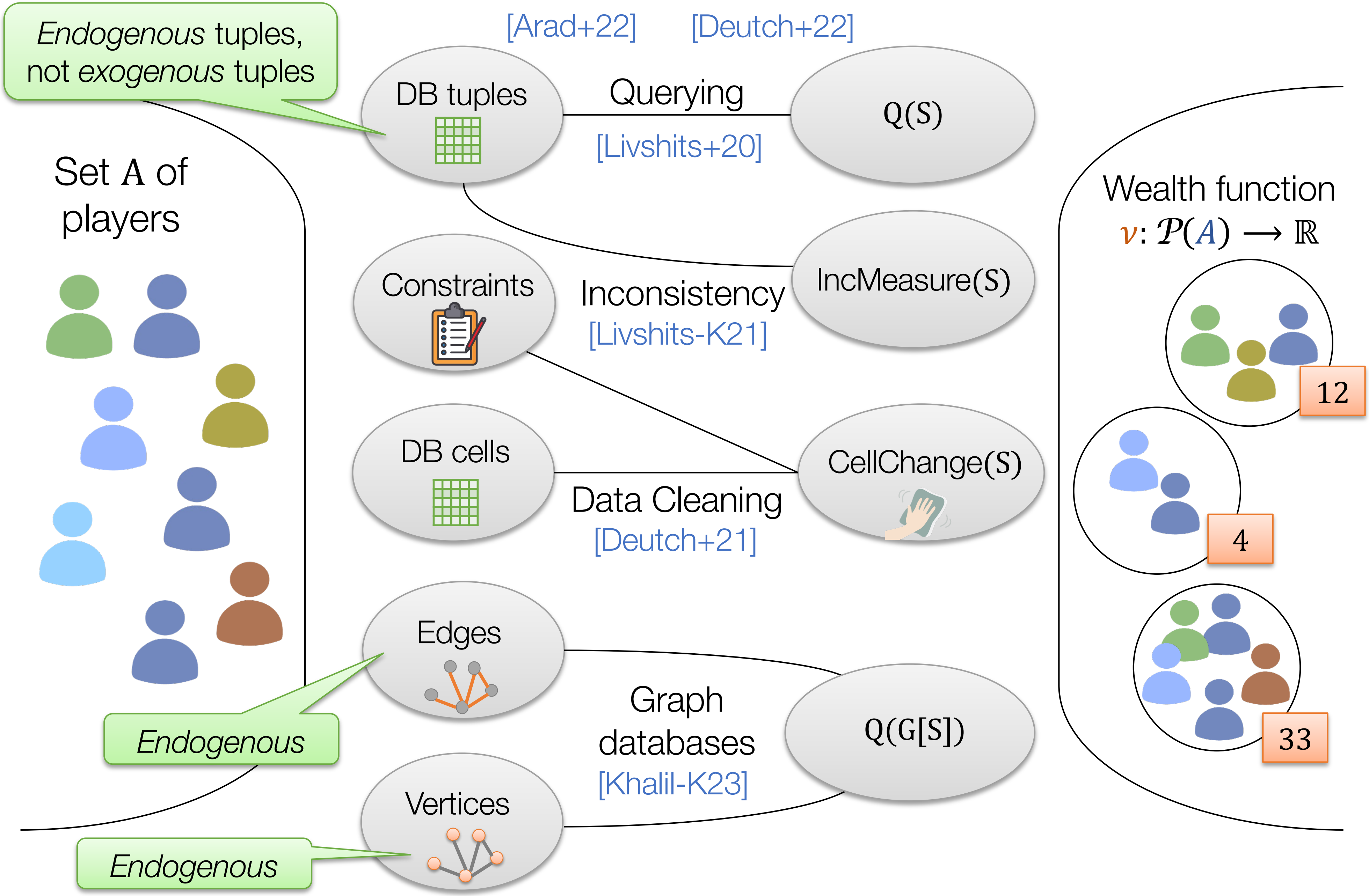




$$\text{Shapley}(A, v, a) = \sum_{B \subseteq A \setminus \{a\}} \frac{|B|!\,(|A| - |B| - 1)!}{|A|!} \big(v(B \cup \{a\}) - v(B)\big)$$

Shapley value: expected δ

Examples of Database Usage



Computation Techniques

- Factorization through linearity of expectation
  - Example: Inconsistency measure #violations under functional dependencies, #problematic tuples [Livshits-K21]
- Reduction to queries over probabilistic DBs
  - General result [Deutch-Frost-K-Monet22]
- Knowledge compilation (to d-DNNF)
  - *Daniel's talk…* [Deutch-Frost-K-Monet22]
- Approximation via sampling [Reshef-K-Livshits20] [Livshits-K21] [Khalil-K23]
  - Additive approx gives multiplicative approx via the *gap property*: the Shapley value is either zero or large

Reduction to PQE

For every Boolean query Q, Shapley[Q] reduces in PTime to Eval[Q] over tuple-independent databases [Deutch-Frost-K-Monet22]

- Proof idea:
  1. Reduce Shapley to the problem of counting the size-k-sets of tuples that satisfy the query
  2. Produce from the database multiple TIDs, each with a different (uniform) probability for the endogenous tuples
  3. Each probability gives a linear combination over the counts of size-k-sets ; all linearly independent (Vandermonde)

  ⇒ Solve equation system to find the counts

- Similar to a known reduction for the SHAP score [VandenBroeck-Lykov-Schleich-Suciu21]

Other Direction?

- The other direction is open: we do not know whether Shapley[Q] and PQE[Q] have the same complexity
- Solved positively for the class CQs w/o self-joins
  - For both, *the tractable CQs are the hierarchical CQs* [Livshits+20]

Importance of Query Parameters

(preliminary work, unpublished yet)

Peter
Lindner

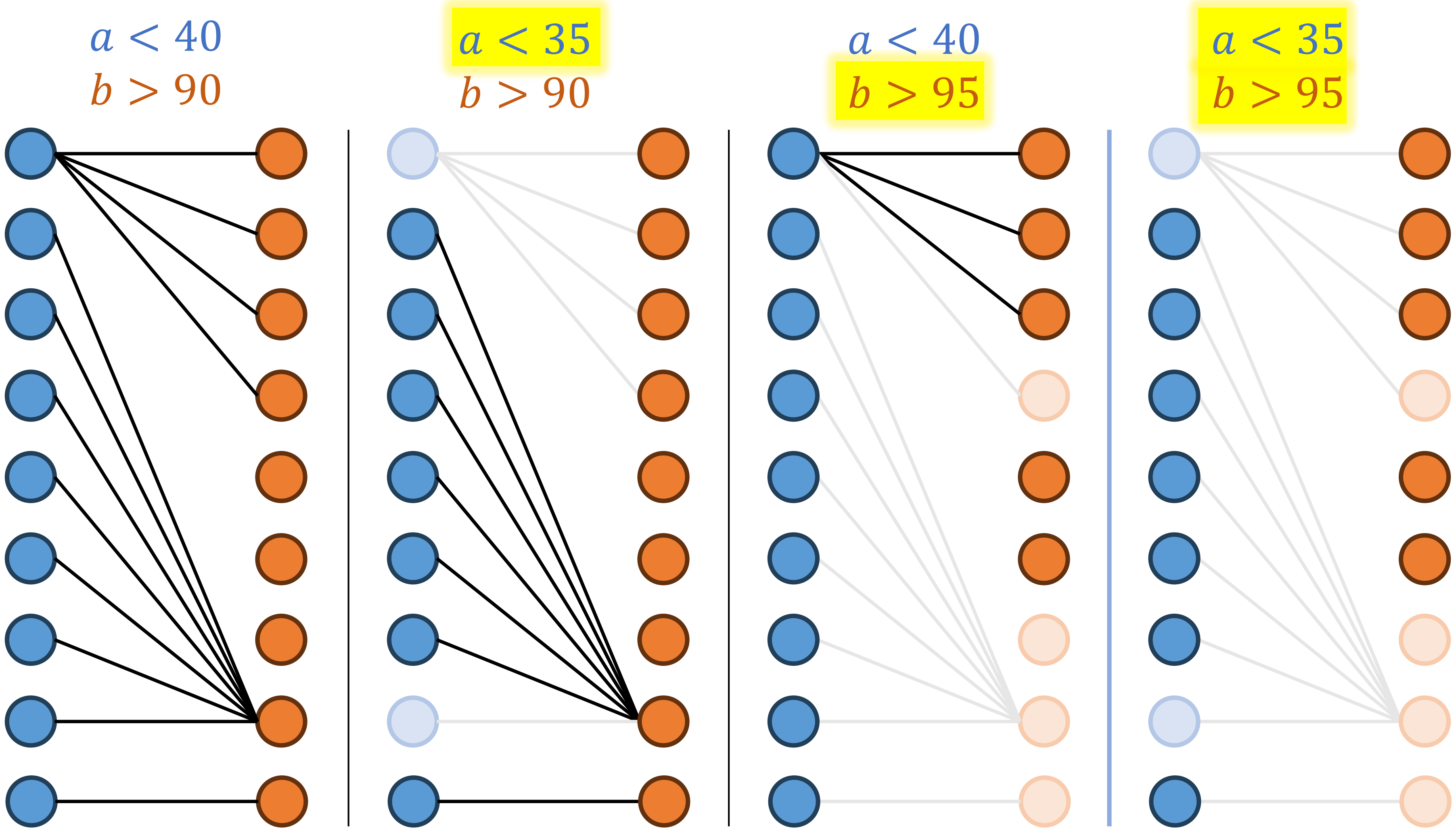
Christoph
Standke

Martin
Grohe

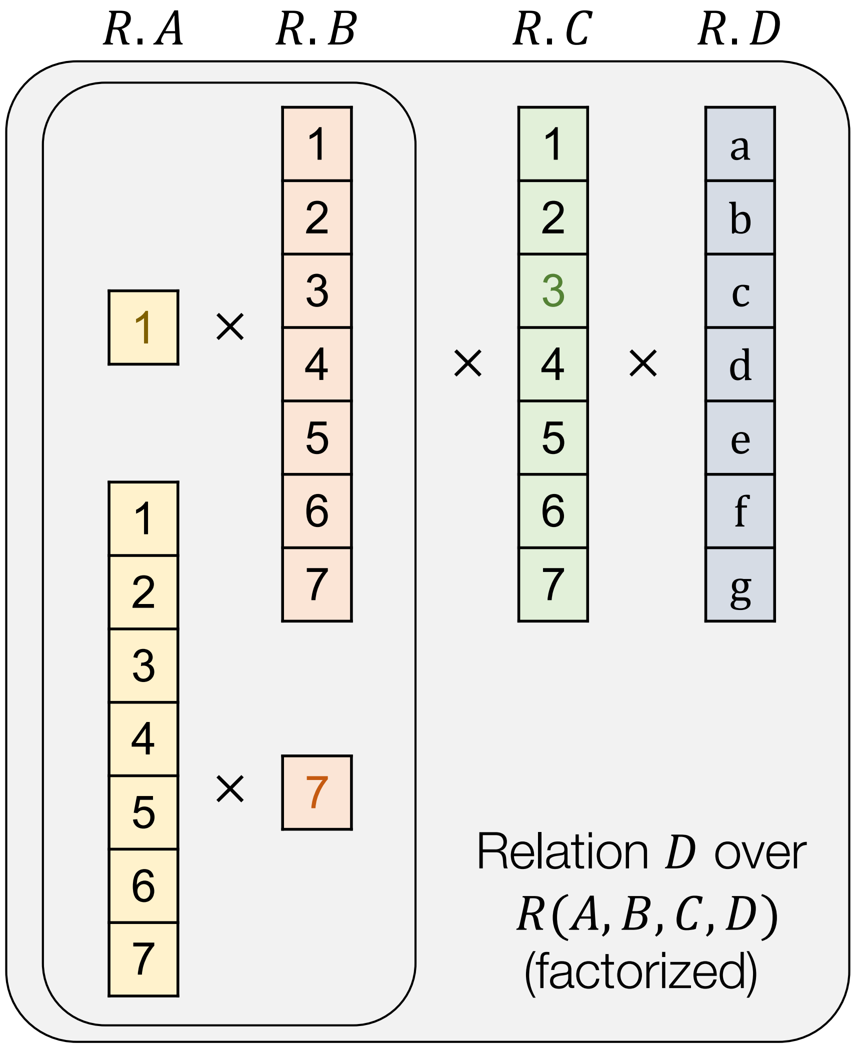
$$\exists x, y \ [\,\text{Salary}(x, a) \wedge \text{Salary}(y, b) \wedge \text{Manages}(x, y) \wedge a < 40 \wedge b > 90\,]$$

How critical are the exact parameter values?

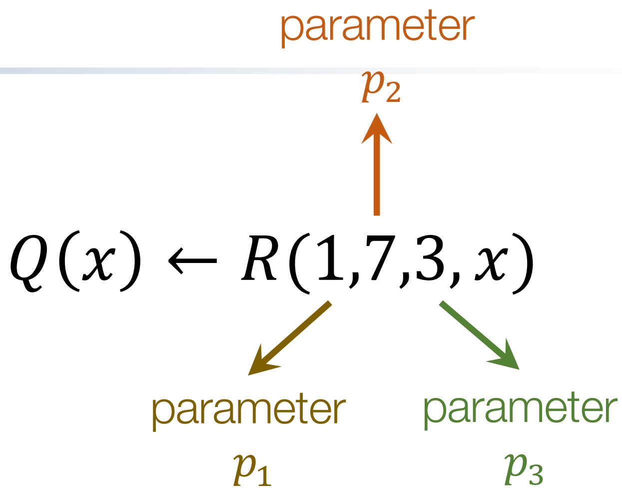
Maybe they are chosen arbitrarily… does it matter?



Another Example



$$Q(D) \coloneqq \{\mathrm{a, b, c, d, e, f, g}\}$$



How arbitrary is the choice of parameter values?

- Changing each of the three alone does not change $Q(D) \coloneqq \{\mathrm{a}, \dots, \mathrm{g}\}$
- The value of p_3 really makes no difference
- What about p_1 and p_2?
  - Changing each *separately* makes no difference
  - … even if p_3 changed in parallel
  - Changing *both* empties the result

Concepts of Sensitivity to Parameters

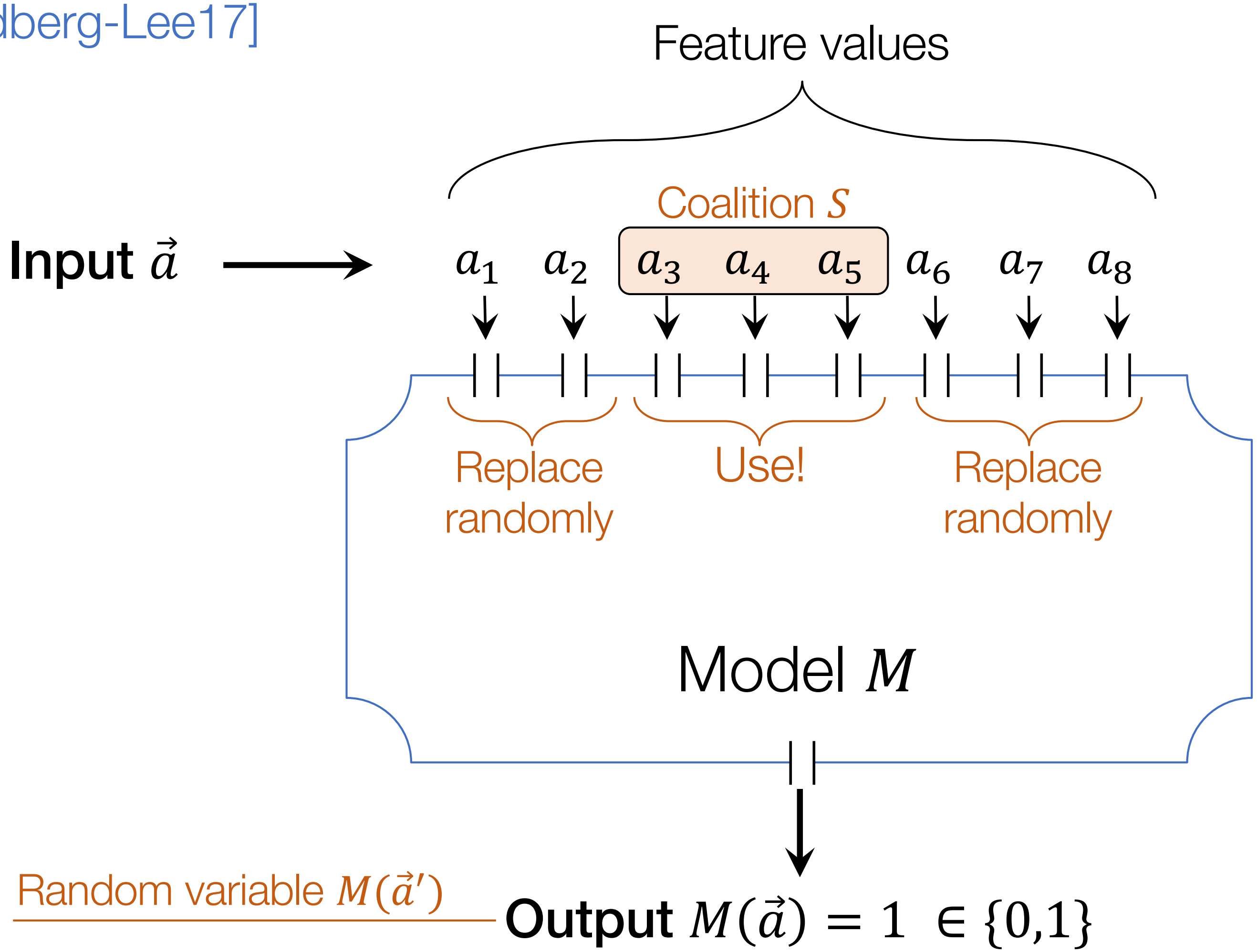
- The *empty-answer* problem: which small param changes cause the result to be nonempty?
  - [Koudas+06] [Mottin+13]
- Parameter perturbations to explain non-answers
  - [Chapman-Jagadish09] [Tran-Chan10]
- Fact checking, cherry-picked queries
  - [Wu+17] [Lin+21]
- We study the application of the Shapley value to assess the contribution of parameters

Parameter Contribution as Coop. Game

- Goal: *assess the contribution of individual parameter values to the outcome*
- What is the cooperative game here?
- Unlike other settings, we cannot just *throw away* parameters outside of the coalition ; what else?
- Similar situation in feature contribution for ML classifiers
  - ⇒ The SHAP score [Lundberg-Lee17]
- We apply a similar approach

The SHAP Score for ML Classifiers

[Lundberg-Lee17]



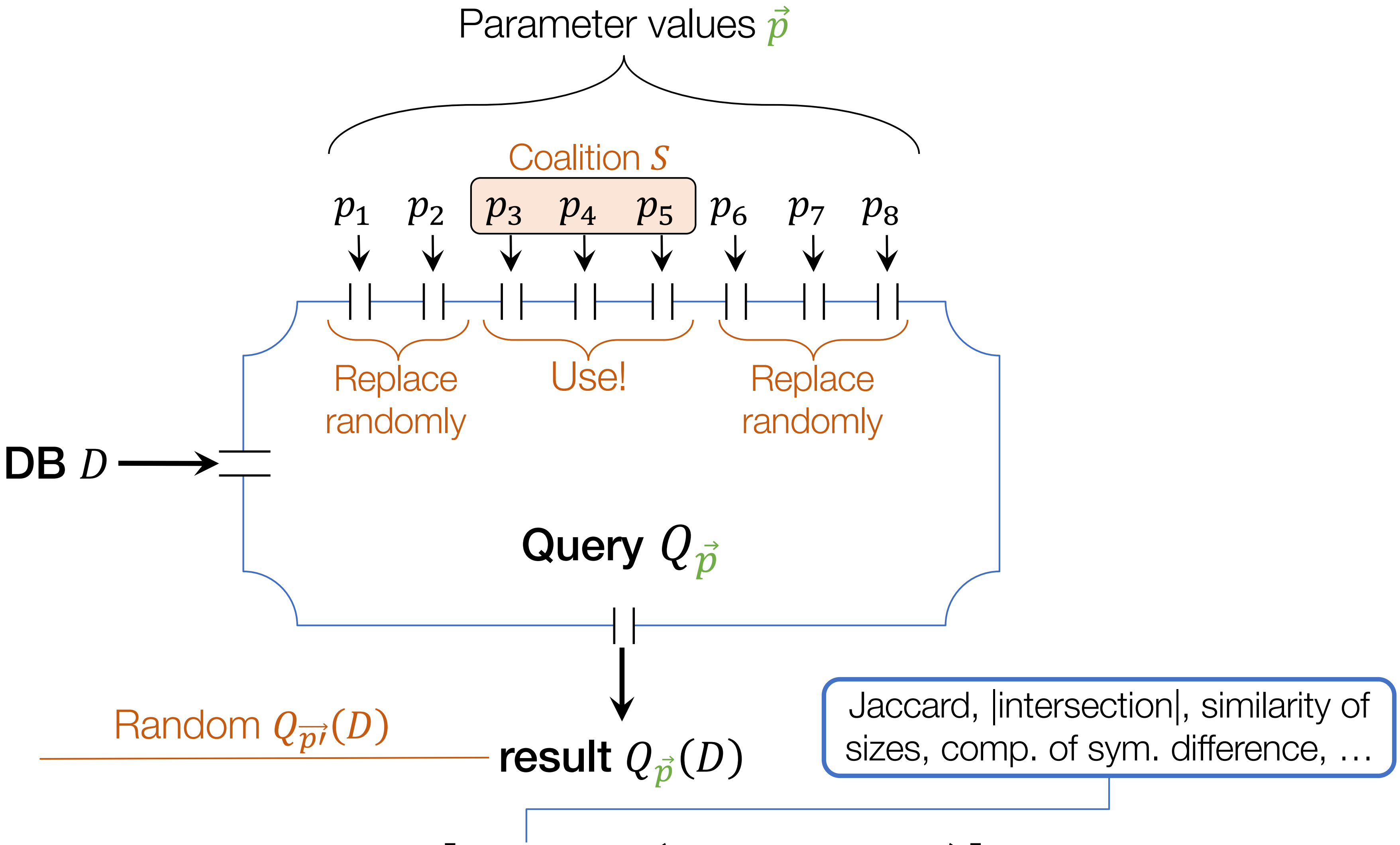
SHAP score: Shapley value for the utility $\nu(S) = \mathbb{E}[M(\vec{a}')]$

Idea: high utility $\Rightarrow$ *values of* S *lead to* $M(x) = 1$ *regardless of the rest*

Adapting SHAP to Query Parameters

- We treat parameters similarly to features
- Assume distributions over parameter values
  - Uniform, perturbations, ad-hoc, …
  - Hence, the query (and result) are random
- Unlike classifiers, the outcome is not binary, but a *set of tuples*
  - Different random changes have different impacts on this set
  - Hence, the utility function compares the random result with the actual result

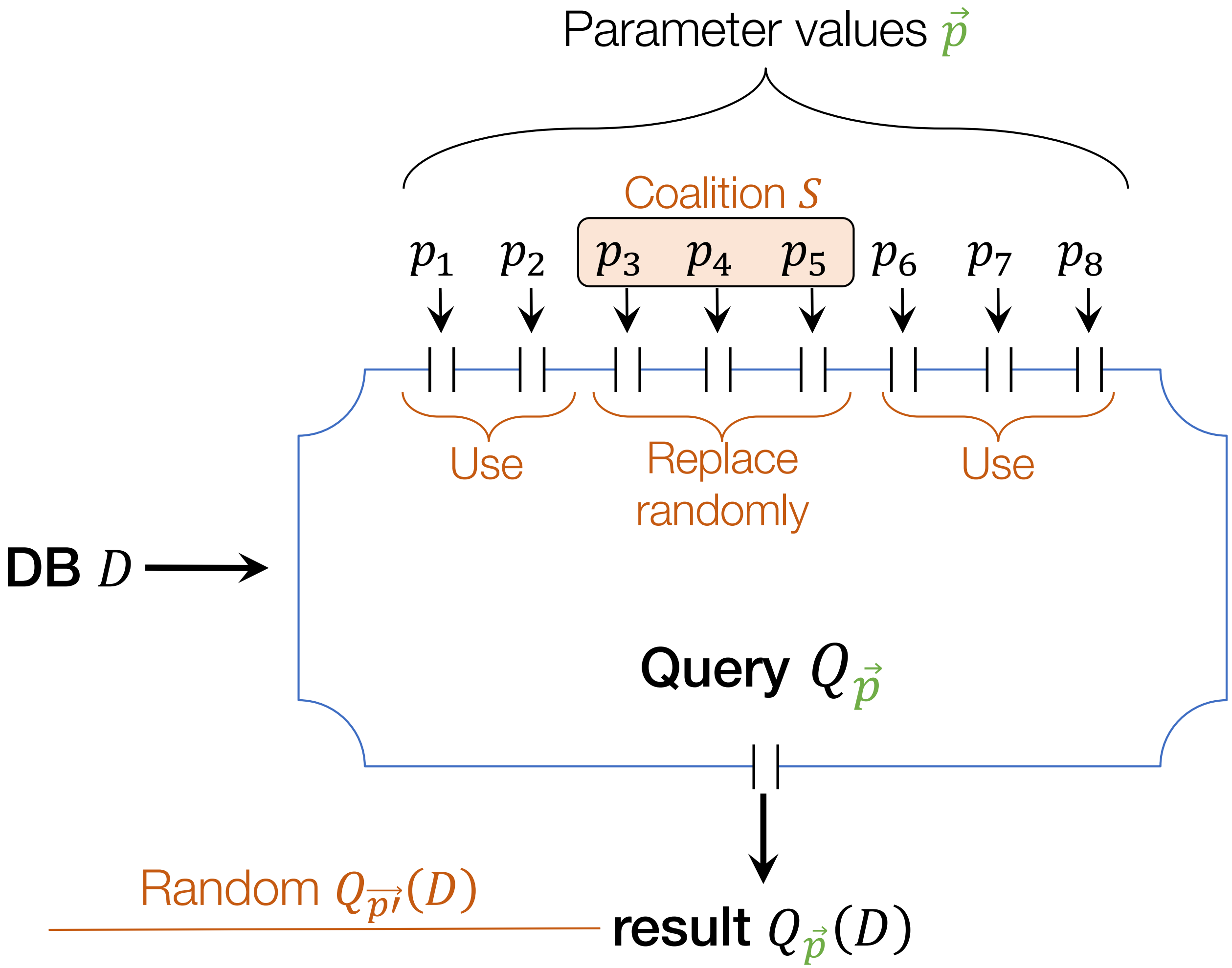
SHAP Score for Query Parameters



$$v(S) = \mathbb{E}\left[\text{similarity}\left(Q_{\vec{p}}(D)\,, Q_{\vec{p'}}(D)\right)\right]$$

Idea: *high utility ⇒ values of S give the actual result, regardless of the rest*

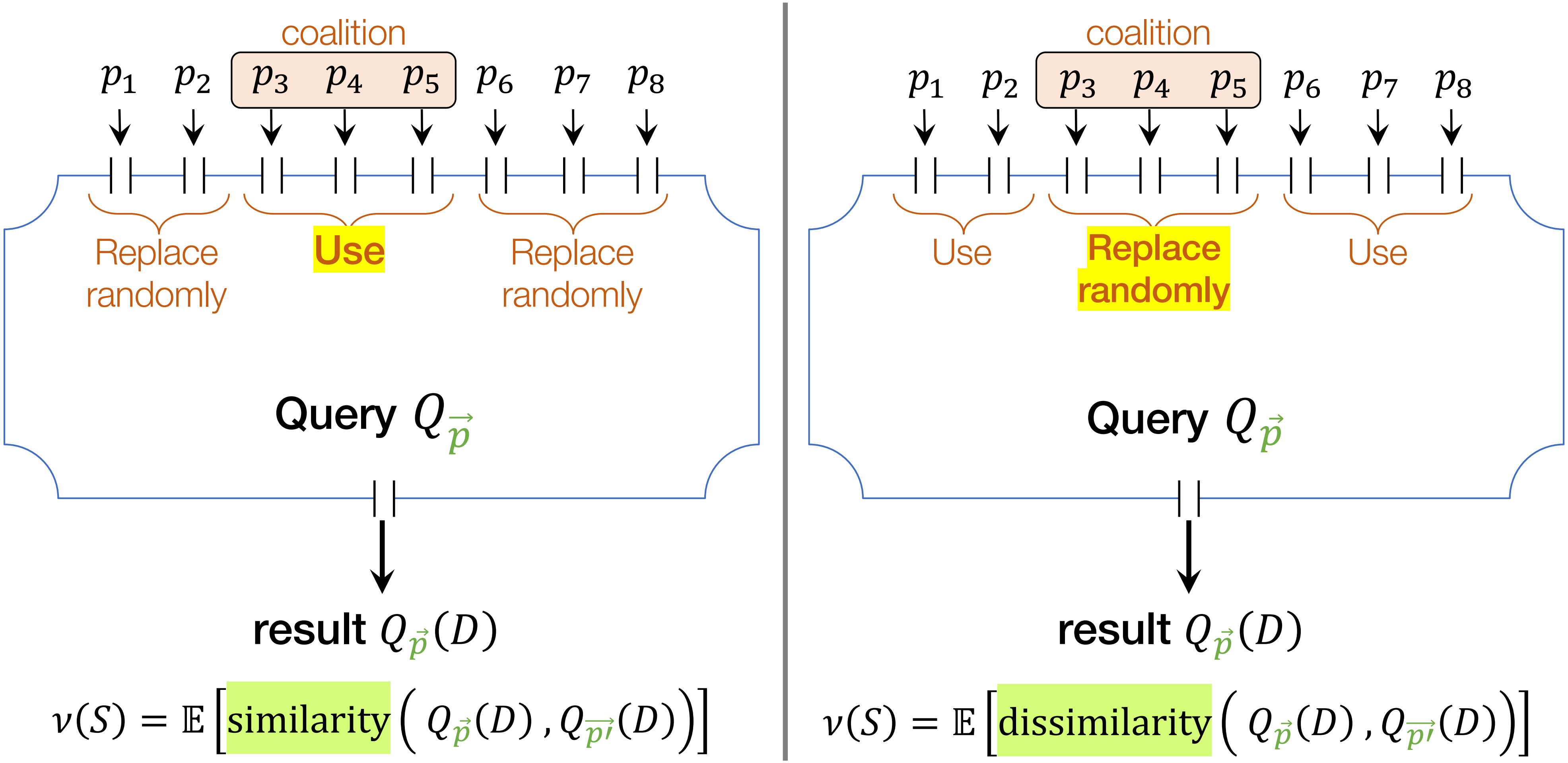
Alternative View



$$v(S) = \mathbb{E}\left[\text{dissimilarity}\left(Q_{\vec{p}}(D)\,, Q_{\vec{p'}}(D)\right)\right] = \mathbb{E}\left[K - \text{similarity}\left(Q_{\vec{p}}(D), Q_{\vec{p'}}(D)\right)\right]$$

Idea: high utility ⇒ changing S greatly impacts the result

Equivalent SHAP Definitions



The two cooperative games lead to the same Shapley value!

Complexity Study

- Algorithms use a general reduction of [Van den Broeck-Lykov-Schleich-Suciu22] from SHAP to expectation calculation
- Polynomial-time algorithms for full acyclic CQs
  - Extends to acyclic CQs with inequalities (e.g., $x < p$)
  - In contrast, even one existential variable can make an acyclic CQ #P-hard
- Efficient approximation scheme under general conditions
  - Conditions – we can efficiently sample parameters, evaluate queries, and calculate similarity

Conclusion

- Contribution measurement in databases: not new (e.g., past proposals based on causality)
- As done in other disciplines, recent efforts to deploy cooperative game theory, specifically the Shapley value
  - Also others, e.g., Banzhaff [Abramovich+23]
- Several deployments: queries, cleaning, …, query design
- Tight connections to probabilistic databases, not fully resolved yet
- Many other directions for future work
  - Database-specific axioms for contribution measures?
  - Non-monotonicity: negation [Reshef+20], non-tuples, non-answers
  - Connection to semiring annotation?
  - Tractability conditions on similarity functions?
  - …

Thank you!