Towards a benefit-based optimizer for Interactive Data Analysis (vision paper)

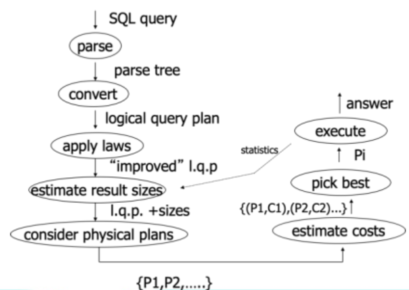
Patrick Marcel, **Nicolas Labroche, Panos Vassiliadis**

Outline

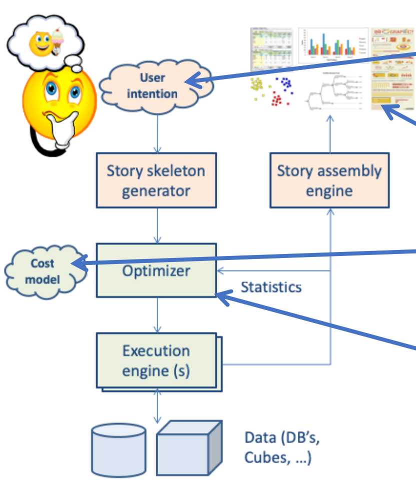
- Challenge
- Vision
- How to
- Perspective

Ten year challenge…

- ❑ **Ten years ago**
  - SQL, MDX queries
  - Tuples as answers
  - TPC-H, SSB
    - ❑ Primary metric: QphH@Size
  - CBO Optimizer
- ❑ **Now**
  - SQL, MDX queries
  - Tuples as answers
  - TPC-H, SSB, TPC-DS
    - ❑ Primary metric: QphH@Size
  - CBO Optimizer



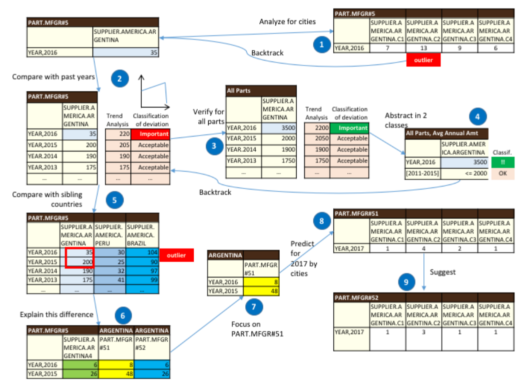
Ten years from now (the vision)



- ❑ **Query: an intention in an high level declarative language**
  - Analyze this, explain that…
- ❑ **Answer: a data story**
  - Set of dashboards with highlights & narratives
- ❑ **Primary metric: the number of insights**
  - Human-digestible pieces of interesting information about the data
- ❑ **Optimizer: concerned with sequences of analytical steps**
  - Select the plan leading to the best insights

Intentions

- **Intentions are non prescriptive**
- **Example**
  - Verify that distribution of sales for mfgr#5 in Argentina from 2011 to 2016 holds in general,
  - build a clustering model for it,
  - compare with sibling countries,
  - explain the highest country-wise difference
- **The optimizer decides**
  - the roll-up(s) for the verification,
  - the algorithm and number of clusters,
  - the way to explain the difference,
  - etc.
- **Each of these degrees of freedom gives rise to a new plan**
  - yielding an answer different from those of the other plans



Insights

- **Insights are diverse**
  - They vary in complexity, value, they are domain-dependent, etc.
- **Insights should be tested for validity**
  - E.g., to avoid the Simpson's paradox [Zhao&al, SIGMOD 2017]
- **Insights are among us**
  - Subjective insights
    - Unexpected values in cubes [Sarawagi, VLDB 2000]
    - Interesting patterns in data [Geng&Hamilton, ACM CompSur. 2006]
    - Surprising patterns in data [De Bie, IDA 2013]
  - Objective insights
    - Statistically significant relationships in datasets [Chirigati&al, SIGMOD 2016]
    - Hidden cause [Sarawagi, VLDB 1999]

Cost model

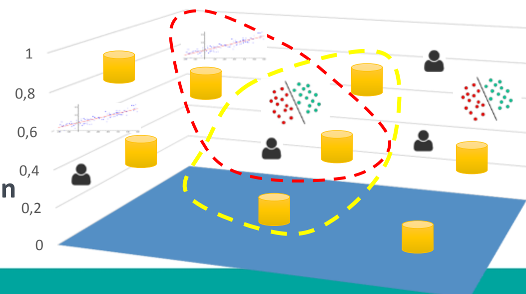
- **Traditional optimizers are concerned with resource consumption**
  - Still needed for "local" optimizations
- **IDA optimizer is concerned with what the user gains from the exploration**
  - It's more a "benefit" model
- **Benefit objective function defined (and learned?) from**
  - the number of insights,
  - the time it takes to obtain them,
  - some properties of insights or sets of insights:
    - their statistical significance
    - their relevance for the user
    - their understandability, diversity, etc.
  - the appropriateness of the insight to the current intention, etc.
- **Traditional optimization schemes still needed**
  - Statistics collection, plan recycling, query re-optimization, etc.

How to generate actions from intentions?

- **Generating queries over data sources**
  - Partly specified by the intention, generated from incomplete specifications [Simitsis&al, VLDBJ 2008], [Vassiliadis&Marcel, DOLAP 2018]
- **Generating ML actions over retrieved sources**
  - Meta-learning [Lemke&al, AIR 2015]
    - How to predict a set of algorithms suitable for a specific problem under study, based on the relationship between data characteristics and algorithm performance
  - Auto-learning [Feurer&al, NIPS 2015]
    - How to choose and parametrize a ML algorithm for a given dataset, at a given cost

How to generate the actual plan?

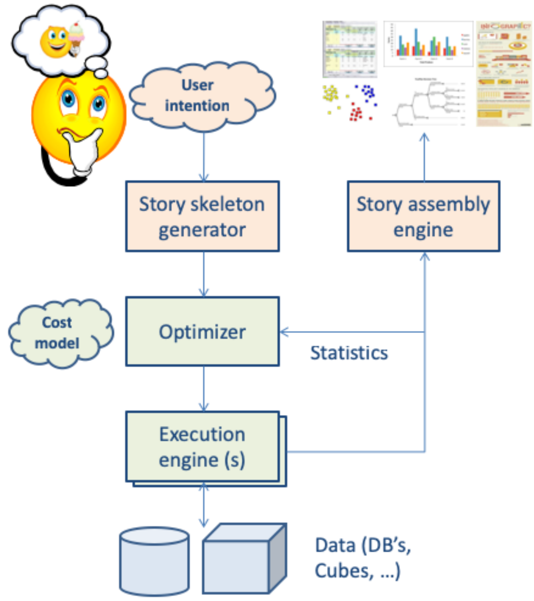
- Generate plan nodes (data sources and actions) from the user intention and current dashboards
- Project nodes in a feature space defined by
  - Data source characteristics
    - As done in meta-learning systems: statistical, information-theoretic and landmarking-based meta-features
  - Actions (queries, ML algorithms) characteristics
    - Complexity, parameters, etc.
- Produce bundles of data sources + actions
  - Using e.g., fuzzy clustering with constraints
    - [Alsayasneh&al, TKDE 2018]
- Prune irrelevant bundles
  - Using e.g., hard constraints on time, number of insights
- Score remaining bundles with the objective function
  - Pick the best one as the plan



Perspectives

- Categorization of insights
- Objective functions
- Mechanisms for statistic collection, user feedback
- Feature space
- Pruning strategy
- …

Thank you! Questions?



The vision:

- **… query via intentions …**
- **… to produce a data story…**
- **… optimized with respect to the best insights!**

http://www.cs.uoi.gr/~pvassil/publications/2018_DOLAP/

References

- [Alsayasneh&al, TKDE 2018] M.Alsayasneh,S.Amer-Yahia,É.Gaussier,V.Leroy,J.Pilourdault,R.M.Bor- romeo, M. Toyama, and J. Renders. Personalized and diverse task composition in crowdsourcing. IEEE Trans. Knowl. Data Eng., 30(1):128–141, 2018.
- [Chirigati&al, SIGMOD 2016] F. Chirigati, H. Doraiswamy, T. Damoulas, and J. Freire. Data polygamy: The many-many relationships among urban spatio-temporal data sets. In SIGMOD, pages 1011–1025. ACM, 2016.
- [De Bie, IDA 2013] T.D.Bie. Subjective interestingness in exploratory data mining.In IDA, pages 19–31, 2013.
- [Eichmann&al, IEEE DEB 2016] P. Eichmann, E. Zgraggen, Z. Zhao, C. Binnig, and T. Kraska. Towards a benchmark for interactive data exploration. IEEE Data Eng. Bull., 39(4):50–61, 2016.
- [Feurer&al, NIPS 2015] M.Feurer,A.Klein,K.Eggensperger,J.T.Springenberg,M.Blum,andF.Hutter. Efficient and robust automated machine learning. In NIPS, pages 2962–2970, 2015.
- [Geng&Hamilton, ACM Comp. Sur. 2006] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. ACM Comput. Surv., 38(3):9, 2006.
- [Lemke&al, AIR 2015] C. Lemke, M. Budka, and B. Gabrys. Metalearning: a survey of trends and technologies. Artif. Intell. Rev., 44(1):117–130, 2015.
- [Milo&Somet, KDD 2018] T. Milo and A. Somech. Next-step suggestions for modern interactive data analysis platforms. In KDD, pages 576–585, 2018.
- [Sarawagi, VLDB 2000] S. Sarawagi. User-adaptive exploration of multidimensional data. In Proceed- ings of VLDB, pages 307–316, 2000.
- [Sarawagi, VLDB 1999] S. Sarawagi. Explaining differences in multidimensional aggregates. In Pro- ceedings of VLDB, pages 42–53, 1999.
- [Simitsis&al, VLDBJ 2008] A. Simitsis, G. Koutrika, and Y. E. Ioannidis. Précis: from unstructured key- words as queries to structured databases as answers. VLDB J., 17(1):117–149, 2008.
- [Vassiliadis&Marcel, DOLAP 2018] P. Vassiliadis and P. Marcel. The road to highlights is paved with good intentions: Envisioning a paradigm shift in OLAP modeling. In DOLAP, 2018.
- [Zhao&al, SIGMOD 2017] Z.Zhao,L.D.Stefani,E.Zgraggen,C.Binnig,E.Upfal,andT.Kraska.Controlling false discoveries during interactive data exploration. In SIGMOD, pages 527–540, 2017.