

Big Data e Ciência de Dados

Prof. André Britto de Carvalho

Introdução

- Dados
 - Fatos que podem ser gravados e que têm significado implícito.
 - Nomes, telefones, endereços, notas, etc.

Introdução

- Banco de Dados → Coleção de dados que descreve as atividades de uma ou mais **organizações** ou **empresas**.

Introdução

- Sem perceber geramos dados a todo momento
 - Uso de cartões fidelidade
 - Compras no cartão de crédito
 - Buscas na internet
 - Ida a uma hospital

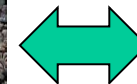
Introdução

■ Passado

- Poucas companhias geravam dados
- Consumidas por companhias, governo, ONG, pessoas

■ Presente

- Todos produzem dados
- Todos consomem dados



Explosão de dados

- Onde os dados são gerados
 - Dados de empresas
 - Dados do governo
 - Dados sociais
 - Dados de sensores
 - Música
 - Imagens
 - Vídeos

Fontes de dados

- Dispositivos eletrônicos
 - Smartphones
 - Logs de servidores de aplicação
 - Jogos e web sites
 - Sensores
 - Dados do clima, reservatórios de água, corpo humano
 - Imagens e vídeos
 - Monitoriamento de tráfego, vigilância.

Fontes de dados

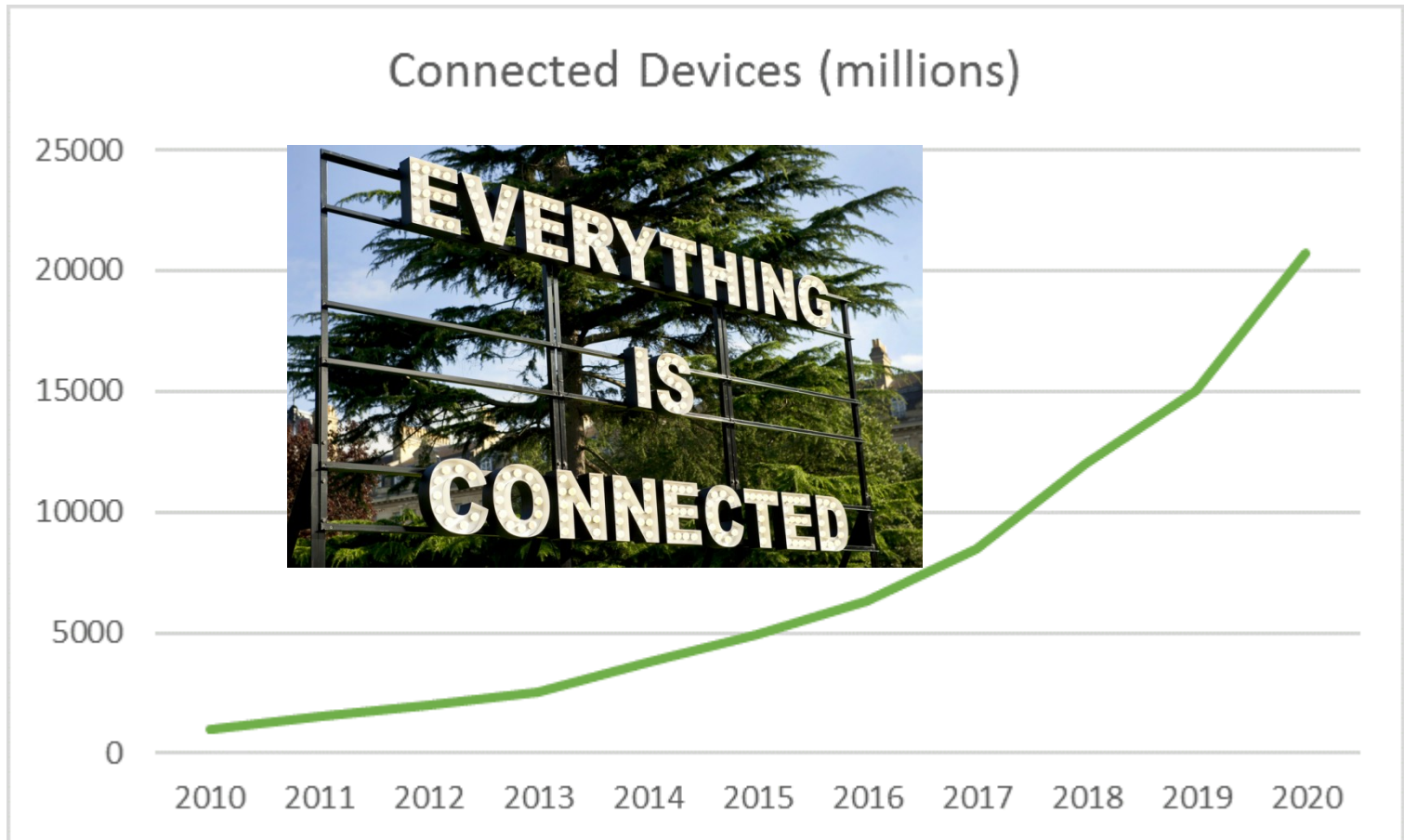
- Atividades humanas
 - Blogs
 - Emails
 - Search and browsing
 - Redes sociais

Por que há essa explosão de dados?

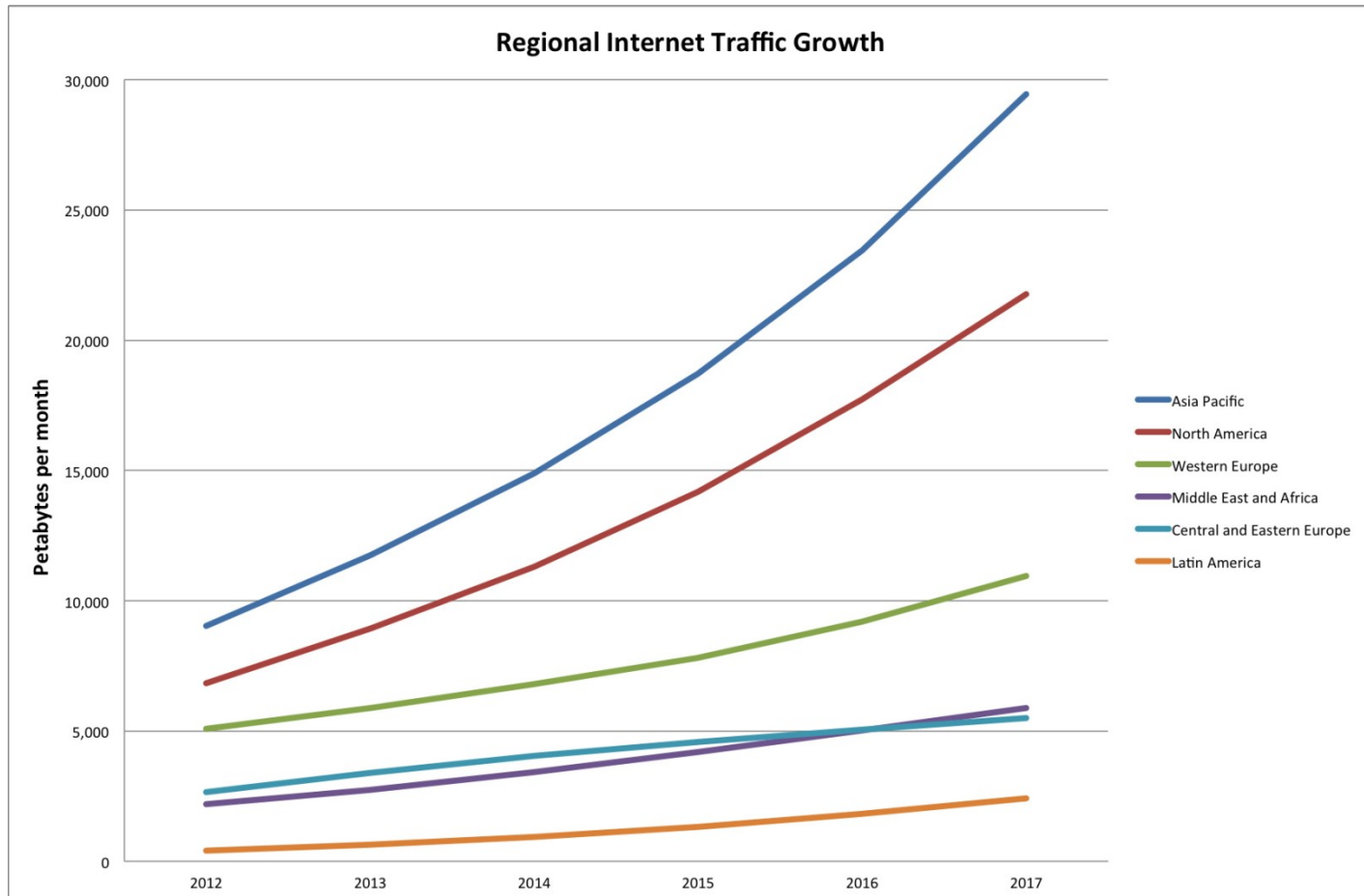
- Avanços da tecnologia
 - Transmissão
 - Processamento
 - Armazenamento
- Mais, rápido e barato



Crescimento na transmissão de dados



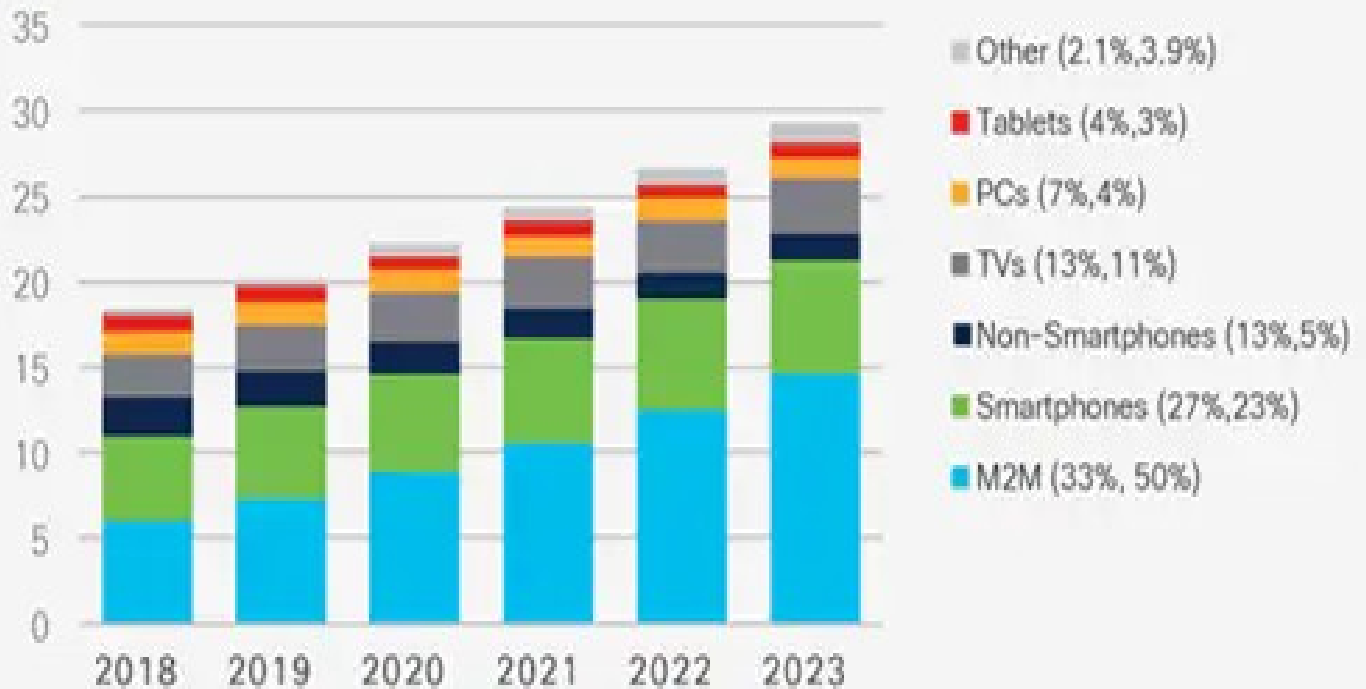
Crescimento na transmissão de dados



De onde os dados são gerados?

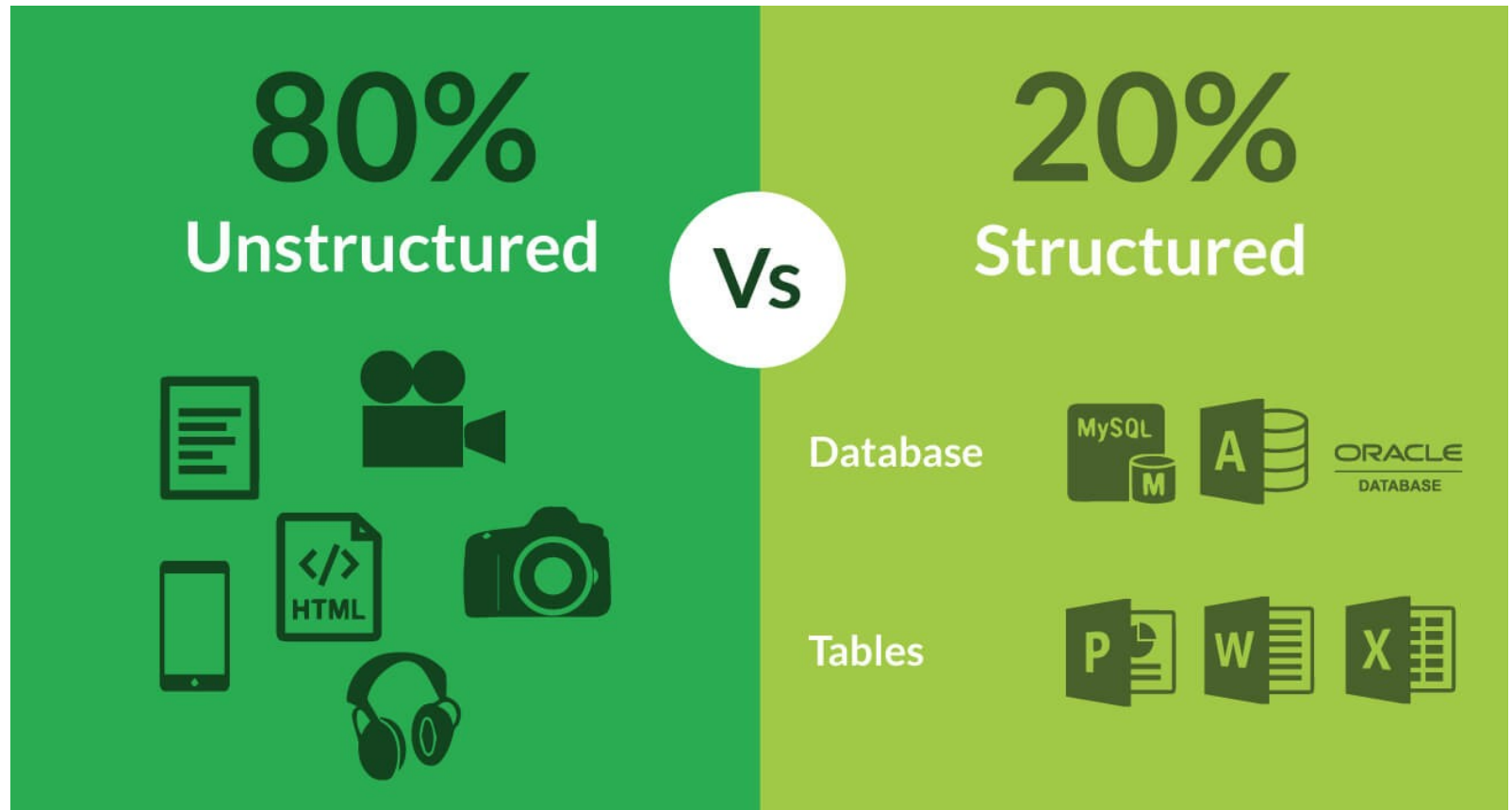
10% CAGR
2018-2023

Billions of
Devices

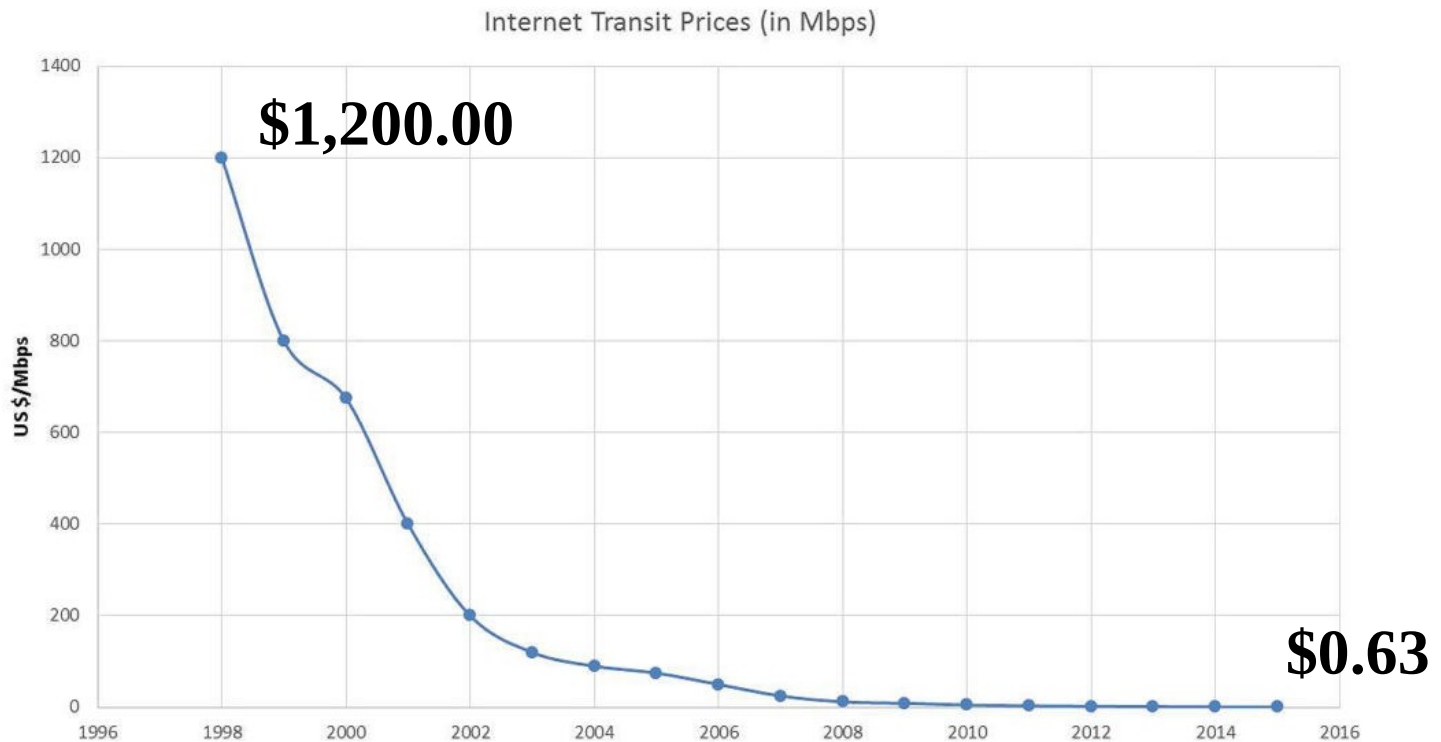


* Figures (n) refer to 2018, 2023 device share

O que são esses dados?

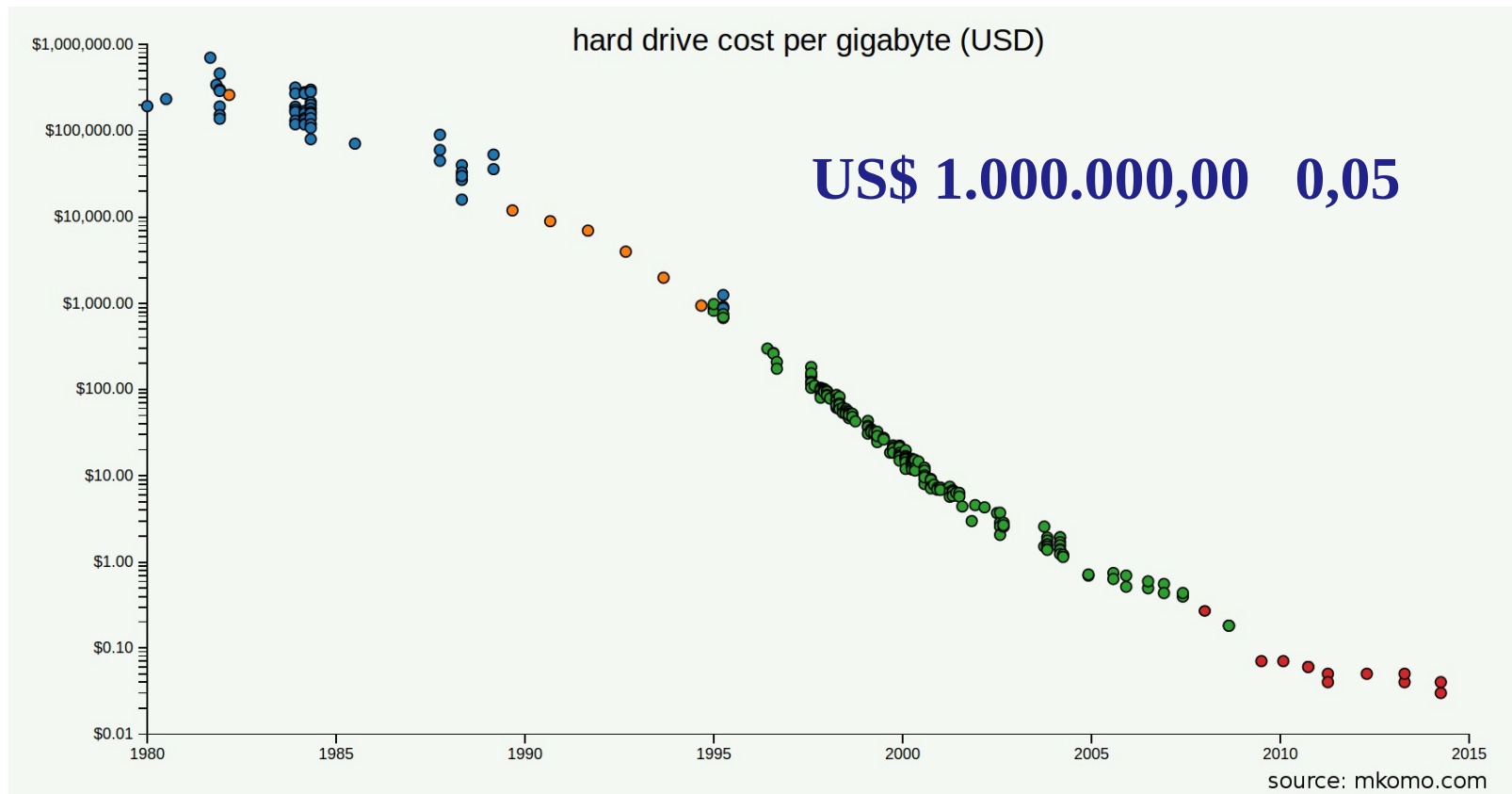


Quanto ao custo da transmissão?

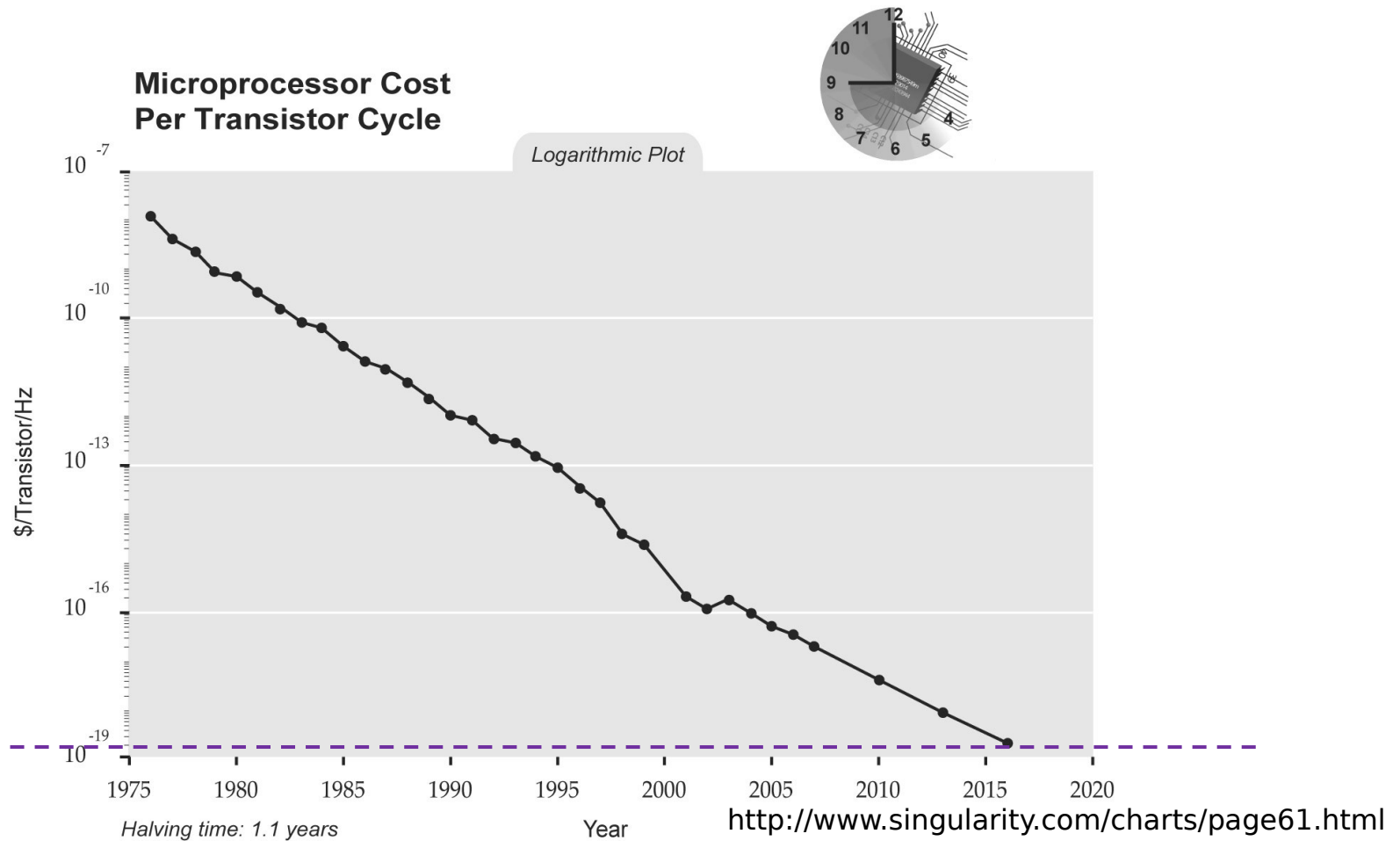


Source: DrPeering.net

Quanto ao custo de armazenamento?



Quanto ao custo do processamento?



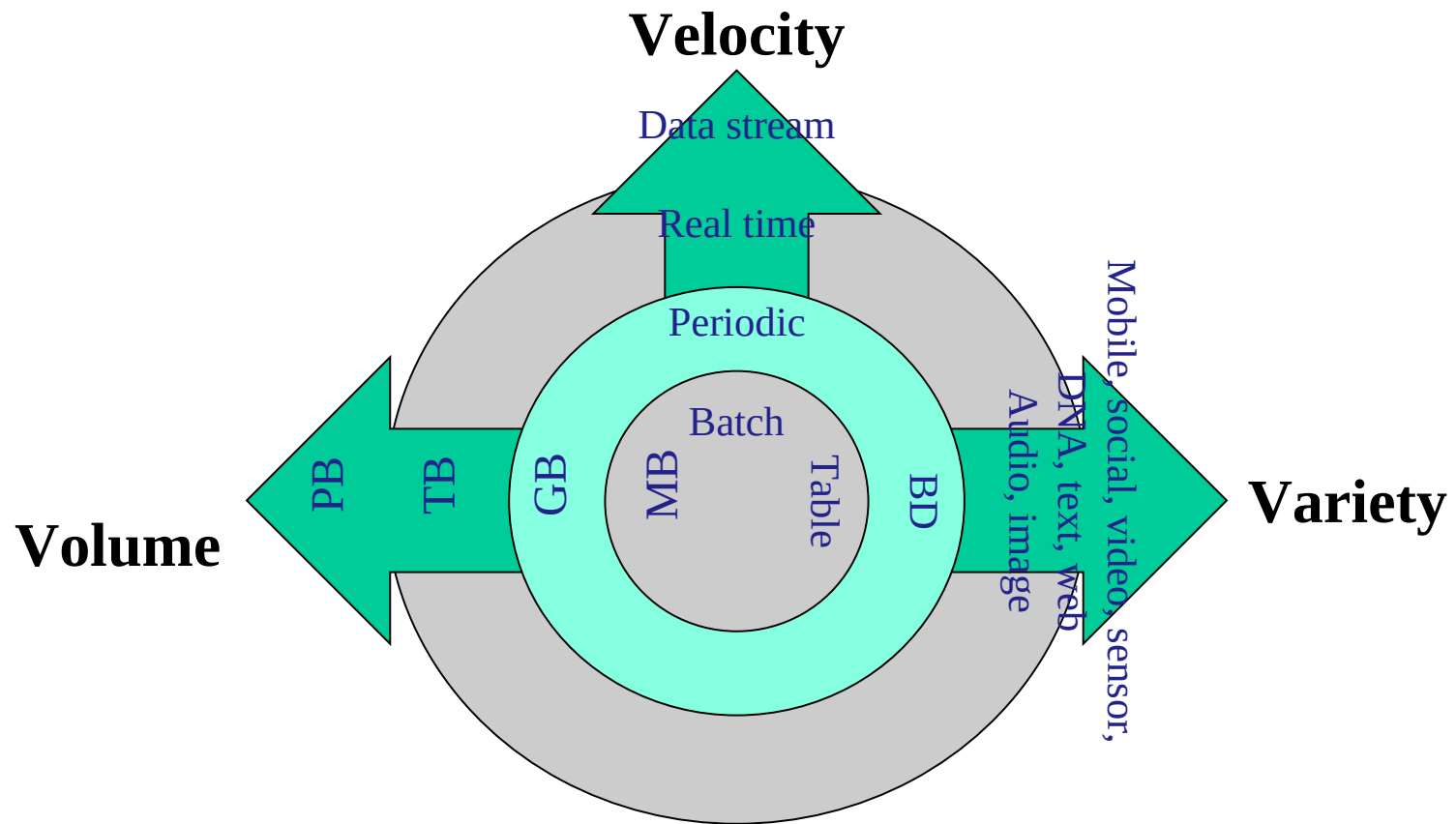
Big Data



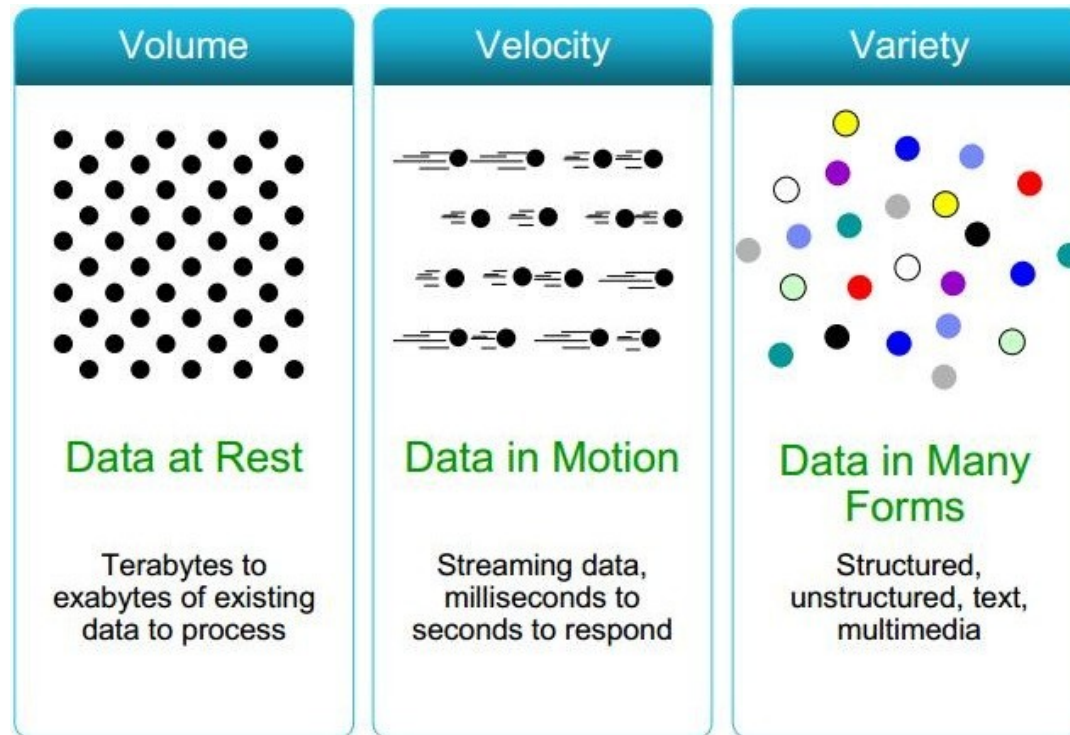
Big Data

- Grande **volume** de dados, gerado em alta taxa de **velocidade** e com uma larga variedade (Vs 3)
 - Volume: estruturados e não estruturados
 - Variety: diferentes fontes
 - Velocity : gerados em streams de dados cada vez mais rápidos

Big Data























3 Vs



Data is money

TOP 10 RANKING CHANGES SIGNIFICANTLY...

Only three brands that appeared in the BrandZ™ Global Top 10 in 2006—Google, Microsoft, and IBM—remain in the Top 10 in 2017.

	2006	Brand Value 2006 \$Mil.	2017	Brand Value 2017 \$Mil.
1	 Microsoft	62,039		245,581
2		55,834		234,671
3		41,406	 Microsoft	143,222
4	 中国移动 China Mobile	39,168		139,286
5		38,510		129,800
6	 Walmart	37,567	 AT&T	115,112
7		37,445		110,999
8		36,084	 Tencent 腾讯	108,292
9		31,028		102,088
10	 TOYOTA	30,201	 McDonald's	97,723

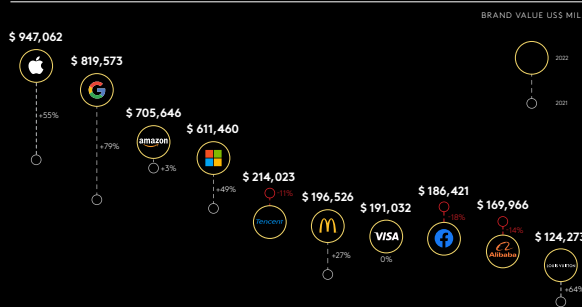
Source: Kantar Millward Brown / BrandZ™ (including data from Bloomberg)

Data is money

THE TOP 100 MOST VALUABLE GLOBAL BRANDS

1	APPLE	51	MEITUAN
2	GOOGLE	52	AMD
3	AMAZON	53	TIKTOK
4	MICROSOFT	54	AMERICAN EXPRESS
5	TENCENT	55	WELLS FARGO
6	MCDONALD'S	56	XBOX
7	VISA	57	RBC
8	FACEBOOK	58	GUCCI
9	ALIBABA	59	J.P. MORGAN
10	LOUIS VUITTON	60	JD
11	NVIDIA	61	HDFC BANK
12	MASTERCARD	62	ICBC
13	NIKE	63	HAIER
14	MOUTAI	64	INFOSYS
15	VERIZON	65	VODAFONE
16	ARAMCO	66	TOYOTA
17	COCA-COLA	67	HUAWEI
18	IBM	68	CHASE
19	ADOBE	69	BANK OF AMERICA
20	INSTAGRAM	70	MERCEDES-BENZ
21	UPS	71	MERCADO LIBRE
22	ORACLE	72	TD
23	AT&T	73	SIEMENS
24	YOUTUBE	74	SNAPCHAT
25	THE HOME DEPOT	75	UNITEDHEALTHCARE
26	ACCENTURE	76	BMW
27	HERMES	77	PING AN
28	PAYPAL	78	DHL
29	TESLA	79	UBER
30	NETFLIX	80	COMMBANK**
31	SAP	81	DELL TECHNOLOGIES
32	TELEKOM/T-MOBILE	82	KUAI SHOU
33	QUALCOMM	83	ZARA
34	INTEL	84	NIT
35	STARBUCKS	85	FEDEX
36	XFINITY	86	LOWE'S
37	WALMART	87	LANCÔME
38	DISNEY	88	CHINA MOBILE
39	MARLBORO	89	ADIDAS
40	LINKEDIN	90	TARGET
41	CISCO	91	IKEA
42	TEXAS INSTRUMENTS	92	LIC
43	SALESFORCE	93	BUDWEISER
44	SAMSUNG	94	AIA
45	CHANEL	95	KFC
46	TCS*	96	ADYEN
47	INTUIT	97	XIAOMI
48	COSTCO	98	ALDI
49	SPECTRUM	99	AIRBNB
50	L'ORÉAL PARIS	100	MORGAN STANLEY

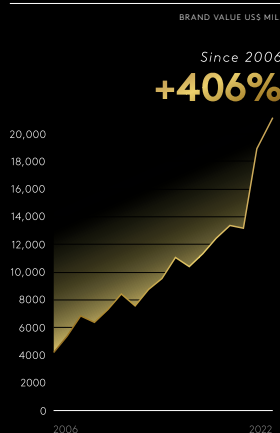
THE TOP 10



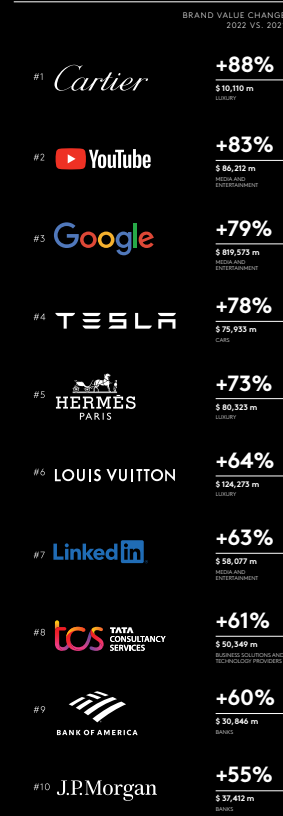
NEWCOMERS



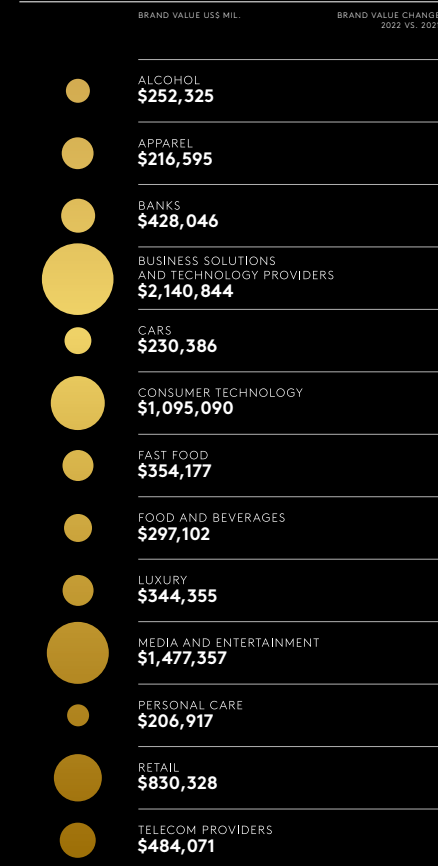
THRESHOLD FOR ENTRY



TOP 10 RISERS



CATEGORY COMPOSITION



Data is money

- 1 ton de ouro vale \$39,289,360.00
- Através da história, por volta de 185,000 tons de ouro foram mineradas
 - \$7.4 trilhões dólares
- Data G7 (Apple, Alibaba, Alphabet, Amazon, Facebook, Microsoft, Tencent) are worth \$3.7 trillion (July 2017)x

Análise de dados

- Como era:
 - As técnicas de análise de dados era validadas em um número pequeno de bases da dados.
 - Predições baseadas em computadores tinham uma baixa taxa de acerto

Data analysis

- Mas as previsões humanas muitas vezes são piores
 - The Americans have need of the telephone, but we do not. We have plenty of messenger boys (1878)
 - William Preece, Post Office Engineering Chief
 - I think there is a world market for maybe five computers (1943)
 - IBM president Thomas Watson
 - There's no chance that the iPhone is going to get any significant market share (2007)
 - Steve Ballmer, Microsoft CEO

Consultant predictions

'Expert' Disruption Forecasts

In the mid-1980s AT&T hired McKinsey & Co to
forecast cell phone adoption by the year 2000

THEIR (15-YEAR) PREDICTION

900,000

SUBSCRIBERS

THE ACTUAL NUMBER WAS

109 million

They were **off**
by a factor of:

120x



Análise de dados

- Dados geralmente contém informação relevante
 - Uma vez analisados, podem trazer benefícios
 - Sociais, políticos e econômicos
 - Aumento do interesse em análise de dados

Ciência de Dados

- A ciência de dados é o estudo dos dados para extrair conhecimento significativo para os negócios.
- Ela é uma abordagem multidisciplinar que combina princípios e práticas das áreas de matemática, estatística, inteligência artificial e engenharia da computação para analisar grandes quantidades de dados.

Ciência de Dados

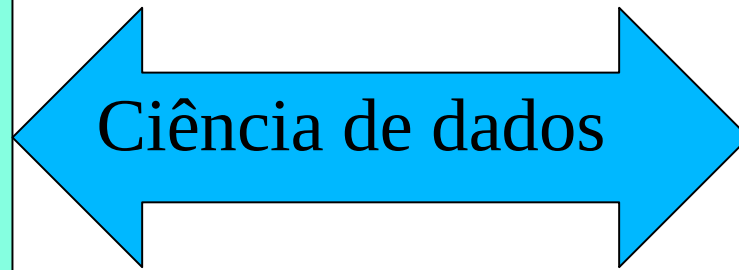
- Essa análise ajuda os cientistas de dados a fazer e responder perguntas como o que aconteceu, por que aconteceu, o que acontecerá e o que pode ser feito com os resultados.

Ciência de Dados x Analytics

- Também usadas como sinônimos
- Analytics
 - Análise de dados
 - Mais relacionada com a atividade de negócios
- Ciência de dados
 - Mais relacionada com pesquisa e inovação

Ciência de Dados

Processamento e armazenamento de dados e tecnologias de transmissão(Big Data)



Tomada de Decisão Baseada em dados

Ciência de dados

- Mineração de dados se aproxima do conceito de Ciência de Dados
 - Mas Ciência de Dados é mais amplo
 - Inclui
 - Planejamento de experimentos
 - Pré-processamento
 - Modelagem
 - Avaliação
- } Data Mining

Big Data x Ciência de Dados

- Muitas vezes são usadas como sinônimos
 - Ciência de dados: criação de soluções de modelagem de dados
 - Aptos a extrair conhecimento de dados reais
 - Big Data: tecnologias para extração e gerenciamento de dados

Engenharia de Dados

- Área que trata da transformação dos dados brutos de uma empresa.
- Projetar e implementar soluções que envolvam dados, especialmente para resolver problemas de processamento de dados em tempo real e manipular quantidades massivas de dados.

Análise de dados

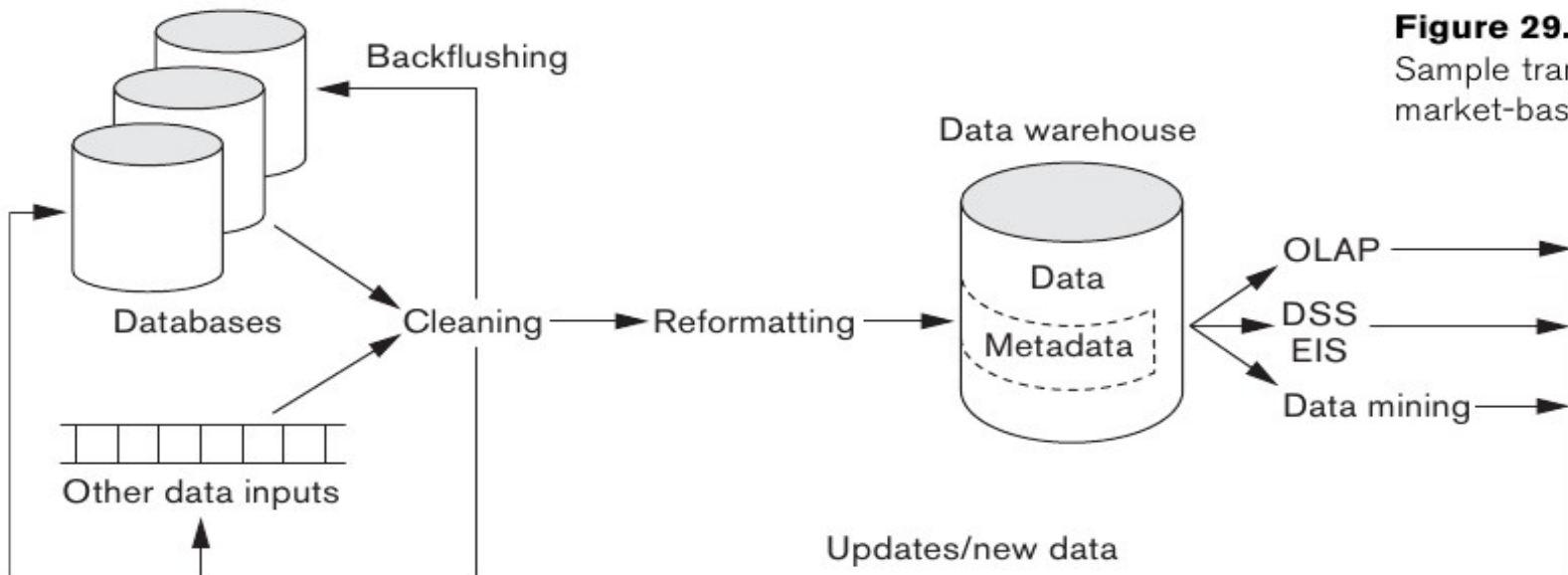


Figure 29.1

Sample transactions in market-basket model.

Referências

- ELMASRI, R; NAVATHE, S.B. **Sistemas de Banco de Dados**, Addison Wesley, 6ª Edição.
 - Capítulo 1 e 2.
- SILBERSCHATZ, A; Korth H.F.; Sudarshan S. **Sistemas de Banco de Dados**, Editora Campus, 6ª Edição.
 - Capítulo 1.
- RAMAKRISHNAN R; GEHRKE J. **Sistemas de Gerenciamento de Banco de Dados**. Mcgraw-Hill Interamericana, 3ª Edição.
 - Capítulo 1.

Material do Prof. Andre Ponce de Leon de Carvalho. USP. São Carlos