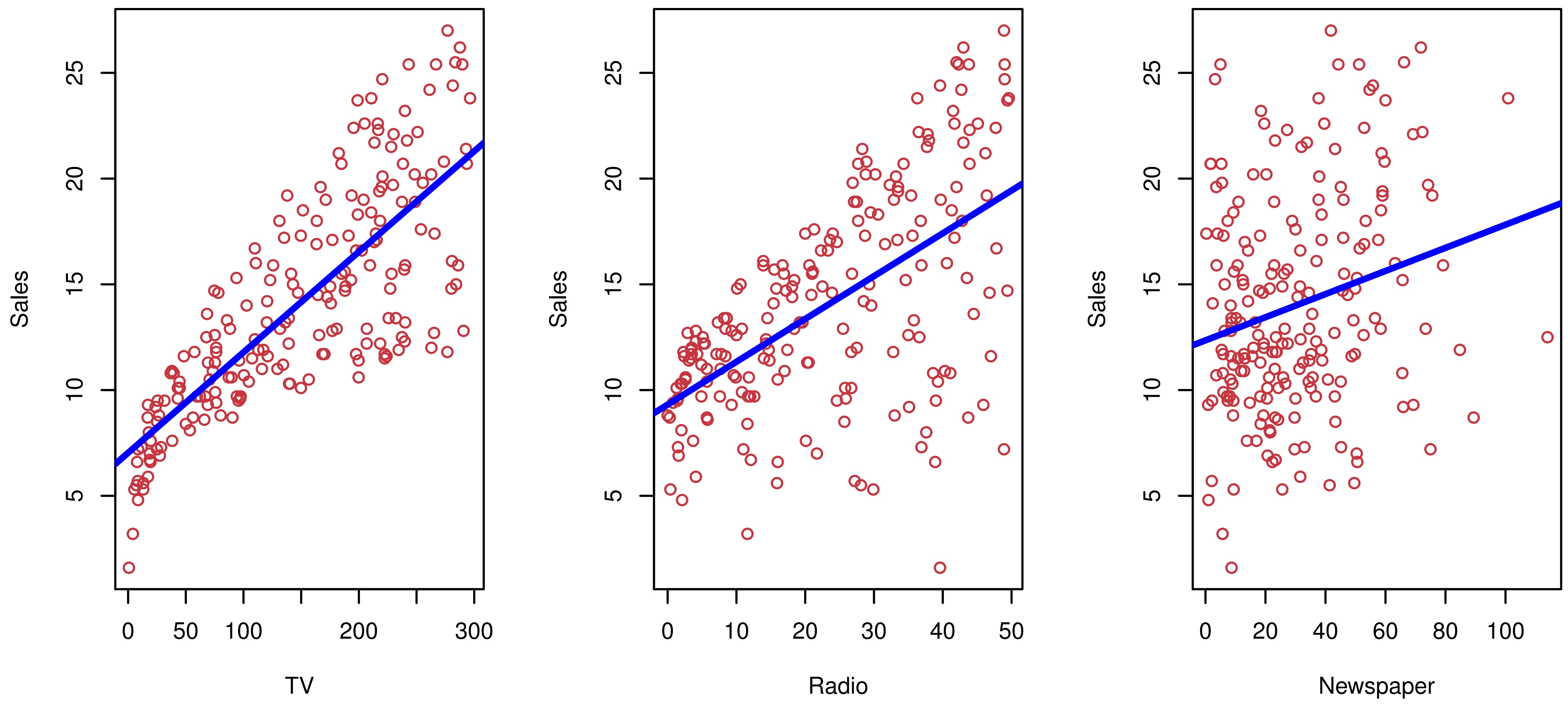
# Machine Learning

Anisio Lacerda

# What is Machine Learning?

- Can we **predict sales** using TV, Radio and Newspaper **budgets**?



200 markets (points) and thousands of units

# Example: Predicting sales

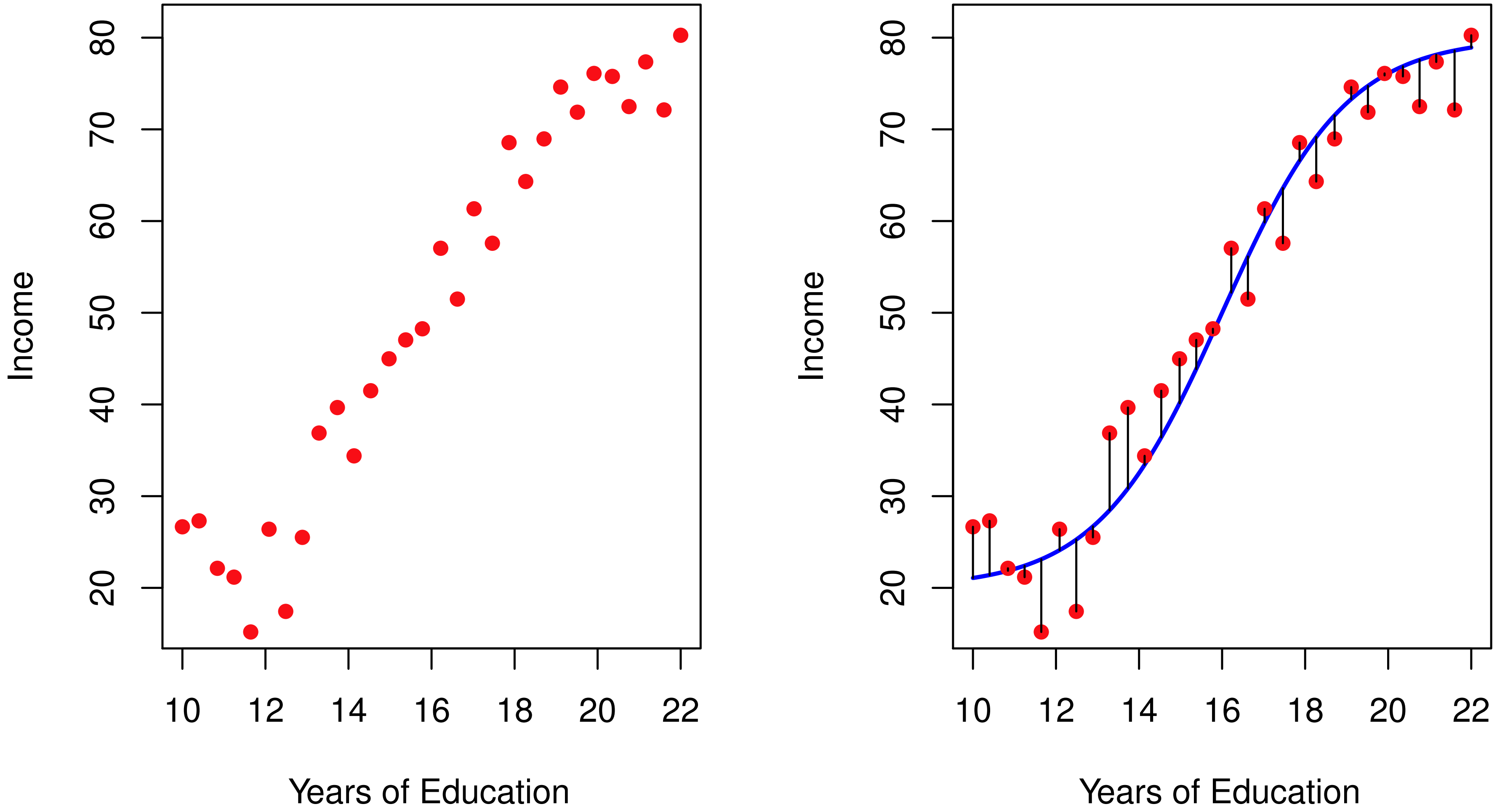
- It is **not** possible to **directly increase sales**
- We **can** control the **advertising expenditure**
- If there is an **association** between **advertising** and **sales**, we are able to **adjust budgets** to increase sales
- Hence, we want to develop an **accurate model** to **predict sales** based on the **three media budgets**

# Notation

- **Output** variable: **sales**
  - Generically refer to response as Y
- **Input** variables (features) X: **advertising budgets**:
  - TV ($X_1$), Radio ($X_2$) and Newspaper ($X_3$)
- We may refer to the input vector as: $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$
- We write our model as: $Y = f(X) + \epsilon$

where $\epsilon$ captures measurement errors and other discrepancies and f is some fixed but unknown function

# Another example: Income



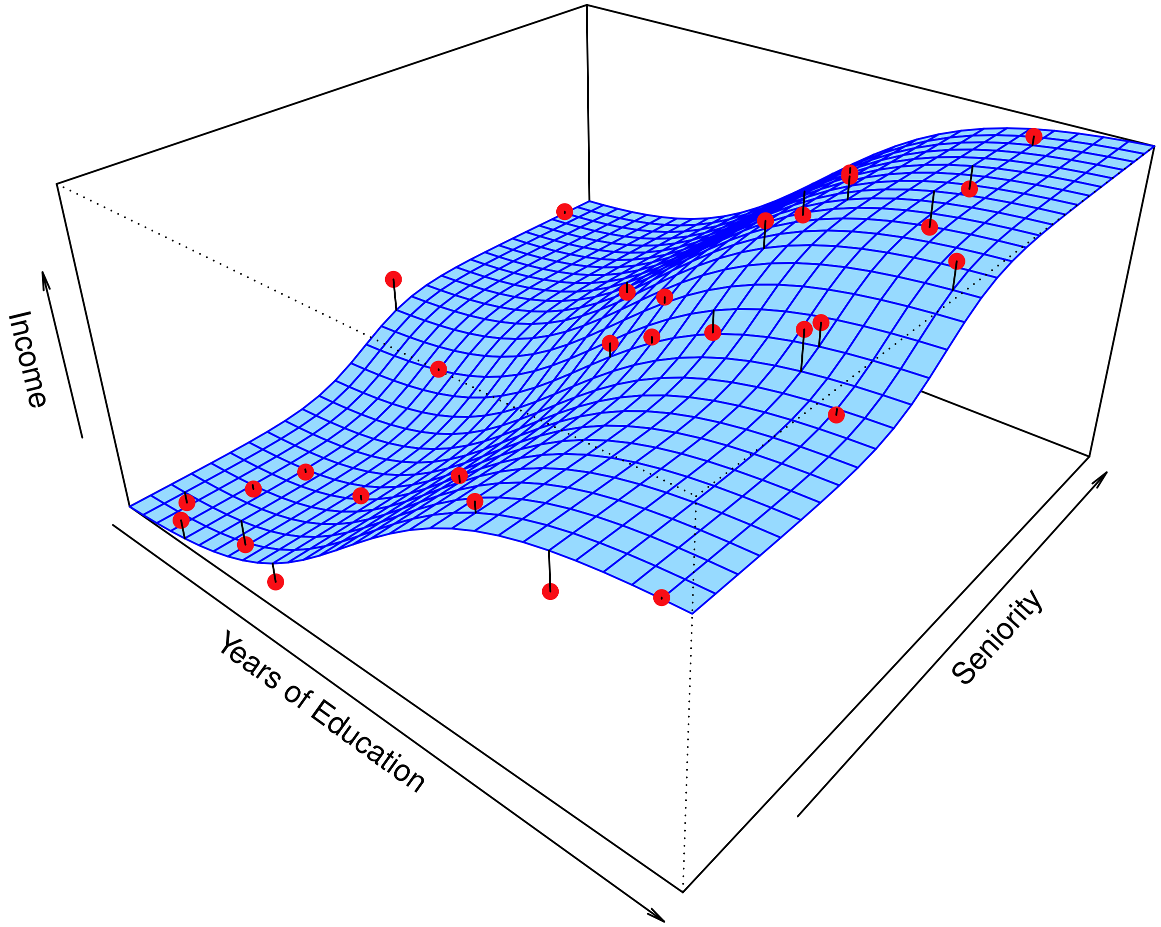
- Can we **predict** **income** using **years of education**?

- Income dataset is simulated, then we know f (blue curve)

- The vertical lines represent the error terms $\epsilon$

# Income prediction

- **Years of education** may be **not enough**
- **Seniority** is another important feature



- Shortly, **machine learning refers to a set of approaches for estimating f**.

# Why estimate f?

- **Prediction**
  - We want to predict Y values at new points X
- **Inference**
  - We want to understand the way Y is affected as X change.

# Prediction

- With a good f we can make **predictions Y** at **new points X**
- Imagine the situation, a set of inputs X are easily available, but the output Y cannot be easily obtained
- We can predict Y using: $\hat{Y} = \hat{f}(X)$

where $\hat{f}$ represents our **estimate** for $f$ , and $\hat{Y}$ represents the **resulting prediction** of $Y$

- Note that, the estimated function $\hat{f}$ is often treated as ***black box***
- We are **not concerned** with its **exact form**, provided it yields **accurate predictions**

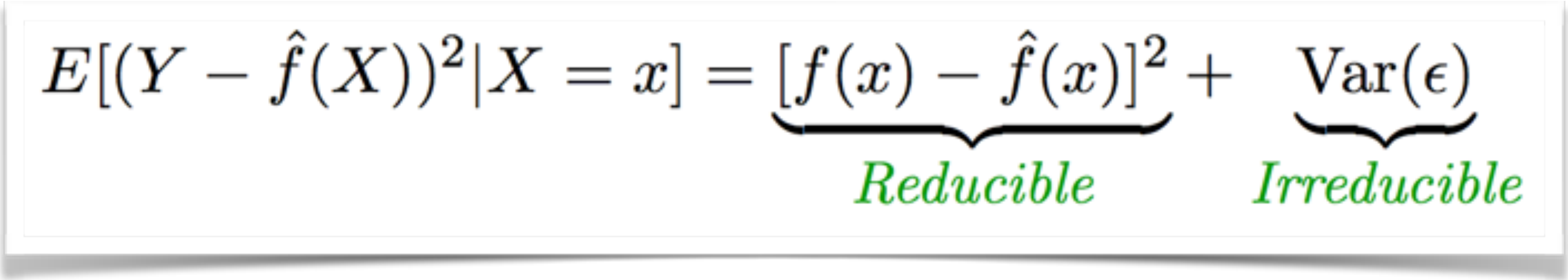
# Prediction: example

- **Input** variables **X**: patient’s blood example
  - easily measured
- **Output** variables **Y**: encodes the patient’s risk for a severe adverse reaction to a particular drug
  - difficultly measured
- If we **correctly predict Y**:
  - we can **avoid** giving the drug at high **risk** of an **adverse reaction** - patients for whom the estimated Y is high

# Errors in predicted: $\hat{Y}$

- Note that, the estimated function $\hat{f}$ is often treated as ***black box***
- We are **not concerned** with its **exact form**, provided it yields **accurate predictions**

# Errors in predicted: $\hat{Y}$



$$E[(Y-\hat{f}(X))^2|X=x] = \underbrace{[f(x)-\hat{f}(x)]^2}_{\mathit{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\mathit{Irreducible}}$$

- *Reducible error*
  - $\hat{f}$ is not a perfect estimate for $f$ , and this inaccuracy introduces some error
- *Irreducible error*
  - If it were possible to form a perfect estimate for $f$ , so that our estimated response took the form $\hat{Y} = f(X)$
  - We still have an error because $Y$ is also function of $\epsilon$ , which cannot be predicted using $X$

# Why estimate f?

- Prediction
  - We want to predict Y values at new points X
- Inference
  - We want to understand the way Y is affected as X change.

# Inference

- We are often interested in **understanding** the **way** that **Y** is **affected** as $\mathbf{X_1, X_2, \ldots, X_p}$

- We want to estimate f, but our goal is **not make features** for Y

- We **need** to **know** the **exact form** of $\hat{f}$

  - Which **features** are **associated** with the **response**?

  - What is the **relationship between** the **response** and each **feature**?

  - Can the **relationship** between **Y** and each **feature** be adequately summarized using a **linear** equation, or is the relationship **more complicated**?

# Why estimate f?

- Prediction
    - We want to predict Y values at new points X
- Inference
    - We want to understand the way Y is affected as X change.

Whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f may be appropriate

# How to estimate f

- We will **always** assume we have **observed** a **set** of ***n*** different data **points**

- These **observations** are called the ***training data*** because we will use them to train, or teach, our method how to estimate f

- Let $\mathbf{x_{ij}}$ represent the value of the ***j*th feature** for **observation *i*,** where $i = 1 \ldots n$ and $j = 1 \ldots p$

- Correspondingly, let $\mathbf{y_i}$ represent the **response** variable for the ***i*th observation**

- Then our **training data** consists of $\mathbf{\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}}$, where $\mathbf{x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T}$

# How to estimate f

- The **goal** is to **apply** a statistical **learning method** to the **training data** to **estimate** the **unknown f**:

$$Y \approx \hat{f}(X) \text{ for any observation } (X, Y)$$

- **Parametric** methods

- **Non-parametric** methods

# Parametric methods

- **Two-step** model-based approach
  - **First**, we make an **assumption** about the functional **form** of **f**
  - For example, f is linear in X

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

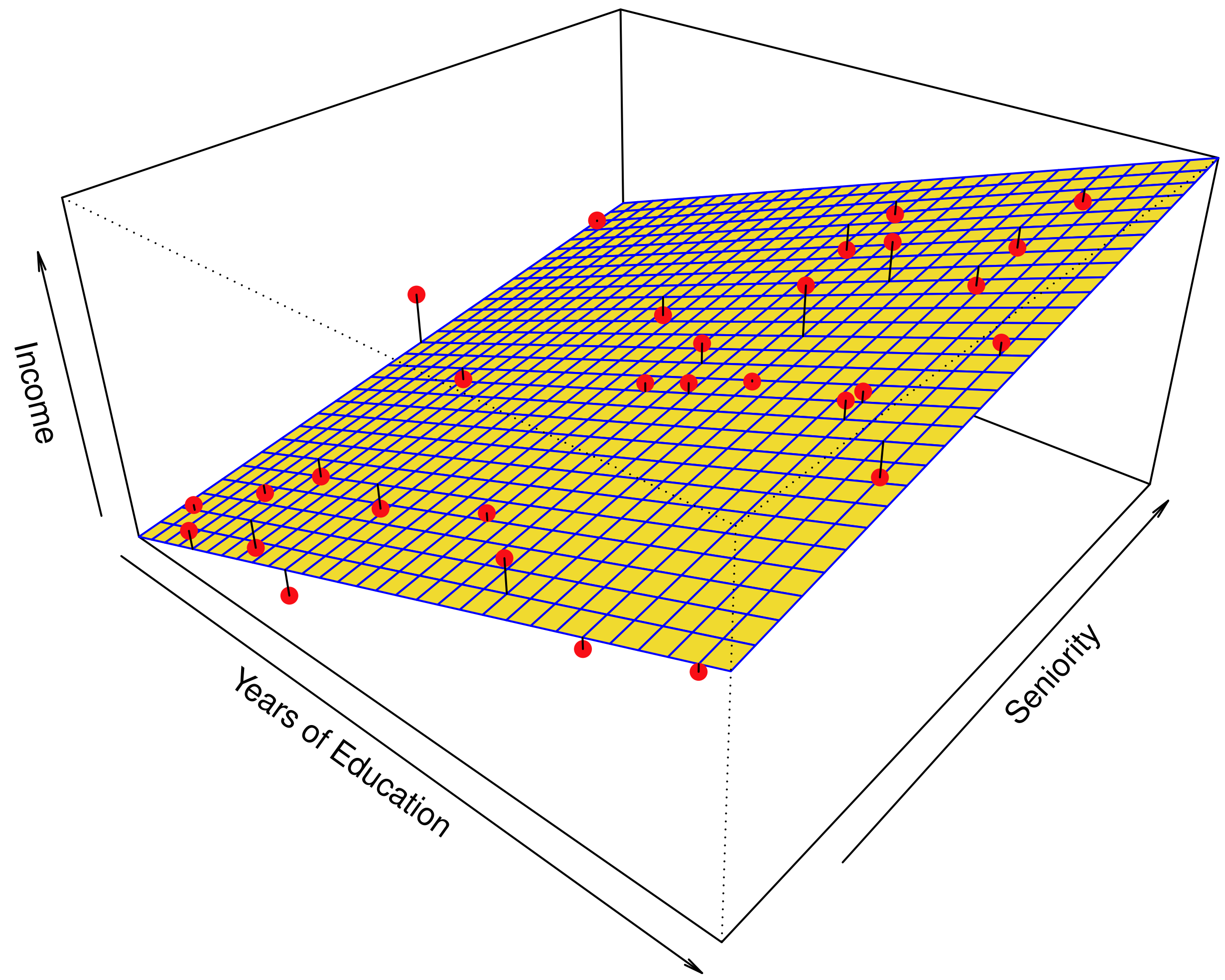
- Given the **selected model**, we need a procedure that **uses** the **training data** to ***fit*** or ***train*** the **model**
  - For example, the ***(ordinary) least squares***
  - We want to find values of parameters such that

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

# Parametric methods

- The model-based approach is referred to as *parametric*
- It **reduces** the **problem** of **estimating f form** to one of **estimating** a **set of parameters**
- Assuming a parametric form for f **simplifies the problem** of estimating f because it is generally much **easier** to estimate a set of **parameters** than it is to fit an entirely **arbitrary function f**

# Parametric methods

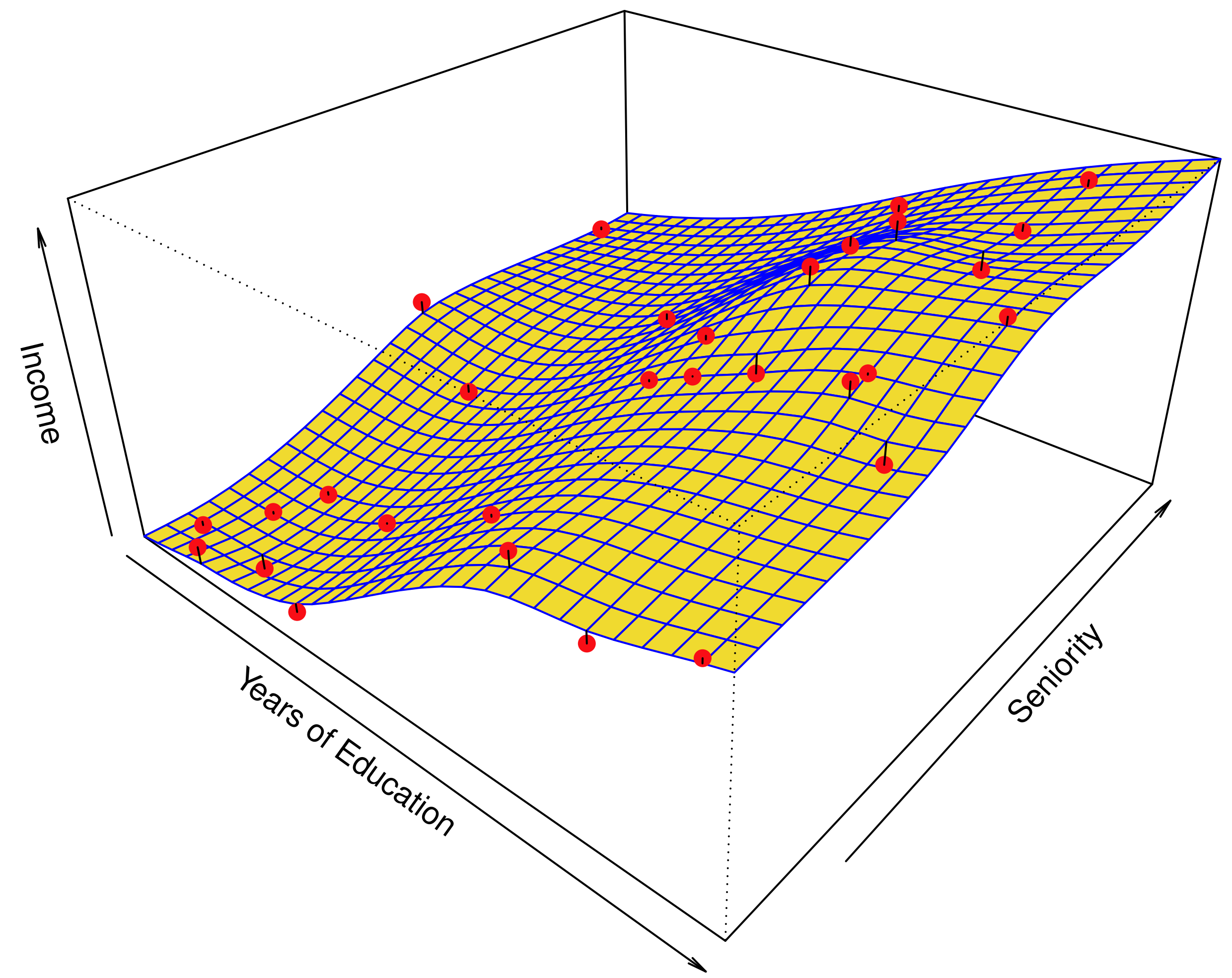


$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}$$

# Non-parametric methods

- Do **not** make **explicit assumptions** about the functional **form of f**
- Seek an **estimated f** that gets **as close to the data points as possible** without being too rigid

# Non-parametric methods



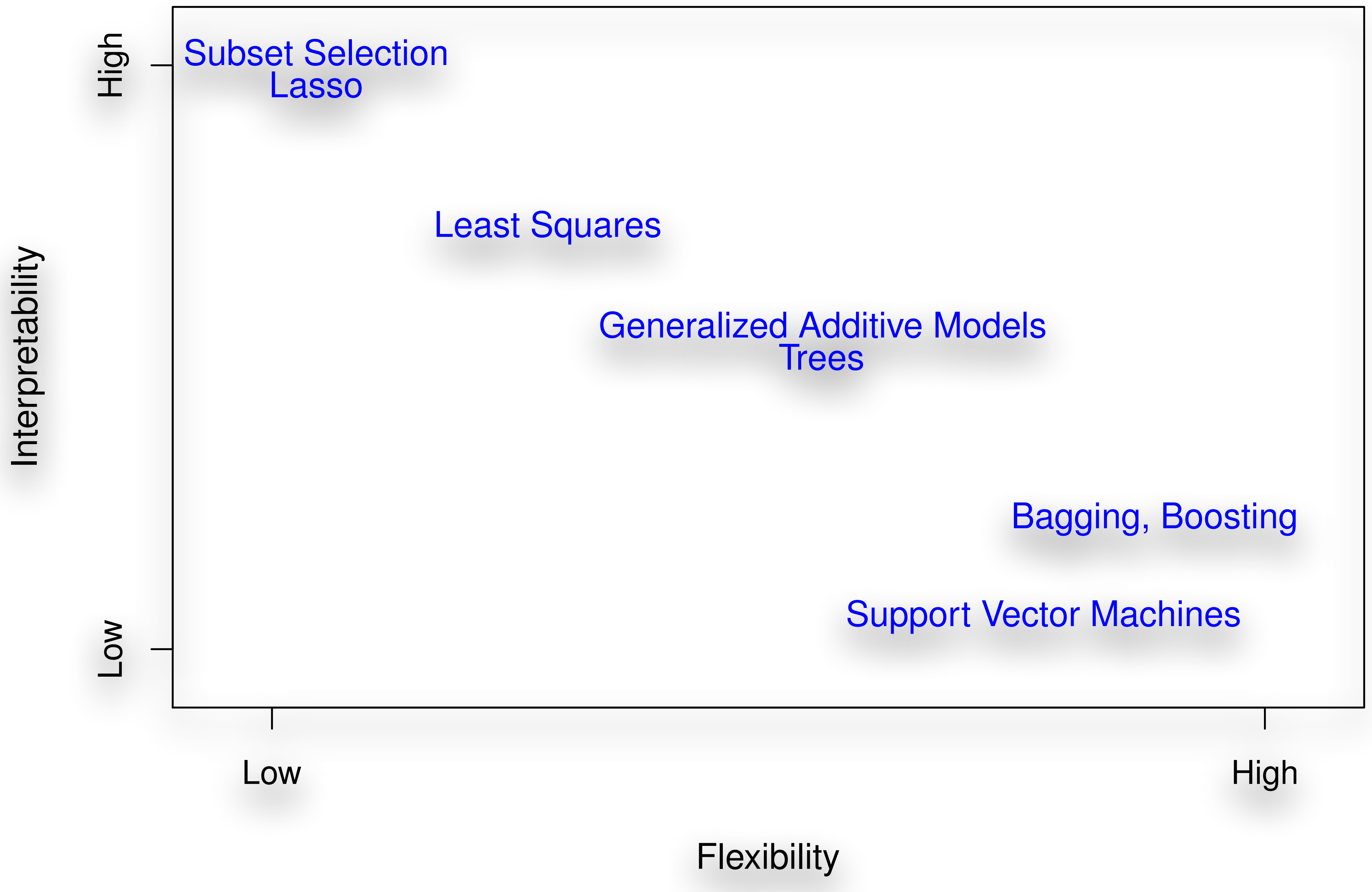
# Parametric vs Non-parametric

| | Pros | Cons |
|---|---|---|
| Parametric | • Reduce the estimation of f to **estimation** of a **set of parameters** | • The chosen model will **usually not match** the true **unknown for of f** |
| Non-parametric | • **No assumption** about the **form of f** | • They do not reduce the estimation of f to a small number of parameters<br>• Requires a **large number of observations** |

# Trade-offs

- Prediction **accuracy** versus **interpretability**
  - **Linear models** are **easy** to **interpret**; **thin-plate splines are not**
- **Good** fit versus **over-fit** or **under-fit**
  - How do we know when the fit is just right?
- **Parsimony** versus **black-box**
  - We often prefer a simpler model involving fewer variables over a black-box predictor involving them all

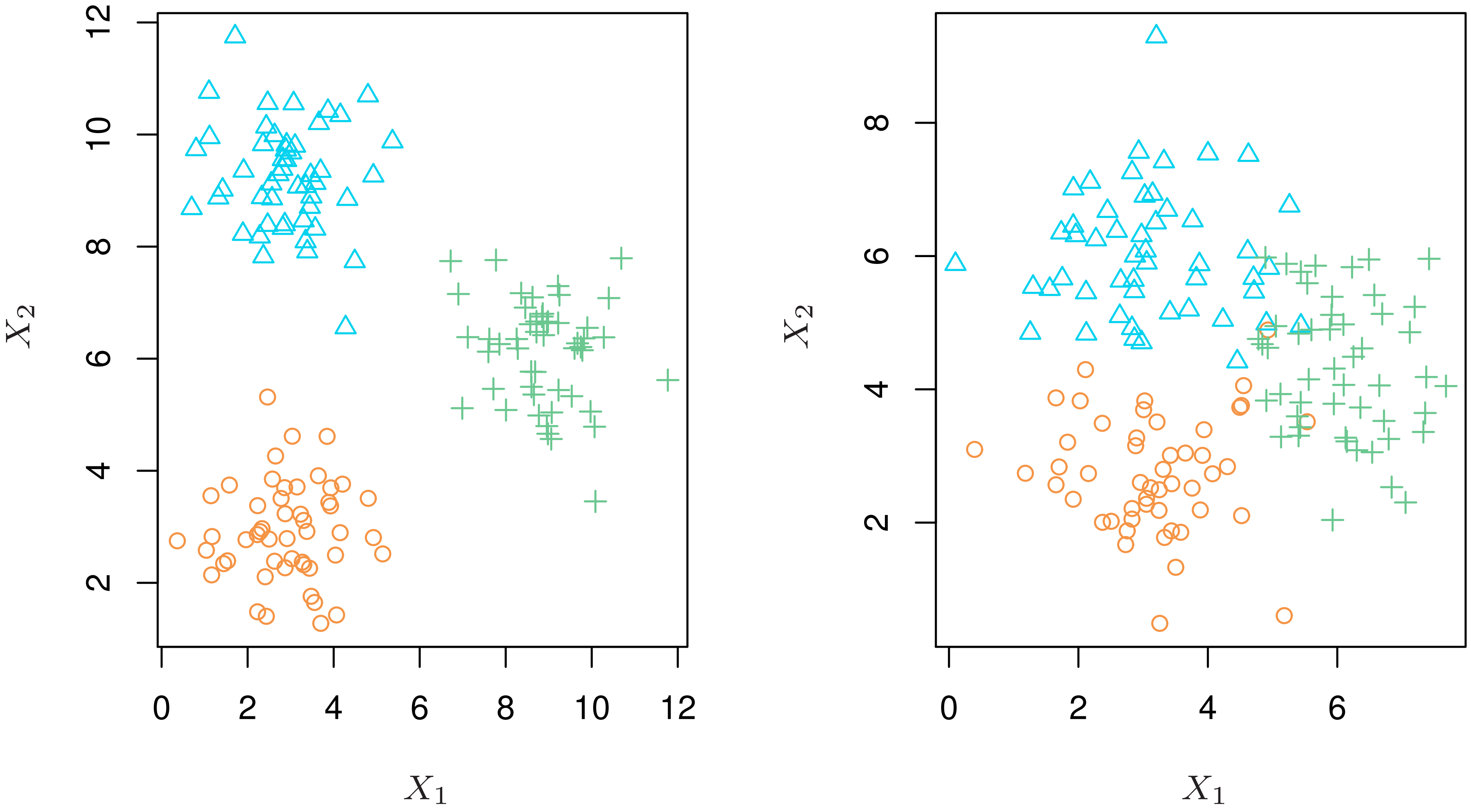
# Trade-offs


High
Low
Interpretability
Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
Low
High
Flexibility

# Supervised vs Unsupervised

- **Supervised**: for each observation there is an associated **label**

- **Unsupervised**: we observe a vector of features but **not** associated **label**



# Regression vs Classification

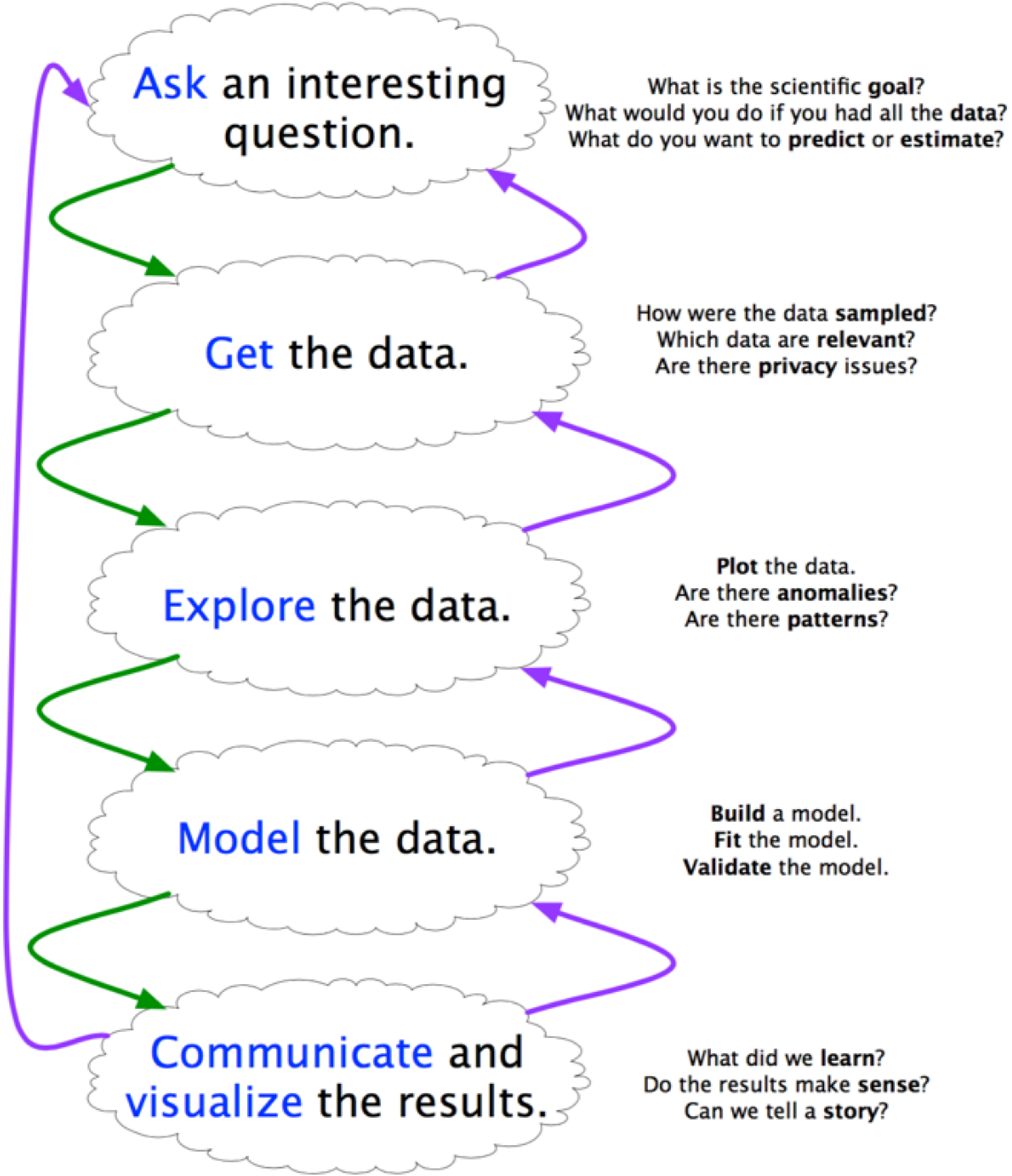
- **Features** can be characterized as either:
  - ***Quantitative****:* numerical values, such as income, the value of a house, the price of a stock, …
  - ***Qualitative*** (also known as *categorical*): person's gender (male or female), the brand of a product, …
- **Qualitative** = **classification**
- **Quantitative** = **regression**

# Assessing model accuracy

- ***There is no free luch in machine learning***:
  - **no method dominates all others over all possible data sets**
- Hence, it is an important **task** to **decide** for any given **set of data** which **methods** produces the **best results**
- Challenge part of machine learning

# Project proposal

- What is the idea of this project?
- Who will participate?
- What data will you use? Will you need time “cleaning up” the data?
- What code will you need to write?
- What existing code are you planning to use?
- What references are relevant? Mention 1-3 related papers
- What are you planning to accomplish by the milestone?
- See suggestions (soon)



Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

**Plot** the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

**Build** a model.
**Fit** the model.
**Validate** the model.

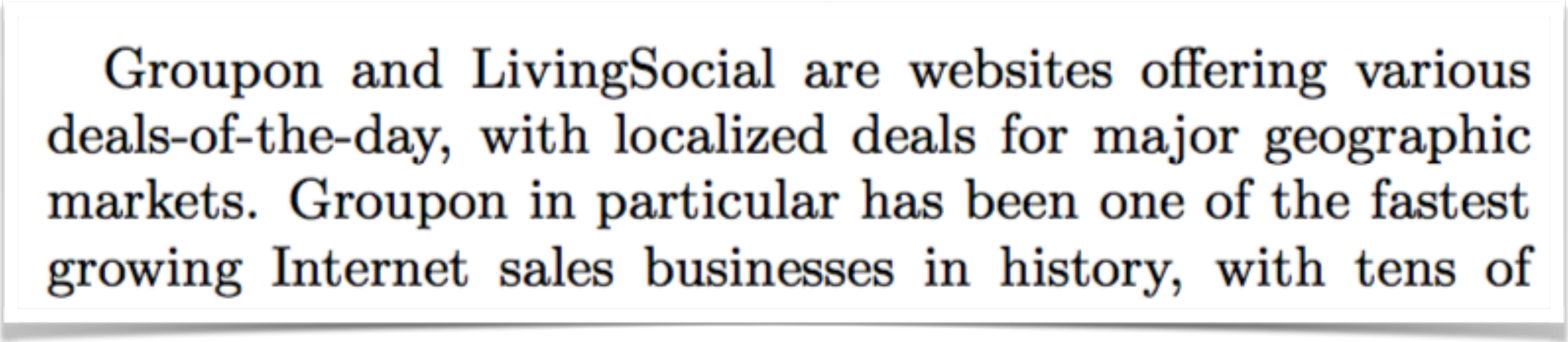
Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

# *Daily Deals: Prediction, Social Diffusion, and Reputational Ramifications*

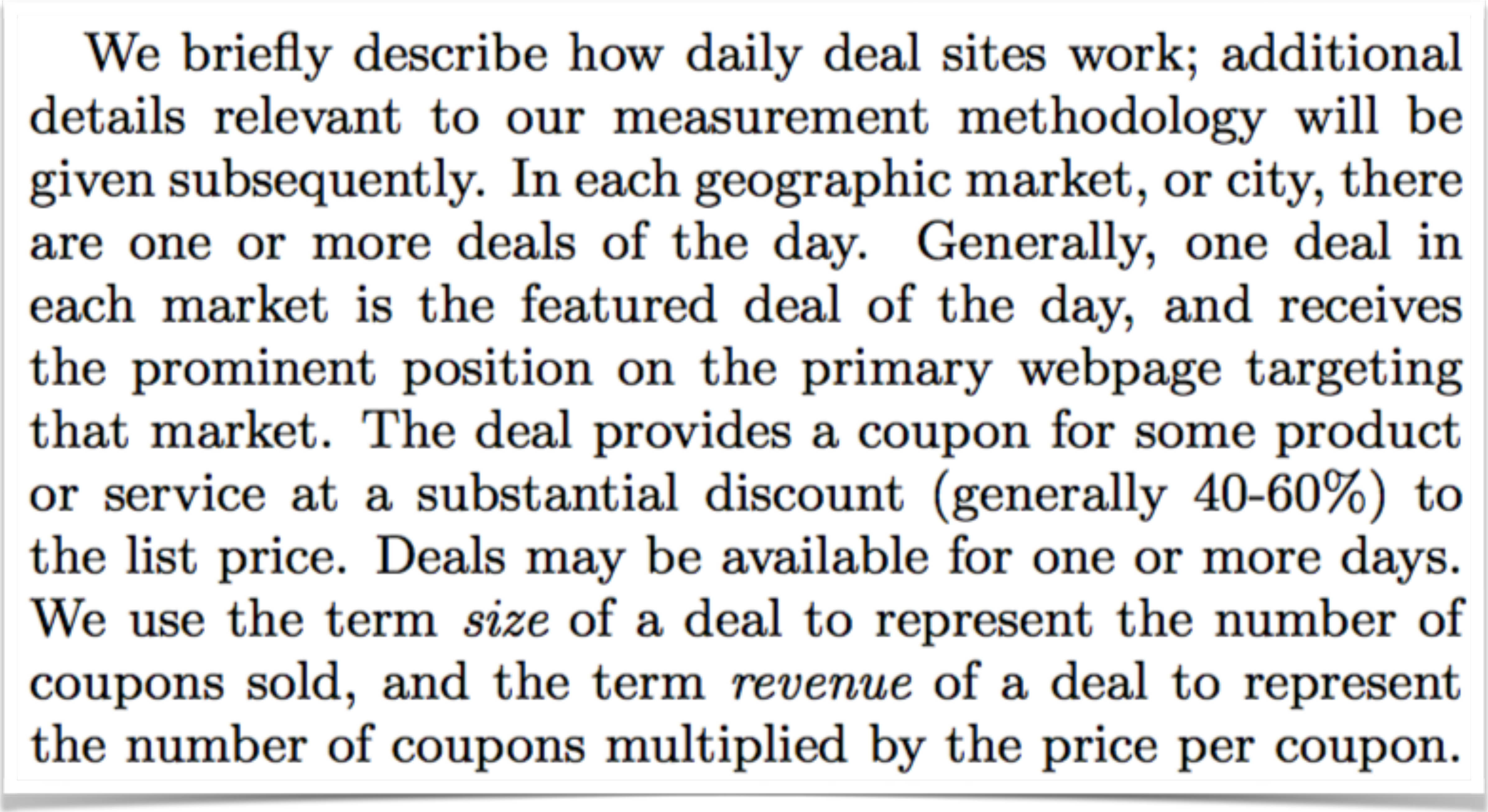
# Daily-deals sites

# **Context:** Daily deals sites



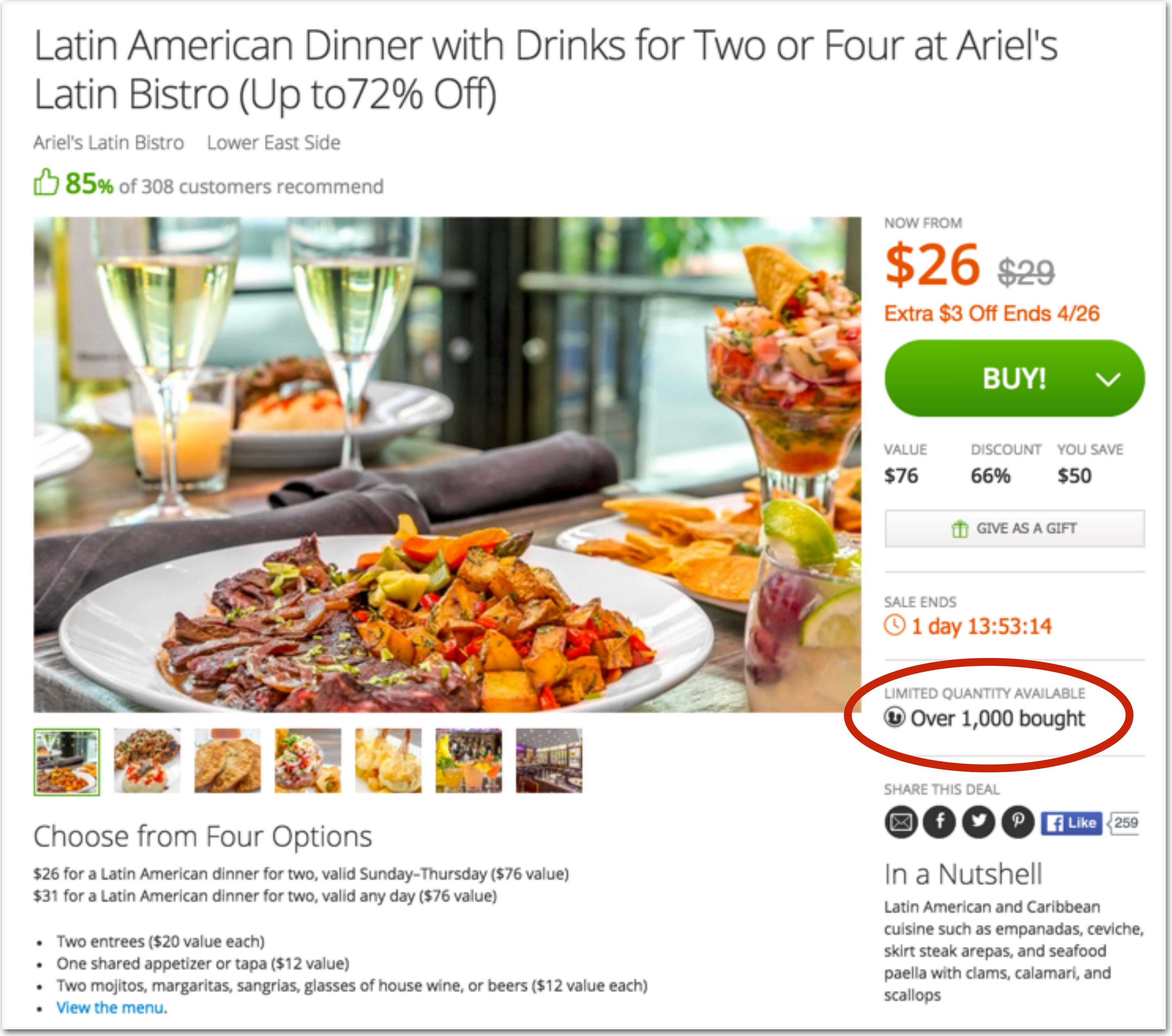
Groupon and LivingSocial are websites offering various deals-of-the-day, with localized deals for major geographic markets. Groupon in particular has been one of the fastest growing Internet sales businesses in history, with tens of

# **Task**: Daily deals sites



We briefly describe how daily deal sites work; additional details relevant to our measurement methodology will be given subsequently. In each geographic market, or city, there are one or more deals of the day. Generally, one deal in each market is the featured deal of the day, and receives the prominent position on the primary webpage targeting that market. The deal provides a coupon for some product or service at a substantial discount (generally 40-60%) to the list price. Deals may be available for one or more days. We use the term *size* of a deal to represent the number of coupons sold, and the term *revenue* of a deal to represent the number of coupons multiplied by the price per coupon.

# Daily deals sites



## Latin American Dinner with Drinks for Two or Four at Ariel's Latin Bistro (Up to72% Off)

Ariel's Latin Bistro    Lower East Side

85% of 308 customers recommend

NOW FROM

$26 ~~$29~~

Extra $3 Off Ends 4/26

| VALUE | DISCOUNT | YOU SAVE |
|---|---|---|
| $76 | 66% | $50 |

SALE ENDS

1 day 13:53:14

LIMITED QUANTITY AVAILABLE

Over 1,000 bought

### Choose from Four Options

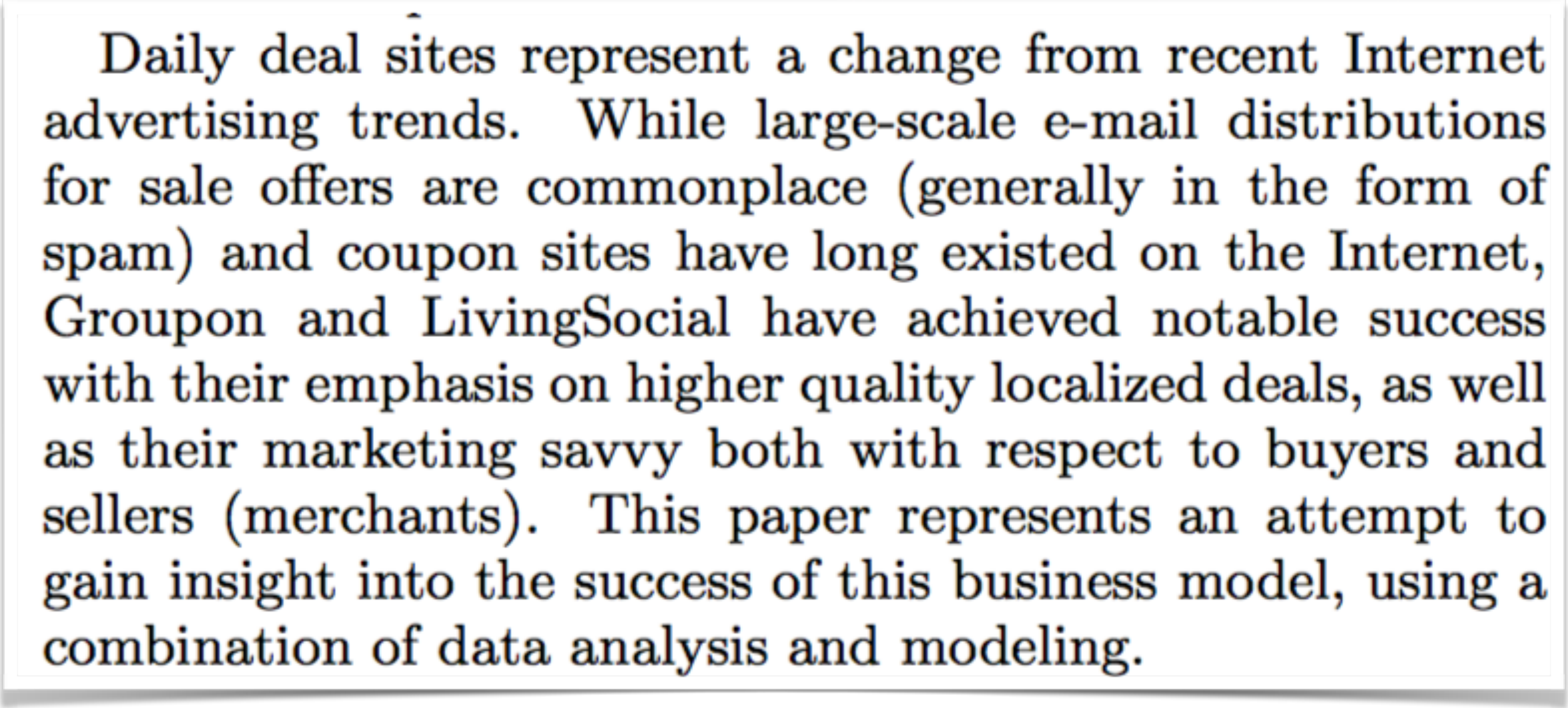
$26 for a Latin American dinner for two, valid Sunday–Thursday ($76 value)
$31 for a Latin American dinner for two, valid any day ($76 value)

- Two entrees ($20 value each)
- One shared appetizer or tapa ($12 value)
- Two mojitos, margaritas, sangrias, glasses of house wine, or beers ($12 value each)
- View the menu.

### In a Nutshell

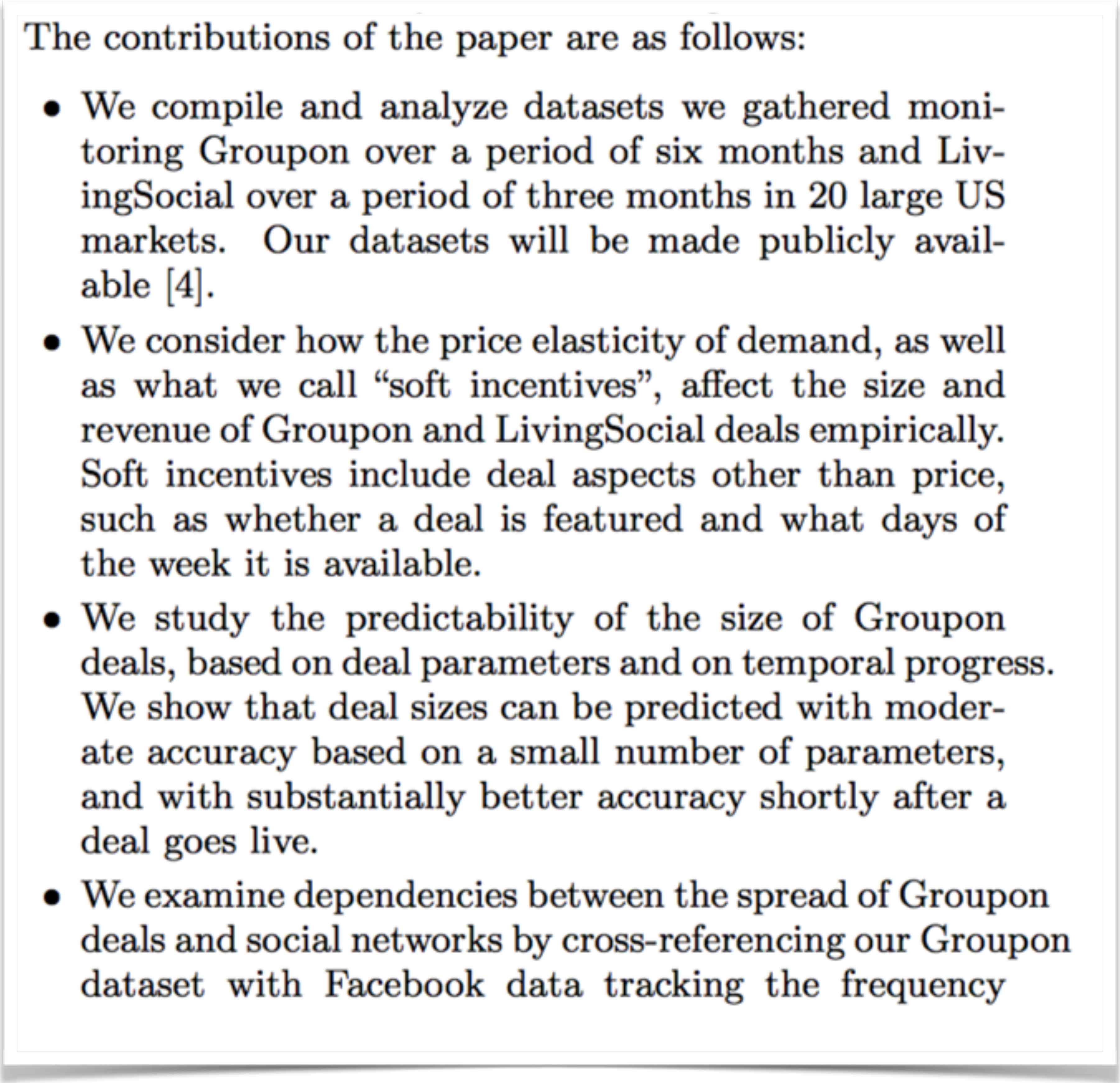
Latin American and Caribbean cuisine such as empanadas, ceviche, skirt steak arepas, and seafood paella with clams, calamari, and scallops

# **Motivation**: Daily deals sites



Daily deal sites represent a change from recent Internet advertising trends. While large-scale e-mail distributions for sale offers are commonplace (generally in the form of spam) and coupon sites have long existed on the Internet, Groupon and LivingSocial have achieved notable success with their emphasis on higher quality localized deals, as well as their marketing savvy both with respect to buyers and sellers (merchants). This paper represents an attempt to gain insight into the success of this business model, using a combination of data analysis and modeling.

# **Contributions**: Daily deals sites



The contributions of the paper are as follows:

- We compile and analyze datasets we gathered monitoring Groupon over a period of six months and LivingSocial over a period of three months in 20 large US markets. Our datasets will be made publicly available [4].
- We consider how the price elasticity of demand, as well as what we call "soft incentives", affect the size and revenue of Groupon and LivingSocial deals empirically. Soft incentives include deal aspects other than price, such as whether a deal is featured and what days of the week it is available.
- We study the predictability of the size of Groupon deals, based on deal parameters and on temporal progress. We show that deal sizes can be predicted with moderate accuracy based on a small number of parameters, and with substantially better accuracy shortly after a deal goes live.
- We examine dependencies between the spread of Groupon deals and social networks by cross-referencing our Groupon dataset with Facebook data tracking the frequency

# Introduction

- Context (recommender systems, etc…)
- Define your task
    - top-N movie recommendation
    - given a conference and a institution, predict the number of papers that will be published
    - given an author, describe its publications
- Motivation (why this problem is important?)
- Challenge (why this problem is difficult?)
- Contributions (what we did?)

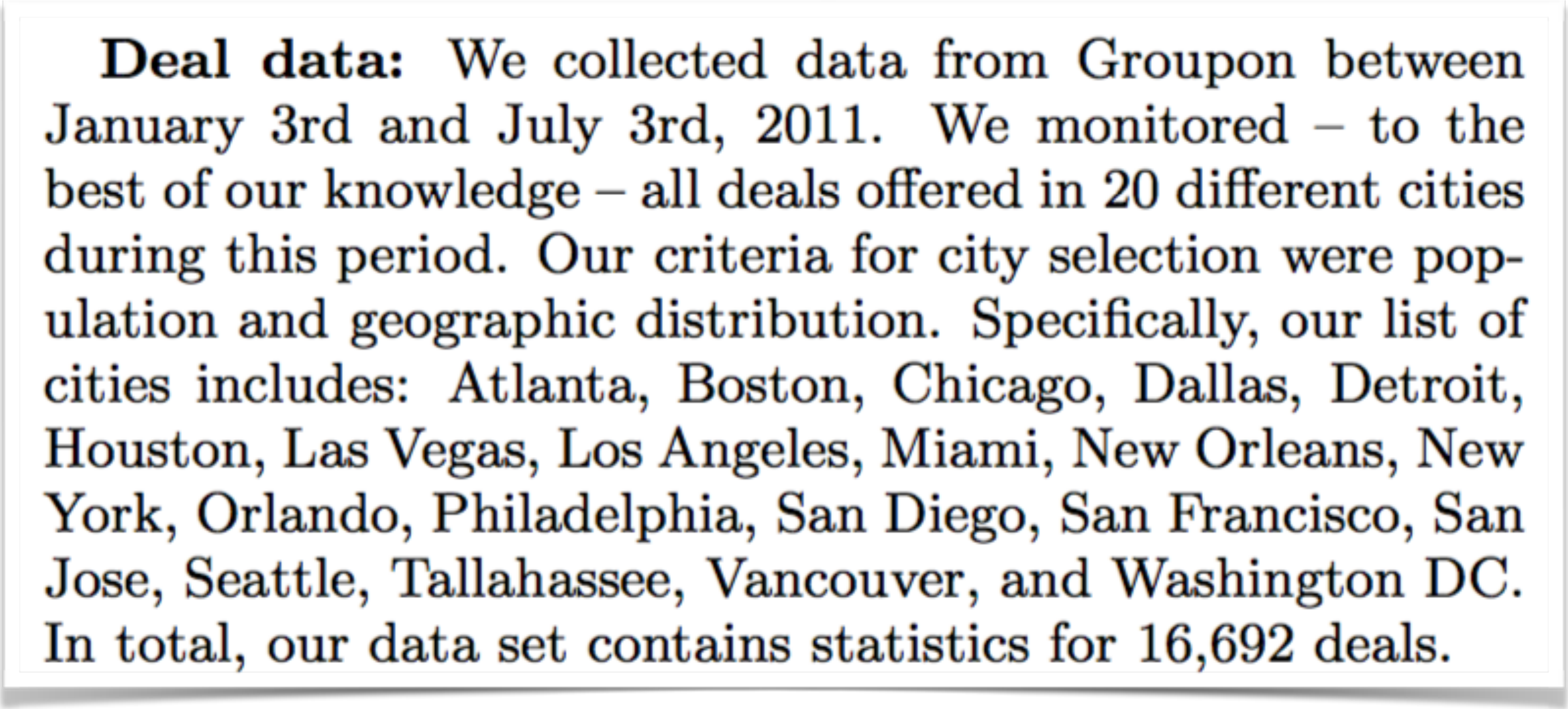
# Problem

- Tasks:
    - Analyze datasets
    - Predict the number of coupons sold based on
        - small number of parameters
        - temporal progress

# Analyze the dataset

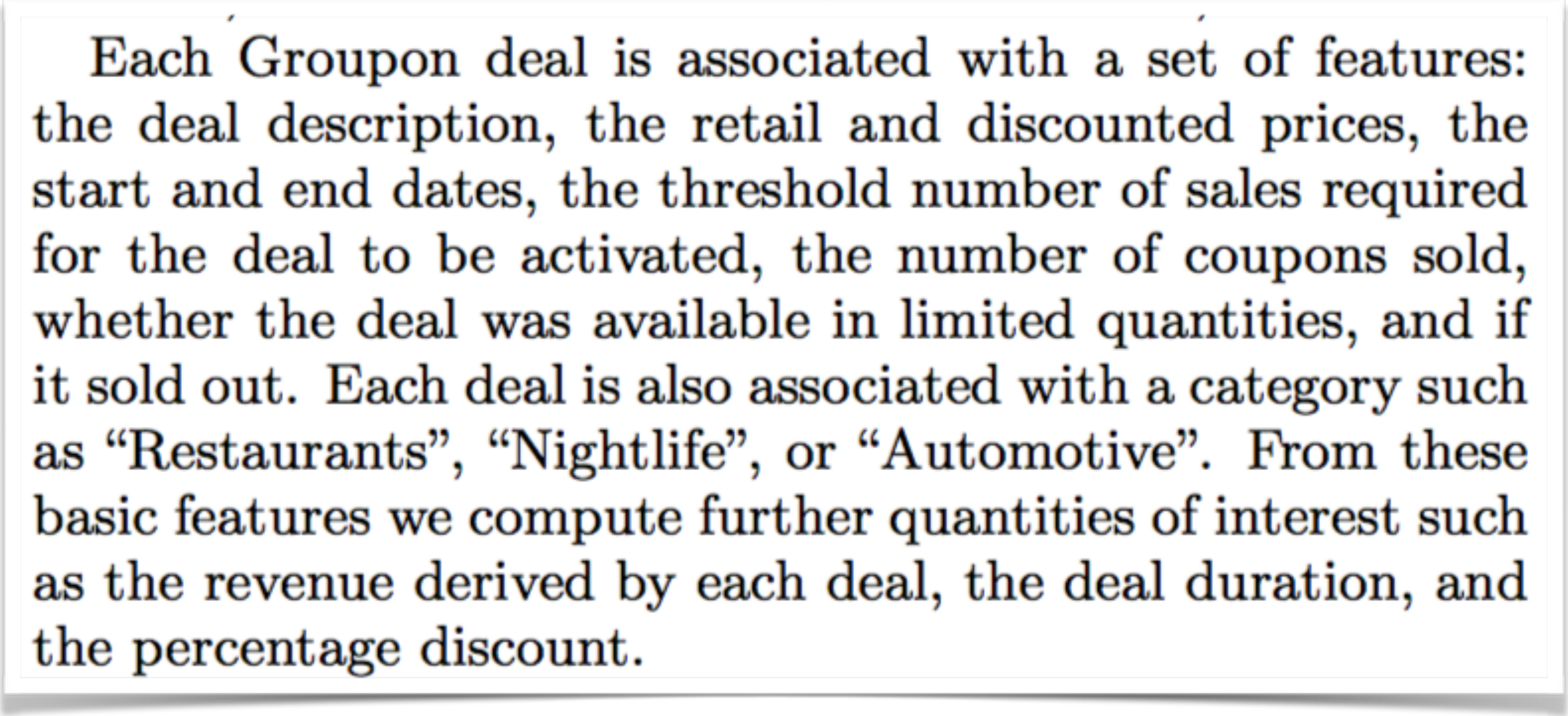
- **Understand  and Explain your data!**
  - Which site? (e.g. Movielens/IMDB)
  - Which conference? (e.g. NIPS)
  - # instances (e.g., users, items, categories, venues, authors, documents, papers)
  - mean number of: (i) items per category, (ii) papers per conference, (iii) authors per paper, etc
  - other measures relevant to your data…

# Describe the dataset



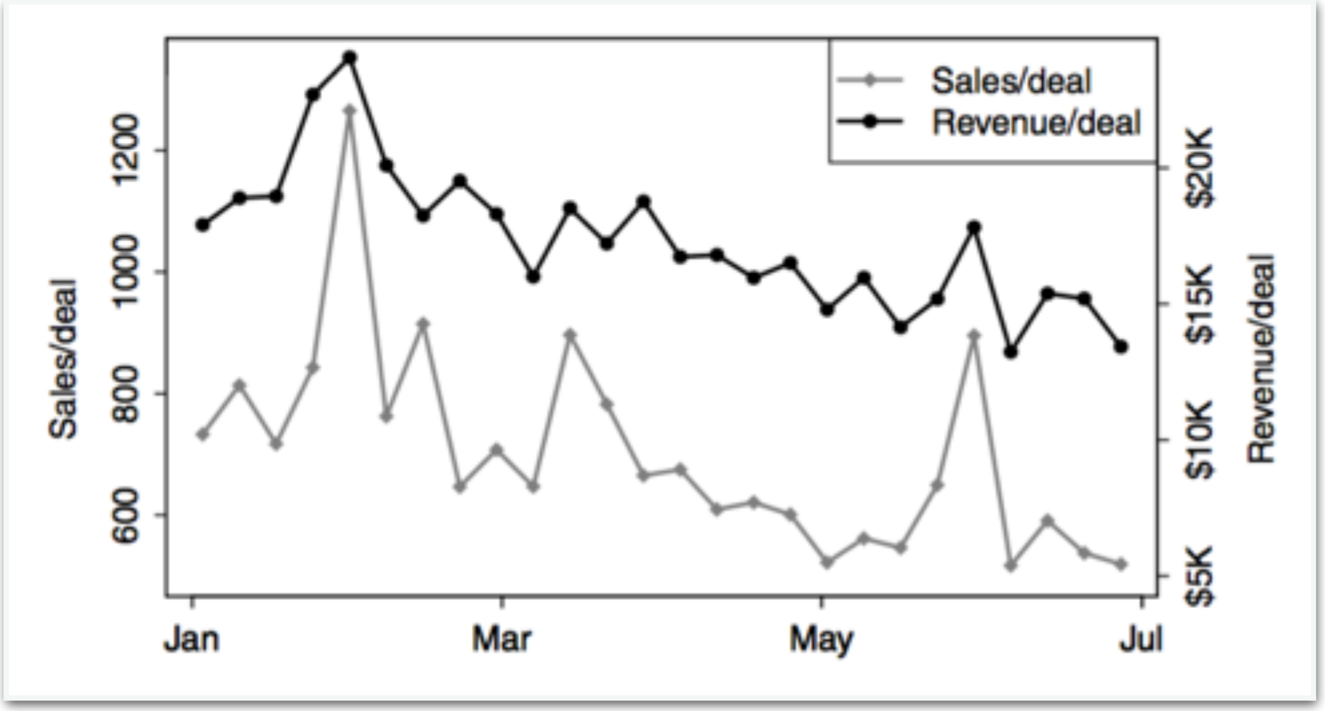
**Deal data:** We collected data from Groupon between January 3rd and July 3rd, 2011. We monitored – to the best of our knowledge – all deals offered in 20 different cities during this period. Our criteria for city selection were population and geographic distribution. Specifically, our list of cities includes: Atlanta, Boston, Chicago, Dallas, Detroit, Houston, Las Vegas, Los Angeles, Miami, New Orleans, New York, Orlando, Philadelphia, San Diego, San Francisco, San Jose, Seattle, Tallahassee, Vancouver, and Washington DC. In total, our data set contains statistics for 16,692 deals.

# Describe the dataset

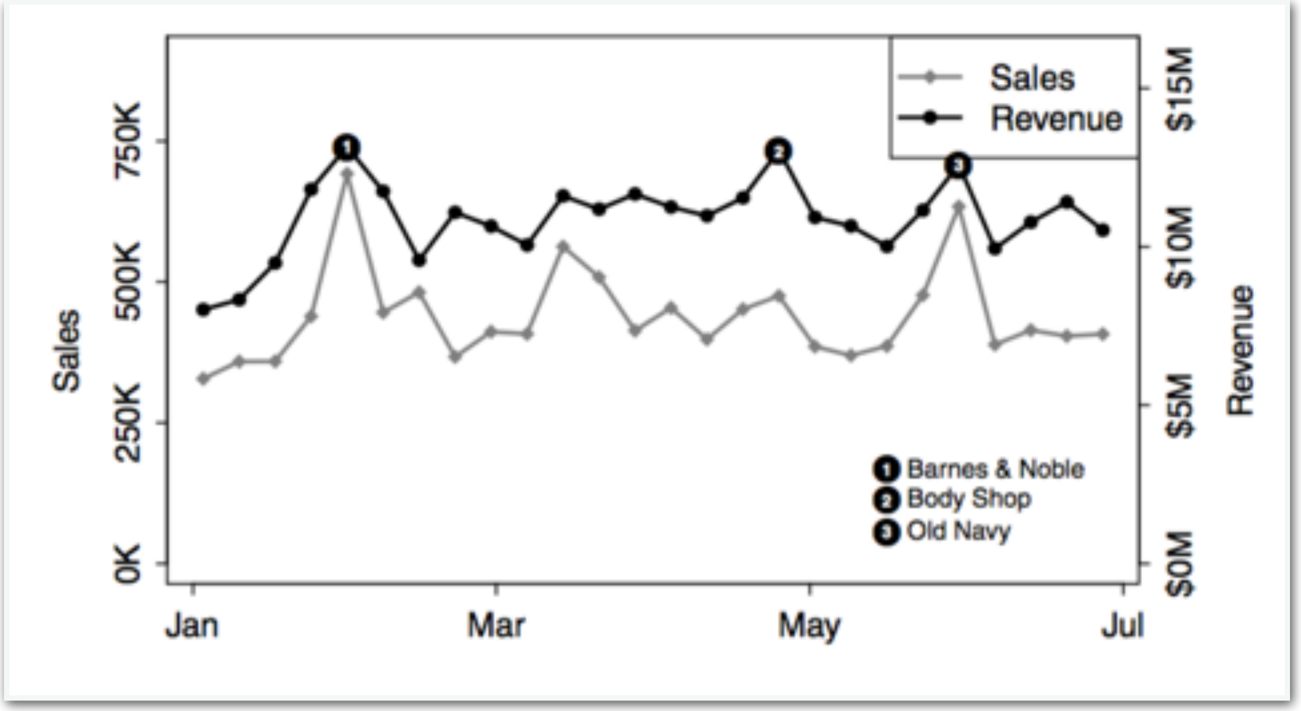


Each Groupon deal is associated with a set of features: the deal description, the retail and discounted prices, the start and end dates, the threshold number of sales required for the deal to be activated, the number of coupons sold, whether the deal was available in limited quantities, and if it sold out. Each deal is also associated with a category such as "Restaurants", "Nightlife", or "Automotive". From these basic features we compute further quantities of interest such as the revenue derived by each deal, the deal duration, and the percentage discount.

# Analyze the dataset

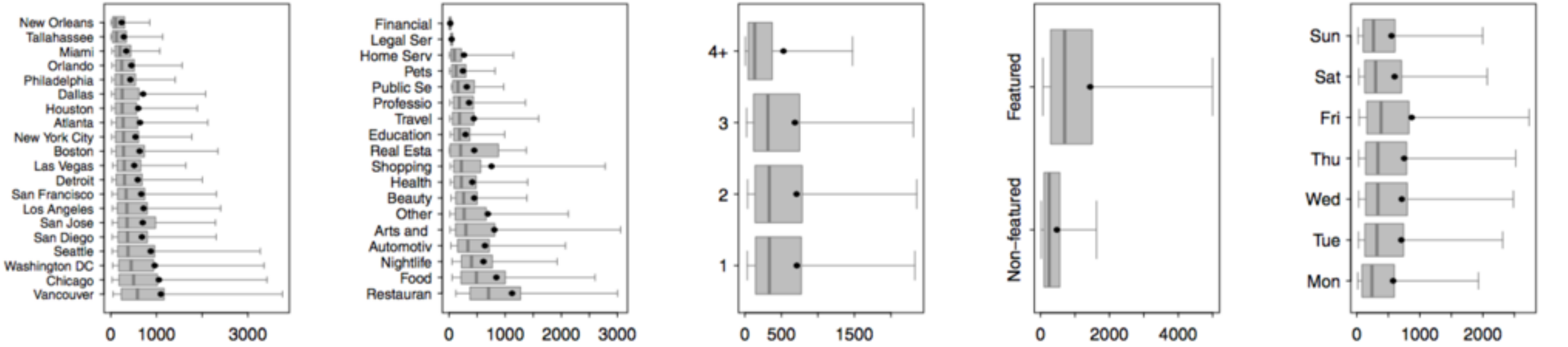


Weekly revenue and sales
in 20 selected cities



Revenue and coupons sold
per deal week-over-week

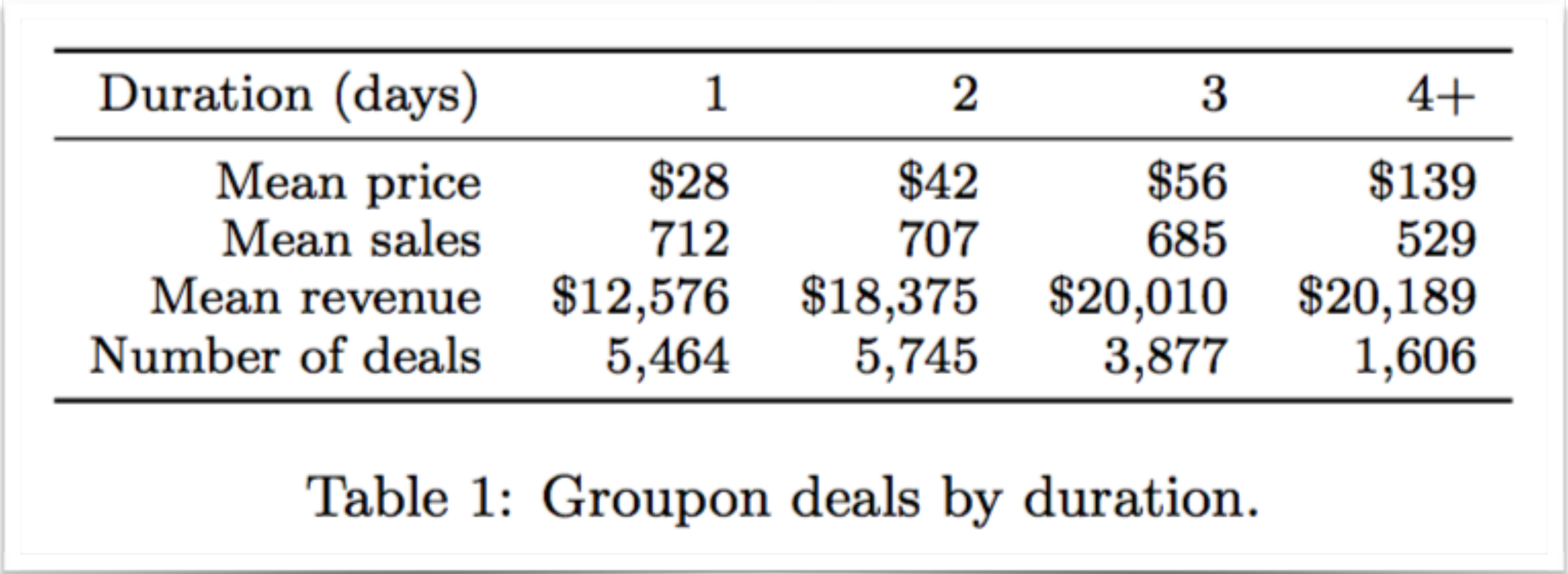
# Analyze the dataset



(a) By city (b) By category (c) By duration (days) (d) By placement (e) By start weekday

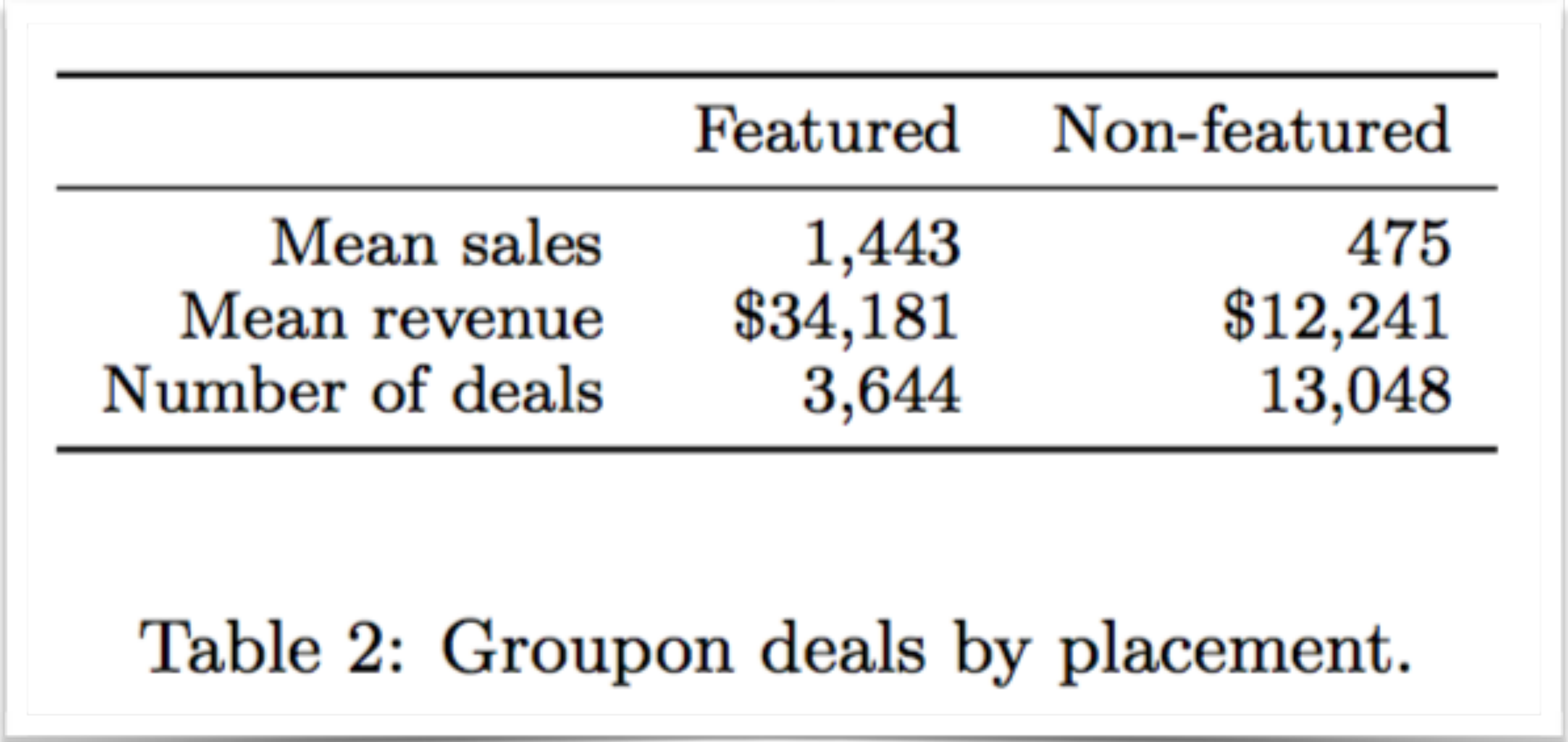
Groupon deal sizes on the x-axis

# Analyze the dataset



| Duration (days) | 1 | 2 | 3 | 4+ |
|---|---|---|---|---|
| Mean price | $28 | $42 | $56 | $139 |
| Mean sales | 712 | 707 | 685 | 529 |
| Mean revenue | $12,576 | $18,375 | $20,010 | $20,189 |
| Number of deals | 5,464 | 5,745 | 3,877 | 1,606 |

Table 1: Groupon deals by duration.



| | Featured | Non-featured |
|---|---|---|
| Mean sales | 1,443 | 475 |
| Mean revenue | $34,181 | $12,241 |
| Number of deals | 3,644 | 13,048 |

Table 2: Groupon deals by placement.

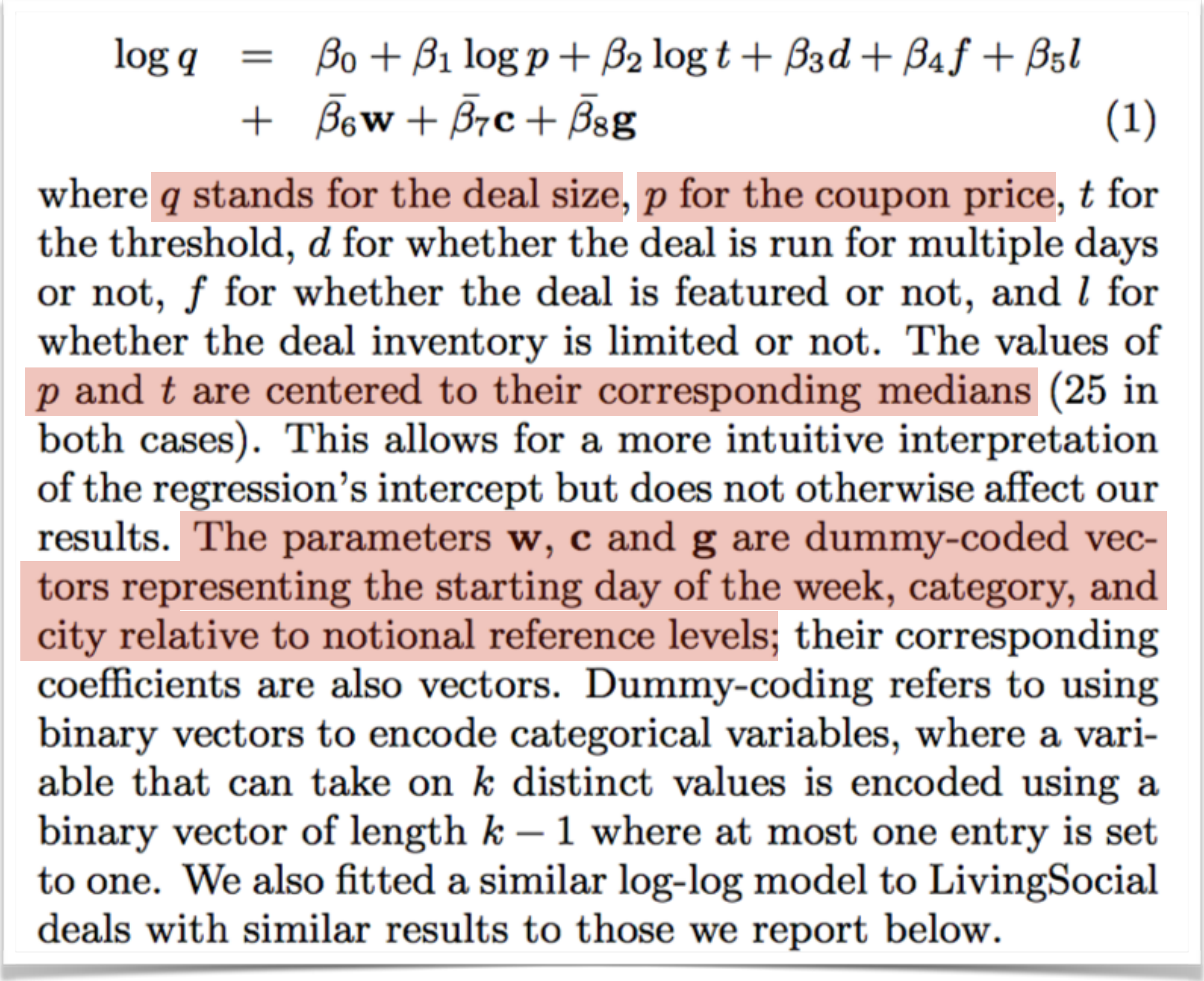
# Problem

- Tasks:
  - Analyze datasets
  - **Predict the number of coupons sold based on**
    - small number of parameters
    - temporal progress

# Task

- Model
- Metric(s)
- Results

# Model (prediction)



$$
\begin{aligned}
\log q \quad &= \quad \beta_0 + \beta_1 \log p + \beta_2 \log t + \beta_3 d + \beta_4 f + \beta_5 l \\
&+ \quad \bar{\beta}_6 \mathbf{w} + \bar{\beta}_7 \mathbf{c} + \bar{\beta}_8 \mathbf{g} \qquad (1)
\end{aligned}
$$

where $q$ stands for the deal size, $p$ for the coupon price, $t$ for the threshold, $d$ for whether the deal is run for multiple days or not, $f$ for whether the deal is featured or not, and $l$ for whether the deal inventory is limited or not. The values of $p$ and $t$ are centered to their corresponding medians (25 in both cases). This allows for a more intuitive interpretation of the regression's intercept but does not otherwise affect our results. The parameters **w**, **c** and **g** are dummy-coded vectors representing the starting day of the week, category, and city relative to notional reference levels; their corresponding coefficients are also vectors. Dummy-coding refers to using binary vectors to encode categorical variables, where a variable that can take on $k$ distinct values is encoded using a binary vector of length $k - 1$ where at most one entry is set to one. We also fitted a similar log-log model to LivingSocial deals with similar results to those we report below.

# Model

- Regression
- Classification
- Clustering
- Recommender systems
- **Use a figure to clarify your proposal!**

# Results

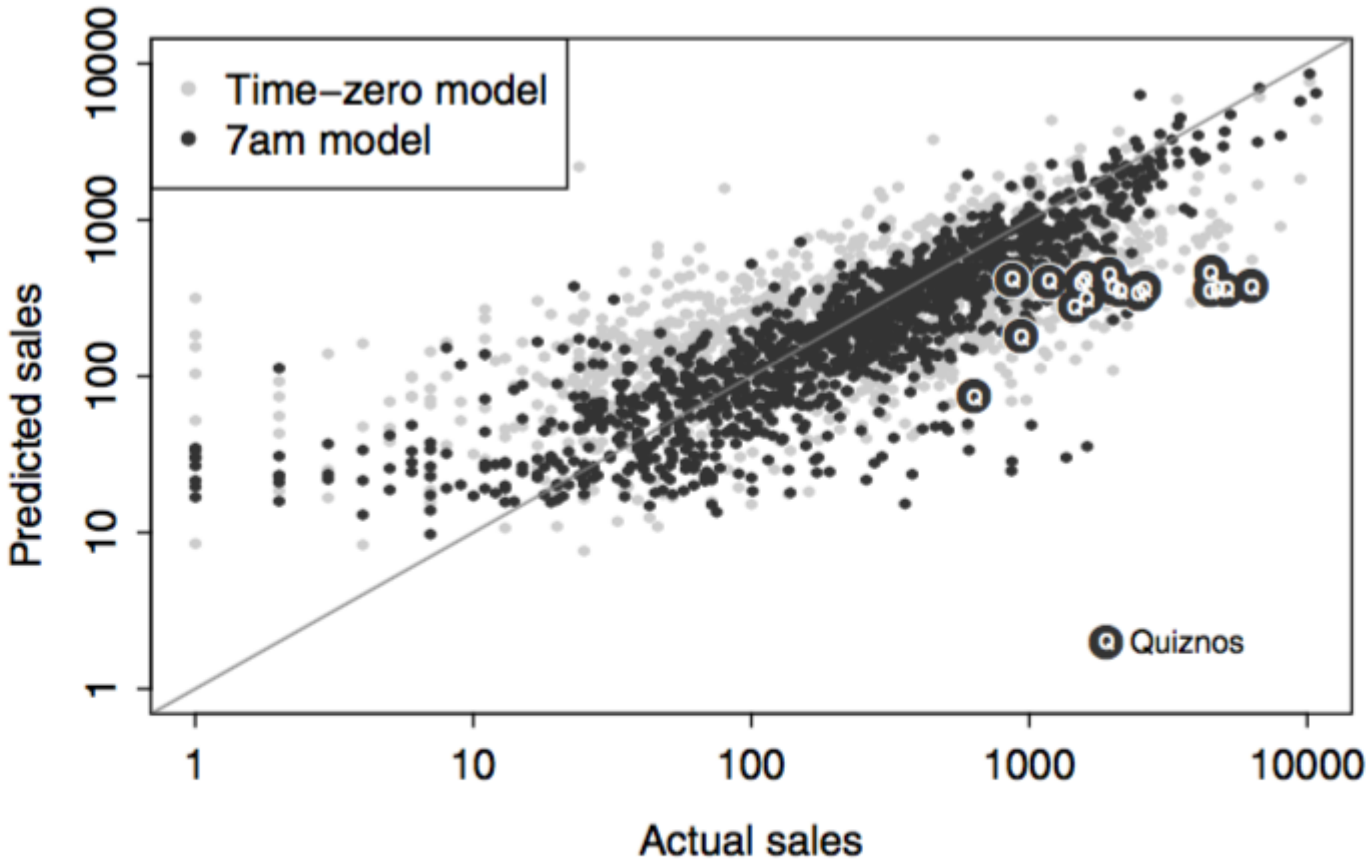
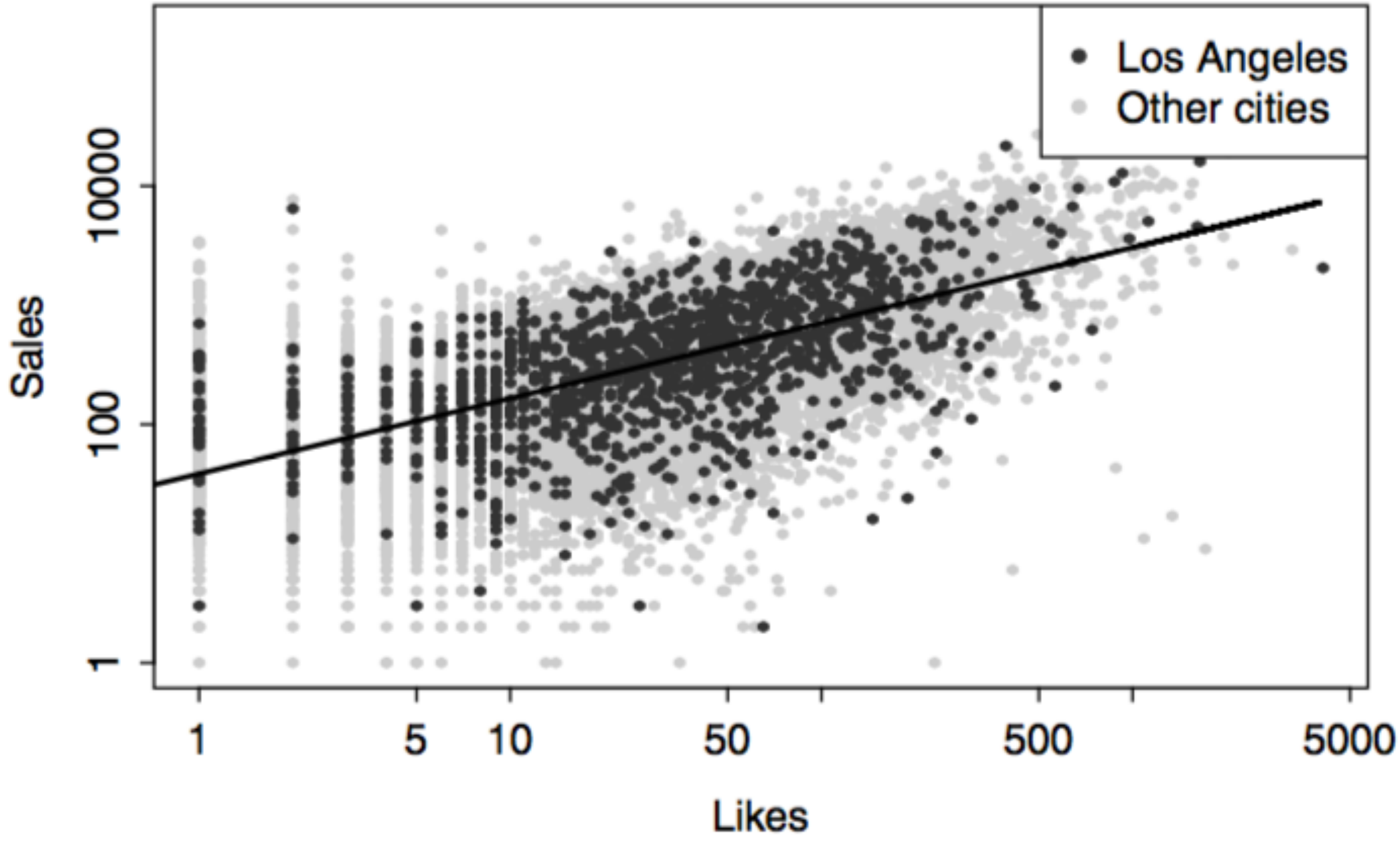


Figure 5: Actual vs. predicted sales for a test set of Groupon deals, in log-log scale for our plain regression model, as well as the the model incorporating early morning sales.

# Results

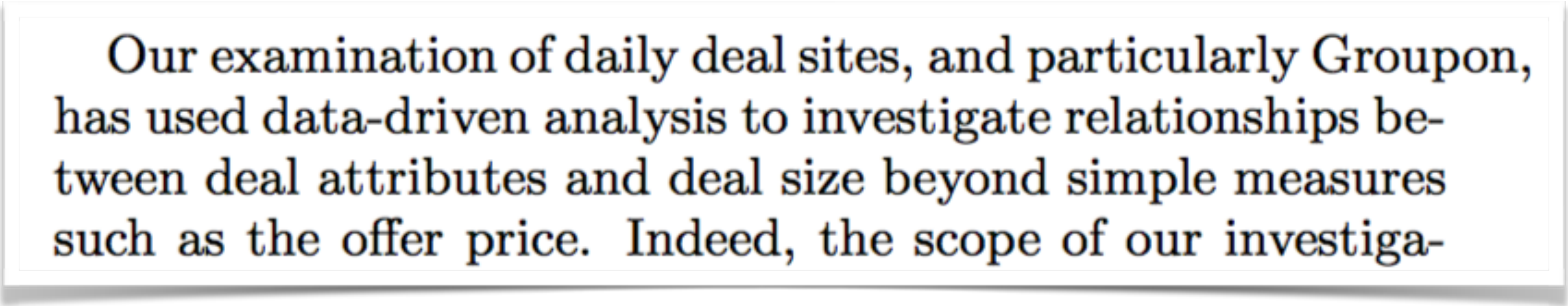


(a) Groupon, slope = 0.63

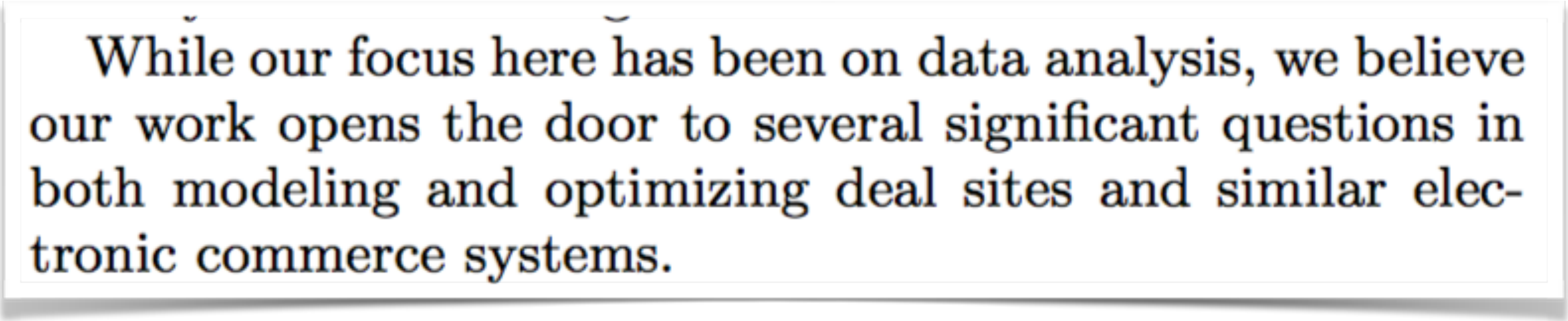
# Results

- Graphs
- Tables

# Conclusion and Future Work



Our examination of daily deal sites, and particularly Groupon, has used data-driven analysis to investigate relationships between deal attributes and deal size beyond simple measures such as the offer price. Indeed, the scope of our investiga-



While our focus here has been on data analysis, we believe our work opens the door to several significant questions in both modeling and optimizing deal sites and similar electronic commerce systems.

# An Empirical Comparison of Supervised Learning Algorithms

# Empirical comparison

- Supervised learning
- 10 algorithms
- Several metrics
  - Accuracy, F-score, Lift, ROC Area, average precision,…
- 11 binary classification problems (datasets)
- “Performance by Metric”
- “Performance by Problem”
- “Bootstrap analysis”

# Summary

- Introduction
- Modelling
- Results
- Conclusion and Future Work