Data Pricing and Data License Agreements in the Cloud

Magdalena Balazinska

Department of Computer Science & Engineering

eScience Institute

University of Washington

http://www.cs.washington.edu/people/faculty/magda

UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Acknowledgments

Research team on project

- Prasang Upadhyaya (UW, lead on licensing)
- Paraschos Koutris (UW, lead on pricing)
- Prof. Dan Suciu (UW)
- Dr. Hakan Hacigumus (NEC Labs)

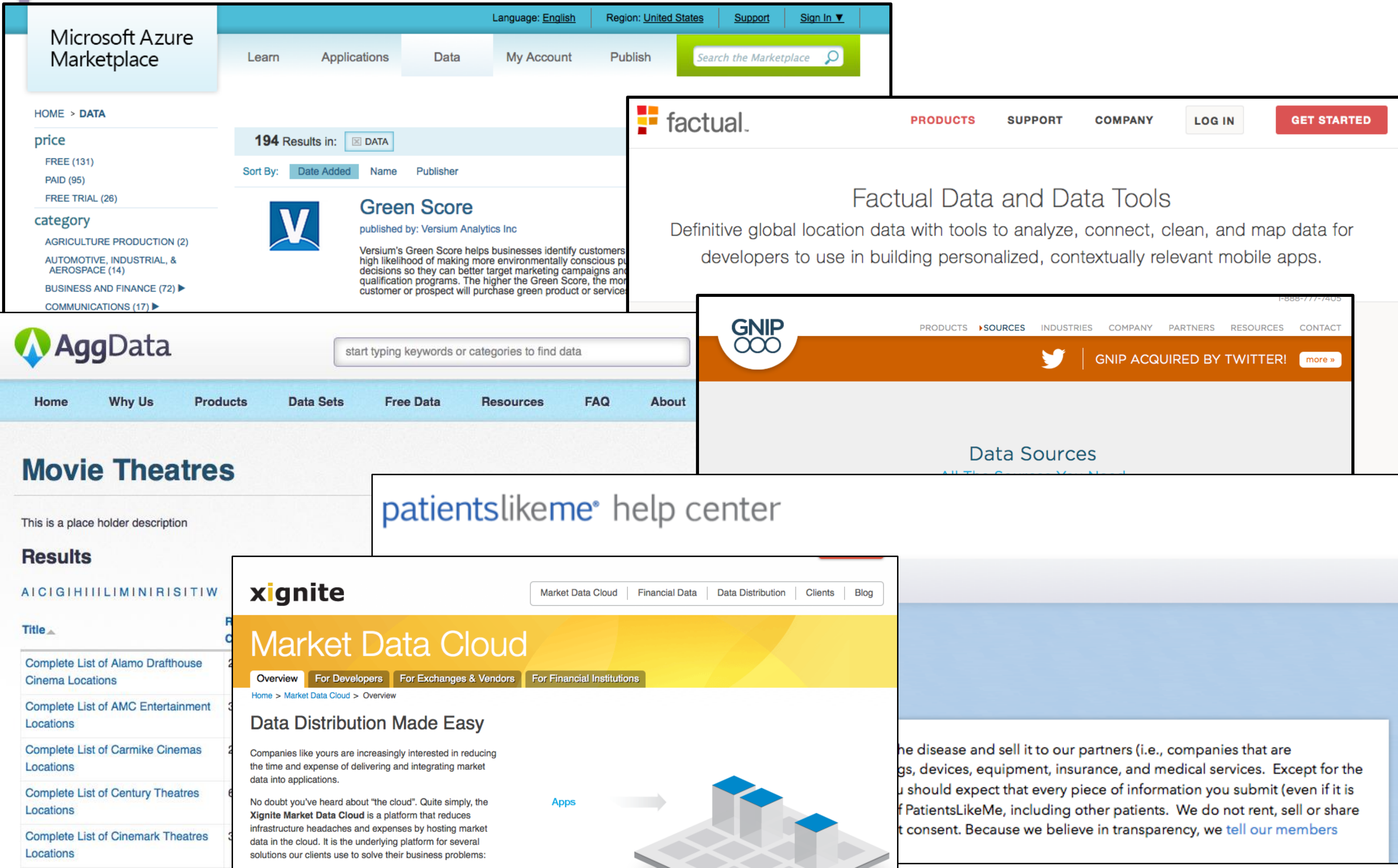
Sponsors

- National Science Foundation
- NEC
- Microsoft Research

Data Has Value

- First wave of computing: value in hardware
  - IBM, Intel, DEC
- Second wave of computing: value in software
  - Microsoft, Oracle, Google
- Third wave of computing: value in data
  - Dun & Bradstreet, Factual, Facebook, Google

Data itself is now a product that is being created, improved, bought and sold on the Web



Green Score

published by: Versium Analytics Inc

Versium's Green Score helps businesses identify customers
high likelihood of making more environmentally conscious pu
decisions so they can better target marketing campaigns and
qualification programs. The higher the Green Score, the mor
customer or prospect will purchase green product or service

Factual Data and Data Tools

Definitive global location data with tools to analyze, connect, clean, and map data for developers to use in building personalized, contextually relevant mobile apps.

Data Sources

Movie Theatres

This is a place holder description

Results

A | C | G | H | I | L | M | N | R | S | T | W

Title

Complete List of Alamo Drafthouse Cinema Locations

Complete List of AMC Entertainment Locations

Complete List of Carmike Cinemas Locations

Complete List of Century Theatres Locations

Complete List of Cinemark Theatres Locations

he disease and sell it to our partners (i.e., companies that are
gs, devices, equipment, insurance, and medical services. Except for the
u should expect that every piece of information you submit (even if it is
f PatientsLikeMe, including other patients. We do not rent, sell or share
t consent. Because we believe in transparency, we tell our members

Market Data Cloud

Data Distribution Made Easy

Companies like yours are increasingly interested in reducing the time and expense of delivering and integrating market data into applications.

No doubt you've heard about "the cloud". Quite simply, the **Xignite Market Data Cloud** is a platform that reduces infrastructure headaches and expenses by hosting market data in the cloud. It is the underlying platform for several solutions our clients use to solve their business problems:

What are the Technical Challenges

- Challenge 1: Data License Agreements
  - All data comes with terms of use
  - Can we automate their enforcement?

- Challenge 2: Data Pricing
  - Existing pricing methods are limited
  - Can we support flexible pricing?

Data Comes with Terms of Use

Medical Data

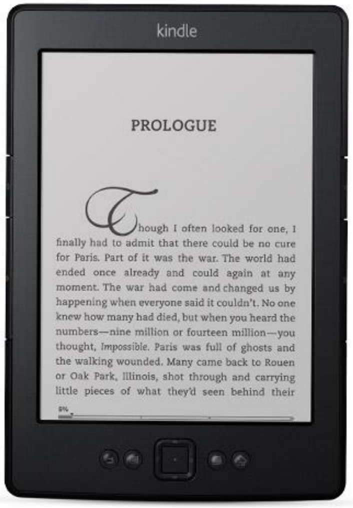
Name	Ailment	Birth date	Sex	Location
John Doe	Asthma	Jan 7th 1972	M	Seattle
Mary Jane	Dislocated shoulder	Mar 21st 1965	F	San Diego
Alice Summer	Flu	May 28th 1986	M	San Francisco
…	…	…	…	…
Bob B	Flu	Oct 14th 2000	M	Miami

Queries that try to identify an individual referenced in the database are prohibited (MIMIC II)

Maps

Overlaying map data with any other data is prohibited (Navteq)

Digital Books



PROLOGUE

Though I often looked for one, I finally had to admit that there could be no cure for Paris. Part of it was the war. The world had ended once already and could again at any moment. The war had come and changed us by happening when everyone said it couldn't. No one knew how many had died, but when you heard the numbers—nine million or fourteen million—you thought, *Impossible*. Paris was full of ghosts and the walking wounded. Many came back to Rouen or Oak Park, Illinois, shot through and carrying little pieces of what they'd seen behind their

Each book may be lent once for 2 weeks while being inaccessible by the lender (Kindle)

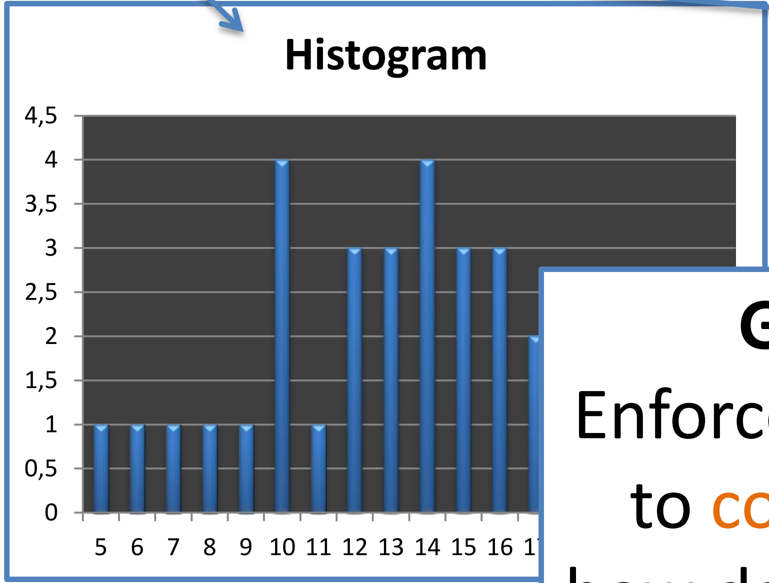
More Examples

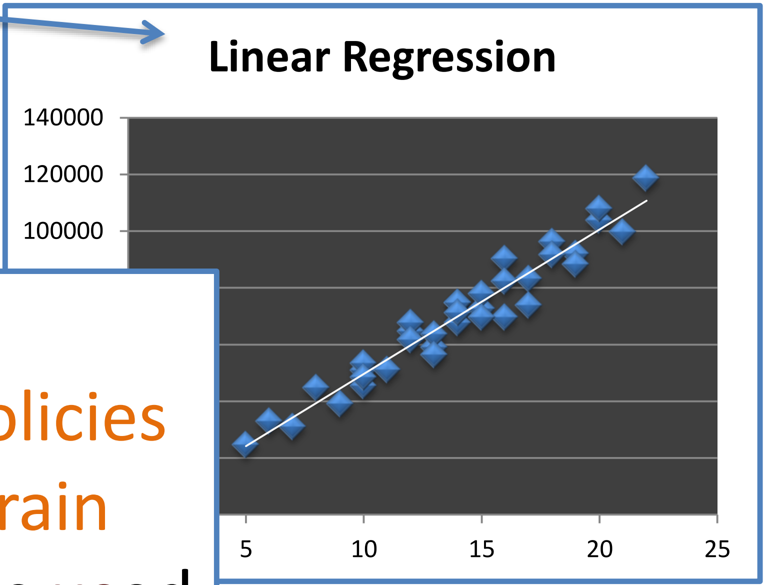
Terms of use	Source
Overlaying Navteq data with any other data is prohibited	Navteq
Each book may be lent once for a duration of 14 days and will not be readable by the lender during the loan period	Amazon Kindle
In a month, all queries may, in total, return up to 2M characters of data at the free tier	Microsoft Translator
OAuth calls are permitted 350 requests per hour	Twitter and Foursquare
Queries that try to identify an individual referenced in the database are prohibited	MIMIC II
You are required to display all attribution information and any proprietary notices associated with the Foursquare Data	Foursquare, Yelp, World Bank
Don’t aggregate or blend our star ratings and review counts with other providers. You may show content from multiple providers, but Yelp data should stand on its own …	Yelp

Terms of Use Control Use, Not Access

Example for medical research data

Permitted


Histogram


Linear Regression

Goal

Enforce policies to constrain how data is used

Augmenting Data Sources

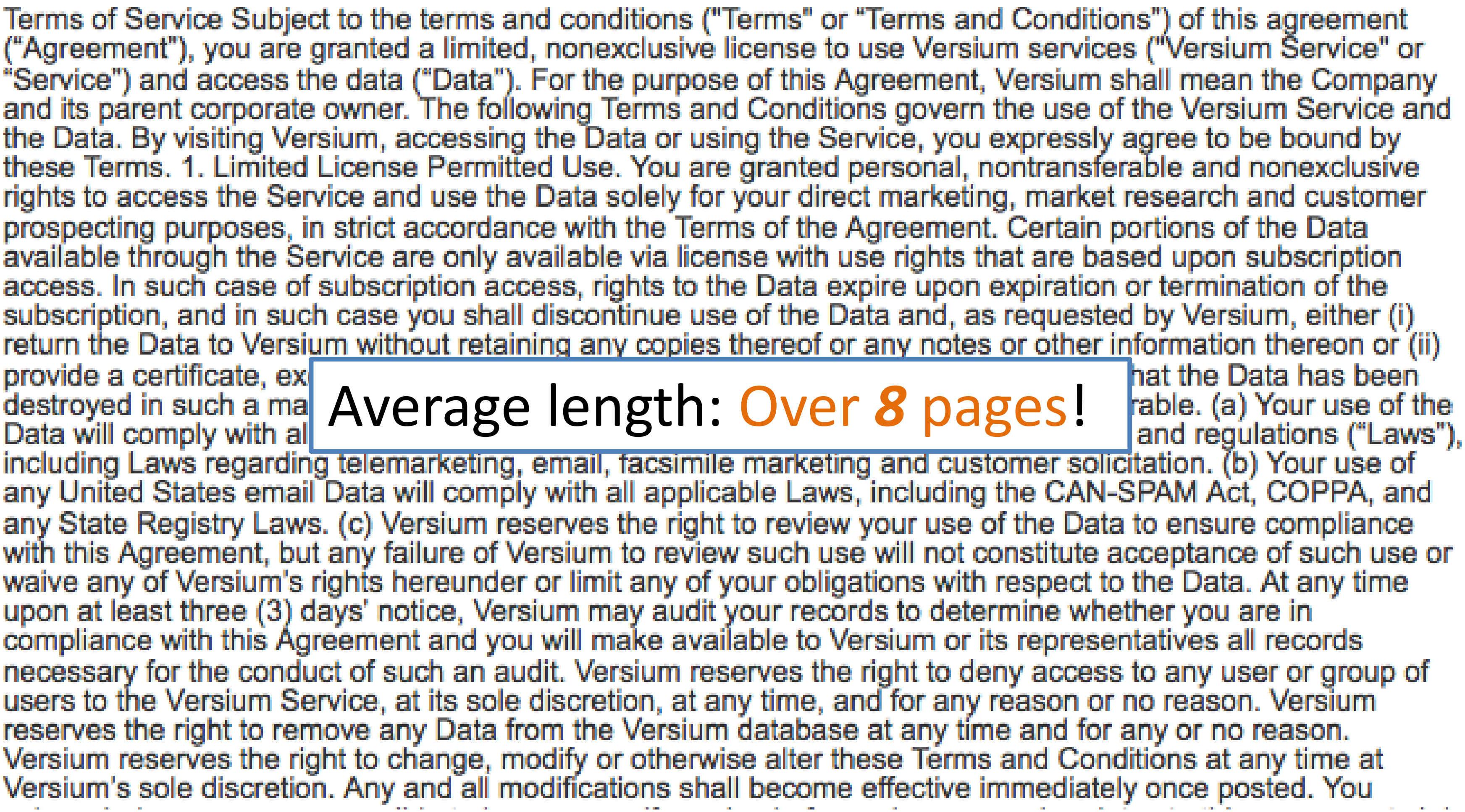
```
Join medical data with
voters registry
```

Fine-grained Access

```
Look up specific patient
```

Denied

Today: Written Agreements

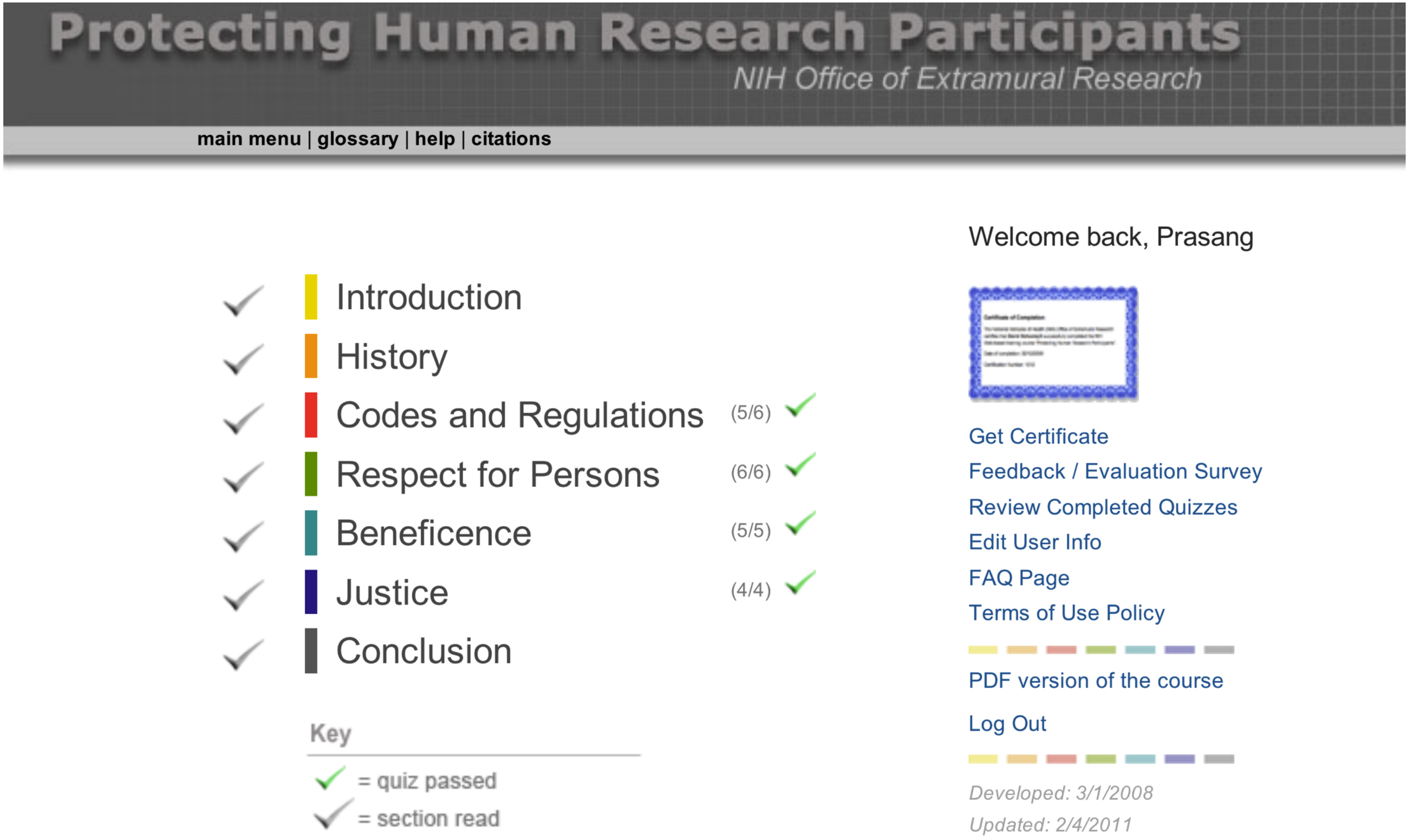


Terms of Service Subject to the terms and conditions ("Terms" or "Terms and Conditions") of this agreement ("Agreement"), you are granted a limited, nonexclusive license to use Versium services ("Versium Service" or "Service") and access the data ("Data"). For the purpose of this Agreement, Versium shall mean the Company and its parent corporate owner. The following Terms and Conditions govern the use of the Versium Service and the Data. By visiting Versium, accessing the Data or using the Service, you expressly agree to be bound by these Terms. 1. Limited License Permitted Use. You are granted personal, nontransferable and nonexclusive rights to access the Service and use the Data solely for your direct marketing, market research and customer prospecting purposes, in strict accordance with the Terms of the Agreement. Certain portions of the Data available through the Service are only available via license with use rights that are based upon subscription access. In such case of subscription access, rights to the Data expire upon expiration or termination of the subscription, and in such case you shall discontinue use of the Data and, as requested by Versium, either (i) return the Data to Versium without retaining any copies thereof or any notes or other information thereon or (ii) provide a certificate, ex hat the Data has been destroyed in such a ma rable. (a) Your use of the Data will comply with al and regulations ("Laws"), including Laws regarding telemarketing, email, facsimile marketing and customer solicitation. (b) Your use of any United States email Data will comply with all applicable Laws, including the CAN-SPAM Act, COPPA, and any State Registry Laws. (c) Versium reserves the right to review your use of the Data to ensure compliance with this Agreement, but any failure of Versium to review such use will not constitute acceptance of such use or waive any of Versium's rights hereunder or limit any of your obligations with respect to the Data. At any time upon at least three (3) days' notice, Versium may audit your records to determine whether you are in compliance with this Agreement and you will make available to Versium or its representatives all records necessary for the conduct of such an audit. Versium reserves the right to deny access to any user or group of users to the Versium Service, at its sole discretion, at any time, and for any reason or no reason. Versium reserves the right to remove any Data from the Versium database at any time and for any or no reason. Versium reserves the right to change, modify or otherwise alter these Terms and Conditions at any time at Versium's sole discretion. Any and all modifications shall become effective immediately once posted. You

Average length: Over ***8*** pages!

…

Or Detailed Courses



Protecting Human Research Participants

NIH Office of Extramural Research

main menu | glossary | help | citations

Welcome back, Prasang

- ✓ Introduction
- ✓ History
- ✓ Codes and Regulations (5/6) ✓
- ✓ Respect for Persons (6/6) ✓
- ✓ Beneficence (5/5) ✓
- ✓ Justice (4/4) ✓
- ✓ Conclusion

Key

✓ = quiz passed

✓ = section read

Get Certificate
Feedback / Evaluation Survey
Review Completed Quizzes
Edit User Info
FAQ Page
Terms of Use Policy
PDF version of the course
Log Out

Developed: 3/1/2008

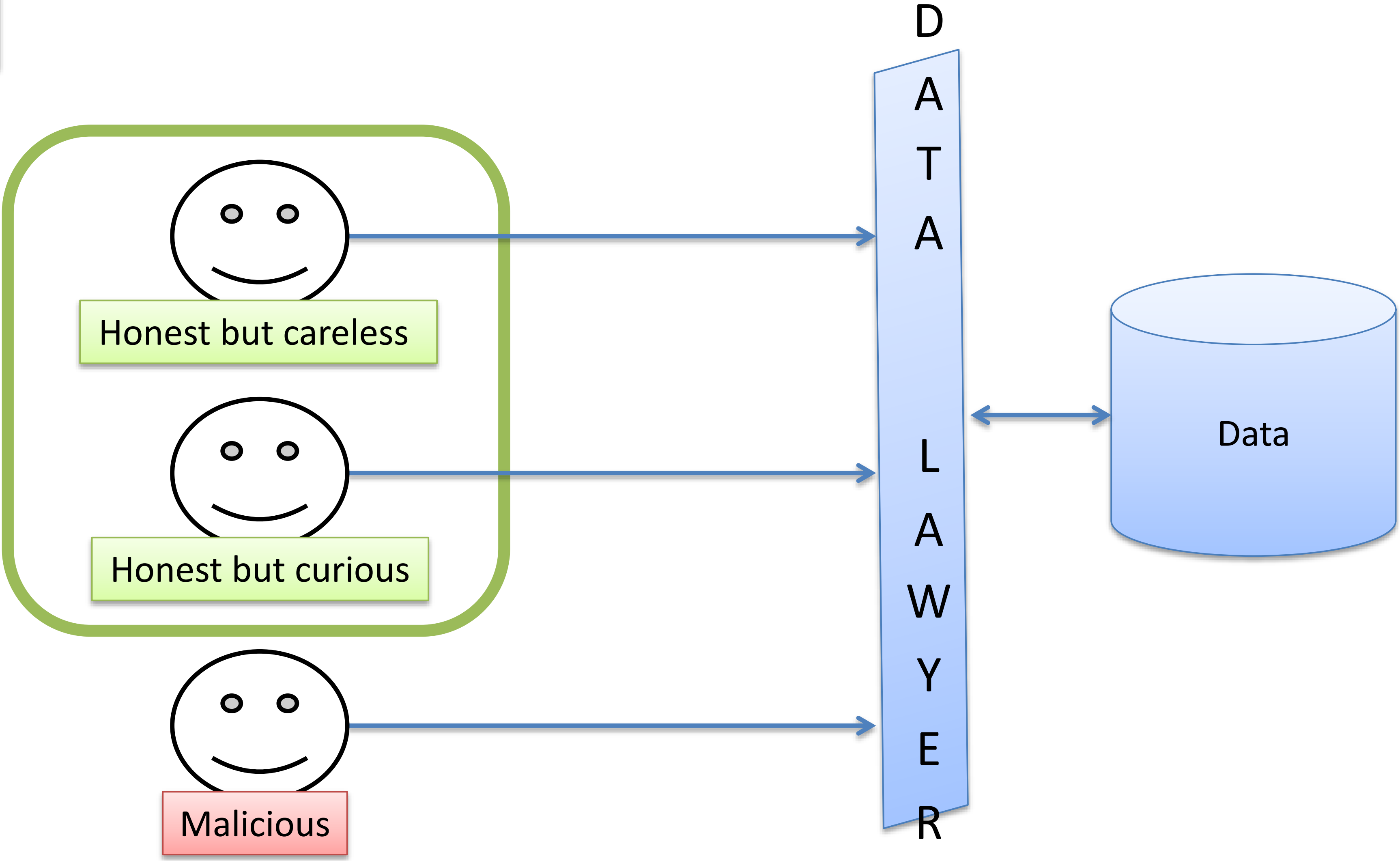
Updated: 2/4/2011

Problem with License Agreements

- Burden users with compliance
- Assume that users will comply

Managing Data License Agreements with DataLawyer

Trust but verify



Approach Overview

- Data seller defines *policies*

- *Data* and *policies* are loaded into a DataLawyer-enabled database system

- Buyer queries the data

- DataLawyer checks all queries before execution

Challenge: Semantics

Example Policy 1: Can access up to 10K records/month.

If the buyer computes a histogram on the data and filters out some buckets, did he use the input tuples from the filtered bucket or not?

Challenge: Performance

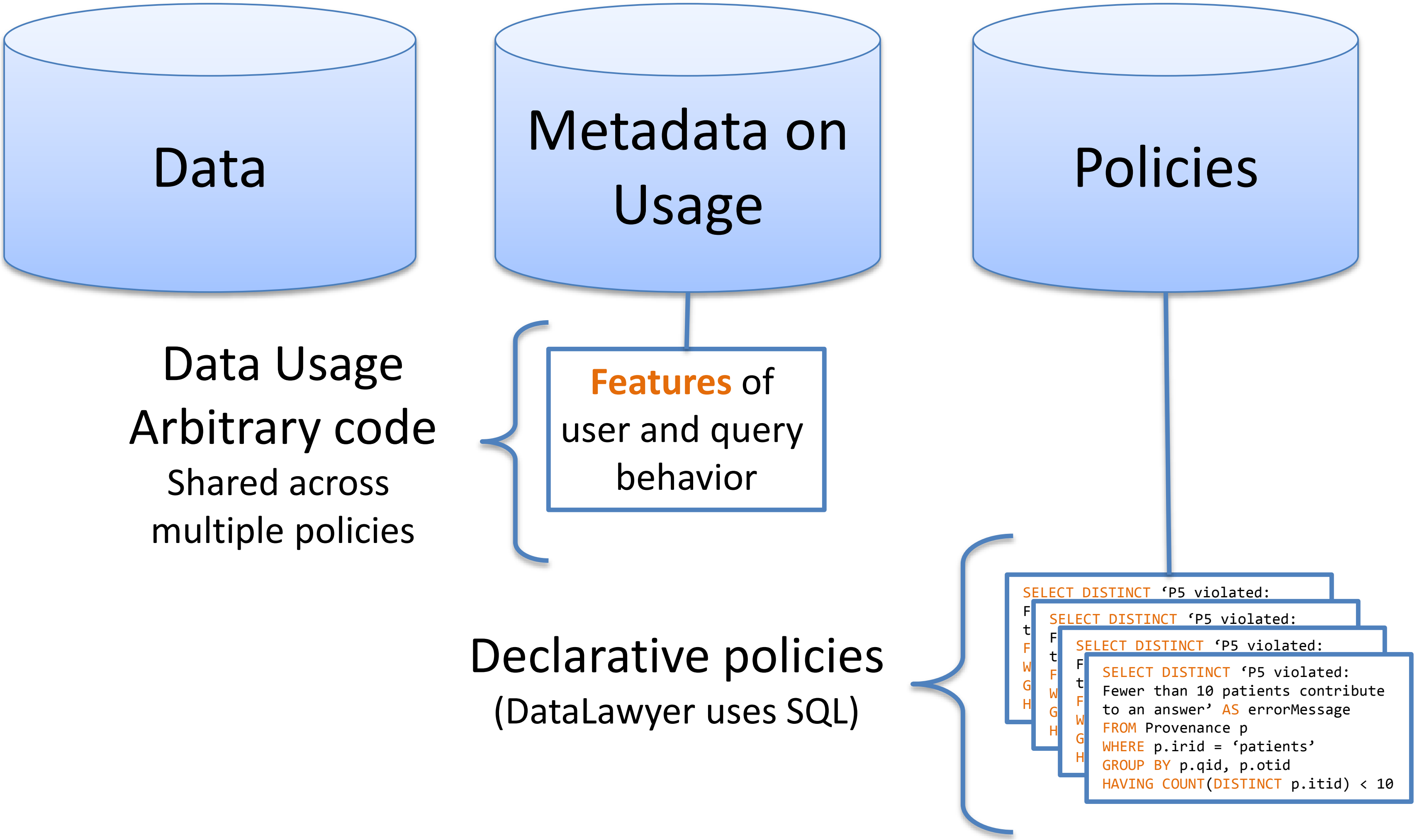
Cheap

Example Policy 2: Only allow aggregate queries where each output tuple must aggregate over at least 10 values.

Expensive

Policies are **expensive** to check online!

DataLawyer Setup



Features of user and query behavior

Data Usage

Arbitrary code

Shared across multiple policies

Declarative policies

(DataLawyer uses SQL)

```
SELECT DISTINCT ‘P5 violated:
Fewer than 10 patients contribute
to an answer’ AS errorMessage
FROM Provenance p
WHERE p.irid = ‘patients’
GROUP BY p.qid, p.otid
HAVING COUNT(DISTINCT p.itid) < 10
```

Usage Log

They capture *features* of a query that are used in the policies

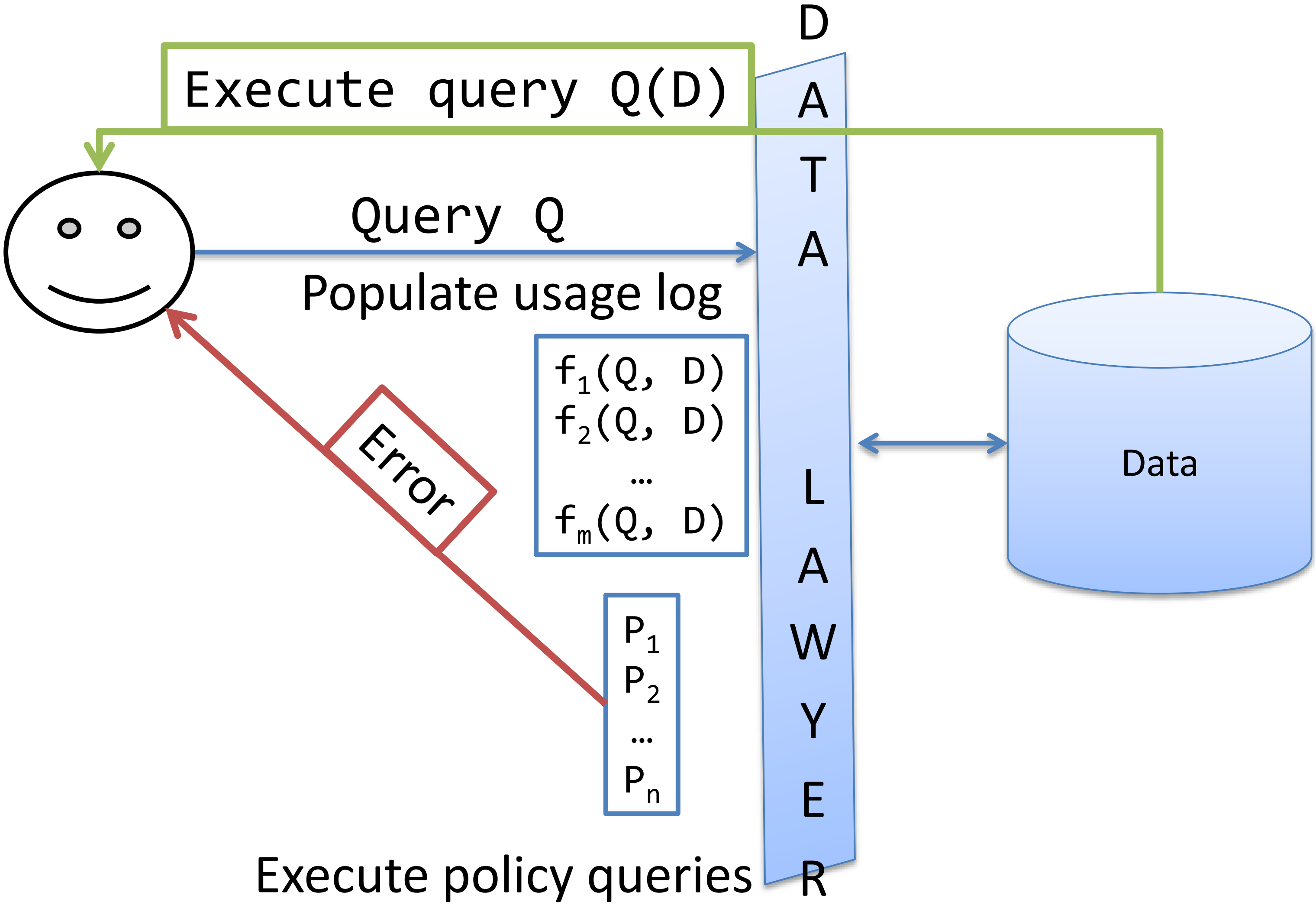
Examples are:

1. Provenance
2. User log
3. Static analysis of the query
4. Pricing
5. …

We require them to:

1. Be deterministic
2. Be append only
3. Contain a timestamp with each tuple

DataLawyer: Operation



Execute query Q(D)

Query Q

Populate usage log

$f_1(Q, D)$
$f_2(Q, D)$
…
$f_m(Q, D)$

Error

P_1
P_2
…
P_n

Execute policy queries

DATA LAWYER

Data

DataLawyer Workflow Example

Data: Patients

pid	disease	treatment	outcome
1	asthma	albuterol	positive
…			

Query: What fraction of asthma patients were treated with albuterol?

DataLawyer: Populates the usage logs

uid	query	table	column
1	1	Patient	treatment
1	1	Patient	outcome

DataLawyer checks policies, which are queries over the usage logs

- Queries are not allowed to access column pid
- Queries must aggregate data from at least 10 rows in Patients

Example Using SQL

Policy: Stop queries where *fewer* than 10 patients contribute to *any* output tuple.

```
SELECT DISTINCT ‘P5 violated: Fewer than 10 patients contribute to an answer’
             AS errorMessage
FROM Provenance p
WHERE p.irid = ‘patients’
GROUP BY p.qid, p.otid
HAVING COUNT(DISTINCT p.itid) < 10
```

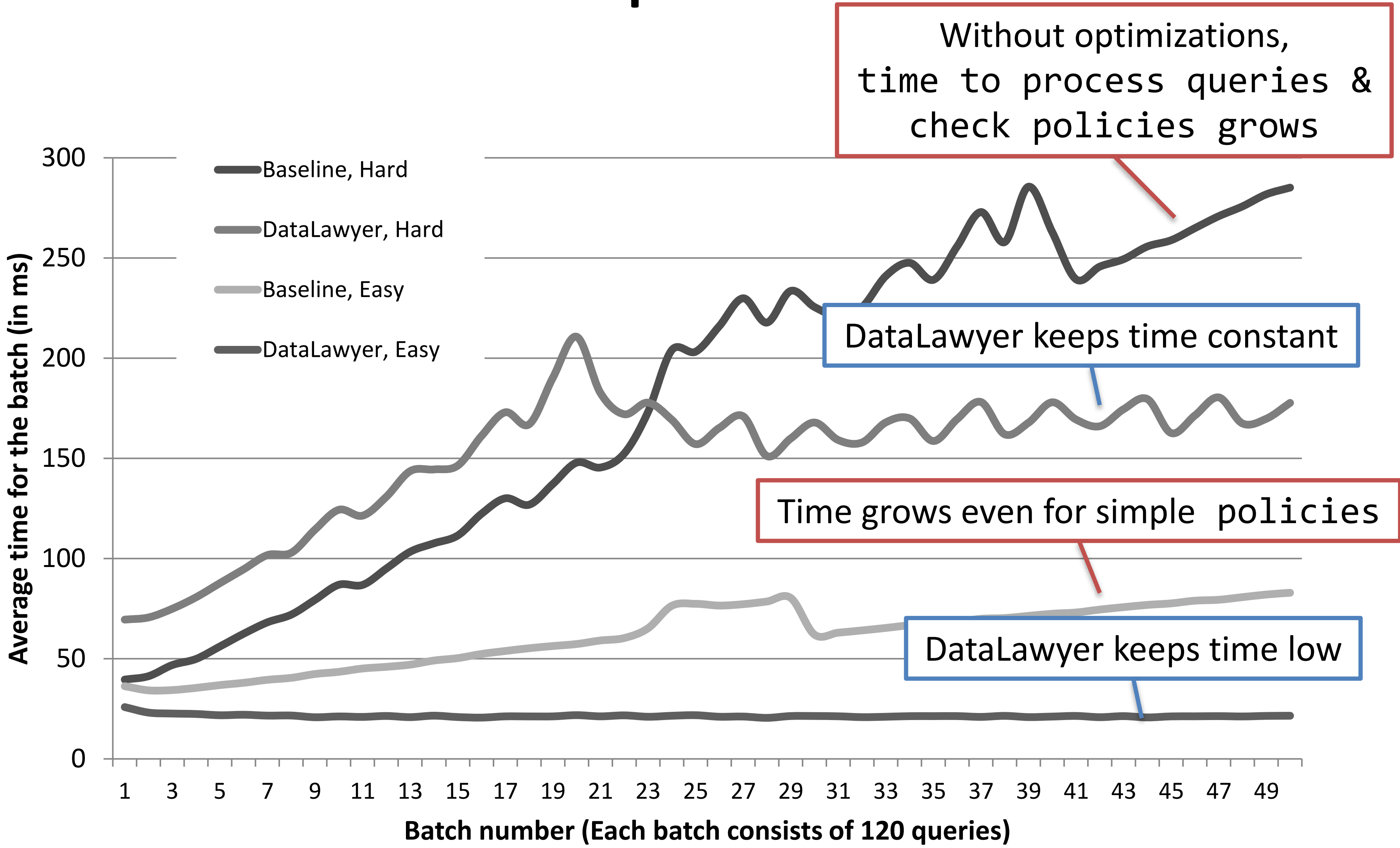
Policy refers to the provenance usage log

If `false`, no violation

If `true`, at least
one example of a violation

```
Usage log that captures how each tuple in query result
was derived from records on disk
Provenance(ts,      // Timestamp
           qid,     // Query id
           otid,    // Output tuple id, a hash of the output tuple
           irid,    // Input relation id, usually the name
           itid     // Input tuple id, usually the name
          )
```

Need for Optimizations



Without optimizations, time to process queries & check policies grows

DataLawyer keeps time constant

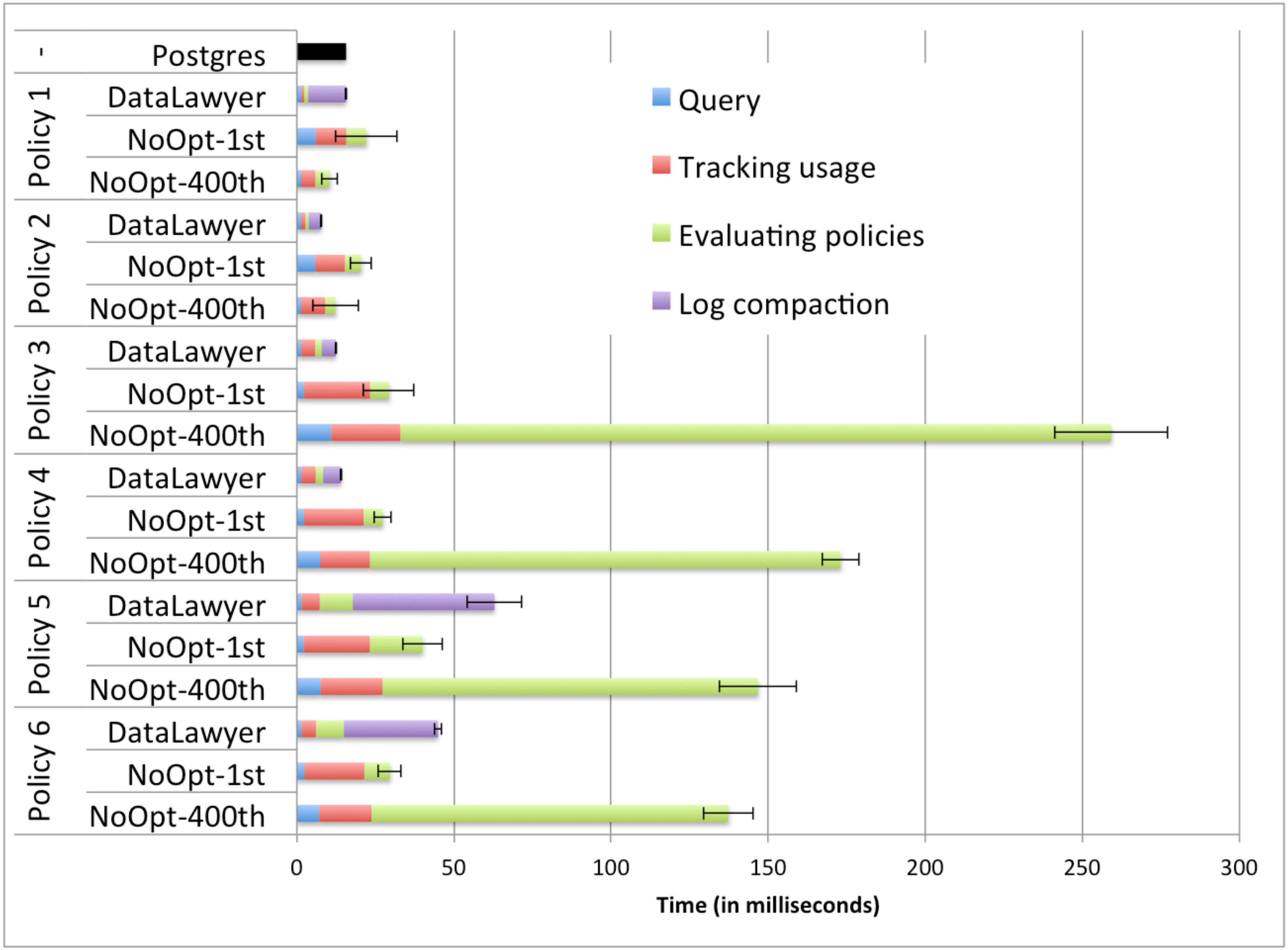
Time grows even for simple policies

DataLawyer keeps time low

Policy Evaluation

- There are three major steps:
  - Generate the usage logs
  - Evaluate policies
  - Write log to disk if everything is okay, else abort
- Our optimizations:
  - Avoid generating the logs
  - Prune the logs by removing data no longer needed
  - Avoid evaluating all policies
  - Try to evaluate cheaper, partial policies first

DataLawyer Performance Illustration



Data License Agreements Summary

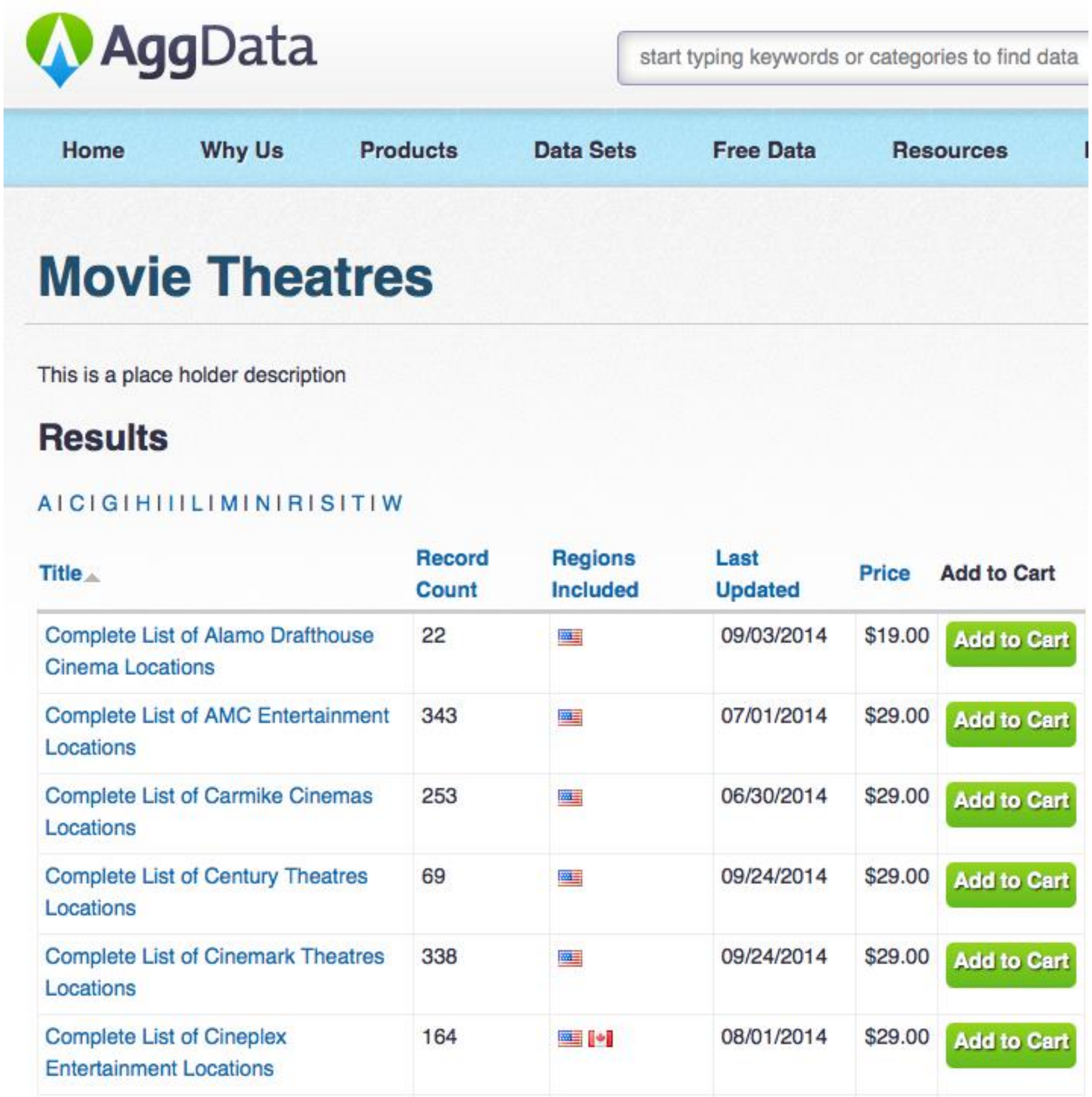
- Data comes with terms of use
  - Even free data often has terms of use
- Today, terms of use are written in natural language
  - Compliance for buyers is tedious and error-prone
- Possible to automate the process: DataLawyer
  - Enables more precise terms of use specification
  - Enables efficient enforcement
- Open problems
  - Malicious users
  - Data leaving database system

What are the Technical Challenges

- Challenge 1: Data License Agreements
  - All data comes with terms of use
  - Can we automate their enforcement?

- Challenge 2: Data Pricing
  - Existing pricing methods are limited
  - Can we support flexible pricing?

Data Pricing Today: Fixed



Movie Theatres

This is a place holder description

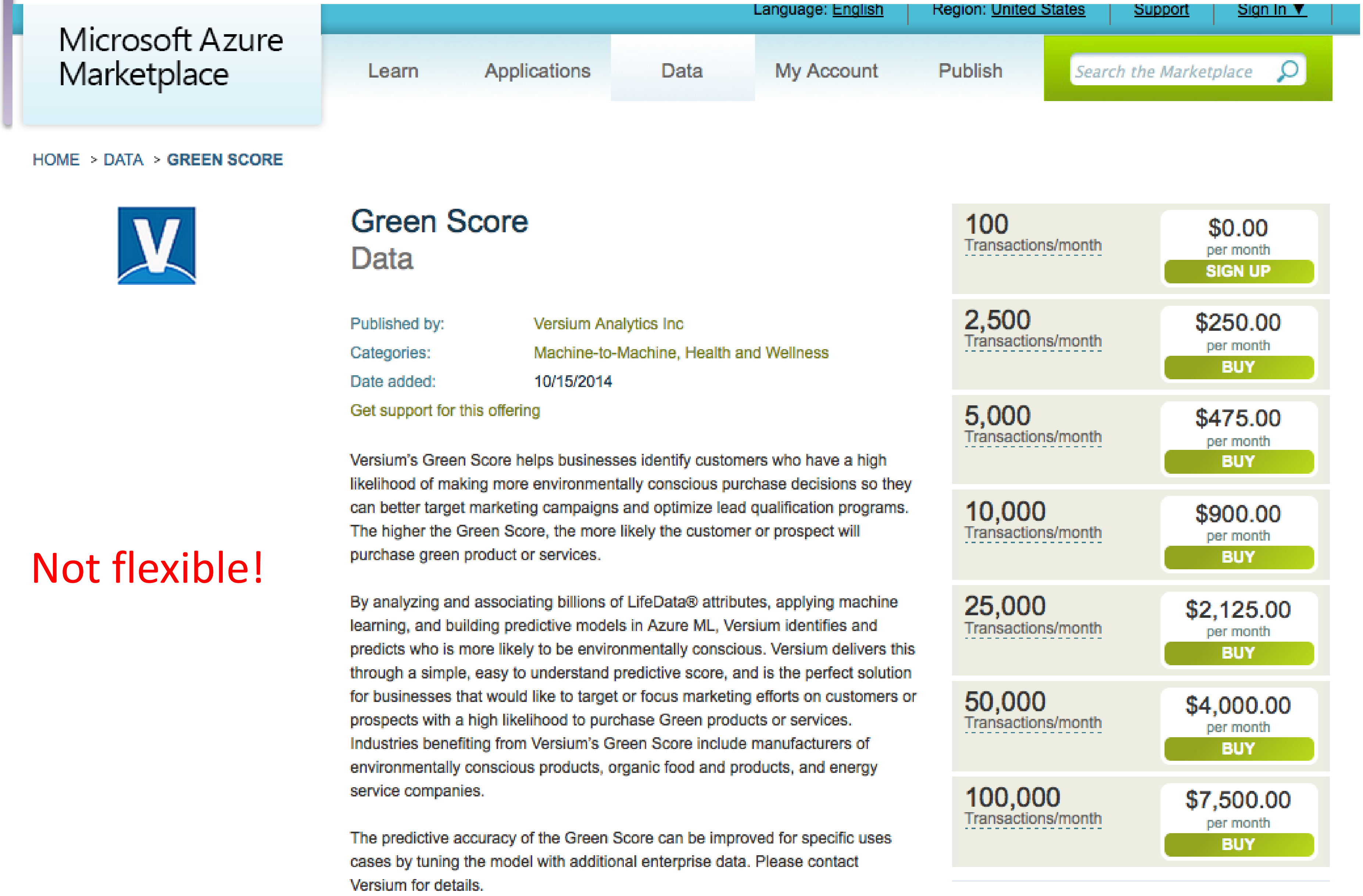
Results

A | C | G | H | I | L | M | N | R | S | T | W

Title	Record Count	Regions Included	Last Updated	Price	Add to Cart
Complete List of Alamo Drafthouse Cinema Locations	22	🇺🇸	09/03/2014	$19.00	Add to Cart
Complete List of AMC Entertainment Locations	343	🇺🇸	07/01/2014	$29.00	Add to Cart
Complete List of Carmike Cinemas Locations	253	🇺🇸	06/30/2014	$29.00	Add to Cart
Complete List of Century Theatres Locations	69	🇺🇸	09/24/2014	$29.00	Add to Cart
Complete List of Cinemark Theatres Locations	338	🇺🇸	09/24/2014	$29.00	Add to Cart
Complete List of Cineplex Entertainment Locations	164	🇺🇸 🇨🇦	08/01/2014	$29.00	Add to Cart

Not flexible!

Data Pricing Today: Subscriptions



HOME > DATA > GREEN SCORE

Green Score
Data

Published by: Versium Analytics Inc

Categories: Machine-to-Machine, Health and Wellness

Date added: 10/15/2014

Get support for this offering

Versium's Green Score helps businesses identify customers who have a high likelihood of making more environmentally conscious purchase decisions so they can better target marketing campaigns and optimize lead qualification programs. The higher the Green Score, the more likely the customer or prospect will purchase green product or services.

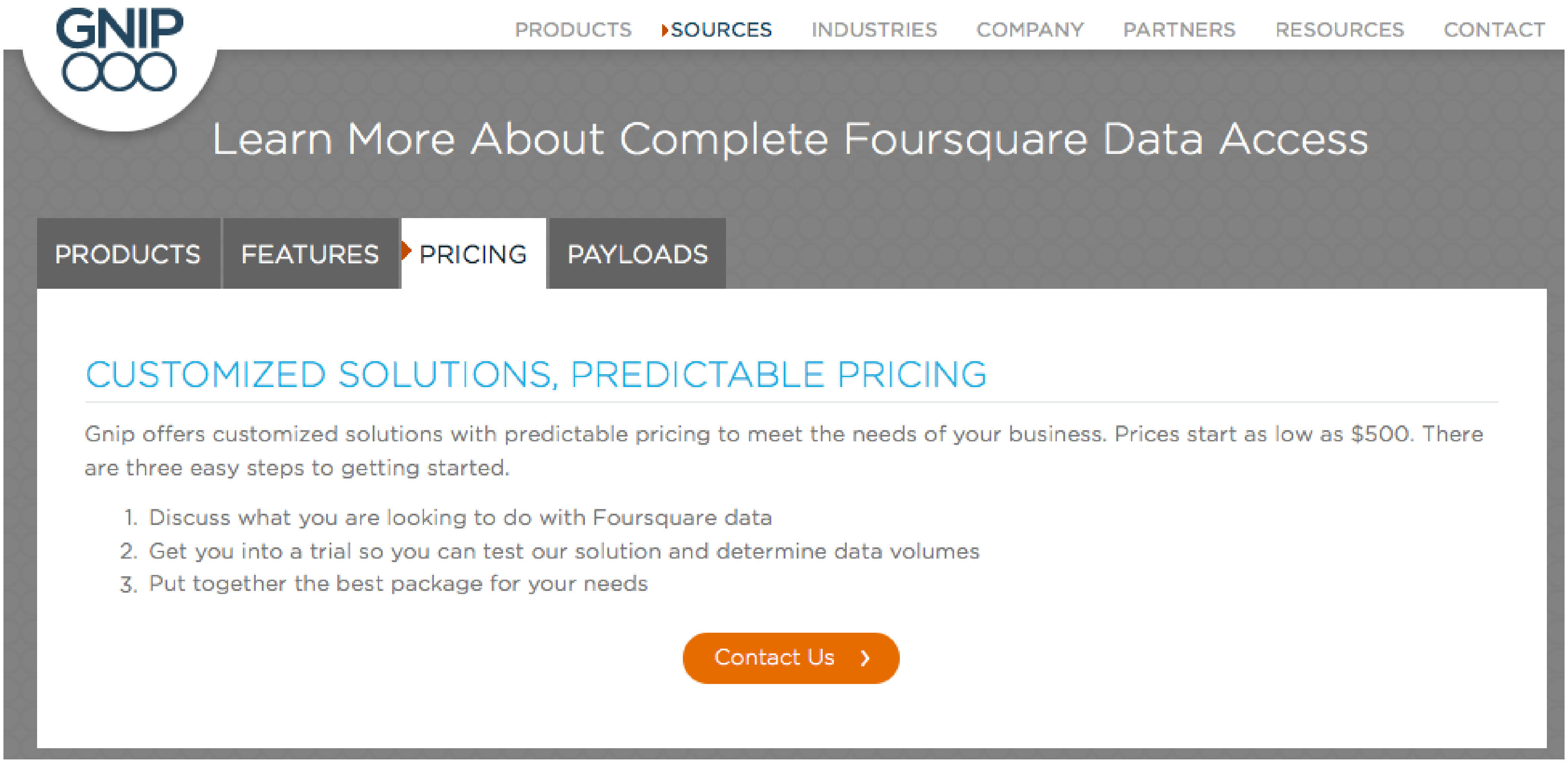
By analyzing and associating billions of LifeData® attributes, applying machine learning, and building predictive models in Azure ML, Versium identifies and predicts who is more likely to be environmentally conscious. Versium delivers this through a simple, easy to understand predictive score, and is the perfect solution for businesses that would like to target or focus marketing efforts on customers or prospects with a high likelihood to purchase Green products or services. Industries benefiting from Versium's Green Score include manufacturers of environmentally conscious products, organic food and products, and energy service companies.

The predictive accuracy of the Green Score can be improved for specific uses cases by tuning the model with additional enterprise data. Please contact Versium for details.

100 Transactions/month	$0.00 per month SIGN UP
2,500 Transactions/month	$250.00 per month BUY
5,000 Transactions/month	$475.00 per month BUY
10,000 Transactions/month	$900.00 per month BUY
25,000 Transactions/month	$2,125.00 per month BUY
50,000 Transactions/month	$4,000.00 per month BUY
100,000 Transactions/month	$7,500.00 per month BUY

Not flexible!

Data Pricing Today: Private Price



Learn More About Complete Foursquare Data Access

CUSTOMIZED SOLUTIONS, PREDICTABLE PRICING

Gnip offers customized solutions with predictable pricing to meet the needs of your business. Prices start as low as $500. There are three easy steps to getting started.

1. Discuss what you are looking to do with Foursquare data
2. Get you into a trial so you can test our solution and determine data volumes
3. Put together the best package for your needs

Not scalable!

Example Scenario

- Seller has a database of cities and business contact information
  - Businesses in one province or state: $300
  - One type of business: $150
  - Cities with given climate: $10

- Buyer:
  - Q1: “Businesses with more than 200 employees” (selection)
  - Q2: “West-coast businesses in cities with high yearly precipitation” (join)

- How to satisfy buyer?

Current Pricing: Fixed Prices

- Fixed price for entire dataset
    - Must create and price views specific to queries Q1 and Q2
    - OR user must buy entire dataset if view not available
    - AND user must perform joins by herself
        - Certainly the case if datasets have different owners

Current Pricing: Subscriptions

- Subscriptions
  - Fixed number of *transactions* per month
  - Must create and price appropriate parameterized queries
  - Today queries are dataset specific (i.e., no joins!)

  - Can satisfy Q1: “Businesses with more than 200 employees”
  - Cannot Q2: “West-coast businesses in cities with high yearly precipitation”

Other Data Pricing Issues

- Today's data pricing can also have **bad properties**

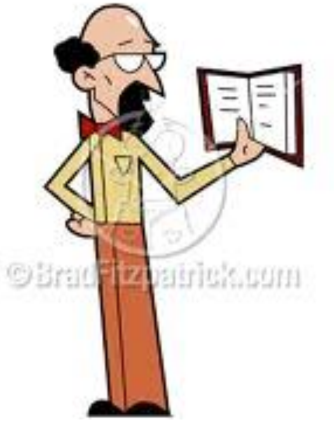
- Example: Weather Imagery on Azure DataMarket
  - 1,000,000 transactions -> $2,400
  - 100,000 -> $600
  - 10,000 -> $120
  - 2,500 -> $0

- **Arbitrage opportunity**:
  - Emulate many users
  - Get as much data as you want for free!

Data Pricing with QueryMarket

Query-Based Pricing

- Seller specifies a set of queries Q_1, … Q_n
- These queries form *views* on the data for sell
- Seller prices the views: price(Q_1), …, price(Q_n)
  - D = all cities and businesses in North America
  - V_1 (businesses in one state) = \$300
  - V_2 (businesses of one type) = \$150
  - V_3 (cities with a given climate) = \$10



Query-Based Pricing

- QueryMarket system computes other query prices
  - Q2: “West-coast businesses in cities with high yearly precipitation”
  - Key idea: Compute least expensive set of views that can be used to answer the query. The sum of the price of these views is the price of the query

- System guarantees price properties
  - Arbitrage-free prices
  - Maximal prices (no unintended discounts)

Conclusion

- Data has value
- Data is bought and sold online
- Supporting modern data markets requires
  - New tools for managing license agreements
  - New methods for pricing data
- Much work remains to be done

http://cloud-data-pricing.cs.washington.edu

Potential Techniques

	Online	**Offline**
Fuzzy Semantics	Privacy Mechanisms (Reduces data utility)	Intrusion Detection System (Assumes the user is malicious. Offline)
Precise Semantics	Access Control (Want full access to data)	Auditing Systems (Offline)

None of these work!