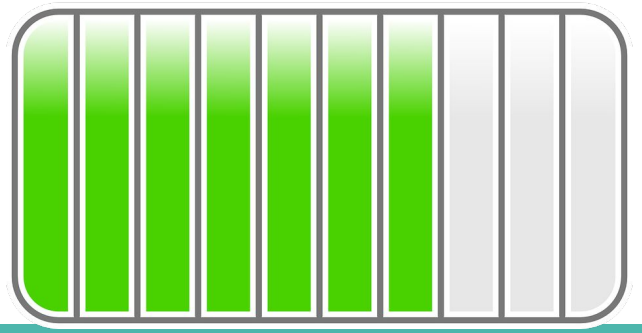
Metodologias para DS

Como funciona um projeto de DS?

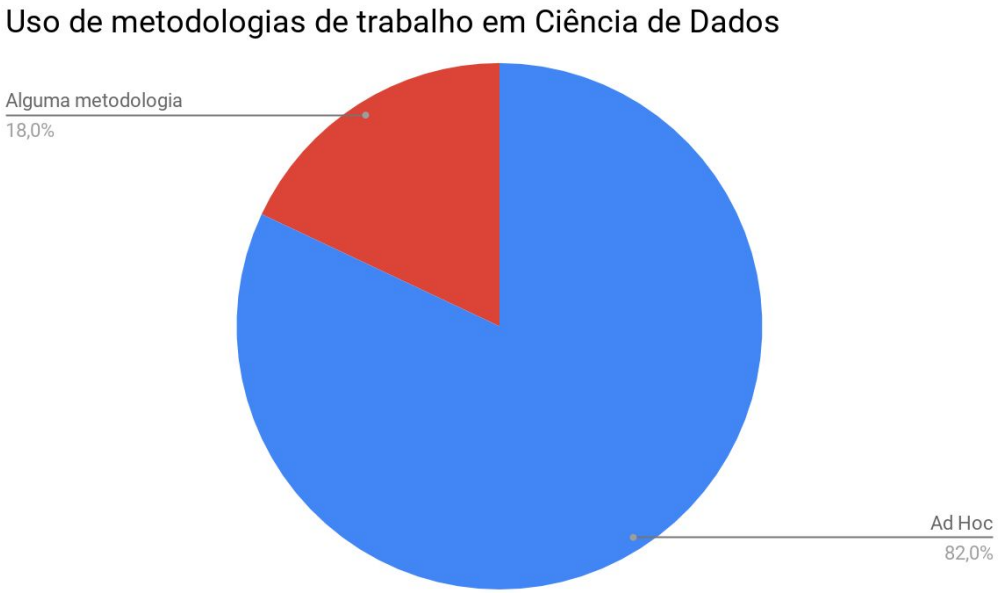
Ferramentas

O que precisa ser feito?

- Fazer a pergunta correta
- Construir ou utilizar os métodos corretos e interpretar os resultados
- Comunicar de forma que as pessoas se importem com o que foi encontrado



Mas como?



https://neptune.ai/blog/data-science-project-management

Metodologias

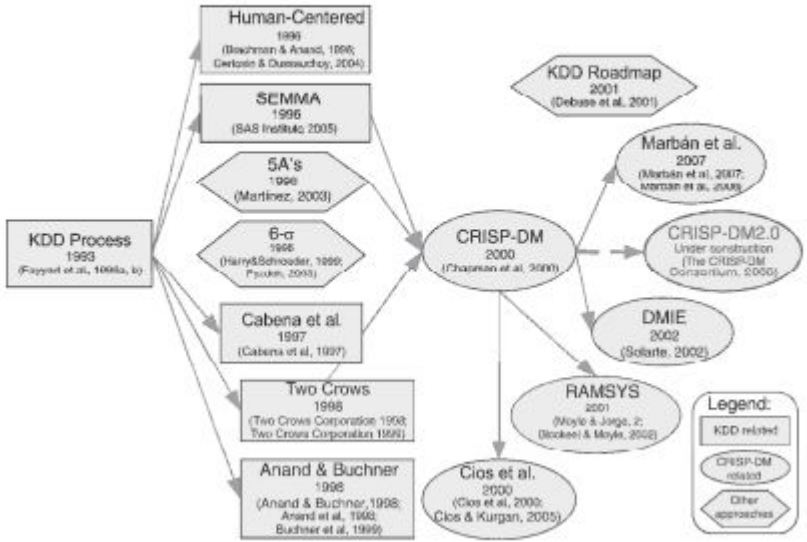
- Uma metodologia é uma estratégia geral que orienta os processos e atividades dentro de um determinado domínio.
- A metodologia não depende de tecnologias ou ferramentas específicas, nem é um conjunto de técnicas ou receitas.
- Em vez disso, uma metodologia fornece aos cientistas de dados uma estrutura para proceder de acordo com os métodos, processos e heurísticas que serão usados para obter respostas ou resultados

Problemas na falta de metodologias

- Dependência da experiência de um analista sênior
- Ineficiência
- Problemas no compartilhamento de informação
- Entregar algo diferente do que foi solicitado
- Projeto irreprodutíveis
- Escopo deficiente

Algumas metodologias

Figure 1. Evolution of Data Mining Methodologies
Source: Mariscal et al. [47]



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7924527/

KDD

https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

KDD

"A Descoberta de Conhecimento em Bases de Dados (KDD) é o processo interativo para identificar nos dados novos padrões que sejam válidos, novos, potencialmente úteis e interpretáveis".(FAYYAD et al.,1996)

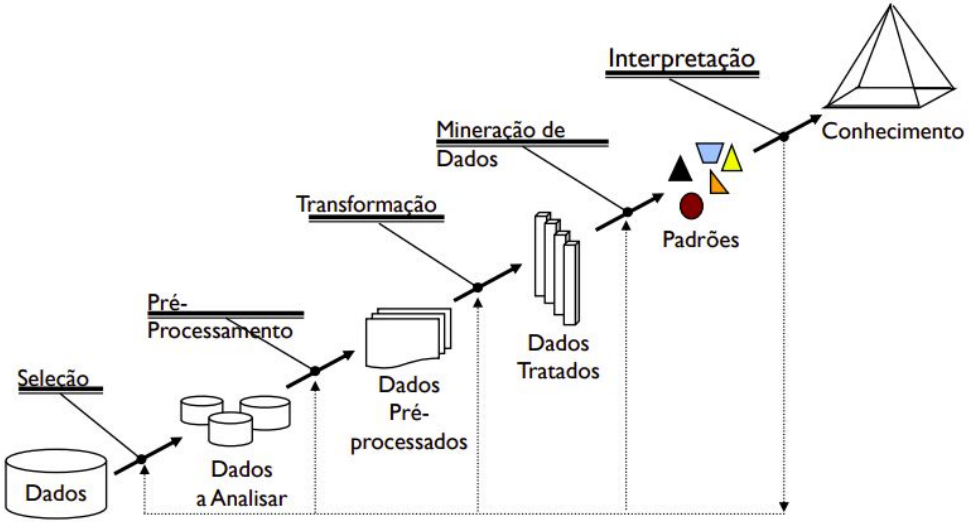
- KDD: Conjunto de passos e ou processos que visam a descoberta de conhecimento útil a partir dos dados
- *Data Mining* é uma etapa desse processo de descobrimento envolvendo a fase de modelagem dos dados.

https://www.linkedin.com/pulse/kdd-knowledge-discovery-databases-quando-o-processo-beatriz/?originalSubdomain=pt

KDD

- **Interpretáveis** - permitindo um entendimento pelo Negócio. Isto implica em uma representação simples que os torna inteligíveis.
- **Válidos** - o novo conhecimento gerado deve ser adequado ao Negócio.
- **Novos** - devem acrescentar novos conhecimentos aos já existentes no Negócio.
- **Úteis** – na sua aplicação devem proporcionar benefícios ao Negócio.

KDD


Interpretação
Conhecimento
Mineração de Dados
Padrões
Transformação
Dados Tratados
Pré-Processamento
Dados Pré-processados
Seleção
Dados a Analisar
Dados

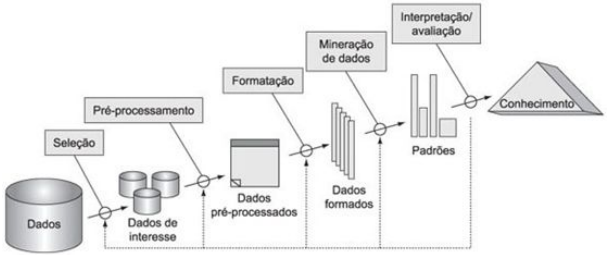
KDD

- **Seleção**: quais dados serão utilizados? quais variáveis serão utilizadas?
- **Pré-Processamento:** existem ruídos? Existem dados omissos?
- **Transformação:** quais atributos são úteis? posso reduzir o número de atributos?
- **Mineração de dados:** adaptação dos dados para a tarefa de ML
- **Interpretação:** o que os dados estão nos dizendo?

KDD

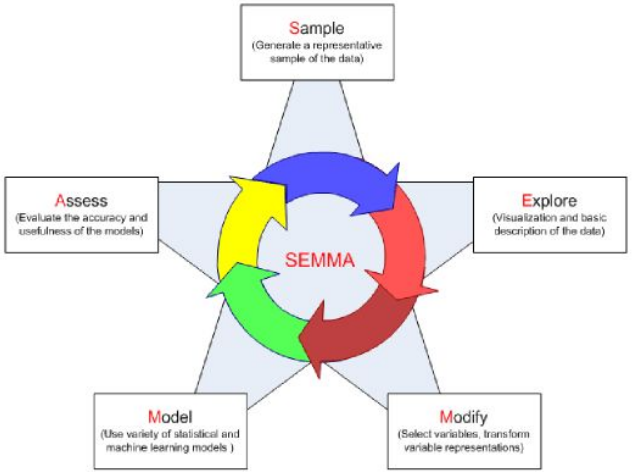
Prós: interativo e iterativo, envolvendo várias etapas com muitas decisões sendo tomadas pelo usuário;

Contras: deve ser precedido pelo desenvolvimento de uma compreensão do domínio da aplicação, o conhecimento prévio relevante e os objetivos do usuário final.



SEMMA

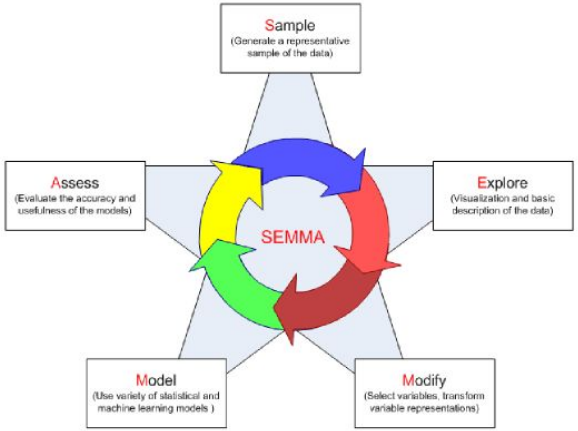
O processo SEMMA foi desenvolvido pelo SAS Institute. O acrónimo SEMMA significa Sample,Explore, Modifique, Modelo, Avalie e faça referência ao processo de realização de um projeto de mineração de dados. O SAS Instituto considera um ciclo com 5 etapas para o processo.



SEMMA

Prós: interativo e iterativo, parecido com o CRISP-DM

Contras: foca mais no modelo do que no negócio



CRISP-DM

Crisp-DM in a Nutshell

O Caos!

80% dos projetos de DS não são colocados em produção! [Gartner]

Causas mais comuns:

Não atende a uma necessidade do cliente

Problemas nos dados descobertos tardiamente

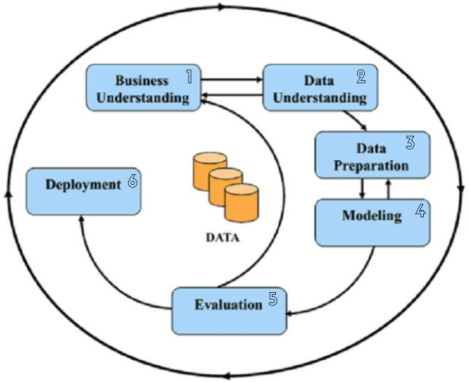
Modelos muito complexos

A solução!

Utilizar uma metodologia para o desenvolvimento do projeto. As metodologias mais conhecidas são o KDD e o Crisp-DM. O primeiro não foca diretamente no negócio nem na criação de um modelo. Além disso não fomenta a iteração entre as etapas.
Por isso vamos de CRISP-DM!

Cross-Industry Standard Process for Data Mining

Guia em 6 passos para planejar, organizar, executar e analisar um projeto de data science



Manter a organização e ter uma linha de raciocínio clara, evitando surpresas desagradáveis ao final do projeto

Iteração entre as etapas!

1 - entender profundamente o negócio, as áreas envolvidas, o objetivo do uso de DS nesse negócio e as pessoas envolvidas

2 - intimamente ligado ao entendimento do negócio e a proposta de aplicação.
- Quais dados são necessários?
- Coletar, explorar e verificar qualidade

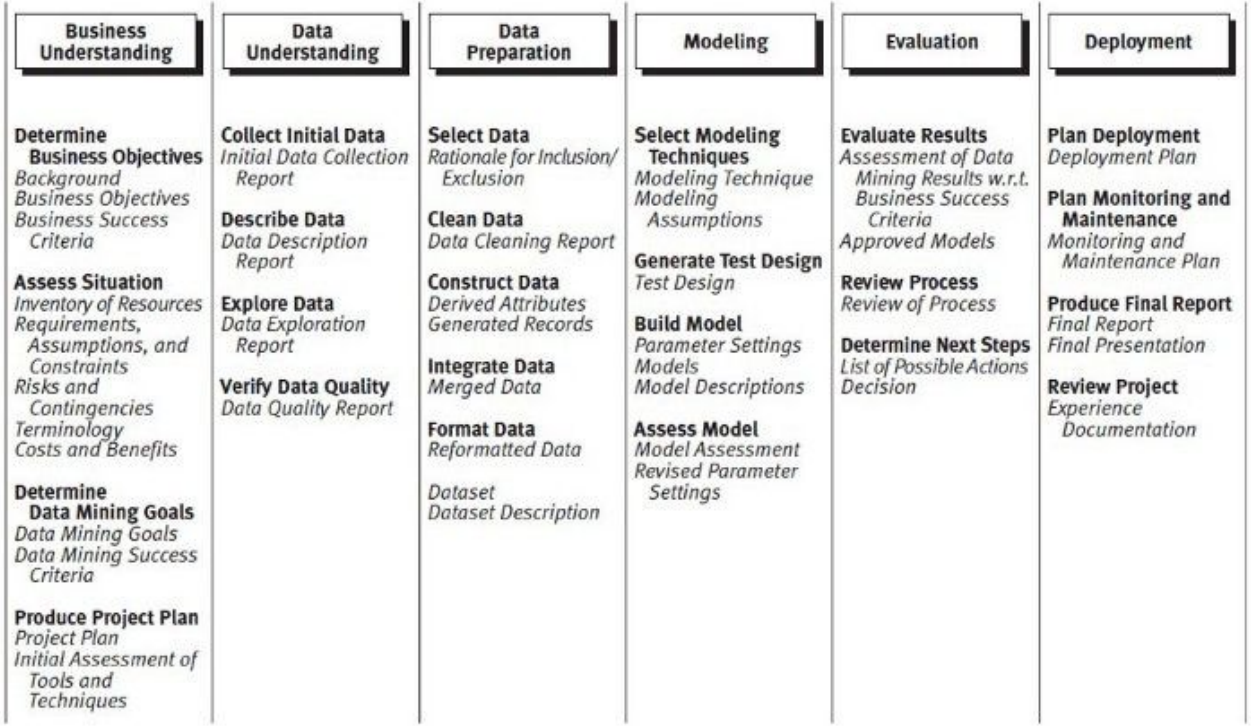
3 - preparar o dado para a modelagem:
- Limpar os dados
- Gerar novos atributos
- Integrar dados

4 - Encontrar respostas para o nosso problema!
- Aplicar a técnica que pode responder o problema do negócio

5 - Avaliar do ponto de vista de DS (quais métricas fazem sentido aqui?) e principalmente avaliar do ponto de vista do negócio

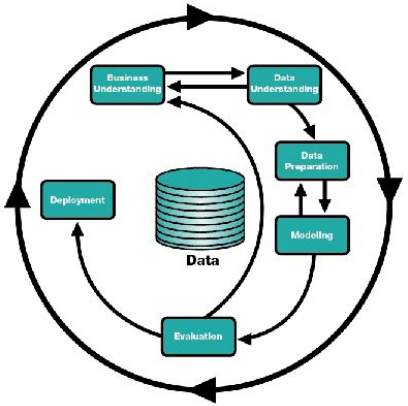
6 - Se a avaliação estiver satisfatória o modelo está pronto para ser utilizado! Caso contrário o ciclo se repete.

CRISP-DM

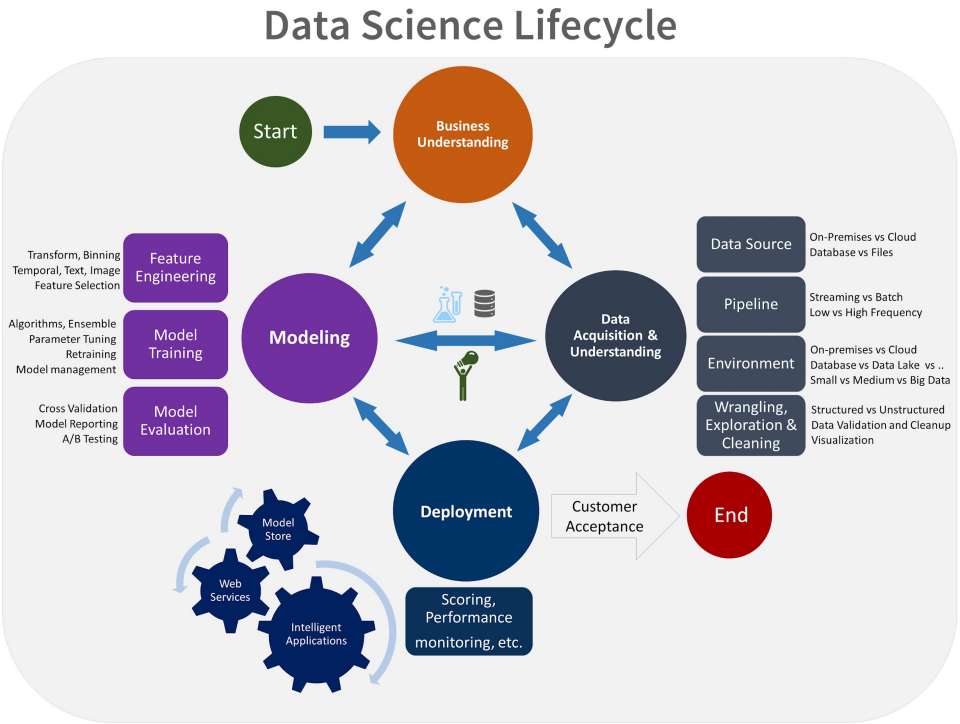

Business Understanding
Determine Business Objectives
Background
Business Objectives
Business Success Criteria
Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
Data Understanding
Collect Initial Data
Initial Data Collection Report
Describe Data
Data Description Report
Explore Data
Data Exploration Report
Verify Data Quality
Data Quality Report
Data Preparation
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data
Dataset
Dataset Description
Modeling
Select Modeling Techniques
Modeling Technique
Modeling Assumptions
Generate Test Design
Test Design
Build Model
Parameter Settings
Models
Model Descriptions
Assess Model
Model Assessment
Revised Parameter Settings
Evaluation
Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models
Review Process
Review of Process
Determine Next Steps
List of Possible Actions
Decision
Deployment
Plan Deployment
Deployment Plan
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Documentation

CRISP-DM

- **Prós**: flexível e cíclico; focado na tarefa e fácil de implementar
- **Contras**: não foi pensado para o trabalho em equipe; não cobre as etapas de comunicação



TDSP



Scrum + CRISP-DM

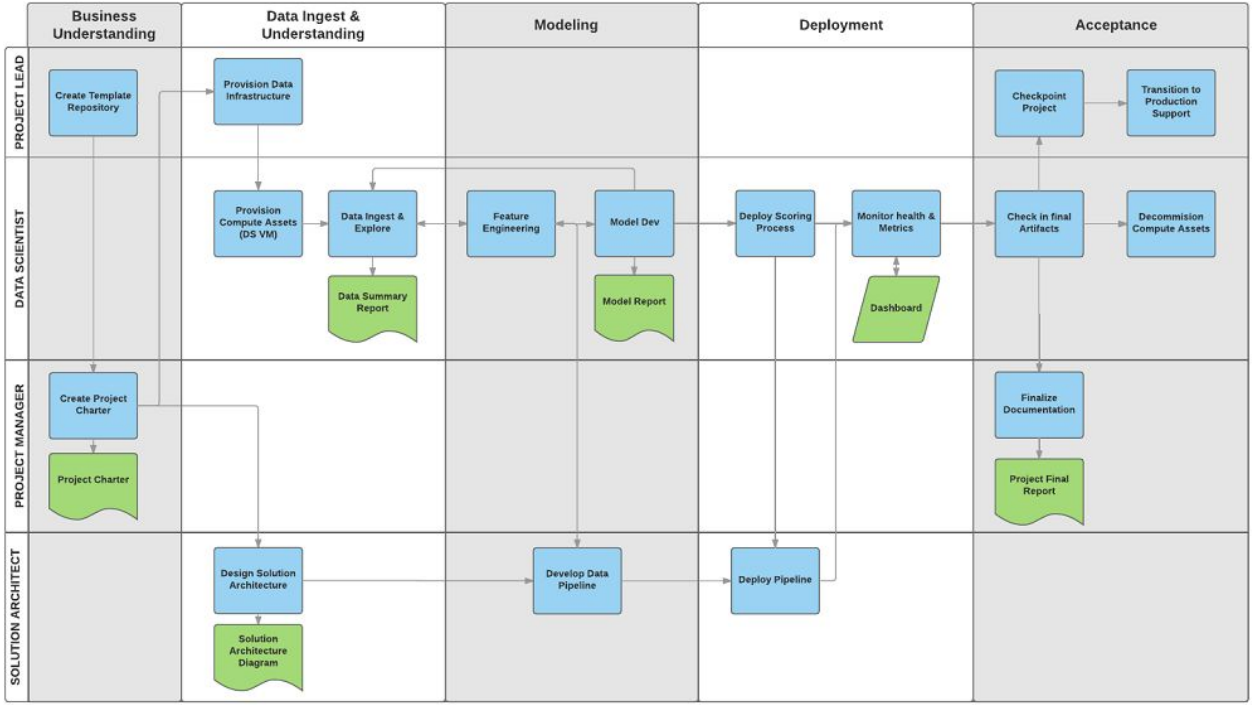
TDSP

- **Compreensão do negócio:** definir objetivos e identificar fontes de dados
- **Aquisição e compreensão de dados:** ingira dados e determine se eles podem responder à pergunta (combina efetivamente o entendimento de dados e a limpeza de dados do CRISP-DM)
- **Modelagem**: engenharia de recursos e treinamento de modelos (combina Modelagem e Avaliação)
- **Implantação**: implantar em um ambiente de produção
- **Aceitação do cliente:** validação do cliente se o sistema atende às necessidades do negócio (fase não coberta explicitamente pelo CRISP-DM)

TDSP

- Gerente de grupo: supervisiona toda a unidade de ciência de dados
- Líder de equipe: gerencia a equipe de ciência de dados
- Líder do projeto: gerencia atividades diárias no projeto especificado
- Colaborador individual do projeto: “membro da equipe de desenvolvimento” do Scrum; inclui cientista de dados, analista de negócios, engenheiro de dados, etc.

TDSP


Business Understanding
Data Ingest & Understanding
Modeling
Deployment
Acceptance
PROJECT LEAD
DATA SCIENTIST
PROJECT MANAGER
SOLUTION ARCHITECT
Create Template Repository
Provision Data Infrastructure
Checkpoint Project
Transition to Production Support
Provision Compute Assets (DS VM)
Data Ingest & Explore
Feature Engineering
Model Dev
Deploy Scoring Process
Monitor health & Metrics
Check in final Artifacts
Decommision Compute Assets
Data Summary Report
Model Report
Dashboard
Create Project Charter
Finalize Documentation
Project Charter
Project Final Report
Design Solution Architecture
Develop Data Pipeline
Deploy Pipeline
Solution Architecture Diagram

TDSP

Prós: Abrangente: mais do que apenas um processo – completo com modelos reutilizáveis no GitHub, definições de função e muito mais. Inclusão Opcional do Scrum: O TDSP pode, opcionalmente, ser usado em conjunto com o Scrum (onde um sprint passa por todas as fases). Mantido: A Microsoft parece atualizar seu guia e repositório a cada poucos meses.

Contras: Alguma inconsistência: a Microsoft parece às vezes esquecer de estender algumas de suas atualizações para toda a sua documentação. Curva de aprendizado acentuada: algumas equipes acham a estrutura abrangente complicada de aprender e com muita estrutura (por exemplo, funções de gerenciamento detalhadas e modelos de documentos específicos). Aspectos específicos da Microsoft: Embora grande parte da estrutura seja independente da pilha técnica da Microsoft, existem outras partes da estrutura, especialmente a infraestrutura, que mencionam especificamente os produtos da Microsoft.

Comparando

KDD	SEMMA	CRISP-DM
---	---	Business Understanding
Selection	Sample	Data Understanding
Preprocessing	Explore	
Transformation	Modify	Data Preparation
Data Mining	Model	Modeling
Interpretation/Evaluation	Assess	Evaluation
---	---	Deployment

Fonte: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb

http://eprints.bournemouth.ac.uk/30415/1/DMKD%20for%20BRIAN-20170802_FINAL.pdf

E agora?

- Produto
- Squad
- Metodologias ágeis
- Metodologias de CD

E agora?

- Produto
- Squad
- Metodologias ágeis
- Metodologias de CD

- **Proponham uma metodologia de trabalho que incorpore as necessidades de um projeto de DS**

Referências

- https://towardsdatascience.com/adapting-project-management-methodologies-to-data-science-a710ac9872ea
- https://neptune.ai/blog/data-science-project-management