OPOWER

Data Science with Hadoop at Opower

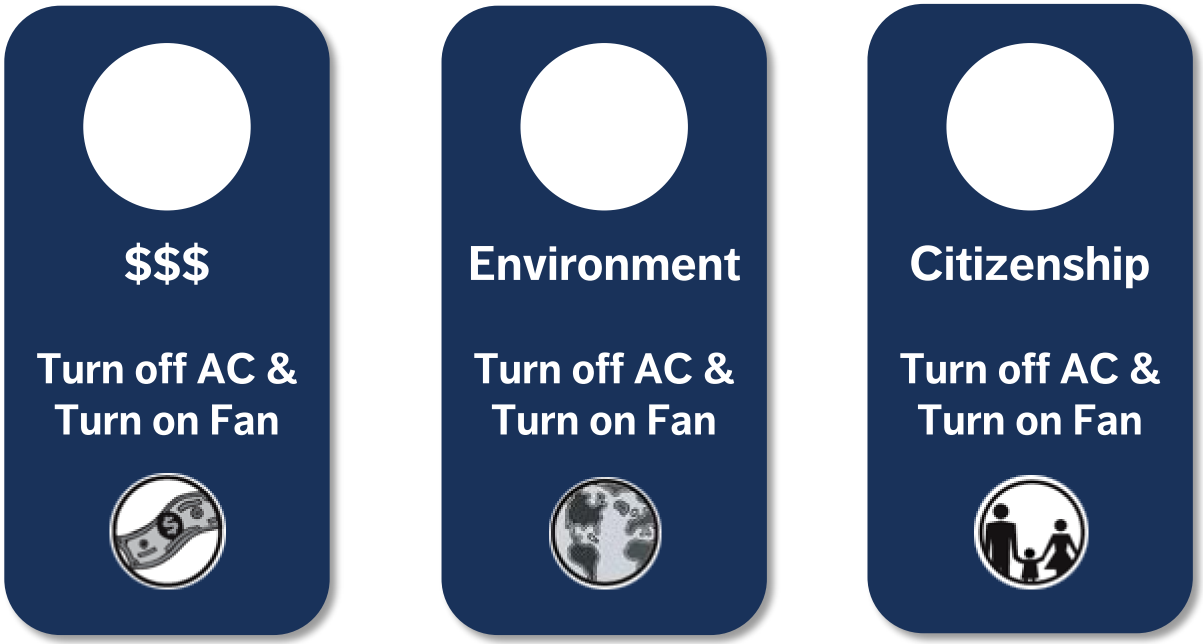
Erik Shilts

Advanced Analytics

erik.shilts@opower.com

What is Opower?

A study:



$$$

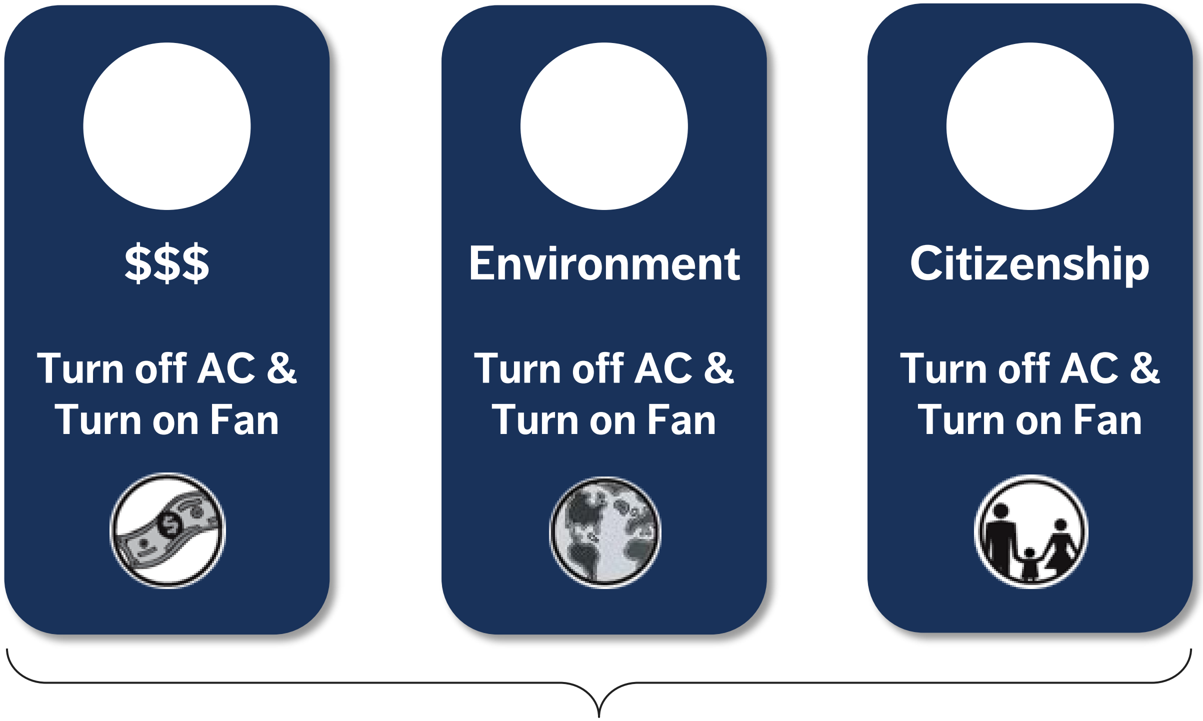
Turn off AC &
Turn on Fan

Environment

Turn off AC &
Turn on Fan

Citizenship

Turn off AC &
Turn on Fan



$$$

Turn off AC &
Turn on Fan

Environment

Turn off AC &
Turn on Fan

Citizenship

Turn off AC &
Turn on Fan

Zero Impact on Consumption

$$$

Turn off AC &
Turn on Fan

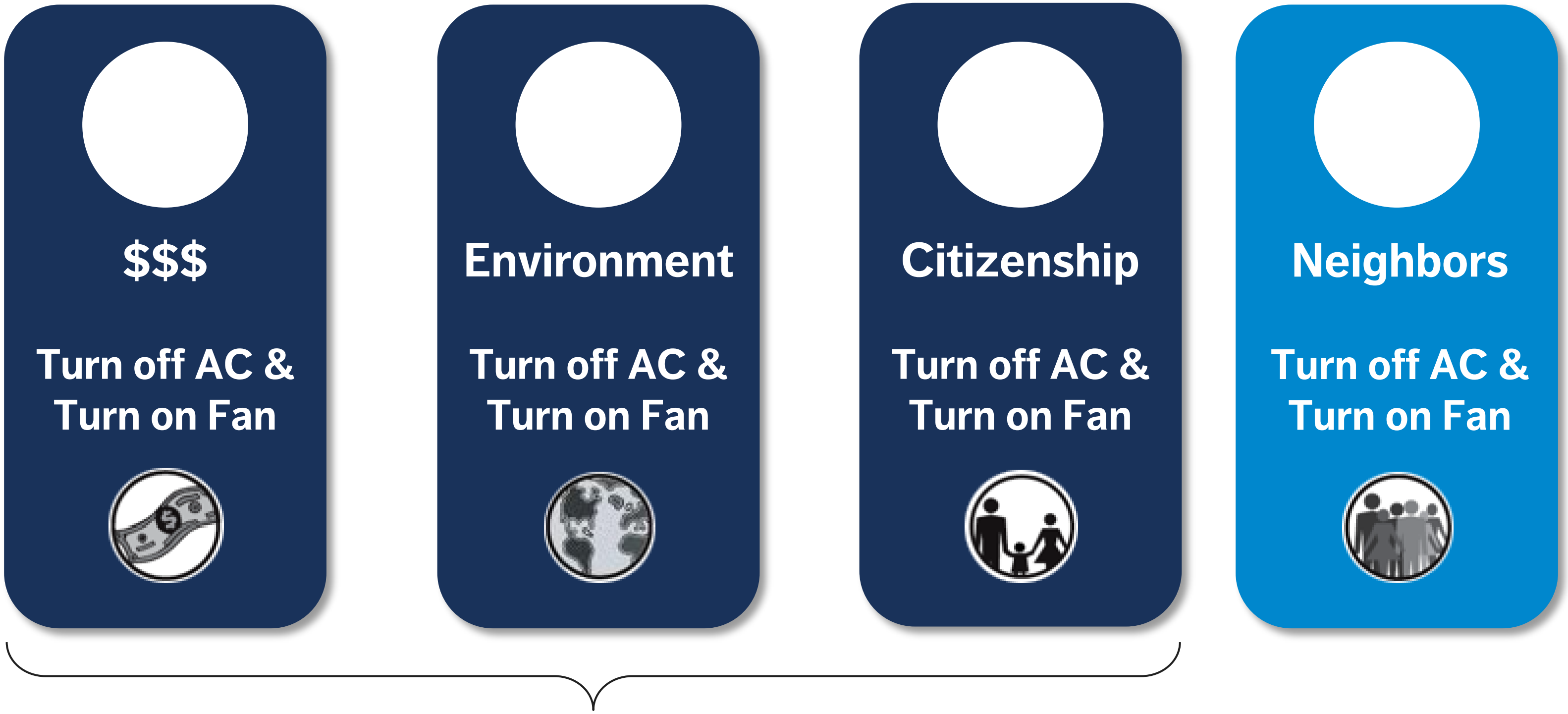
Environment

Turn off AC &
Turn on Fan

Citizenship

Turn off AC &
Turn on Fan

Zero Impact on Consumption



$$$

Turn off AC & Turn on Fan

Environment

Turn off AC & Turn on Fan

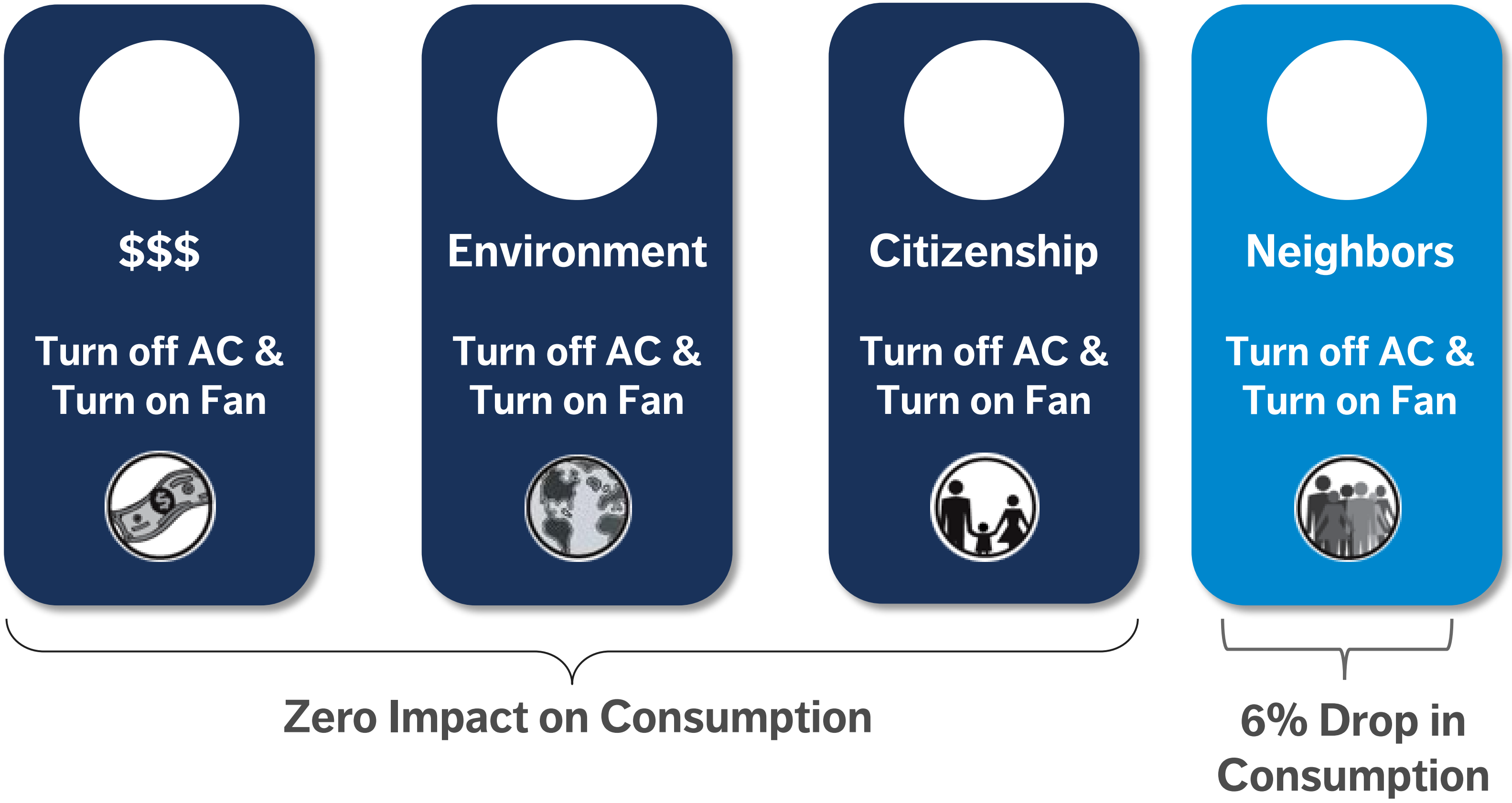
Citizenship

Turn off AC & Turn on Fan

Neighbors

Turn off AC & Turn on Fan

Zero Impact on Consumption



$$$

Turn off AC & Turn on Fan

Environment

Turn off AC & Turn on Fan

Citizenship

Turn off AC & Turn on Fan

Neighbors

Turn off AC & Turn on Fan

Zero Impact on Consumption

6% Drop in Consumption

Opower Details

Customer Engagement Platform
for Utilities

Company

- ~300 employees
- Cleantech Company of the Year 2012!
- 75 utility partners covering > 50M households
- > 1.5 Terawatt hours saved

Our DNA

- **Data analytics**
- Behavioral science

What is Opower?

What is Opower?

One giant big data problem

Advanced Analytics

Advanced Analytics provides consumer insights

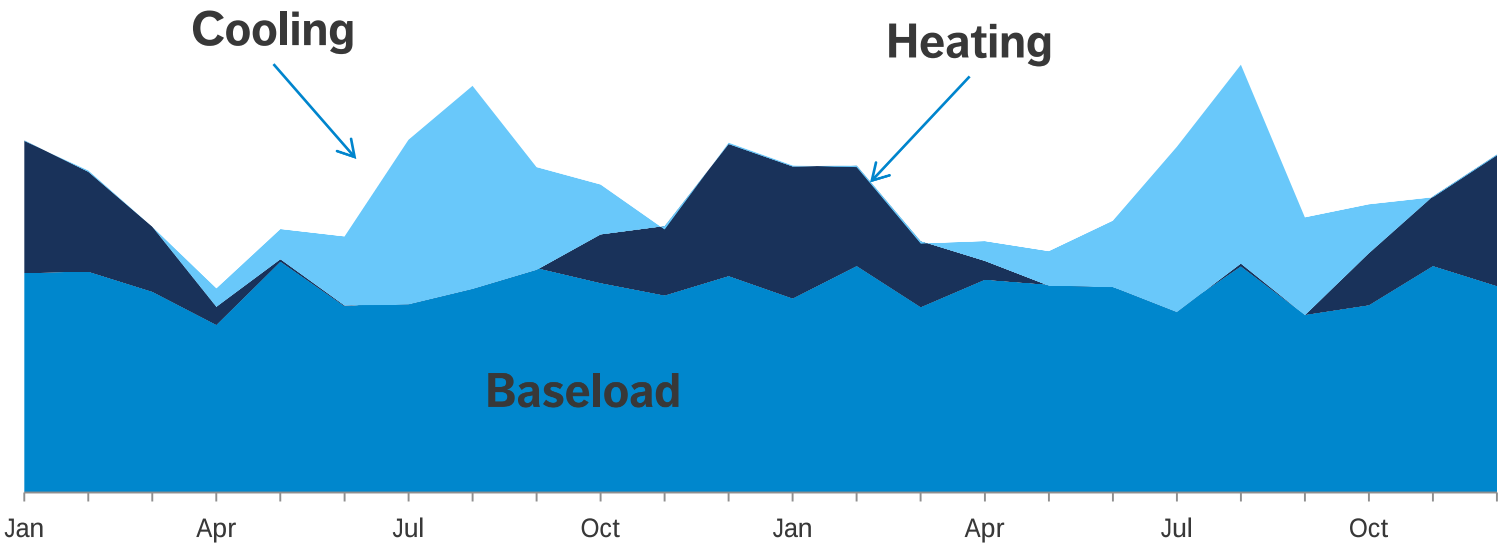
Our charter is to provide consumers with **insights** that give **context** and **control** over how they use energy.

We use machine learning and predictive modeling

Our charter is to provide consumers with **insights** that give **context** and **control** over how they use energy.

Use **machine learning**, signal processing, and **predictive modeling** to provide energy usage insights.

We provide insights into individual energy use



Data science

Data scientists extract meaning

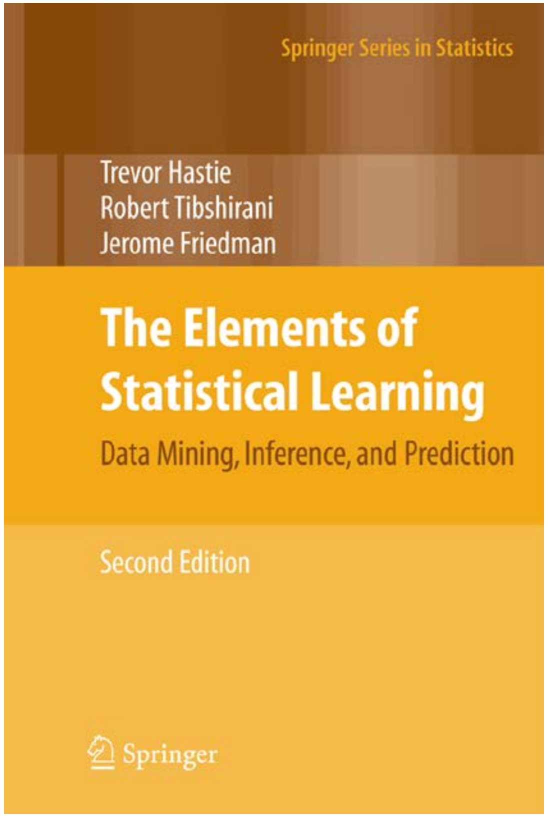
Data science is a discipline … with the goal of **extracting meaning from data** and creating data products.

Wikipedia: http://en.wikipedia.org/wiki/Data_science

Data scientists are statisticians

Data science is a discipline … with the goal of **extracting meaning from data** and creating data products.

In other words, **machine learning**, **statistics**, and **pretty charts**.

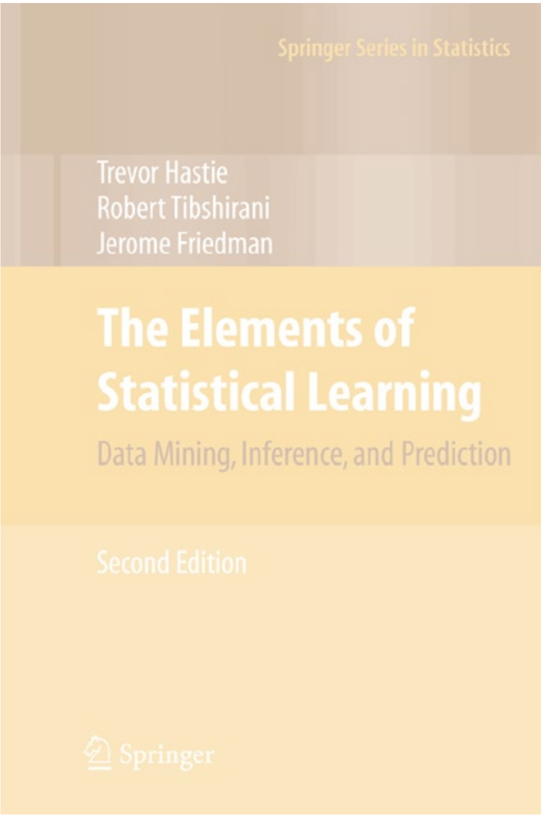

Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Wikipedia: http://en.wikipedia.org/wiki/Data_science

Data scientists want to extract meaning

Data science is a discipline … with the goal of **extracting meaning from data** and creating data products.

In other words, **machine learning**, **statistics**, and **pretty charts**.



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer

Wikipedia: http://en.wikipedia.org/wiki/Data_science

Data scientists are data mungers

Data science is a discipline … of **data munging**.

Data scientists prepare data

Data science is a discipline … of **data munging**.

Data munging is the process of **converting data** from one form into another for more **convenient consumption**.

Wikipedia: http://en.wikipedia.org/wiki/Data_wrangling

Data scientists are plumbers

Data science is a discipline … of **plumbing**.

Plumbing is **difficult**.

It’s temporary, I swear!

Data science is a discipline … of **plumbing**.

Move data from here to there.

Hack to get the data how you want it.

http://funmeme.com/post/2009/08/02/Plumbing-FAIL-e28093-Funny-Pic.aspx

It works. For now.

Data science is a discipline … of **plumbing**.

Multiple sources are tricky to handle.

Construct a **series of tubes**.

http://www.ontimeplumber.com.au/plumbing_disasters/plumbing_disasters.html

Needs user testing

Data science is a discipline … of **plumbing**.

Sometimes you have to **start over** when you think you're done.

http://www.funnyjunk.com/funny_pictures/234485/Awkward/

Data science is mostly plumbing

Data science is a discipline … of **plumbing**.

It’s where we spend all of our time

Data science is a discipline … of plumbing.

We spend **80%** of our time on data munging and other **infrastructure** work.

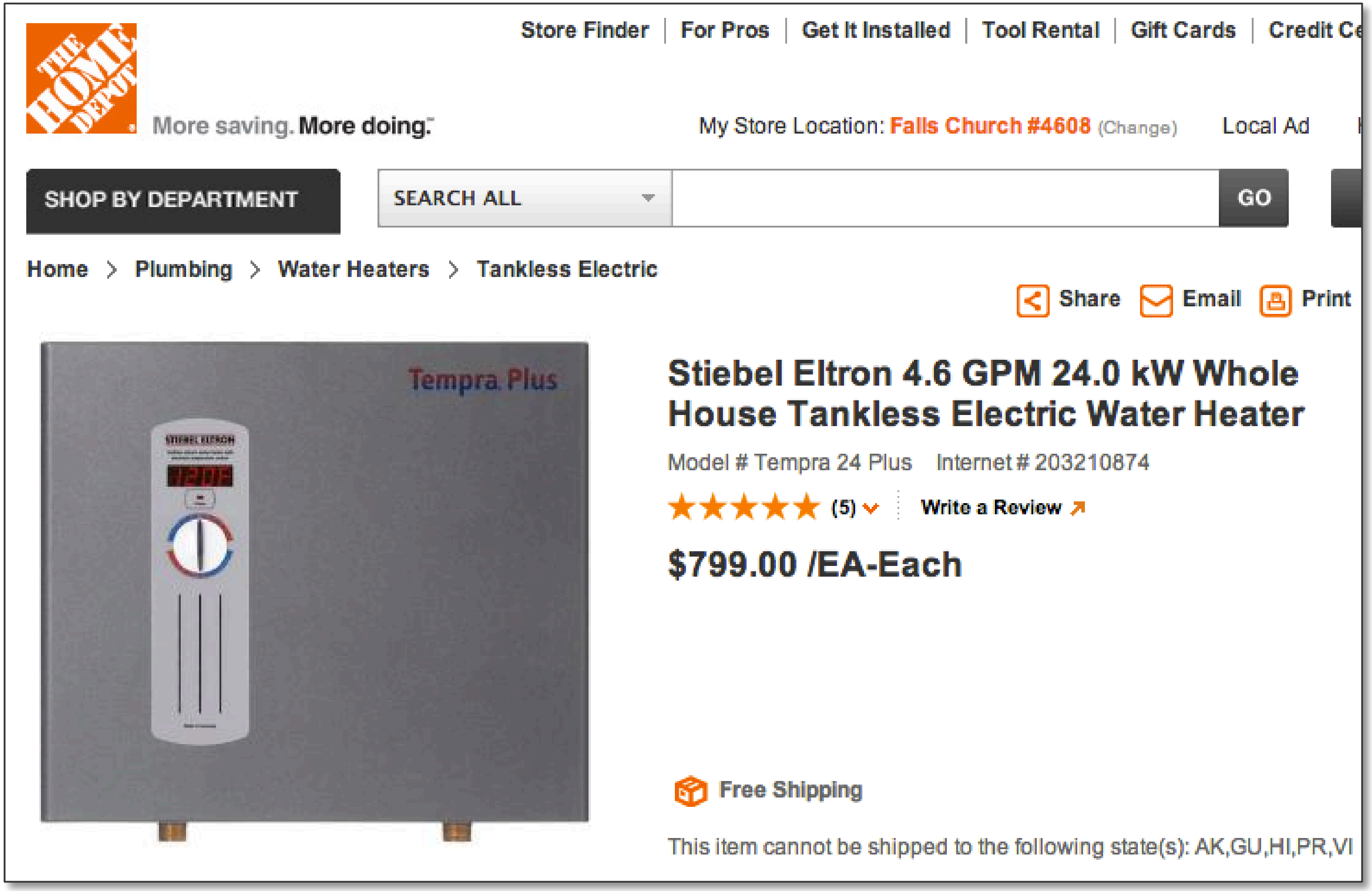
Fun stuff only 20% of the time

Data science is a discipline … of **plumbing**.

- We spend **80%** of our time on data munging and other **infrastructure** work.
- Sprinkle on some **modeling** and **charts** for the other **20%**.

Data science in practice

Electric tankless water heater 10% off


Store Finder | For Pros | Get It Installed | Tool Rental | Gift Cards | Credit Ce
THE HOME DEPOT
More saving. More doing.
My Store Location: Falls Church #4608 (Change)
Local Ad
SHOP BY DEPARTMENT
SEARCH ALL
GO
Home > Plumbing > Water Heaters > Tankless Electric
Share
Email
Print
Tempra Plus
STIEBEL ELTRON
120F

Stiebel Eltron 4.6 GPM 24.0 kW Whole House Tankless Electric Water Heater

Model # Tempra 24 Plus Internet # 203210874

★★★★★ (5) Write a Review

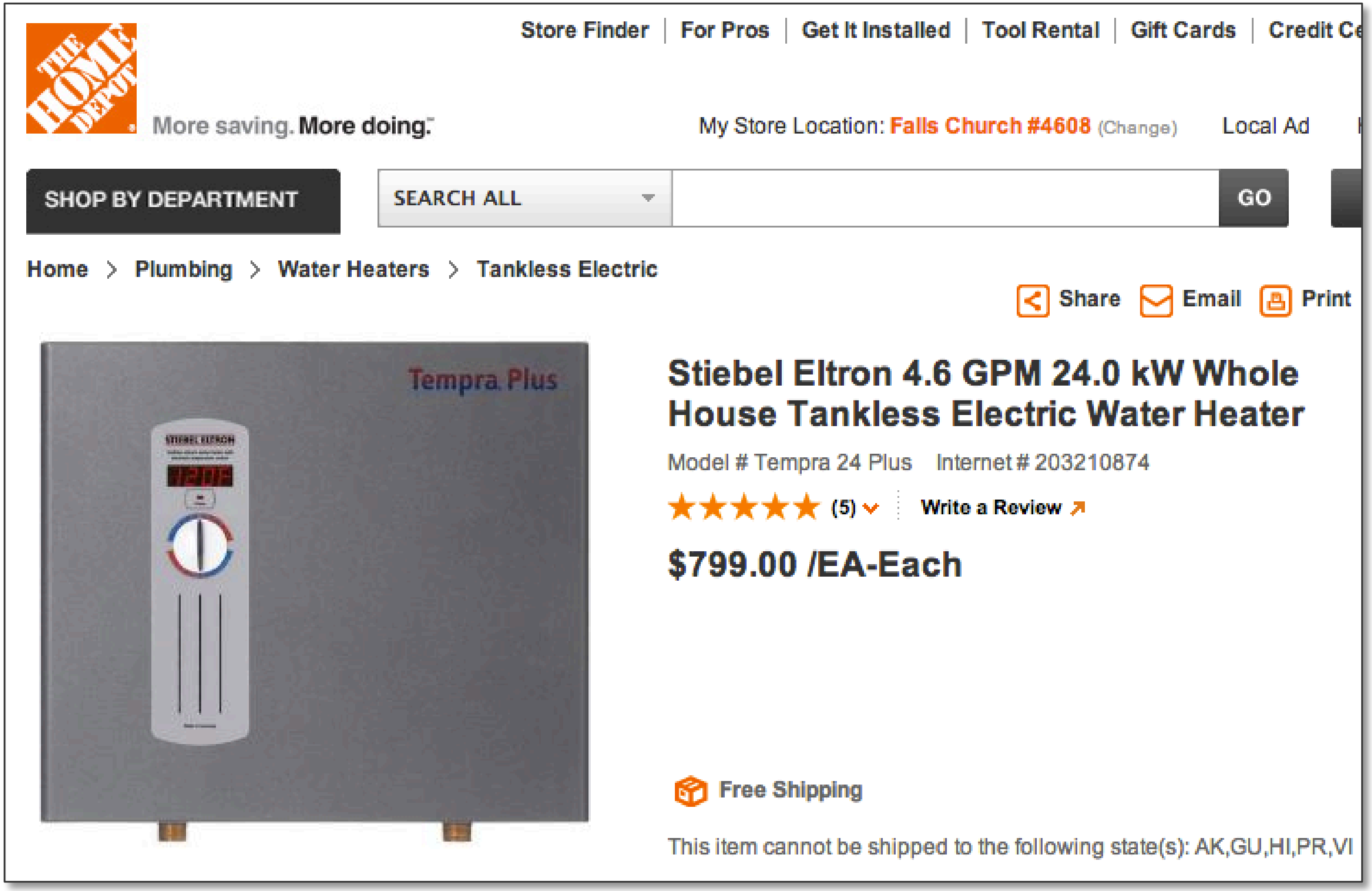
$799.00 /EA-Each

Free Shipping

This item cannot be shipped to the following state(s): AK,GU,HI,PR,VI

http://www.homedepot.com/Plumbing-Water-Heaters-Tankless-Electric/h_d1/N-5yc1vZc1ty/R-203210874/h_d2/ProductDisplay?catalogId=10053&langId=-1&storeId=10051

Who should get this promotion?



Store Finder | For Pros | Get It Installed | Tool Rental | Gift Cards | Credit Ce
THE HOME DEPOT
More saving. More doing.
My Store Location: Falls Church #4608 (Change)
Local Ad
SHOP BY DEPARTMENT
SEARCH ALL
GO
Share
Email
Print
Tempra Plus

Home > Plumbing > Water Heaters > Tankless Electric

Stiebel Eltron 4.6 GPM 24.0 kW Whole House Tankless Electric Water Heater

Model # Tempra 24 Plus Internet # 203210874

★★★★★ (5) Write a Review

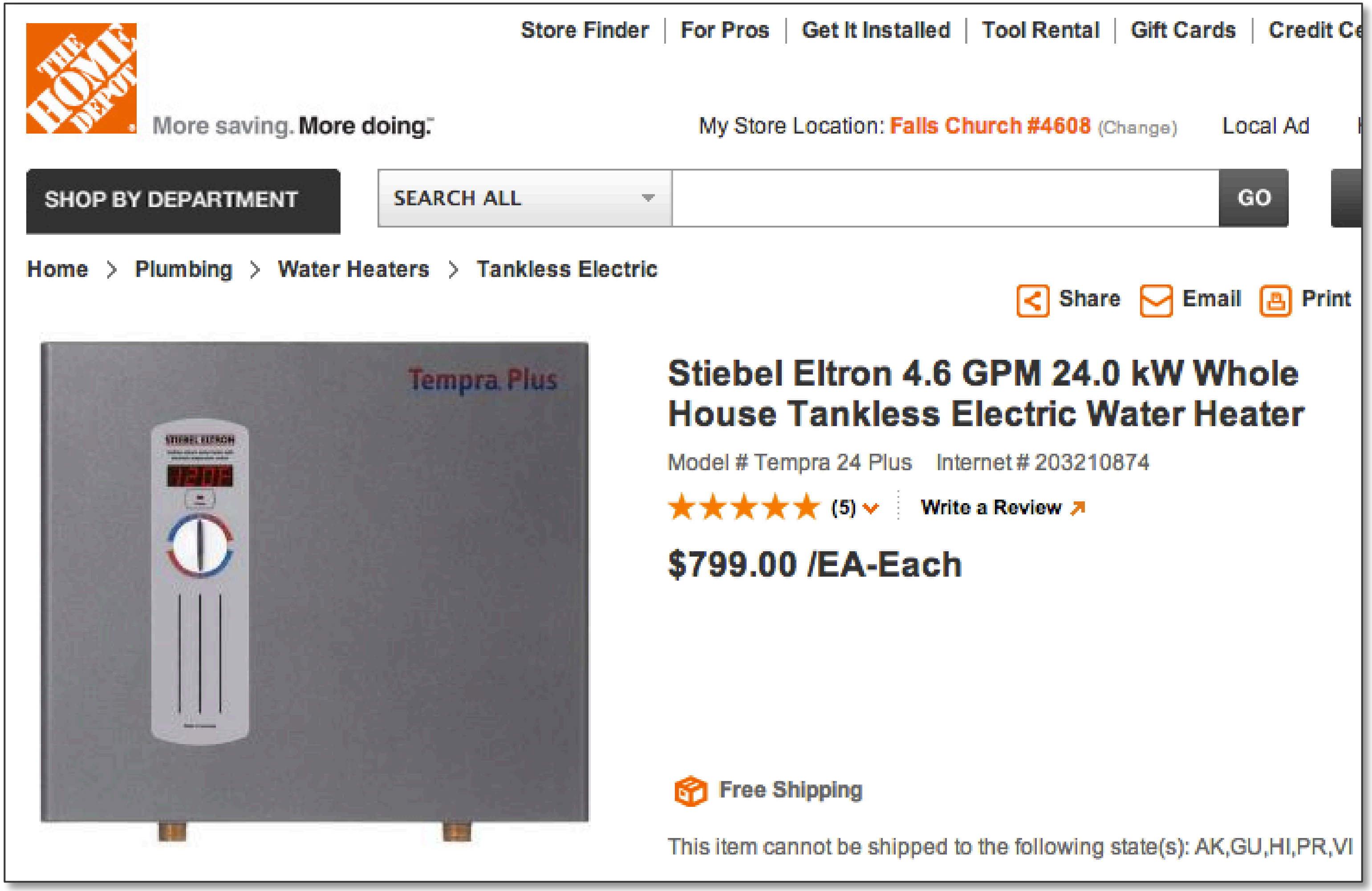
$799.00 /EA-Each

Free Shipping

This item cannot be shipped to the following state(s): AK,GU,HI,PR,VI

http://www.homedepot.com/Plumbing-Water-Heaters-Tankless-Electric/h_d1/N-5yc1vZc1ty/R-203210874/h_d2/ProductDisplay?catalogId=10053&langId=-1&storeId=10051

Maximize take-up rate



Store Finder | For Pros | Get It Installed | Tool Rental | Gift Cards | Credit Ce

More saving. More doing.

My Store Location: Falls Church #4608 (Change)

Local Ad

SHOP BY DEPARTMENT

SEARCH ALL

GO

Home > Plumbing > Water Heaters > Tankless Electric

Share Email Print

Stiebel Eltron 4.6 GPM 24.0 kW Whole House Tankless Electric Water Heater

Model # Tempra 24 Plus Internet # 203210874

★★★★★ (5) Write a Review

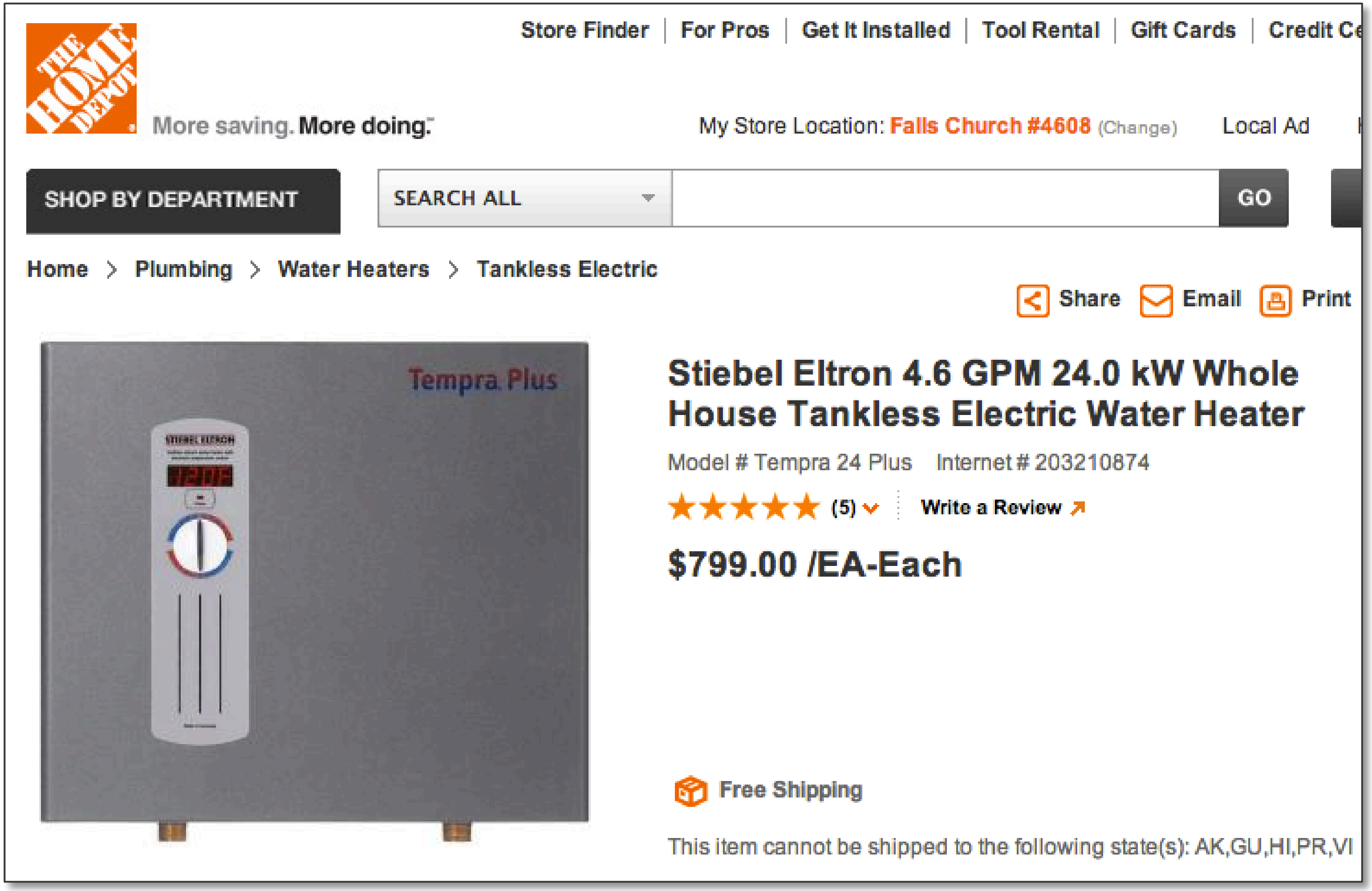
$799.00 /EA-Each

Free Shipping

This item cannot be shipped to the following state(s): AK,GU,HI,PR,VI

http://www.homedepot.com/Plumbing-Water-Heaters-Tankless-Electric/h_d1/N-5yc1vZc1ty/R-203210874/h_d2/ProductDisplay?catalogId=10053&langId=-1&storeId=10051

Minimize marketing cost



Store Finder | For Pros | Get It Installed | Tool Rental | Gift Cards | Credit Ce

My Store Location: Falls Church #4608 (Change)

Home > Plumbing > Water Heaters > Tankless Electric

Stiebel Eltron 4.6 GPM 24.0 kW Whole House Tankless Electric Water Heater

Model # Tempra 24 Plus Internet # 203210874

★★★★★ (5) Write a Review

$799.00 /EA-Each

Free Shipping

This item cannot be shipped to the following state(s): AK,GU,HI,PR,VI

http://www.homedepot.com/Plumbing-Water-Heaters-Tankless-Electric/h_d1/N-5yc1vZc1ty/R-203210874/h_d2/ProductDisplay?catalogId=10053&langId=-1&storeId=10051

Data science in practice

Identify likely purchasers

Data science in the past

How would we have solved this **before Hadoop**?

Past is same as the present: construct a model

How would we have solved this **before Hadoop**?

Construct a **model** of likely purchasers.

Predict purchase behavior with a model

Probability(purchase) =

β_1 Electric Heat +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

β_4 Response Rate +

β_5 Type Of Message

We can **model purchase behavior** at the consumer level.

Include predictors that indicate heavy winter electric usage, neighbor influences, and responsiveness to past communications.

Housing heat type correlates with water heat type

Probability(purchase) =

β_1 **Electric Heat** +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

β_4 Response Rate +

β_5 Type Of Message

Does the consumer use electric heat?

Households with gas heat are unlikely to purchase an electric water heater. (Natural gas is cheap.)

Willingness to invest in efficient products

Probability(purchase) =

β_1 Electric Heat +

β_2 **Similar Purchases** +

β_3 Neighbors Purchased +

β_4 Response Rate +

β_5 Type Of Message

Has the consumer participated in similar program promotions?

Past purchase behavior is a good predictor of future behavior.

Neighbor effects can be powerful

Probability(purchase) =

β_1 Electric Heat +

β_2 Similar Purchases +

β_3 **Neighbors Purchased** +

β_4 Response Rate +

β_5 Type Of Message

Is the product popular about their neighbors?

Neighbor effects may influence purchase behavior.

Responsiveness proxies engagement

Probability(purchase) =

β_1 Electric Heat +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

β_4 **Response Rate** +

β_5 Type Of Message

Has the consumer responded to past communications?

Past responsiveness indicates high engagement.

Home Energy Reports influence usage perceptions

Probability(purchase) =

β_1 Electric Heat +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

β_4 Response Rate +

β_5 **Type Of Message**

What type of message has the consumer received on their Home Energy Reports?

The relative positioning of past energy usage may influence willingness to invest in future lower usage.

We have a model. Let’s get the data.

Probability(purchase) =

β_1 Electric Heat +

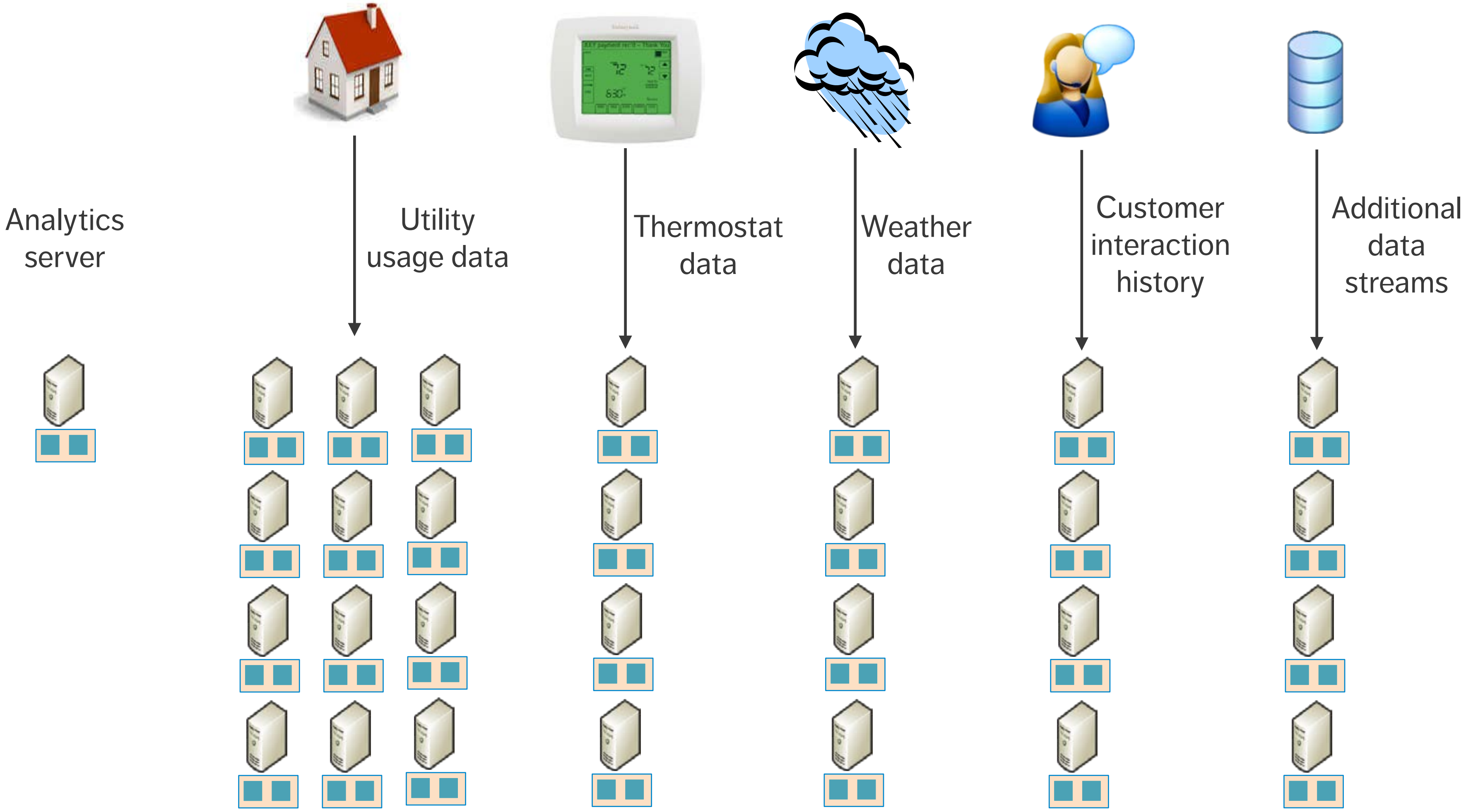
β_2 Similar Purchases +

β_3 Neighbors Purchased +

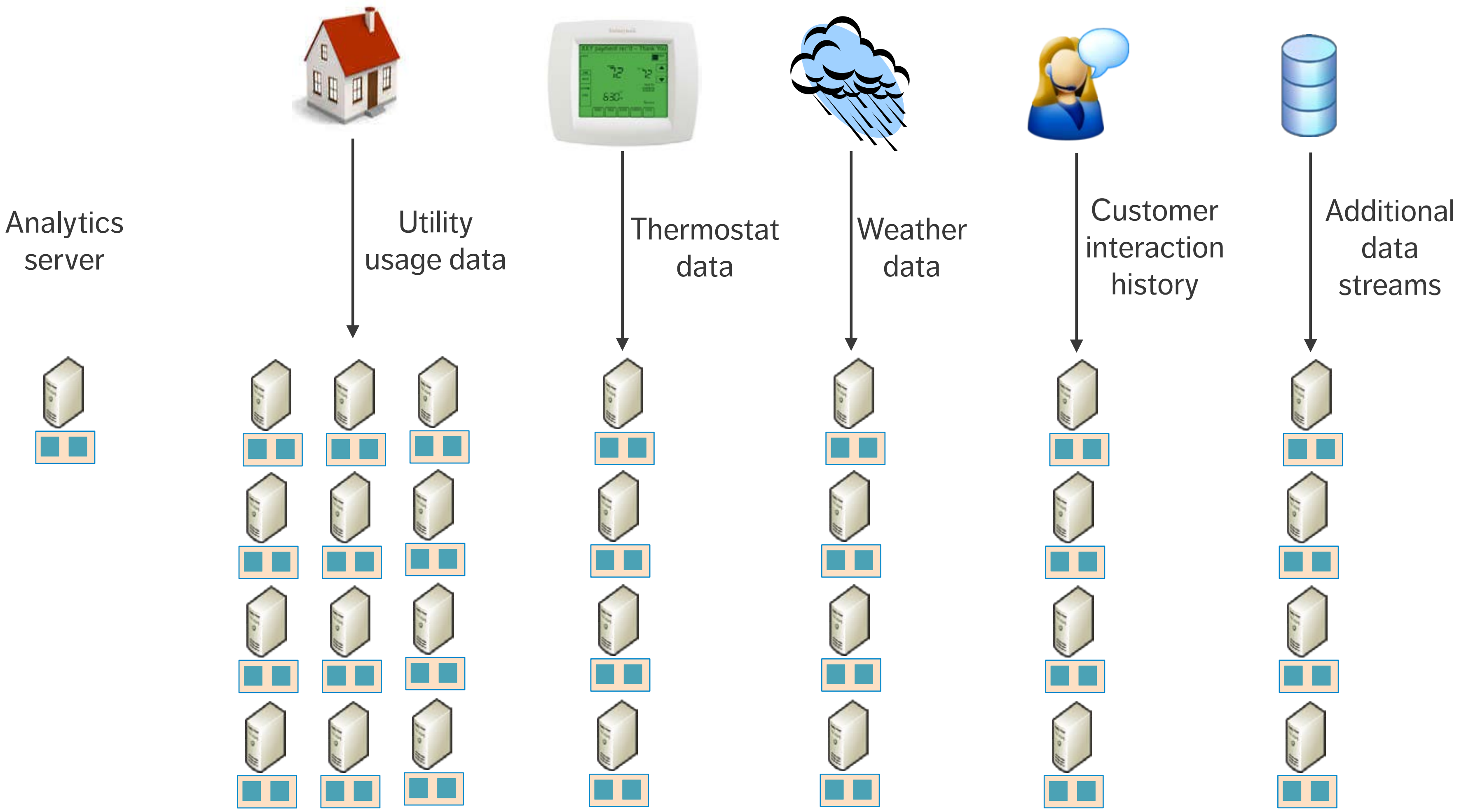
β_4 Response Rate +

β_5 Type Of Message

Disparate data sources



Let's start plumbing



Analytics server

Utility usage data

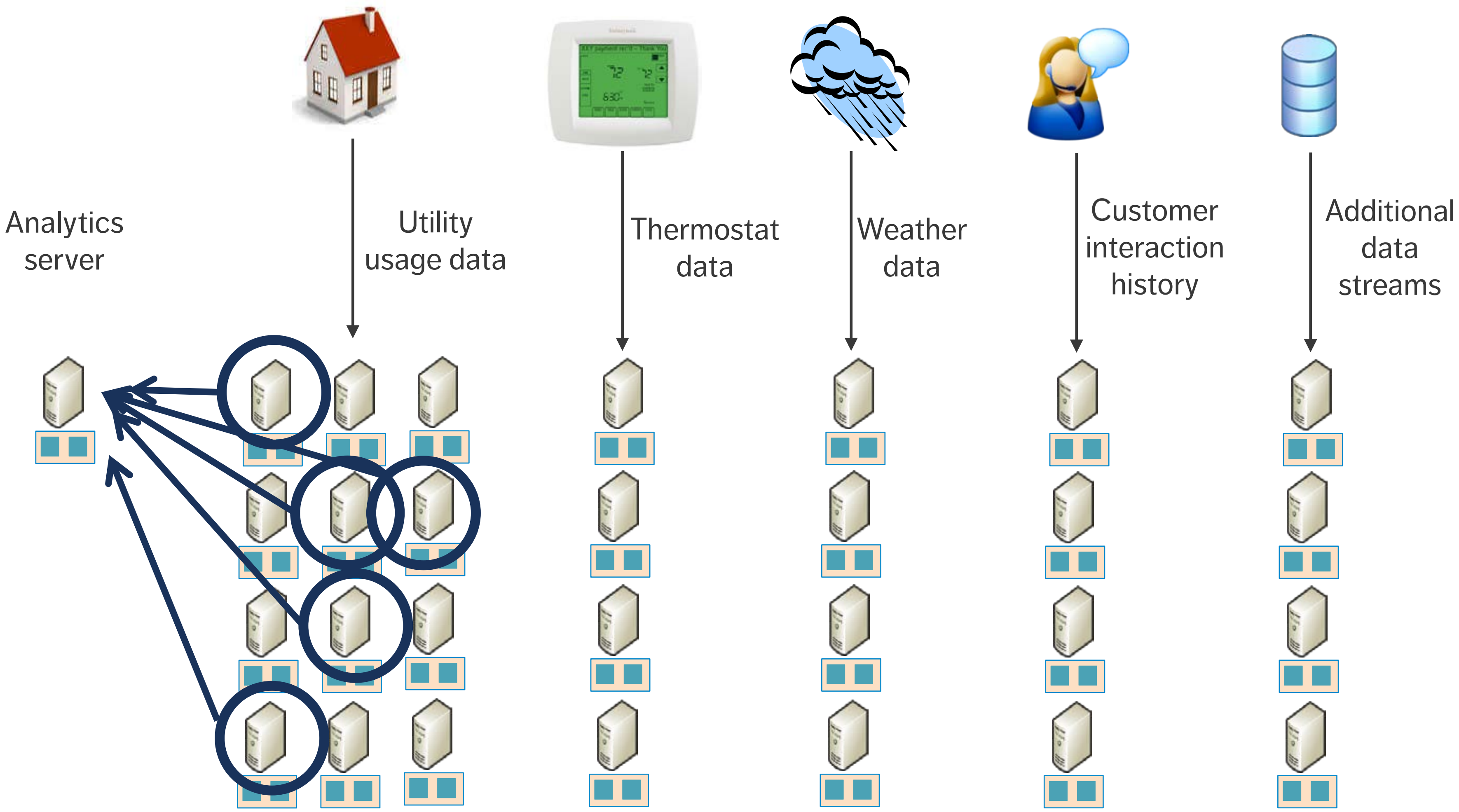
Thermostat data

Weather data

Customer interaction history

Additional data streams

Pipe utility data



Analytics server

Utility usage data

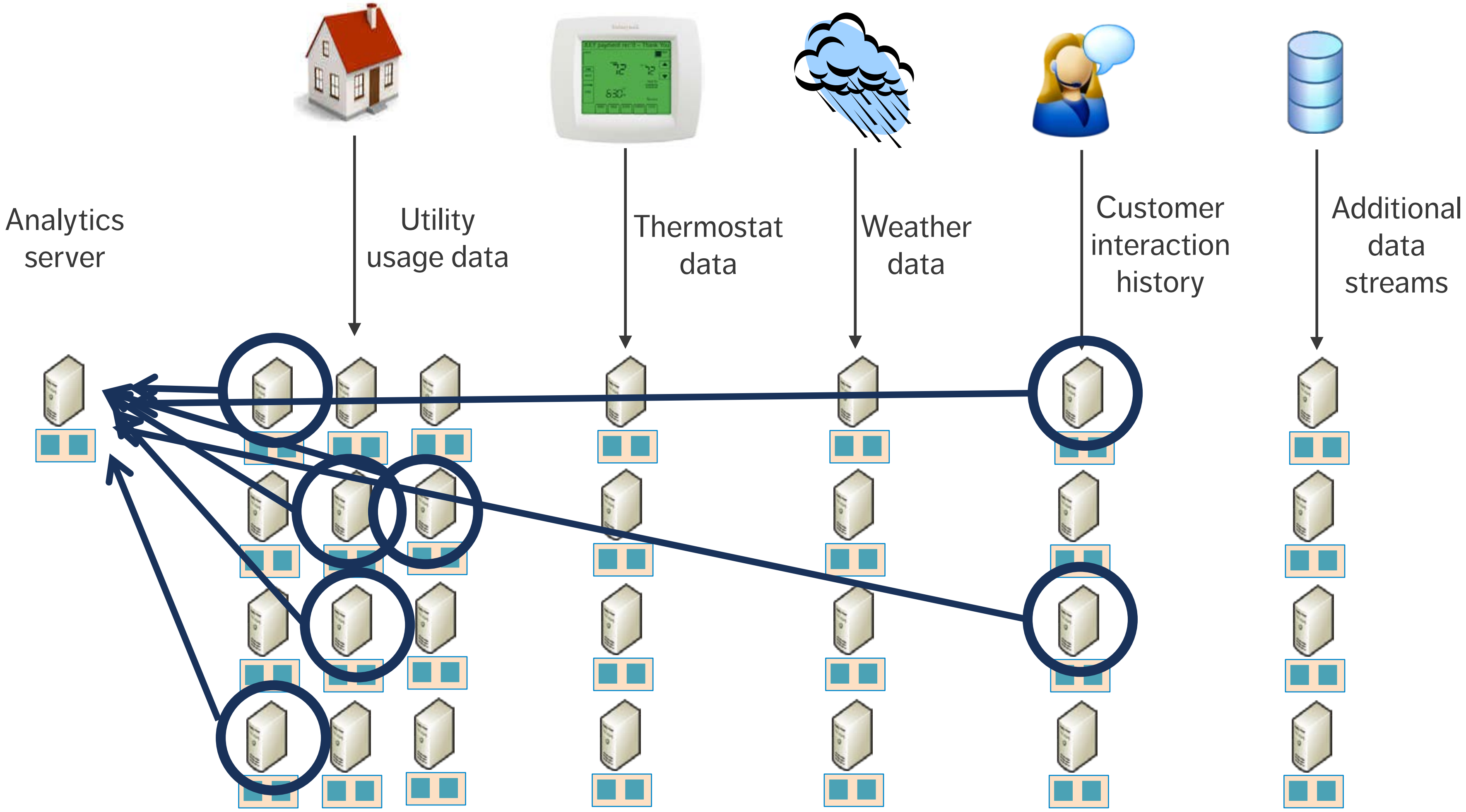
Thermostat data

Weather data

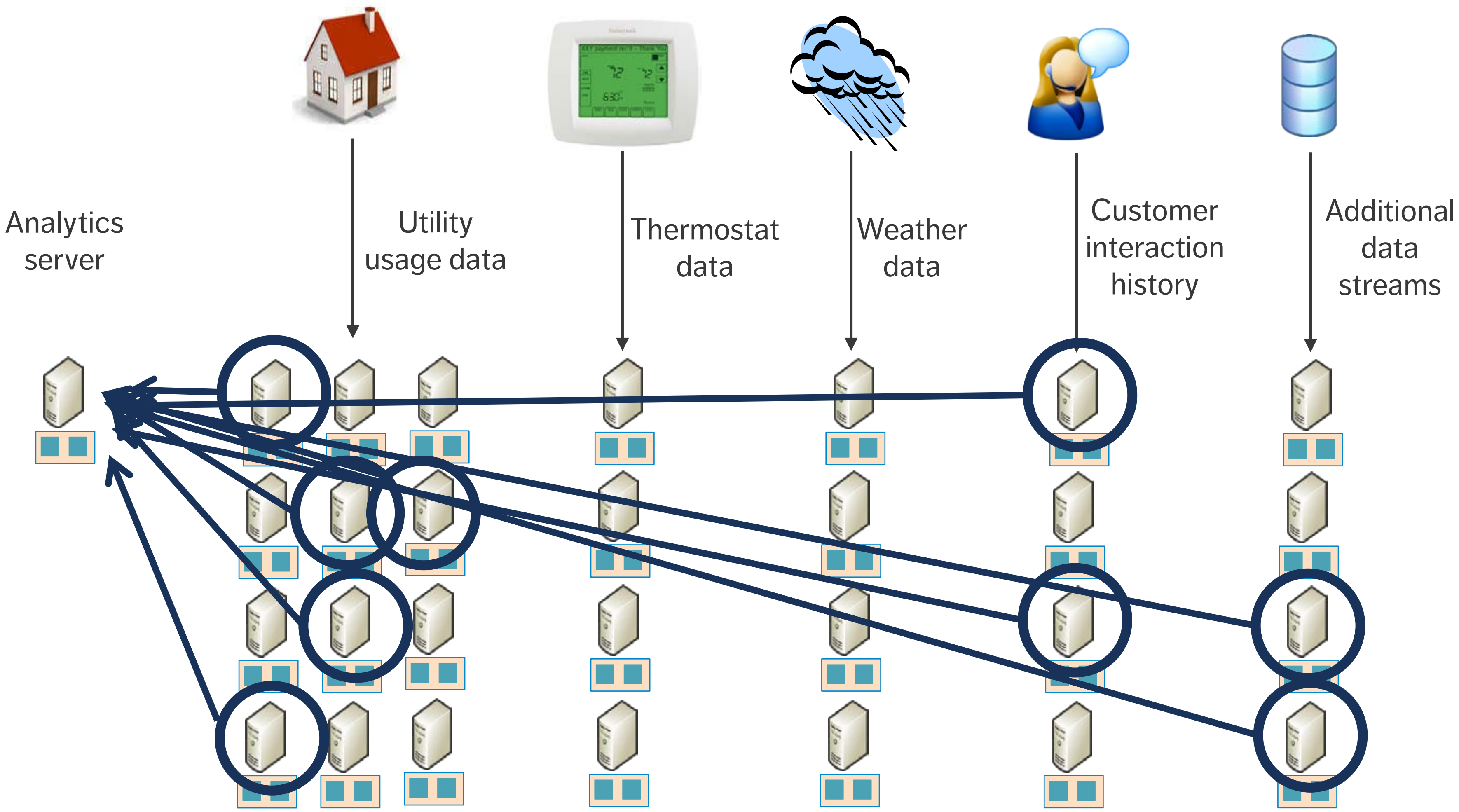
Customer interaction history

Additional data streams

Pipe customer interaction data



Finally, pipe Home Energy Report data



Now we're ready to model

Probability(purchase) =

β_1 Electric Heat +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

β_4 Response Rate +

β_5 Type Of Message

There's a problem

Probability(purchase) =

β_1 Electric Heat +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

β_4 Response Rate +

β_5 Type Of Message

We know these predictors

Heat type is sparse and inaccurate

Probability(purchase) =

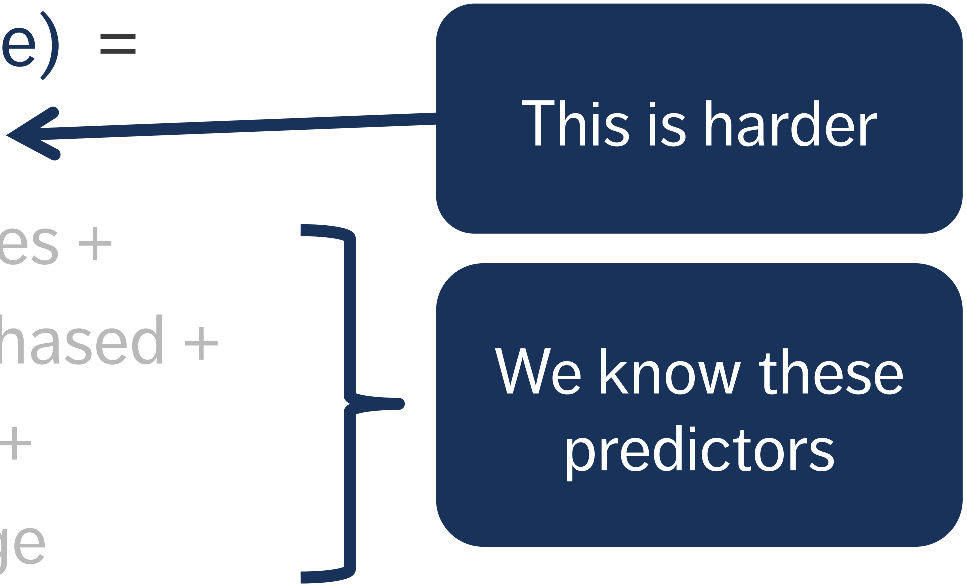
β_1 **Electric Heat** +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

β_4 Response Rate +

β_5 Type Of Message



This is harder

We know these predictors

Model electric heat to compensate for bad data

Probability(purchase) =

β_1 **Electric Heat** +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

β_4 Response Rate +

β_5 Type Of Message

Parcel data coverage of heat type is **sparse** and **inaccurate**.

We need another data source for heat type.

We construct a model to predict heat type

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

δ_1 Weather Sensitivity +

δ_2 Neighbors Heat +

δ_3 Natural Gas Price

We can **model** the **presence of electric heat**.

Include predictors of weather sensitivity, area prevalence, and local natural gas price.

Sensitivity of electricity usage to cold weather

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

δ_1 **Weather Sensitivity** +

δ_2 Neighbors Heat +

δ_3 Natural Gas Price

How sensitive is the consumer's electricity usage to cold weather?

High sensitivity to cold weather is our best indicator of electric heat.

Heat Type Is Related to Geography

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

δ_1 Weather Sensitivity +

δ_2 **Neighbors Heat** +

δ_3 Natural Gas Price

Is electric heat popular in the consumer's area?

Heat type tends to have specific geographic distributions.

Gas Prices May Affect Heat Type Adoption

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

δ_1 Weather Sensitivity +

δ_2 Neighbors Heat +

δ_3 **Natural Gas Price**

How expensive is the alternative?

Natural gas may be hard to get in certain areas.

We have another model. Let’s get the data.

Probability(purchase) =

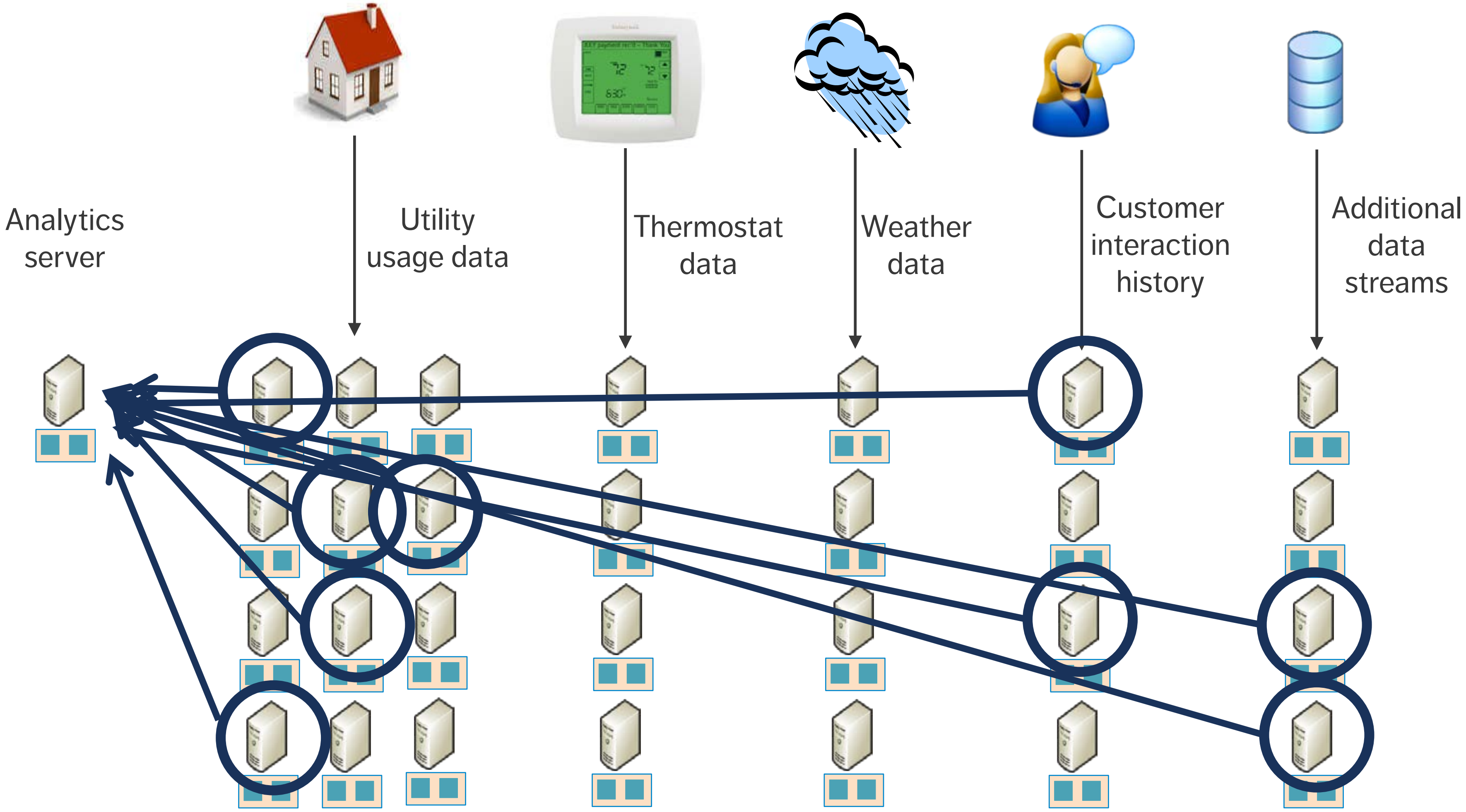
β_1 Pr(**Electric Heat**) =

δ_1 Weather Sensitivity +

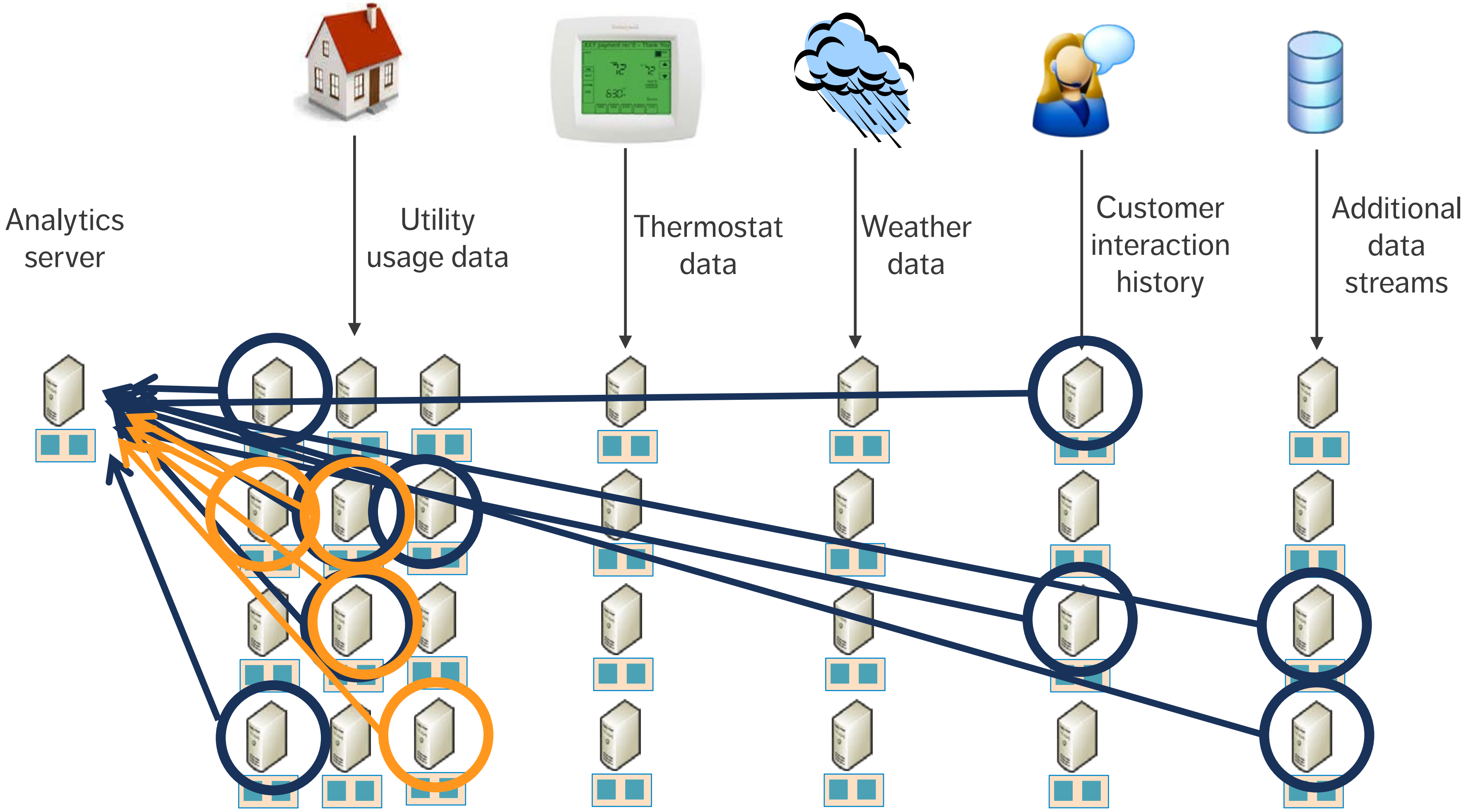
δ_2 Neighbors Heat +

δ_3 Natural Gas Price

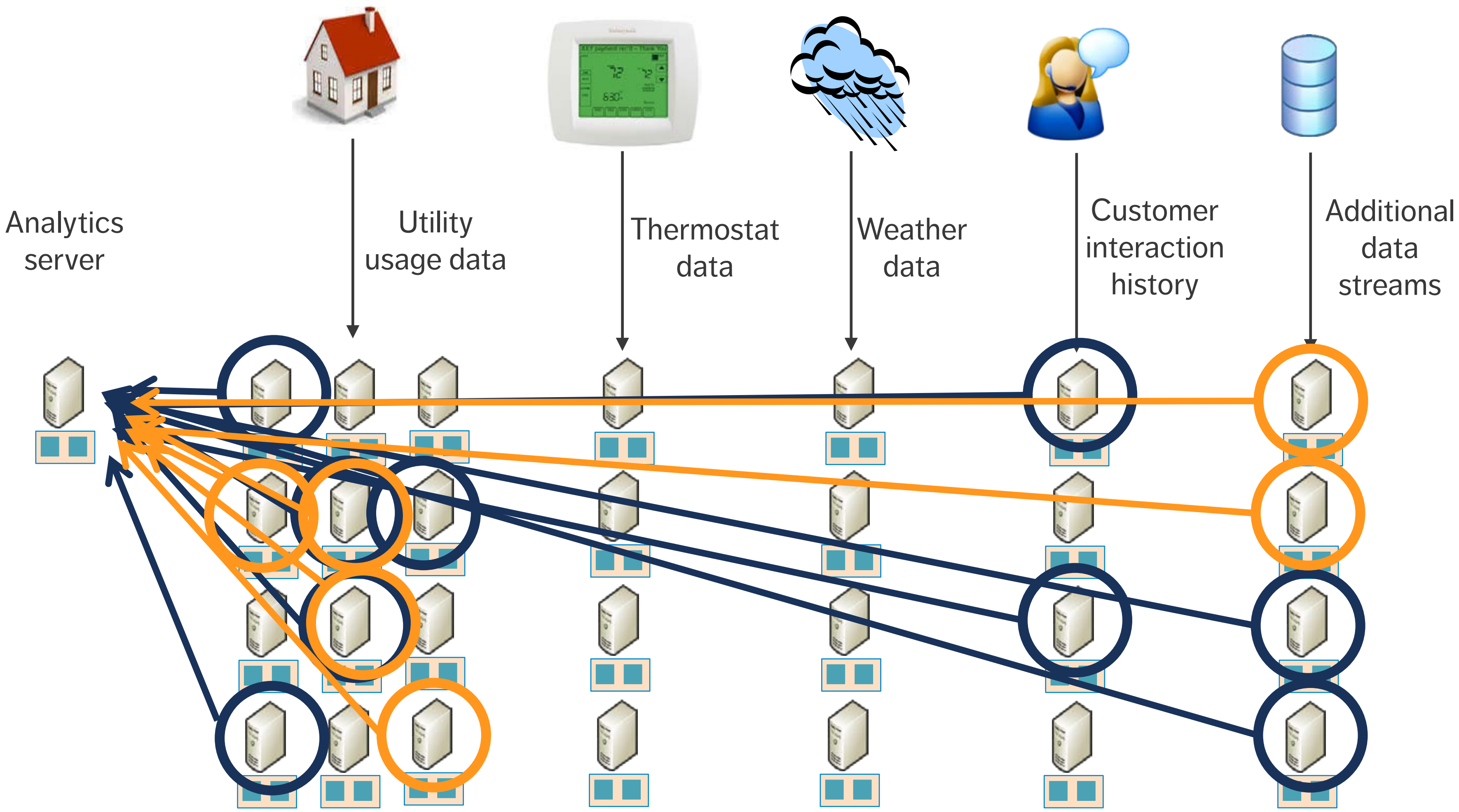
Our plumbing so far



Pipe neighbor heat type



Pipe natural gas prices



Now we're ready to model (x2)

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

δ_1 Weather Sensitivity +

δ_2 Neighbors Heat +

δ_3 Natural Gas Price

There’s a problem (x2)

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

δ_1 Weather Sensitivity +

δ_2 Neighbors Heat +

δ_3 Natural Gas Price

We know these predictors

We don’t know weather sensitivity

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

δ_1 **Weather Sensitivity** +

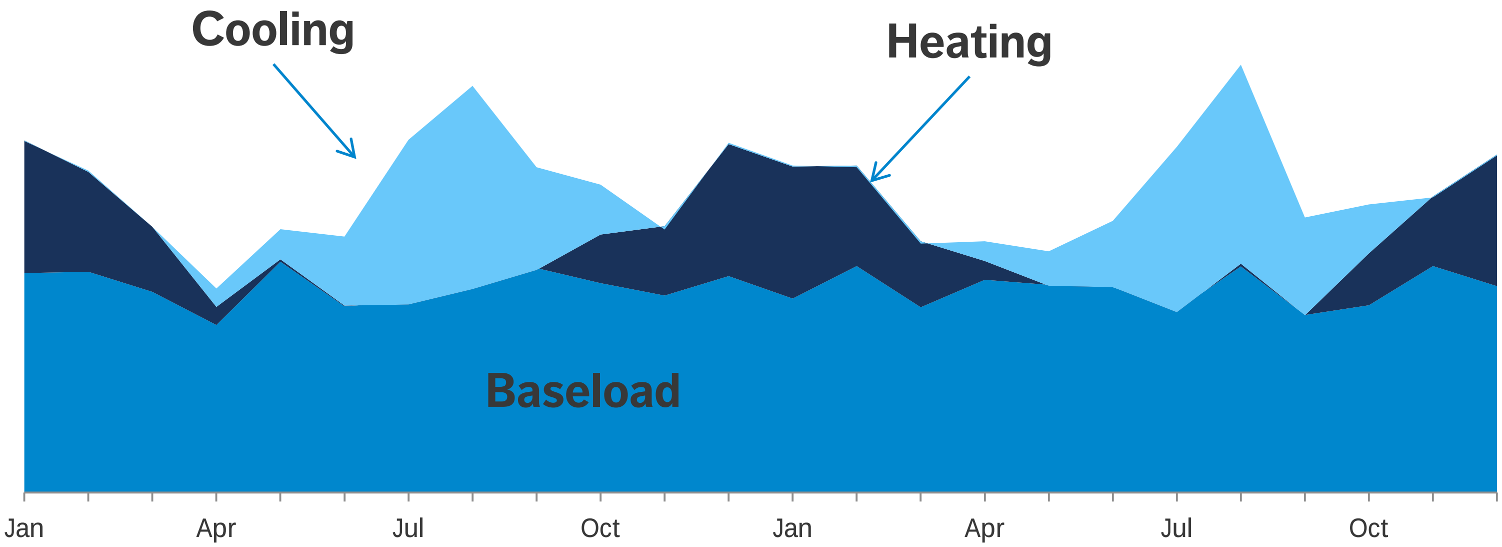
δ_2 Neighbors Heat +

δ_3 Natural Gas Price

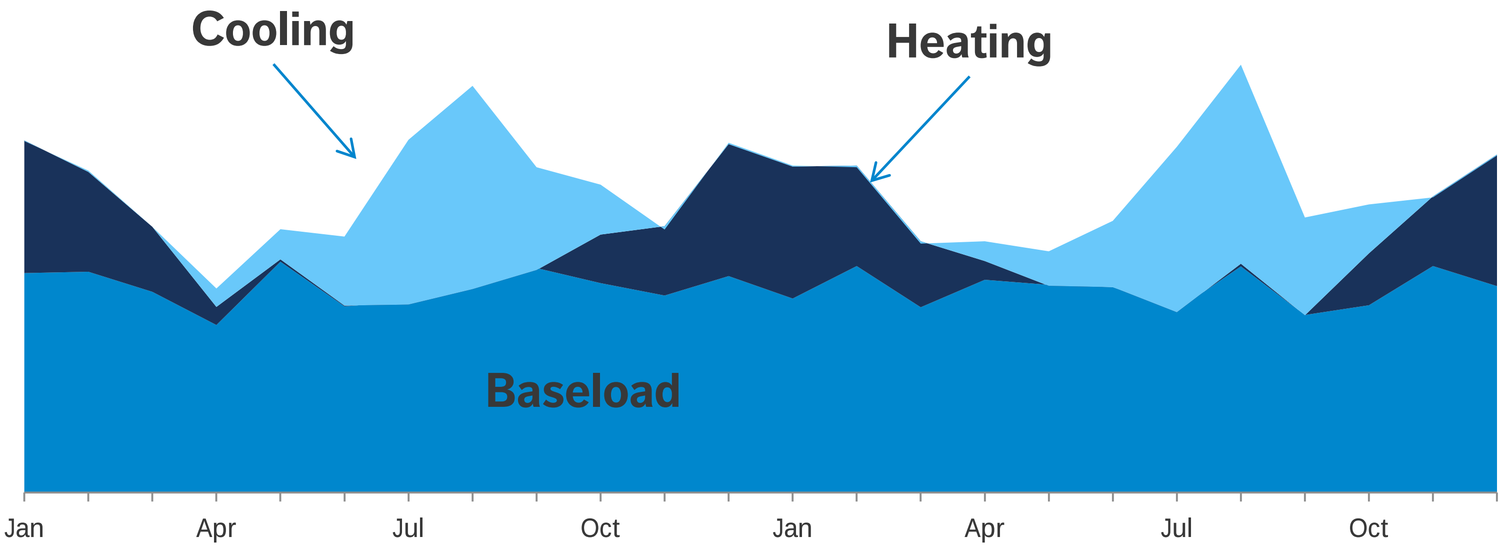
This is harder

We know these predictors

Luckily, we know how to do this



We have a disaggregation algorithm. Let's get the data.

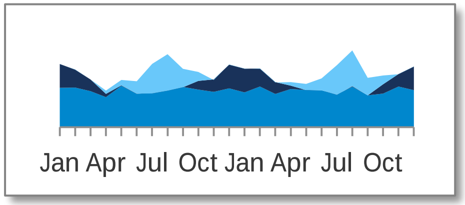


Disaggregate heating and cooling

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

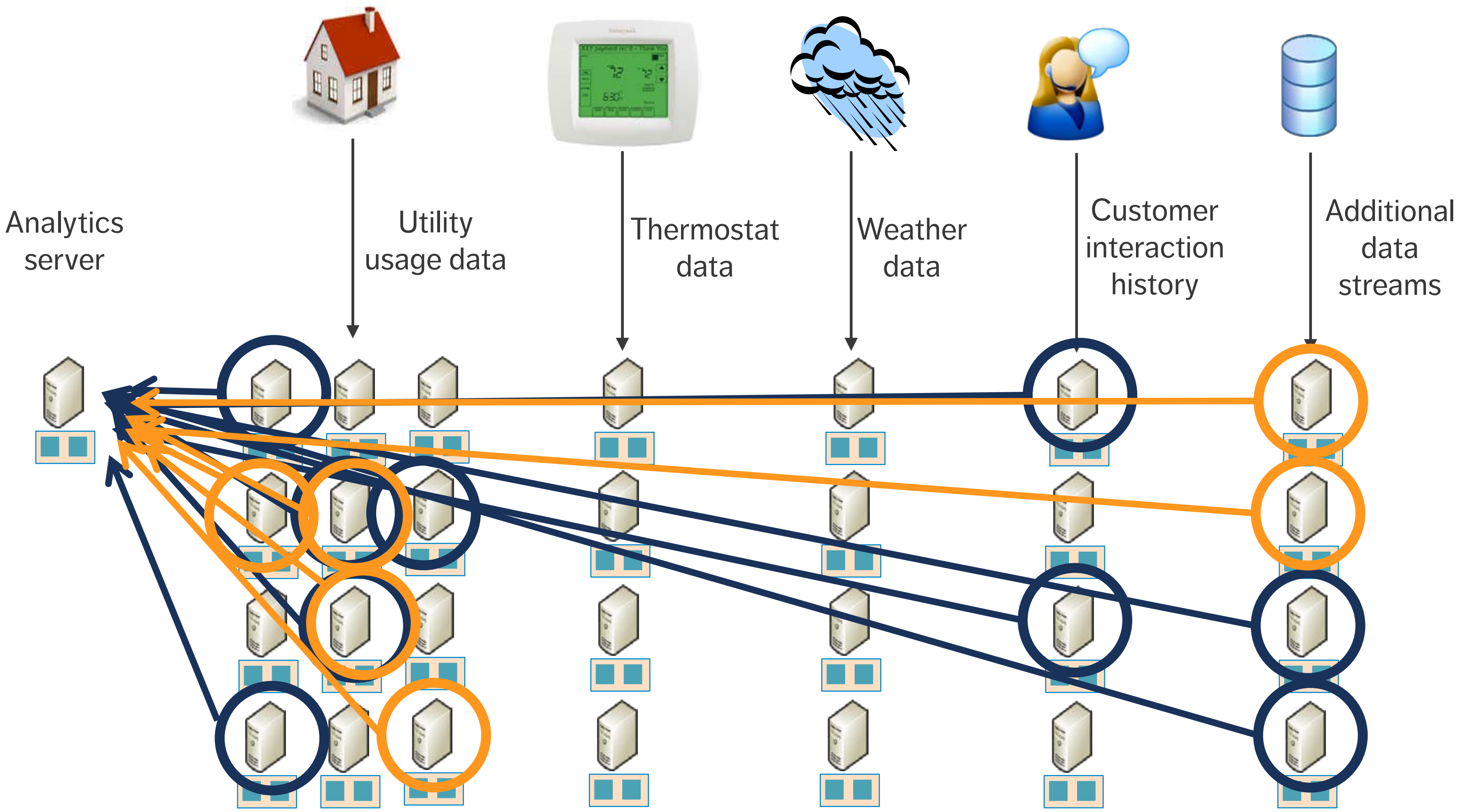
δ_1 **Weather Sensitivity** =



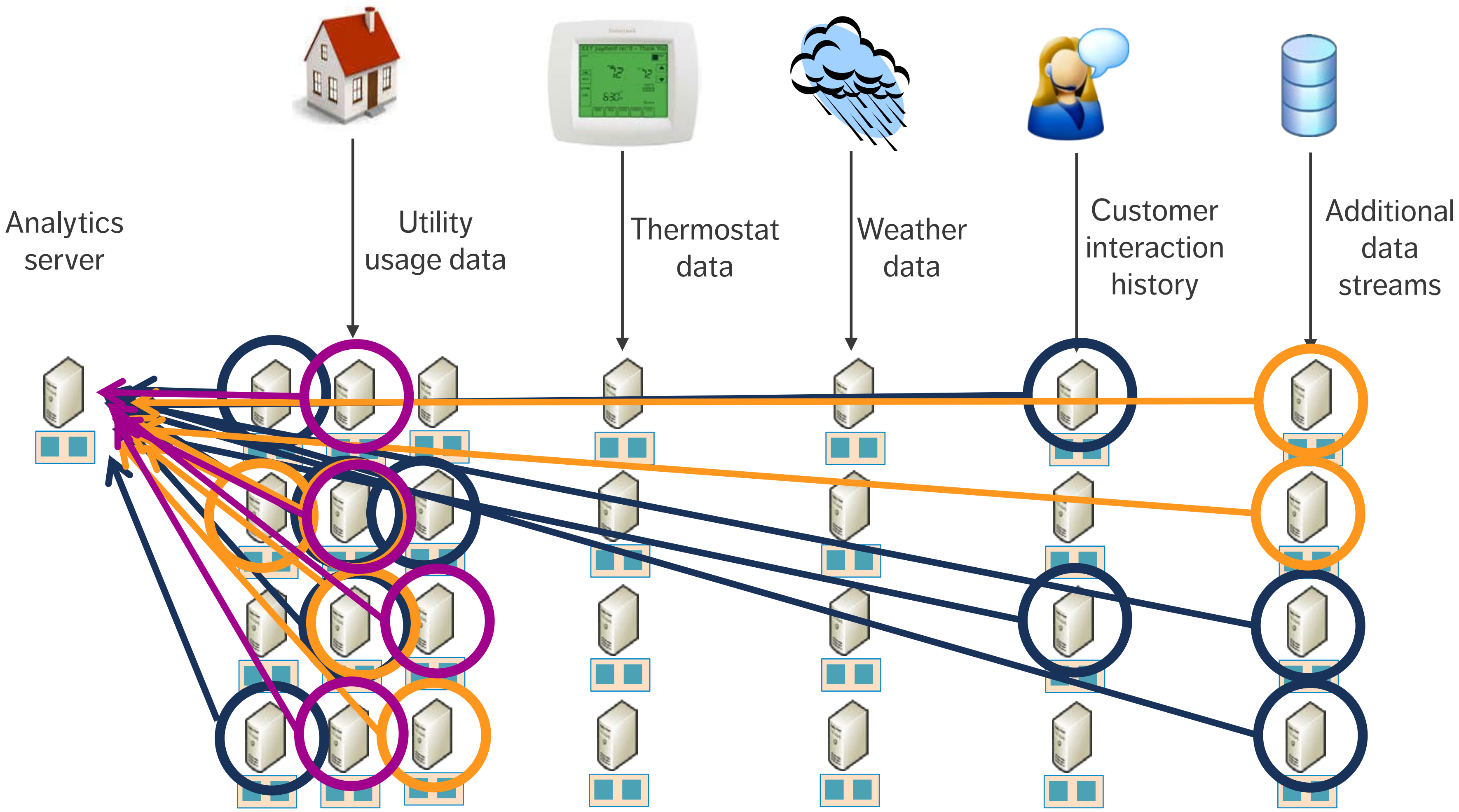
Correlate electricity usage with weather.

Let’s grab the data.

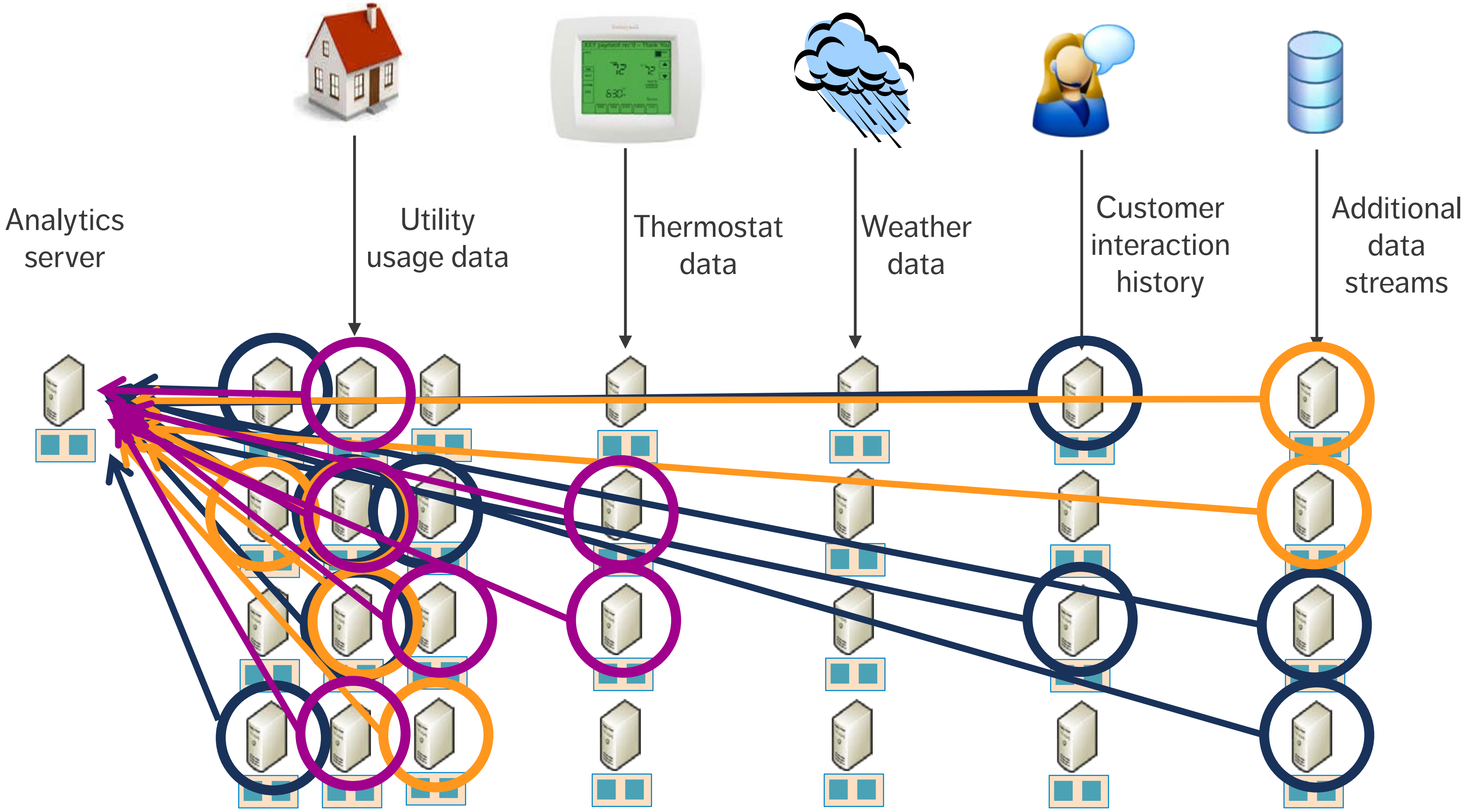
Our plumbing so far (x2)



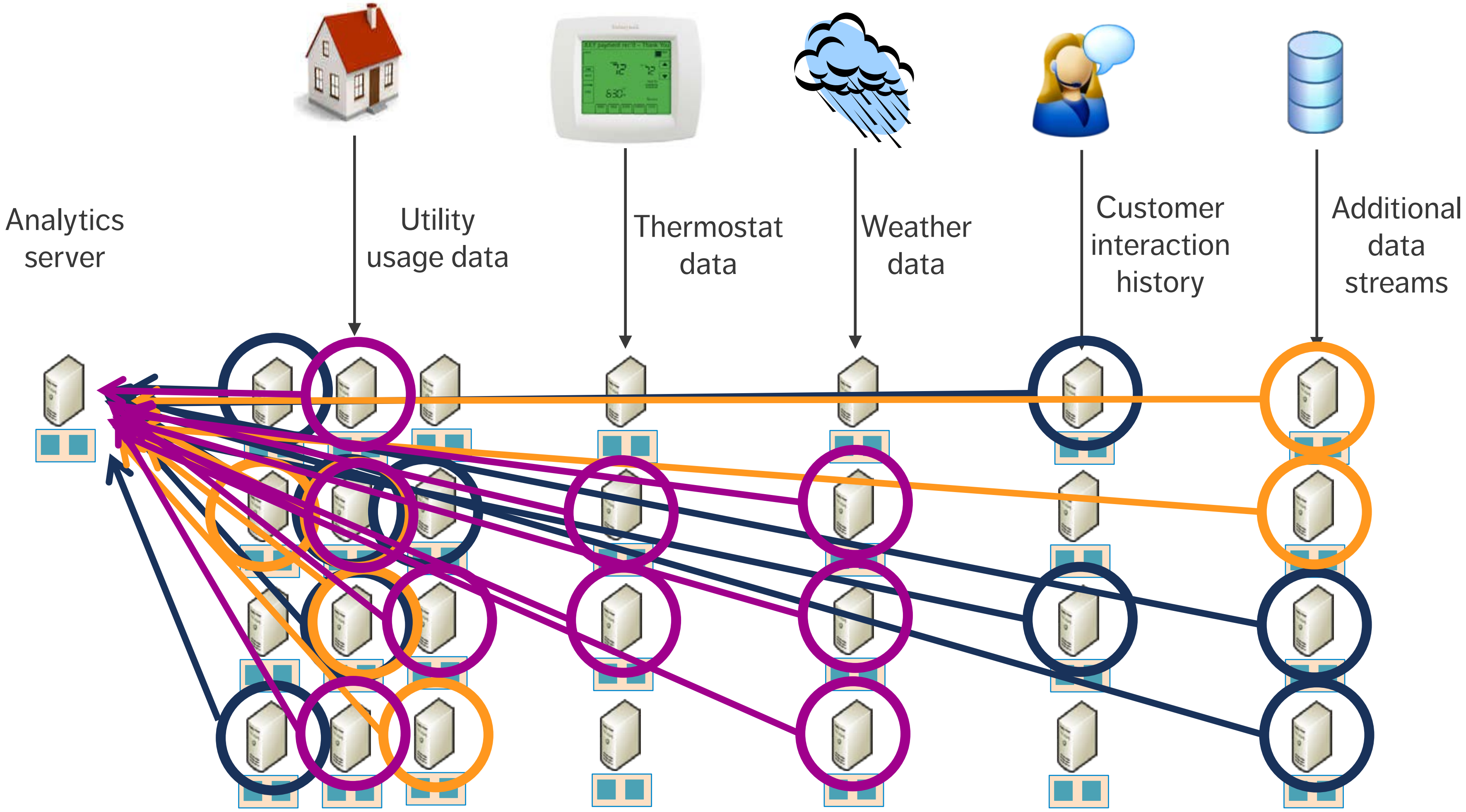
Pipe electricity usage data



Pipe thermostat data



Pipe weather data



Analytics server

Utility usage data

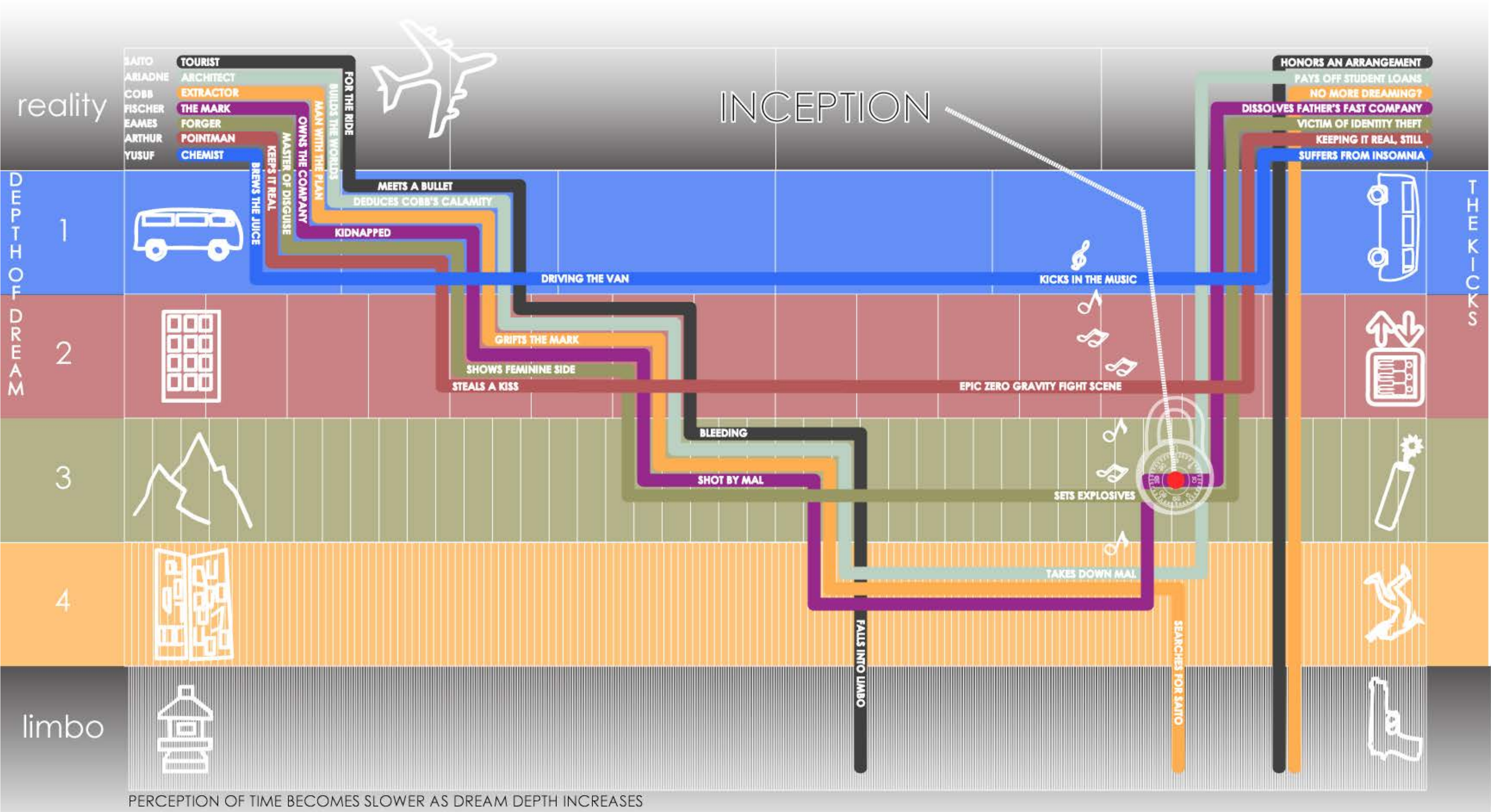
Thermostat data

Weather data

Customer interaction history

Additional data streams

Starting to feel like Inception



2010 INFOGRAPHIC
BY DANIEL WANG

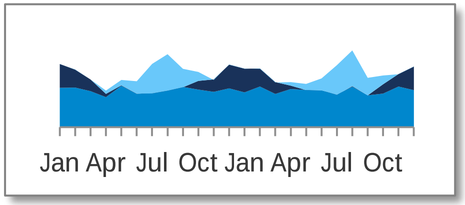
http://www.chartgeek.com/wp-content/uploads/2012/04/inception-explained-chart.jpg

Now we're ready to model (finally)

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

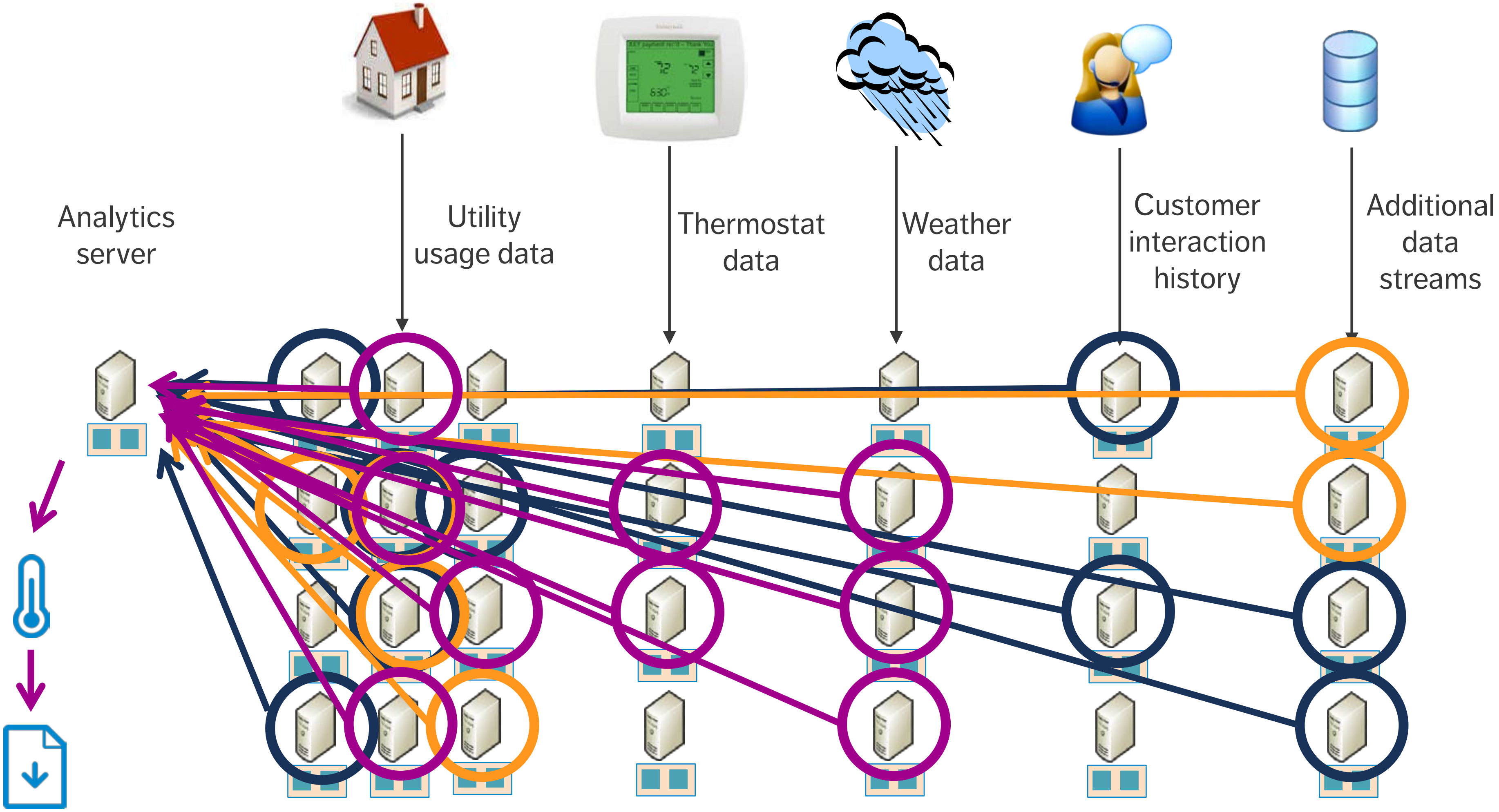
δ_1 **Weather Sensitivity** =



Construct **disaggregation** algorithms.

Calculate sensitivity for all households.

Disaggregate and store results

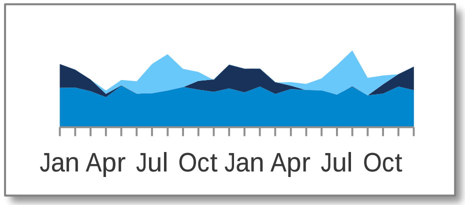


We know each customer’s heating sensitivity

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

✓ **Weather Sensitivity** =



Let’s continue with our electric heat model.

We have the data to finish our heat type model

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

δ_1 Weather Sensitivity +

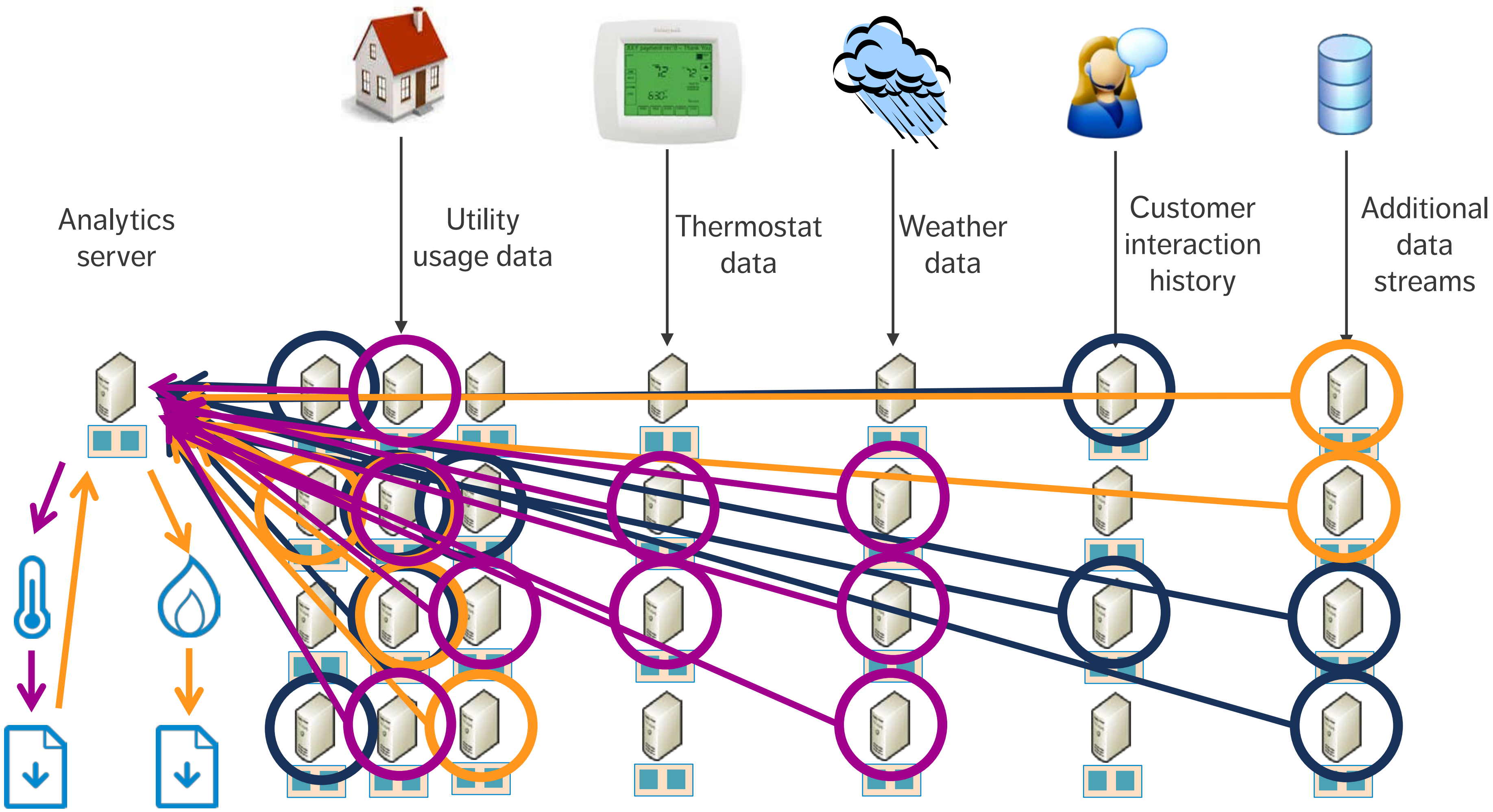
δ_2 Neighbors Heat +

δ_3 Natural Gas Price

Construct **electric heat model**.

Impute heat type for all households.

Impute heat type and store results



We know each customer’s heat type

Probability(purchase) =

✓ Pr(**Electric Heat**) =

δ_1 Weather Sensitivity +

δ_2 Neighbors Heat +

δ_3 Natural Gas Price

Let’s continue with our water heater purchase model.

We now have the data to finish our purchase model

Probability(purchase) =

β_1 Electric Heat +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

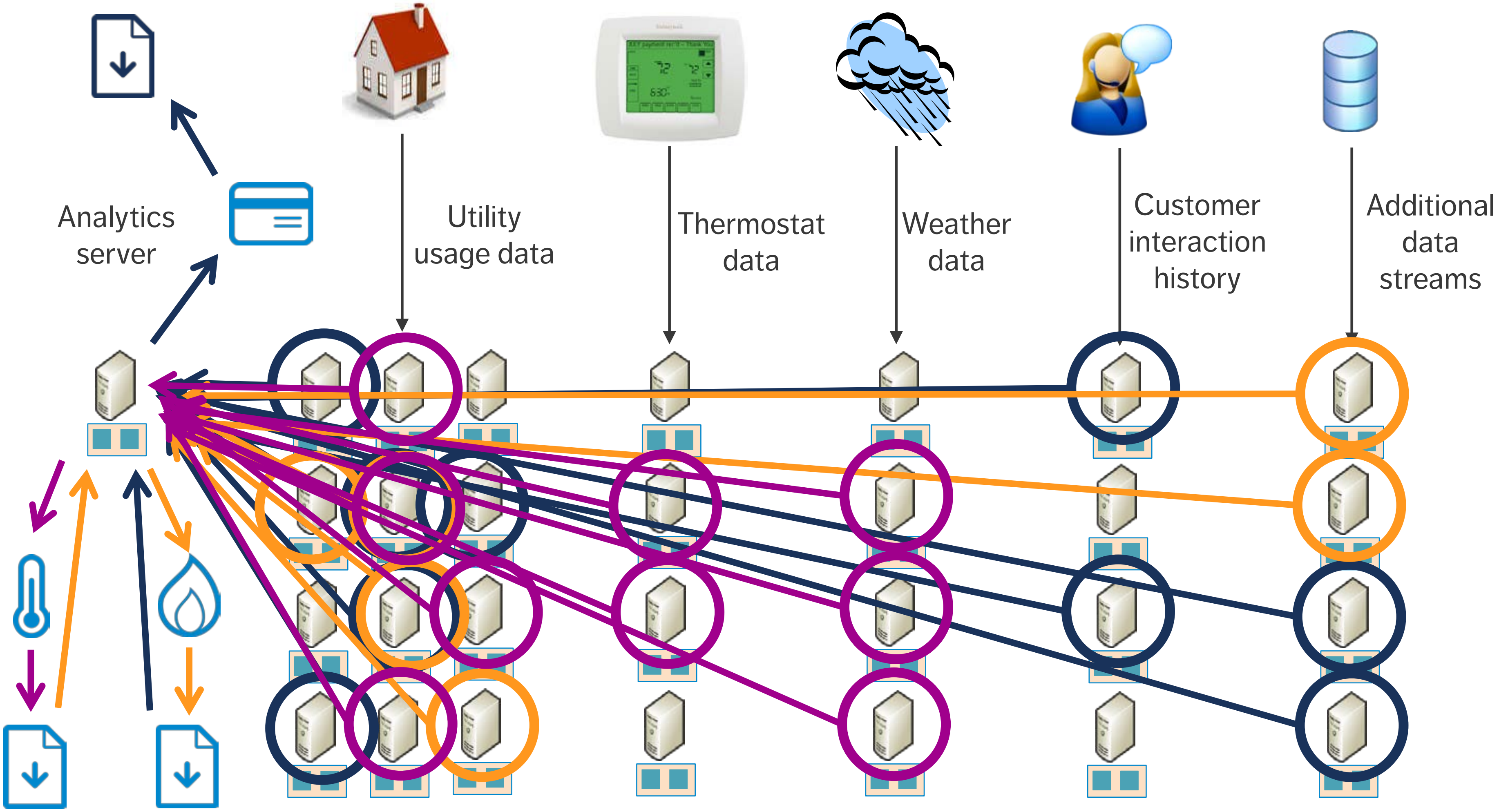
β_4 Response Rate +

β_5 Type Of Message

Construct **purchase behavior model**.

Calculate **likelihood to purchase** for all households.

Calculate likelihood to purchase and store results



Analytics server

Utility usage data

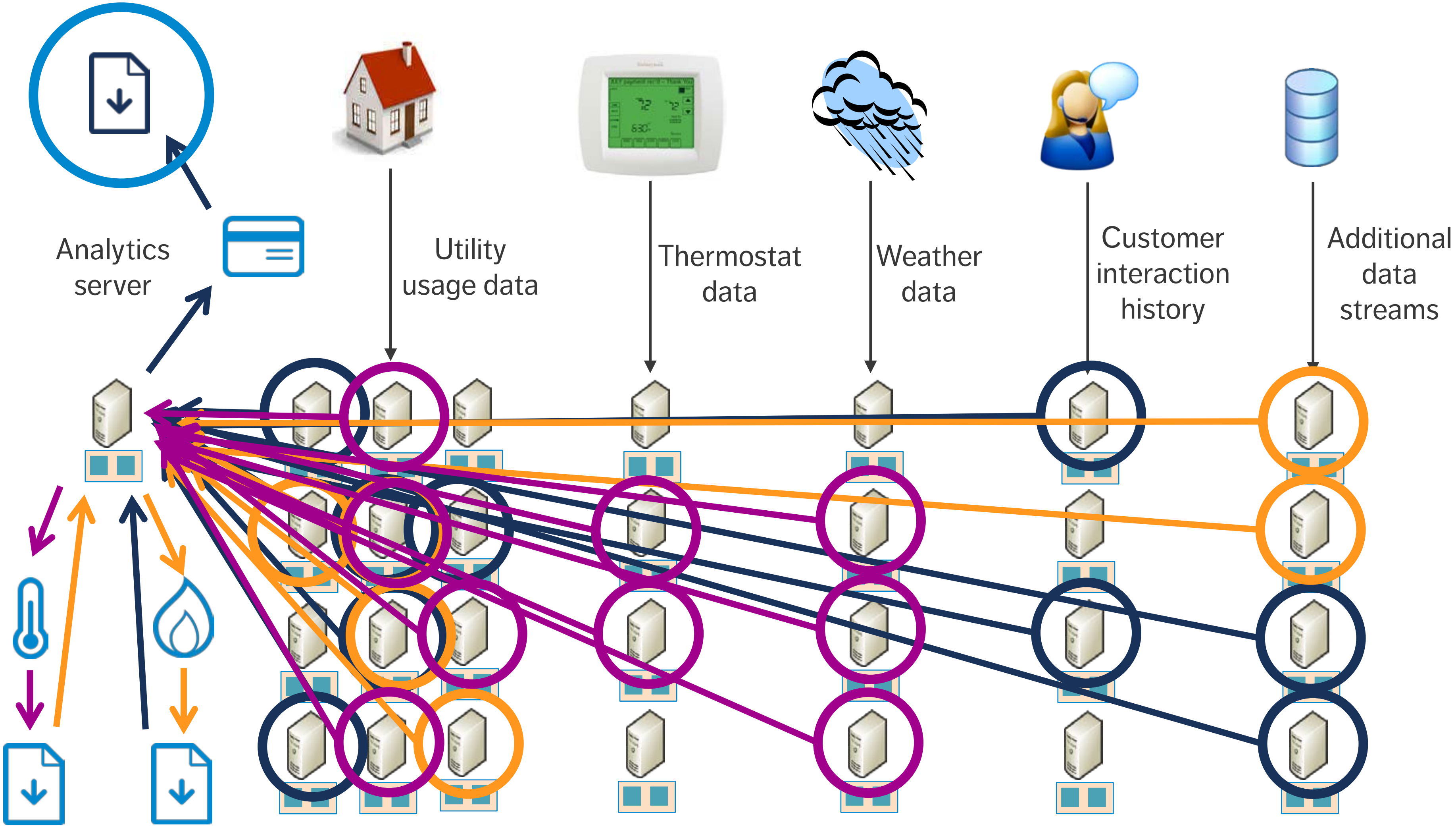
Thermostat data

Weather data

Customer interaction history

Additional data streams

We have our desired result



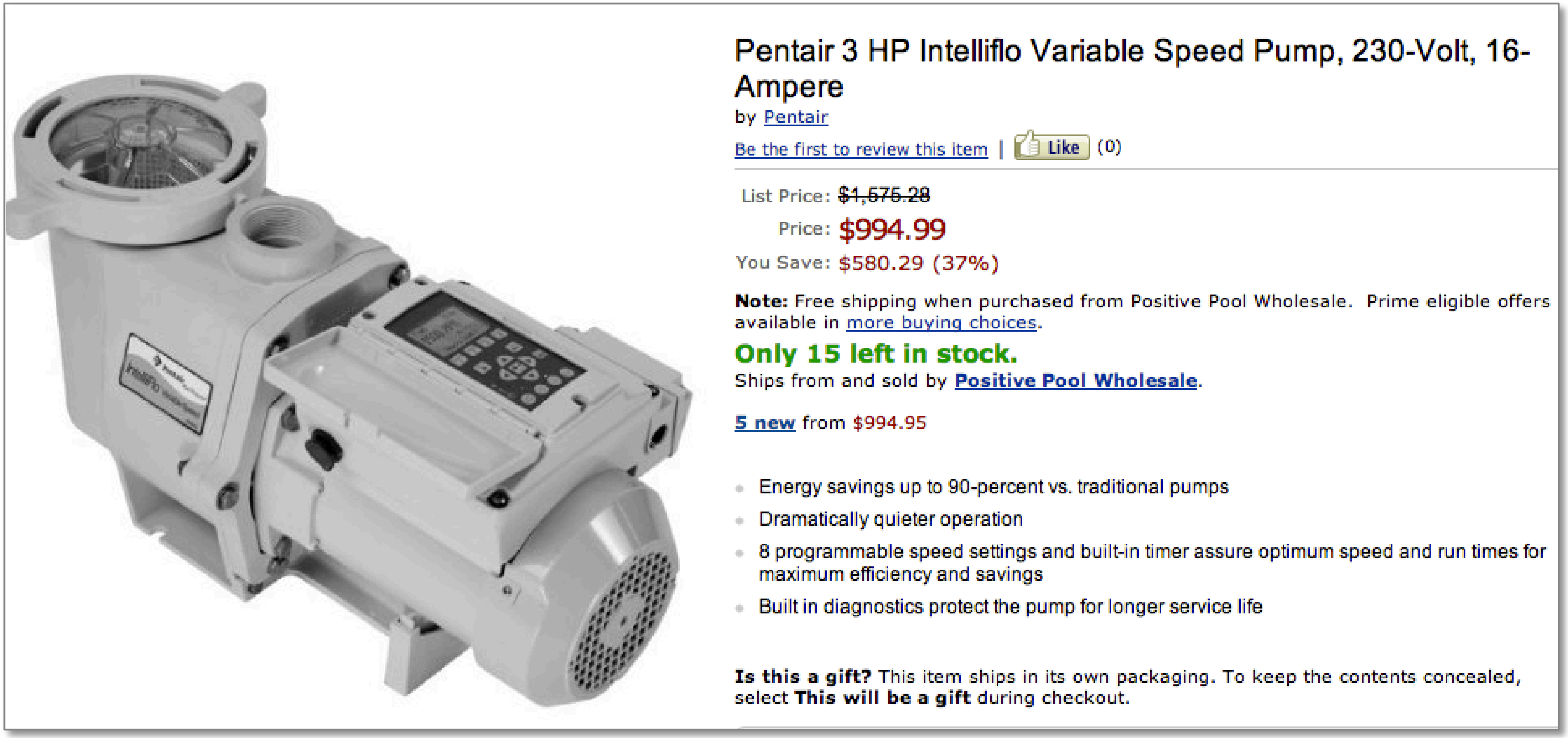
Data science is plumbing

Analytics server

Additional data streams

http://www.ontimeplumber.com.au/plumbing_disasters/plumbing_disasters.html

New request: Who would buy an efficient pool pump for 10% off?



Pentair 3 HP Intelliflo Variable Speed Pump, 230-Volt, 16-Ampere

by Pentair

Be the first to review this item | Like (0)

List Price: ~~$1,575.28~~

Price: $994.99

You Save: $580.29 (37%)

Note: Free shipping when purchased from Positive Pool Wholesale. Prime eligible offers available in more buying choices.

Only 15 left in stock.

Ships from and sold by **Positive Pool Wholesale**.

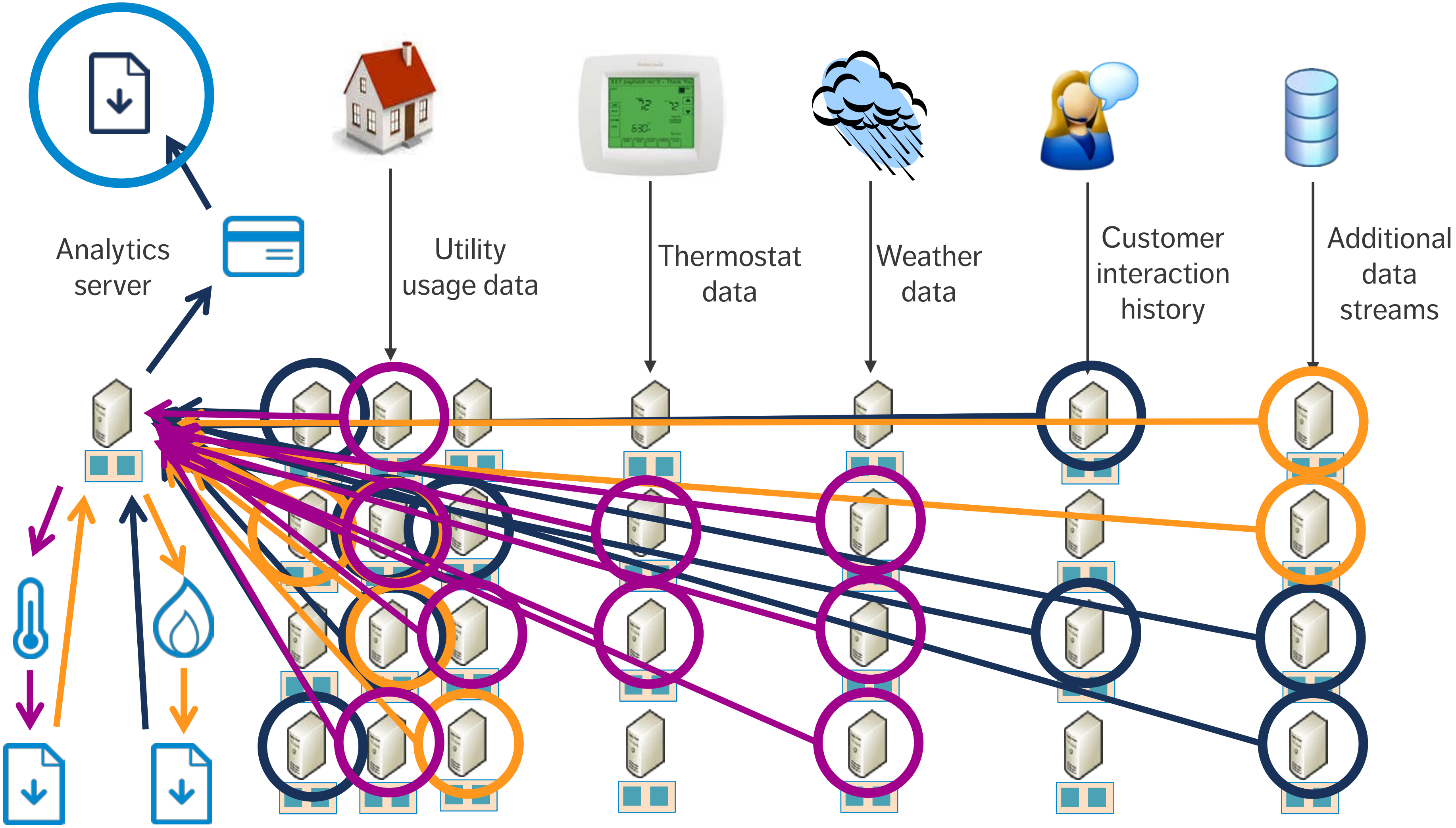
5 new from $994.95

- Energy savings up to 90-percent vs. traditional pumps
- Dramatically quieter operation
- 8 programmable speed settings and built-in timer assure optimum speed and run times for maximum efficiency and savings
- Built in diagnostics protect the pump for longer service life

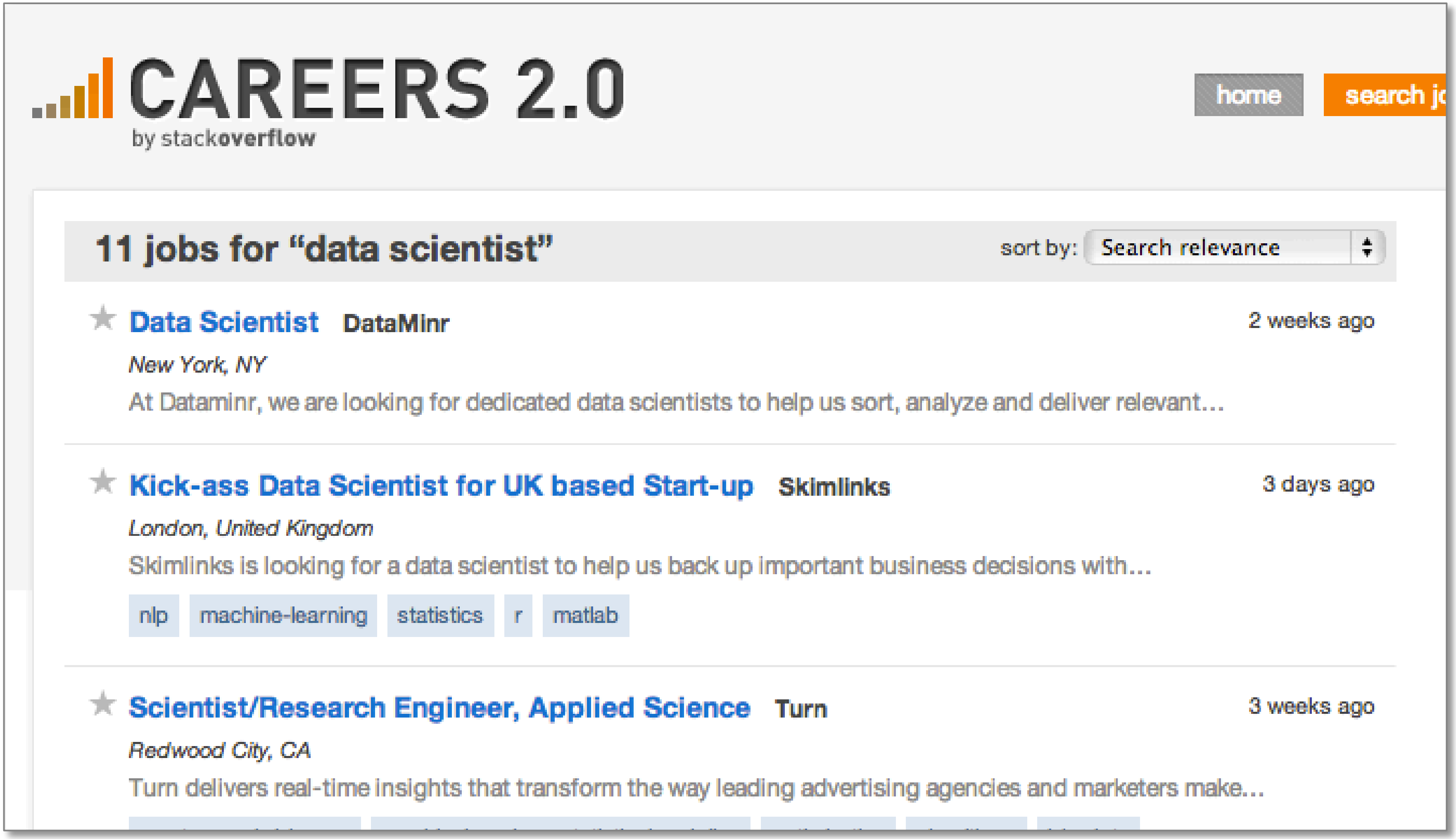
Is this a gift? This item ships in its own packaging. To keep the contents concealed, select **This will be a gift** during checkout.

http://www.amazon.com/Pentair-Intelliflo-Variable-230-Volt-16-Ampere/dp/B007E4VWNO/ref=sr_1_3?ie=UTF8&qid=1350601695&sr=8-3&keywords=variable+speed+pool+pump

I remember what the last model took…



… and I start searching the want-ads



CAREERS 2.0
by stackoverflow

11 jobs for "data scientist"

sort by: Search relevance

☆ **Data Scientist** DataMinr — 2 weeks ago

New York, NY

At Dataminr, we are looking for dedicated data scientists to help us sort, analyze and deliver relevant…

☆ **Kick-ass Data Scientist for UK based Start-up** Skimlinks — 3 days ago

London, United Kingdom

Skimlinks is looking for a data scientist to help us back up important business decisions with…

nlp machine-learning statistics r matlab

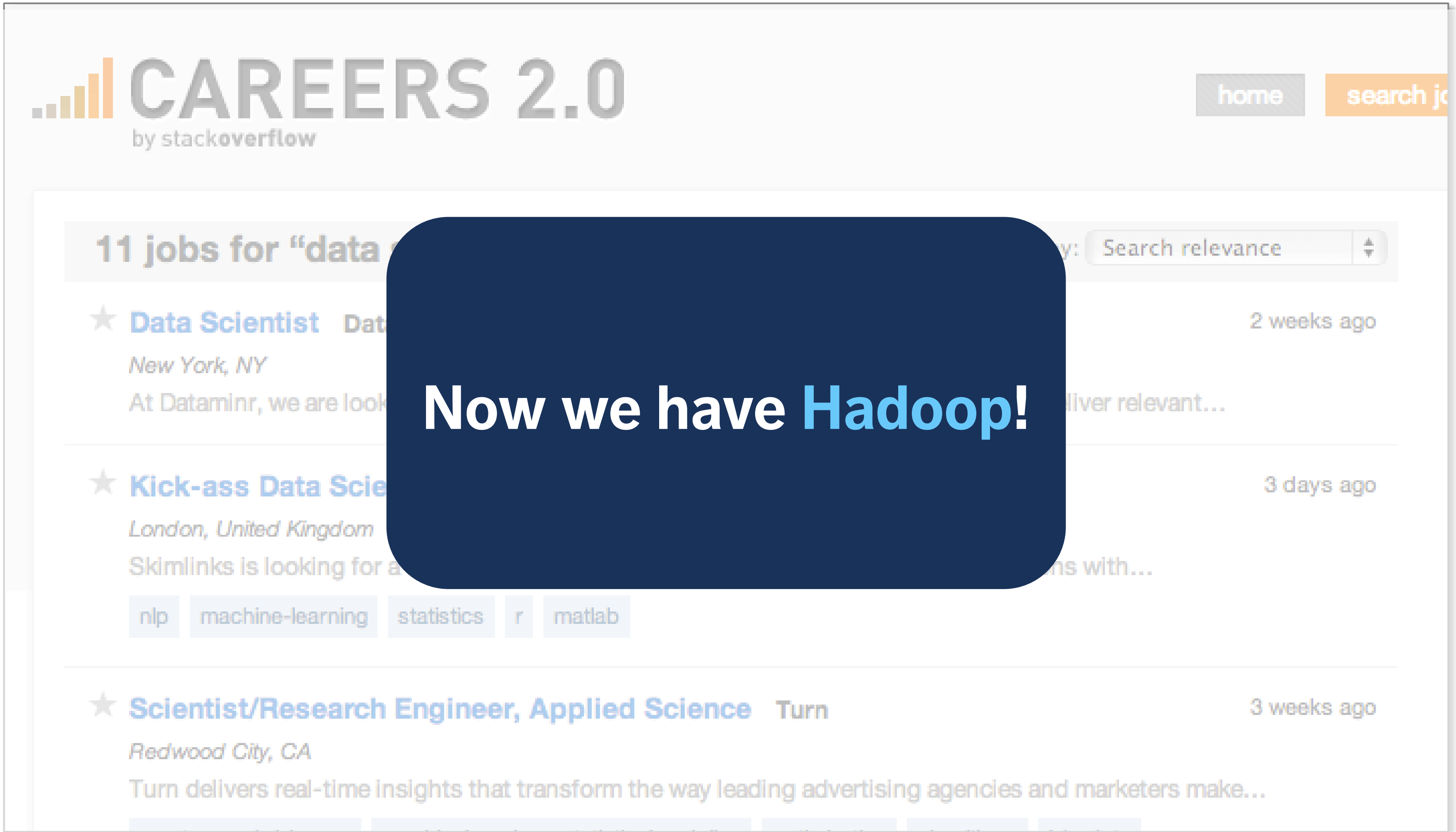
☆ **Scientist/Research Engineer, Applied Science** Turn — 3 weeks ago

Redwood City, CA

Turn delivers real-time insights that transform the way leading advertising agencies and marketers make…

http://careers.stackoverflow.com/jobs?searchTerm=data+scientist&location=

But it gets better



11 jobs for "data

Search relevance

★ Data Scientist Dat

New York, NY

At Dataminr, we are look

liver relevant...

2 weeks ago

★ Kick-ass Data Scie

London, United Kingdom

Skimlinks is looking for a

ns with...

3 days ago

nlp machine-learning statistics r matlab

★ Scientist/Research Engineer, Applied Science Turn

Redwood City, CA

Turn delivers real-time insights that transform the way leading advertising agencies and marketers make...

3 weeks ago

Now we have Hadoop!

http://careers.stackoverflow.com/jobs?searchTerm=data+scientist&location=

Past is same as the present: construct a model

How would we have solved this with **Hadoop**?

Construct a **model** of likely purchasers.

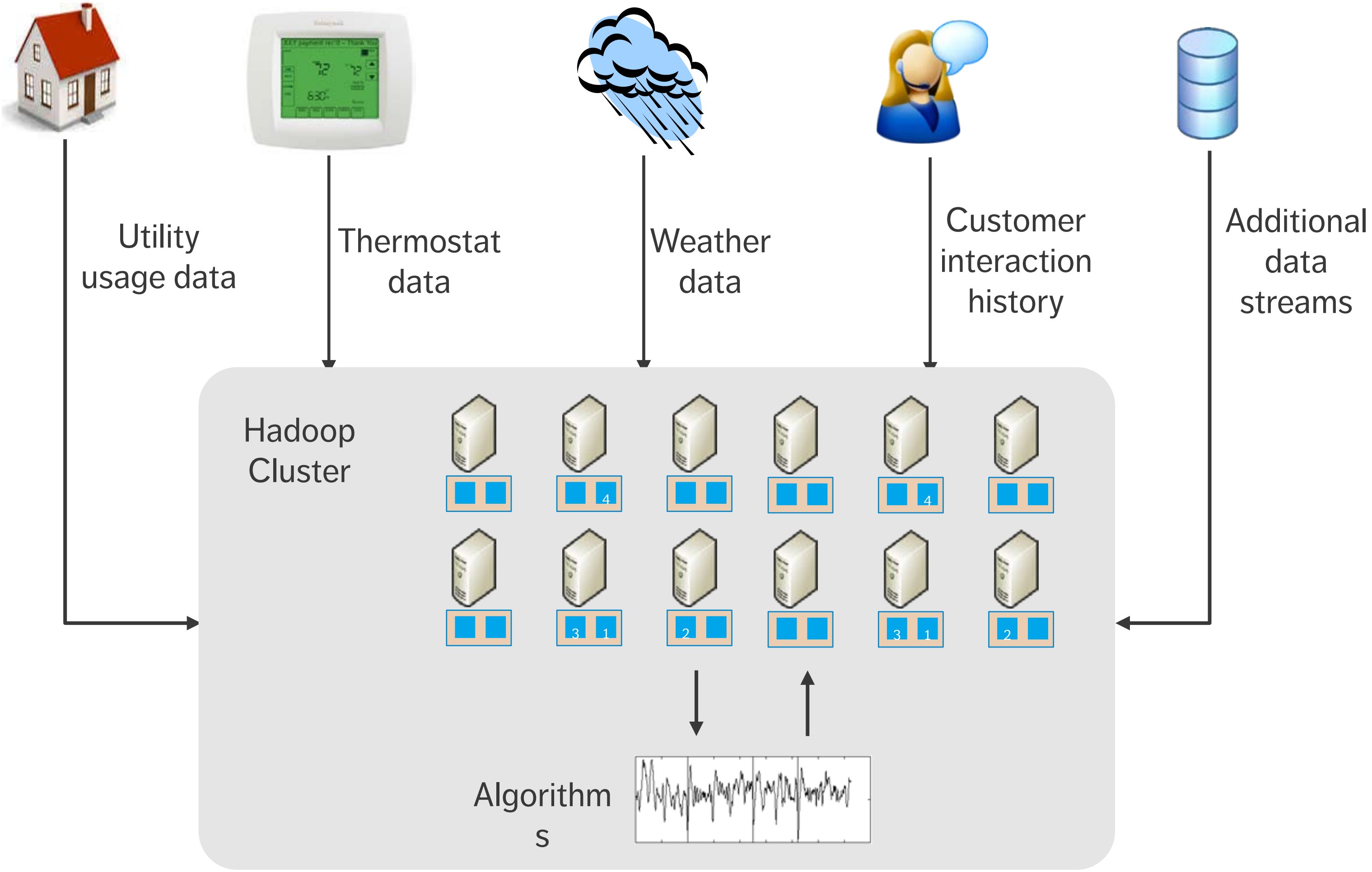
Hadoop has a key advantage

How would we have solved this with **Hadoop**?

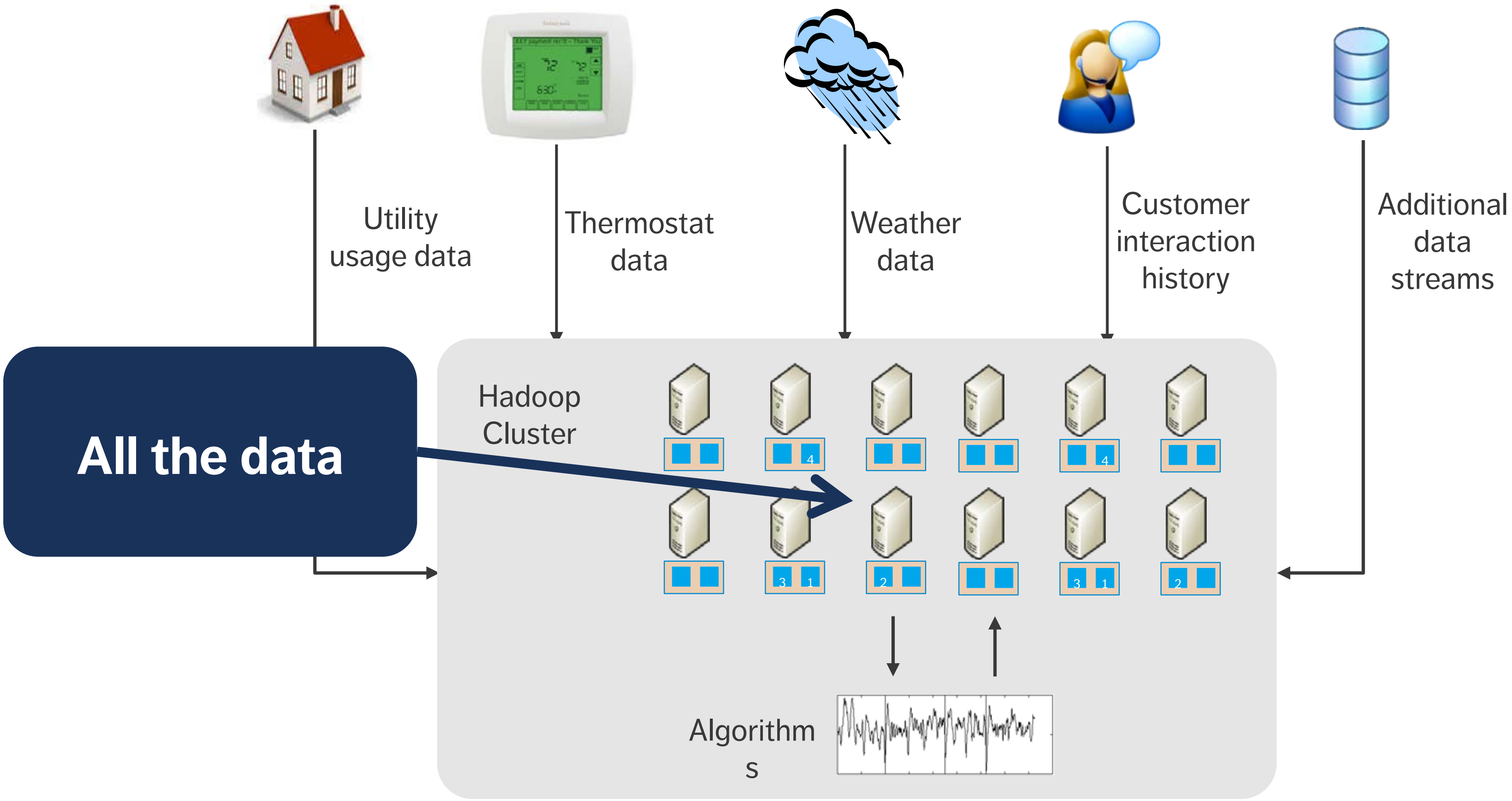
Construct a **model** of likely purchasers.

Integrated data warehousing and data crunching

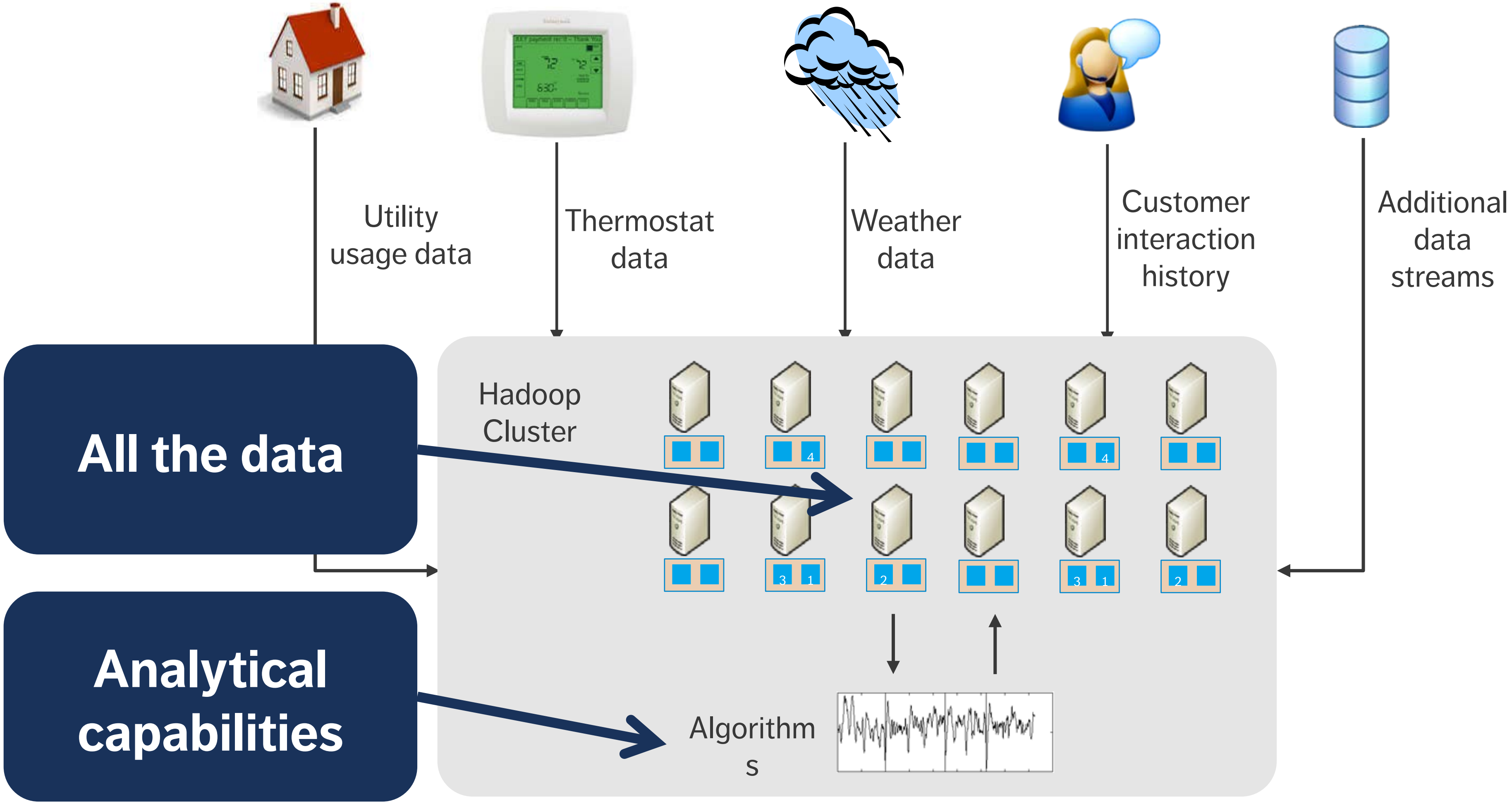
Data and analytical capabilities in a single place



Hadoop solves plumbing problem



Fully integrated data crunching



Utility usage data

Thermostat data

Weather data

Customer interaction history

Additional data streams

Hadoop Cluster

Algorithm s

All the data

Analytical capabilities

Our model is the same. Let's start building it.

Probability(purchase) =

β_1 Electric Heat +

β_2 Similar Purchases +

β_3 Neighbors Purchased +

β_4 Response Rate +

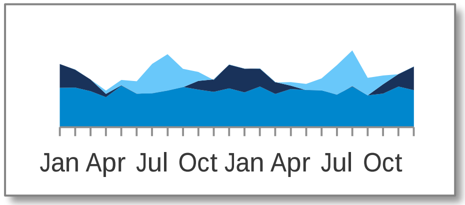
β_5 Type Of Message

Still need weather sensitivity

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

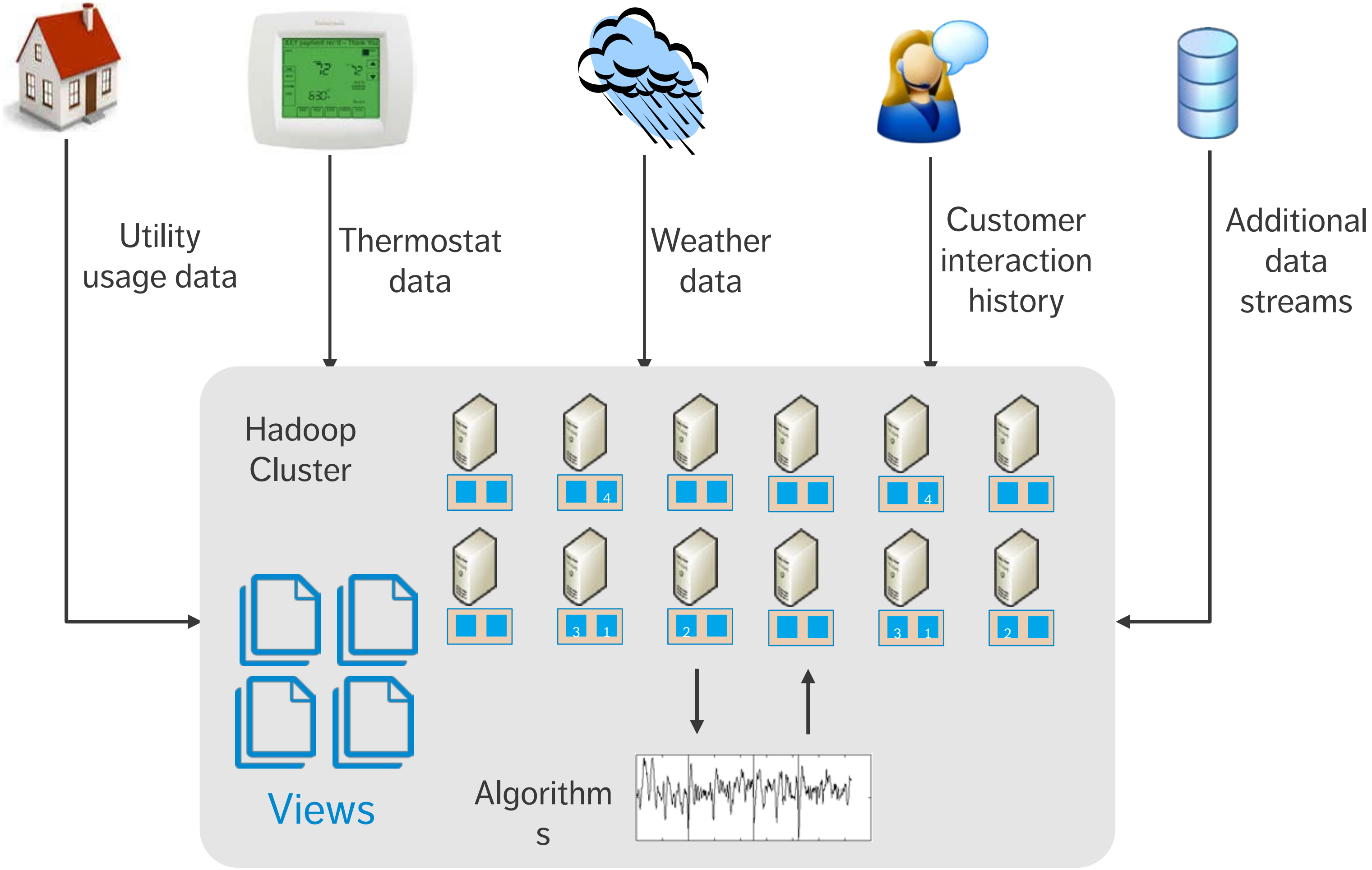
δ_1 **Weather Sensitivity** =



Calculating sensitivity is much **easier** with Hadoop.

Let’s get the data.

Fetch your data with Hive views



Views provide fresh data on demand

Hive is a SQL-like interface to Hadoop.

Hive views are **saved queries** that you treat like a table.

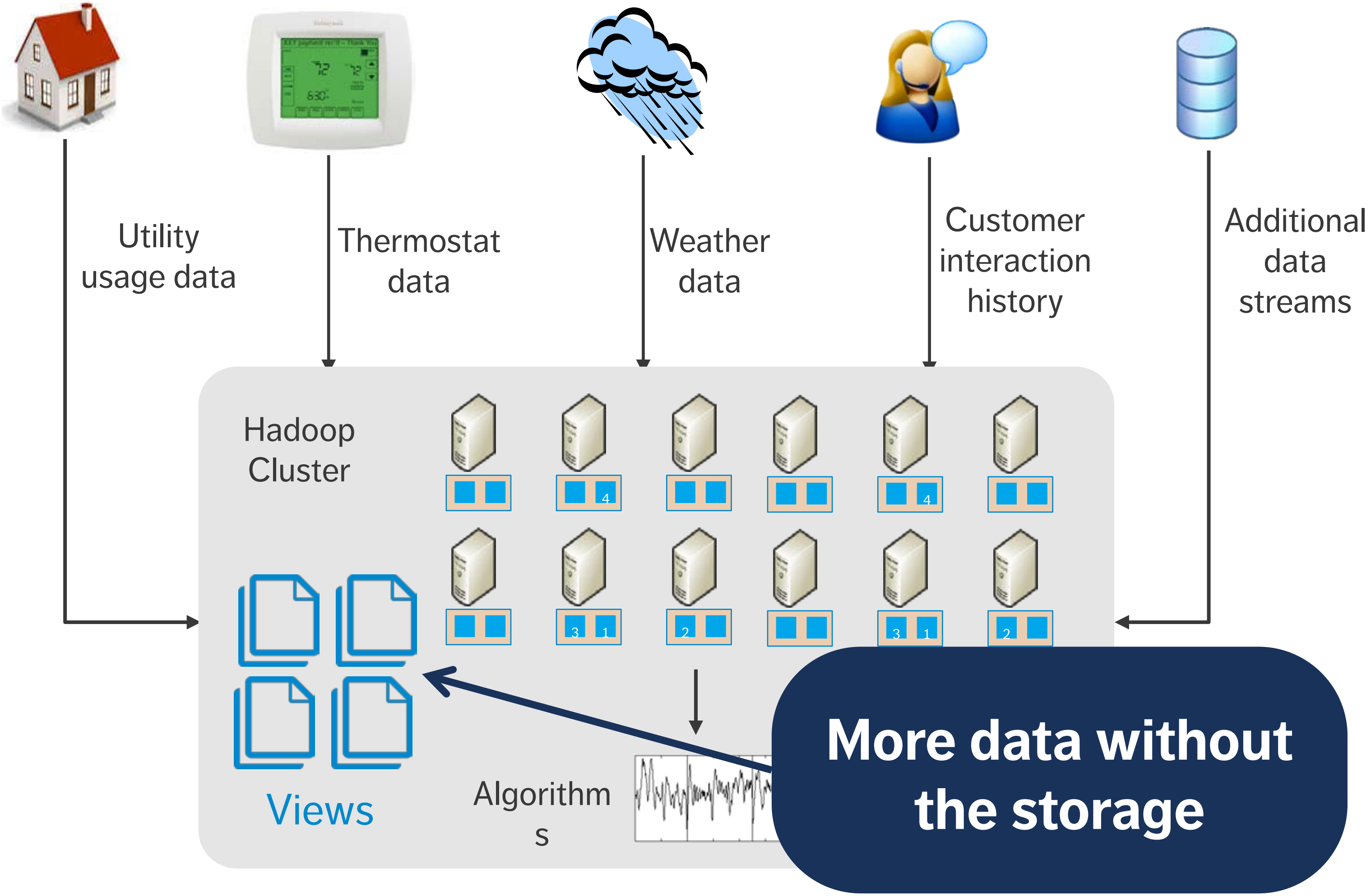
Build views on top of views to setup **complex** analyses.

Querying a view takes **longer to execute**, but it ensures **fresh** data.

View syntax is plain SQL

```
CREATE VIEW
  analytics.disaggregation_inputs_view
AS
SELECT
  w.temperature,
  r.usage_value
FROM
  analytics.weather w
  JOIN analytics.reads r on w.zip_code = r.zip_code
;
```

Views are data on demand



Utility usage data

Thermostat data

Weather data

Customer interaction history

Additional data streams

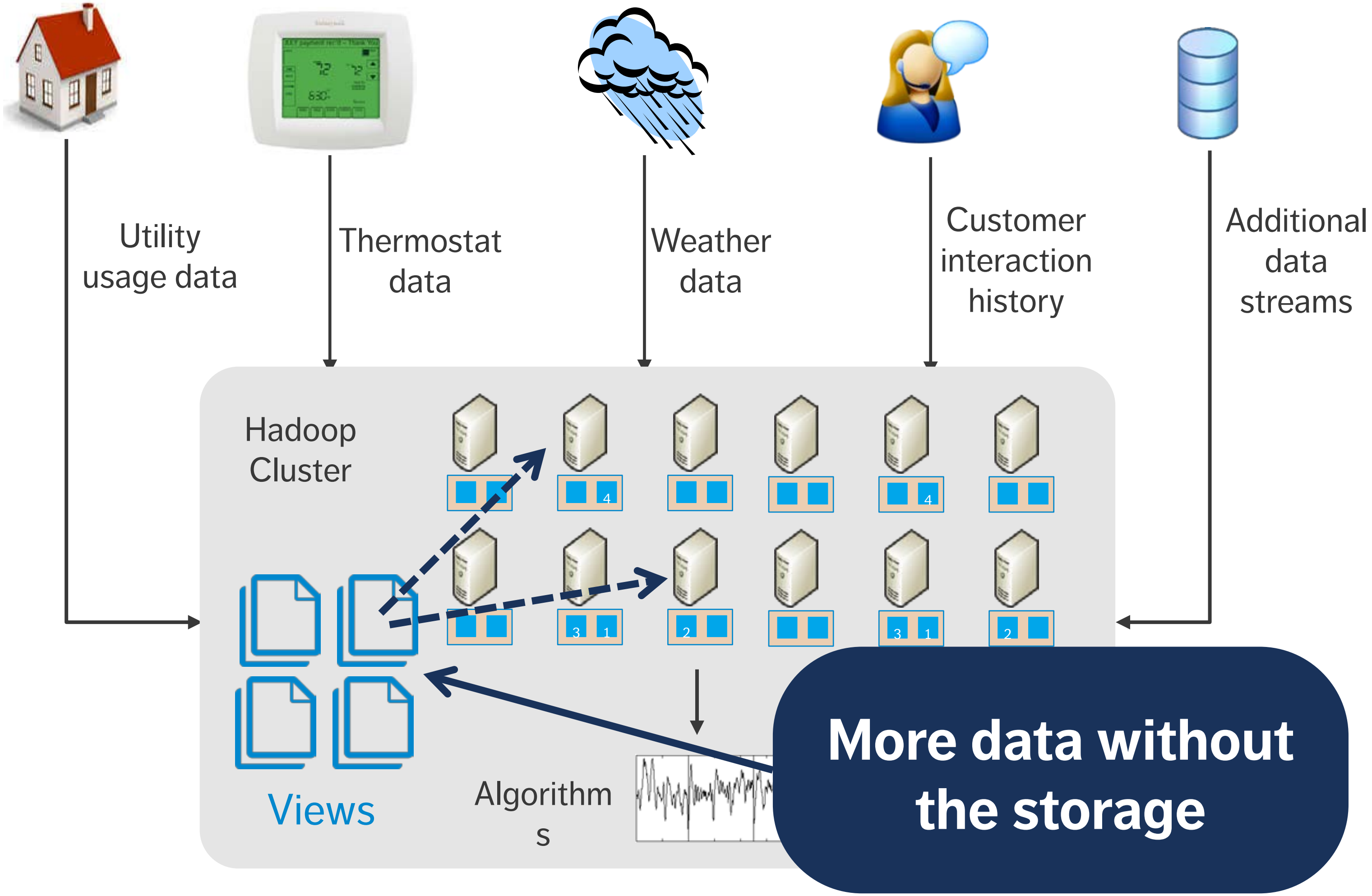
Hadoop Cluster

Views

Algorithm s

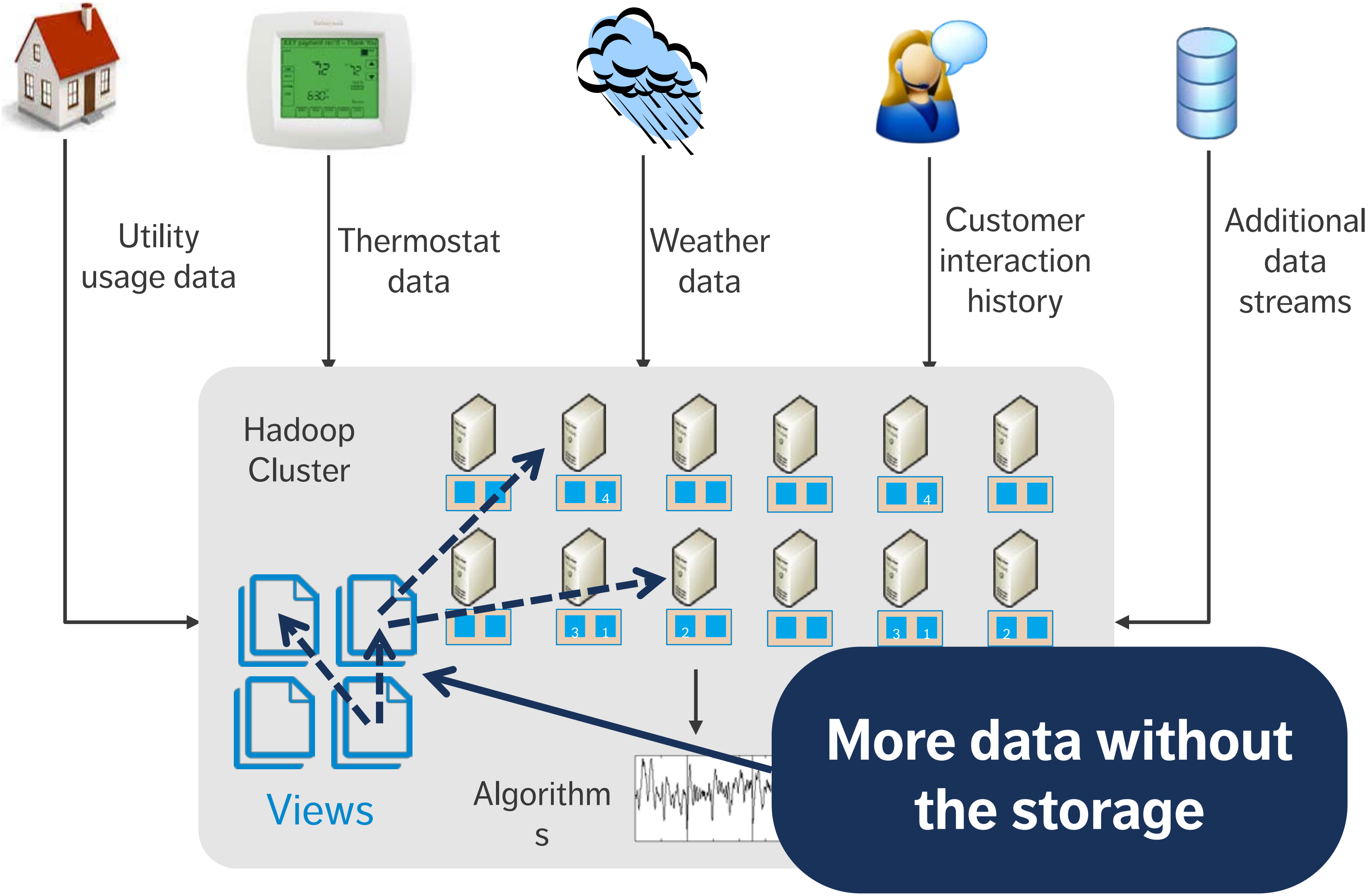
More data without the storage

Views point at data without storing it



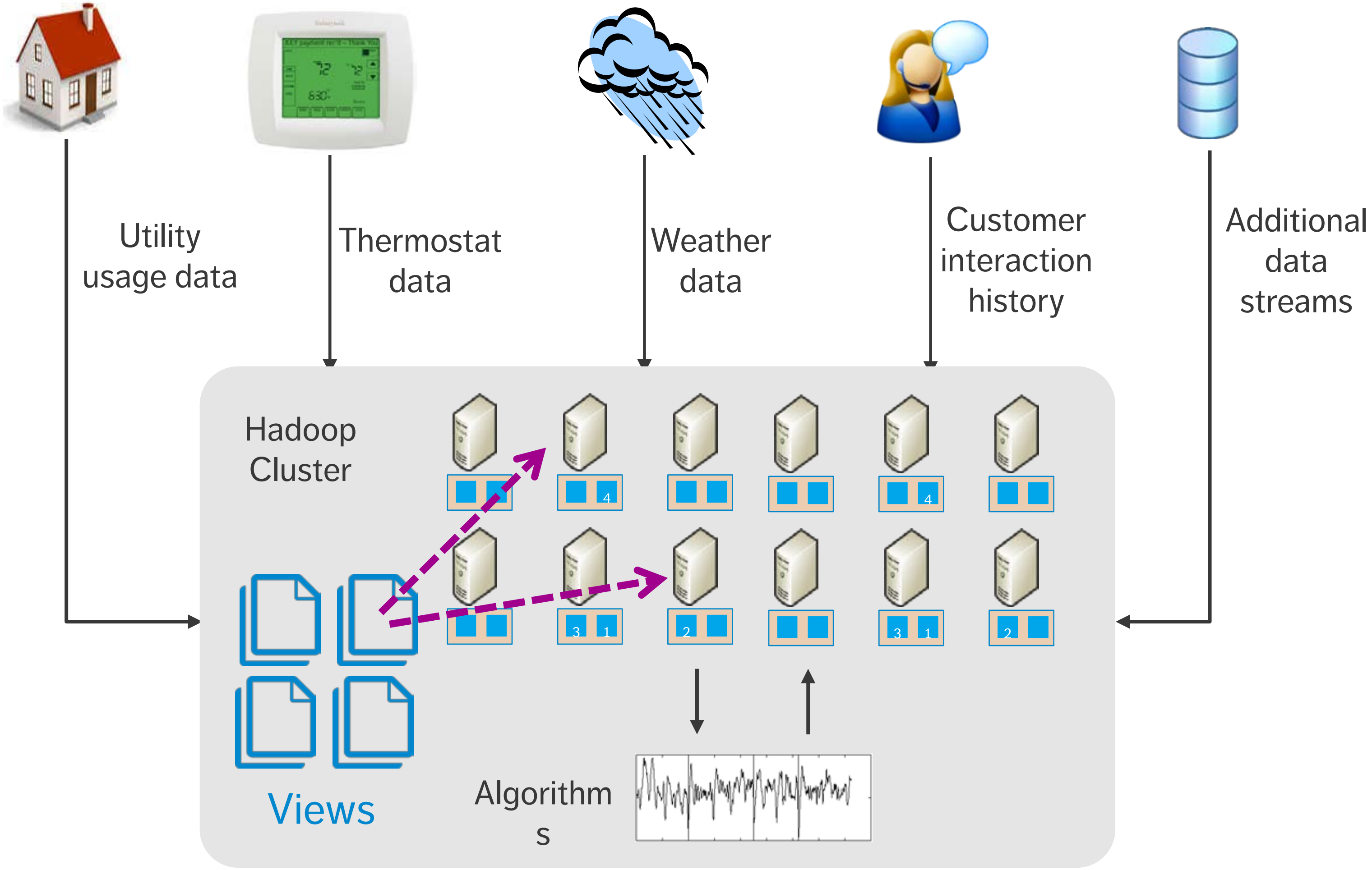
More data without the storage

Views on top of views for complex analyses



More data without the storage

Setup a view to get disaggregation data

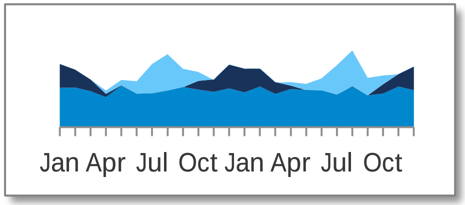


We have our disaggregation data

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

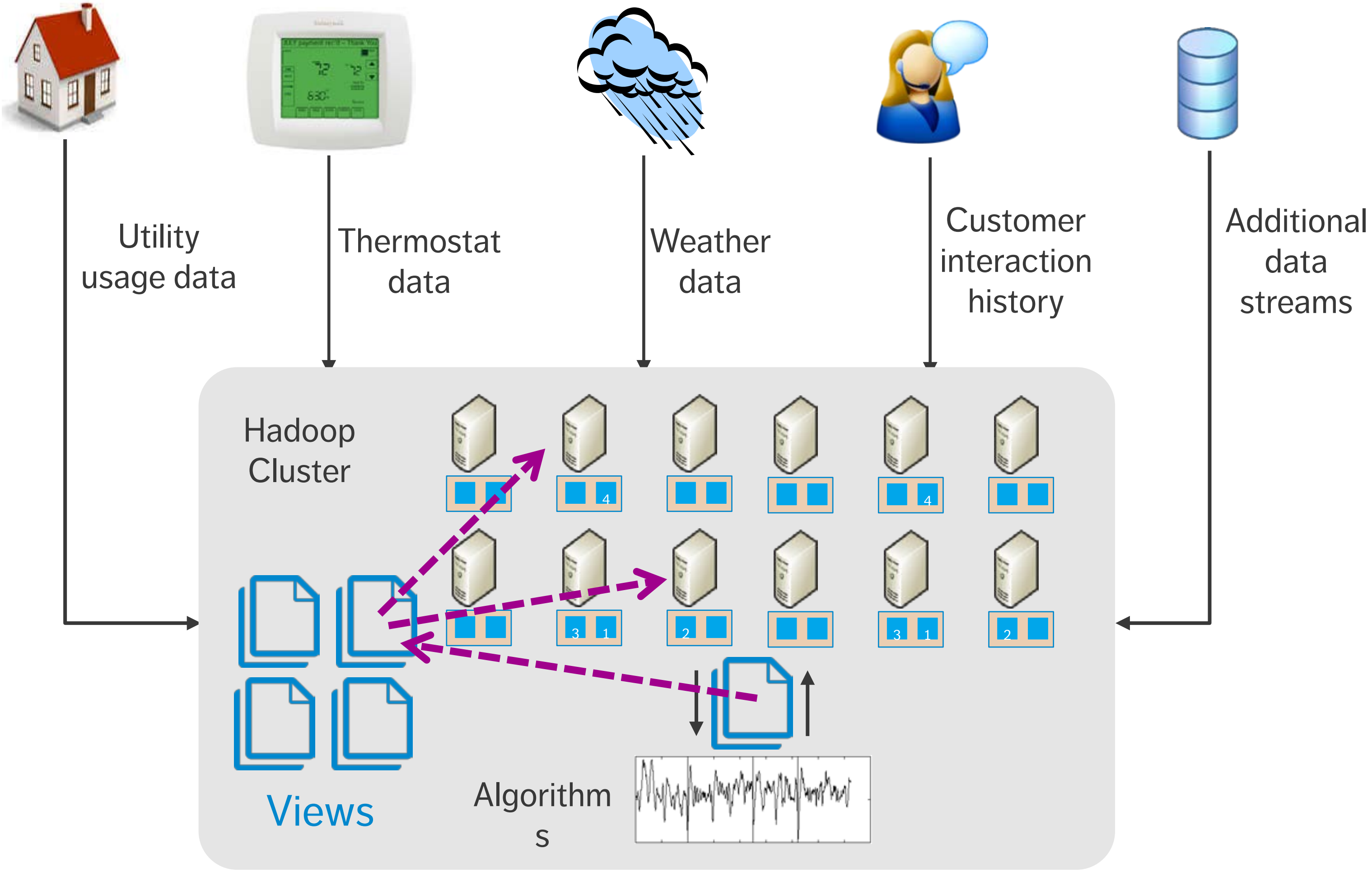
δ_1 **Weather Sensitivity** =



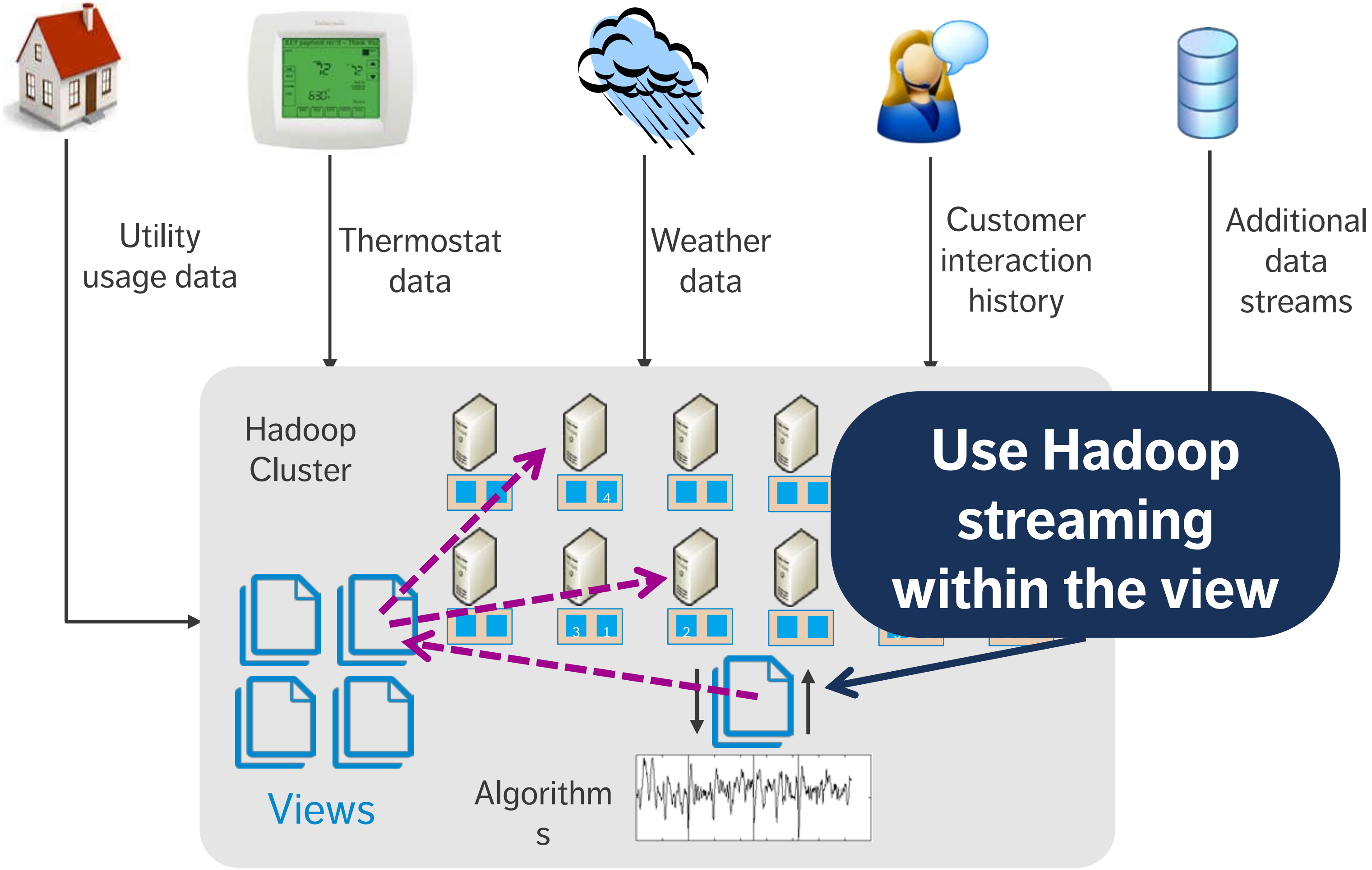
We need to **calculate** the model and **store** the results.

Hadoop is built to do both.

Setup a view to run disaggregation algorithms



Hadoop streaming + Views = Power



Use Hadoop streaming within the view

Hadoop streaming can calculate anything

Stream data through **any script**.

Pipe any data through **standard input** and send any data to **standard output**.

Integrate with **any language**: R, Python, Ruby, Bash, Java, etc.

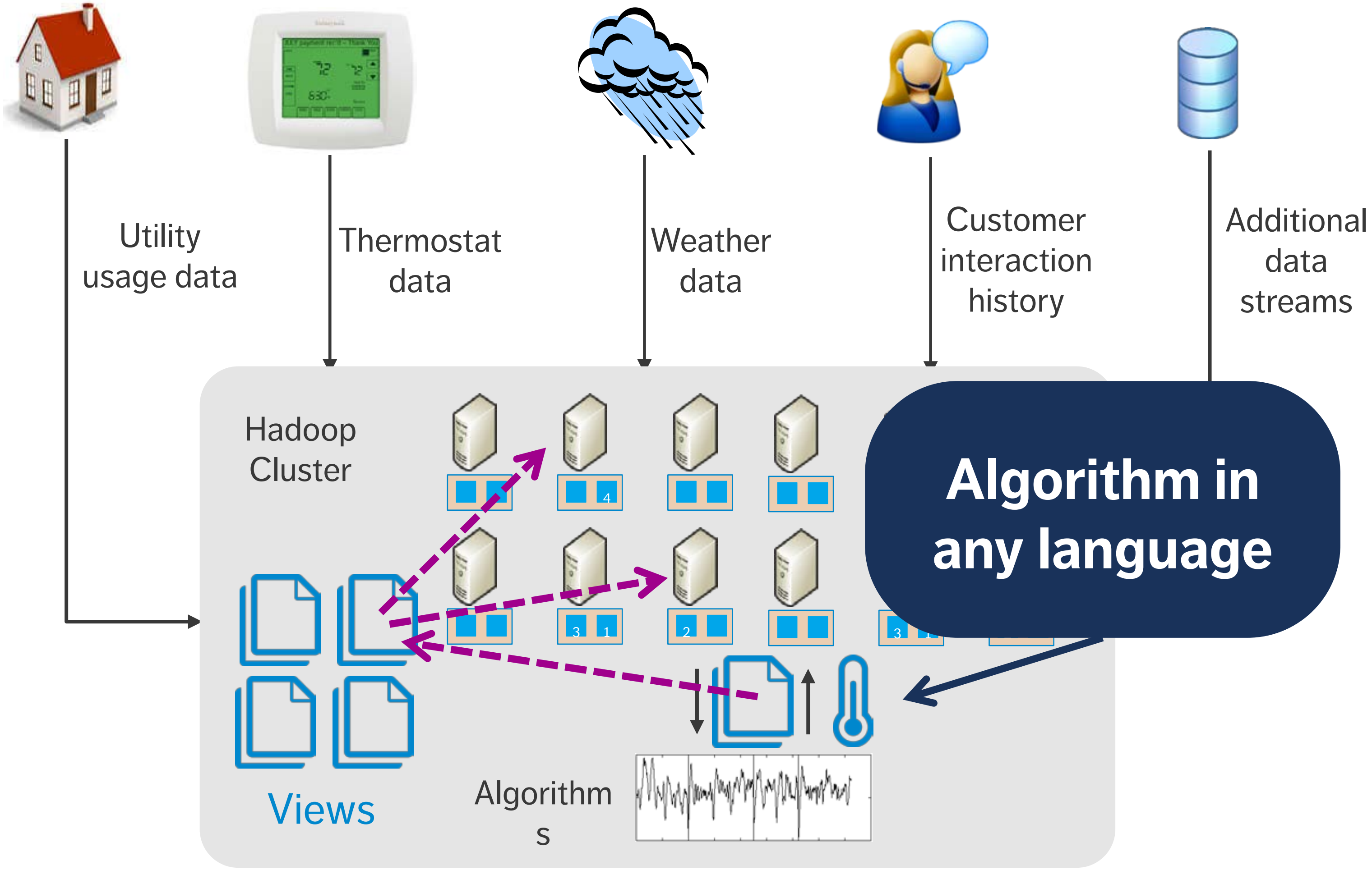
SELECT TRANSFORM command in Hive is an easy way to use Hadoop streaming.

Hadoop streaming is easy to implement in Hive

```
CREATE VIEW
  analytics.disaggregation_outputs_view
AS
SELECT
  TRANSFORM (
    diw.temperature,
    diw.usage_value
  )
USING
  ‘weather_disaggregation.R’
FROM
  analytics.disaggregation_inputs_view diw
;
```

Executable reads from stdin and writes to stdout

Simple SQL syntax to produce any result



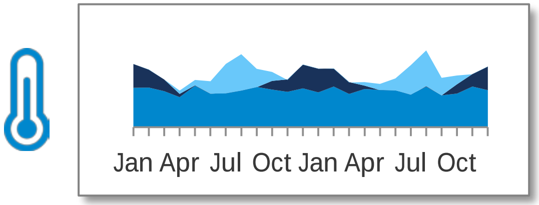
Algorithm in any language

We know each customer’s heading sensitivity

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

✓ **Weather Sensitivity** =



Let’s continue with our electric heat model.

We’re ready to model electric heat

Probability(purchase) =

β_1 Pr(**Electric Heat**) =

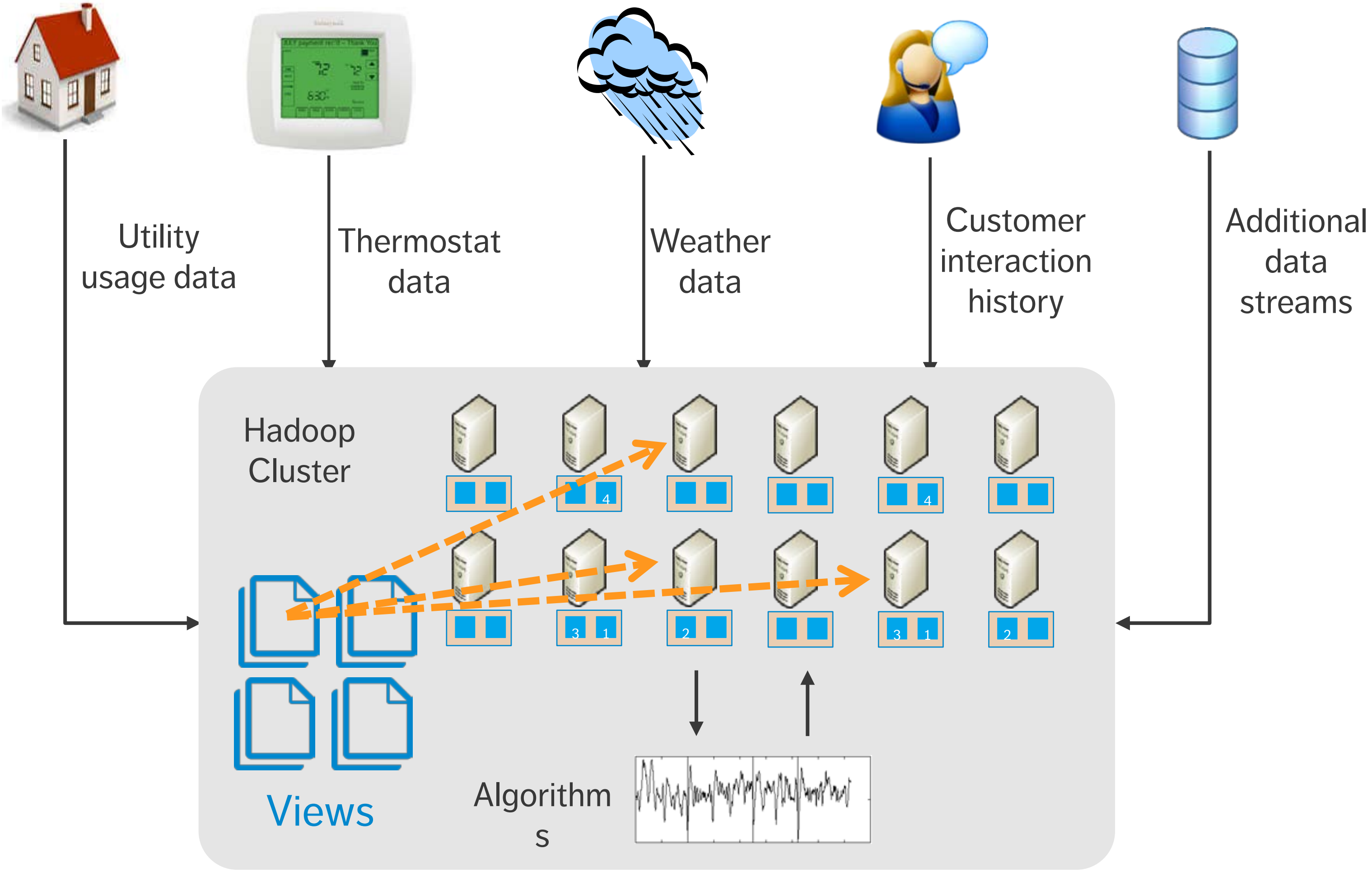
δ_1 Weather Sensitivity +

δ_2 Neighbors Heat +

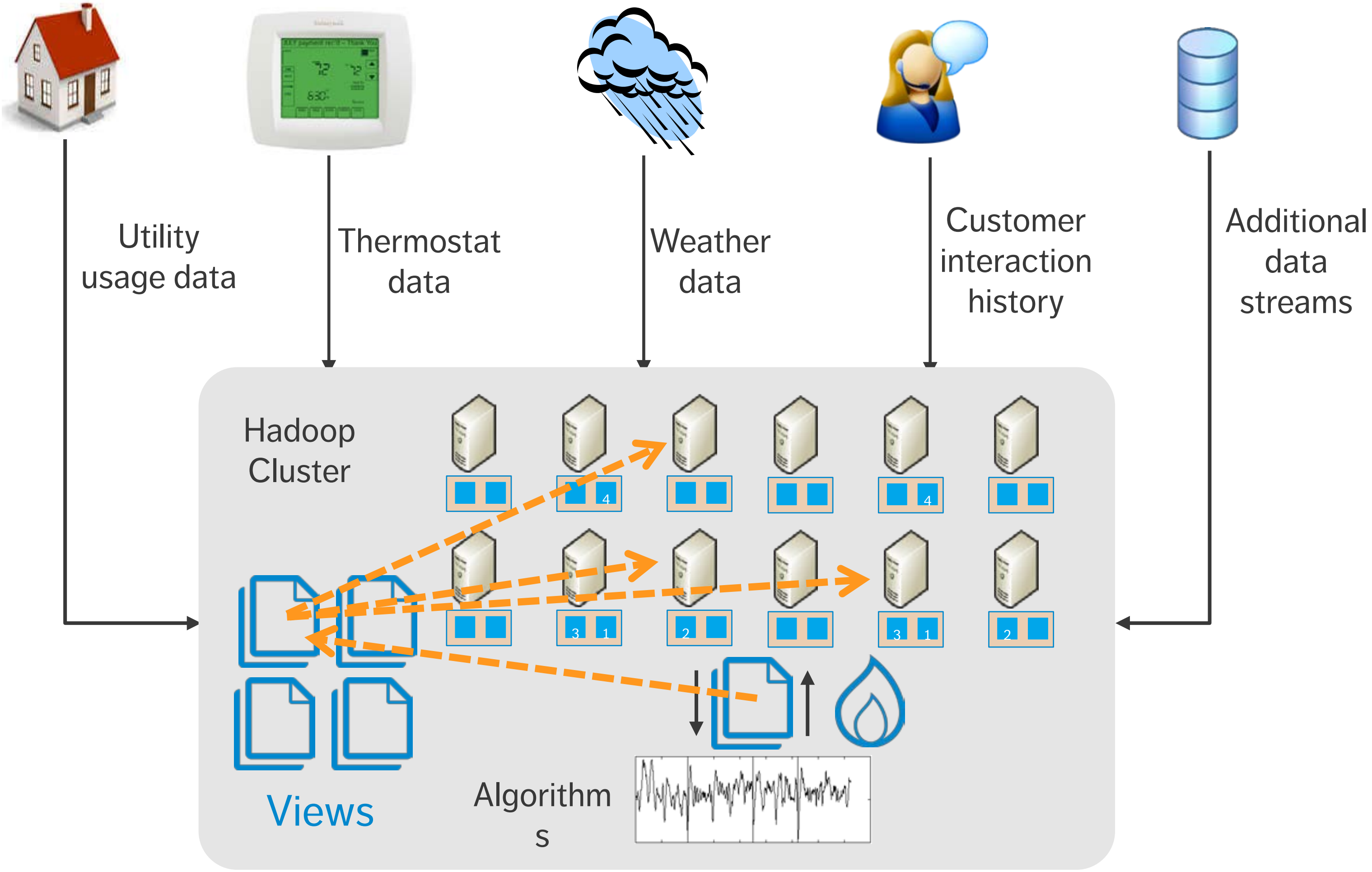
δ_3 Natural Gas Price

Let’s get our data.

Setup a view to fetch data for electric heat model



Implement electric heat model in a view



We know each customer's heat type

Probability(purchase) =

✓ Pr(**Electric Heat**) =

δ_1 Weather Sensitivity +

δ_2 Neighbors Heat +

δ_3 Natural Gas Price

Let's continue with our water heater purchase model.

We’re ready to model purchase behavior

Probability(purchase) =

β_1 Electric Heat +

β_2 Similar Purchases +

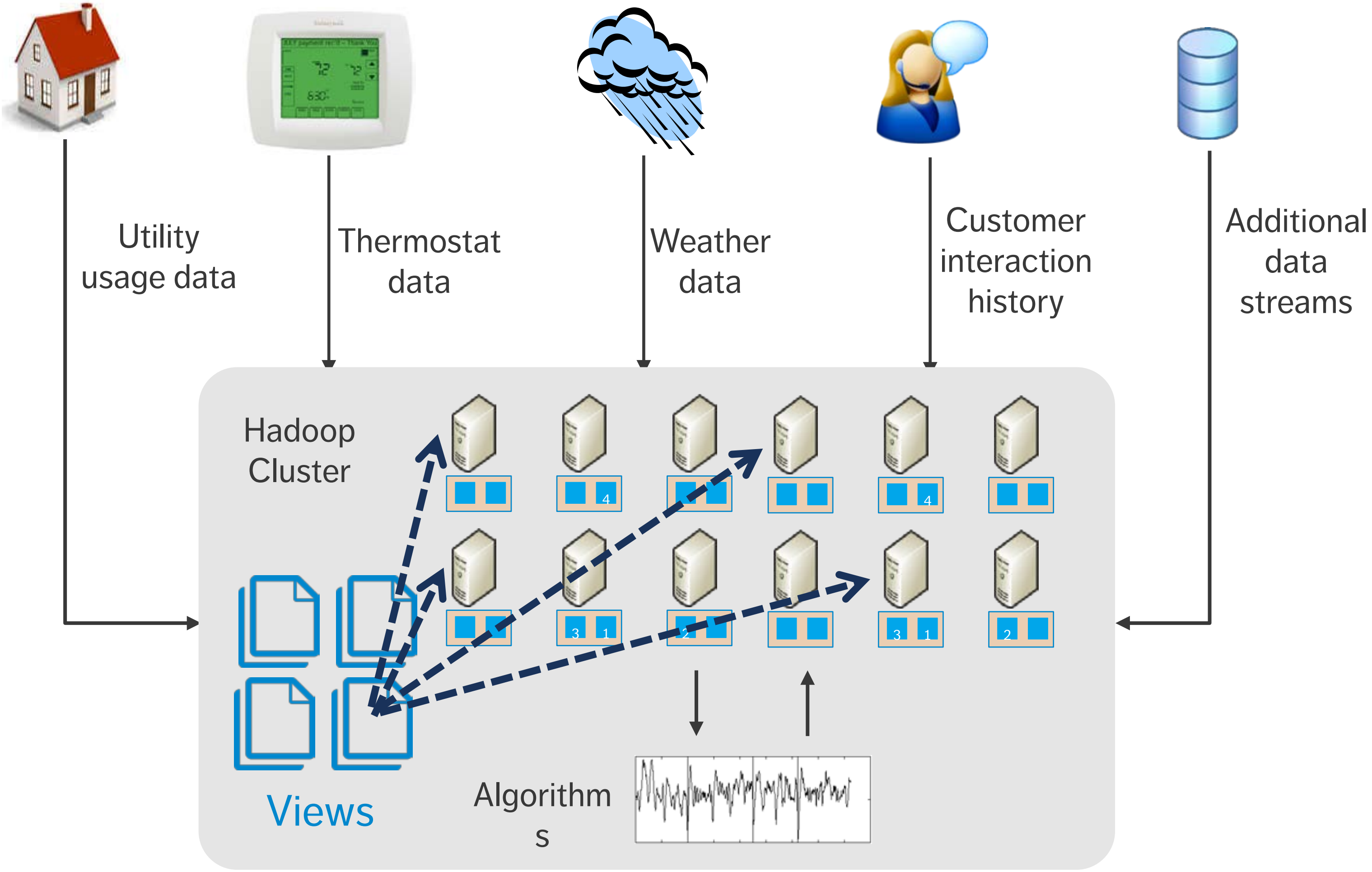
β_3 Neighbors Purchased +

β_4 Response Rate +

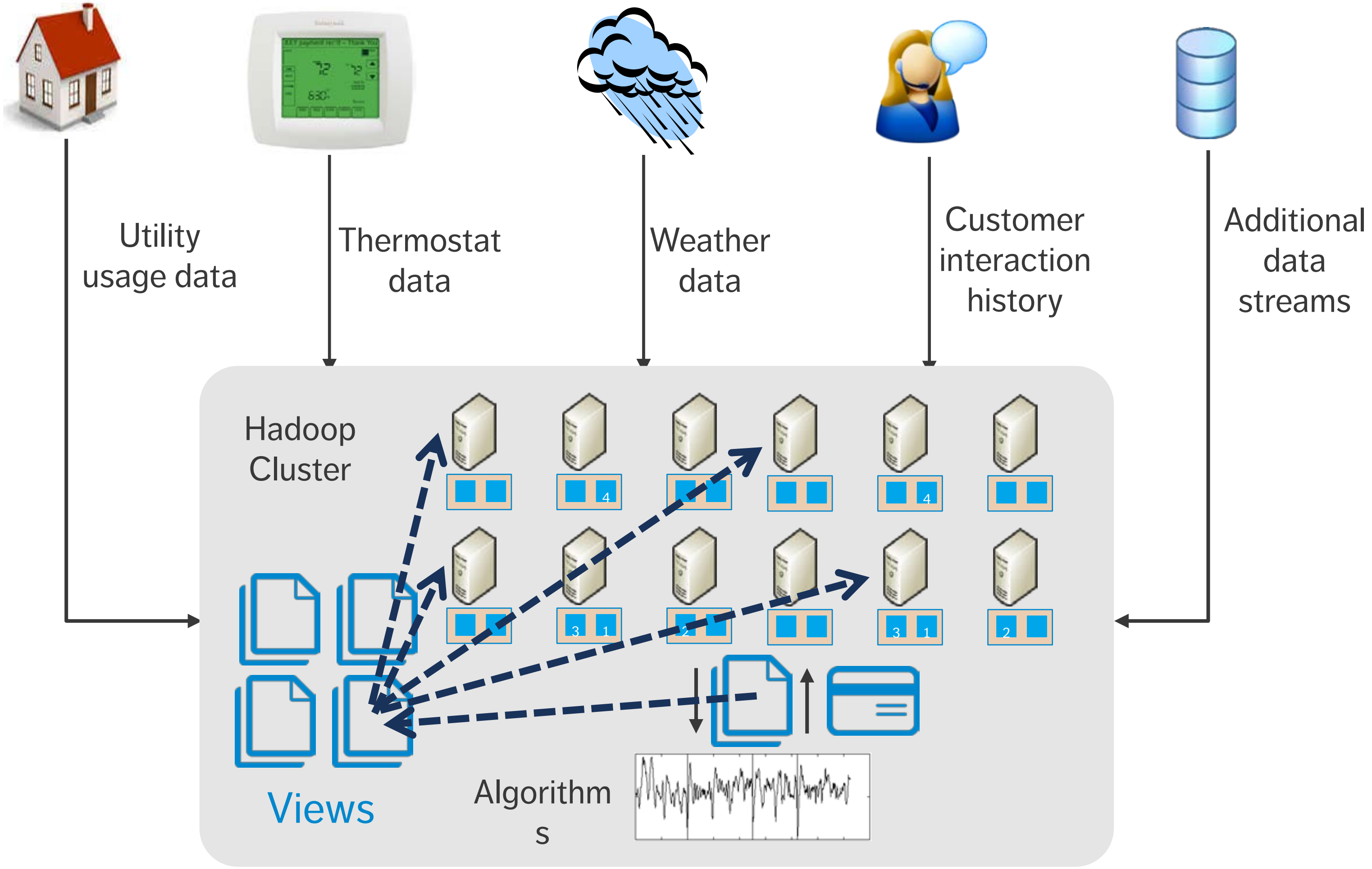
β_5 Type Of Message

Let’s get our data.

Setup a view to fetch data for purchase behavior model



Implement purchase behavior model



Utility usage data

Thermostat data

Weather data

Customer interaction history

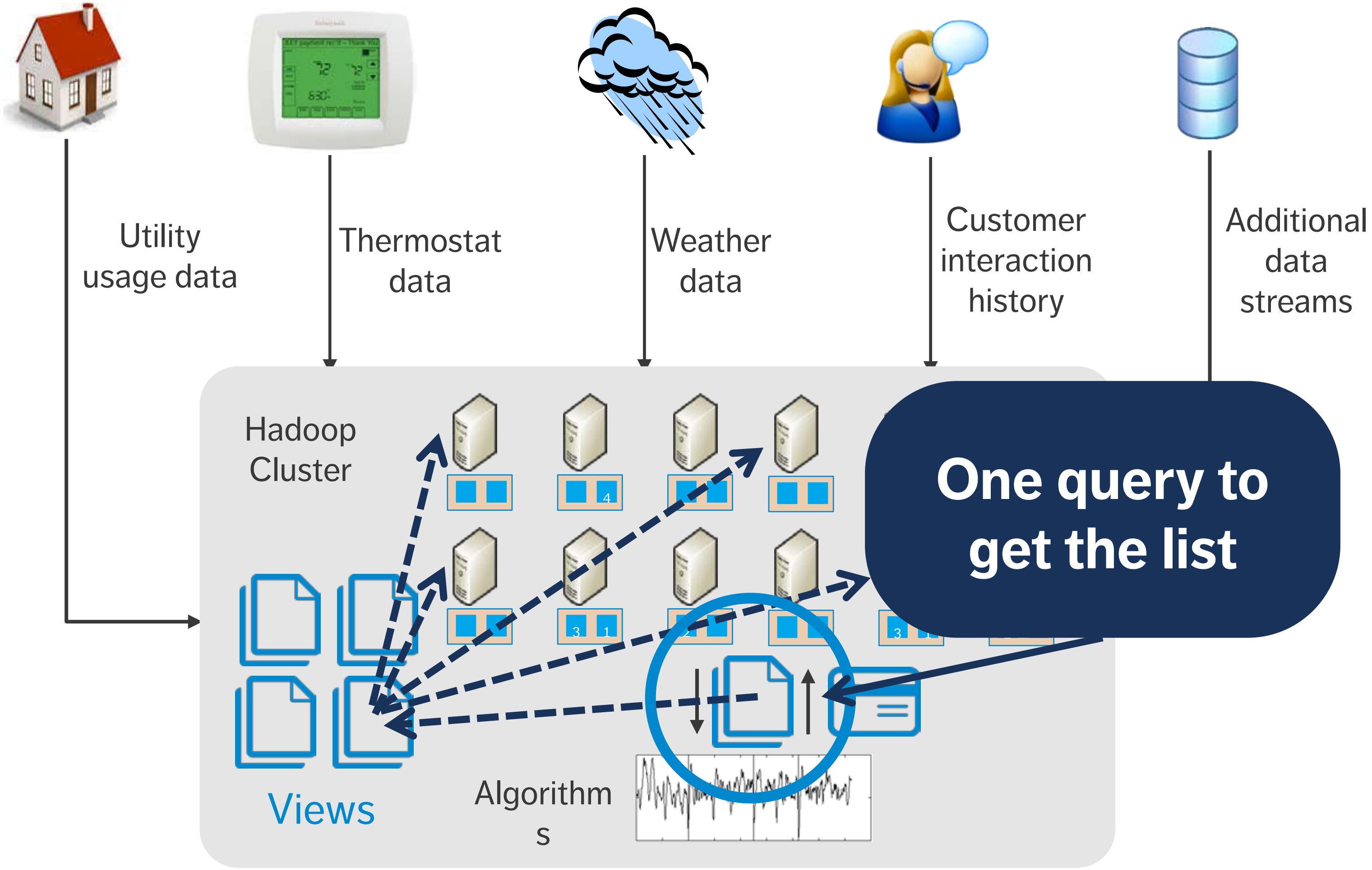
Additional data streams

Hadoop Cluster

Views

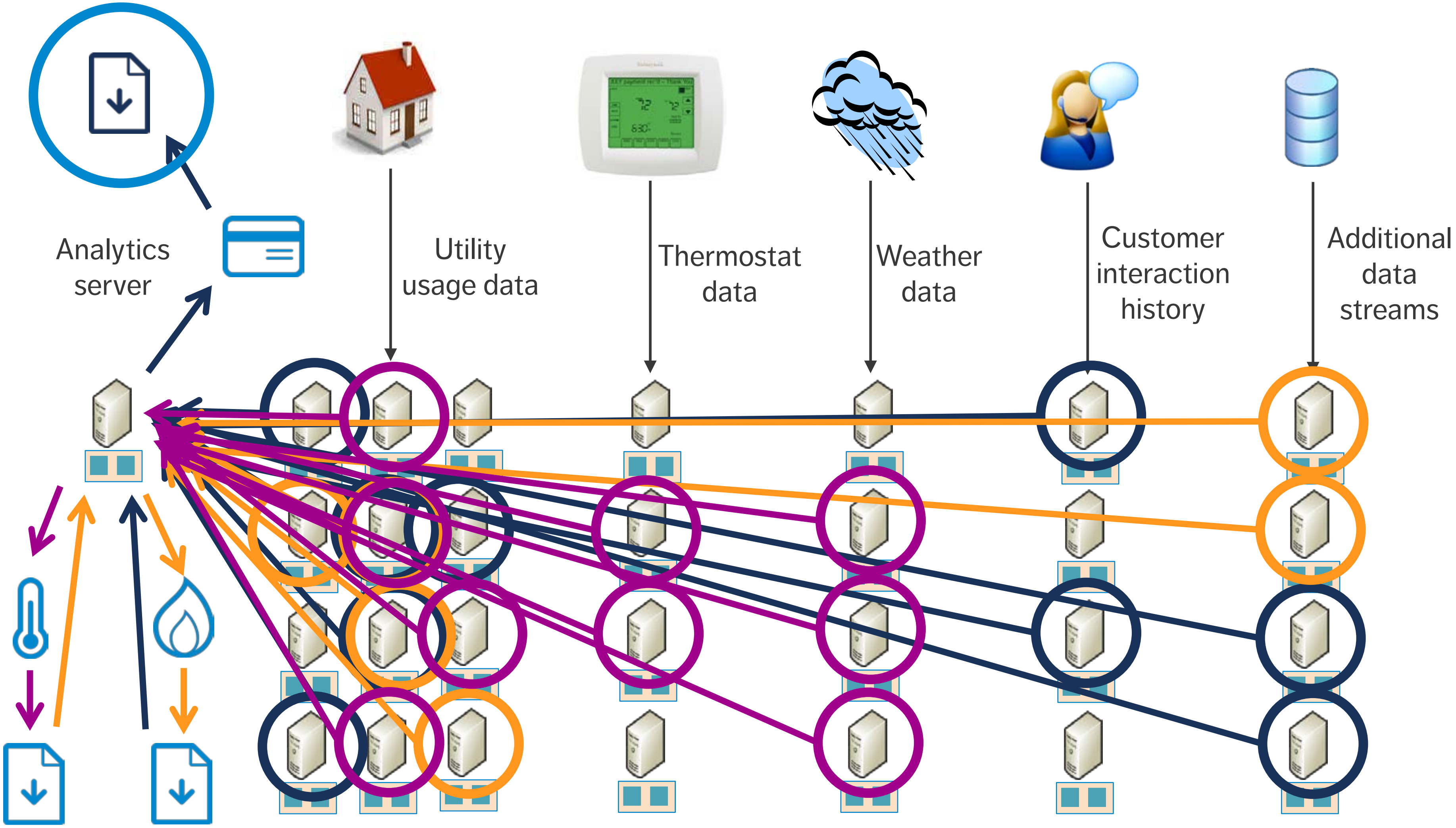
Algorithm
s

We have our desired result



One query to get the list

Major plumbing in the old world



Some considerations on the past vs now

	Past ⚠	Now ✓
Refresh data		
Score new households		
Add new data source		
Build new model		

Refreshing data is a breeze

	Past ⚠	Now ✓
Refresh data	Major plumbing	Single query
Score new households		
Add new data source		
Build new model		

Easy to calculate insights for new households

	Past ⚠	Now ✓
Refresh data	Major plumbing	Single query
Score new households	Major plumbing	Single query
Add new data source		
Build new model		

New data? No problem.

	Past ⚠	Now ✓
Refresh data	Major plumbing	Single query
Score new households	Major plumbing	Single query
Add new data source	Major plumbing	Couple lines of SQL
Build new model		

Re-use previous work for new models

	Past ⚠	Now ✓
Refresh data	Major plumbing	Single query
Score new households	Major plumbing	Single query
Add new data source	Major plumbing	Couple lines of SQL
Build new model	Major plumbing	Re-use views

Hadoop radically reduces plumbing

	Past ⚠	Now ✓
Refresh data	Major plumbing	Single query
Score new households	Major plumbing	Single query
Add new data source	Major plumbing	Couple lines of SQL
Build new model	Major plumbing	Re-use views

Big data

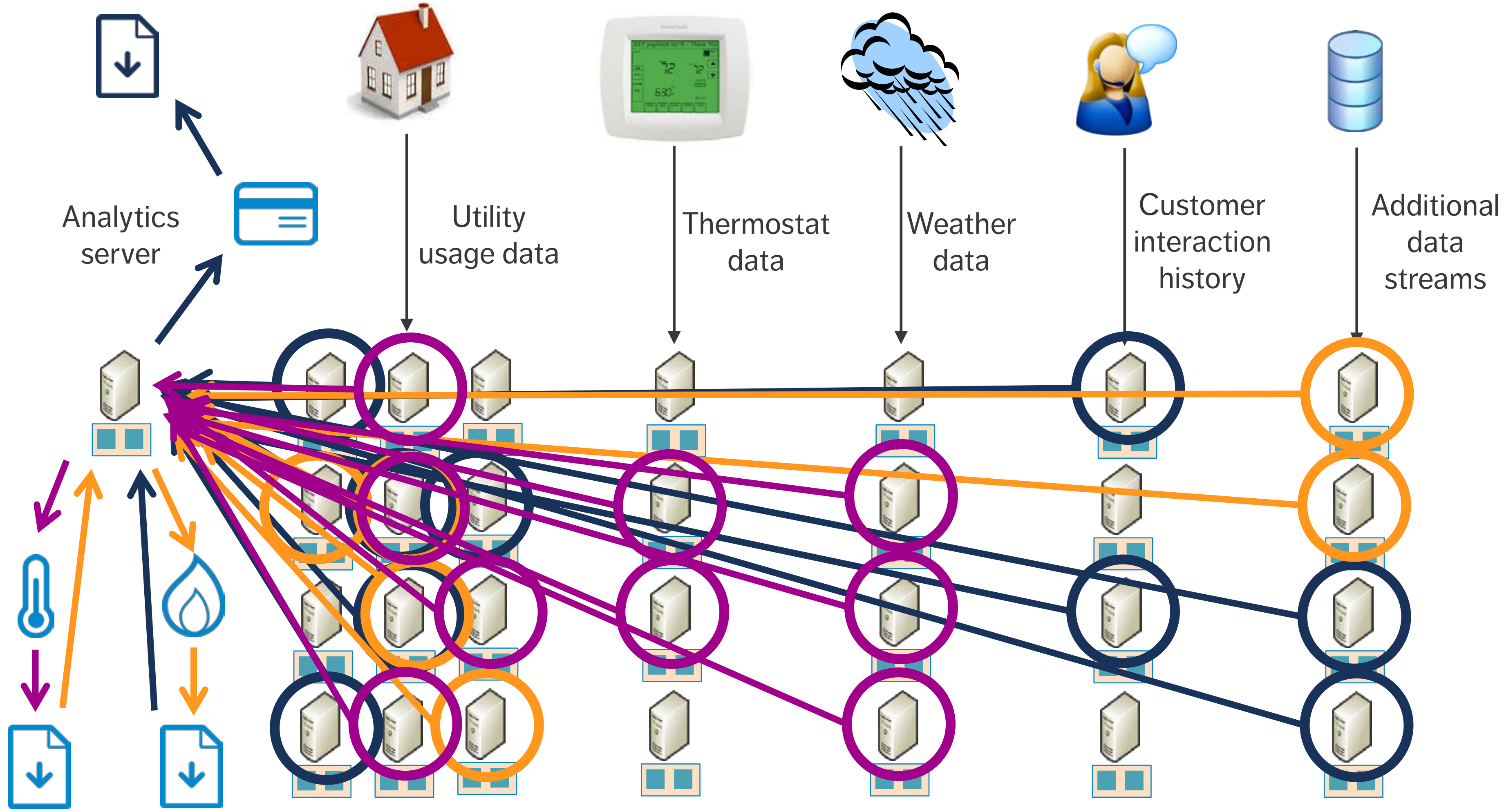
Big data

Quantity

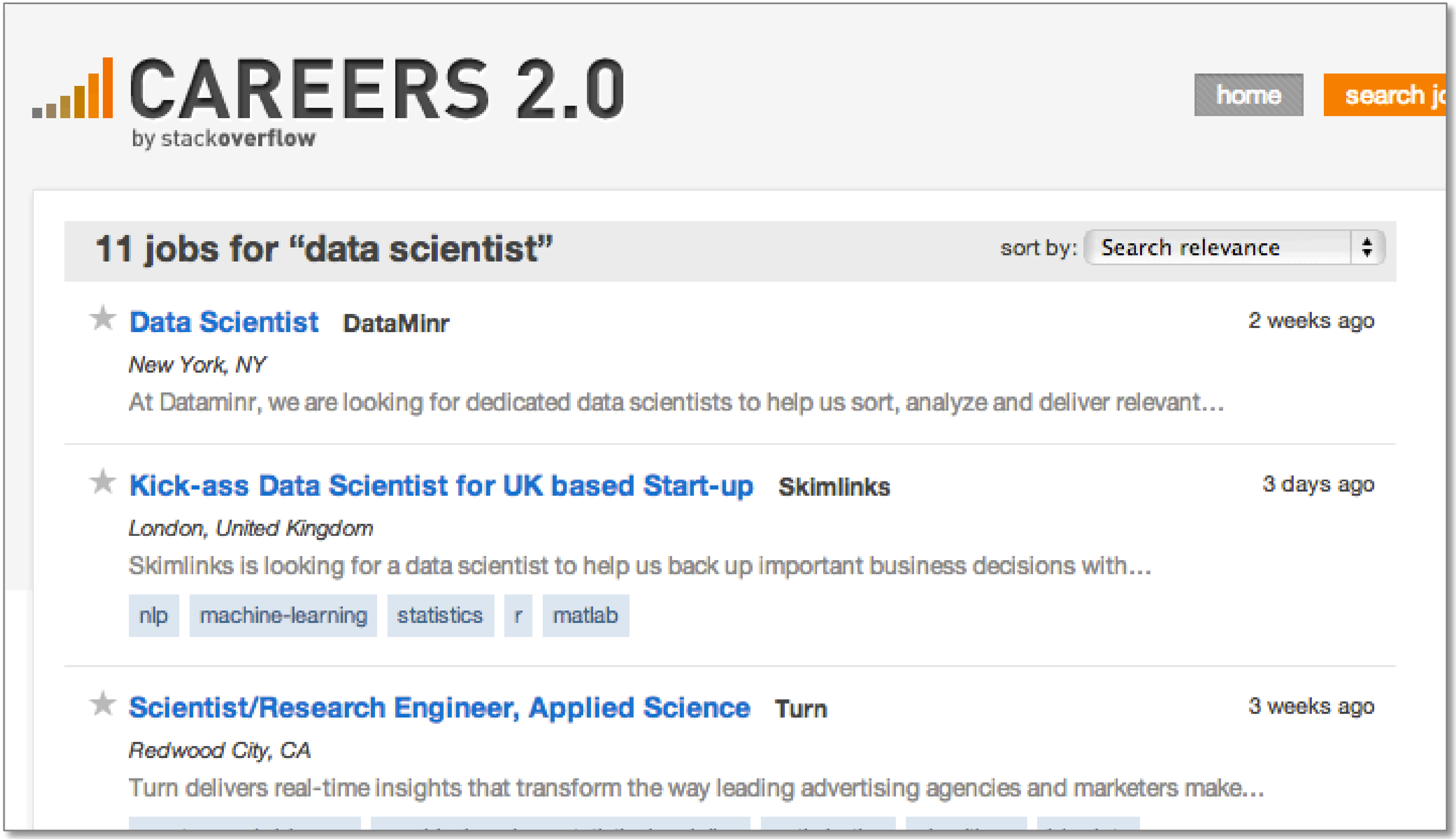
Big data

Variety + Quantity

It doesn't have to be like this



You could look for a new job



11 jobs for “data scientist”

sort by: Search relevance

★ Data Scientist DataMinr — 2 weeks ago

New York, NY

At Dataminr, we are looking for dedicated data scientists to help us sort, analyze and deliver relevant…

★ Kick-ass Data Scientist for UK based Start-up Skimlinks — 3 days ago

London, United Kingdom

Skimlinks is looking for a data scientist to help us back up important business decisions with…

nlp machine-learning statistics r matlab

★ Scientist/Research Engineer, Applied Science Turn — 3 weeks ago

Redwood City, CA

Turn delivers real-time insights that transform the way leading advertising agencies and marketers make…

http://careers.stackoverflow.com/jobs?searchTerm=data+scientist&location=

Hadoop

Big data plumbing

OPOWER

Happy plumbing!

Erik Shilts
Advanced Analytics

erik.shilts@opower.com