



# Data Pricing From Economics to Data Science

Jian Pei

[jpei@cs.sfu.ca](mailto:jpei@cs.sfu.ca)

# Outline

- Introduction
- Economics of data pricing
- Fundamental principles of data pricing
- Pricing digital products
- Pricing data products
- Future directions

---



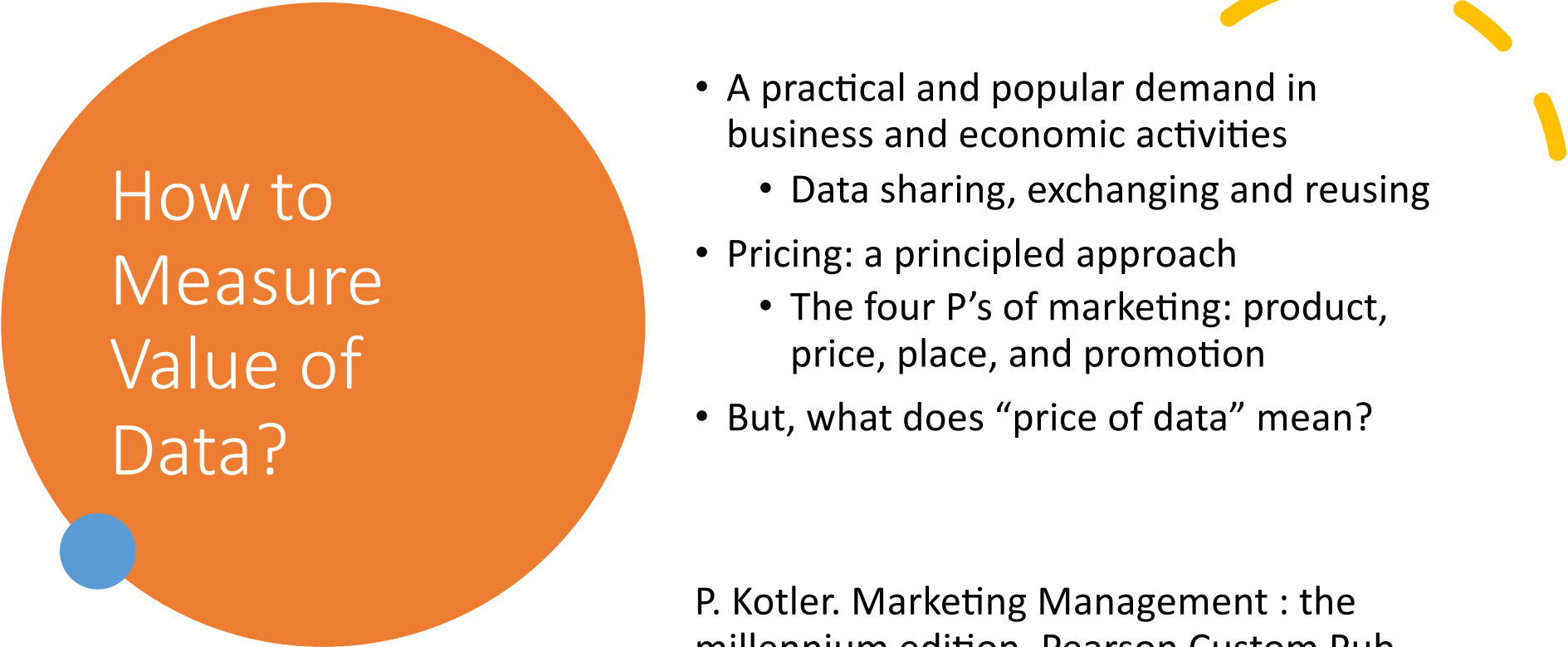
# Introduction

# Data as Fundamental Resource

- Products and services delivered in digital forms
- Second use of big data
- Sharing and reusing data has profound implications to economy
  - Case study: Nagaraj finds that mining activities, particularly by smaller firms with less resources, were strongly benefited by open maps or maps sponsored by governments
- Fairness and innovation: universal availability of data often helps minority parties and emerging initiatives

A. Nagaraj. The private impact of public information: Landsat satellite maps and gold exploration. Unpublished, 07 2016.





# How to Measure Value of Data?

- A practical and popular demand in business and economic activities
  - Data sharing, exchanging and reusing
- Pricing: a principled approach
  - The four P's of marketing: product, price, place, and promotion
- But, what does “price of data” mean?

P. Kotler. Marketing Management : the millennium edition. Pearson Custom Pub., Boston, MA, 2000.

# Scenario 1: Data Transmission

- Example: a mobile service provider offers a smart phone user the price of its data package
- Price: data transmission service
  - Factors: data amount quota, location, transmission speed, ...
  - Often independent from data content, data quality, data collection, data storage, and data processing





## Scenario 2: Digital Products

- Example: watch a movie at home, where the movie is delivered as a stream of data
- Price: content product/service
  - Factors: content, resolutions, ...
  - Often independent from data transmission service

# Scenario 3: Data Products

- Example: use weather forecasting services in business operations
- Price: subscription/data product/services
  - Factors: granularity, long/short term, analysis requirements, business vertical relevance
  - Often independent from data delivery, standard information for public





# Data Pricing Is Far From Trivial

- Many scenarios exist
- Highly interdisciplinary
  - Economics, computation and data science are fundamental
- This tutorial provides a comprehensive survey

# Data as Products/Services: Categorization

- Data and information as goods: distributed purely in digital form
- Digital products: intangible goods but can be consumed through electronics
  - Examples: e-books, downloadable music, online ads, and internet coupons
  - Many digital products have physical correspondences in one way or another, though not absolutely necessary
- Data products: data sets as products and information services derived from data sets
- Information goods include digital products and data products

# Roadmap (1)

- Economics of data pricing
  - Cost reduction in information goods and impact on pricing
  - Differences between digital products and data products
- Fundamental principles of data pricing
  - Versioning: a general framework for pricing information goods
  - Desirable properties in data pricing, including trustfulness, fairness, revenue-maximization, arbitrage-freeness, privacy preservation, and computational efficiency
- Pricing digital products
- Pricing data products
- Challenges and future directions

# Roadmap (2)

- Economics of data pricing
- Fundamental principles of data pricing
- Pricing digital products
  - Three major streams of revenues for digital products
  - Bundling and subscription planning pricing models
  - Auctions
- Pricing data products
  - Overview
  - Arbitrage-free pricing
  - Revenue maximization pricing
  - Fair and truthful pricing
  - Privacy preserving pricing
  - Pricing dynamic data and online pricing
- Challenges and future directions





# Economics of Data Pricing

# What Is Pricing?

- The practice that a business sets a price at which a product or a service can be sold
- Often part of the marketing plan of a business
- Objectives in pricing
  - Profitability
  - Fitness in marketplace
  - Market positioning
  - Price consistency across categories and products
  - Meeting or preventing competitions
  - ...

# Major Pricing Strategies

- Operation-oriented pricing
- Revenue-oriented pricing
- Customer-oriented pricing
- Value-oriented pricing
- Relationship-oriented pricing
- Subject in economics and marketing research
  - Beyond the scope of this tutorial

R. Brennan, L. Canning, and R. Mcdowell. Business-to-business market- ing. 01 2013.

J. Nagle, T.T. & Hogan. The Strategy and Tactics of Pricing: A Guide to Growing More Profitably. Prentice Hall, prentice hall edition, 2010.

# “Technology changes. Economic laws do not.”

- Cost reduction is the core in production, distribution and consumption of information goods comparing to physical products
- Cost reduction in information goods
  - Search costs
  - Production costs
  - Replication costs
  - Transportation costs
  - Tracking and verification costs
- Digital and data economics: investigation of how standard economic models adjust when major costs are reduced dramatically.

A. Goldfarb and C. Tucker. Digital economics. *Journal of Economic Literature*, 57(1):3–43, March 2019.

# Search Costs

- The costs of looking for information
- Information goods allow more effective and efficient online search
  - Easier for users to discover digital products and data products
  - Easier for price comparison
  - Example: online prices of books and CDs are clearly cheaper than offline
- Low search costs facilitate the sales of rare and long tail products
  - Often more variety in information goods and services
  - Example: online media consumption is more diverse than offline
  - Recommender systems play an important role in deciding the degree of variety
  - Echo chamber effect: customers may tend to consume more that aligns with their viewpoints



## Low Search Costs Lead to Platforms

- Platform businesses: provide extensive matching services to customers and improve trade efficiency
- Strategies for building platforms and running platform businesses
  - Interoperability
  - Compatibility
  - Standards

H. Halaburda and Y. Yehezkel. Platform competition under asymmetric information. *American Economic Journal: Microeconomics*, 5(3):22–68, 2013.

# Production Costs

- A wide spectrum of production costs in traditional products are substantially reduced in information goods
  - Some essential major costs in traditional production are dramatically reduced or even approach zero in producing information goods, such as materials, semi-finished products and their transportation
  - Unit costs of information goods can approach zero through sharing
  - Information goods often can reduce the costs of customization to extreme
- The substantial reduction in production costs gives rise to a series of innovative business models, such as economics of sharing, pay-as-you-go, and query-based data consumption
- Facilitate innovation and long tail products

# Replication Costs

- Information goods are nonrival
- Bundling is often used for zero-marginal cost, non-rival information goods, often in the scale of thousands of products
- Making data public, such as Wikipedia and open source software
  - Individual contributors: demonstrate their professional skills to potential employers
  - Companies support those products to complement their sales on other products



# Implications of Low Replication Costs

- Benefit: low replication costs, though may reduce revenue, help supplies and demands, and thus boost quality
- Challenge: the zero marginal costs and non-rivalrous property pose challenges to copyright policies and enforcement
  - The protection of intellectual properties indeed has negative impact on follow-on innovation in gene sequencing
- Governments mandate “open data” may lead to data leakages and privacy breaches that affect citizens’ offline welfare
- The zero marginal costs or non-rivalrous nature also ease the way for spamming and online crime

# Transportation Costs

- Thanks to the Internet, the costs of transporting information goods approach zero
  - The effect of flat world: local communities may not affect adoptions and consumptions of information goods
  - At the same time, some studies demonstrate that tastes may still be local in music and content consumption
- Regulation may put sophisticated constraints on locations
  - Example: when Wikipedia was blocked in China in October 2005, more contributors from outside China were motivated to contribute
- Copyright policies may affect the availability and consumption of information goods, such as news media, in different regions, and thus may be reflected by price

# Tracking and Verification Costs

- Information good providers can track users with relatively low costs
- Implications: extensive personalized markets and possible price discrimination
- Example: behavioral price discrimination
  - Set prices according to customers' previous behavior
  - if customers are well aware of the benefits of tracking information to a monopoly, they may likely choose to be privacy sensitive and hold the information
- Example: versioning
  - Sell information at different prices to different customers using different versions, more discussion later

# Personalized Advertising

- Taking advantage of low tracking and verification costs
- Challenge: how should a company set prices for many advertisements that may be shown to massive customers?
  - Auctions can be used to discover prices for information goods
  - Auctions may be less useful when online marketplaces become mature



# Privacy Concerns and Trust

- Should privacy be treated as goods?
- How should privacy be priced?
- Privacy regulation and the impact on welfare
- The low verification costs facilitate online transactions extensively and lower the costs of trust dramatically

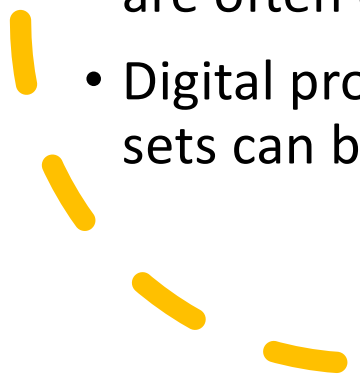
# Digital Products and Data Products: Differences

- The units of digital products are often well defined and fixed
- The consumption of a digital product is often independent from each other
- One individual unit of data products at the lowest granularity may not be valuable
- More often than not, many basic units of data are combined, aggregated and consumed together



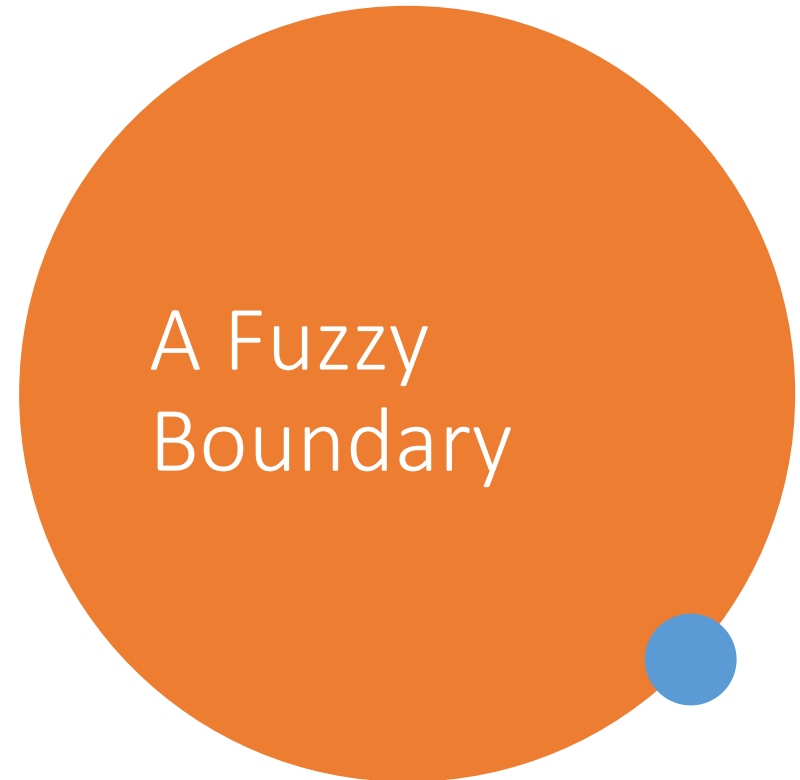
# Digital Products and Data Products: Differences

- Data sets as data products have very strong and flexible aggregate-ability
  - The aggregate-ability enables many opportunities for innovations in data business and posts many technical and business challenges
- Typically digital products are consumed directly by people, data sets are often consumed by computers
- Digital products are easy to be consumed by others in whole, data sets can be reused by others in different ways





- The same information can be regarded as digital products in some situations and as data products in some other situations
- Example: social media
  - As digital products when a customer reads them online
  - Be collected and processed in batch as data products





# Summary

- Information goods distinguish themselves from the traditional physical products in significant cost reductions
- Major cost reduction: search costs, production costs, replication costs, transportation costs, and tracking and verification costs
- Major differences between digital products and data products: consumption units, aggregate-ability, means of consumption, and reusing and reselling
- The boundary between digital products and data products is sometimes fuzzy

---

# Fundamental Principles of Data Pricing



# Fundamental Principles

- Versioning: a fundamental framework
- Important desiderata
  - Truthfulness
  - Revenue maximization
  - Fairness
  - Arbitrage-free Pricing
  - Privacy-preservation
  - Computational Efficiency



## Challenges in Pricing Information Goods

- Replication costs very low or even approaching zero → the price of an information good tends to be very low
- Advantage: information goods economically appealing
- Disadvantage: information goods economically dangerous
  - Competitors may easily enter the market

# Linking Price to Value

- Setting the price reflecting the value that a customer places on the information
- Versioning strategy: making different versions to appeal to different types of customers
- Examples: software, movies, ...
- Versioning divides customers into subgroups so that each subgroup may regard some features highly valuable and some other features of little value

C. Shapiro and H. R. Varian. Versioning: The smart way to sell information. Harvard Business Review, pages 106–114, November-December 1998.

# How to Produce Different Versions?

- Delay in information delivery
- Access convenience
- Comprehensiveness of information
- Information manipulation
- User community
- Annoyance
- Customer support
- ...
- Most versions of information goods are created by subtracting value from the most technologically advanced and complete version

# Free Versions

- Customers may not realize the value of an information good unless they try it
- Free versions can provide opportunities to potential customers to test out
  - Building awareness
  - Gaining follow-on sales
  - Creating a customer network
  - Attracting attentions
  - Gaining competitive advantages



## How Many Versions Should Be Made?

- Depending on the characteristics of the information
  - How many different ways an information good may be used?
- Depending on the value that different customers may place on it
- Example: views of relational data



# Important Desiderata in Data Pricing

- Truthfulness
- Revenue maximization
- Fairness
- Arbitrage-free Pricing
- Privacy-preservation
- Computational Efficiency

# Truthfulness

- A market is trustful if every buyer is selfish and only offers the price that maximizes the buyer's true utility value
- In a trustful market, no buyer pays more than sufficient to purchase a product
- Truthfulness can facilitate a wide spectrum of pricing mechanisms, such as many kinds of auctions

# Revenue Maximization

- Rationale: for a business to be successful long term, a more immediate and important requirement is to win over as many customers as possible
- Maximizing revenue instead of optimizing cost, profit or sales
- Traditional physical products: revenue is maximized when marginal revenue becomes zero
- For information goods, replication costs are very low, revenue maximization and profit maximization for information products become quite different

# Fairness

- A market is fair if each seller gets the fair share of the revenue in coalition
- Shapley fairness:  $k$  sellers cooperatively participate in a transaction that leads to a payment  $v$ 
  - Balance: the payment is fully distributed to all sellers
  - Symmetry: the same contribution to utility should be paid the same
  - Zero element: no contribution, no payment
  - Additivity: if the goods can be used for two tasks  $T1$  and  $T2$  with payment  $v1$  and  $v2$ , respectively, then the payment to complete both tasks  $T1 + T2$  is  $v1 + v2$

L. S. Shapley. A Value for  $n$ -Person Games. Technical Report P-295, RAND Corporation, Santa Monica, CA, 1952.

# Shapley Value

- The Shapley value is the unique allocation of payment that satisfies all requirements in Shapley fairness

$$\psi(s) = \sum_{S \subseteq D \setminus \{s\}} \frac{\mathcal{U}(S \cup \{s\}) - \mathcal{U}(S)}{\binom{n-1}{|S|}}$$

- Equivalently 
$$\psi(s) = \frac{1}{N!} \sum_{\pi \in \Pi(D)} (\mathcal{U}(P_s^\pi \cup \{s\}) - \mathcal{U}(P_s^\pi))$$

- Challenges in information goods: the marginal costs of production are close to zero, a seller can produce more units of the same information good to obtain a larger Shapley value

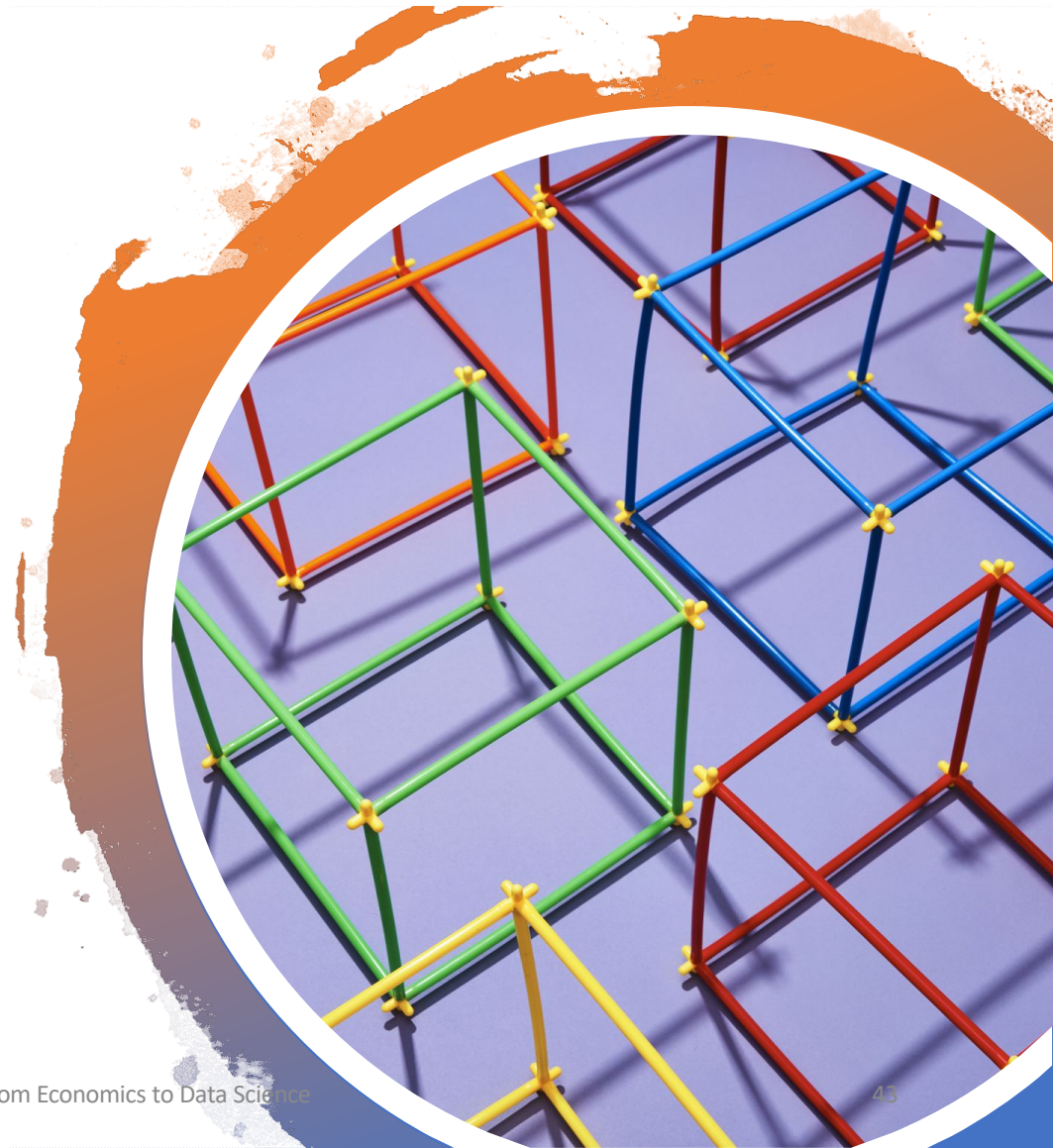
# Arbitrage-freeness

- Arbitrage is the activities that take advantage of price differences between two or more markets or channels
  - Example: an article list price: \$35, monthly subscription rate \$25
- Arbitrage is often undesirable in pricing models
- At least it should be able to check whether a pricing model is arbitrage-free
  - Example: a data service provider sells query results with prices based on variance, a variance of 10 for \$5 each query result and a variance of 1 for \$100 each query result. Each answer is perturbed independently. A user can purchase the result of a query 10 times and get the average

C. Li, D. Y. Li, G. Miklau, and D. Suciu. A theory of pricing private data. *ACM Trans. Database Syst.*, 39(4), Dec. 2015.

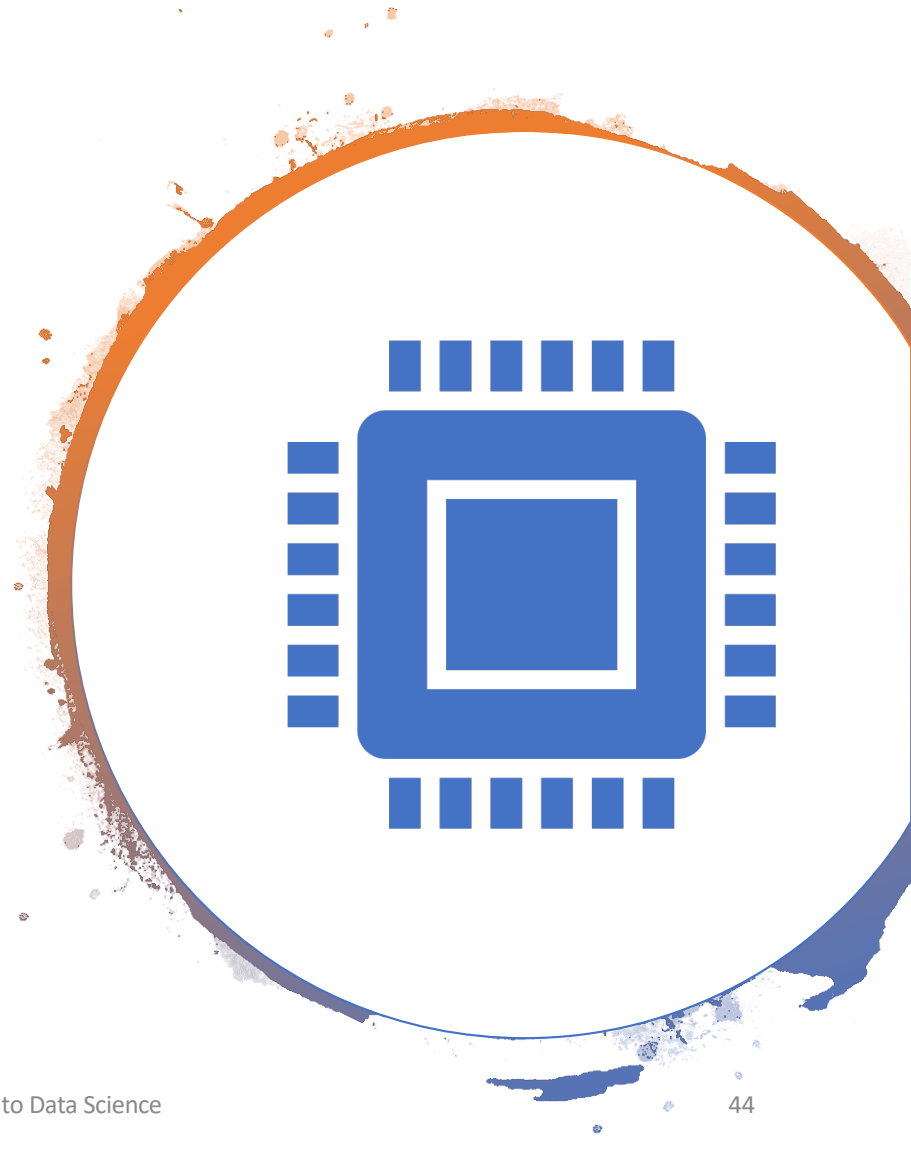
# Privacy-Preservation

- Highly desirable to preserve privacy in marketplaces of information goods
- Transactions in a marketplace may disclose privacy of various parties in many different ways
  - Privacy of buyers
  - Privacy of providers of information goods
  - Privacy of a third party involved
- Many privacy-preservation approaches have been tried



# Computational Efficiency

- Computing prices efficiently with respect to a large number of goods and a large number of buyers
- The complexity of computing prices has to be polynomial with respect to the number of sellers, and cannot grow with respect to the number of goods/buyers when prices are updated
- Auction efficiency





---

# Summary

- Versioning is a common mechanism in designing and pricing information goods
- A series of important requirements on pricing information goods, including truthfulness, revenue maximization, fairness, arbitrage-free pricing, privacy preservation, and computational efficiency



# Pricing Digital Products

J. Pei: Data Pricing -- from Economics to Data Science

46

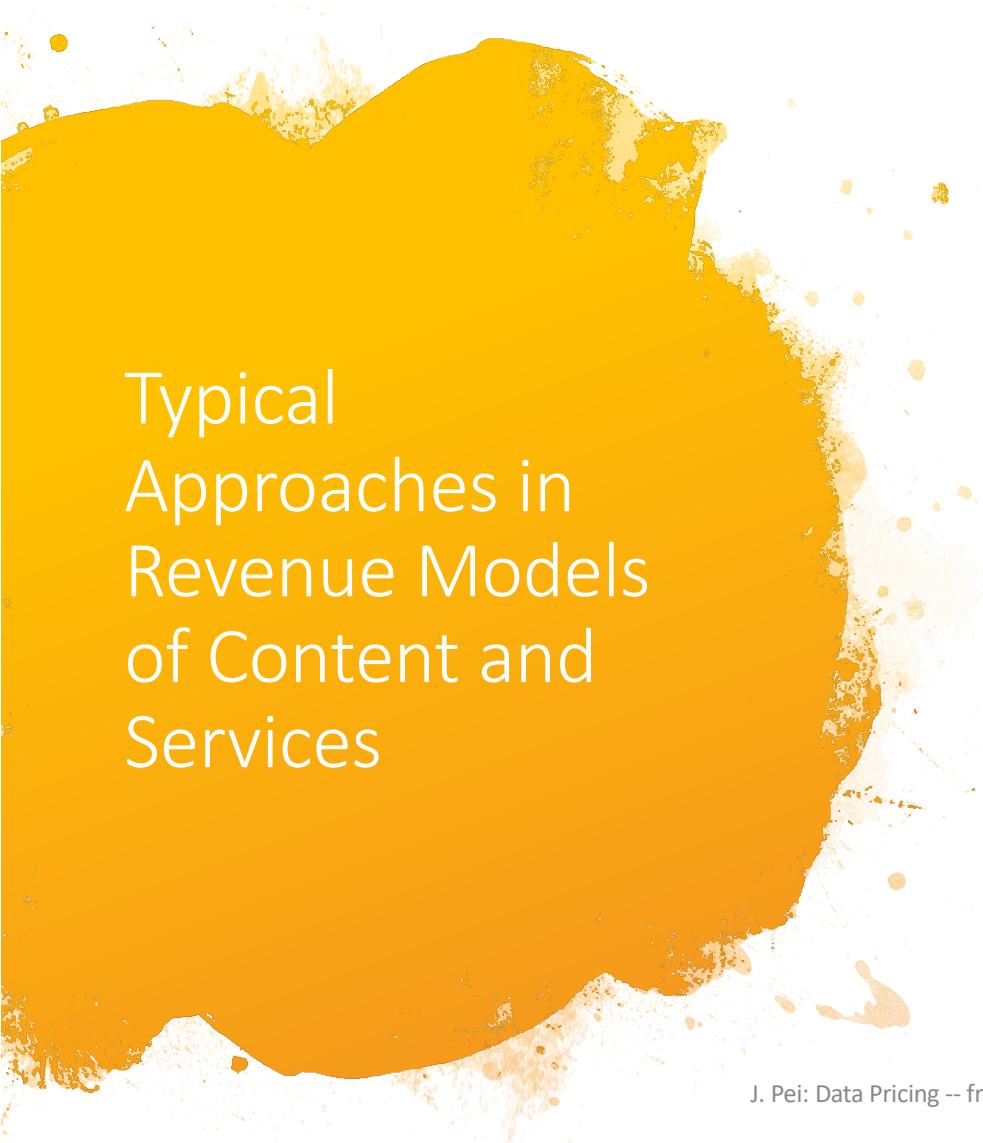
# A Brief Review on Pricing Digital Products

- Three major streams of revenues for digital products
- Two major types of pricing models
- Some general ideas in pricing digital products can be borrowed and extended to data products

# Streams of Revenues

- Revenue maximization often serves as the basic objective in pricing mechanisms
- Three streams of revenues for digital products that are delivered online
  - Money: a provider can sell to customers content, or more broadly services, such as movies and e-books
  - Information/privacy: a provider can collect customer information by tracking (e.g., using cookies) and sell the information about customers to generate revenues
  - Time/attention: a provider can sell space in their digital products to advertisers to produce revenue

A. Lambrecht, A. Goldfarb, A. Bonatti, A. Ghose, D. Goldstein, R. Lewis, A. Rao, N. Sahni, and S. Yao. How do firms make money selling digital goods online? *Marketing Letters*, 25:331–341, 09 2014.



## Typical Approaches in Revenue Models of Content and Services

- Rigid pricing
- Designing pricing tiers
- Setting up duration of subscription plans
- Designing freemium models

A. Rao. Online Content Pricing: Purchase and Rental Markets. *Marketing Science*, 34(3):430–451, May 2015.

# Micropayments

- A unique feature in digital product consumption
- A customer can pay a very small amount
  - Typically impractical in traditional transactions using standard credit cards due to the network service fees
- Micropayments and subscriptions have different effects on consumer behavior

S. Athey, E. Calvano, and J. Gans. The impact of the internet on advertising markets for news media. Working Paper 19419, National Bureau of Economic Research, September 2013.

# Example: Pricing Software Products

- S. Lehmann and P. Buxmann. Pricing strategies of software vendors. Business & Information Systems Engineering, 1:452–462, 12 2009.

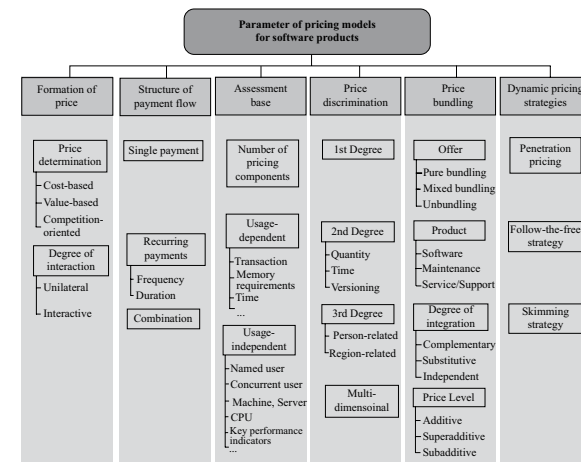


Fig. 1 Parameters of pricing models for software products

# Information/privacy Stream: Selling Customer Data

- A company can generate revenues from information/privacy stream
  - Example: customers' identities, behavior patterns, preferences and needs
- There are various ways to sell customer data
- Example: a website can
  - Provide direct marketing companies user activity information
  - Collaborate with data management platforms (DMP, for advertising) and produce revenues by facilitating businesses to identify audience segments
- Pricing model examples
  - Design customized discounts in marketing campaigns based on social networks
  - pricing customer-level information such that the data about each customer is sold individually and individual queries to the database are priced linearly





## Time/Attention Stream

- Embed advertisements in products
- Challenge: hard to accurately measure advertising effects
- Combine user information and advertising opportunities, such as retargeted advertising



# Bundling

- Product bundling organizes products or services into bundles, such that a bundle of products or services are for sale as one combined product or service package
  - A common marketing practice, particularly in the traditional industry like telecommunication services, financial services, health-care, and consumer electronics
- Designing product bundles essentially is a combinatorial optimization problem

# Grand Bundle: Settings and Assumptions

- Settings:  $n$  heterogeneous products for one buyer
- Objective: maximize expected revenue
- The independent product value distribution assumption: for each product  $x_i$ , the price that the buyer would like to pay for is an arbitrary distribution  $D_i$  in range  $[a_i, b_i]$ , where  $0 \leq a_i \leq b_i < \infty$ , and those distributions  $D_1, \dots, D_i$  are independent from each other
- Additive buyer: the buyer's value for a set of products is the sum of the buyer's values of those individual products in the set

R. B. Myerson. Optimal auction design. Math. Oper. Res., 6(1):58–73, Feb. 1981.

# Grand Bundle: Pricing Model

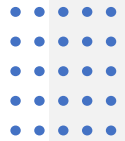
- Theoretical result: either selling each item separately or selling all items together as a grand bundle produces at least a constant fraction of the optimal revenue
- Pricing model: either pricing each product individually or pricing the grand bundle in the expected price
- Examples: Hulu and Amazon Prime Video offer
- Grand bundle is optimal if customers with higher values for the grand bundle indeed have higher relative values for smaller bundles compared to the grand bundle
- It is easier to price a bundle of a larger number of products

C. Daskalakis, A. Deckelbaum, and C. Tzamos. Strong duality for a multiple-good monopolist. *Econometrica*, 85(3):735–767, 2017.

# Subscription

- Price the interactions between customers and a platform over a period of time
- Subscribing customers are in general heterogeneous in both usage rate and value of products
- Subscription model: select a subscription fee and the period for each set of products and also set the rental price for each product
- grand subscription: a single rental price for the set that includes all products
- Subscription fees can be set proportional to the cardinality of a set of products and can achieve  $\frac{1}{4\log 2m + \log n}$  of the optimal revenue for  $n$  types of customers and  $m$  types of products

S. Alaei, A. Makhdoumi, and A. Malekian. Optimal subscription planning for digital goods. SSRN Electronic Journal, 01 2019.



# Auctions

- A long history back to the Babylonian and Roman empires
- Many excellent surveys
- Our focus here: the important role of auctions as a pricing mechanism for digital products


# Four Basic Types of Auctions

- Ascending-bid auction (aka English auction): the price is raised successively until only one bidder remains, who wins the object at the final price
- Descending auction (aka the Dutch auction): start at a very high price and lower the price continuously, until the first bidder calls out and accepts the current price
- First-price sealed-bid auction: every bidder submits a bid without knowing the others' bids, and the one making the highest bid wins and pays at the named price
- Second-price sealed-bid auction (aka the Vickrey auction): every bidder submits a bid without knowing the others' bids, and the one making the highest bid wins and pays only the second highest bid



# Value Information in Auctions



- Private-value model: every bidder has an independent value on the object for sale
  - Pure common-value model: the actual value of the object for sale is the same for all bidders, but bidders have different private information about that actual value
  - There are also models considering both values private to individual bidders and common to all bidders
- 



# Revenue Equivalence Theorem

- A fundamental principle in auction theory
- for a set of risk-neutral bidders with independent private valuation of an object drawn from a common cumulative distribution that is strictly increasing and atomless on  $[v_{\min}, v_{\max}]$ , any auction mechanism yields the same expected revenue and thus any bidder with valuation  $v$  making the same expected payment if (1) the object is allocated to the bidder with the highest valuation; and (2) any bidder with valuation  $v_{\min}$  has an expected utility of 0
- Based on the revenue equivalence theorem, the four basic types of auctions lead to the same payment by the winner and the same revenue

W. Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961.

# Sponsored Search Auctions

- Sponsored search: content providers pay search engines for traffic to their websites
  - Content providers bid for keywords in search engines, and search engines decide which ad to display in which position to answer a query from a user
- Many pricing models, such as pay-per mille/pay-per impression (PPM), pay-per-click (PPC), and pay-per- action (PPA)
- Early days a generalized first price auction is used
- Google generalizes the second price auction mechanism and enhances the ranking of bids by additional information

# Recent Research on Sponsored Search

- Analysis of auction mechanisms based on assumptions about rationality, budget constraints and CTR distributions
- Practical sponsored search systems and auction mechanisms when the standard assumptions do not hold
- Empirical studies to understand bidding behavior and statics
- Deep learning approaches to develop auction strategies in sponsored search

# Auctions on Digital Products with Unlimited Supplies

- Digital products may have unlimited supply
- Challenge: how to ensure the bids are truthful?
- Generalize the second price auction: the top  $k$  highest bidders win and each pays the  $(k + 1)$ -th bidding price
- Denote by  $B$  the set of bidders, and by  $b_1, b_2, \dots$  the bidding prices in descending order
- maximize  $k \cdot b_{k+1}$

# Competitive Auctions

- An auction is competitive if it yields revenue within a constant factor of optimal fixed pricing
- When there is unlimited supply, the Vickrey auction is
  - Not competitive if the seller chooses the number of products to sell before knowing the bids
  - Not truthful if the seller chooses after knowing the bids

A. V. Goldberg, J. D. Hartline, and A. Wright. Competitive auctions and digital goods. In Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'01, pages 735–744, USA, 2001. Society for Industrial and Applied Mathematics.

# Random Sampling Auction

- An auction is bid-independent if bidder  $i$ 's bid value should only determine whether the bidder wins the auction, but not the price
- Auction mechanism
  - Select a sample  $B'$  of  $B$  at random, independent from the bid values
  - Use the bids in  $B'$  to compute the optimal bid threshold  $f_{B'}$  that maximizes the revenue in  $B'$
  - Every bidder in  $B - B'$  whose bid value is over  $f_{B'}$  wins
  - Symmetrically, use the bids in  $B - B'$  to compute the optimal bid threshold  $f_{B-B'}$  that maximizes the revenue in  $B-B'$ , and every bidder in  $B'$  whose bid value is higher than  $f_{B-B'}$  wins
- Random sampling auctions are competitive

A. V. Goldberg, J. D. Hartline, and A. Wright. Competitive auctions and digital goods. In Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'01, pages 735–744, USA, 2001. Society for Industrial and Applied Mathematics.

# Multiple Products with Unlimited Supplies

- The bidder-optimal product assignment given the bids and the optimal sale prices can be determined by solving the integer programming problem

$$\begin{aligned} & \max && \sum_j \sum_i x_{ij} r_j \\ & \text{subject to} && \\ & && r_m = 0 \\ & && \sum_j x_{ij} \leq 1 && 1 \leq i \leq n \\ & && x_{ij} \geq 0 && 1 \leq i \leq n, 1 \leq j \leq m \\ & && p_i + r_j \geq a_{ij} && 1 \leq i \leq n, 1 \leq j \leq m \\ & && \sum_i p_i = \sum_j \sum_i x_{ij} (a_{ij} - r_j) \end{aligned}$$

where  $x_{ij}$  is the assignment of product  $j$  to bidder  $i$ ,  $r_j$  is the optimal price for product  $j$ ,  $p_i$  is the profit of bidder  $i$ , and  $a_{ij}$  is bid of bidder  $i$  on product  $j$ .

# Multiple Products with Unlimited Supplies

- Optimal pricing in random sampling auctions
  - Let  $B$  be the set of bidders
  - Obtain a sample  $B'$  of bidders
  - Compute the optimal sale prices for  $B'$
  - Run the fixed-price auction on  $B - B'$  using the sale prices computed in the previous slide
  - All bidders in  $B'$  lose the auction
- Competitive and truthful

A. V. Goldberg and J. D. Hartline. Competitive auctions for multiple digital goods. In Proceedings of the 9th Annual European Symposium on Algorithms, ESA'01, pages 416–427, Berlin, Heidelberg, 2001. Springer-Verlag.



# Asymmetric Deterministic Auctions

- No deterministic auction can be competitive
- Asymmetric auction: each buyer has the same information about the product but a different opportunity cost of obtaining the product
  - Bidders' valuations are drawn from different distributions
- An asymmetric deterministic auction can approximate the revenue of any optimal single-price sale in the worst case
  - A general derandomization technique to transform any randomized auction into an asymmetric deterministic auction with approximately the same revenue

G. Aggarwal, A. Fiat, A. V. Goldberg, J. D. Hartline, N. Immorlica, and M. Sudan. Derandomization of auctions. In Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing, STOC'05, pages 619–625, New York, NY, USA, 2005. Association for Computing Machinery.

# Envy-free Auctions

- In random sampling auctions, some bidders may lose even they make bids higher than some winning bidders do
- An auction cannot be truthful, competitive and envy-free at the same time
- Possible tradeoffs between truthfulness and envy-freeness based on the consensus revenue estimate (CORE) technique

A. V. Goldberg and J. D. Hartline. Envy-free auctions for digital goods. In Proceedings of the 4th ACM Conference on Electronic Commerce, EC'03, pages 29–35, New York, NY, USA, 2003. Association for Computing Machinery.

# Online Auctions

- A digital good may be sold repetitively
- Auctions on digital goods may run continuously
- Customers may want to have a prompt answer to their bids
- Online auctions: different customers bid at different times

R. Lavi and N. Nisan. Competitive analysis of incentive compatible on- line auctions. In Proceedings of the 2nd ACM Conference on Electronic Commerce, EC'00, pages 233–241, New York, NY, USA, 2000. Association for Computing Machinery.

# Incentive Compatible Auctions

- An (online) auction is incentive compatible if the bidders are rationally motivated to reveal their true valuations of the object
- An online auction is incentive compatible if and only if it is based on supply curves under the assumption of limited supply
- Before it receives the  $i$ -th bid  $b_i(q)$ , it fixes the supply curve  $p_i(q)$  based on the previous bids, and
  - The quantity  $q_i$  sold to customer  $i$  is the quantity  $q$  that maximizes the sum  $\sum_{j=1}^q (b_i(j) - p_i(j))$ ; and
  - The price paid by  $i$  is  $\sum_{j=1}^q p_i(j)$

# When Supply Curves Are Not Available

- Incentive-compatible randomized online auction for unlimited supply
- Each bidder  $i$  picks a random number  $t \in \{0, \dots, \lfloor \log h \rfloor\}$  and set the price threshold to  $s_i = 2^t$ , where  $h$  is the ratio of the highest valuation against the lowest valuation among all bidders
- This auction is  $O(\log h)$ -competitive

Z. Bar-Yossef, K. Hildrum, and F. Wu. Incentive-compatible online auctions for digital goods. In Proceedings of the Thirteenth Annual ACM- SIAM Symposium on Discrete Algorithms, SODA'02, pages 964–970, USA, 2002. Society for Industrial and Applied Mathematics.

# Further Improvement

- Divide a sequence of bids  $b_1, b_2, \dots$  into  $l = (\lfloor \log h \rfloor + 1)$  buckets, such that bucket  $B_j$  contains the bids with indexes in range  $[2^j, 2^{j+1})$ .
- The weight of bucket  $B_j$  is the sum of bids within  $I_j$ , that is,  $w_j = \sum_{i \in B_j} i$ .
- A new bidder can choose one of the buckets at random with the probability proportional to the bucket weight and pays the price of the lowest bid of the bucket
- The price  $s_i$  that bidder  $i$  pays follows the probability distribution

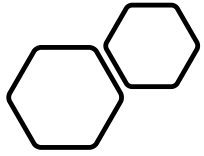
$$Pr[s_i = 2^j] = \left( \frac{w_j}{\sum_{r=0}^{l-1} w_r} \right)^d$$

- The auction is  $O(3^d (\sqrt{\log h})^{\overline{d+1}})$ -competitive. By setting  $d = \sqrt{\log \log h}$ , the auction is  $O(e^{\sqrt{\log \log h}})$ -competitive
- Z. Bar-Yossef, K. Hildrum, and F. Wu. Incentive-compatible online auctions for digital goods. In Proceedings of the Thirteenth Annual ACM- SIAM Symposium on Discrete Algorithms, SODA'02, pages 964–970, USA, 2002.



# Summary

- Revenue maximization plays a fundamental role in pricing digital products
- Three major streams of revenues for digital products, namely money information/privacy, and time/attention
- Bundling and subscription planning for digital products
- Basic types of auctions and their applications in digital products



# Pricing Data Products



# Pricing Data Products

- Overview: data markets and major players
- Arbitrage-free pricing
- Fair and truthful pricing
- Privacy preservation in data marketplaces
- Pricing dynamic data and online pricing

# Economic Analysis of Data Taxonomy as a Market Mechanism

- Data and databases are legally protected by either copyright or database right
  - Copyright protects expression and significant creative effort that creates and organizes data
  - Database right protects a whole database
- Challenge: both copyright and database right are hard to enforce due to the non-rivalrous nature of data
- Three types of data: open, public, and private data

K. Pantelis and L. Aija. Understanding the value of (big) data. In 2013 IEEE International Conference on Big Data, pages 38–42, 2013.

# Data Markets

- Two types of queries
  - Estimate the value of a “thing” or compare the values of “things”
  - Show all about a “thing”
- 7 categories of beneficiaries
  - Analysts, application vendors, data processing algorithm developers, data providers, consultants, licensing and certification entities, and data market owners
- 3 types of market structures
  - Monopoly
  - Oligopoly
  - Strong competition markets

A. Muschalle, F. Stahl, A. Löser, and G. Vossen. Pricing approaches for data markets. In M. Castellanos, U. Dayal, and E. A. Rundensteiner, editors, *Enabling Real-Time Business Intelligence*, pages 129–144, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

# Pricing Strategies and Models

- Free data
- Usage-based pricing
- Package pricing
- Flat fee tariff model
- Two-part tariff
- Freemium model

A. Muschalle, F. Stahl, A. Löser, and G. Vossen. Pricing approaches for data markets. In M. Castellanos, U. Dayal, and E. A. Rundensteiner, editors, *Enabling Real-Time Business Intelligence*, pages 129–144, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

# Machine Learning Models as Data Products

- Powerful deep models heavily rely on large amounts of training data
- Data utility for model building and the associated pricing, particularly considering privacy

# Optimal Data Market Mechanisms

- Optimal mechanisms for a monopoly data provider to sell her/his data
  - Feasible to achieve optimal revenue by a simple one-round protocol
- Optimal design for data buyers to purchase data estimators with different variance and combine the estimators to meet a required quality guarantee on variance with the lowest total cost

M. Babaioff, R. Kleinberg, and R. Paes Leme. Optimal mechanisms for selling information. In Proceedings of the 13th ACM Conference on Electronic Commerce, EC'12, pages 92–109, New York, NY, USA, 2012. Association for Computing Machinery.

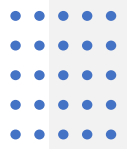
R. Cummings, K. Ligett, A. Roth, Z. S. Wu, and J. Ziani. Accuracy for sale: Aggregating data with a variance constraint. In Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, ITCS'15, pages 317–324, New York, NY, USA, 2015. Association for Computing Machinery.

# Privacy Compensation

- Truthful and approximately optimal auctions for data buyers to obtain accurate estimates on data from owners who are compensated for privacy loss
- Modeling data owners' costs of privacy loss is very difficult
- A take-it-or-leave-it mechanism

A. Ghosh and A. Roth. Selling privacy at auction. In Proceedings of the 12th ACM Conference on Electronic Commerce, EC'11, pages 199–208, New York, NY, USA, 2011. Association for Computing Machinery.

K. Ligett and A. Roth. Take it or leave it: Running a survey when privacy comes at a cost. In P. W. Goldberg and M. Guo, editors, Proceedings of the Eighth International Workshop on Internet and Network Economics (WINE'12), volume 7695 of Lecture Notes in Computer Science, pages 378–391, Berlin, Heidelberg, 2012. Springer.



# Two Dimensions for Versioning

- Data quality
- Data position

D. Bergemann, A. Bonatti, and A. Smolin. The design and price of information. *American Economic Review*, 108(1):1–48, January 2018.



# Modeling Data Quality

- A linear multi-factor model Value of data = fixed cost +  $\sum_i w_i \cdot \text{factor}_i$
- A two-level model
  - Data platform
  - Customers who want to maximize the data utility
  - The whole model is a bi-level programming problem, which is NP-hard

J. Heckman, E. Peters, N. G. Kurup, E. Boehmer, and M. Davaloo. A pricing model for data markets. In iConference 2015 Proceedings. iSchools, 2015.

H. Yu and M. Zhang. Data pricing strategy based on data quality. Computers & Industrial Engineering, 112:1 – 10, 2017.

# Versioning through Charging by Queries

- A view of a data set is a version
- The seller only needs to specify on a few views, and then the prices of other views can be decided algorithmically

P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. Query- based data pricing. In Proceedings of the 31st ACM SIGMOD-SIGACT- SIGAI Symposium on Principles of Database Systems, PODS'12, pages 167–178, New York, NY, USA, 2012. Association for Computing Machinery.

P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. Query- based data pricing. J. ACM, 62(5), Nov. 2015.

# Arbitrage-free Pricing

- Arbitrage is the activities that take advantage of price differences between two or more markets or channels
- Arbitrage is undesirable but not uncommon at all
  - Subscription based pricing possibly with a query limit allows arbitrage
  - The pricing model of charging users a certain amount of API calls for a fixed rate may potentially allow arbitrage, depending on the package size

M. Balazinska, B. Howe, and D. Suciu. Data markets in the cloud: An opportunity for the database community. *PVLDB*, 4(12):1482–1485, 2011.

A. Muschalle, F. Stahl, A. Löser, and G. Vossen. Pricing approaches for data markets. In M. Castellanos, U. Dayal, and E. A. Rundensteiner, editors, *Enabling Real-Time Business Intelligence*, pages 129–144, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

# Arbitrage-free Pricing of Linear Aggregate Queries

Li *et al.* [124] consider linear queries. Given a data set of  $n$  tuples  $x_1, \dots, x_n$ , a linear query  $\mathbf{q} = (q_1, \dots, q_N)$  is a real-valued vector, and the answer  $\mathbf{q}(\mathbf{x}) = \sum_{i=1}^n q_i x_i$ . For a multiset of queries  $\mathbf{S} = \{\mathbf{Q}_1, \dots, \mathbf{Q}_k\}$  and query  $\mathbf{Q}$ , if the answer to  $\mathbf{Q}$  can be linearly derived from the answers to the queries in  $\mathbf{S}$ , then  $\mathbf{Q}$  is said to be determined by  $\mathbf{S}$ , denoted by  $\mathbf{S} \rightarrow \mathbf{Q}$ . A pricing function  $\pi(\mathbf{Q})$  is arbitrage-free if for any multiset  $\mathbf{S}$  and query  $\mathbf{Q}$  such that  $\mathbf{S} \rightarrow \mathbf{Q}$ ,  $\pi(\mathbf{Q}) \leq \sum_{i=1}^k \pi(\mathbf{Q}_i)$ .

C. Li, D. Y. Li, G. Miklau, and D. Suciu. A theory of pricing private data. *ACM Trans. Database Syst.*, 39(4), Dec. 2015.

## Three Challenges

- Arbitrage is still possible to derive answers to a bundle of queries from another bundle of queries and their answers
- Arbitrage is still possible on biased estimators for statistical queries
- Can we can obtain arbitrage-free pricing maximizing profit given the distribution of buyer demands?

A. Roth. Technical perspective: Pricing information (and its implications). Commun. ACM, 60(12):78, Nov. 2017.

# Arbitrage-free Pricing for General Queries

- Three types of pricing models for query bundles
  - Instance-independent pricing function depends on the query bundle but not the database instance
  - Up-front dependent pricing function depends on both the query bundle and the database instance
  - Delayed pricing function depends on both the query bundle and the answer computed by the query bundle on the current database instance

B.-R. Lin and D. Kifer. On arbitrage-free pricing for general data queries. Proc. VLDB Endow., 7(9):757–768, May 2014.

# Five Types of Arbitrage

- Price-based arbitrage: if prices are quoted by queries, in order to avoid, answers to queries should not be deduced from prices along
- Separate account arbitrage: a buyer may use multiple accounts to derive answers to a query bundle
- Post-processing arbitrage: if the answers to a query bundle  $q'$  can always be deduced from answers to another query bundle  $q$ , the price of  $q$  should be no cheaper than that of  $q'$
- Serendipitous arbitrage: for a specific database instance, the answers to  $q$  may be derived from the answers to  $q'$
- Almost-certain arbitrage: two queries behave almost identical but their prices are dramatically different

B.-R. Lin and D. Kifer. On arbitrage-free pricing for general data queries. Proc. VLDB Endow., 7(9):757–768, May 2014.

# Qirana: Efficient and Scalable Pricing

- Regard a query as an uncertainty reduction mechanism
- A buyer faces a set of possible databases  $I$  defined by a database schema, primary keys and predefined constraints
- Once a buyer obtains the answer  $E$  to a query  $Q$ , all possible databases  $D$  such that  $E \neq Q(D)$  are eliminated

S. Deep and P. Koutris. Qirana: A framework for scalable query pricing. In Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD'17, pages 699–713, New York, NY, USA, 2017. Association for Computing Machinery.

S. Deep, P. Koutris, and Y. Bidasaria. Qirana demonstration: Real time scalable query pricing. Proc. VLDB Endow., 10(12):1949–1952, Aug. 2017.



# Arbitrage-free Pricing of User-based Markets

- Denote by  $q_i$  a selection query over user attributes, by  $U_i$  the set of all users satisfying  $q_i$ , and by  $p_i$  the price of each user in  $U_i$
- If a buyer purchases  $n$  users ( $1 \leq n \leq |U_i|$ ) in  $U_i$ , he has to pay  $n \cdot p_i$
- Uniform pricing is arbitrage-free, but is a logarithmic approximation to the maximum revenue arbitrage-free pricing solution
- A greedy non-uniform pricing design
  - Start with the optimal uniform pricing
  - Iteratively updates the pricing function

C. Xia and S. Muthukrishnan. Arbitrage-free pricing in user-based markets. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS'18, pages 327–335, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

# Arbitrage-free Pricing for Multiple Versions of Machine Learning Models

- The broker trains the optimal model on the complete raw data
- Random Gaussian noises are added to the optimal model to produce different versions for different buyers
  - Rationale: the error of a machine learning model instance is monotonic with respect to the variance of the noise injected into the model
- A pricing function is arbitrage-free if and only if the price of a randomized model instance is monotonically increasing and subadditive with respect to the inverse of the variance

L. Chen, P. Koutris, and A. Kumar. Towards model-based pricing for machine learning in a data marketplace. In Proceedings of the 2019 International Conference on Management of Data, SIGMOD'19, pages 1535– 1552, New York, NY, USA, 2019. Association for Computing Machinery.

# Revenue Maximization Pricing

- A buyer is single-minded if the buyer wants to purchase the answer to a single set of queries
- Three types of pricing models
  - Uniform bundle pricing: set the price of every bundle identical
  - Additive or item pricing: price each item and charges a bundle the sum of prices for the items in the bundle
  - Fractionally subadditive pricing or XOS: set  $k$  weights  $w_j^1, \dots, w_j^k$  for each item  $j$ , and for a bundle  $e$ , the price is set to  $\max_{i=1}^k \sum_{j \in e} w_j^i$

S. Chawla, S. Deep, P. Koutrisw, and Y. Teng. Revenue maximization for query pricing. Proc. VLDB Endow., 13(1):1–14, Sept. 2019.

# Some Major Results

- There exists uniform bundle pricing that is  $O(\log m)$  approximation of revenue maximization
- Item pricing can achieve an  $O(\log B)$  approximation of maximum revenue, where  $B$  is the maximum number of bundles an item can involve
- Uniform bundle pricing, item pricing and XOS pricing combining a constant number of item pricing functions are still  $\Omega(\log m)$  away from maximum revenue

C. Swamy and M. Cheung. Approximation algorithms for single-minded envy-free profit-maximization problems with limited supply. In 2008 IEEE 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pages 35–44, Los Alamitos, CA, USA, oct 2008. IEEE Computer Society.

S. Chawla, S. Deep, P. Koutrisw, and Y. Teng. Revenue maximization for query pricing. Proc. VLDB Endow., 13(1):1–14, Sept. 2019.

# Revenue Maximization for Machine Learning Models

- The optimization problem is coNP-hard
- Relax the subadditive constraint  $p(x + y) \leq p(x) + p(y)$  by  $\frac{q(x)}{x} \geq \frac{q(y)}{y}$  for every  $0 < x \leq y$
- For every well standing pricing function  $p()$ , there exists a pricing function  $q(x)$  with the relaxed subadditive constraint such that  $\frac{p(x)}{2} \leq q(x) \leq p(x)$ , and  $q(x)$  can be computed using dynamic programming in  $O(n^2)$  time, where  $n$  is the number of interpolated price points

L. Chen, P. Koutris, and A. Kumar. Towards model-based pricing for machine learning in a data marketplace. In Proceedings of the 2019 International Conference on Management of Data, SIGMOD'19, pages 1535– 1552, New York, NY, USA, 2019. Association for Computing Machinery.

# Fair and Truthful Pricing

- Each seller  $j$  supplies a data stream  $X_j$
- Each buyer  $n$  conducts a prediction task  $Y_n$ ,  $X_j, Y_n \in R^T$
- Taking a prediction task  $Y_n$  and an estimate  $\hat{Y}_n$ , a prediction gain function  $G_n: R^{2T} \rightarrow [0, 1]$  measures the quality of the prediction
- The value buyer  $n$  gets from estimate  $\hat{Y}_n$  is  $\mu_n \cdot G(Y_n, \hat{Y}_n)$
- A machine learning model  $M: R^{MT} \rightarrow R^T$  uses data from  $M$  sellers to produce an estimate for buyer  $n$ 's prediction task  $Y_n$

# Fair and Truthful Pricing

- Let  $p_n$  and  $b_n$  be the price and the bid, respectively
- Allocation function  $AF: (p_n, b_n; X_M) \rightarrow \widetilde{X}_M$  measures the quality at which buyer  $n$  obtains that  $M$  is allocated to the sellers
- Revenue function  $RF: (p_n, b_n, Y_n; M, G, X_M) \rightarrow r_n$
- The utility buyer  $n$  receives by bidding  $n_n$  for  $Y_n$  is  $U(b_n, Y_n) = \mu_n \cdot \mathcal{G}(Y_n, \hat{Y}_n) - \mathcal{RF}(p_n, b_n, Y_n)$ :
- A market is truthful if for all prediction tasks  $Y_n$ ,  $\mu_n = \operatorname{argmax}_{z \in R^+} U(z, Y_n)$
- The data market is truthful if and only if function  $AF^*$  is monotonic: an increase in the difference between price rate  $p_n$  and bid  $b_n$  leads to a decrease in predication gain

A. Agarwal, M. Dahleh, and T. Sarkar. A marketplace for data: An algorithmic solution. In Proceedings of the 2019 ACM Conference on Economics and Computation, EC'19, pages 701–726, New York, NY, USA, 2019. Association for Computing Machinery.

# Efficient Shapley Value Approximation

- Computing Shapley value is exponential
- A permutation sampling method that approximates Shapley value for any bounded utility functions

- Use 
$$\psi(s) = \frac{1}{N!} \sum_{\pi \in \Pi(D)} (\mathcal{U}(P_s^\pi \cup \{s\}) - \mathcal{U}(P_s^\pi))$$

- Estimate  $\psi(s) = E[\mathcal{U}(P_s^\pi \cup \{s\}) - \mathcal{U}(P_s^\pi)]$  by sample mean
- To achieve  $(\delta, \epsilon)$ -approximation

$P(|\hat{s} - s|_p \leq \epsilon) \geq 1 - \delta$ , where  $\hat{s}$  is the estimate, we need  $\frac{2r^2N}{\epsilon^2} \log \frac{2N}{\delta}$  samples and evaluate the utility function  $O(N^2 \log N)$  times, where  $r$  is the range of the utility function  $\mathcal{U}$ .

X. Deng and C. H. Papadimitriou. On the complexity of cooperative solution concepts. *Mathematics of Operations Research*, 19(2):257–266, 1994.



# Reducing Utility Function Evaluation

- Apply the idea of feature selection using group testing
- For user  $s$ , let  $\beta_s$  be the random variable that  $s$  appears in a random sample of sellers
- For sellers  $s_i$  and  $s_j$ , the difference in Shapley values between  $s_i$  and  $s_j$  is

$$\begin{aligned} \psi(s_i) - \psi(s_j) &= \frac{1}{N-1} \sum_{S \in D \setminus \{s_i, s_j\}} \frac{\mathcal{U}(S \cup \{s_i\}) - \mathcal{U}(S \cup \{s_j\})}{\binom{N-2}{|S|}} \\ &= E[(\beta_{s_i} - \beta_{s_j}) \mathcal{U}(\beta_{s_1}, \dots, \beta_{s_j})] \end{aligned}$$

- $(\epsilon, \delta)$ -approximation  
 $O(\sqrt{N} (\log N)^2)$  times

R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gu'rel, B. Li, C. Zhang, D. Song, and C. J. Spanos. Towards efficient data valuation based on the shapley value. In K. Chaudhuri and M. Sugiyama, editors, volume 89 of Proceedings of Machine Learning Research, pages 1167–1176. PMLR, 16–18 Apr 2019.

# Approximation for Supervised Learning Models

- Monte-Carlo method
  - Generate Monte-Carlo estimates until the average empirically converges
  - In practice, it is sufficient to estimate the Shapley value up to the intrinsic noise in the predictive performance on the test data set, thus, truncation can be used in practice based on the bootstrap variation on the test set
- Gradient Shapley method
  - Train a model using one “epoch” of the training data
  - Update the model by gradient descent on one data point at a time

A. Ghorbani and J. Zou. Data shapley: Equitable valuation of data for machine learning. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 2242–2251, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

# Shapley Values in kNN Classifiers

- For some specific model, the computation may be reduced dramatically
- For unweighted kNN classifiers, the exact computation needs only  $O(N \log N)$  time, and  $(\epsilon, \delta)$ -approximation can be achieved in  $O(N^{h(\epsilon, k)} \log N)$  time when  $\epsilon$  is not too small and  $k$  is not too large

- Key enabler 
$$\nu(S) = \frac{1}{k} \sum_{i=1}^{\min\{k, |S|\}} \mathbb{1}[y_{\alpha_i(S)} = y_{\text{test}}]$$

- Using locality sensitive hashing

R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gurel, B. Li, C. Zhang, C. Spanos, and D. Song. Efficient task-specific data valuation for nearest neighbor algorithms. Proc. VLDB Endow., 12(11):1610–1623, July 2019.

# Privacy Preserving Marketplaces of Data

- When a user shares her/his data with some others, more often than not, the user may disclose her/his privacy to some extent
- Privacy protection
- Privacy compensation
- Many designs use differential privacy

C. Dwork. Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, and A. Li, editors, *Theory and Applications of Models of Computation*, pages 1–19, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

# A Truthful and Privacy Preserving Marketplace

- Assumption: there is only one query and the individual evaluation of their data are private
- Data owners are asked to report the costs for the use of their data
- Transform the problem into variants of multi-unit procurement auction
- When a buyer holds an accuracy goal, the classic Vickrey auction can minimize the buyer's total cost and guarantee the accuracy
- When the buyer has a budget, they give an approximation algorithm to maximize the accuracy under the budget constraint
- May not work well when the costs and the data are correlated

A. Ghosh and A. Roth. Selling privacy at auction. In Proceedings of the 12th ACM Conference on Electronic Commerce, EC'11, pages 199–208, New York, NY, USA, 2011. Association for Computing Machinery.

# A Posted-price-like Mechanism

- A set of data sellers categorized into different types and the associated distributions of costs
- Offer each user a contract with the expected payment corresponding to the type
  - If a seller takes the offer, the payment is determined by the seller's verifiable type and the associated payment in the contract
- The sellers are truthful
- The mechanism is Bayesian incentive compatible, ex-interim individually rational,  $O(\epsilon^{-1})$ -accurate, perfectly data private, and  $\epsilon$ -differentially private

L. K. Fleischer and Y.-H. Lyu. Approximately optimal auctions for selling privacy when costs are correlated with data. In Proceedings of the 13th ACM Conference on Electronic Commerce, EC'12, pages 568–585, New York, NY, USA, 2012. Association for Computing Machinery.

# More Methods

- Assume that individual valuations are public, return unbiased estimations and pricing multiple queries consistently
- Multiple sellers' data are correlated
- Time series data with temporal correlation, use Pufferfish privacy
- In general, it is impossible for any mechanism to compensate individuals for privacy loss properly if correlations between their private data and their cost functions are unknown beforehand

C. Li, D. Y. Li, G. Miklau, and D. Suci. A theory of pricing private data. *ACM Trans. Database Syst.*, 39(4), Dec. 2015.

C. Niu, Z. Zheng, F. Wu, S. Tang, X. Gao, and G. Chen. Unlocking the value of privacy: Trading aggregate statistics over private correlated data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD'18*, pages 2031–2040, New York, NY, USA, 2018. Association for Computing Machinery.

C. Niu, Z. Zheng, S. Tang, X. Gao, and F. Wu. Making big money from small sensors: Trading time-series data under pufferfish privacy. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pages 568–576, April 2019.

# Contract Design

- For a query bundle  $\{Q_1, \dots, Q_k\}$ , a contract is a tuple  $(p, \epsilon, s)$ , where  $p > 0$  is the price paid by a buyer,  $\epsilon$  is the privacy budget, such that a buyer can get answer to query  $Q_i$  with  $\epsilon_i$ -differential privacy guarantee, and  $\epsilon \geq \sum_{i=1}^k \epsilon_i$ , and  $p$  is the post-hoc fine to be paid if a buyer is found misusing the query answers
- If there are  $n$  types of honest buyers and one type of adversarial buyers, up to  $n$  contracts are sufficient -- the seller should adjust the contracts' pricing to account for the risks from adversarial users

P. Naghizadeh and A. Sinha. Adversarial contract design for private data commercialization. In Proceedings of the 2019 ACM Conference on Economics and Computation, EC'19, pages 681–699, New York, NY, USA, 2019. Association for Computing Machinery.



# Tradeoff between Privacy and Data Utility

- Let two-part tariff pricing function  $R(s, x) = \alpha_s + \beta_s x$  be the price for  $x$  amount of data with sensitivity level  $s$
- Two types of data users
  - One type for aggregate information and patterns in data
  - The other type for individual identity and personal information
- Core idea: the data owner can identify the sensitive attributes in the data and offer a low price for data without sensitive attributes, and charge for a high price for data with sensitive attributes.

X.-B. Li and S. Raghunathan. Pricing and disseminating customer data with privacy awareness. *Decision Support Systems*, 59:63 – 73, 2014.

# Collect Data or Not?

- Due to the privacy concerns, when a company may have opportunities to collect data about its customers, should it do it (i.e., collecting and revealing the data) or not (i.e., a blanket policy of never collecting)?
- Should not collect customer data if the total gains from trading the data cannot cover the privacy loss
  - In practice, there is an increasing tendency for consumers to overestimate their loss of privacy, particularly when the use of the private data is uncertain
- Should offer two contracts on their services and products
  - One of the contracts collects the customer data at a certain price
  - The other contract does not collect any customer data at a different price

J. Jaisingh, J. Barron, S. Mehta, and A. Chaturvedi. Privacy and pricing personal information. *European Journal of Operational Research*, 187(3):857 – 870, 2008.

# Preserving Privacy of Buyers

- Transactions may also disclose privacy of data buyers, such as what, when, and how much they buy
- After making an initial deposit and maintaining a sufficient balance, a buyer can engage in an unlimited number of priced oblivious-transfer protocols where the sellers and broker cannot know anything other than the amount of interaction and the initial deposit amount
  - The broker even cannot know the buyer's current balance and when the buyer's balance runs out
- Major technique: adapting conditional disclosure to the two-party setting

W. Aiello, Y. Ishai, and O. Reingold. Priced oblivious transfer: How to sell digital goods. In *Advances in Cryptology - EUROCRYPT 2001, International Conference on the Theory and Application of Cryptographic Techniques, Innsbruck, Austria, May 6-10, 2001, Proceedings*, volume 2045 of *Lecture Notes in Computer Science*, pages 119–135. Springer, 2001.

# Sterling: a Decentralized Market-place for Private Data

- Support privacy-preserving distribution and use of data
- Major techniques
  - Privacy-preserving smart contracts
  - Trusted execution environments
  - Differential privacy

N. Hynes, D. Dao, D. Yan, R. Cheng, and D. Song. A demonstration of sterling: A privacy-preserving data marketplace. Proc. VLDB Endow., 11(12):2086–2089, Aug. 2018.

# Pricing Dynamic Data and Online Pricing

- How to price views on data streams properly?
- Estimate and optimize the operational costs
- Example:
  - Two data buyers  $b_1$  and  $b_2$  purchase two queries  $q_1$  and  $q_2$ ,  $q_1$  is about all customers in North America, while  $q_2$  keeps all the same as  $q_1$  but focuses on only customers in Canada
  - The optimal pricing of  $q_1$  and  $q_2$  should take the advantage of the overlap between the two queries so that the sharing can save the operational cost, and, at the same time, be fair to  $b_1$  and  $b_2$
- A greedy method that enumerates all possible sharing plans and selects the one with the minimum additional cost

S. Al-Kiswany, H. Hacigümüş, Z. Liu, and J. Sankaranarayanan. Cost exploration of data sharings in the cloud. In Proceedings of the 16th International Conference on Extending Database Technology, EDBT'13, pages 601–612, New York, NY, USA, 2013. Association for Computing Machinery.

# Taking Risk in Cost

- If the costs of the previous sharings are already cumulated to a high level, and the additional cost of a new sharing (i.e., the risk) is moderate and can be amortized well by the previous sharing, then the new sharing may be taken
- Five rules to ensure fair pricing, let  $AC(S)$  be the cost attributed to a sharing  $S$ 
  - For two identical sharings  $S1 = S2$ ,  $AC(S1) = AC(S2)$  should hold
  - $AC(S)$  should be no higher than the lowest cost of  $S$  if no other sharing exists
  - If the query of  $S1$  is contained by the query of  $S2$ , and the lowest cost of  $S1$  is smaller than the lowest cost of  $S2$  if no other sharing exists, then  $AC(S1) \leq AC(S2)$
  - A sharing plan with common subexpressions with other sharings should be compensated
  - The cost of the global plan should be equal to the sum of costs attributed to all sharings

Z. Liu and H. Hacigümüş. Online optimization and fair costing for dynamic data sharing in a cloud data market. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD'14, pages 1359–1370, New York, NY, USA, 2014. Association for Computing Machinery.

# History-aware Pricing

- Buyers are charged only once for data purchased and not updated
- Refund – a user can ask for refunds of data they she/he has bought before
- For each query, the seller issues a coupon in addition to the query result, where the coupon records the identity information of the data in the query result
  - Coupon  $c = ((id, uid, v), \tau, H(id \oplus \tau \oplus \kappa))$ , where  $id$  is the tuple identifier,  $uid$  is the user-id,  $v$  is the version-id that is monotonically increasing,  $\tau$  is a query identifier that is monotonically increasing,  $H$  is a cryptographic hash function, and  $\kappa$  is a secret key only known to the seller
  - If a buyer gets two coupons  $c_1$  and  $c_2$  in two different purchases such that  $c_1[(tid, uid, v)] = c_2[(tid, uid, v)]$ , then the buyer can ask the seller for a refund by showing the two coupons
  - No arbitrage-free guarantee

P. Upadhyaya, M. Balazinska, and D. Suciu. Price-optimal querying with data apis. Proc. VLDB Endow., 9(14):1695–1706, Oct. 2016.

# Arbitrage-free History-aware Pricing

- A buyer already purchases queries  $Q = Q_1, \dots, Q_k$  and pays for a total of  $p(Q, D)$  so far
- When a new query  $Q_{k+1}$  comes, let the support set  $S_{k+1} = \{D_i \in S \mid Q(D_i) = Q(D), Q_{k+1}(D_i) \neq Q_{k+1}(D)\}$
- The new total price  $p((Q_1, \dots, Q_k, Q_{k+1}), D) = p(Q, D) + \sum_{D_i \in S_{k+1}} w_i$
- Arbitrage-free

S. Deep and P. Koutris. Qirana: A framework for scalable query pricing. In Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD'17, pages 699–713, New York, NY, USA, 2017. Association for Computing Machinery.



# Online Pricing for Mobile Crowd-sensing Data Markets

- Data providers distributed in space
- Three types of spatial queries from buyers, single-data query, multi-data query and range query
- The vendor uses the raw data from data providers and produces a statistical model through Gaussian process to answer queries
- A randomized online pricing strategy so that the price can be adaptive from the historical queries
- Arbitrage-free
- A constant factor approximation of revenue maximization

Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen. An online pricing mechanism for mobile crowdsensing data markets. In Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing, Mobihoc'17, New York, NY, USA, 2017. Association for Computing Machinery.

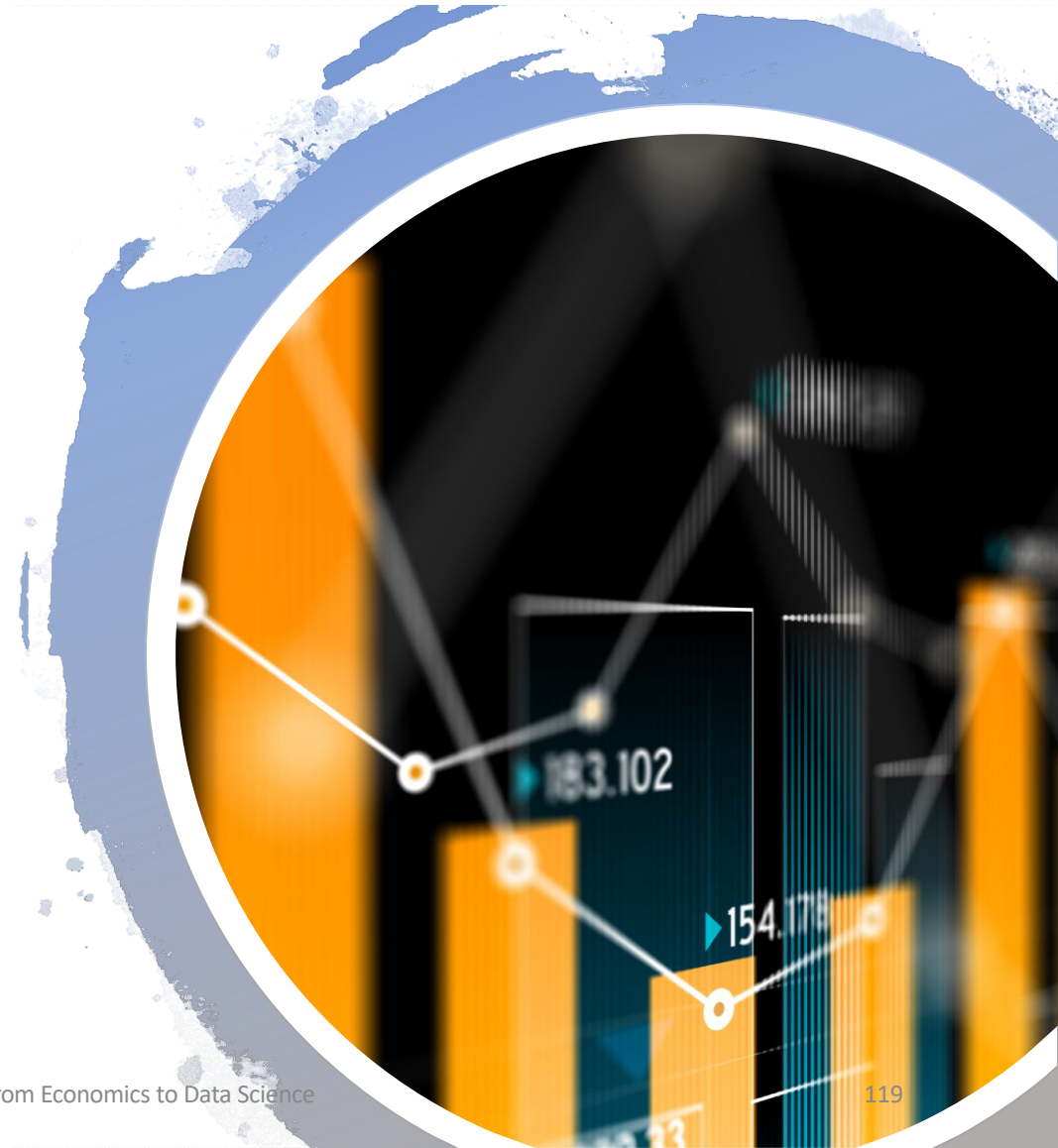
# Adjustable Prices

- A query may be sold to different buyers at different time
- The broker can adjust prices over time
- The objective is to maximize the broker's cumulative revenue by posting reasonable prices for sequential queries
- A contextual dynamic pricing mechanism with the reserve price constraint
- Central idea: use the properties of ellipsoid for efficient online optimization
  - Can support both linear and non-linear market value models with uncertainty

C. Niu, Z. Zheng, F. Wu, S. Tang, and G. Chen. Online pricing with reserve price constraint for personal data markets. ArXiv, abs/1911.12598, 2019.

# Summary

- Structure, players, and ways to produce data products in data marketplaces
- Several important areas in pricing data products, including arbitrage-free pricing, revenue maximization pricing, and fair and truthful pricing
- How to price dynamic data and online pricing
- When pricing data products in a data marketplace, those several considerations are typically incorporated and integrated in one way or another



# Challenges and Future Directions

# Six Challenges

- Create information extraction systems that can attach structured, semantically meaningful labels to text data with little human effort
  - Given that labeled text data, enable an analyst OLAP operations on Web data with the same simplicity as a Web search
- Enable analysts to create domain specific data processing flows with little costs. Enable and optimize black-box user defined functions in these flows
- Build systems that can reliably answer brand monitoring queries to indexes, with heavy read and write access, sticking to ACID constraints on a scalable infrastructure in order to enable near real-time brand monitoring

# Six Challenges

- Given a list of entities and their properties, identify a set of mining algorithms that have been optimized to a very high degree for exactly this data set
  - Provide a user with an execution sample and recommend an algorithm with regard to execution time, precision and recall
- Develop incentives for suppliers for price transparency
- Collect transactional data from customers and leverage data for solving the above challenges

A. Muschalle, F. Stahl, A. Löser, and G. Vossen. Pricing approaches for data markets. In M. Castellanos, U. Dayal, and E. A. Rundensteiner, editors, *Enabling Real-Time Business Intelligence*, pages 129–144, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg



## More Challenges

- Data supply chain
- End-to-end solutions
- Valuation of data and data usage from different angles, data providers, users, brokers
- Accounting and auditing
- Fairness, truthfulness, and privacy preservation in specific applications