

Engenharia e Ciência de Dados: Dos Bastidores ao Palco na Gestão de Dados de Pesquisa

Altigran Soares da Silva (alti)

alti@icomcomp.ufam.edu.br



Sumário

- ▶ **Ciência de Dados:** Enorme interesse da indústria e da academia
- ▶ No entanto, nada menos que **80%** do tempo e esforço são gastos com tarefas relacionadas à preparação dos dados a serem analisados
- ▶ Aquisição, extração, deduplicação, integração, limpeza, proteção, ...
- ▶ Enorme demanda, poucas pessoas
- ▶ Trabalho pesado, propenso a erros.
- ▶ Precisa ser realizado automaticamente
 - ▶ Grande desafio para a comunidade de **Engenharia de Dados**
- ▶ Exemplos concreto
 - ▶ Resultados recentes de pesquisa na área



Ciência de Dados: Buzzword de mais de uma década

- ▶ “Information Platforms and the Rise of the Data Scientist”
 - ▶ Jeff Hammerbacher. Beautiful Data: The Stories Behind Elegant Data Solutions (2009)
- ▶ Super chique! Todos querem aprender sobre **Ciência de Dados**!
- ▶ Senso comum: **Ciência de Dados** tem a ver com aprendizagem de máquina e modelagem estatística



October 2012 Issue



October 2012 Issue

DATA Data Scientist: The Sexiest Job of the 21st Century

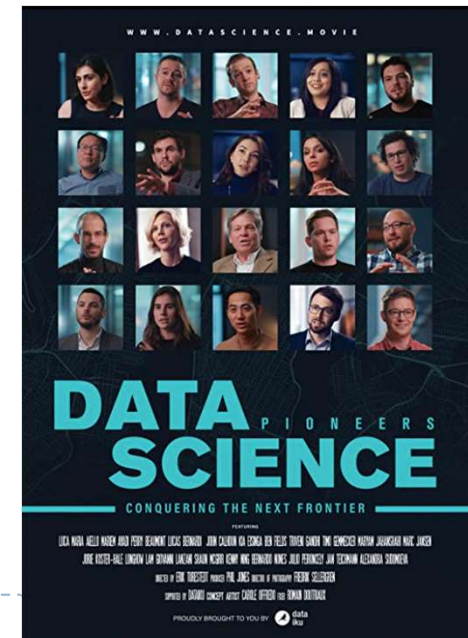
by Thomas H. Davenport and D.J. Patil
FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

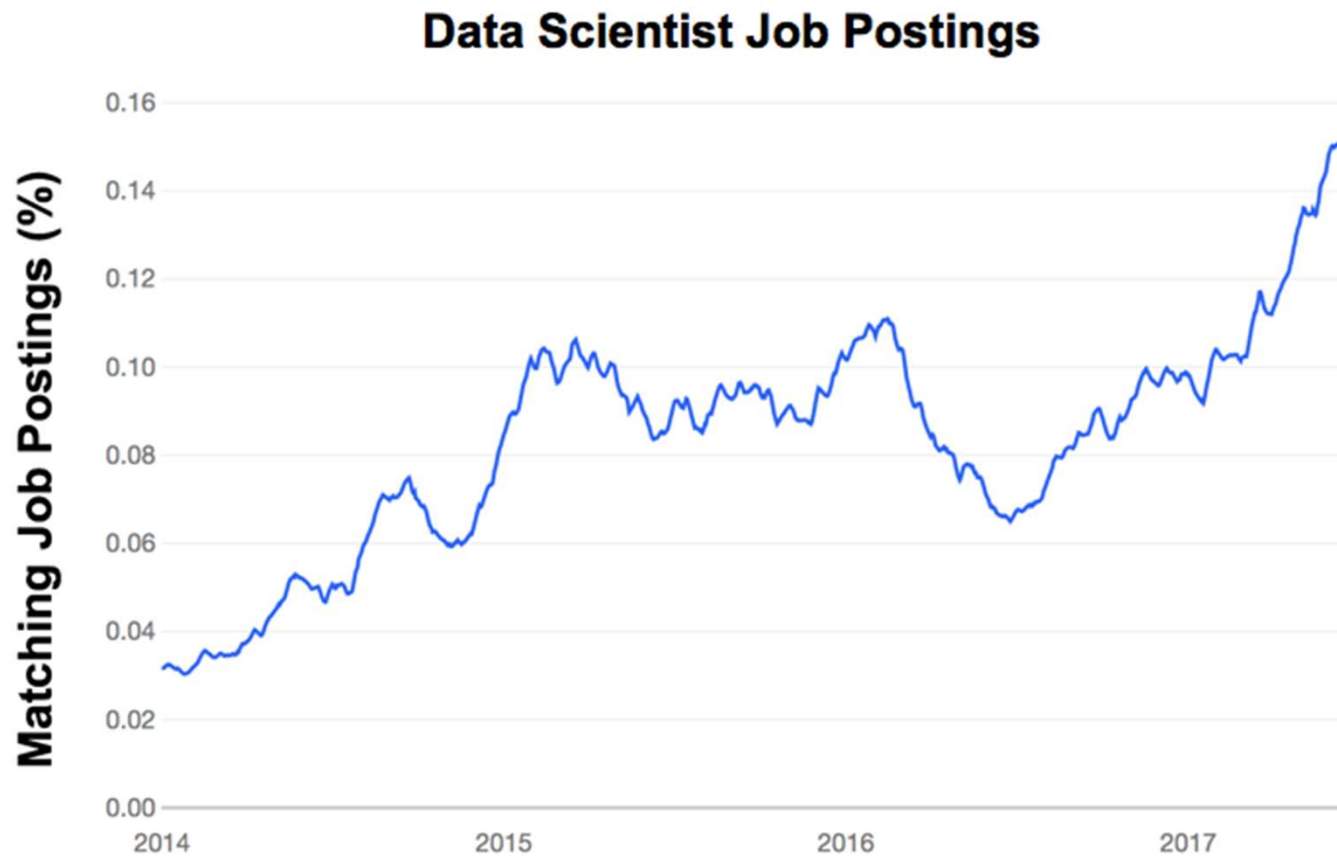
DATA Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil
FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

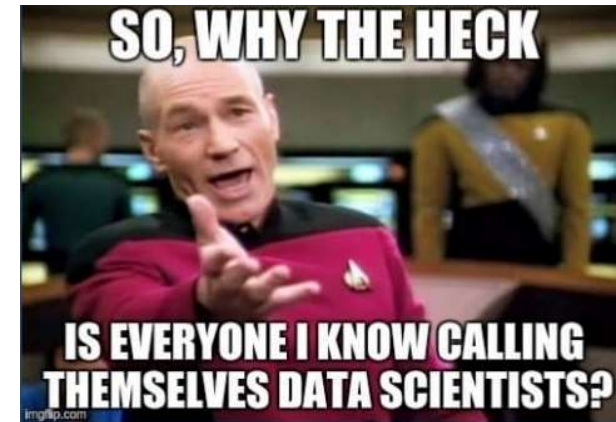


The “*sexiest job of the 21st century*”



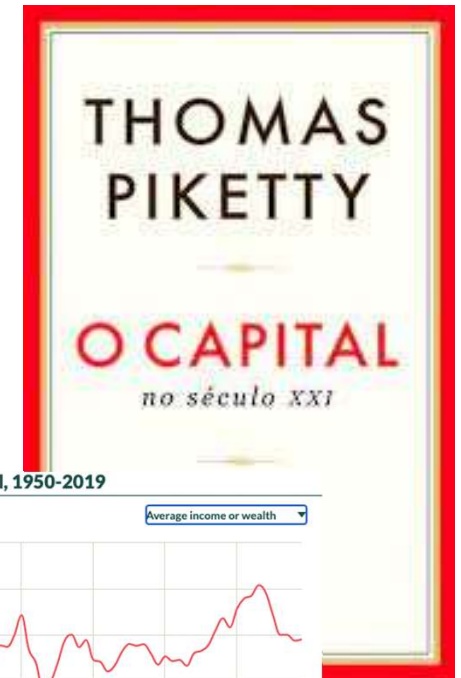
Ciência de Dados – Uma Definição (não me culpem!)

- ▶ Berman et al. (Susan Davidson, Michael Franklin, etc.)
 - ▶ *Realizing the Potential of Data Science, CACM 61(4), 2018*
- ▶ Processos e sistemas para **extrair** conhecimento ou *insights* de dados de várias formas e **traduzi-los** em ação.
- ▶ Campo interdisciplinar que integra abordagens da **estatística**, **mineração de dados** e **análise preditiva**
- ▶ Incorpora avanços da computação de alto **desempenho** e de **gerenciamento de dados**.

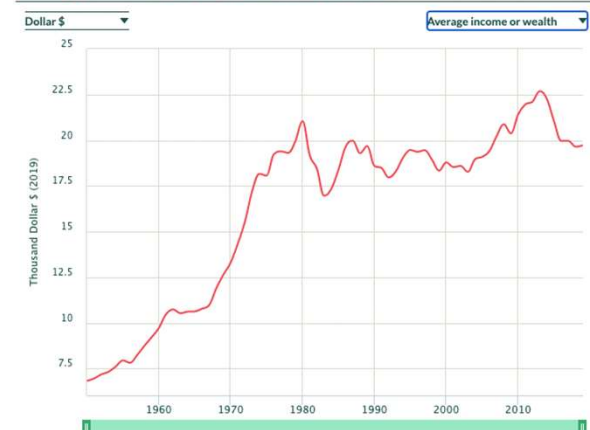


Ciência de Dados – Um dos meus exemplos favoritos

- ▶ O Capital no Século 21, Thomas Piketty (2015)
 - ▶ Baseia seu estudo sobre a distribuição de riqueza no **World Top Incomes Database (WTID)**,
 - ▶ Mais ampla base de dados históricos disponível sobre a **evolução da desigualdade de renda**.
 - ▶ Trabalho conjunto de cerca de 30 pesquisadores do mundo todo.
 - ▶ Argumenta que de modo geral todos os estudos anteriores sobre assunto, incluindo os trabalhos de Marx, **não se basearam em dados concretos** e sim em **ideologia** e **observações parciais da realidade**.



Average national income, Brazil, 1950-2019



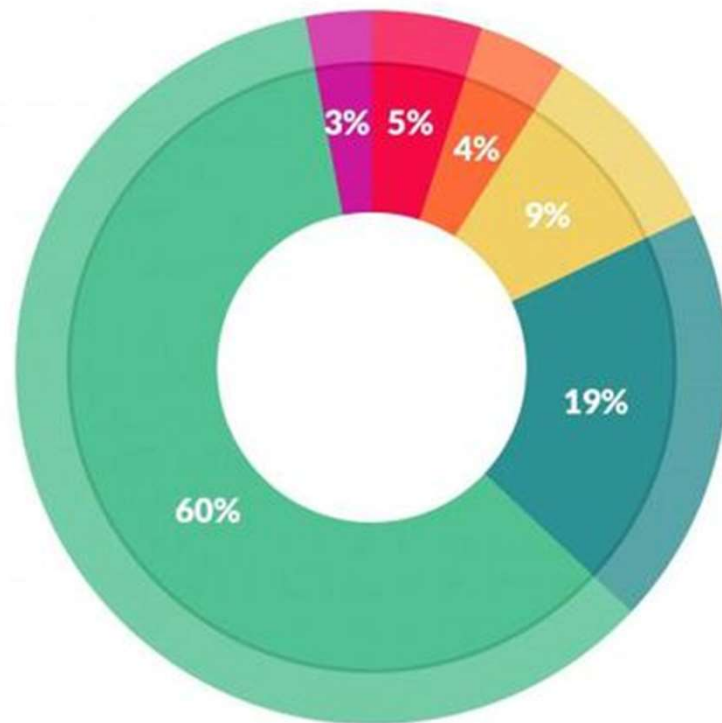
Ciência de Dados – Um dos meus exemplos favoritos

- ▶ “Durante muito tempo, os debates intelectuais e políticos sobre a distribuição da riqueza se alimentaram da abundância de preconceitos e da pobreza de fatos.”
- ▶ “Na falta de fontes, métodos e conceitos bem definidos, é possível dizer qualquer coisa e, da mesma forma, o seu oposto.”
- ▶ “As complicações de ordem tecnológica absorveram boa parte da energia dos autores e, muitas vezes, se sobrepujaram à análise e à interpretação”
- ▶ “E este livro deve muito a essa melhoria recente das condições de trabalho do pesquisador”



Ciência de Dados e Realidade (FORBES 2016)

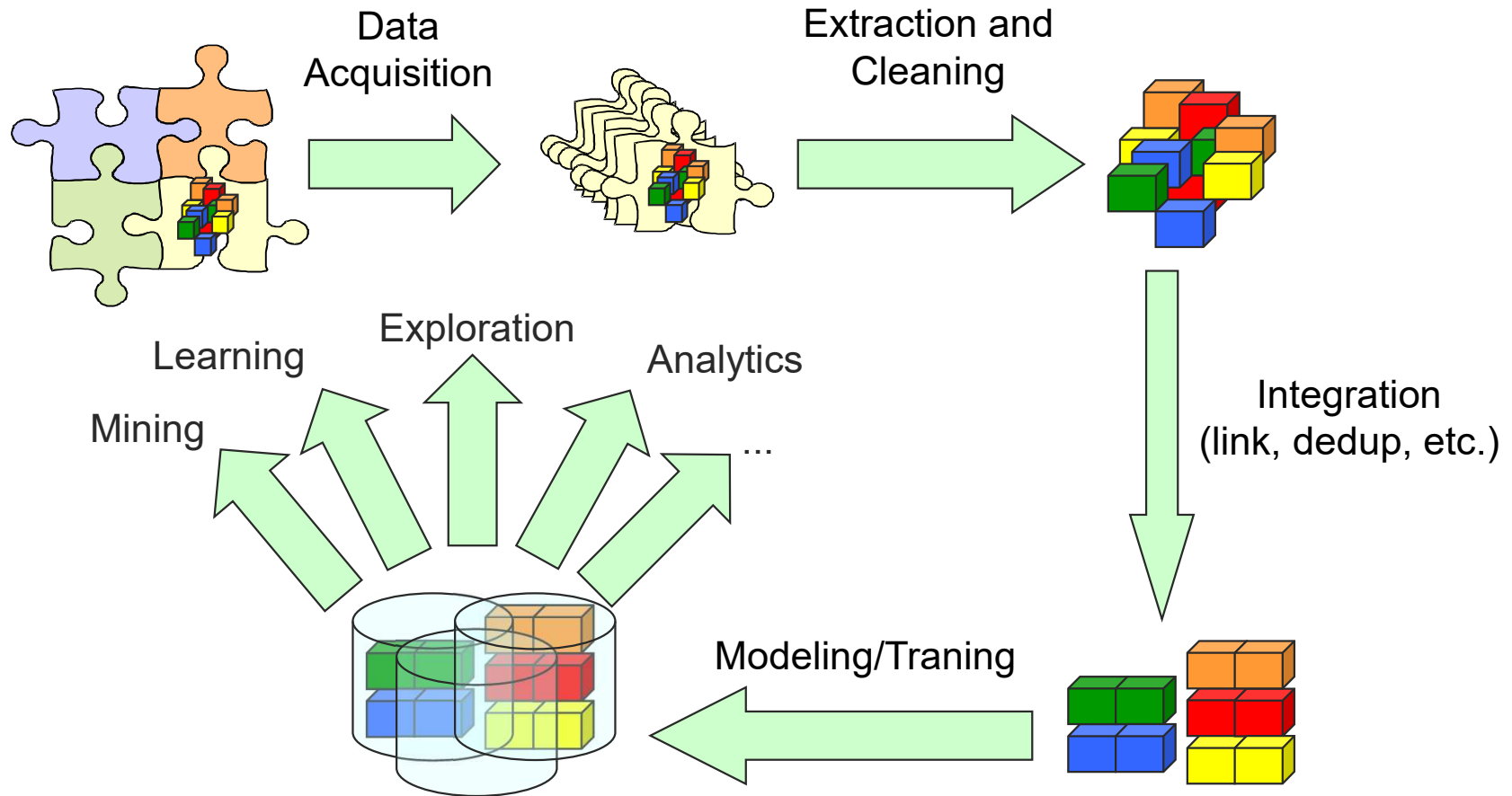
- ▶ 80% do tempo dos cientistas de dados é gasto em pré-processamento, limpeza, etc.



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Data Pipeline



Data Scientist



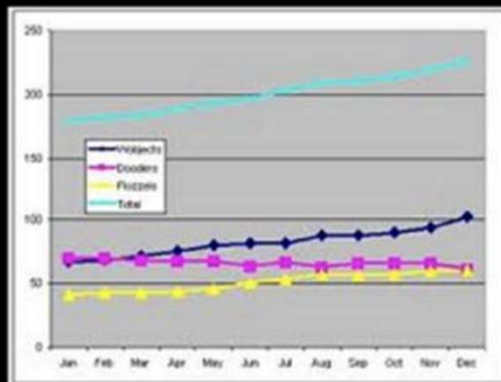
What my friends think I do



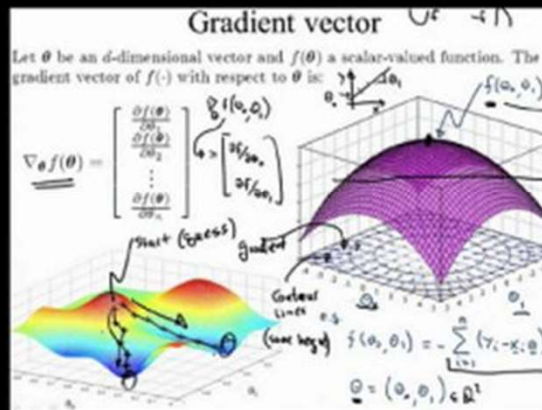
What my mom thinks I do



What society thinks I do



What my boss thinks I do



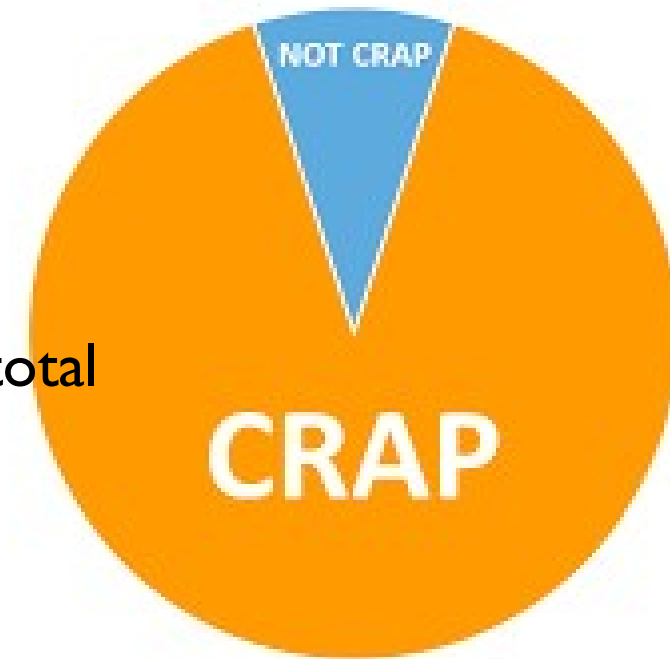
What I think I do



What I actually do

É trabalho demais ...

- ▶ O subconjunto dos dados interessantes e úteis não é conhecido previamente
- ▶ Por isso, a fase de preparação deve incluir **todo o conjunto dos dados** que são possivelmente relevantes
- ▶ Tipicamente, não mais do que **10% a 12%** do volume total de dados é realmente necessário
- ▶ O processo demanda muito tempo e recursos
- ▶ O custo aumenta exponencialmente em função do volume dados
- ▶ Complexo, propenso a erros, muitas vezes artesanal



Então, por que ter esse trabalho?

- ▶ Invariavelmente, quando mais dados, melhores os resultados
- ▶ **The Unreasonable Effectiveness of Data**
 - ▶ Halevy, Norvig & Pereira IEEE Intelligent Systems 24, 2 (March 2009)
- ▶ "Invariavelmente, **modelos simples** e um **monte de dados** vencem modelos **mais elaborados** baseados em **menos dados**"
- ▶ Alguns fenômenos são bem explicados por fórmulas elegantes
 - ▶ Física, química, astronomia
- ▶ Outros nem tanto
 - ▶ Psicologia, economia, genética, etc.
- ▶ Dados podem ajudar
 - ▶ Grandes avanços em vários campos:
 - ▶ Ex. Reconhecimento da voz, tradução automática.

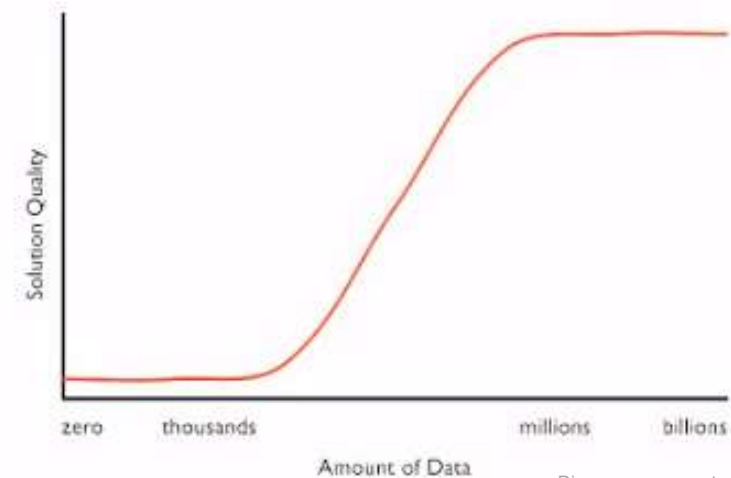
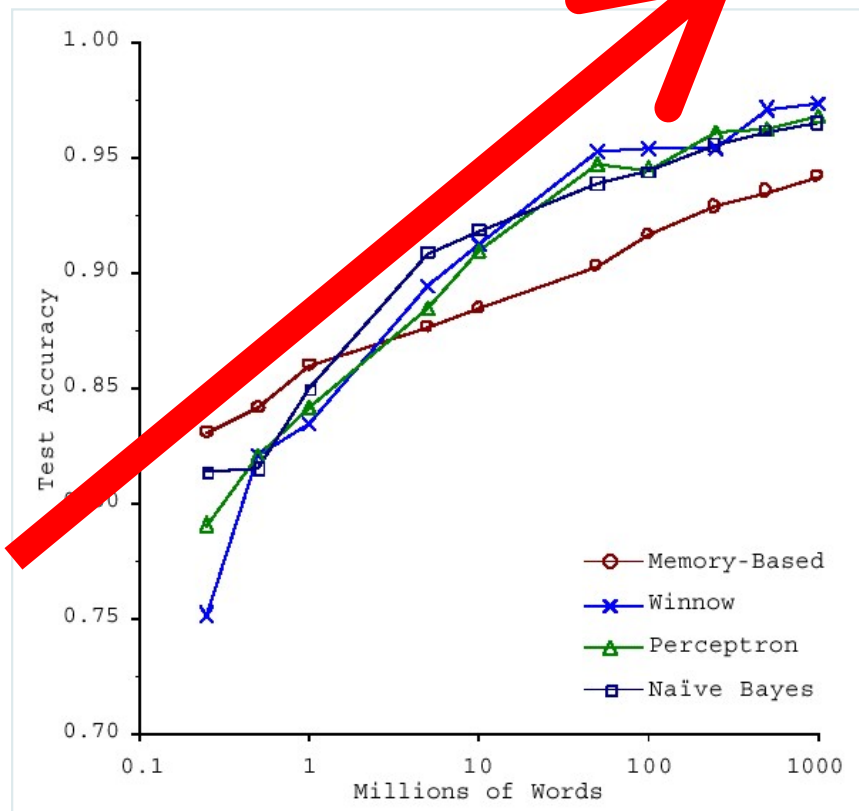
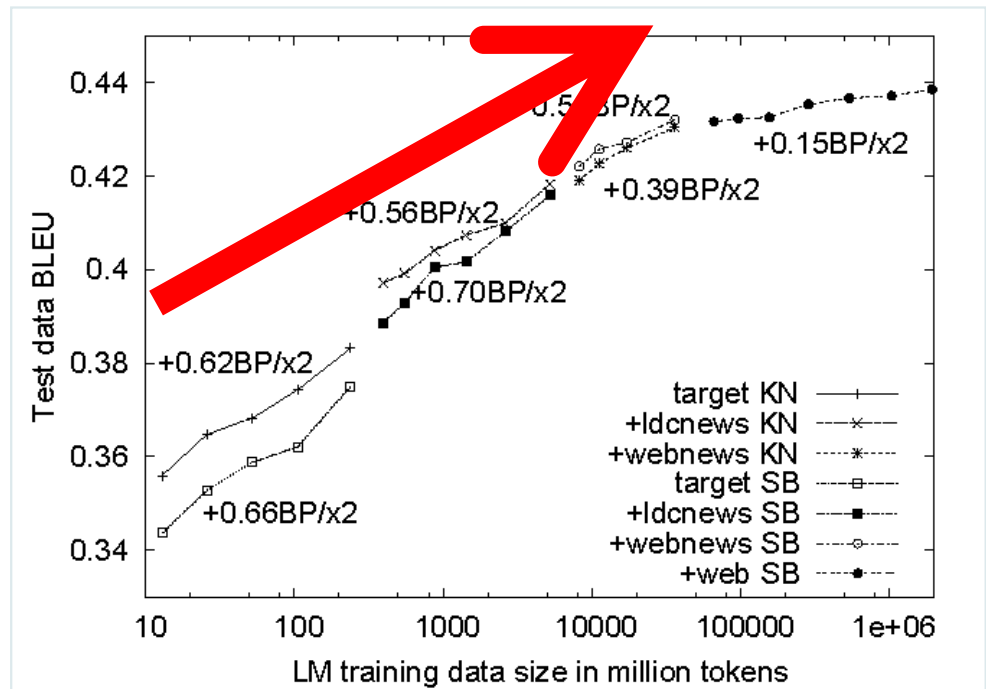


Diagrama apresentado por Peter Norvig

Mais dados fazem diferença !



Natural Language Disambiguation
(Banko and Brill, ACL 2001)



Machine Translation:
Kernesey-Ney x Stupid Backoff Smoothing
(Brants et al., EMNLP 2007)

O Que é Necessário ?

- ▶ 95% das 120.000 empresas do setor tem menos que 10 funcionários
- ▶ As organizações que podem se beneficiar de processos baseados em dados não poderão gastar recursos substanciais para realizar esses processos
- ▶ É necessário a **automação maciça de processos de Engenharia de Dados**
- ▶ A intervenção manual, se houver, deve ser limitada ao feedback de alto nível e à especificação de exceções

The Data Science Education panel @ ICDE 2017

- ▶ **Data Science Education: We're Missing the Boat, Again**
- ▶ “There is a **black art** to making our systems **sing** and **dance** at scale, even though we like to **pretend everything happens automatically**.“
- ▶ “How can we **stop pretending** and start **teaching the black art** in a principled way?”
- ▶ A second wave of data science:
 - ▶ Ethics and Legal compliance,
 - ▶ Scientific reproducibility
 - ▶ Data quality
 - ▶ Algorithmic bias

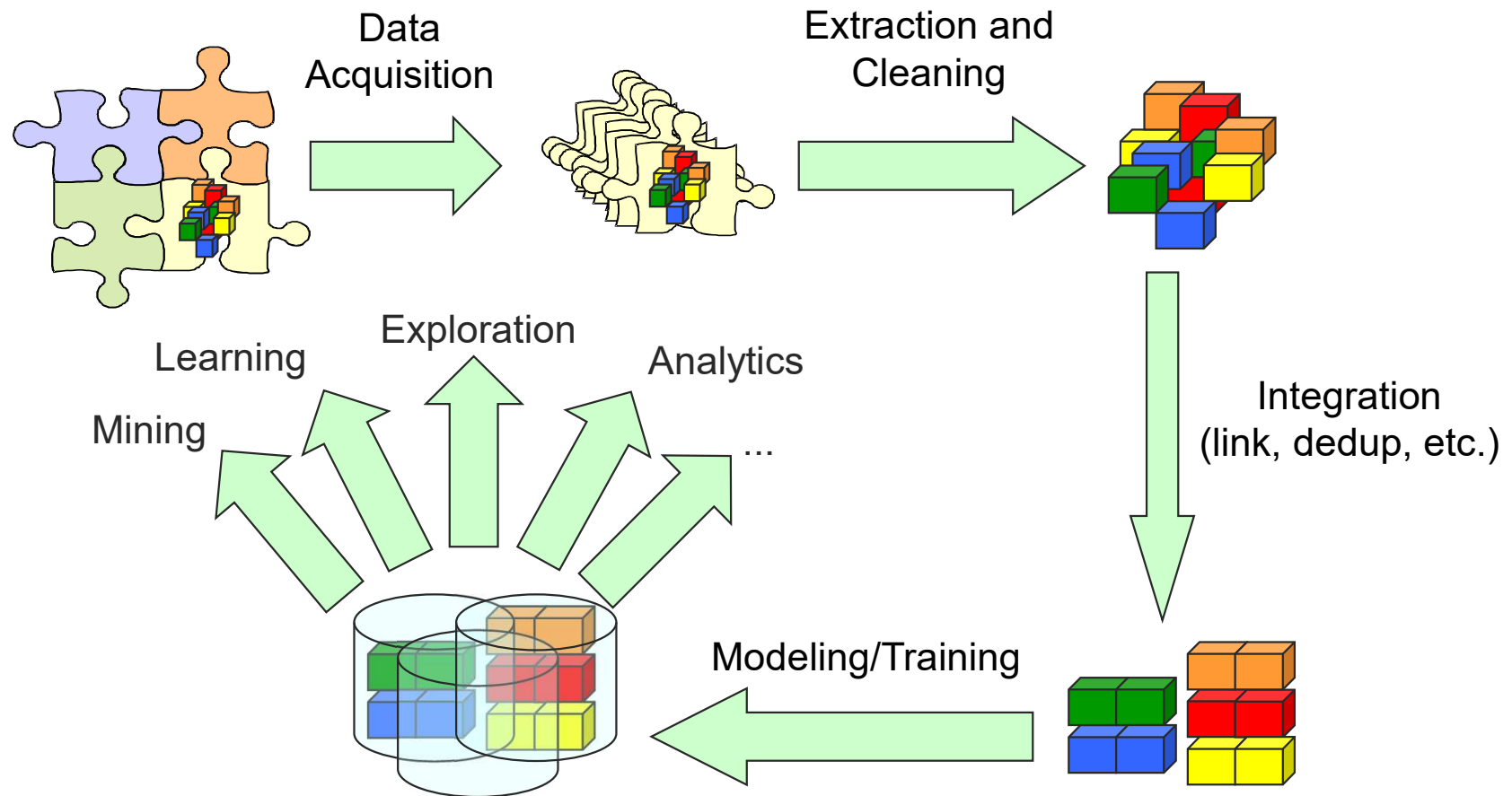


Conclusão: Ciência de Dados x Engenharia de Dados

- ▶ Para coisas acontecerem, a boa **ciência de dados** precisa de boa **engenharia de dados**
- ▶ **Preparar dados** é tão importante, relevante e difícil quanto **extrair conhecimento**
- ▶ É necessário construir **métodos**, **técnicas** e **ferramentas** para ajudar nesta tarefa
- ▶ Como no cinema, tudo acontece nos bastidores



Data Pipeline



Repositórios de Dados

- ▶ “One size doesn't fit all” → Organizações cuja atividade depende de dados geralmente não utilizam um único formato ou sistema de gerência de dados
- ▶ Cenário Comum:
 - ▶ Uma **coleção** de conjuntos de dados gerados independentemente.
 - ▶ Organização pouco **rígida** e muitas vezes **informal**.
 - ▶ Tipicamente: Catálogo de **metadados** com alguma **curadoria**.
 - ▶ É necessário ‘pescar’ os conjuntos de **dados úteis**
 - ▶ Poucas informações sobre **finalidade, valor** ou **origem**
- ▶ Dados em “silos”. Grupos que têm o “**conhecimento tribal**” sobre os dados
- ▶ Problemas: perdas **produtividade** e **oportunidades**, duplicação de trabalho e manuseio incorreto de dados.



Repositórios de Dados **Tabulares**

- ▶ Tipicamente
 - ▶ Dados estruturados de **bancos de dados relacionais**
 - ▶ **Arquivos** (semi) estruturados: CSV, logs, XML, JSON, ...
- ▶ Dados em “estado bruto” de uma organização
 - ▶ Não curados ou parcialmente curados
 - ▶ Cópias “brutas” dos dados de um sistema de origem
 - ▶ Dados transformados usados para tarefas como elaboração de relatórios, visualização e análise
- ▶ Ainda assim, podem ter um **conteúdo valioso**



Repositórios com Dados Tabulares

- ▶ **Havard Dataverse Repository** : Dados públicos de várias universidades e instituições de pesquisa em diversas disciplinas acadêmicas.
 - ▶ 75k datasets, 350 k arquivos individuais, 1,8 milhão de acessos de 150 países (Nov/2023)
- ▶ **US Government Data Lake (Data.gov):**
 - Vasto repositório de dados públicos cobrindo tudo, desde censos até dados ambientais, financeiros e de transporte.
- ▶ **AWS Open Data Lake**
- ▶ **Google Cloud Public Datasets**
- ▶ **Microsoft Azure Open Datasets**

Alguns Desafios

▶ **Heterogeneidade e Complexidade dos Dados:**

- ▶ Dados armazenados de forma “bruta” e não padronizada → dificulta o seu uso direto em análises
- ▶ Sem definições explícitas sobre como os dados de diferentes fontes se relacionam → complica a tarefa de identificar informações correlatas, essencial para a análise integrada.

▶ **Problemas de Integração e Normalização**

- ▶ Fontes distintas podem utilizar diferentes nomenclaturas, formatos e convenções, o que gera barreiras na integração de dados potencialmente relacionados.
- ▶ Frequentemente, são necessárias adaptações nos dados para garantir que elementos correspondentes de diferentes tabelas possam ser comparados ou utilizados de forma conjunta.

▶ **Implicações para a Qualidade das Análises**

- ▶ Dados não estruturados e desconectados reduzem a capacidade de realizar análises holísticas e compreensivas, limitando a extração de conhecimento.

Descoberta de Junções

- ▶ Como encontrar dados relevantes para seu propósito?
- ▶ Caso de uso comum: **descoberta de junções**
 - ▶ Descobrir tabelas que podem ser associadas a uma determinada tabela.
 - ▶ Encontrar tabelas "semanticamente" relacionadas, ou seja, com colunas que podem se associadas na forma original ou depois de alguma transformação.
- ▶ Junção ou cruzamento de dados: operação onipresente na análise de dados
- ▶ Possíveis formas de junção são obscuras

Descoberta de Junções

- ▶ Como encontrar dados relevantes para seu propósito?



Table A:

Name	Mode of Travel	Purpose	Destination	Day	Month	Year	Expense
Philip Duffy	Air	Regional Meeting	London	10	April	2019	189.06
Jeremy Oppenheim	Taxi	Exchange Visit	Ottawa	30	Jul	2019	8.08
Mark Sedwill	Air	Evening Meal	Bristol	02	September	2019	50

Table B:

Name	Date	Destination	Purpose
Clark	23/07	France	Discuss EU
Gyimah	03/09	Belgium	Build Relations
Harrington	05/08	China	Discuss Productivity

Table C:

Bird Name	Scientific Name	Date	Location
Pine Siskin	Carduelis Pinus	2019	Ottawa
American Robin	Turdus migratorius	2019	Ottawa
Northern Flicker	Colaptes auratus	2019	London



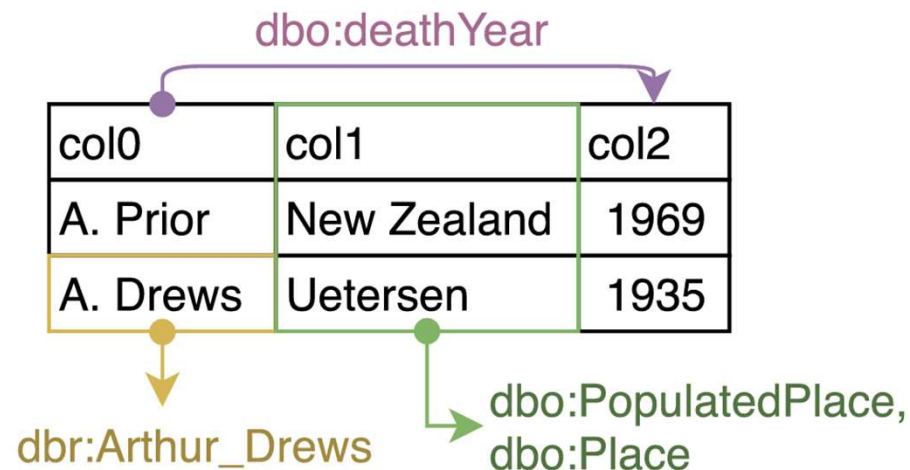
Receivers formos de junção são obscuros

Fan et al. **Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning**. PVLDB 16(7) 2023



Semantic Table Annotation (STA)

- ▶ Enriquecimento semântico dos meta-dados de tabelas
- ▶ Anotação semântica das colunas de uma tabela, atribuindo anotações semânticas aos elementos da tabela com base nos valores que ela contém.
- ▶ Usa um vocabulário predefinido: DBpedia, Schema.org, ou Wikidata.
- ▶ Aplicações:
- ▶ Integração de dados
 - ▶ Busca
 - ▶ Qualidade
 - ▶ Acessibilidade



Tarefas Associadas: CEA, CTA, CPA

- ▶ **Cell Entity Annotation (CEA):** Vincula células individuais a entidades de um grafo de conhecimento.
- ▶ **Column Type Annotation (CTA):** Atribui tipos semânticos (classes) a colunas inteiras.
- ▶ **Column-Column Property Annotation (CPA):** Identifica relações semânticas entre colunas.

Egypt	1,010,408	Cairo
Germany	357,386	Berlin

wd:Q79 ("Egypt")
wd:Q183 ("Germany")

(a) CEA

Egypt	1,010,408	Cairo
Germany	357,386	Berlin

wd:Q6256 ("country")

(b) CTA

Egypt	1,010,408	Cairo
Germany	357,386	Berlin

wdt:P36 ("capital")

(c) CPA

Desafios e Aplicações

▶ Desafios:

- ▶ Erros ortográficos e inconsistências nos dados.
- ▶ Ambiguidade: o mesmo valor pode ter diferentes significados.
- ▶ Tipologia Variada: Dados tabulares frequentemente têm diferentes estruturas e cabeçalhos pouco claros.

▶ Casos de Uso:

- ▶ Integração de dados: Facilitar a integração de dados de fontes diversas
- ▶ Busca, Q&A: Melhor a precisão de consultas em tabelas.
- ▶ Análise de Dados: Tornar dados “brutos” utilizáveis em análises mais avançadas através da semântica.

Exemplo

Tabela C1.1: Pacientes e Diagnósticos

PacienteID	Nome	Idade	Sexo	CID	DataDiagnostico	Localidade
P001	Ana Silva	45	F	J45	03/05/23	Manaus, AM
P002	Bruno Souza	60	M	I10	14/04/23	Rio de Janeiro, RJ
P003	Carla Dias	38	F	E11	19/06/23	Fortaleza, CE
P004	Diego Lima	50	M	K21	12/07/23	Salvador, BA

Tabela C1.2: Tratamentos e Prescrições

PacienteID	Medicamento	Dosagem	Frequencia	DataInicio	DataFim
P001	Salbutamol	2 mg	3x ao dia	03/05/23	17/05/23
P002	Losartana	50 mg	1x ao dia	14/04/23	14/10/23
P003	Metformina	500 mg	2x ao dia	19/06/23	19/12/23
P004	Omeprazol	20 mg	1x ao dia	12/07/23	12/01/24

Tabela C1.3: Visitas Médicas

PacienteID	DataVisita	Especialidade	Diagnostico	Observacoes
P001	10/05/23	Pneumologia	J45	Melhorou com o tratamento
P002	20/04/23	Cardiologia	I10	Necessita monitoramento
P003	25/06/23	Endocrinologia	E11	Controle glicêmico adequado
P004	15/07/23	Gastroenterologia	K21	Sintomas controlados

Tabela C2.1: Estações Meteorológicas

NomeEstacao	Tipo	Coordenadas
Norte	Urbana	-3.1, -60.0
Sudeste	Rural	-22.9,-43.2
Centro	Urbana	-15.8,-47.9
Sul	Rural	-25.4, -49.3

Tabela C2.2: Medições Diárias

ID	Estacao	Data	TMax	TMin	Precipitacao	QualidadeDoAr
M001	Norte	01/05/23	32°C	23°C	10 mm	Boa
M002	Sul	15/04/23	28°C	19°C	5 mm	Moderada
M003	Sul	20/06/23	35°C	25°C	0 mm	Ruim
M004	Centro	10/07/23	22°C	14°C	2 mm	Boa

Exemplo - CTA

PacienteID	Nome	Idade	Sexo	CID	DataDiagnostico	Localidade
P001	Ana Silva	45	F	J45	03/05/23	Manaus, AM
P002	Bruno Souza	60	M	I10	14/04/23	Rio de Janeiro, RJ
P003	Carla Dias	38	F	E11	19/06/23	Fortaleza, CE
P004	Diego Lima	50	M	K21	12/07/23	Salvador, BA

NomeEstacao	Tipo	Coordenadas
Norte	Urbana	-3.1, -60.0
Sudeste	Rural	-22.9,-43.2
Centro	Urbana	-15.8,-47.9
Sul	Rural	-25.4, -49.3

PacienteID	Medicamento	Dose	Frequencia	Observacoes
P001	Salbutamol	200µg	2x/dia	Resposta satisfatória
P002	Losartana	50mg	1x/dia	Controle de pressão
P003	Metformina	850mg	2x/dia	Controle glicêmico
P004	Omeprazol	20mg	1x/dia	Controle de refluxo

<semanticTag>geo:location</semanticTag>
 <semanticTag>geo:city</semanticTag>
 <semanticTag>geo:state</semanticTag>
 <association>
 <equivalentTo>geo:coordinates</equivalentTo>
 </association>

Estacao	Data	TemperaturaMax	TemperaturaMin	UmidadeRelativa	VelocidadeVento	QualidadeAr
Norte	01/05/23	32°C	28°C	75%	15 km/h	Bom
Sul	15/06/23	35°C	25°C	80%	20 km/h	Ruim
Sul	20/06/23	35°C	25°C	80%	20 km/h	Ruim
Centro	10/07/23	22°C	14°C	60%	10 km/h	Boa

<semanticTag>geo:coordinates</semanticTag>
 <semanticTag>geo:latitude</semanticTag>
 <semanticTag>geo:longitude</semanticTag>
 <association>
 <equivalentTo>geo:location</equivalentTo>
 </association>

PacienteID	DataVisita	Especialidade	Diagnostico	Observacoes
P001	10/05/23	Pneumologia	J45	Melhorou com o tratamento
P002	20/04/23	Cardiologia	I10	Necessita monitoramento
P003	25/06/23	Endocrinologia	E11	Controle glicêmico adequado
P004	15/07/23	Gastroenterologia	K21	Sintomas controlados

CTA usando Grandes Modelos de Linguagem

▶ **Métodos Tradicionais de CTA**

- ▶ Modelos de IA pré-treinados como **BERT** e **RoBERTa** requerem grandes volumes de dados de treinamento específicos da tarefa.
- ▶ Métodos baseados em grafos de conhecimento vinculam as entidades da tabela a um grafo para determinar o tipo de coluna.

▶ **LLMs para CTA**

- ▶ LLMs como ChatGPT têm sido usados como alternativa aos métodos tradicionais
- ▶ Vantagem: menor necessidade de dados específicos da tarefa (*zero-shot* e *few-shot*).
- ▶ Desafio: Design de prompts eficazes é fundamental para o sucesso.
- ▶ Muitos artigos recentes nos principais veículos da área escritos por vários pesquisadores importantes

Korini & Bizer: CTA using ChatGPT. VLDB (2023)

- ▶ Métodos Propostos
 - ▶ Prompt Simples (Zero-shot)
 - ▶ Instruções Explícitas
 - ▶ Uso de "*Message Roles*"
 - ▶ Aprendizagem *In-Context* (Few-shot)
 - ▶ Pipeline em Duas Etapas

Prompt Simples (Zero-shot)

- ▶ **Coluna:** Este prompt instrui o modelo a classificar uma coluna com base nos seus cinco primeiros valores

```
Classifique a coluna
fornecida em um dos
seguintes tipos:
RestaurantName, PriceRange,
Telephone, Time, Review.
```

Coluna:

```
Friends Pizza
Luigi's Pizzeria
Joe's Italian
Restaurant Bella
Napoli Mama's Kitchen
```

Tipo:

- ▶ **Texto:** Neste formato, o modelo recebe os valores da coluna como texto e é solicitado a classificá-lo como uma das classes.

```
Classifique o texto
fornecido em uma das
seguintes classes:
RestaurantName,
PriceRange, Telephone,
Time, Review.
```

Texto:

```
Friends Pizza
Luigi's Pizzeria
Joe's Italian Restaurant
Bella Napoli
Mama's Kitchen
```

Classe:

- ▶ **Tabela:** Neste formato, todo o conteúdo da tabela é dado como entrada para que o modelo faça a classificação de todas as colunas

```
Classifique as colunas da
tabela fornecida em uma das
seguintes classes:
RestaurantName, PriceRange,
Telephone, Time, Review.
```

Tabela:

```
Friends Pizza || 2525 ||
Cash || 7:30 AM || Great
pizza!
Luigi's Pizzeria || 1250 ||
Visa || 12:00 PM || Cozy
atmosphere
Joe's Italian Restaurant ||
3300
```

...

Instruções Explícitas

- ▶ Neste método, o prompt inclui instruções claras e passo a passo sobre como o modelo deve abordar a tarefa de classificação.

```
1. Considere a coluna e os tipos fornecidos.  
2. Examine os valores da coluna.  
3. Selecione o tipo que melhor representa o significado da coluna.  
4. Responda com o tipo selecionado.
```

```
Classifique a coluna fornecida em um dos seguintes tipos: RestaurantName,  
PriceRange, Telephone, Time, Review.
```

```
Coluna:
```

```
Friends Pizza  
Luigi's Pizzeria  
Joe's Italian  
Restaurant Bella
```

Uso de "*Message Roles*"

- ▶ Neste método, o prompt é dividido em diferentes papéis de message: **Sistema** e **Usuário**

O modelo deve classificar as colunas de uma tabela nos tipos semânticos apropriados com base nos dados fornecidos.

Classifique as colunas da tabela a seguir:

Friends Pizza || 2525 || Cash || 7:30 AM || Great pizza!

Luigi's Pizzeria || 1250 || Visa || 12:00 PM || Cozy atmosphere

...

Pipeline em Duas Etapas

- ▶ Neste método, o modelo é primeiro solicitado a prever o domínio da tabela e depois a classificar as colunas.

- ▶ **Passo 1:**

Classifique o domínio da tabela a seguir (exemplo: música, restaurantes, hotéis, eventos).

```
Friends Pizza      || 2525 || Cash   || 7:30 AM   || Great pizza!  
Luigi's Pizzeria  || 1250 || Visa    || 12:00 PM  || Cozy atmosphere
```

...

- ▶ **Passo 2**

O domínio é restaurante. Classifique as colunas da tabela a seguir nos seguintes tipos:

RestaurantName, PriceRange, Telephone, Time, Review.

```
Friends Pizza      || 2525 || Cash   || 7:30 AM   || Great pizza! ...  
Luigi's Pizzeria  || 1250 || Visa    || 12:00 PM  || Cozy atmosphere
```

Resultados do Experimento

► Comparação com Baselines

- **ChatGPT** (zero-shot): 89.47% de micro-F1.
- **RoBERTa** (com 356 exemplos): 89.73% de micro-F1.
- **Random Forest**: Desempenho significativamente inferior, mesmo com mais exemplos de treinamento.

► Conclusão

- **Eficiência**: ChatGPT demonstrou ser muito eficiente em termos de dados de treinamento necessários.
- **Futuro**: Explorar pipelines multi-etapas e o uso de *fine-tuning* em LLMs para melhorar ainda mais o desempenho.

Conclusão

- ▶ **Ciência de Dados**: Enorme interesse da indústria e da academia
- ▶ No entanto, nada menos que **80%** do tempo e esforço são gastos com tarefas relacionadas à preparação dos dados a serem analisados
- ▶ Aquisição, extração, deduplicação, integração, limpeza, proteção
- ▶ Enorme demanda, poucas pessoas
- ▶ Trabalho pesado, propenso a erros.
- ▶ Grande desafio para a comunidade de **Engenharia de Dados**

Agradecimentos



Pra saber mais ...

- ▶ J. Haritsa "Data Science and Astrology: Is there a Difference?".ACM India and iSIGCSE Chapter Education Webinar
- ▶ H.V. Jagadish Big Data: It's Not Just the Analytics - ACM SIGMOD Blog 2012
- ▶ F. Berman et al. Realizing the potential of data science. CACM 2018
- ▶ B. Howe et al. Data Science Education: We're Missing the Boat, Again. ICDE 2017
- ▶ D. Deng et al. The Data Civilizer System. CIDR 2017
- ▶ N. Konstantinou et al. The VADA Architecture for Cost-Effective Data Wrangling. SIGMOD 2017
- ▶ A. Halevy et al. The Unreasonable Effectiveness of Data. IEEE Intell. Syst. 2009

Pra saber mais ...

- ▶ Koutras OmniMatch: Effective Self-Supervised Any-Join Discovery in Tabular Data Repositories. CoRR abs/2403.07653 (2024)
- ▶ Cong, WarpGate: A Semantic Join Discovery System for Cloud DataWarehouses CIDR, 2023.
- ▶ Fan, Jin Wang, Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning. PVLDB (2023)
- ▶ Korini & Bizer: Column Type Annotation using ChatGPT. VLDB Workshops 2023
- ▶ Feuer ArcheType: A Novel Framework for Open-Source Column Type Annotation using Large Language Models. PVLDB (2024)
- ▶ Chen, Learning Semantic Annotations for Tabular Data. IJCAI 2019