AutoML e BigML

2023

<lapti>

Programa

- Auto ML
- BigML – histórico e empresa
- Intuição e capacidade humana
- Estrutura do BigML

AutoML

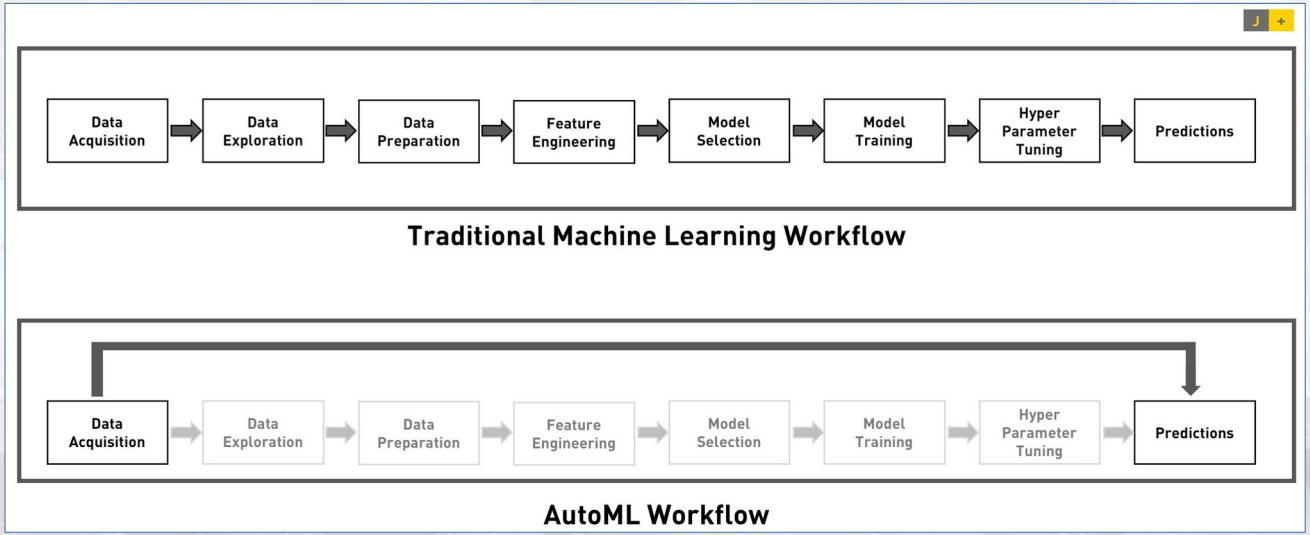
- Problemas em processos de ML
  - Seleção e/ou formação de profissionais adequados
  - Infra-estrutura necessária para execução
  - Complexidade da programação dos artefatos
  - Variação de soluções e frameworks existentes

AutoML

- Automatização da aplicação de ML a problemas existentes
  - Ou ao menos a tentativa de se fazer isso
  - *Automated Machine Learning - Methods, Systems, Challenges, The Springer Series on Challenges in Machine Learning, 2020*
- Artigo inicial em 2013, em um simpósio de Data Mining, mostrando a ferramenta Auto-WEKA
  - ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
  - *Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms, 2013*
- Criou uma tendência, dando origem ao AutoML 2014 Workshop
  - Deste então, atraiu grande interesse da indústria

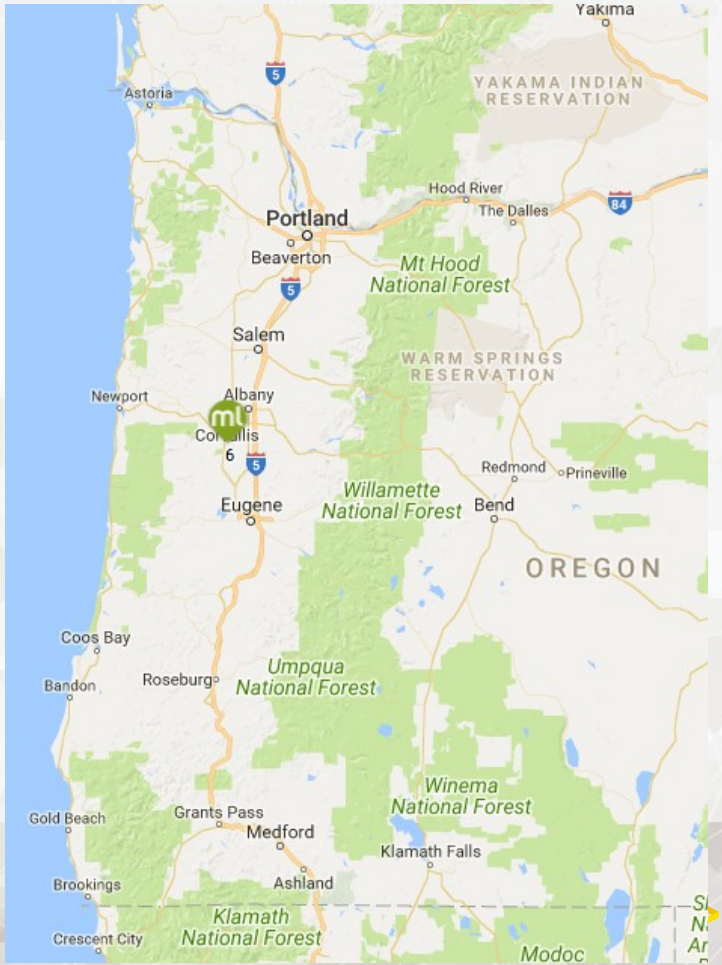
AutoML

- Simplificação no processo de desenvolvimento de modelos
  - Ainda com parametrização, quando necessário
    - Lembrando da importância de DADOS x META-PARÂMETROS
  - Eventualmente exigindo decisões do programador
    - Com conhecimentos específicos
- Ainda se aperfeiçoando, mas já considerado uma tendência importante em IA
  - Tanto em pesquisa, quando em desenvolvimento de indústria
  - *Why AutoML Is Set To Become The Future Of Artificial Intelligence, Janakiram MSV* - Senior Contributor, Forbes Magazine



BigML - Histórico

- Fundada em 2011
  - Antes de ML se tornar “cool”
    - Aproveitando o que os artigos científicos já mostravam
  - Fundador já vinha de empresas de data science e processamento de dados
    - Francisco J Martin
      - PhD em IA (Universitat Politècnica de Catalunya)
      - iSOCO – Intelligence Software Components (1999)
        - IA e web semântica
      - Strands (2004)
        - Recomendação
  - Sede oficial em Corvallis, Oregon
    - Escritório de pesquisa e desenvolvimento em Valencia



BigML - Histórico

www.gazettetimes.com/news/local/statistics-show-corvallis-no-for-patents/article_e2de00a4-2858-11e0-8594-001cc4c03286.html

MENU Gazette-Times News Sports Photo/Video Buy & Sell Obits Opinion 56°

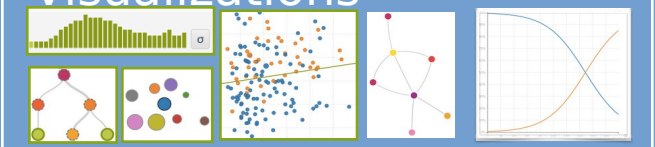
Statistics show Corvallis No. 1 for patents

By BENNETT HALL, Gazette-Times reporter Jan 25, 2011 0

BigML - Histórico

Plataforma BigML

Visualizations

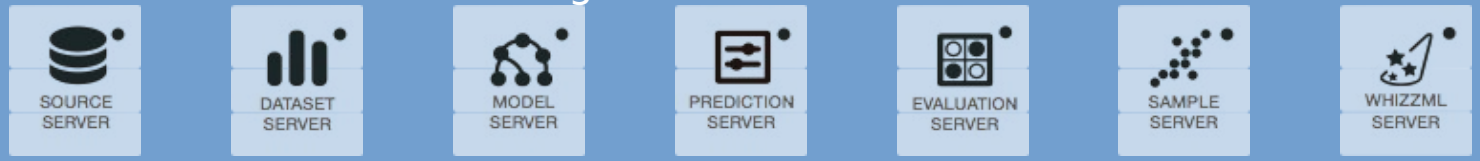


Tools

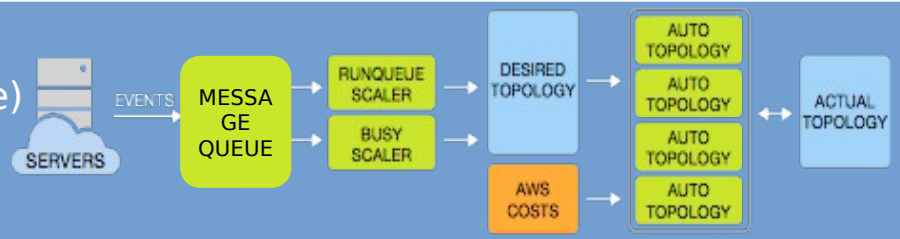
Web-based Frontend

REST

Distributed Machine Learning Backend



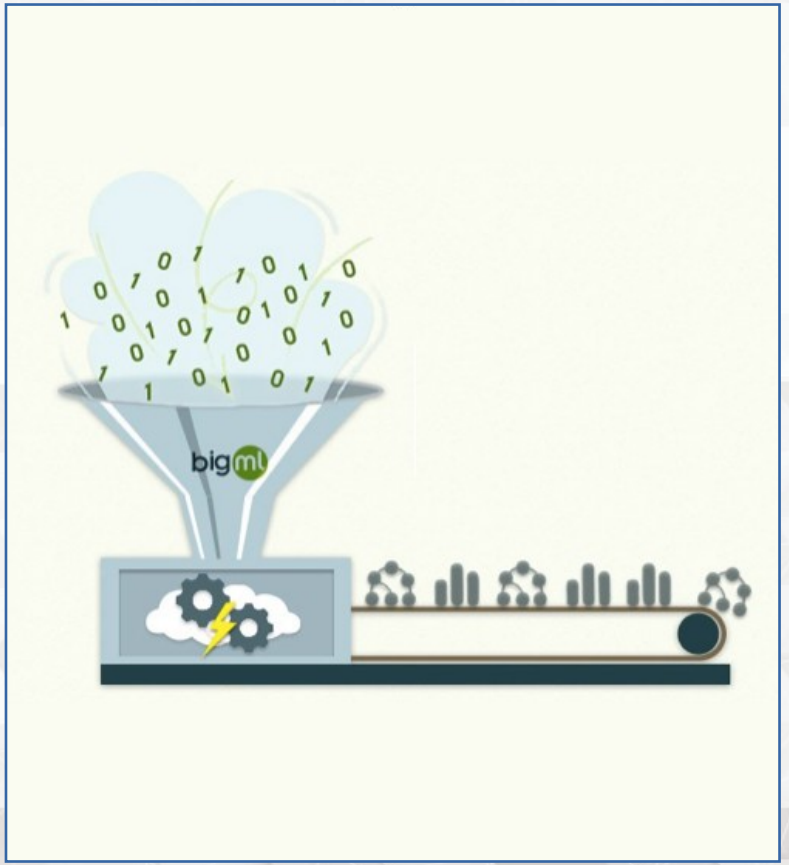
Smart Infrastructure
(auto-deployable, auto-scalable)



A capacidade humana é limitada

O que é Machine Learning?

ML transforma automaticamente grandes volumes de dados em “insights acionáveis”



Processo de ML

- Imaginando…
  - Você está procurando uma casa para comprar
  - Achou uma casa que gosta
  - Mas o preço é justo?

Processo de ML

- Porque não perguntar a um especialista do campo?
  - Especialistas são raros e/ou caros
  - Muito difícil validar essa tal experiência
    - Experiência com locais semelhantes?
    - Todas as variáveis relevantes estão sendo consideradas?
    - O conhecimento de mercado é atualizado?
  - Difícil validar as respostas…
    - Quantas vezes esse especialista acertou?
      - E quantas errou?
    - Provavelmente não conseguem explicar suas decisões em detalhe
  - Humanos não são muito bons em estatísticas intuitivas

Intuição humana

- Considerando as duas seguintes cidades:

Cidade nebulosa	Cidade ensolarada
• 350 dias com chuva ou nebulosidade • 15 dias ensolarados	• 15 dias com chuva ou nebulosidade • 350 dias ensolarados

Questão:
Em qual cidade a venda de óculos de sol (per capita) é maior?

Intuição humana

- Considerando as duas seguintes cidades:

Cidade nebulosa	Cidade ensolarada
• 350 dias com chuva ou nebulosidade • 15 dias ensolarados	• 15 dias com chuva ou nebulosidade • 350 dias ensolarados

Questão:
Em qual cidade a venda de óculos de sol (per capita) é maior?

Intuição humana

- Considerando as duas seguintes cidades:

Cidade nebulosa	Cidade ensolarada
• 350 dias com chuva ou nebulosidade • 15 dias ensolarados	• 15 dias com chuva ou nebulosidade • 350 dias ensolarados

Questão:
Em qual cidade a venda de óculos de sol (per capita) é maior?

Senso comum:
Pessoas na cidade nebulosa não precisam de óculos de sol, já que não tem muito sol mesmo...

Intuição humana

- Considerando as duas seguintes cidades:

Cidade nebulosa	Cidade ensolarada
• 350 dias com chuva ou nebulosidade • 15 dias ensolarados	• 15 dias com chuva ou nebulosidade • 350 dias ensolarados

Questão:
Em qual cidade a venda de óculos de sol (per capita) é maior?

Senso comum:
Pessoas na cidade nebulosa não precisam de óculos de sol, já que não tem muito sol mesmo...

Ocorreu a você que:
Pessoas na cidade nebulosa quase Nunca usam seus óculos de sol, e constantemente perdem eles...

Intuição humana

- Imagine o Sr. Fernández abaixo, selecionado randomicamente:

Sr. Fernández

É mais provável que o Sr. Fernández seja um bibliotecário ou um fazendeiro?

Intuição humana

- Imagine o Sr. Fernández abaixo, selecionado randomicamente:

Sr. Fernández

É mais provável que o Sr. Fernández seja um bibliotecário ou um fazendeiro?

Ocorreu a você que mundialmente existe um número estimado de mais de 1 bilhão de pessoas oficialmente empregadas na agricultura?

Paradoxo de Simpson

- Reversão de Simpson, efeito Yule-Simpson, paradoxo de amalgamação ou paradoxo de reversão
  - Fenômeno onde uma tendência aparece em diversos grupos de dados, mas desaparece ou reverte quando os grupos são combinados

Desempenho escolar (8th grade)		
	Wisconsin	Texas
2015	159	156
2011	159	153
2009	157	150

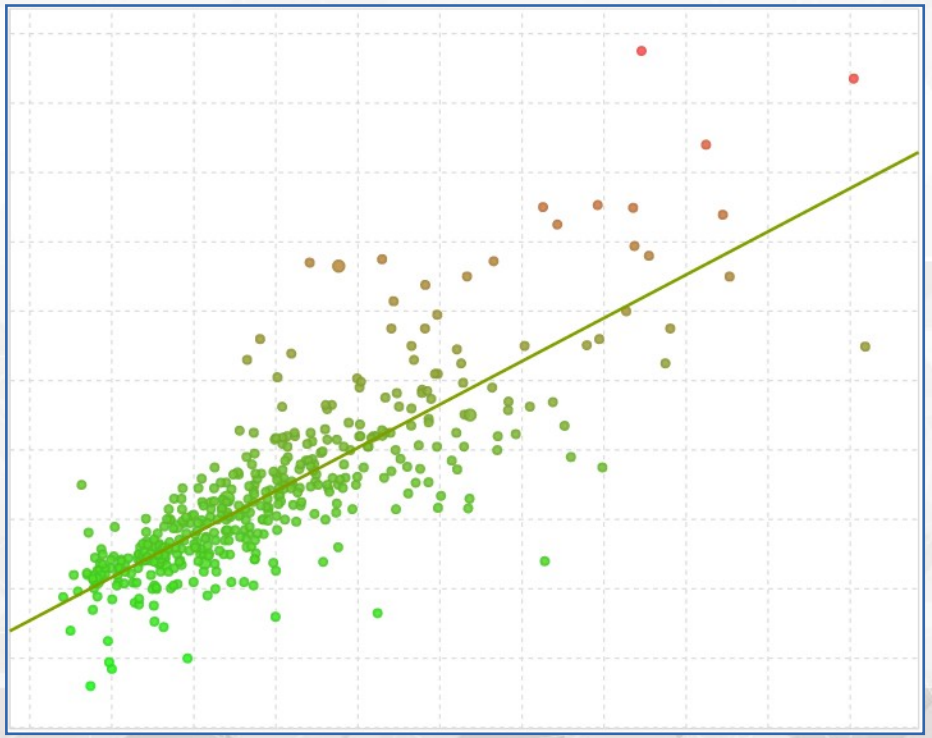
Por raça / etnicidade		
	Wisconsin	Texas
Afro-americanos	120	137
Hispânicos	138	145
Caucasianos	166	169

Dados x Especialista

- Substituir o especialista por dados?
  - Intuição: área relacionada com preço
  - Coletar dados de vendas passadas

área	preço	previsão
2424	360000	400262
1785	307500	320195
1003	185000	222211
4135	600000	614651
1676	328500	306538
1012	247000	223339
3352	420000	516541
2825	435350	450508

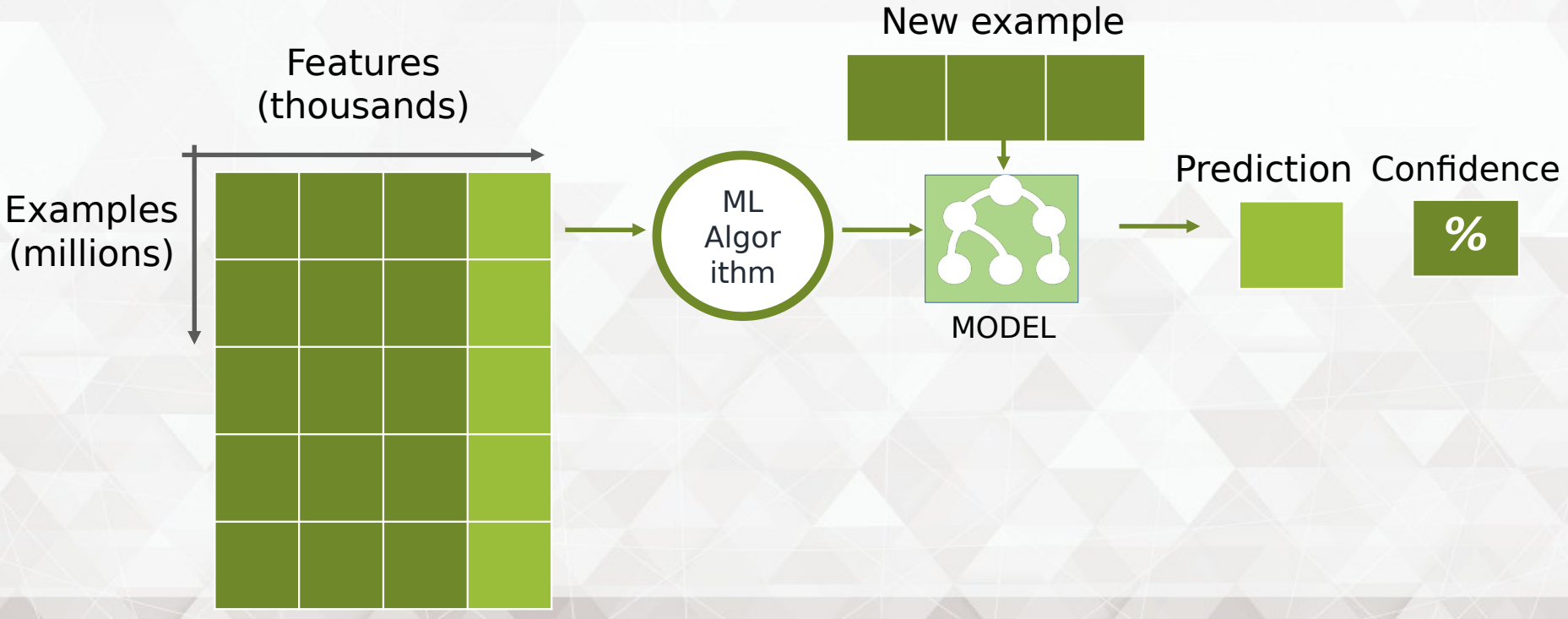
PRECO = 125.3 * AREA + 96535



Mais dados!

SQFT	BEDS	BATHS	ADDRESS	LOCATION	LOT SIZE	YEAR BUILT	PARKING SPOTS	LATITUDE	LONGITUDE	SOLD
2424	4	3,0	1522 NW Jonquil	Timberhill SE 2nd	5227	1991	2	44,594828	-123,269328	360000
1785	3	2,0	7360 NW Valley Vw	Country Estates	25700	1979	2	44,643876	-123,238189	307500
1003	2	1,0	2620 NW Chinaberry	Tamarack Village	4792	1978	2	44,593704	-123,295424	185000
4135	5	3,5	4748 NW Veronica	Suncrest	6098	2004	3	44,5929659	-123,306916	600000
1676	3	2,0	2842 NW Monterey	Corvallis	8712	1975	2	44,5945279	-123,291523	328500
1012	3	1,0	2320 NW Highland	Corvallis	9583	1959	2	44,591476	-123,262841	247000
3352	4	3,0	1205 NW Ridgewood	Ridgewood 2	60113	1975	2	44,579439	-123,333888	420000
2825		3,0	411 NW 16th	Wilkins Addition	4792	1938	1	44,570883	-123,272113	435350

Modelo de ML (dados + algoritmo)

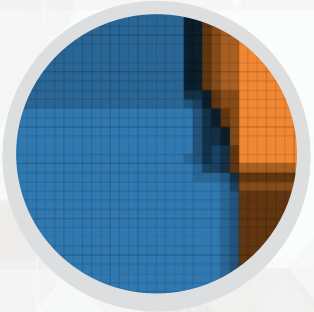


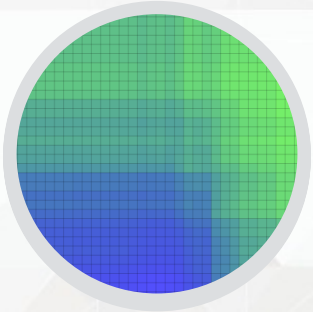
BigML - tasks

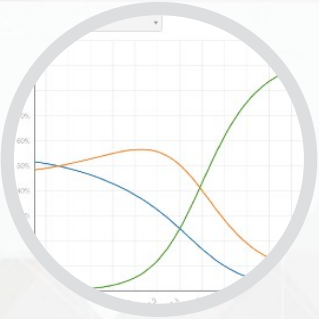
CLASSIFICATION AND REGRESSION

TIME SERIES

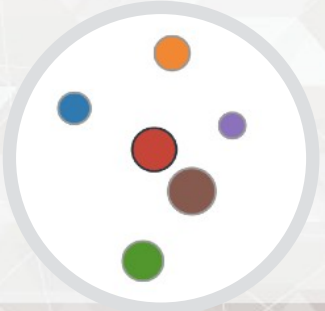
SUPERVISED

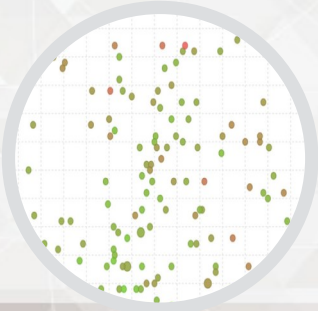


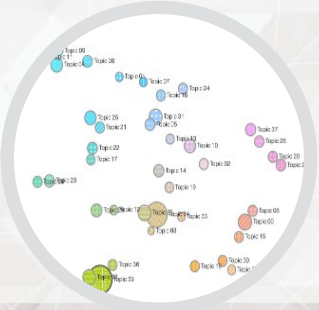




UNSUPERVISED







CLUSTER ANALYSIS

ANOMALY DETECTION

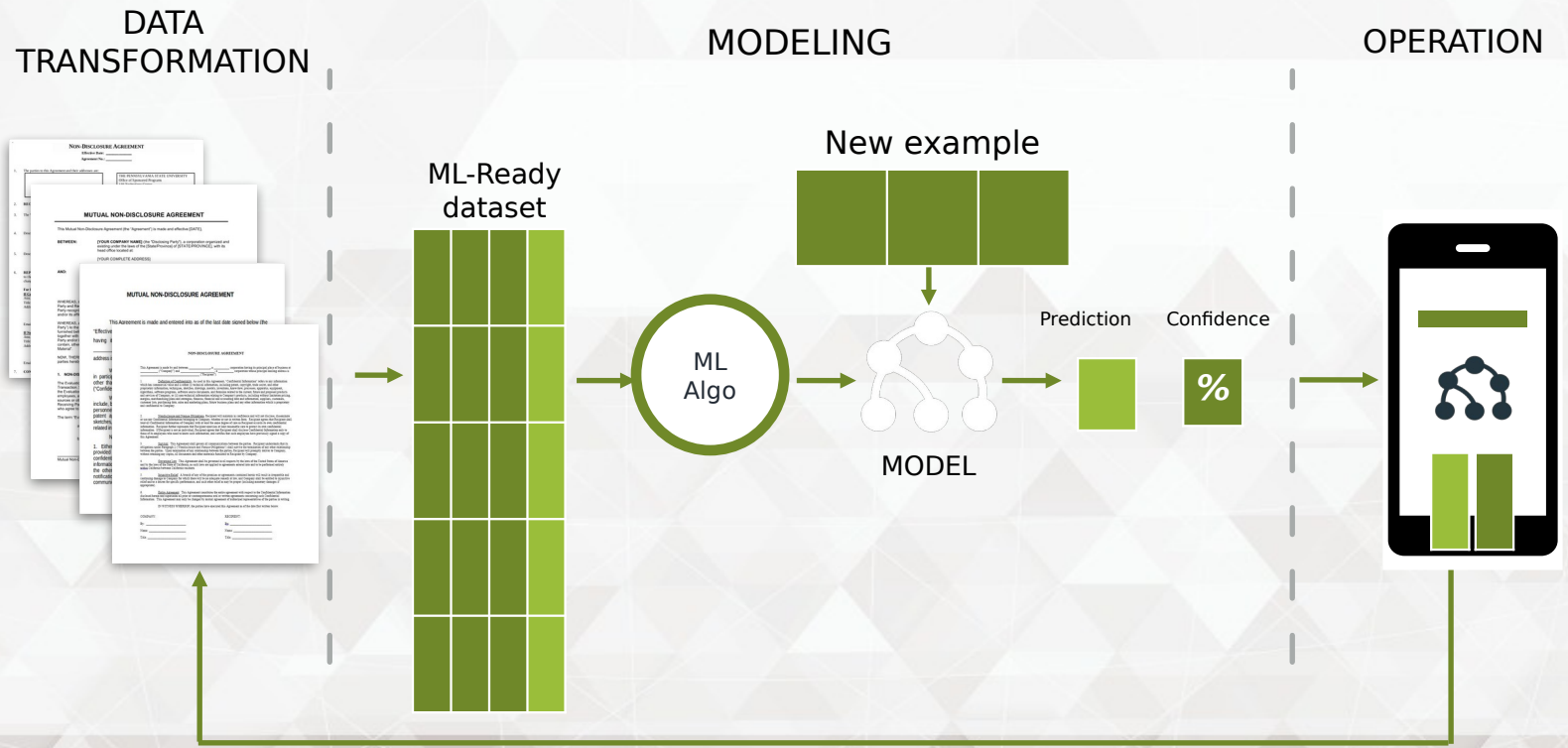
ASSOCIATION DISCOVERY

TOPIC MODELING

UTFPR UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

<lapti>

End-to-end Machine Learning



Obrigado

leandro@utfpr.edu.br
http://lapti.ct.utfpr.edu.br

<lapti>