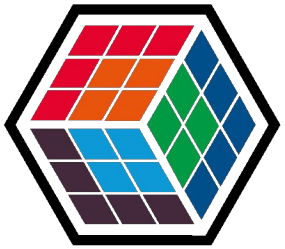


THE DEVELOPER'S CONFERENCE

Data Science - Impacto em Ciência de Dados

Fernando Tadao Ito
Senior Data Scientist

Everton Alvares Cherman
CTO

@ birdie.ai

Agenda

THE DEVELOPER'S CONFERENCE

- Introdução
- Definição de Impacto
- Metrificar Impacto
- Detratores:
  - Complexidade Desnecessária
  - Falta de Reuso
  - Arquitetura sem Objetivo Final
- Boas Práticas para Criação de Projetos de Dados
- Conclusão

Introdução

THE DEVELOPER'S CONFERENCE

Fernando Tadao Ito

Crawling Tech Lead / Senior Data Scientist

Mestrado em Inteligência Artificial pela UFSCar

Everton Alvares Cherman

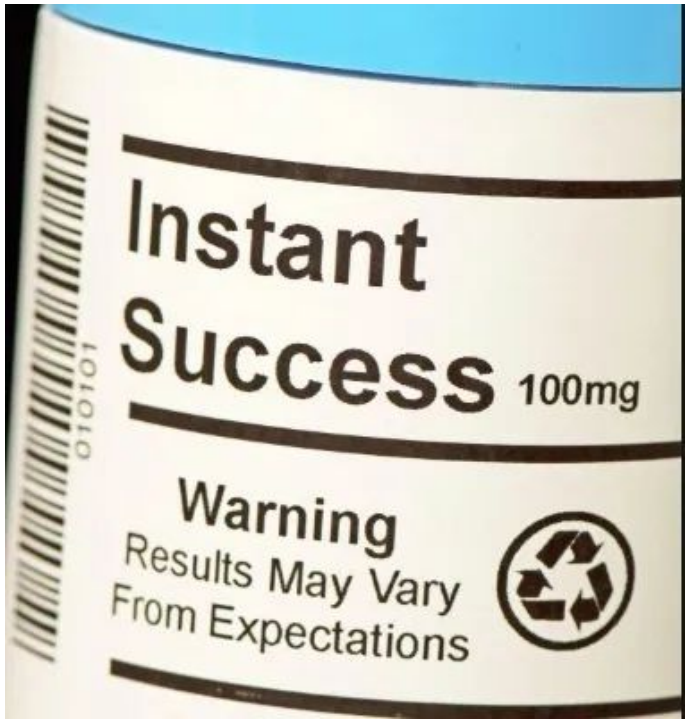
CTO / Data Tech Lead

Doutorado em Aprendizado de Máquina pela USP

Disclaimer

THE DEVELOPER'S CONFERENCE

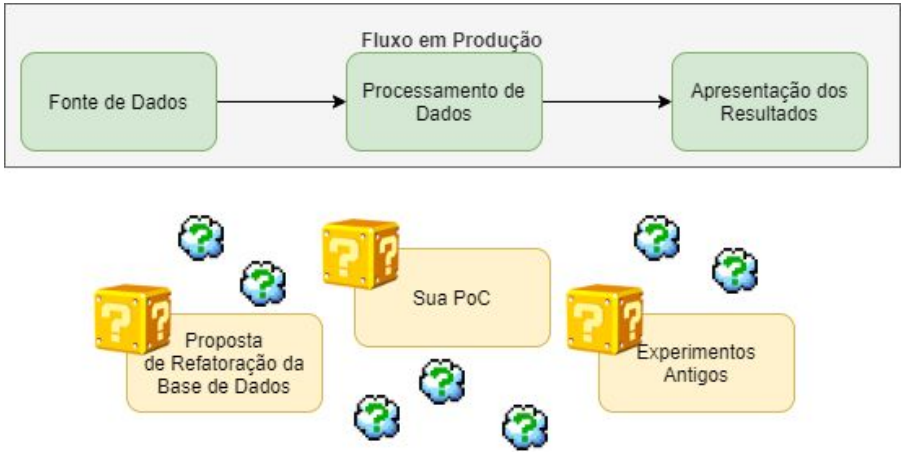
- *Impacto* tem diferentes significados para cada negócio!
- Não é um guia, mas um convite a uma discussão.



O que é Impacto?

THE DEVELOPER'S CONFERENCE

- *Impacto* é o quanto o seu trabalho afeta a sua empresa ou produto como um todo.
  - Modelos em Jupyter, PoCs isoladas: zero impacto!



Por quê Medir Impacto?

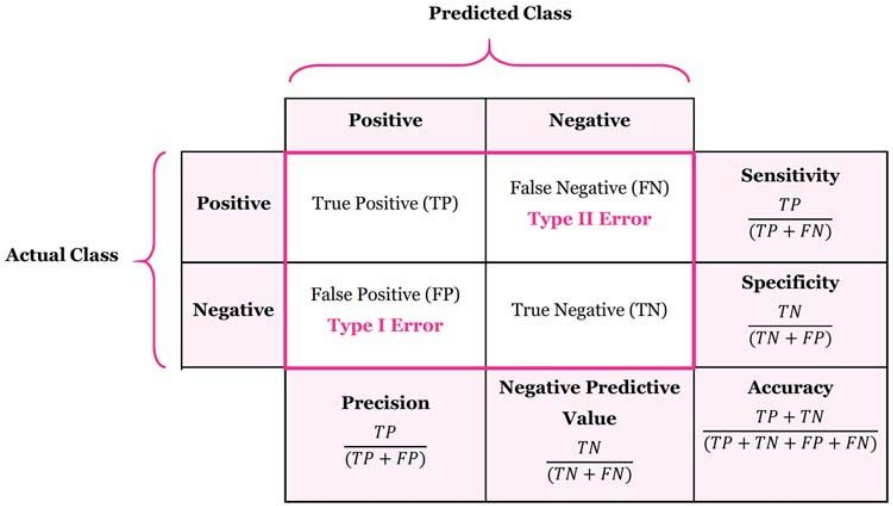
THE DEVELOPER'S CONFERENCE

- Projetos de Dados são custosos!
  - Infraestrutura cara
  - Mão-de-obra especializada
  - Risco inerente
- Projetos de Dados podem ser sutis.
  - Como o cliente vai perceber uma melhora de 2% na sua acurácia?

Como Medir Impacto?

THE DEVELOPER'S CONFERENCE

- Cientistas de Dados tem infinitas métricas para seus modelos.
  - F-score, P-value, coeficiente de Pearson...
- Mas isso é só uma parte infinitesimal do problema.



Fonte:https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

Como Medir Impacto?

THE DEVELOPER'S CONFERENCE

- **Métricas de Projeto**
  - *Quanto tempo e pessoas para desenvolver?*
  - *Quanto tempo para colocar em Produção?*
- **Métricas de Engenharia**
  - *Qual a performance média da inferência do modelo?*
  - *Qual o custo por dado inferido em infraestrutura?*
- **Métricas de Satisfação**
  - *Usuário está satisfeito com o produto?*

Detratores de Impacto

THE DEVELOPER'S CONFERENCE

- Ignorar métricas externas à modelagem danifica o seu impacto como Cientista de Dados.
  - Sem métricas de projeto, **nada é terminado**
  - Sem métricas de engenharia, **colocar em produção é muito difícil**
  - Sem métricas de satisfação, **um projeto não tem utilidade**

- Existem alguns sinais que mostram a falta de otimização multi-objetivo de seu projeto.

THE DEVELOPER'S CONFERENCE

Identificando Detratores de Impacto

Quem são eles?

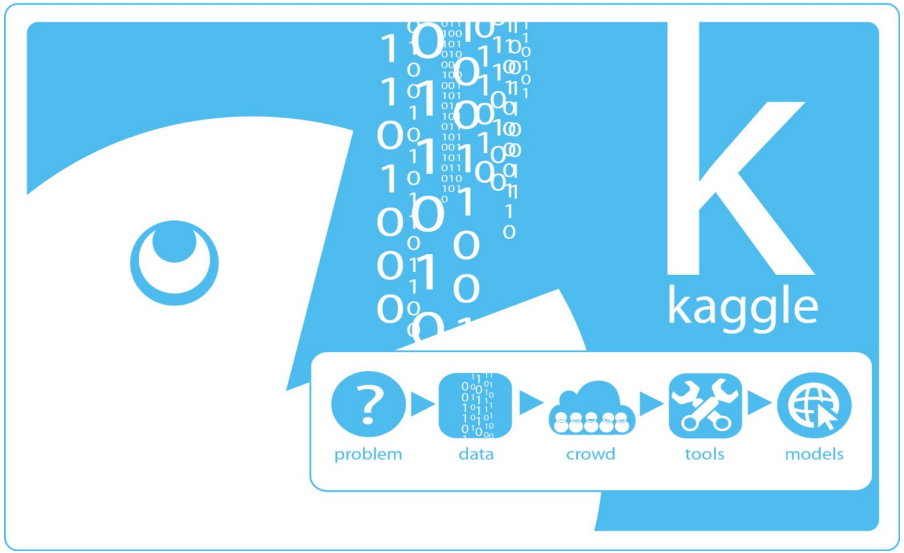
THE DEVELOPER'S CONFERENCE

- Complexidade Desnecessária
  - *Usar canhão pra matar formiga*
- Falta de Reuso
  - *Encostar modelos e recursos subjacentes*
- Arquitetura sem Objetivo Final
  - *Fazer PoC sem saber onde no Produto ela se encaixa*

Complexidade

THE DEVELOPER'S CONFERENCE

- *Over-engineering* de dados.
- Resolver o problema != ter a melhor acurácia.
  - Síndrome do *Kaggle*

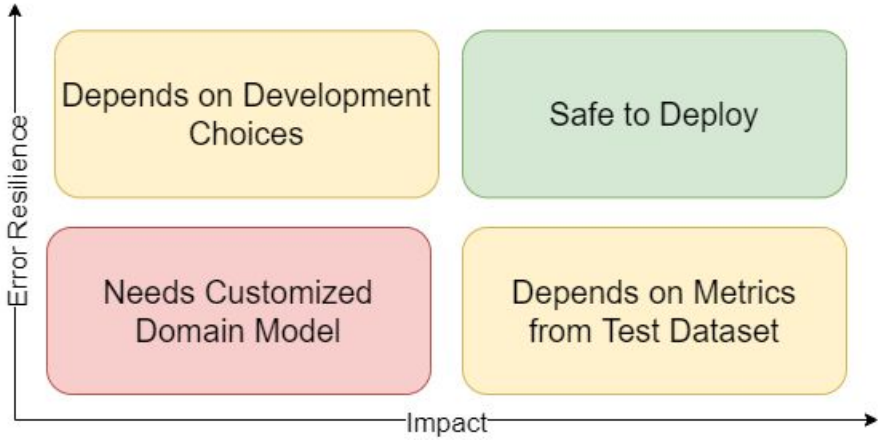


Fonte:
http://www.hub101.la/events/2018/3/13/learning-data-science-through-kaggle-competitions

Complexidade

THE DEVELOPER'S CONFERENCE

- Um problema de **alto impacto** e com alta **resiliência a erros** pode ser potencialmente resolvido com um modelo de baixa acurácia.

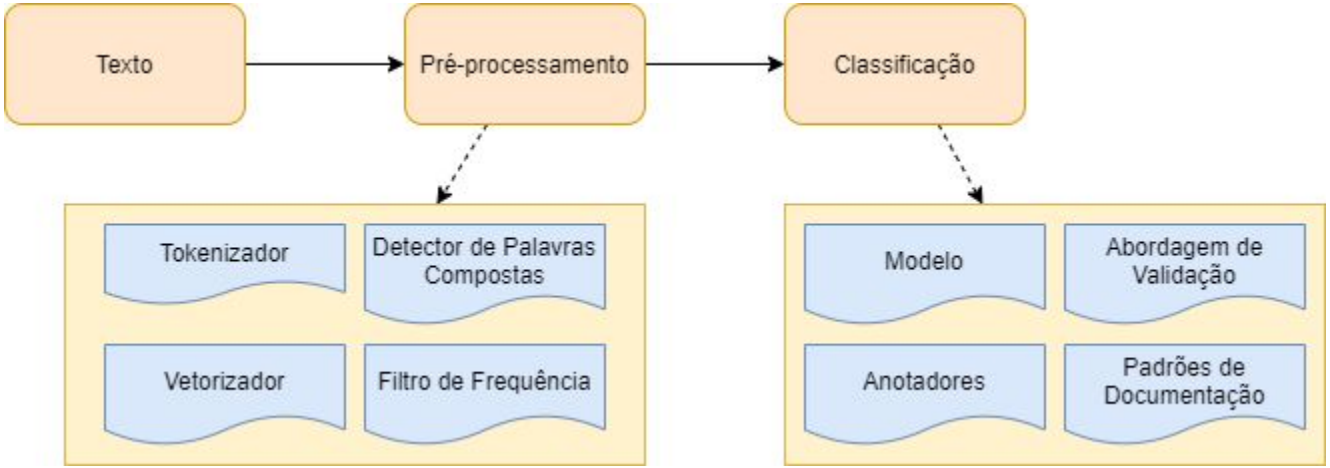


Fonte:
https://medium.com/birdie-ai/should-i-deploy-pre-trained-models-cbe5e884dadc

Reusabilidade

THE DEVELOPER'S CONFERENCE

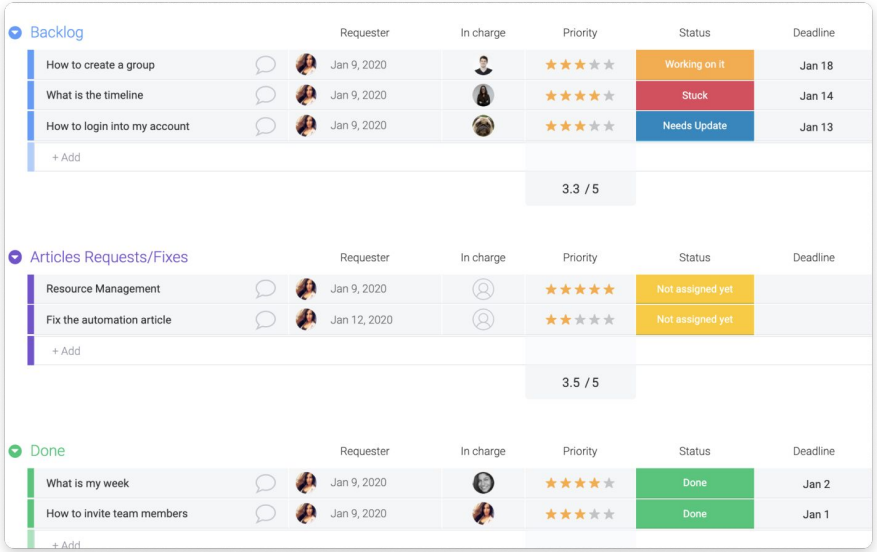
- Projetos de Dados geram diversas ferramentas auxiliares para resolver um problema.
  - Pré-processamento, representação de palavras/documentos, tokenizadores, etc.



Reusabilidade

THE DEVELOPER'S CONFERENCE

- Saber catalogar e separar cada artefato é essencial para a sua reutilização em projetos futuros.
- Ter uma **base de conhecimento** atualizada ajuda nessa parte!

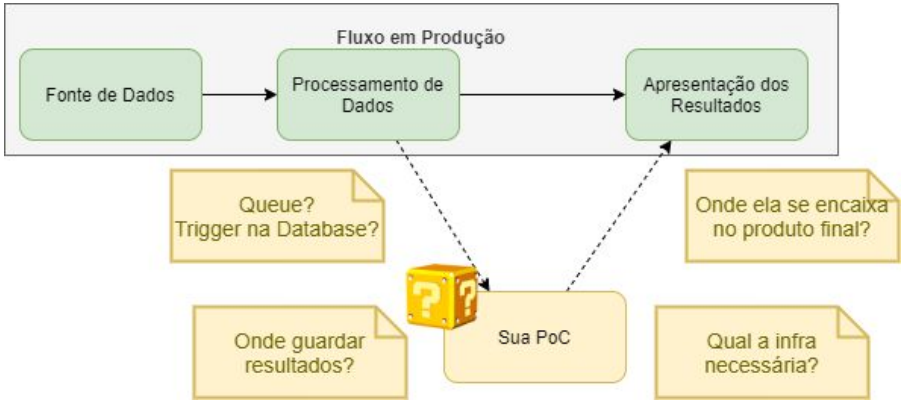


Fonte: https://support.monday.com/hc/en-us/articles/360011232799-monday-com-for-Knowledge-Management-

Integração com Produto

THE DEVELOPER'S CONFERENCE

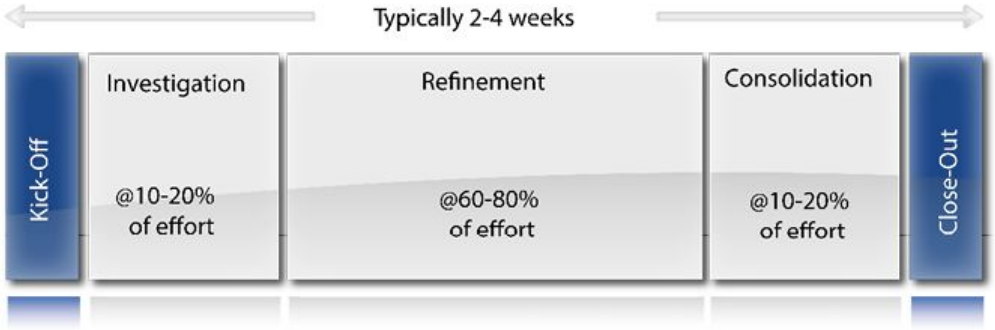
- O valor real de uma PoC não está em sua inovação, e sim em como ela pode ser colocada em Produção!



Integração com Produto

THE DEVELOPER'S CONFERENCE

- PoCs de Dados são exploratórias…
  - Mas é preciso ter um objetivo.
- Um projeto não precisa *nascer* com um objetivo claro, mas precisa *desenvolver* um.

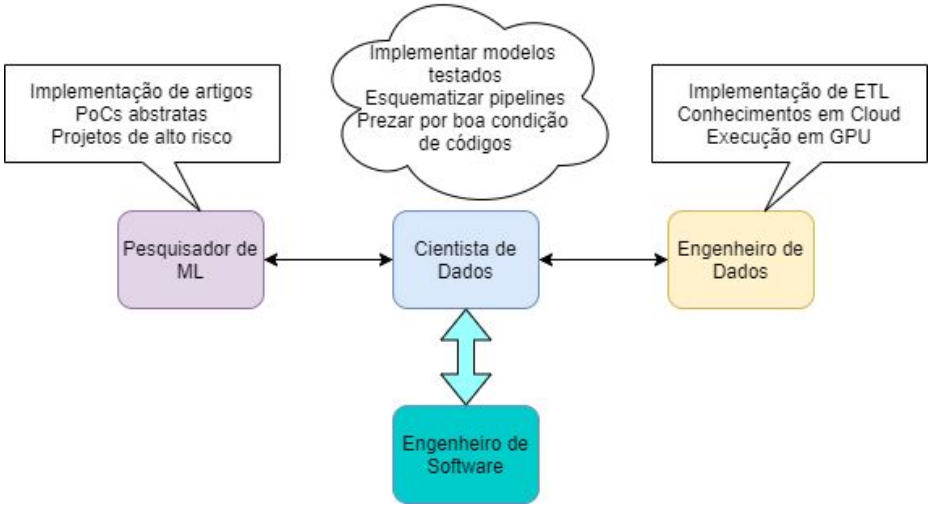


Fonte: https://www.be-positive.co.uk/boxing-clever-timeboxing/

Conclusão

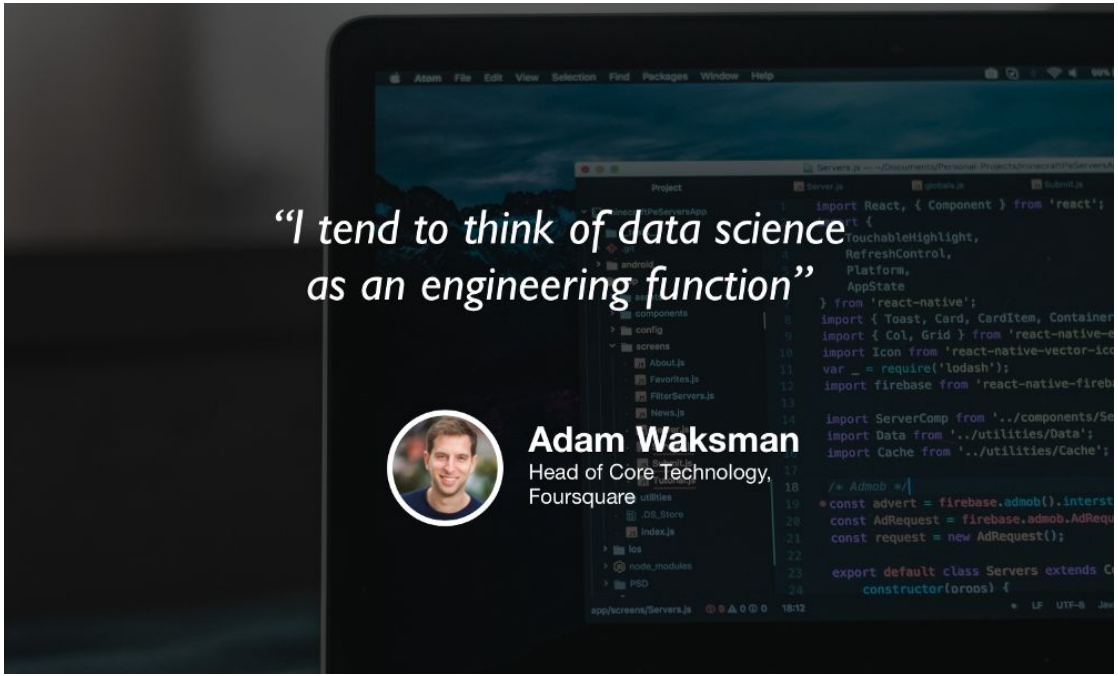
THE DEVELOPER'S CONFERENCE

- Todos esses detratores são sinais de falta de cuidado com outras métricas!
- Cientista de Dados != Engenheiro de Dados.



Conclusão

THE DEVELOPER'S CONFERENCE



https://towardsdatascience.com/data-science-is-becoming-software-engineering-53e31314939a

THE DEVELOPER'S CONFERENCE

Boas Práticas para Projetos de Dados

Estruturando do zero

THE DEVELOPER'S CONFERENCE

O que precisa um projeto de dados?

DEFINIÇÃO CLARA DO PROBLEMA

Estruturando do zero

THE DEVELOPER'S CONFERENCE

Criação de um documento

para descrever o problema e a potencial solução

Por que essa "*burocracia*"?!?!

- Força a pensar com mais detalhe e precisão
- Receber inputs de outros profissionais para melhoria

Request for Comments

THE DEVELOPER'S CONFERENCE

- RFC
  - Documento para solicitar revisão a outros pares
  - Google Docs / Página no Notion

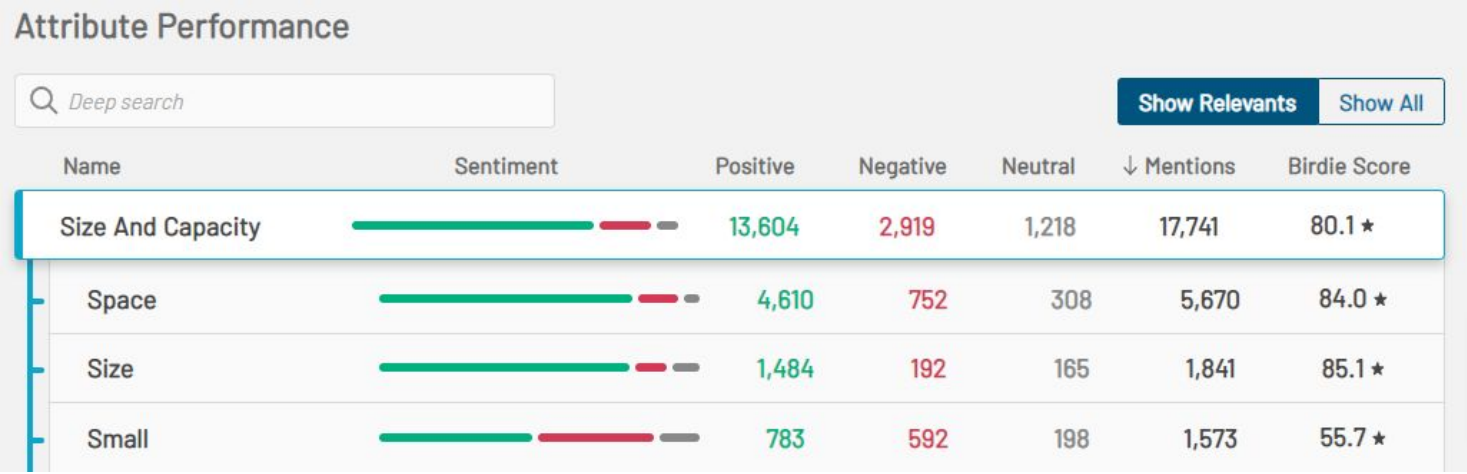
- Estrutura sugerida
  - Resumo
  - Contexto
  - Problema
  - Definição de sucesso
  - Solução

https://pt.wikipedia.org/wiki/Request_for_Comments

Estruturando do zero

THE DEVELOPER'S CONFERENCE

- *Case*: Birdie
  - Análise de Sentimento de Palavras em Sentenças



Contexto

THE DEVELOPER'S CONFERENCE

- Onde o projeto está inserido?
- Por que ele é necessário?
- Qual problema de negócio e/ou técnico ele irá resolver?

Considerando reviews de e-commerces, o cliente (nosso usuário) quer identificar quais aspectos de seus produtos está se destacando em relação à concorrência e quais não estão bem, em outras palavras, quais são os pontos fracos e pontos fortes de seus produtos.

Considere a categoria de Refrigeradores como exemplo. O usuário quer saber a opinião de seus consumidores sobre os principais aspectos desse tipo de produto, como o tamanho, o espaço interno, o barulho e o quanto bem ela mantém os alimentos congelados ou refrigerados, a eficiência energética, entre outros possíveis aspectos mencionados pelo consumidor em textos de reviews em e-commerces.

Problema

THE DEVELOPER'S CONFERENCE

- Redução do problema de negócio para um problema computacional
- Qual é o input e o output esperados?
- O escopo pode ser decomposto em projetos menores que já entregam valor?

Este projeto tem como objetivo criar um analisador de sentimento baseado em aspectos sobre dados textuais de opinião de consumidores coletados de e-commerces descritos em Inglês.

Mais especificamente, o problema será reduzido em um algoritmo que recebe um trecho de uma opinião (snippet) e uma ou mais palavras subsequentes que representam um aspecto e deverá retornar qual é o subjetividade / julgamento / sentimento do consumidor quanto ao aspecto no contexto do snippet. O sentimento pode ter o valores positivo ou negativo, bem como o valor neutro se não houve um julgamento sobre o aspecto em questão.

Input

```
{
  "snippet" : "The internal space of this refrigerators is great, but the ice maker stop wo
  "aspect" : "ice maker"
}
```

Output

```
{
  "sentiment" : "Negative"  // possible values are "negative", "neutral" and "positive"
}
```

Definição de sucesso

THE DEVELOPER'S CONFERENCE

- Como saberemos que o problema foi bem resolvido?
- Quais são os entregáveis?
- Qual é a configuração experimental para aferir a qualidade?
- Acurácia e Precisão vs Cobertura?

Definição de sucesso

THE DEVELOPER'S CONFERENCE

Este projeto deve entregar um algoritmo de análise de sentimento e sua integração ao pipeline de enriquecimento de forma que a classificação do sentimento possa ser realizada para todos os reviews coletados e utilizados posteriormente para análises.

O sucesso do algoritmo de classificação de sentimento deverá ser medido por sua acurácia sobre um conjunto de dados rotulado de referência (conjunto de teste). Espera-se inicialmente um algoritmo com **acurácia acima de 70%** para que se possamos realizar análises utilizando essas classificações.

No entanto, este projeto deve definir um **processo** para que permita realizar com facilidade **futuras melhorias** do algoritmo.

Esse projeto assume que já temos (e consequentemente está **fora de seu escopo**):

- Reviews de e-commerces coletados e armazenados em nosso banco de dados
- Aspectos importantes de cada categoria de produto já extraídos
- Metadados dos reviews identificando a marca e a categoria do produto para possibilitar posterior análise comparativa

Solução

THE DEVELOPER'S CONFERENCE

- Quais são os componentes estruturais da solução?

- Existe uma solução baseline / simples a considerar?

Baseline

THE DEVELOPER'S CONFERENCE

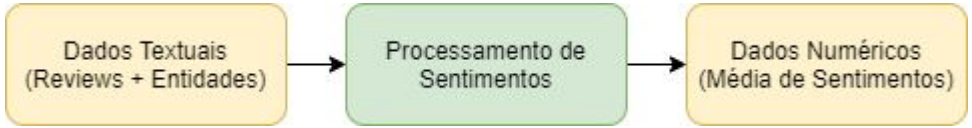
- Utilizar as estrelinhas para determinar sentimento?
  - 4 e 5 - Positivo
  - 3 - Neutro
  - 1 e 2 - Negativo

- Utilizar um dicionário de palavras que carregam sentimento?

Solução baseline

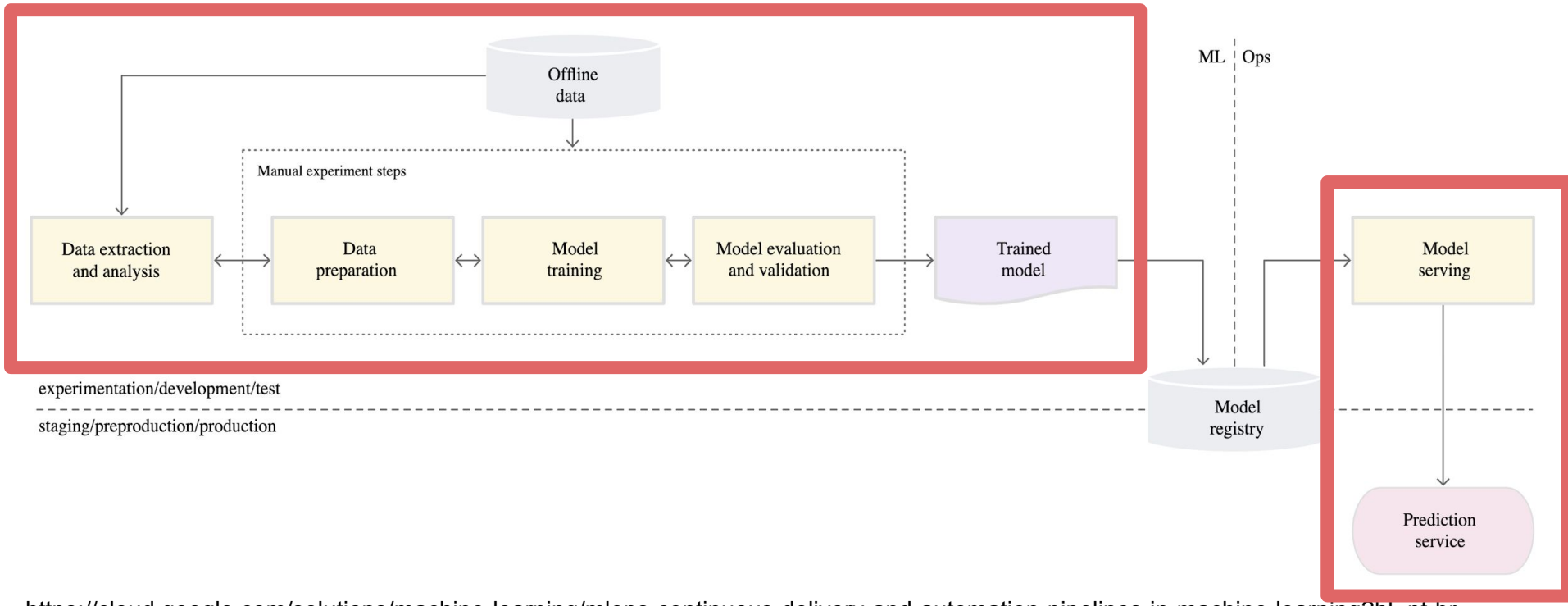
THE DEVELOPER'S CONFERENCE

- Objetivo da PoC inicial era entregar valor **ASAP.**
  - Primeira entrega com uma *pipeline* usando um modelo pré-treinado para provar o valor da solução.
  - Melhorias foram imediatamente colocadas em desenvolvimento.
  - Foco foi dado em criar uma **arquitetura** que permitisse rápida troca de modelos.



Componentes - MLOps

THE DEVELOPER'S CONFERENCE



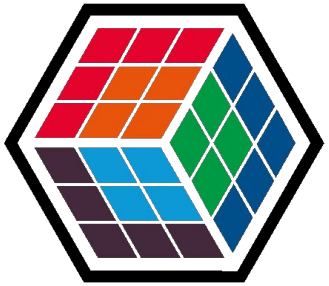
https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning?hl=pt-br

THE DEVELOPER'S CONFERENCE

Obrigado pela atenção!

Perguntas, questões, réplicas, reclamações?

https://www.birdie.ai/careers



THE DEVELOPER'S CONFERENCE