InforAbERTA – IV Jornadas de Informática

Data Science e *Big Data*

Luís Cavique, Porto, março 2014

Agenda

1. Definições: padrões micro e Macro

2. Novos padrões para velhos problemas: *Similis*, *Ramex*, *Process Mining*

3. Oportunidades e Ameaças do *Big Data*

4. Conclusões

1. Definições

- **Bases de Dados**
  - Identifique os clientes que compraram mais de 1.000 euros
  - Identifique os dois produtos mais vencidos
  - Identifique os 10 clientes com mais reclamações

- ***Data Mining/Data Science***
  - Identifique os grupos de clientes com hábitos de compra idênticos (*clustering*)
  - Encontre o produto X que é frequentemente adquirido com o produto Y (regras associativas)
  - O que leva os clientes a reclamar? (classificação)

1. Definições

- **Bases de Dados**
  - **Que dados satisfazem um padrão (consulta)?**

```
SELECT Name,Salary
FROM EMPLOYEE
```

Name	Salary
XXX	1200
YYY	1000
ZZZ	1300
XXX	1200
ABC	1500

- ***Data Mining/Data Science***
  - **Que padrões satisfazem os dados?**

Patient	Age	#Medications	Complication
1	52	7	Yes
2	57	9	Yes
3	43	6	Yes
4	33	6	No
5	35	8	No
6	49	8	Yes
7	58	4	No
8	62	3	No
9	48	0	No
10	37	6	Yes

Age >= 37
AND
#Medications >= 6
→
Complication = Yes (100% confidence)

Patient	Age	#Medications	Complication
1	52	7	Yes
2	57	9	Yes
3	43	6	Yes
4	33	6	No
5	35	8	No
6	49	8	Yes
7	58	4	No
8	62	3	No
9	48	0	No
10	37	6	Yes

Age >= 37
AND
#Medications >= 6
→
Complication = Yes (100% confidence)

1. Definições

- **Regra Associativa: E => D**

- ***support*** = count (D & E) / total_count
- i.e. a probabilidade de comprar D & E em conjunto

- ***confidence*** = support (D & E) / support(E)
- i.e. a probabilidade condicionada de comprar D se comprou E

- Exemplo: Fraldas => Cerveja

1. Definitions

- *Data Mining*: micro e Macro padrões

- **micro padrão**: corresponde a uma pequena percentagem de dados; ex: support >= 5% nas regras de associação;

- **Macro padrão**: corresponde a uma grande percentagem de dados; ex: regressão linear

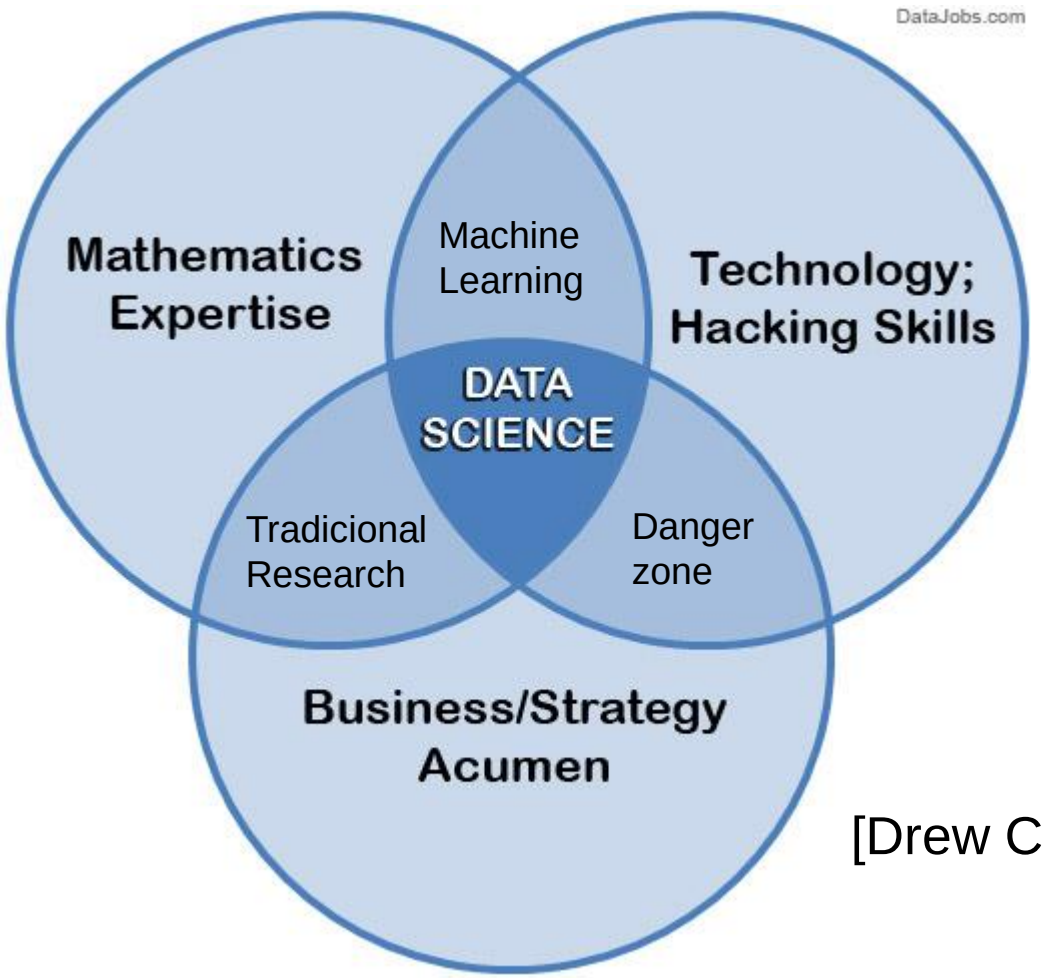
1. Definições: exemplos

micro padrões	Macro padrões
Market Basket Analysis Algoritmo Apriori, Support >= 5%	Análise Clusters
Sequence Mining Algoritmo AprioriAll, Support >= 1%	Teste Hipóteses
Classificação *Decision tree*	Regressão
Classificação K-vizinhos mais próximos	Seleção de Atributos

1. Definições

- Inicialmente, o termo “data mining” tinha uma conotação negativa (“data snooping” e “data fishing”).

1. Definições



[Drew Conway 2010]

1. Definições: sumário

micro padrões	Macro padrões
corresponde a uma pequena percentagem de dados	corresponde a uma grande percentagem de dados
Confidence elevado	*Support elevado*
Machine Learning Hacking Skills	Statistics Mathematics
orçamento = $1,000,000	orçamento = $50,000

"Statistical Modeling: The Two Cultures" paper by Leo Breiman in 2001

2. Novos padrões para velhos problemas

- Market Basket Analysis: Apriori, Silimis
- Sequence Mining: AprioriAll, Ramex, Process Mining

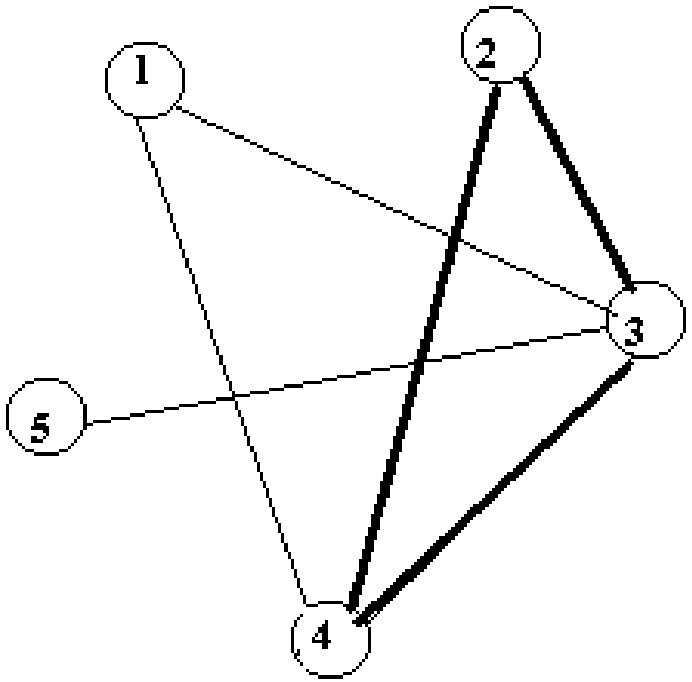
2. Novos padrões para velhos problemas

Apriori [Agrawal et al.1996]

SIZE	COUNT	SUP	CONF	RULE
4	147	5,78	50,00	p29 & p26 ==> p30 & p18
4	147	5,78	47,27	p29 & p18 ==> p30 & p26
4	147	5,78	64,19	p26 & p18 ==> p30 & p29
4	147	5,78	63,36	P30 & p29 & p26 ==> p18
4	147	5,78	63,64	p30 & p29 & p18 ==> p26
4	147	5,78	83,52	p30 & p26 & p18 ==> p29
4	147	5,78	84,48	p29 & p26 & p18 ==> p30

2. Novos padrões para velhos problemas

	1	2	3	4	5
1		25	44	50	33
2			66	50	33
3				66	57
4					33
5					



Peso_Clique (1,4)=50
Peso_Clique (2,3)=66
Peso_Clique (1,3,4)=160
Peso_Clique (2,3,4)=182

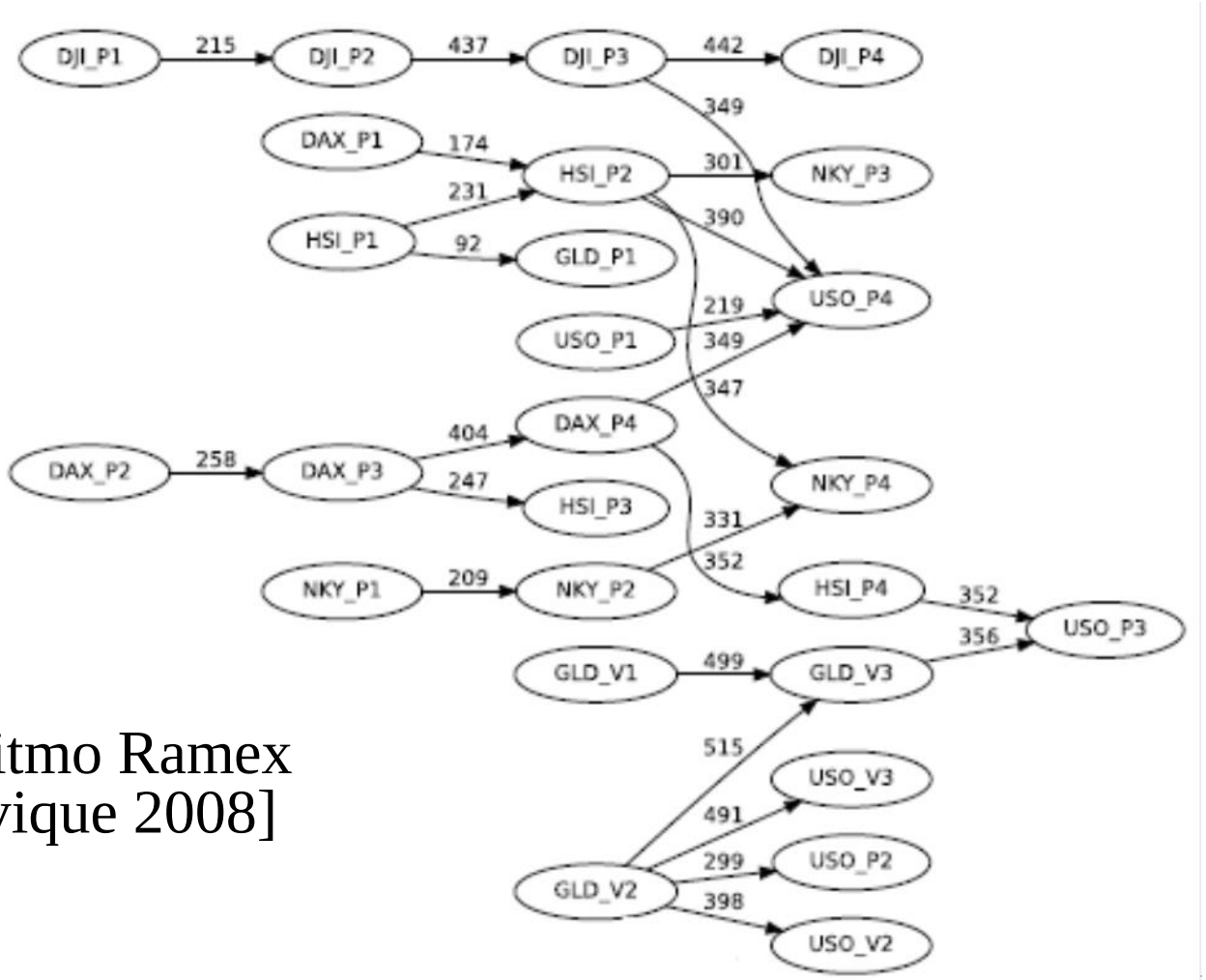
Algoritmo Similis
[Cavique 2007]

2. Novos padrões para velhos problemas

- Sequence Mining
  - ApriorAll [Agrawal 1995]

Rule	Rule_Probability
Indiana Jones and the Temple of Doom = Existing, Star Wars: Episode II - Atta...	0.9
Indiana Jones and the Last Crusade = Existing, Star Wars: Episode II - Attack o...	0.89130434782...
Indiana Jones and the Temple of Doom = Existing, Star Wars: Episode I - The ...	0.88888888888...
Indiana Jones and the Last Crusade = Existing, Star Wars: Episode I - The Pha...	0.87179487179...
Indiana Jones and the Temple of Doom = Existing, Return of the Jedi = Existing...	0.85714285714...
Indiana Jones and the Last Crusade = Existing, Return of the Jedi = Existing -> ...	0.82666666666...
Blade Runner = Existing, Return of the Jedi = Existing -> Star Wars = Existing	0.81578947368...
Indiana Jones and the Temple of Doom = Existing, The Empire Strikes Back = ...	0.80952380952...
Indiana Jones and the Last Crusade = Existing, The Empire Strikes Back = Exis...	0.79012345679...
Blade Runner = Existing, The Empire Strikes Back = Existing -> Star Wars = Exi...	0.66153846153...
Indiana Jones and the Last Crusade = Existing, The Lord of the Rings: The Fell...	0.56140350877...
Indiana Jones and the Last Crusade = Existing, Raiders of the Lost Ark = Existi...	0.50617283950...
Indiana Jones and the Temple of Doom = Existing -> Star Wars = Existing	0.5
Indiana Jones and the Temple of Doom = Existing, Raiders of the Lost Ark = Ex...	0.49586776859...
Indiana Jones and the Temple of Doom = Existing, Indiana Jones and the Last ...	0.49137931034...
Blade Runner = Existing, Raiders of the Lost Ark = Existing -> Star Wars = Existi...	0.484375
Indiana Jones and the Last Crusade = Existing -> Star Wars = Existing	0.47802197802...
Blade Runner = Existing -> Star Wars = Existing	0.29831932773...

2. Novos padrões para velhos problemas



Algoritmo Ramex
[Cavique 2008]

2. Novos padrões para velhos problemas

Process Mining

[van der Aalst, W. 2011]

2. Novos padrões para velhos problemas

- Manifesto Process Mining (IEEE)

C5: melhorar a representação dos resultados do processo de descoberta

C6: equilibrar os critérios de qualidade tais como *fitness*, simplicidade, precisão e generalização

C10: melhorar a usabilidade para os não-especialistas

C11: melhorar compreensão para os não-especialistas

2. Novos padrões para velhos problemas

Algoritmos de Macro-padrões :

1. transformar dados brutos em estruturas condensadas usando funções de acumulação (grafos, cadeias de Markov, redes de Petri)

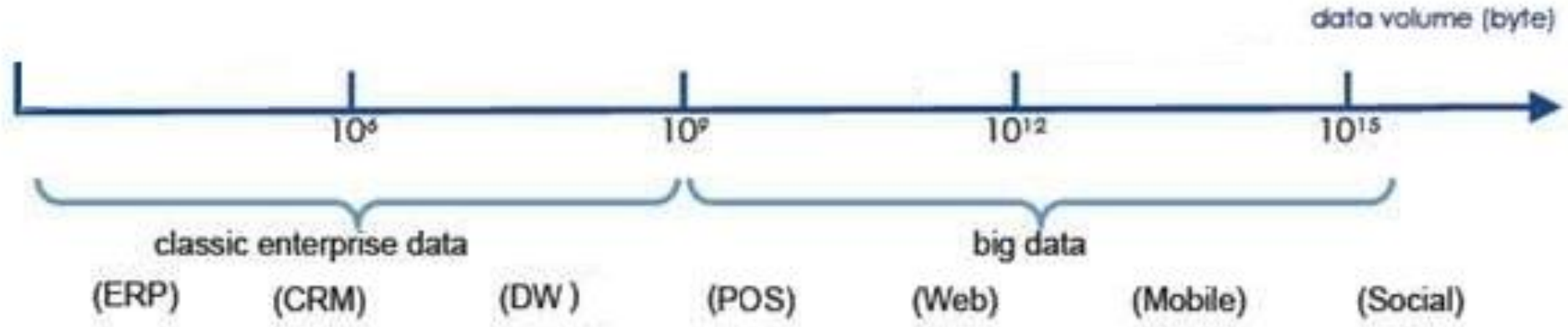
2. pesquisar a estrutura condensada para obter macro-padrões

2. Novos padrões para velhos problemas: resumo

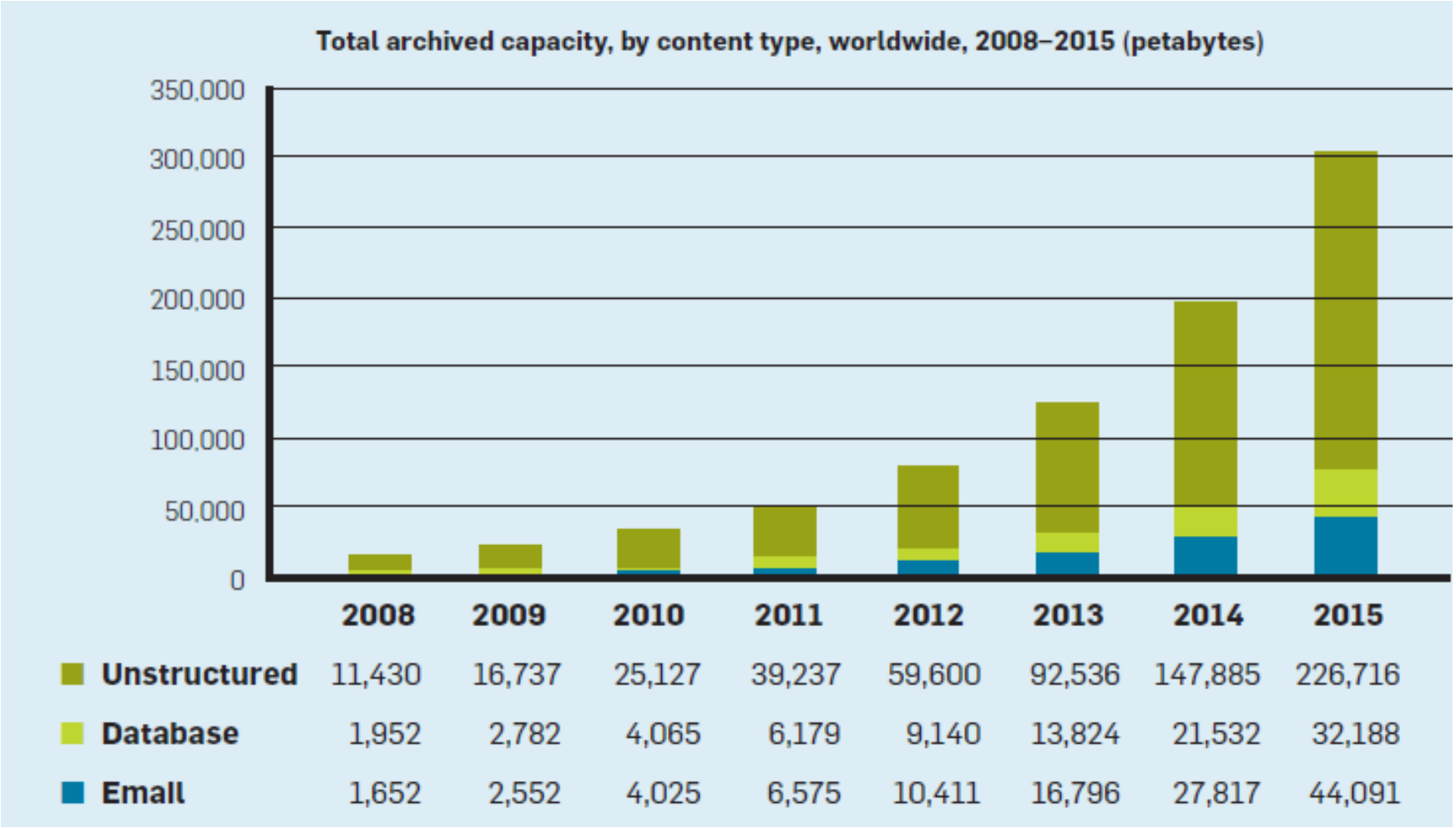
problema	velho padrão (micro padrão)	novo padrão (Macro padrão)
Market Basket Analysis	*Apriori*	*Similis*
Sequence Mining	*AprioriAll*	*Ramex Process Mining*

3. Oportunidades e ameaças do *Big Data*

3. Oportunidades e ameaças do *Big Data*



3. Oportunidades e ameaças do *Big Data*

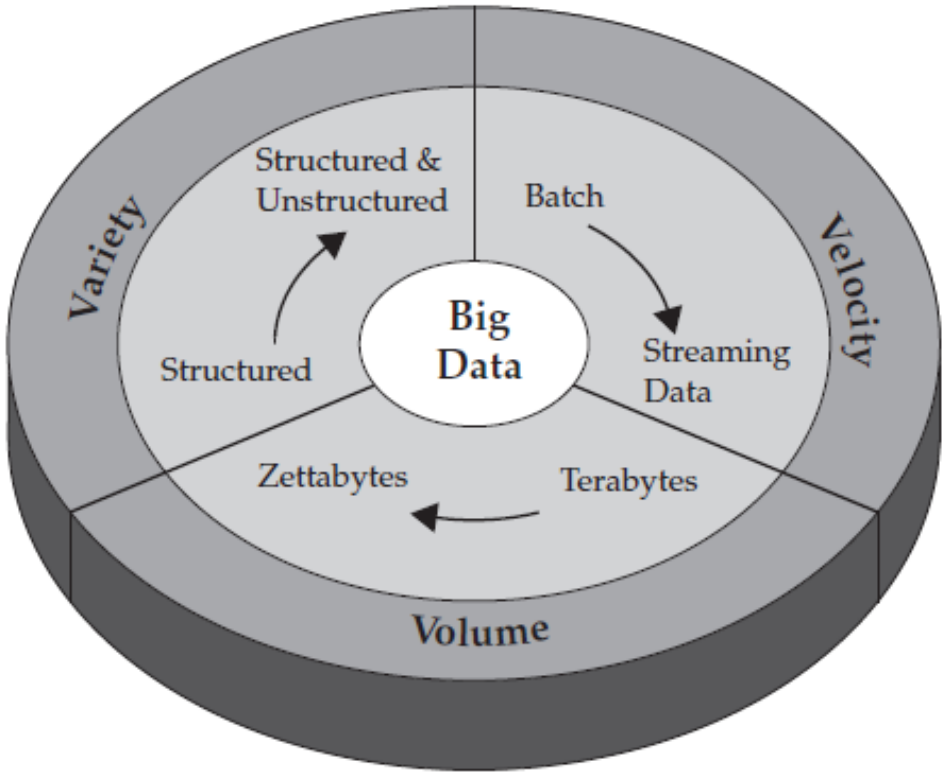


	2008	2009	2010	2011	2012	2013	2014	2015
Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
Email	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

3. Oportunidades e ameaças do *Big Data: 3V+V*

- 3 V:
  - Volume
  - Velocidade
  - Variedade

  - Valor



3. Oportunidades e ameaças do *Big Data: Dark Data*

- “Se souber a pergunta que me quer fazer, eu poderei dar-lhe várias respostas interessantes baseadas nos dados disponíveis”

3. Oportunidades e ameaças do *Big Data: NoSQL*

Type	Typical usage	Examples
Key-value store—A simple data storage system that uses a key to access a value	· Image stores · Key-based filesystems · Object cache · Systems designed to scale	· Berkeley DB · Memcache · Redis · Riak · DynamoDB
Column family store—A sparse matrix system that uses a row and a column as keys	· Web crawler results · Big data problems that can relax consistency rules	· Apache HBase · Apache Cassandra · Hypertable · Apache Accumulo
Graph store—For relationship-intensive problems	· Social networks · Fraud detection · Relationship-heavy data	· Neo4j · AllegroGraph · Bigdata (RDF data store) · InfiniteGraph (Objectivity)
Document store—Storing hierarchical data structures directly in the database	· High-variability data · Document search · Integration hubs · Web content management · Publishing	· MongoDB (10Gen) · CouchDB · Couchbase · MarkLogic · eXist-db · Berkeley DB XML

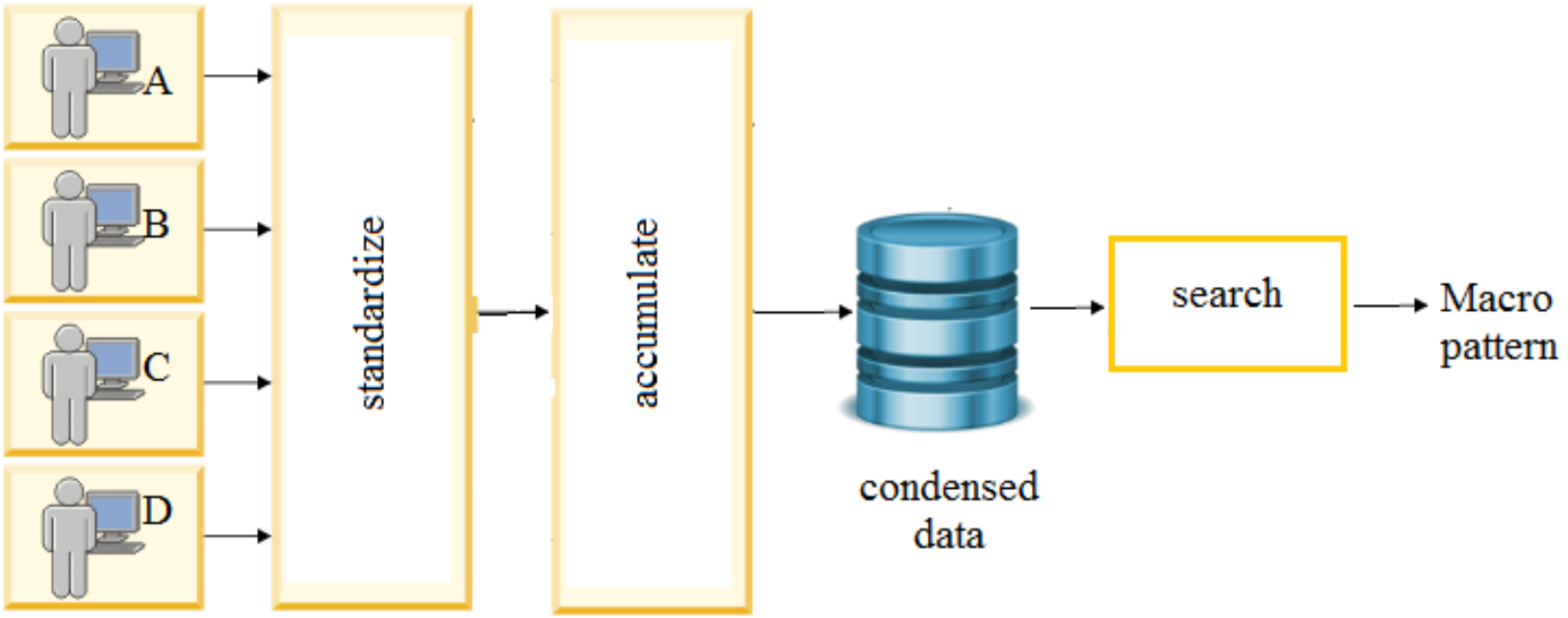
3. Oportunidades e ameaças do *Big Data*

Variância atualmente implementada no Hadoop

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i^2-\overline{y}^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}y_i^2-\frac{n}{n-1}\overline{y}^2.
\end{aligned}
$$

3. Oportunidades e ameaças do *Big Data*

- dados bruto= normalizar (variedade de formatos);
- dados condensados = acumular (dados bruto);
- Macro padrões = procurar (dados condensados);



4. Conclusões

- Padrões de bases de dados vs *data mining;*
- Padrões micro vs Macro;

- Algoritmos que geram Macro padrões, como o Process Mining, obtêm equilíbrios entre a visualização e medidas de qualidade;

- Algoritmos com estruturas condensadas criam grandes oportunidades em *Big Data.*