

# Introdução ao Data Cleaning

Eduardo Pena (ehmpena@inf.ufpr.br)

Graduate program in Informatics  
Department of Informatics  
Federal University of Paraná

3rd July 2018

# Outline

## 1 Motivação

## 2 Data Cleaning Revisited

# Section 1

## **Motivação**

Data is **Dirty!**

# Data is dirty

Dados limpos poupam dinheiro . . . e tempo

# Data is dirty

## Dados limpos poupam dinheiro ... e tempo

1. Dados sujos → bilhões de dolares perdidos todos os anos



S. Kandel and A. Paepcke and J. M. Hellerstein and J. Heer. Enterprise Data Analysis and Visualization: An Interview Study. IEEE TVCG, 2012.

# Data is dirty

## Dados limpos poupam dinheiro ... e tempo

1. Dados sujos → bilhões de dolares perdidos todos os anos
2. “Janitor work” → 30-80% do tempo gasto por analistas



S. Kandel and A. Paepcke and J. M. Hellerstein and J. Heer. Enterprise Data Analysis and Visualization: An Interview Study. IEEE TVCG, 2012.

# Data is dirty

## Dados limpos poupam dinheiro ... e tempo

1. Dados sujos → bilhões de dolares perdidos todos os anos
2. “Janitor work” → 30-80% do tempo gasto por analistas

# Big clean data is the new oil!



S. Kandel and A. Paepcke and J. M. Hellerstein and J. Heer. Enterprise Data Analysis and Visualization: An Interview Study. IEEE TVCG, 2012.



# Qualidade de dados

Esperamos dados com alta qualidade para tomada de decisão

1. Acurácia
2. Completude
3. Consistência
4. Atualização

# Qualidade de dados

Esperamos dados com alta qualidade para tomada de decisão

1. Acurácia
2. Completude
3. Consistência
4. Atualização

Dados devem atender as necessidades dos usuários.

# Qualidade de dados

Esperamos dados com alta qualidade para tomada de decisão

1. Acurácia
2. Completude
3. Consistência
4. Atualização

Dados devem atender as necessidades dos usuários.

# Problemas de data cleaning

	PNome	SNome	Salario	Bonus	Idade	Cep	End	Phone	Produto	Objetivo	Vendas
$t_0$	John	Miller	\$1000	\$300	48	5081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_1$	Brad	Fuhrmann	\$1000	\$400	40	5082	12 Canyon Road	3323	Bread	\$12000	\$14000
$t_2$	Julio	Lopez	\$3000	\$1100	60	5083	15 Bourbon Street	3326	Yogurt	\$20000	\$30000
$t_3$	Paul	Allen	\$1200	\$400	400	5001	20 Calle Ocho	3250	Soda	\$12000	\$13000
$t_4$	Greg	Miller	\$1000	\$300	05-05-1970	4081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_5$	Brad	Furman	\$1000	\$400	40	5082	80 Ocean Drive	3323	Bread	\$12000	\$14000
$t_6$	Jeff	Jones	\$3000	\$2000	36	5056	100 Worth Avenue	4260	Yogurt	\$20000	\$20000

# Problemas de data cleaning

	PNome	SNome	Salario	Bonus	Idade	Cep	End	Phone	Produto	Objetivo	Vendas
$t_0$	John	Miller	\$1000	\$300	48	5081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_1$	Brad	Fuhrmann	\$1000	\$400	40	5082	12 Canyon Road	3323	Bread	\$12000	\$14000
$t_2$	Julio	Lopez	\$3000	\$1100	60	5083	15 Bourbon Street	3326	Yogurt	\$20000	\$30000
$t_3$	Paul	Allen	\$1200	\$400	400	5001	20 Calle Ocho	3250	Soda	\$12000	\$13000
$t_4$	Greg	Miller	\$1000	\$300	05-05-1970	4081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_5$	Brad	Furman	\$1000	\$400	40	5082	80 Ocean Drive	3323	Bread	\$12000	\$14000
$t_6$	Jeff	Jones	\$3000	\$2000	36	5056	100 Worth Avenue	4260	Yogurt	\$20000	\$20000



- ▶ Ser humano não vive 400 anos
- ▶ 05-05-1970 precisa de um casting

 Domínio incorreto

# Problemas de data cleaning

	PNome	SNome	Salario	Bonus	Idade	Cep	End	Phone	Produto	Objetivo	Vendas
$t_0$	John	Miller	\$1000	\$300	48	5081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_1$	Brad	Fuhrmann	\$1000	\$400	40	5082	12 Canyon Road	3323	Bread	\$12000	\$14000
$t_2$	Julio	Lopez	\$3000	\$1100	60	5083	15 Bourbon Street	3326	Yogurt	\$20000	\$30000
$t_3$	Paul	Allen	\$1200	\$400	400	5001	20 Calle Ocho	3250	Soda	\$12000	\$13000
$t_4$	Greg	Miller	\$1000	\$300	05-05-1970	4081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_5$	Brad	Furman	\$1000	\$400	40	5082	80 Ocean Drive	3323	Bread	\$12000	\$14000
$t_6$	Jeff	Jones	\$3000	\$2000	36	5056	100 Worth Avenue	4260	Yogurt	\$20000	\$20000

- Brad Fuhrmann e Brad Furman são a mesma pessoa

 Domínio incorreto  
 Registro duplicado

# Problemas de data cleaning

	PNome	SNome	Salario	Bonus	Idade	Cep	End	Phone	Produto	Objetivo	Vendas
$t_0$	John	Miller	\$1000	\$300	48	5081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_1$	Brad	Fuhrmann	\$1000	\$400	40	5082	12 Canyon Road	3323	Bread	\$12000	\$14000
$t_2$	Julio	Lopez	\$3000	\$1100	60	5083	15 Bourbon Street	3326	Yogurt	\$20000	\$30000
$t_3$	Paul	Allen	\$1200	\$400	400	5001	20 Calle Ocho	3250	Soda	\$12000	\$13000
$t_4$	Greg	Miller	\$1000	\$300	05-05-1970	4081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_5$	Brad	Furman	\$1000	\$400	40	5082	80 Ocean Drive	3323	Bread	\$12000	\$14000
$t_6$	Jeff	Jones	\$3000	\$2000	36	5056	100 Worth Avenue	4260	Yogurt	\$20000	\$20000





- Regra 1: *End* determina funcionalmente *Cep*.
  - Tuplas  $t_0$  and  $t_4$  violam essa dependência de atributo (i.e.,  $t_0$  e  $t_4$  não estão **consistentes** com a regra 1).

<span style="color: red;">■</span>	Domínio incorreto
<span style="color: gray;">■</span>	Registro duplicado
<span style="color: blue;">■</span>	Dependência de atributo

# Problemas de data cleaning

	PNome	SNome	Salario	Bonus	Idade	Cep	End	Phone	Produto	Objetivo	Vendas
$t_0$	John	Miller	\$1000	\$300	48	5081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_1$	Brad	Fuhrmann	\$1000	\$400	40	5082	12 Canyon Road	3323	Bread	\$12000	\$14000
$t_2$	Julio	Lopez	\$3000	\$1100	60	5083	15 Bourbon Street	3326	Yogurt	\$20000	\$30000
$t_3$	Paul	Allen	\$1200	\$400	400	5001	20 Calle Ocho	3250	Soda	\$12000	\$13000
$t_4$	Greg	Miller	\$1000	\$300	05-05-1970	4081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_5$	Brad	Furman	\$1000	\$400	40	5082	80 Ocean Drive	3323	Bread	\$12000	\$14000
$t_6$	Jeff	Jones	\$3000	\$2000	36	5056	100 Worth Avenue	4260	Yogurt	\$20000	\$20000

- ▶ Rule 2: “Se dois vendedores vendem o mesmo produto e têm o mesmo salário, aquele com a menor venda não pode ter o maior bônus”.
  - ▶ Tuplas  $t_2$  e  $t_6$  violam tal regra (i.e., elas são **inconsistentes** com relação a regra 2).

	Domínio incorreto
	Registro duplicado
	Dependência de atributo
	Regra de negócio para valores de atributo



# Data cleaning

## Definição

Data Cleaning é o processo de detecção e correção de registros incorretos em um banco de dados.

## Operação

Data cleaning usa métodos computacionais que ajudam na definição, identificação, e reparo de uma variedade de erros.

# Data cleaning

## Definição

Data Cleaning é o processo de detecção e correção de registros incorretos em um banco de dados.

## Operação

Data cleaning usa métodos computacionais que ajudam na definição, identificação, e reparo de uma variedade de erros.

## Section 2

# **Data Cleaning Revisited**

# Abordagens em data cleaning

- ▶ **Restrições de integridade.** Expressam naturalmente regras de qualidade de dados; sua violação indica que dados estão sujos.
- ▶ **Data cleaning quantitativo.** Assume uma visão estatística dos dados para identificar outliers.
- ▶ **Deduplicação.** Consolida registros em uma ou mais relações que se referem a mesma entidade do mundo real.
- ▶ **Outras abordagens.** Todo o resto, de scripts que transformam dados até master data management.

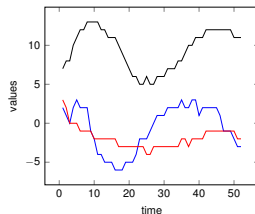
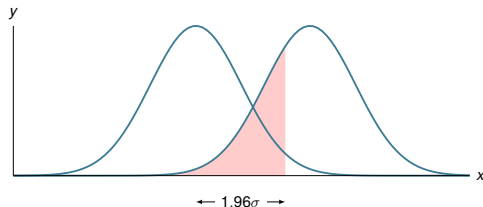
# Data cleaning quantitativo

## ► A visão com relação ao dado

- Configuração univariada
- Configuração multivariada
- Séries temporais

## ► Perspectivas de um SGBD

- Estatística e processamento de consulta
- Time Series Databases (TSDBs)



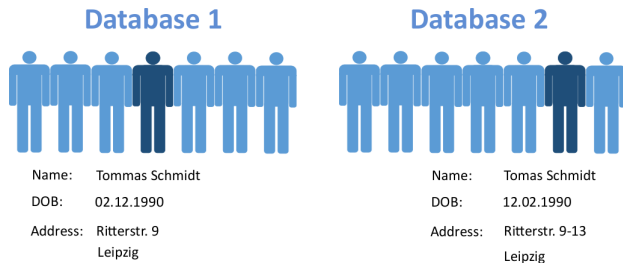
J. M. Hellerstein. Quantitative data cleaning for large databases. Tech report, 2008.

# Deduplicação

- ▶ É o processo de combinar os registros de diversos bancos de dados, onde os registros se referem a mesma entidade do mundo real.

- ▶ **Abordagem básica**

- ▶ Blocking keys
  - ▶ Indexação
  - ▶ Funções de similaridade
  - ▶ Modelos de decisão



P. Christen. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. IEEE TKDE, 2012.

# Outras abordagens

- ▶ Data integration
- ▶ Data transformations (e.g., data Wrangling)
- ▶ Master data management

# Cleaning para consistência

## Restrições de integridade (RIs)

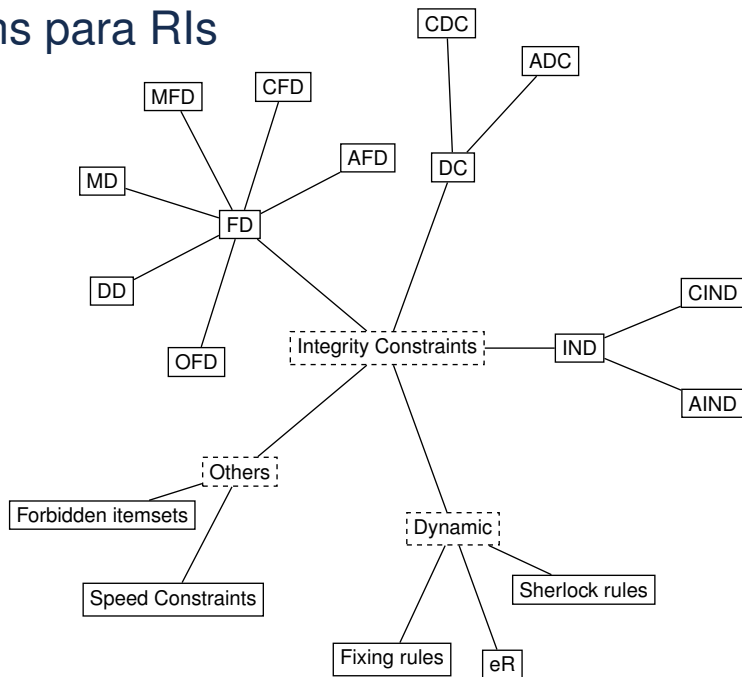
1. Uso de uma linguagem apropriada para expressar regras de qualidade de dados.
2. Detectar violações de RIs nos dados
3. Reparar violações de RIs, se possível
4. Descobrir automaticamente RIs dos dados



Chu, Xu and Ilyas, Ihab F. Qualitative Data Cleaning. PVLDB, 2016.



# Linguagens para RIs



# Expressando regras de qualidade de dados com uma restrição de negação(DCs)

Para verificar a regra “Se dois vendedores vendem o mesmo produto e têm o mesmo salário, aquele com a menor venda não pode ter o maior bônus”

	PNome	SNome	Salario	Bonus	Idade	Cep	End	Phone	Produto	Objetivo	Vendas
$t_0$	John	Miller	\$1000	\$300	48	5081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_1$	Brad	Fuhrmann	\$1000	\$400	40	5082	12 Canyon Road	3323	Bread	\$12000	\$14000
$t_2$	Julio	Lopez	\$3000	\$1100	60	5083	15 Bourbon Street	3326	Yogurt	\$20000	\$30000
$t_3$	Paul	Allen	\$1200	\$400	400	5001	20 Calle Ocho	3250	Soda	\$12000	\$13000
$t_4$	Greg	Miller	\$1000	\$300	05-05-1970	4081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_5$	Brad	Furman	\$1000	\$400	40	5082	80 Ocean Drive	3323	Bread	\$12000	\$14000
$t_6$	Jeff	Jones	\$3000	\$2000	36	5056	100 Worth Avenue	4260	Yogurt	\$20000	\$20000

$$\neg(t_x.Produto = t_y.Produto \wedge t_x.Salario = t_y.Salario \wedge t_x.Vendas > t_y.Vendas \wedge t_x.Bonus < t_y.Bonus)$$

# Restrição de negação/Denial constraints (DCs)

## Intuição

- ▶ Definem um conjunto de predicados que um banco de dados precisa satisfazer para prevenir que atributos recebam valores considerados semanticamente inconsistentes
  - ▶ Generalizam muitas outras RIs (FDs, CFDs, check constraints, etc)

# Detectar violações de RIs nos dados

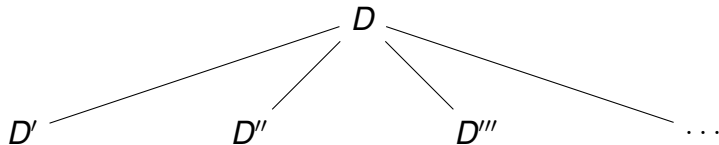
$$\varphi : \neg(t_x.Produto = t_y.Produto \wedge t_x.Salario = t_y.Salario \wedge t_x.Vendas > t_y.Vendas \wedge t_x.Bonus < t_y.Bonus)$$

	PNome	SNome	Salario	Bonus	Idade	Cep	End	Phone	Produto	Objetivo	Vendas
$t_0$	John	Miller	\$1000	\$300	48	5081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_1$	Brad	Fuhrmann	\$1000	\$400	40	5082	12 Canyon Road	3323	Bread	\$12000	\$14000
$t_2$	Julio	Lopez	\$3000	\$1100	60	5083	15 Bourbon Street	3326	Yogurt	\$20000	\$30000
$t_3$	Paul	Allen	\$1200	\$400	400	5001	20 Calle Ocho	3250	Soda	\$12000	\$13000
$t_4$	Greg	Miller	\$1000	\$300	05-05-1970	4081	48 6th avenue	3324	Soda	\$10000	\$10000
$t_5$	Brad	Furman	\$1000	\$400	40	5082	80 Ocean Drive	3323	Bread	\$12000	\$14000
$t_6$	Jeff	Jones	\$3000	\$2000	36	5056	100 Worth Avenue	4260	Yogurt	\$20000	\$20000

- Como reparar os valores de atributo  $t_2.Bonus$  e  $t_6.Bonus$ ?

# Reparar violações de RIs

- ▶ Dado um conjunto de RIs  $\Sigma$  definidas para um banco de dados sujo  $D$ 
  - ▶ Encontre um banco de dados  $D'$  que seja consistente com  $\Sigma$  e que difira minimamente de  $D$
  - ▶  $dist(D, D')$  deve ser mínima



# Reparando violações de RIs

$$\varphi : \neg(t_x.Product = t_y.Product \wedge t_x.Salary = t_y.Salary \wedge t_x.Sales > t_y.Sales \wedge t_x.Bonus < t_y.Bonus)$$

	...	Bonus	...
$t_0$		\$300	
$t_1$		\$400	
$t_2$		\$1100	
$t_3$		\$400	
$t_4$		\$300	
$t_5$		\$400	
$t_5$		\$2000	

...

	...	Bonus	...
$t_0$		\$300	
$t_1$		\$400	
$t_2$		\$2100	
$t_3$		\$400	
$t_4$		\$300	
$t_5$		\$400	
$t_5$		\$2000	

...

	...	Bonus	...
$t_0$		\$300	
$t_1$		\$400	
$t_2$		\$1100	
$t_3$		\$400	
$t_4$		\$300	
$t_5$		\$400	
$t_5$		\$2000	

...

	...	Bonus	...
$t_0$		\$300	
$t_1$		\$400	
$t_2$		\$1100	
$t_3$		\$400	
$t_4$		\$300	
$t_5$		\$400	
$t_5$		\$1000	

...

# Reparando violações de RIs

$$\varphi : \neg(t_x.Product = t_y.Product \wedge t_x.Salary = t_y.Salary \wedge t_x.Sales > t_y.Sales \wedge t_x.Bonus < t_y.Bonus)$$

	...	Bonus	...
$t_0$		\$300	
$t_1$		\$400	
$t_2$		\$1100	
$t_3$		\$400	
$t_4$		\$300	
$t_5$		\$400	
$t_5$		\$2000	

→ ?

# Reparando violações de RIs

$$\varphi : \neg(t_x.Product = t_y.Product \wedge t_x.Salary = t_y.Salary \wedge t_x.Sales > t_y.Sales \wedge t_x.Bonus < t_y.Bonus)$$

	...	Bonus	...
$t_0$		\$300	
$t_1$		\$400	
$t_2$		\$1100	
$t_3$		\$400	
$t_4$		\$300	
$t_5$		\$400	
$t_5$		\$2000	

→ ?

- ▶ Equivalence classes
- ▶ Vertex covers
- ▶ **Probabilistic models**
- ▶ ...



# Descoberta de RIs

Dado uma instância  $r$  de esquema  $R$ , encontre todas RIs que são válidas em  $r$

## Desafio

- ▶ Large search space
  - ▶ Descoberta de FDs em tabela com 100 colunas  $\rightarrow 2^{100} - 1 \rightarrow 1.3$  nonillion combinações de colunas

## Abordagens

- ▶ Schema Driven
  - ▶ Sensível ao tamanho do esquema
- ▶ Instance Driven
  - ▶ Sensível ao tamanho da instância
- ▶ Hybrid

# Desafios de data cleaning

- ▶ Keeping track of data errors
- ▶ Big data cleaning
- ▶ Holistic approaches
- ▶ Trusting the IC discovery

# Thank You

“If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. With adequate metadata, there is some hope, but even so, challenges will remain . . .”

---

*Agrawal et. al.* Challenges and opportunities with Big Data. Technical report, Computing Community Consortium, 2012.