

Sistemas de Recomendação




Quem sou?

- **Carreira em desenvolvimento na área de TI, com foco na liderança de equipes de projetos de desenvolvimento de aplicativos de CTI, URA, gravadores digitais e software de reconhecimento de fala**
- **Desenvolvimento de soluções de IA e plataforma de assistentes virtuais**
- **Profissional com certificações da FGV e USP**
- **Coordenador de Arquitetura TI - Telefônica**

Visão Geral

A decorative network diagram in the top right corner, featuring a complex web of interconnected nodes and lines, rendered in a light blue color.


O TensorFlow é uma biblioteca criada pela equipe **Google Brain Team** junto com o Grupo de pesquisa **Google's Machine Learning Intelligence** com o propósito de conduzir pesquisa relacionadas as áreas de Aprendizado de Máquina e Aprendizado Profundo.

A decorative network diagram in the bottom left corner, featuring a complex web of interconnected nodes and lines, rendered in a light blue color.

Visão Geral

A decorative network diagram in the top right corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue.

O TensorFlow combina a álgebra computacional com técnicas de otimização de compilação, facilitando o cálculo de muitas expressões matemáticas em que o problema é o tempo necessário para executar a computação.

A decorative network diagram in the bottom left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue.

Conceitos Básicos

- Uso transparente da computação em GPU, automatizando o gerenciamento e otimização da mesma memória e dos dados utilizados. Você pode escrever o mesmo código e executá-lo em **CPUs** ou **GPUs**. Mais especificamente, o TensorFlow irá descobrir quais partes do cálculo devem ser movidas para a GPU

Run TF at scale with CMLE



Cloud ML Engine

tf.estimator

High-level API for distributed training

tf.layers, tf.losses, tf.metrics

Components useful when building custom NN models

Core TensorFlow (Python)

Python API gives you full control

Core TensorFlow (C++)

C++ API is quite low level

CPU

GPU

TPU

Android

TF runs on different hardware

Quero saber...



Quero fazer...

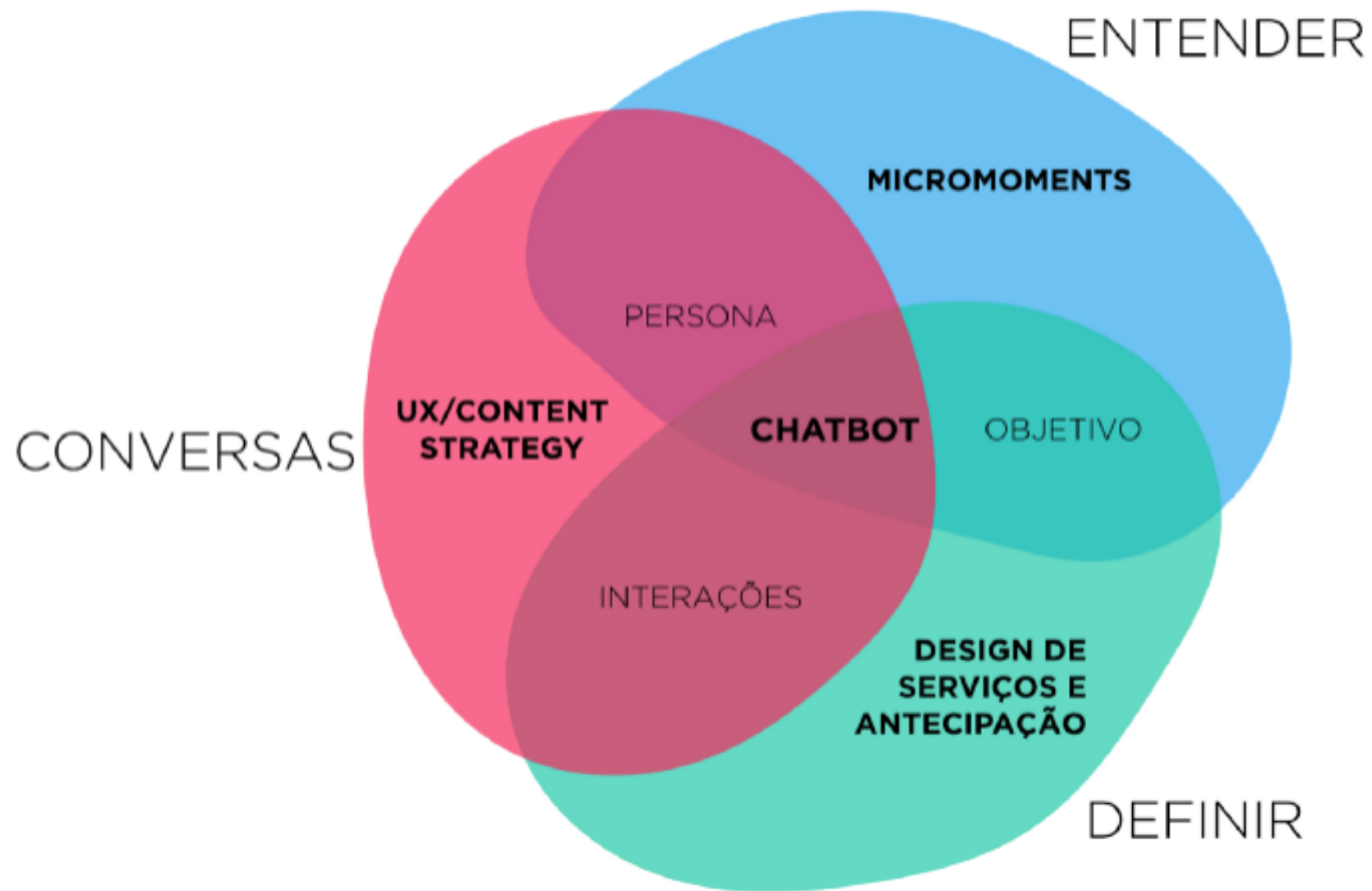


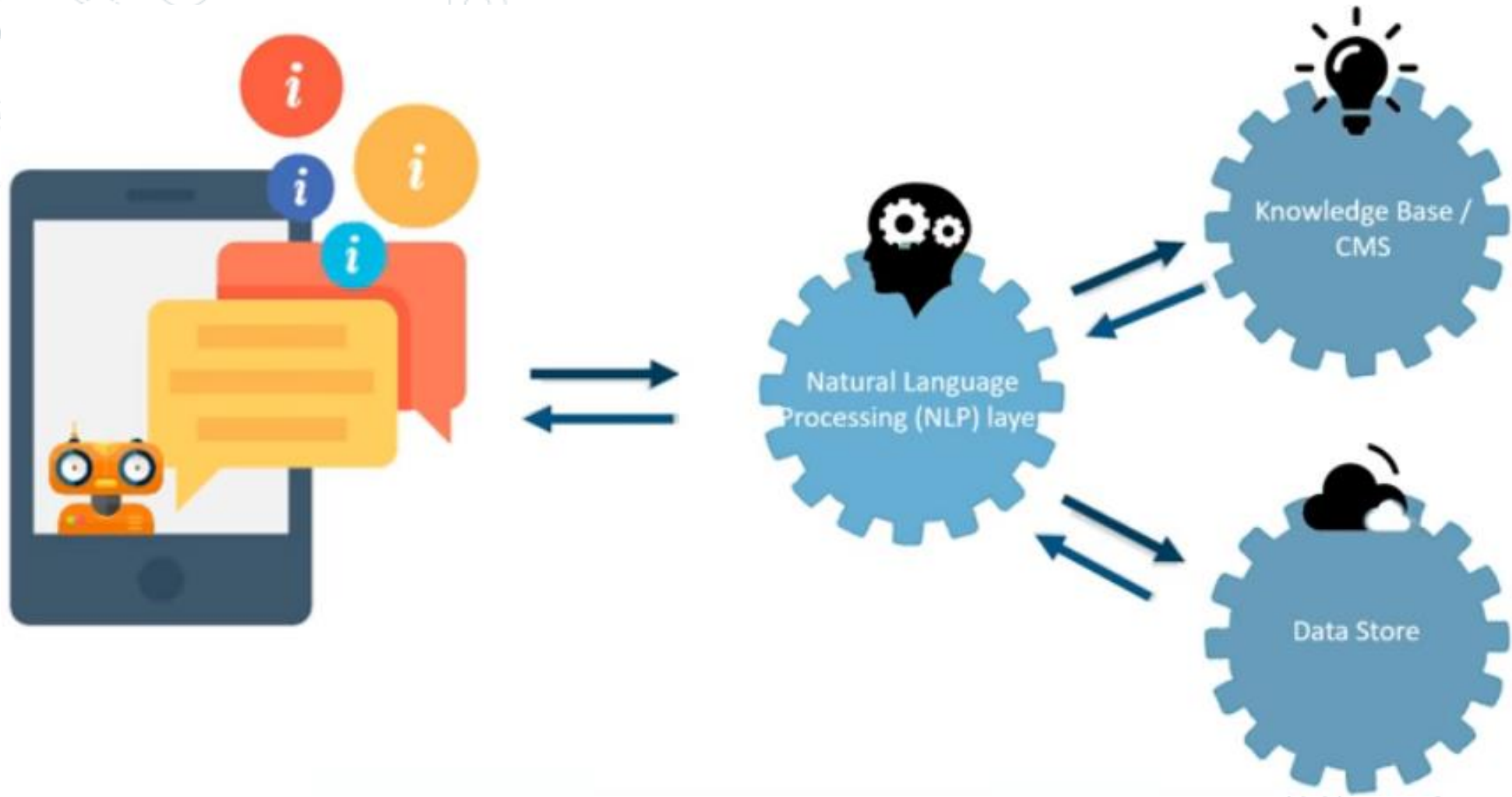
Quero comprar...



Quero ir...







Natural Language Processing

Processamento de Linguagem Natural:

Intenção: é o desejo que o chatbot perceberá que o usuário possui ao enviar uma mensagem específica. Por exemplo: ao enviar um “quero saber meu saldo” a intenção do usuário é “saldo”.

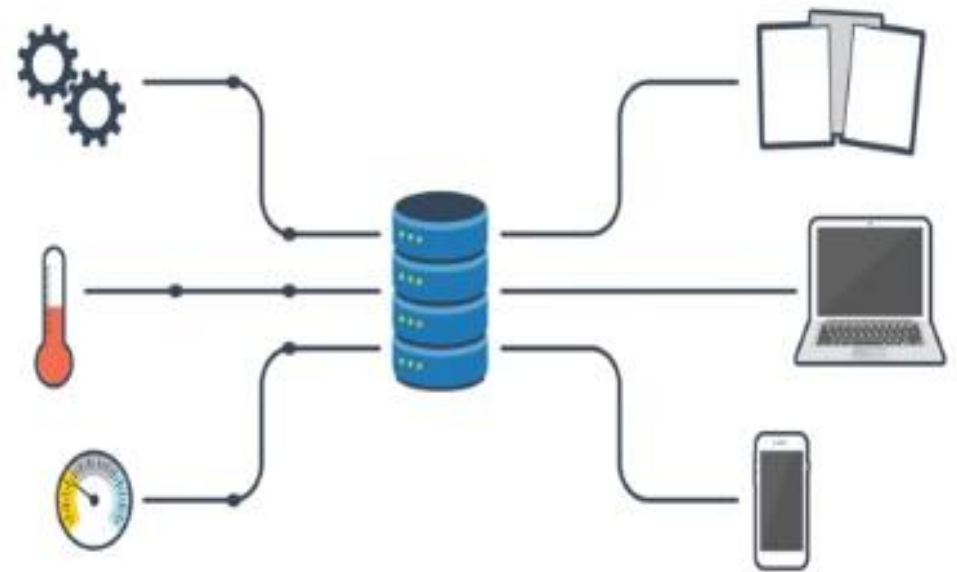
Entidade: é o substantivo relacionado ao desejo que foi detectado pelo chatbot. Por exemplo: Se a frase enviada pelo usuário for “saldo de conta corrente”, a entidade poderia ser “conta corrente” ou “poupança”

Diálogo: é onde se cria as respostas que o chatbot retornará ao detectar uma intenção e/ou uma entidade. Por exemplo: se o chatbot detectar que a intenção é “saldo” ele poderia responder “o valor do saldo é R\$123,00”.

Data Store



- Data required to train the bot
- User's chat comes to bot once it's deployed.



Content Management System

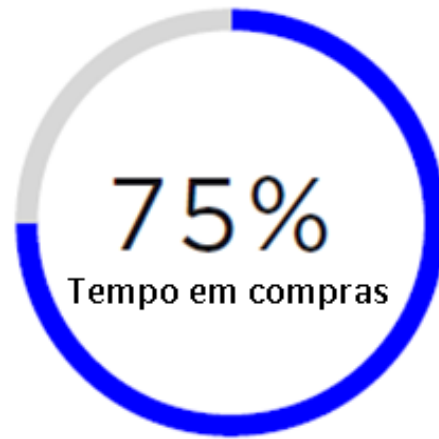
- Real-time, personalized customer experiences ●
- Universal content accessibility & personalization ●
- Ability to reach and retain customers ●



Pesquisas mostram que a experiência do consumidor é importante



dissem que a personalização tem algum impacto na decisão de compra



é gasto na procura de produtos e pesquisa on-line por 50% dos clientes



querem melhor tempo de resposta



dentro das empresas continuam inexplorados

Source : <http://www.nextopia.com/wp-content/uploads/2015/01/personalization-ecommerce-infographic.png>
<https://blog.hubspot.com/blog/tabid/6307/bid/23996/Half-of-Shoppers-Spend-75-of-Time-Conducting-Online-Research-Data.aspx>
<http://possible.mindtree.com/rs/574-LHH-431/images/Mindtree%20Shopper%20Survey%20Report.pdf>
<http://www.getelastic.com/using-big-data-for-big-personalization-infographic/>

RecSys: Motivações e Aplicações

- Motivos para implantar um Sistema de Recomendação
 - Aumentar o número de itens vendidos
 - Vender itens mais diversificados
 - Aumentar a satisfação dos usuários
 - Aumentar a fidelidade dos usuários
 - Melhorar o gerenciamento dos itens
- SRs estão sujeitos à falhas
- Problema está na consistência dos dados!



Métricas

- Identificar o melhor algoritmo de recomendação é um desafio
 - Discordância sobre os atributos e métricas
- Problemas ao avaliar os algoritmos:
 - Algoritmos dependem do conjunto de dados
 - Objetivos da avaliação podem variar

Algoritmos

- Utilizam como base informações e atributos de usuários e itens para recomendar que estão disponíveis no sistema. Estes são os principais componentes avaliados com base em diferentes critérios, cada qual com sua abordagem.



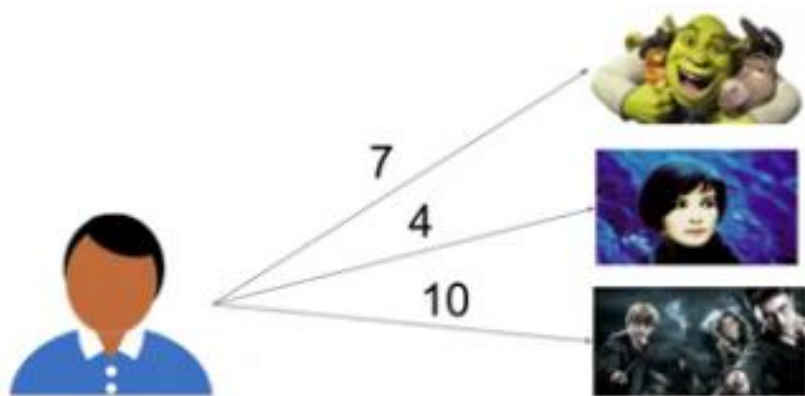
Netflix: Sistema de Recomendação

- Sistema de Recomendação Netflix
 - Tudo se torna uma recomendação!
 - Recomendações arranjadas em grupos colocados em linhas, e cada coluna é um item do grupo.



Netflix: Parâmetros Analisados

- Principais Parâmetros de Recomendação:
 - Semelhança (ou Similaridade);
 - Amigos (social);
 - Popularidade;
 - Gênero (ou Categoria);
 - Outros parâmetros podem incluir localização geográfica do usuário ou dados retirados de seu perfil ou outros acessos.
- Algoritmos observam estes parâmetros em conjunto, não separados.



Fantasy Action Cartoon Drama Comedy

		1		1
			1	
1				1



Fantasy	Action	Cartoon	Drama	Comedy
10	0	7	4	17



Fantasy	Action	Cartoon	Drama	Comedy
0.26	0	0.18	0.11	0.45

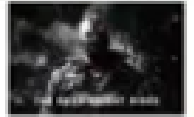
user feature vector



0.26	0	0.18	0.11	0.45
------	---	------	------	------



1	1	0	0	1
---	---	---	---	---



1	1	0	1	0
---	---	---	---	---



0	1	1	0	1
---	---	---	---	---



0	0	0	1	0
---	---	---	---	---



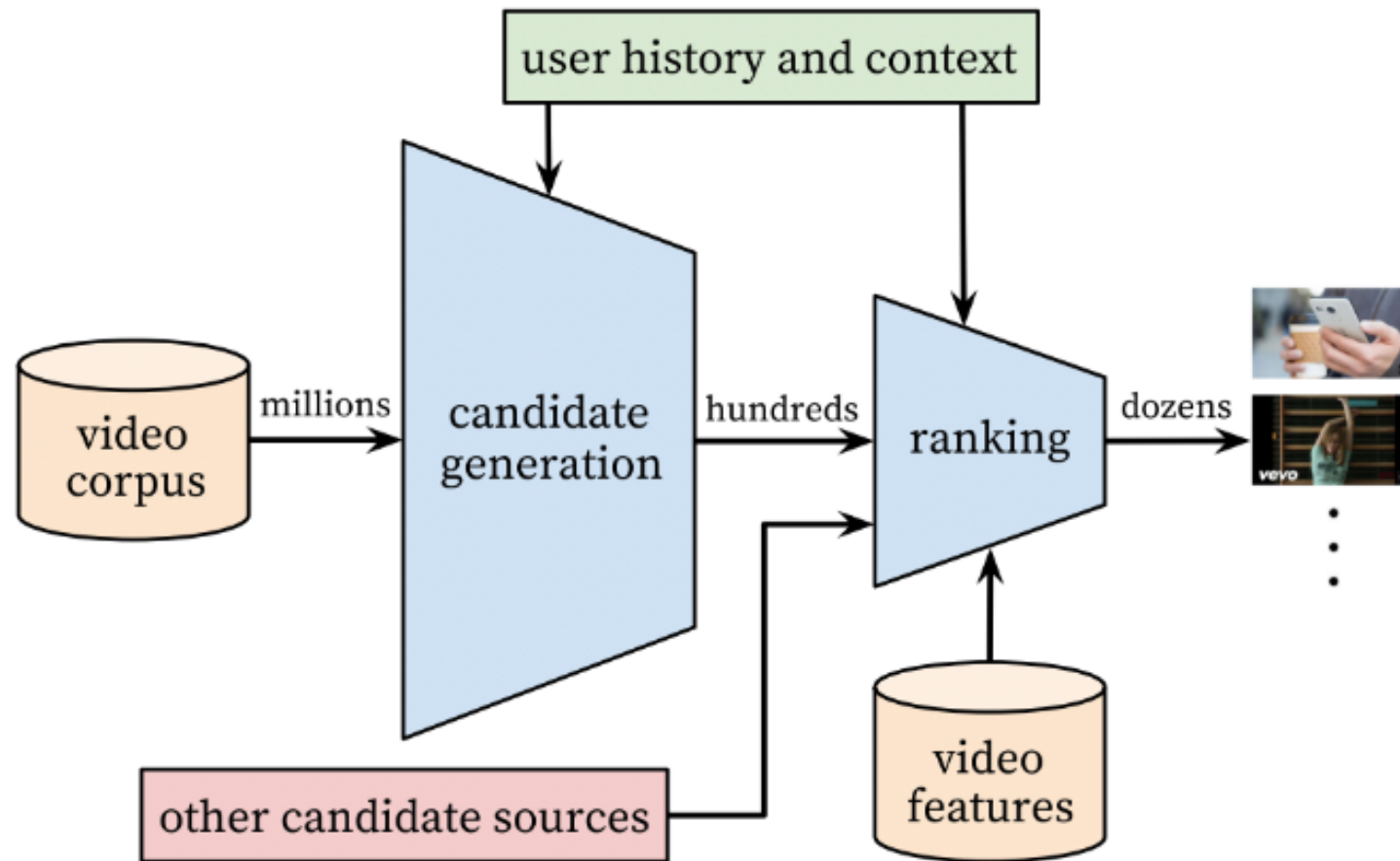
0.26	0	0	0	0.45
0.26	0	0	0.11	0
0	0	0.18	0	0.45
0	0	0	0.11	0

Σ

0.71
0.37
0.63
0.11

user movie ratings

• YouTube – 2 redes neurais

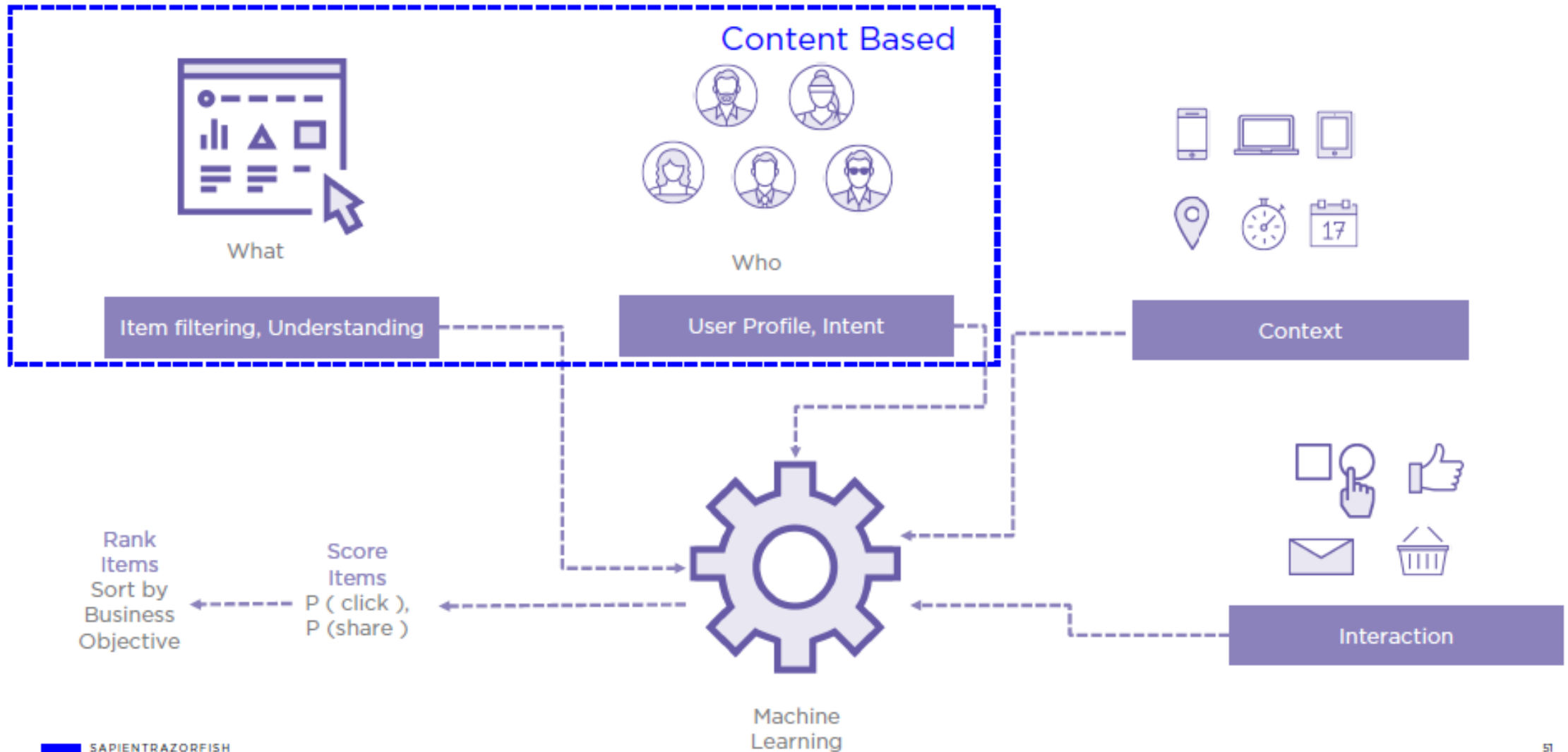




• YouTube – 2 redes neurais

- A rede de geração de candidatos leva o histórico de atividades do usuário (por exemplo, IDs de vídeos sendo assistidos, histórico de pesquisa e dados demográficos no nível do usuário) e gera algumas centenas de vídeos que podem ser amplamente aplicáveis ao usuário.
- Em contraste, a rede de classificação usa um conjunto mais rico de recursos para cada vídeo e pontua cada item da rede de geração de candidatos. Para essa rede, é importante ter um recall alto. Não há problema em algumas recomendações não serem muito relevantes, desde que você não perca os itens mais relevantes.

RecSys baseado em Conteúdo





RecSys baseado em Conteúdo

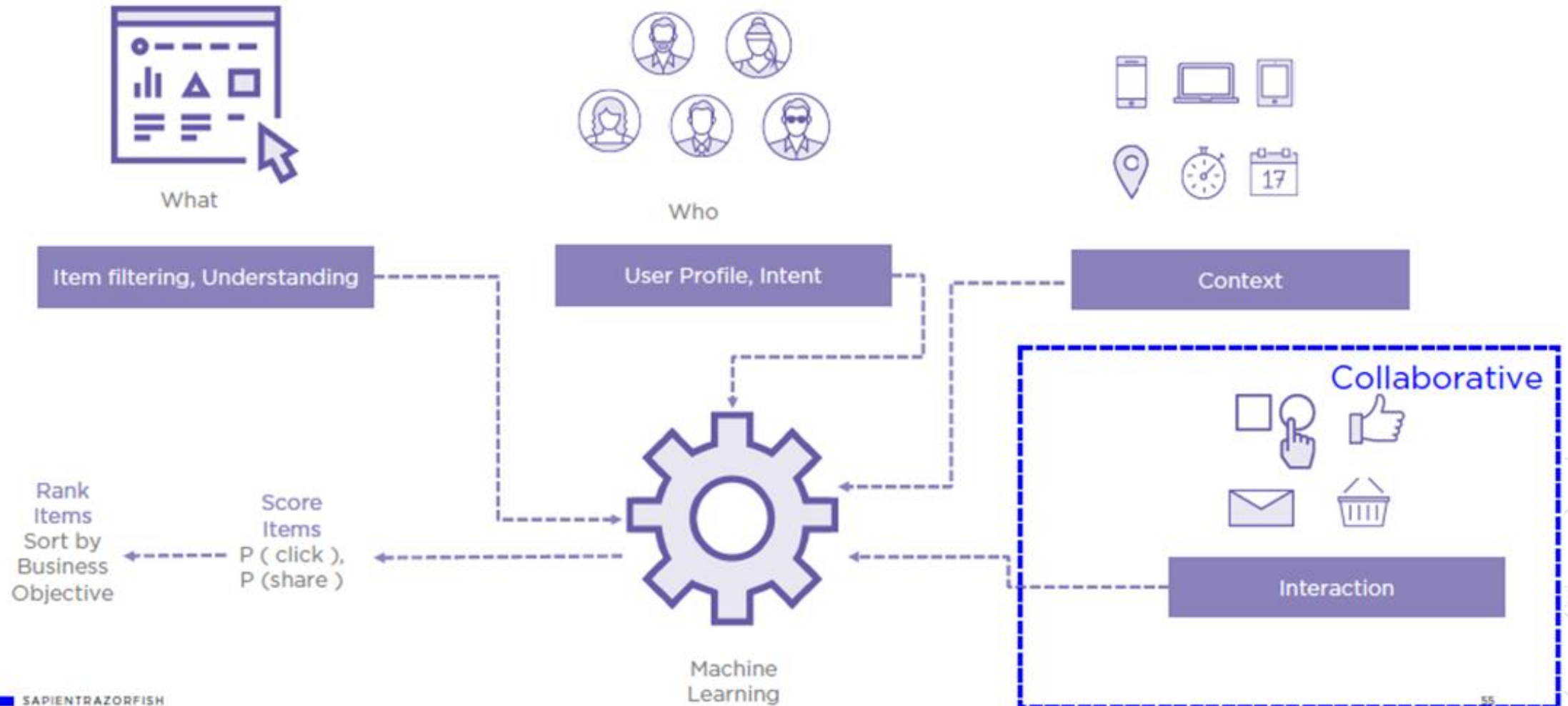
PRÓS

- Não há necessidade de dados de outros usuários
- Fácil de entender a razão por trás da recomendação
- Capaz de recomendar itens novos e desconhecidos

CONTRAS

- Só pode ser eficaz em circunstâncias limitadas
- Nenhuma sugestão adequada se o conteúdo não tiver informações suficientes
- Dependem inteiramente dos itens selecionados anteriormente e, portanto, não podem fazer previsões sobre futuros interesses dos usuários

RecSys Colaborativo



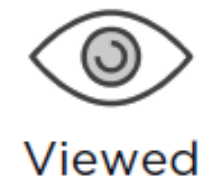
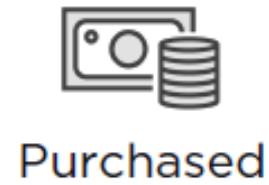
RecSys 101 : Collaborative Filtering : Interactions / Feedback



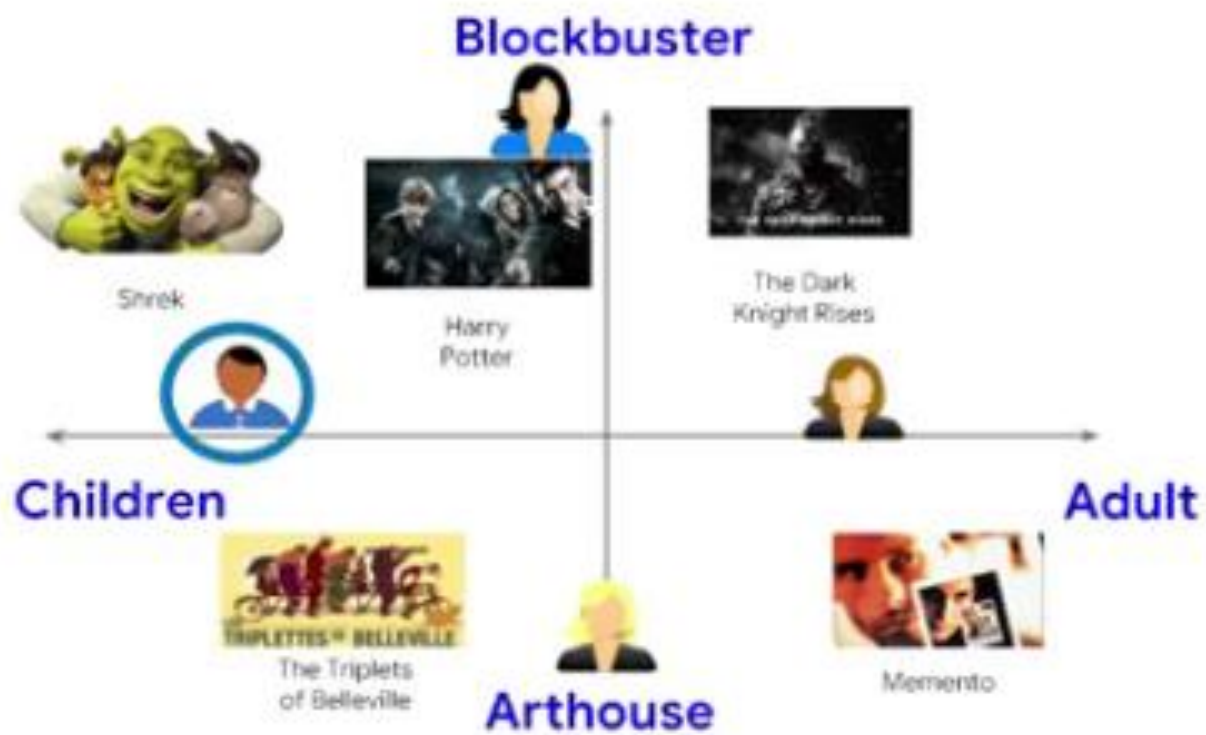
Explicit



Implicit



Item retrieval in two dimensions based on user





RecSys Colaborativo

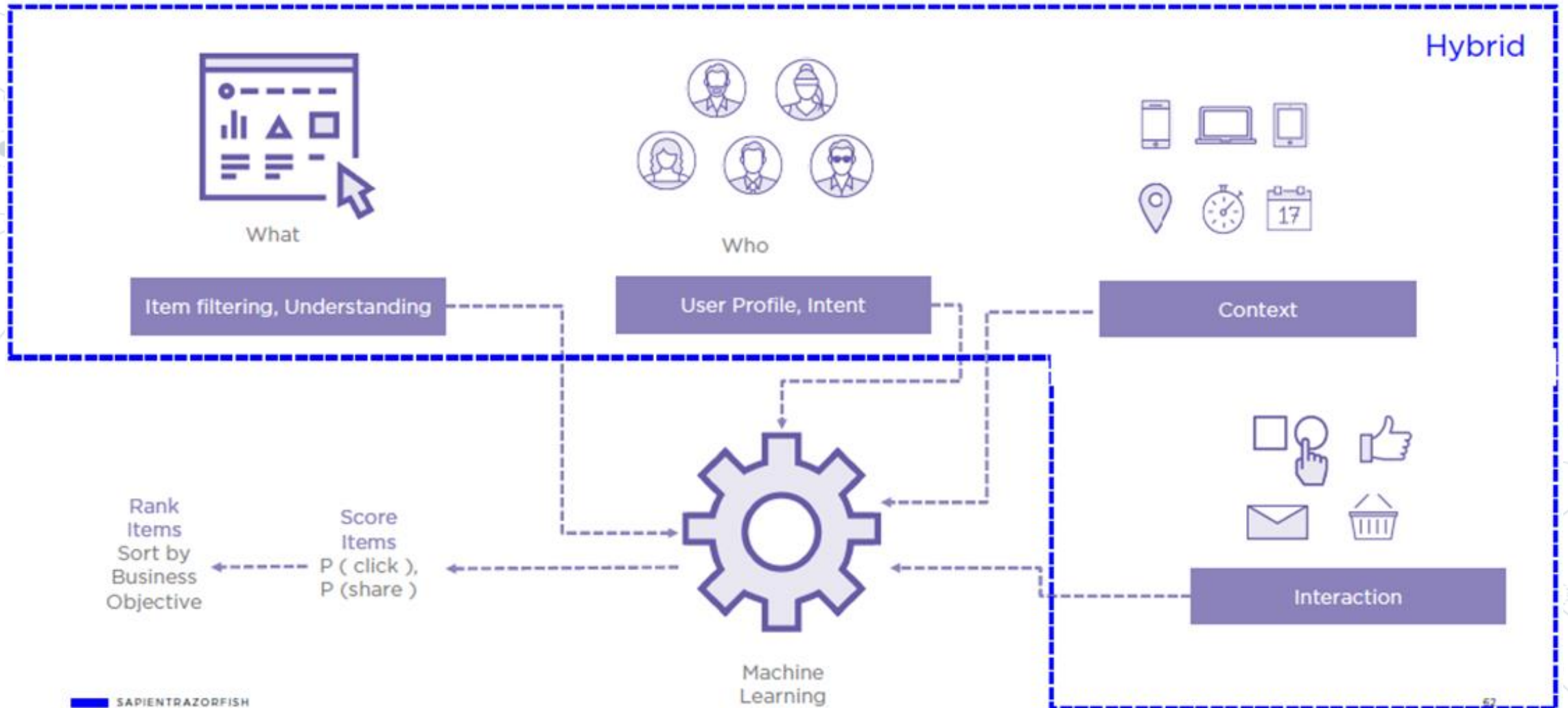
PRÓS

- Informações de conteúdo não são necessárias para usuários ou itens
- Recomendações personalizadas usando a experiência do usuário
- Nenhuma experiência de domínio necessária

CONTRAS

- Não pode produzir recomendações se não houver dados de interação disponíveis (problema de partida a frio)
- Geralmente demonstram baixa precisão quando há poucos dados sobre as classificações do usuário (escassez)
- Itens populares recebem mais feedback (viés de popularidade)

RecSys Híbrido





RecSys Híbrido

PRÓS

- resolver a questão do cold start unificando recomendações por conteúdo e colaboração
- o uso de feedback implícito reduz em grande parte os problemas de dispersão

CONTRAS

- Dificuldade para implementar

Por que Deep Learning para RecSys?

- **Grande quantidade de dados.** Técnicas de Filtragem Colaborativa são limitadas pelo tamanho da matriz de interação que também são extremamente esparsas. DL trabalha muito bem com grandes volumes de dados, consegue reduzir dimensionalidade sem perder a representatividade da informação original. Redes Convolucionais e Autoencoders são bastante utilizados nesse sentido.
- **Dados Heterogêneos e Extrator de Features.** A depender do domínio, os dados que representam o conteúdo podem variar bastante, sendo dados categóricos, numéricos, texto, imagem, áudio.. etc. Conseguir representar todas essas informações de forma unificada traz um grande impacto na similaridade do conteúdo.

Por que Deep Learning para RecSys?

- **Comportamento dinâmico.** A depender do domínio, o comportamento de preferência é muito dinâmico ou de curto prazo. Algoritmos clássicos tem dificuldades em extrair padrões quando esse comportamento muda muito rápido ou não está presente no histórico. Arquiteturas de DL como Redes Recorrentes trabalham muito bem com essa dinâmica e principalmente com o comportamento sequencial no consumo do conteúdo.
- **Melhor representação da relação de Usuário X Conteúdo.** A força de interação entre esses dois atores é a base da filtragem colaborativa, métodos clássicos acabam modelando essa interação de forma linear (ex Fatoração de Matrizes..), o que limita a generalização. Redes Neurais são conhecidas pela representação não-linear da informação e podem representar melhor essa interação.

Material adicional no Google

- ◎ <https://cloud.google.com/solutions/machine-learning/recommendation-system-tensorflow-overview>

Veja os tutoriais que acompanham esta visão geral:

- [Criar o modelo \(parte 1\)](#) mostra como usar o algoritmo WALS no TensorFlow para fazer previsões de classificação para o conhecido conjunto de dados [MovieLens](#).
- [Treinar e ajustar no Cloud Machine Learning Engine \(parte 2\)](#) mostra como usar o [Cloud Machine Learning Engine](#) para treinar o modelo e empregar o recurso de ajuste de hiperparâmetro para otimizá-lo.
- [Aplicar a dados do Google Analytics \(parte 3\)](#) mostra como aplicar o sistema de recomendação a dados importados diretamente do [Google Analytics 360](#) para realizar recomendações para sites que usam o Google Analytics.
- [Implantar o sistema de recomendação \(parte 4\)](#) mostra como implantar um sistema de produção no GCP para fazer recomendações em tempo real para um site.

Outros conteúdos



<https://www.slideshare.net/justinbasilico/deep-learning-for-recommender-systems-92331718>



Muito obrigado !

Eduardo Heitor

<https://linkedin.com/in/eduardoheitor/>

<https://www.facebook.com/eduardo.heitor>

