

# Unique in the data

---

Unique in the Crowd: The privacy bounds of human mobility

Unique in the Shopping Mall: On the reidentifiability of credit card

# Roteiro

A seguir...

- Apresentação
  - Datasets
  - Metodologia
  - Resultados
-

# Apresentação

- Yves-Alexandre de Montjoye
- Os artigos seguem a mesma linha de pesquisa
  - Analisar a unicidade dos dados em um dataset.
- Unicidade é o risco de re-identificação de um usuário em um dataset anônimo
  - A partir de dados externos.
- Os datasets utilizados são anônimos
  - Nenhum identificador: nome, endereço, números...

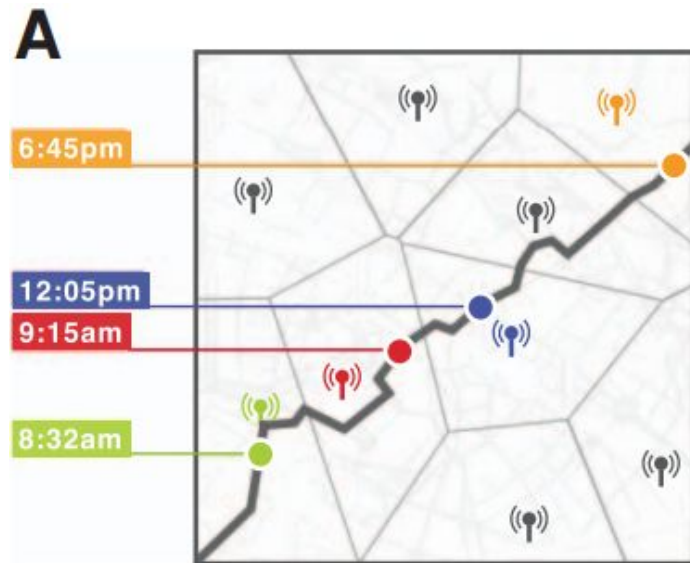
# Dataset

Unique in the Crowd

---

# Dataset

- Dados de mobilidade do usuários
- ~1.5 milhão de usuários de uma operadora telefônica
- Dados coletados entre o período de Abril 2006 a Junho 2007
- Pontos espaço-temporais:
  - Hora e localização da antena.
- Diagrama de Voronoi
  - A metade da distância máxima entre as antenas.



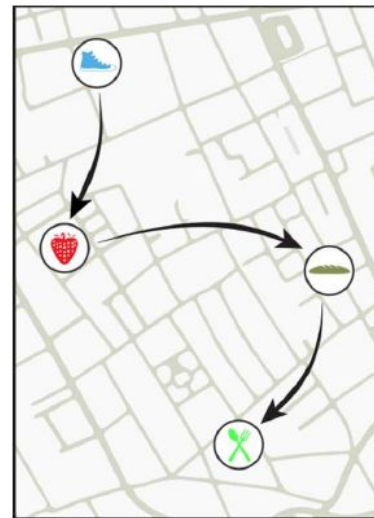
# Dataset

**Unique in the Shopping Mall**

---

# Dataset

- Dados de transações com cartão de crédito
- ~1.1 milhão de pessoas de um país da OECD<sup>1</sup>
  - Gênero (24% mulheres)
  - Renda (39% baixa, 35% média, 22% alta, 4% UNK)
- Dados coletados entre o período de Janeiro a Março
- Pré-processamento:
  - 138 transações com valores maiores que \$22.800
- Pontos espaço-temporais:
  - Dia e loja (tupla)
  - Dia, loja e preço (tripla)\*

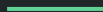


<sup>1</sup>Organisation for Economic Co-operation and Development (Organização para a Cooperação e Desenvolvimento Econômico)

# Metodologia

Avaliar a unicidade  $\varepsilon$

Utilizado em ambos



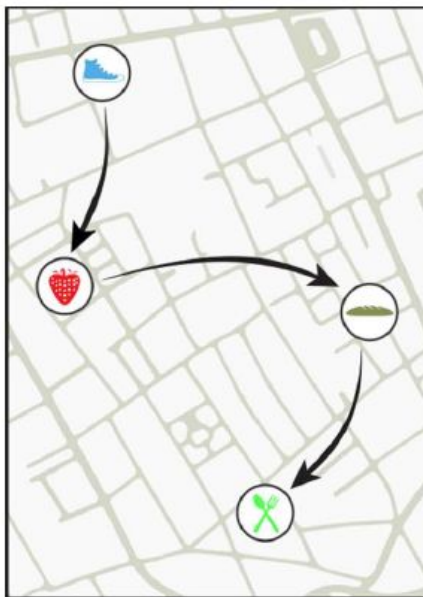


# Metodologia

- Dataset (D)
- Dado um conjunto de  $p$  pontos espaço-temporais  $I_p$ 
  - Selecionados aleatoriamente e previamente conhecidos
- Para cada usuário, extrair de D um subconjunto  $S(I_p)$ 
  - Que contém exatamente os  $p$  pontos de  $I_p$
- Um usuário é identificado se  $|S(I_p)| = 1$ 
  - Ou seja, se apenas ele corresponde ao “rastro” de  $p$  pontos
- $\epsilon_p$  é a porcentagem de usuários identificáveis com  $p$  pontos

# Exemplo<sup>1</sup>

- Sabemos que João:
  - Frequentou uma padaria em 23/09
  - Frequentou um restaurante em 24/09
- Dado 2 pontos  $p$ , apenas um subconjunto  $S(l_p)$ 
  - $|S(l_p)| = 1$
- Logo...
  - 7abc1a23 = João
- Agora sabemos tudo...



shop	user_id	time	price	price_bin
	7abc1a23	09/23	\$97.30	\$49 – \$146
	7abc1a23	09/23	\$15.13	\$5 – \$16
	3092fc10	09/23	\$43.78	\$16 – \$49
	7abc1a23	09/23	\$4.33	\$2 – \$5
	4c7af72a	09/23	\$12.29	\$5 – \$16
	89c0829c	09/24	\$3.66	\$2 – \$5
	7abc1a23	09/24	\$35.81	\$16 – \$49

<sup>1</sup>Retirado do artigo: Unique in the Shopping Mall: On the reidentifiability of credit card

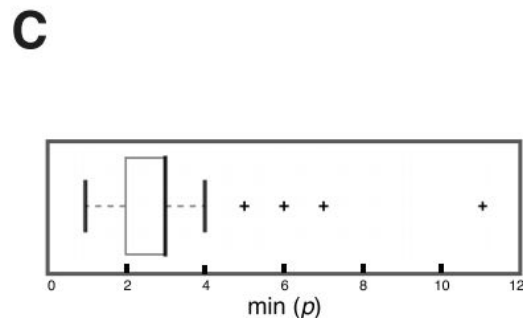
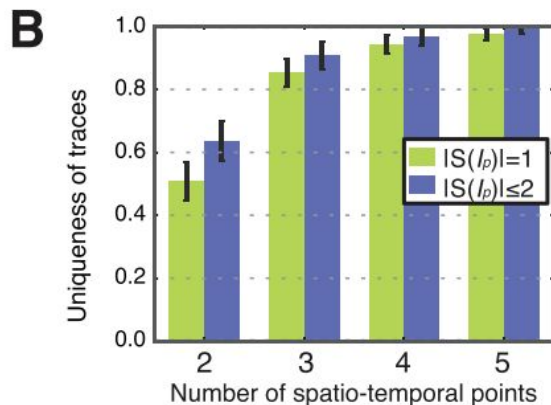
# Resultados

Unique in the Crowd

---

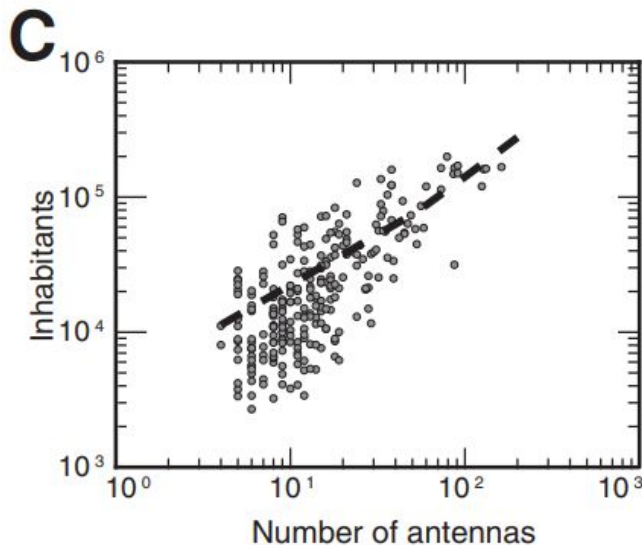
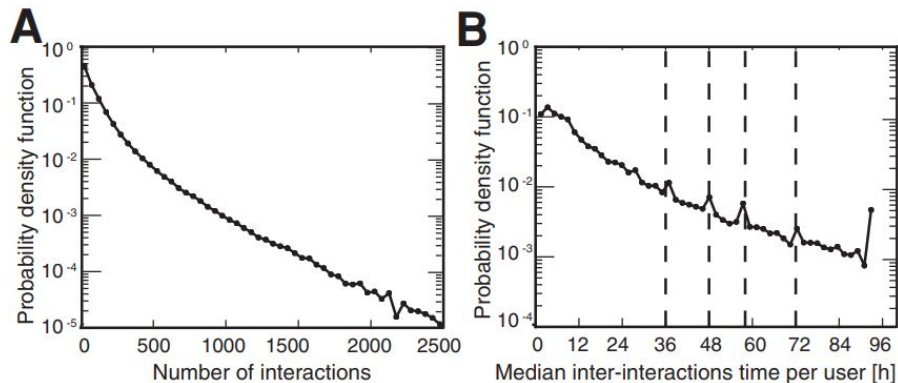
# Resultados

- Imagem A representa um exemplo similar ao anterior
- Imagem B mostra a razão entre a unicidade e o número de pontos
  - Com apenas 4 pontos aleatórios, é possível identificar 95% dos indivíduos
  - 2 pontos é o suficiente para identificar metade dos indivíduos
- Imagem C diz que até no máximo 11 pontos é suficiente para identificar todos os indivíduos



# Resultados

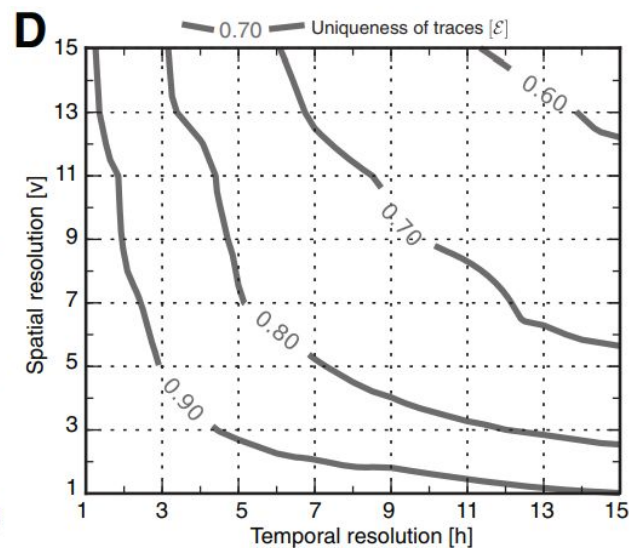
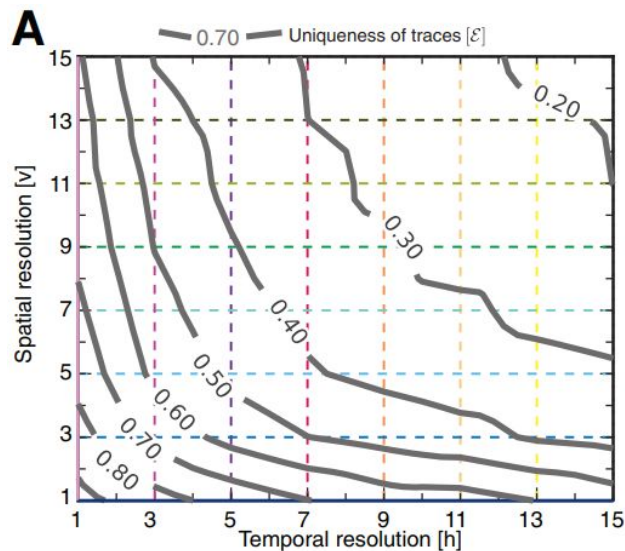
- 114 interações por usuário por mês
- 6500 antenas
  - ~2000 habitantes por antena
- Forte correlação entre o número de antenas e a densidade da população ( $R^2 = 0.6426$ )



Imagens A e B não foram discutidas no artigo

# Resultados

- $\varepsilon$  depende da resolução espacial e temporal dos dados
- Redução na resolução dos dados
  - Maior granularidade
  - Agrupamento de antenas e aumento da janela de  $h$  horas
- A ( $p = 4$ )
- B ( $p = 10$ )

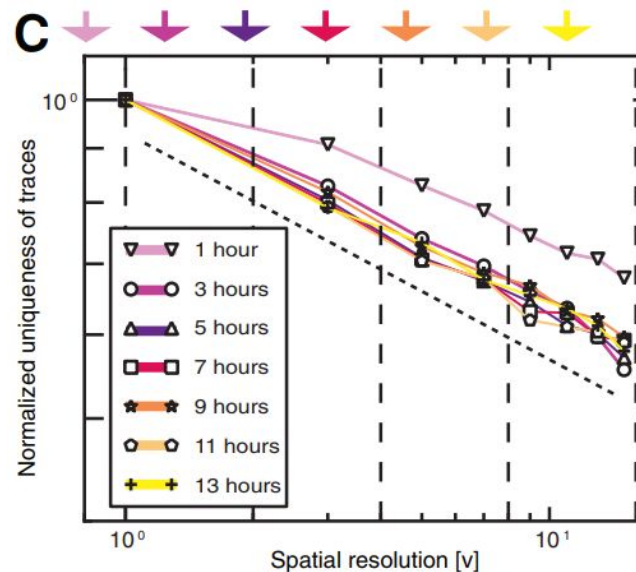
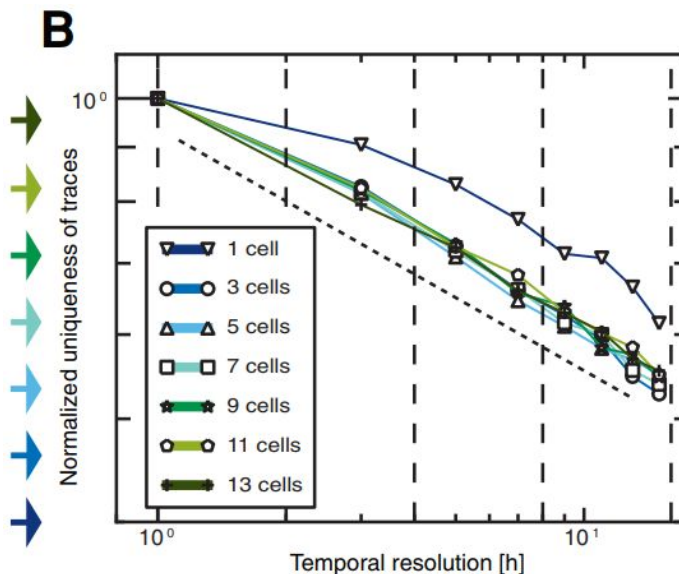


# Resultados

- A unicidade de um “rastros” diminui conforme a função de potência

$$\varepsilon = \alpha - x^\beta$$

- Ambos os estimadores para  $\alpha$  e  $\beta$  são altamente significativos ( $p < 0,001$ )



# Resultados

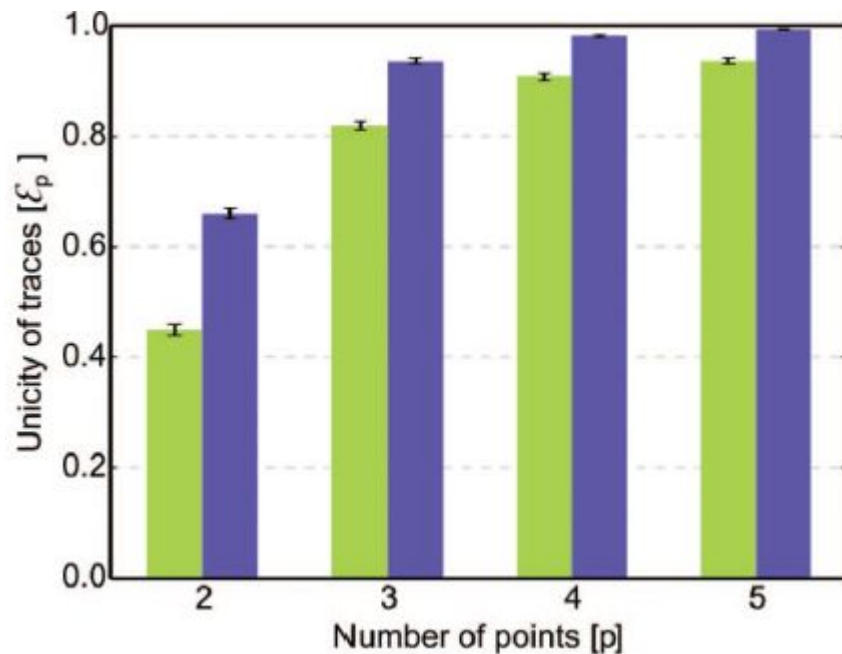
**Unique in the Shopping Mall**

---



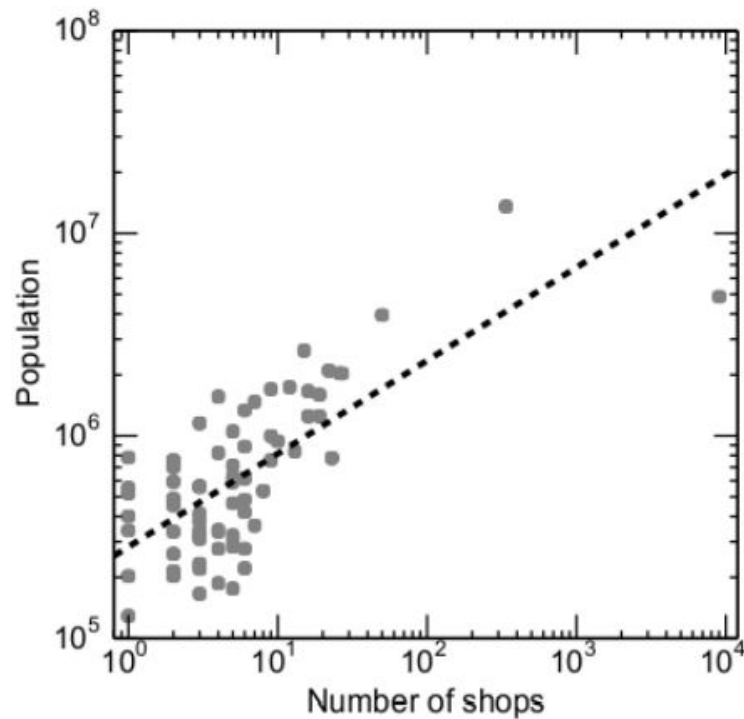
# Resultados

- Resultados semelhantes ao trabalho anterior
- Apenas 4 pontos identificam cerca de 90% dos indivíduos
- Barra verde = (dia, loja)
- Barra azul = (dia, loja, preço)



# Resultados

- Lojas distribuídas ao longo do país
- Correlação com a densidade da população  
( $R^2 = 0.51$ ,  $P < 0.001$ )



# Resultados

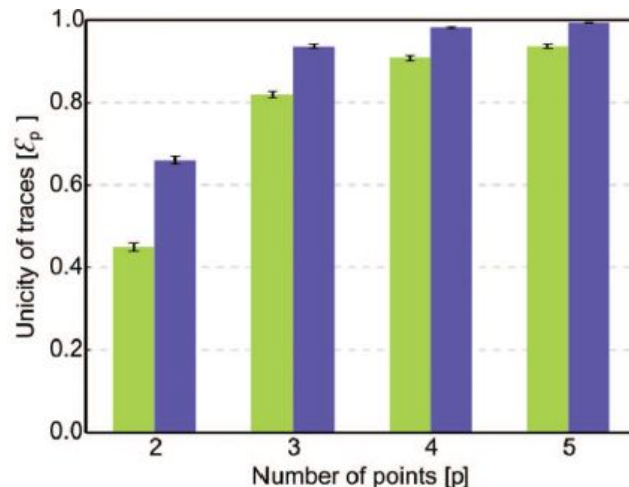
- O preço da transação pode ser usado na reidentificação
- Tripla (dia, loja, preço)
- Preço intervalado, seguido por uma resolução  $\alpha$
- Aumento de 22%, em média

**A**

Bin #	Range
0	]0.2, 0.6]
1	]0.6, 1.8]
2	]1.8, 5.4]
3	]5.4, 16.2]
4	]16.2, 48.6]
5	]48.6, 145.8]
6	]145.8, 437.4]
7	]437.4, 1312.2]
8	]1312.2, 3936.6]
9	]3936.6, 11809.8]
10	]11809.8, 35429.4]

**B**

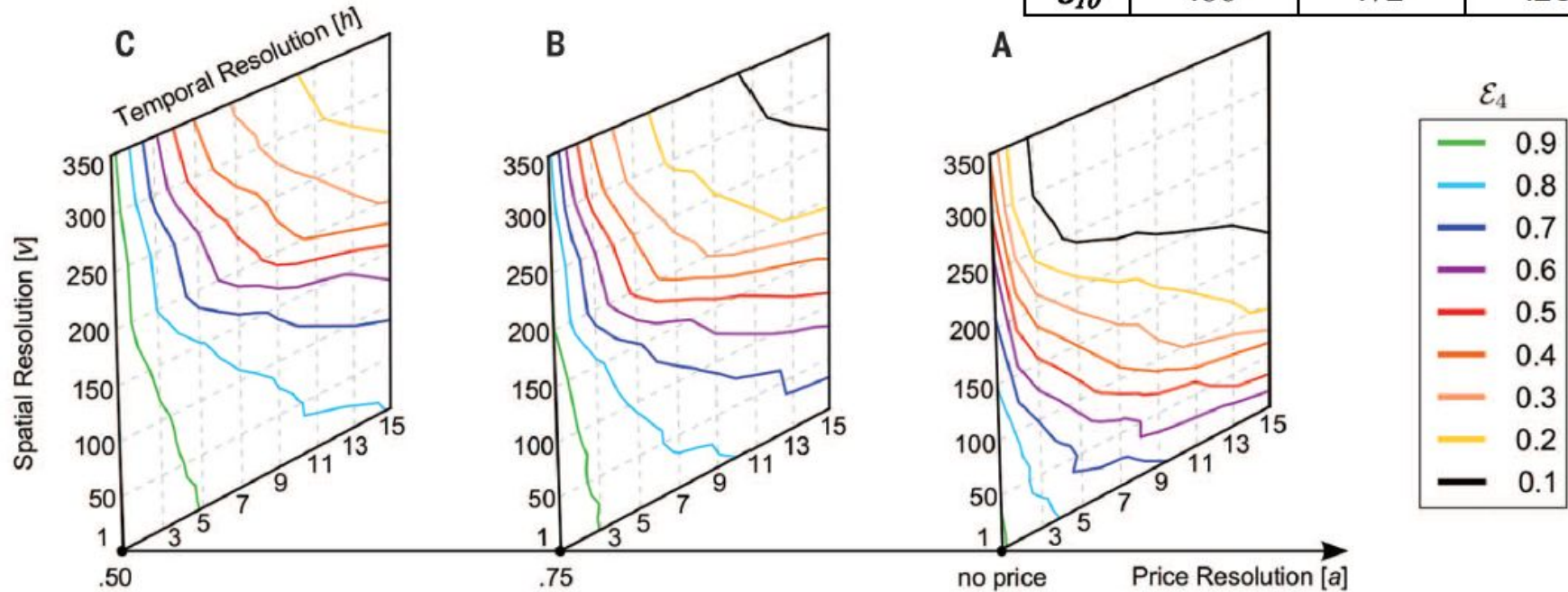
Bin #	Range
0	]0.1, 0.7]
1	]0.7, 4.9]
2	]4.9, 34.3]
3	]34.3, 240.1]
4	]240.1, 1680.7]
5	]1680.7, 11764.9]
6	]11764.9, 82354.3]



# Resultados

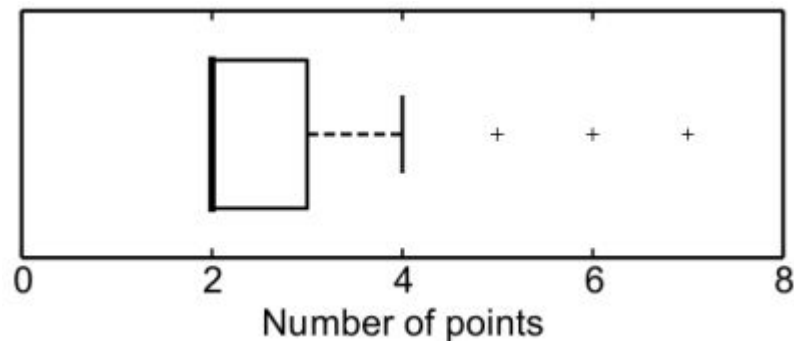
- Resolução dos dados

	Price Resolution [a]		
	.50	.75	no price
$\epsilon_4$	.13	.06	.00
$\epsilon_6$	.40	.25	.03
$\epsilon_{10}$	.86	.72	.21



# Resultados

- A seleção de pontos é aleatória
- Maior concentração de pontos
- No máximo 7 pontos, até para o pior caso



# Fim.

Obrigado!

Perguntas?