Aprendizado por Reforço

AULA - 6

Off-Policy Learning

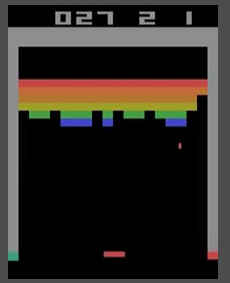
Anteriormente...

- Diferença Temporal

$$V(s) \leftarrow V(s) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s)]$$

- Deep Q-Network

$$L(\theta) = \left(r + \gamma \max_{a'} Q(s', a'; \phi) - Q(s, a; \theta) \right)^2$$



Anteriormente...

- Gradiente de Política

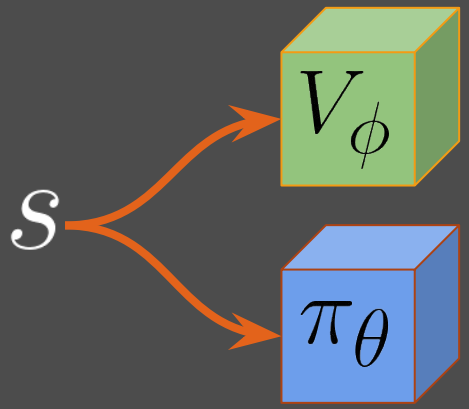
$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) R_{i,t}$$

- Com Baseline

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) [R_{i,t} - V_\phi(s_t)]$$

Anteriormente...

- Actor-Critic



$$V_\phi(s) \quad L(\phi) = (R - V_\phi(s))$$

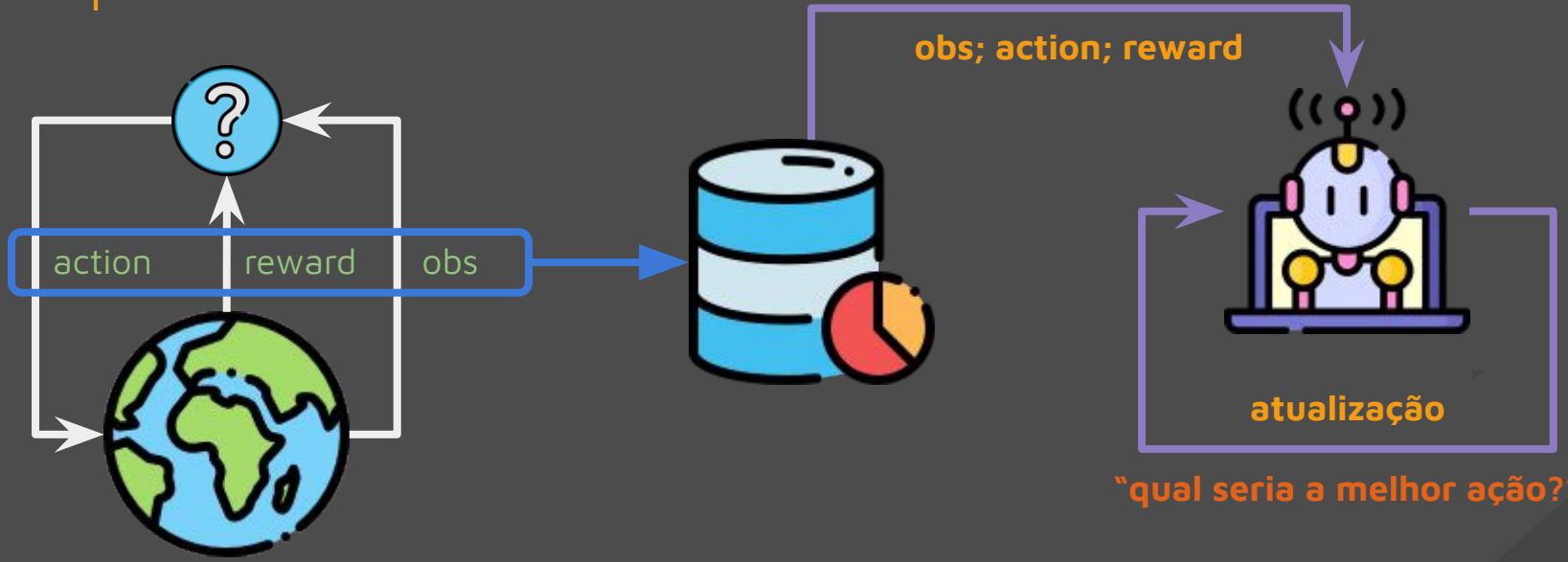
$$\pi_\theta(a|s) \quad L(\theta) = \log \pi_\theta(a|s) \hat{A}(s,a)$$

A3C - PPO - SAC

Off-Policy

O que é aprender *Off-Policy*?

- Como o nome diz "fora da política"
- Com experiências de outras políticas
- A partir de um Dataset



Off-Policy vs Imitation Learning

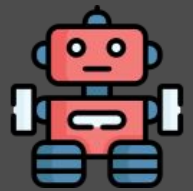
- Aprende com as experiências
- Maximizar recompensa em situações do dataset
- Aprende com experiências de qualquer política

- Imita comportamento
- Minimizar erro entre ações do dataset e do modelo
- Política que coleta experiências deve ser expert

O que é Offline e Online Learning?

- **Offline**: Não há acesso ao ambiente, agente aprende com experiências previamente coletadas
- **Online**: Há acesso ao ambiente, agente coleta novas experiências com uma certa frequência






Por que Off-Policy/Offline RL é Importante?

- O mundo possui uma quantidade enorme de dados
- Podemos usar esses dados para:
  - Minimizar Custos
  - Minimizar Riscos
  - Aproveitar Experiências

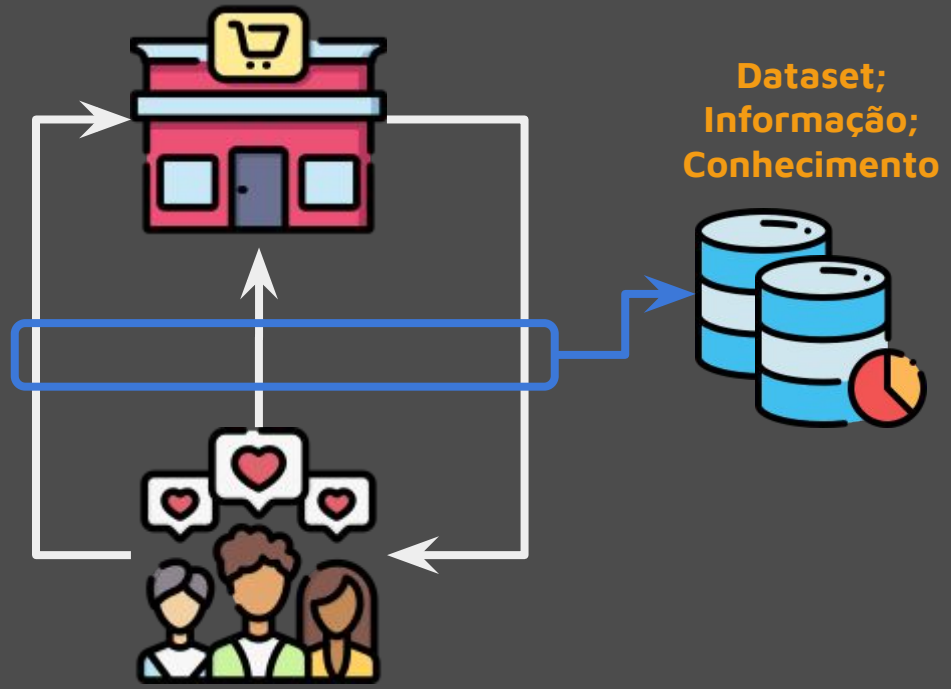
No Mundo Real...

Marketing

Ofertas de Produtos

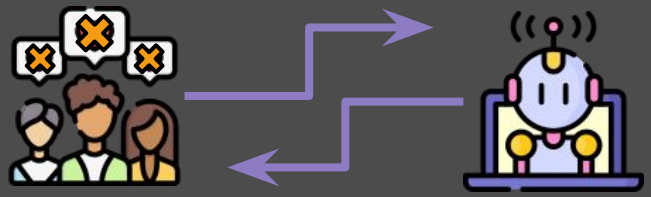
Alocação de Recursos

Horários de Funcionamento

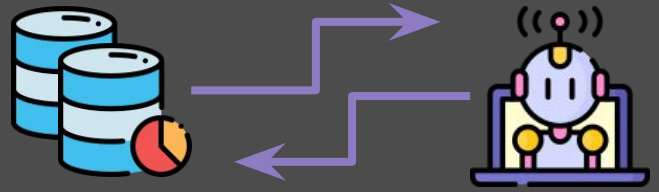


Dataset; Informação; Conhecimento

Dados ou Ambiente?



- Interagir com o ambiente pode ser caro e arriscado, pois erros precisam ser cometidos para explorar e aprender
- Permite exploração
- Otimização direta do ambiente

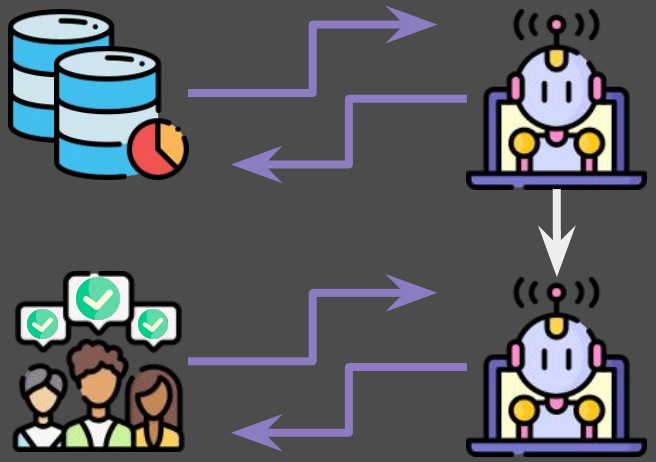


- Interagir com os dados é seguro, e, muitas vezes, mais rápido.
- Aproveita o conhecimento de políticas de coleta (humano ou modelos)
- Não há espaço para exploração
- Otimização limitada pelos dados

Podemos ter Ambos?

1

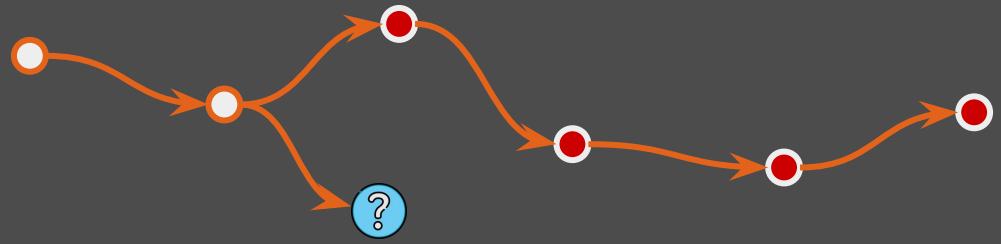
2



- Se a situação permitir
- Treino com dados primeiro
- Atuação no ambiente depois
- Aprender no ambiente
- Modelo menos aleatório
- Aplicação com menos riscos

Problema de Superestimação

- Treinar com dados nem sempre é fácil
- Se os dados apresentam um caminho ruim, é fácil para o modelo superestimar o caminho não percorrido



- Quanto mais representativo, menor esse problema
- Dados muito representativos são incomuns, porque geralmente alguma política está sendo seguida para coletá-los

Conservative Q-Learning (CQL)

- Adiciona o valor de Q na função de perda
- Para minimizar a função, Q deve ser pequeno
- Cria um limite inferior
  - Menor Q possível que ainda faça boas predições
- Subtrair o valor de ações que vêm do dataset
  - Minimiza apenas o valor de ações que estão fora do dataset

Conservative Q-Learning for Offline Reinforcement Learning
https://arxiv.org/abs/2006.04779

Model-Based

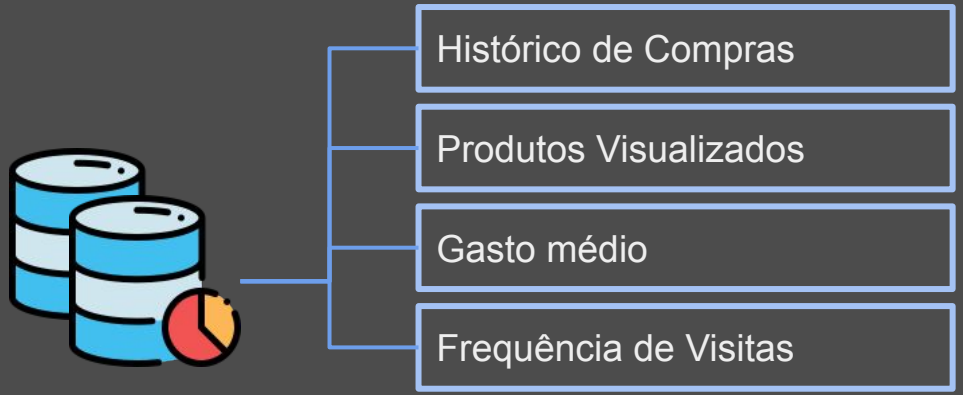
- MOPO
  - Cria um ensemble de modelos do ambiente
  - Iterativamente treina modelos e política
  - Política é treinada com estados e recompensas inferidos pelos modelos
  - No treino da política é adicionado um fator de discordância dos modelos
- MOReL
  - Treina um modelo a partir das experiências
  - Usa o modelo para emular o ambiente, prevendo estado e recompensas
  - Treino com o ambiente "falso"

MOPO: Model-based Offline Policy Optimization
https://arxiv.org/abs/2005.13239

MOReL : Model-Based Offline Reinforcement Learning
https://arxiv.org/abs/2005.05951

Dificuldades com os Dados

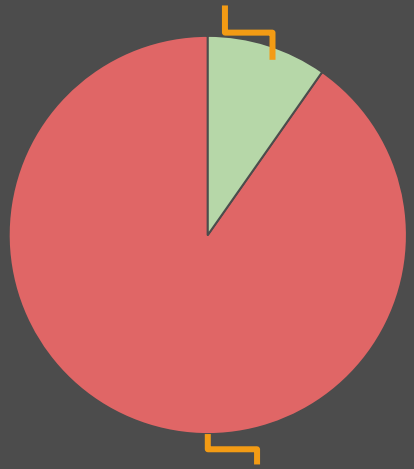
- Informações Disponíveis
- Há informações suficientes para tomar decisões?
- Construção da representação do estado
- Construção da Recompensa



Dificuldades com os Dados

- Desproporcionalidade
- Talvez é necessário repensar as recompensas ou filtrar os dados

Sequências que geram Recompensas Positivas



Sequências que geram Recompensas Negativas/Zero

Dificuldades de Avaliação

- Como avaliar uma política sem acesso ao ambiente?
- *Off-Policy Evaluation*

- Direct Estimator
- Importance Sampling
- Weighted Importance Sampling
- Doubly Robust

Off-Policy Estimators

- Tão bons quanto os dados disponíveis

Projeto com a Acordo Certo

- **Problema**
- Trazer usuários para a plataforma para quitar suas dívidas através de mensagens SMS

- Não se pode acionar usuários com menos de 3 dias do último contato
- Não se pode acionar todos os usuários (são muitos)

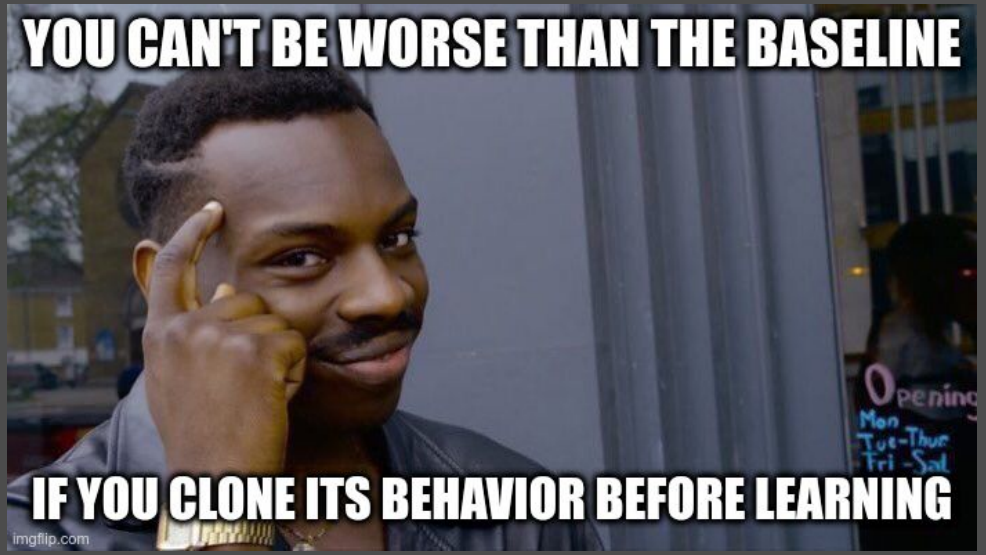
Dados Disponíveis

- Montar Recompensa
  - cadastro +
  - acordo ++
  - acionamento -
- Montar Estados
  - Estado Markoviano
  - Histórico
  - Último acionamento?
  - Quantos acionamentos?

- Histórico de Acionamentos
- Ações no Site (cadastro/acordo)
- Profissão
- Valor da Dívida
- Renda média

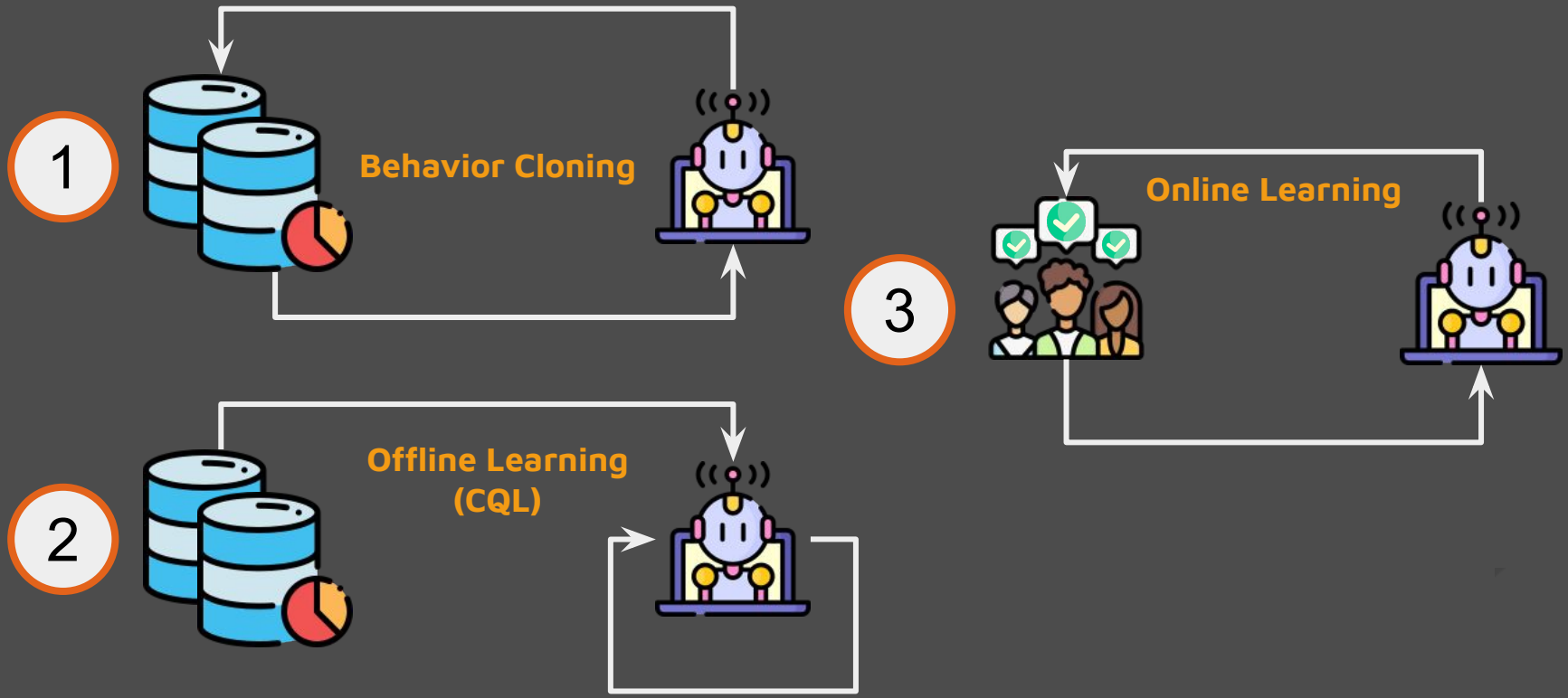
Quem Coletou os Dados?

- Modelos previamente desenvolvidos pela empresa
- Há inteligência nos dados / nas trajetórias

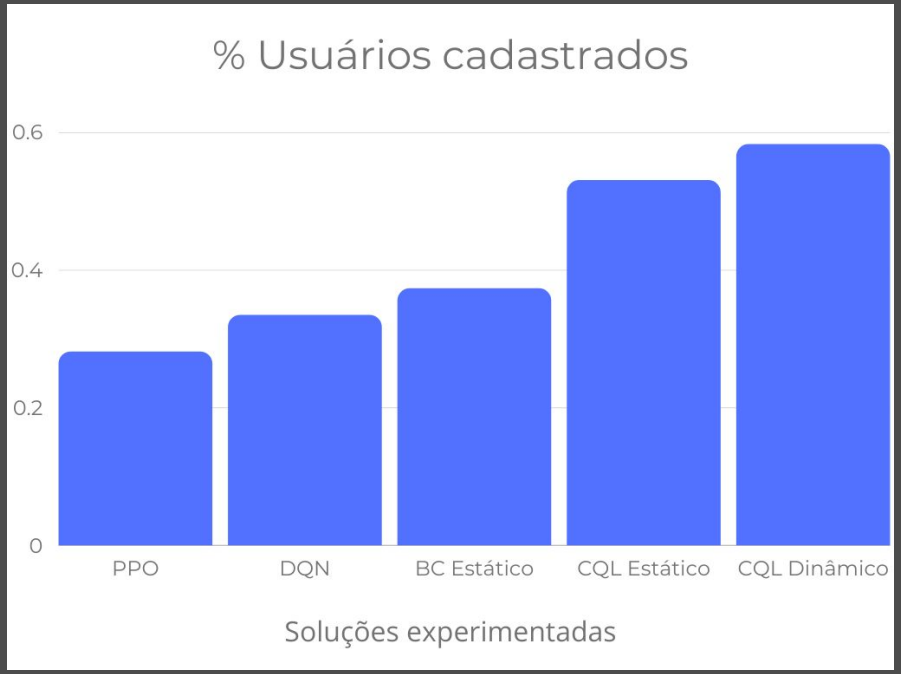


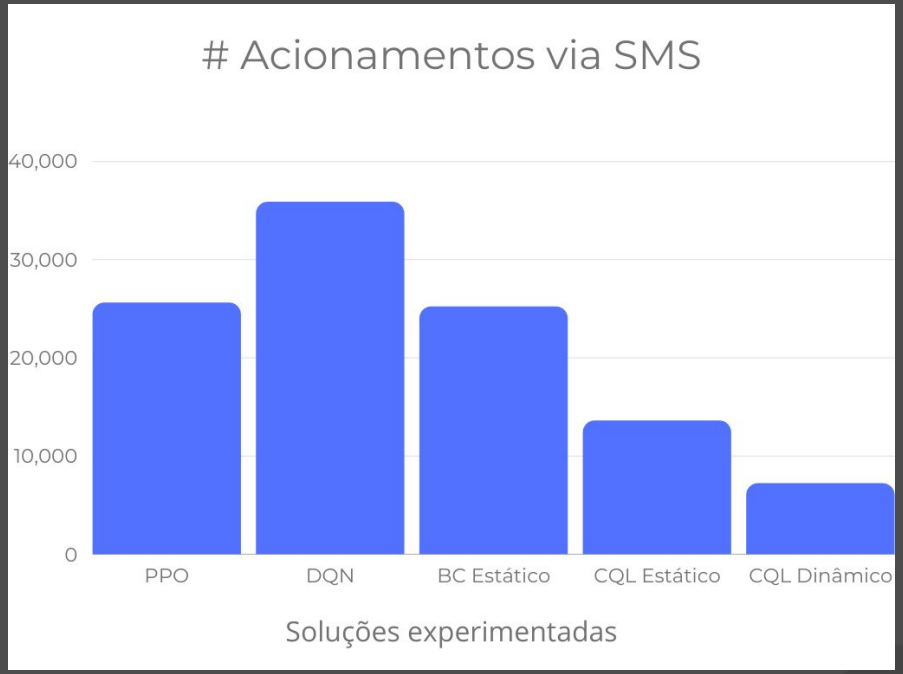
*quase verdade

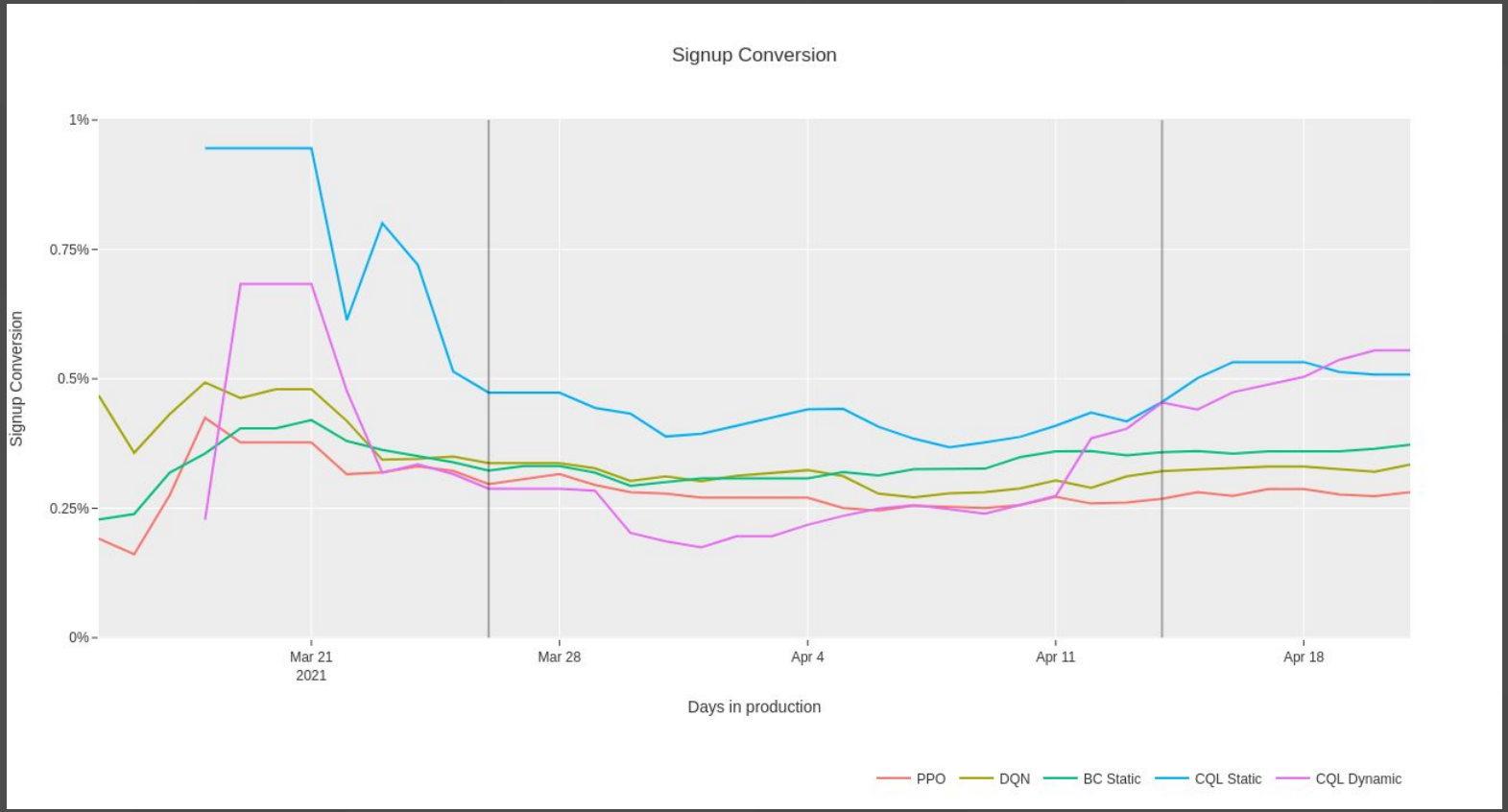
Fluxo de Treino



Resultados


% Usuários cadastrados
0.6
0.4
0.2
0
PPO
DQN
BC Estático
CQL Estático
CQL Dinâmico
Soluções experimentadas


Acionamentos via SMS
40,000
30,000
20,000
10,000
0
PPO
DQN
BC Estático
CQL Estático
CQL Dinâmico
Soluções experimentadas



Conversão em Cadastros

É isso aí . . .

Próxima Aula: Tópicos Avançados