ThoughtWorks®

Andressa Sivolella

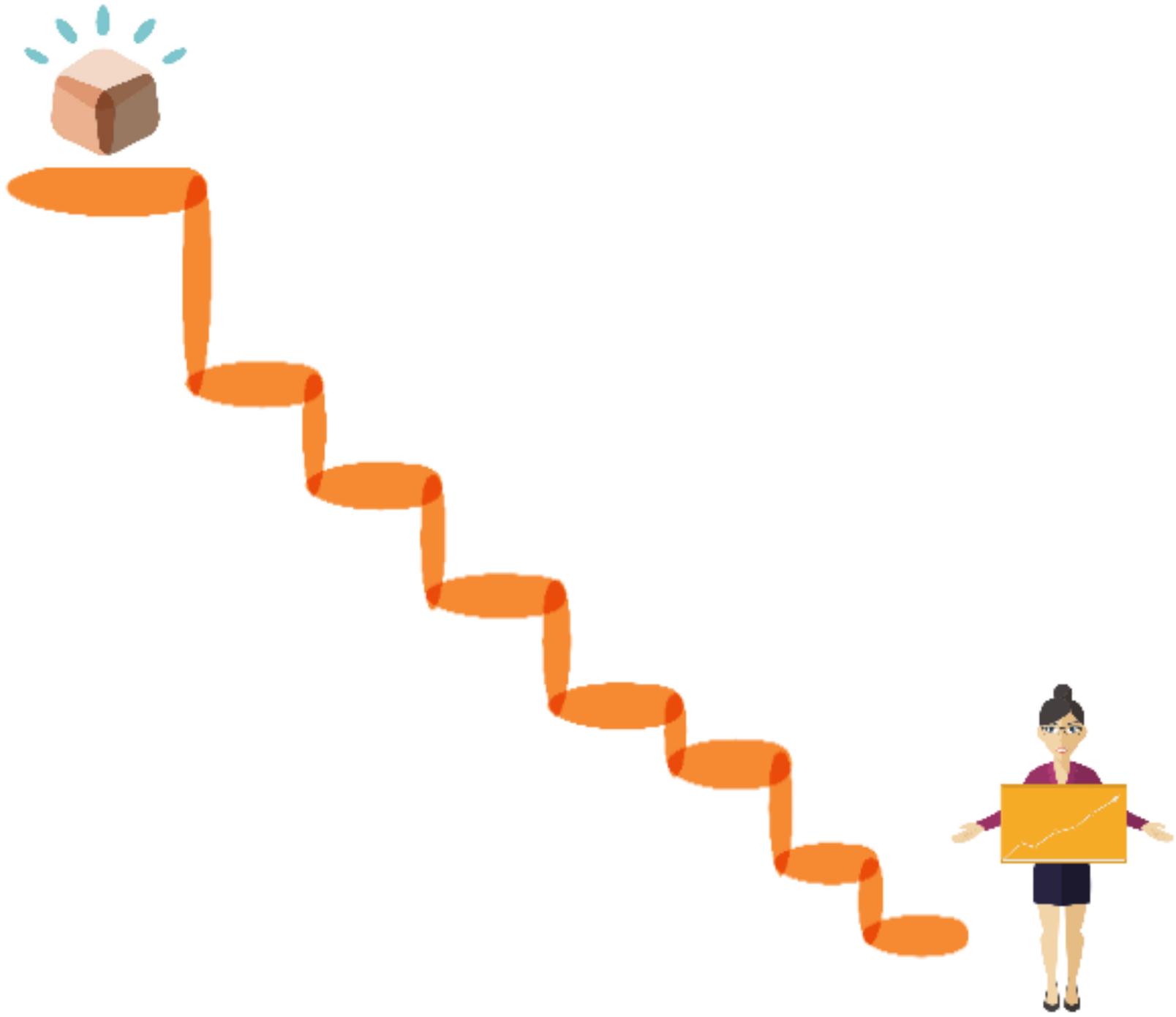
# AGILE ANALYTICS

*Uma abordagem de BI e Analytics Avançado*
*Guiada por Valor de Negócio*

# ESTA É A MARÍLIA...

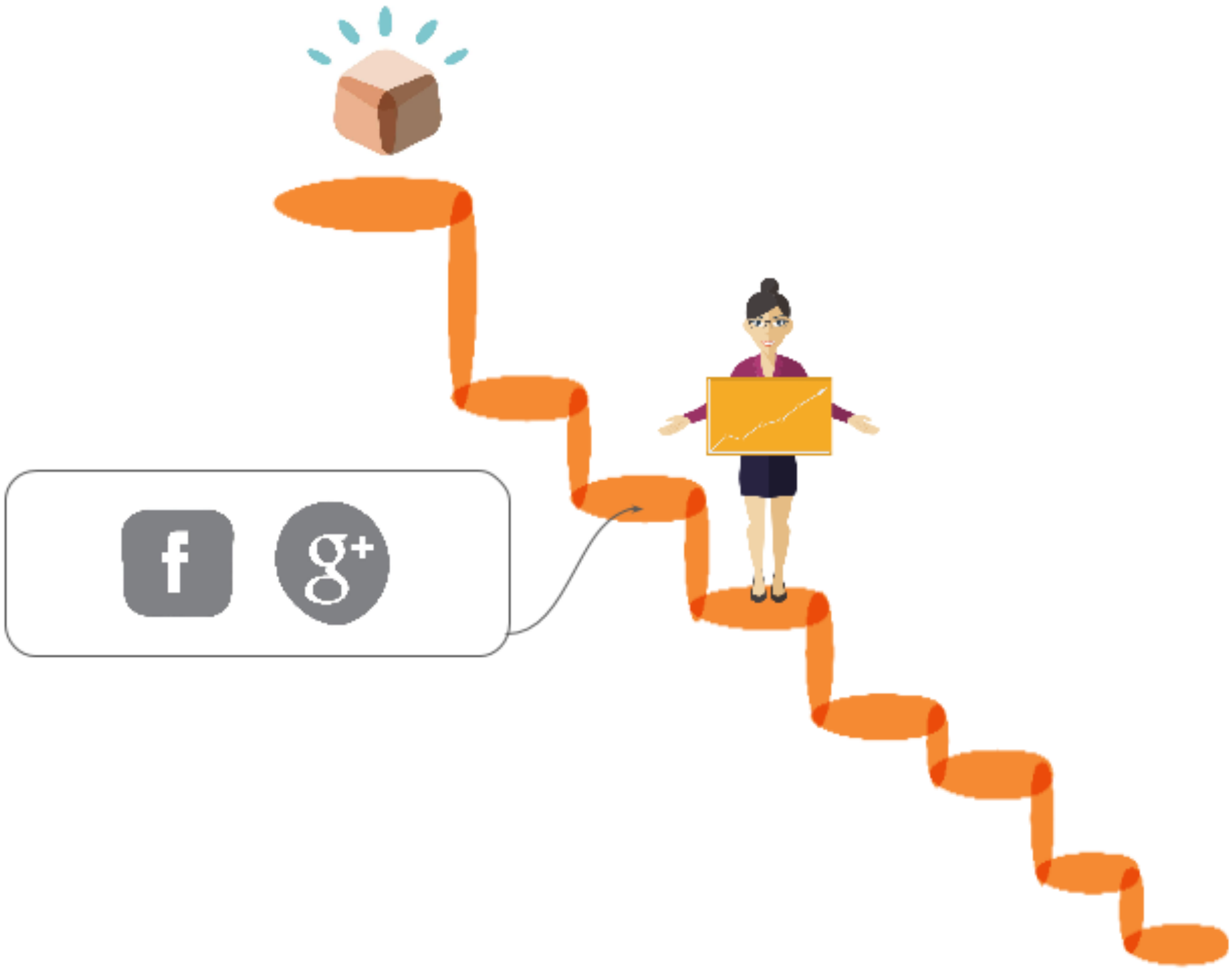
**Gerente de Projetos**

# OUTUBRO DE 2015...

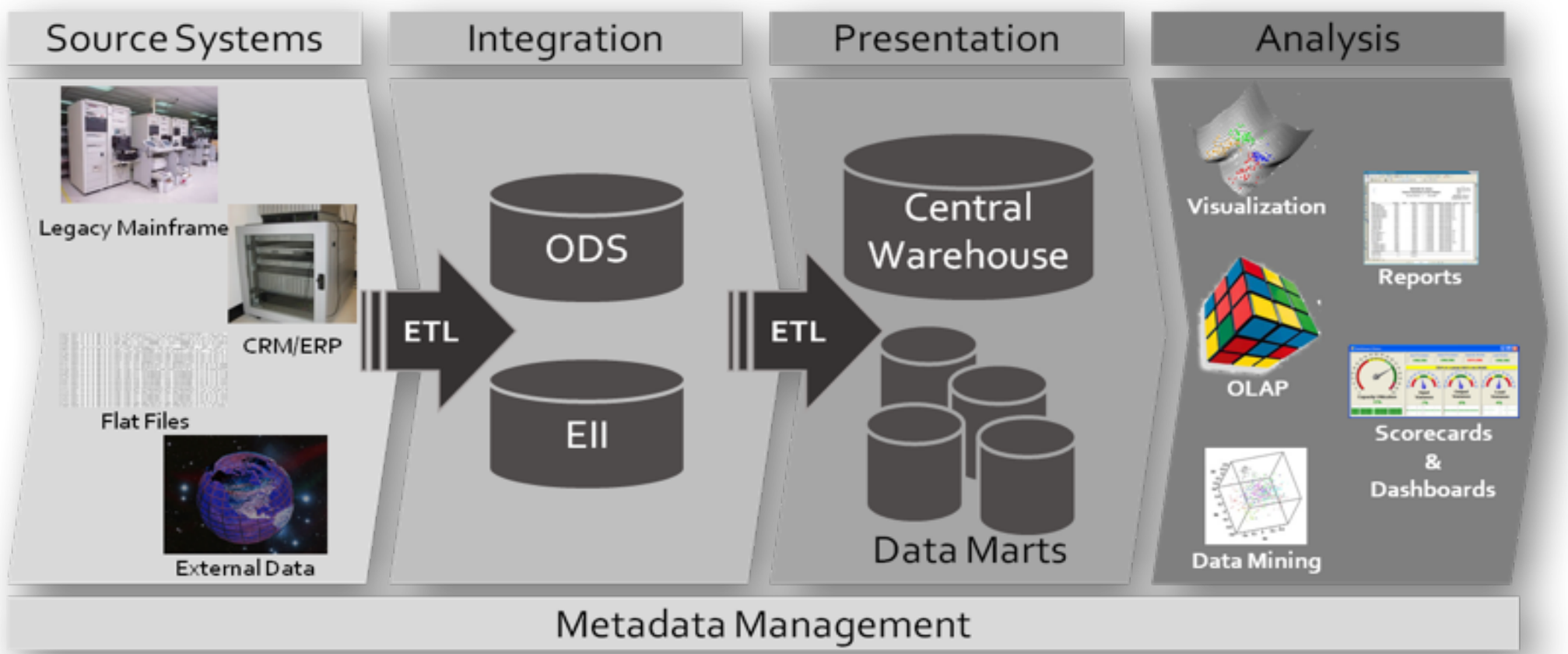


# JANEIRO DE 2016

# JULHO DE 2016

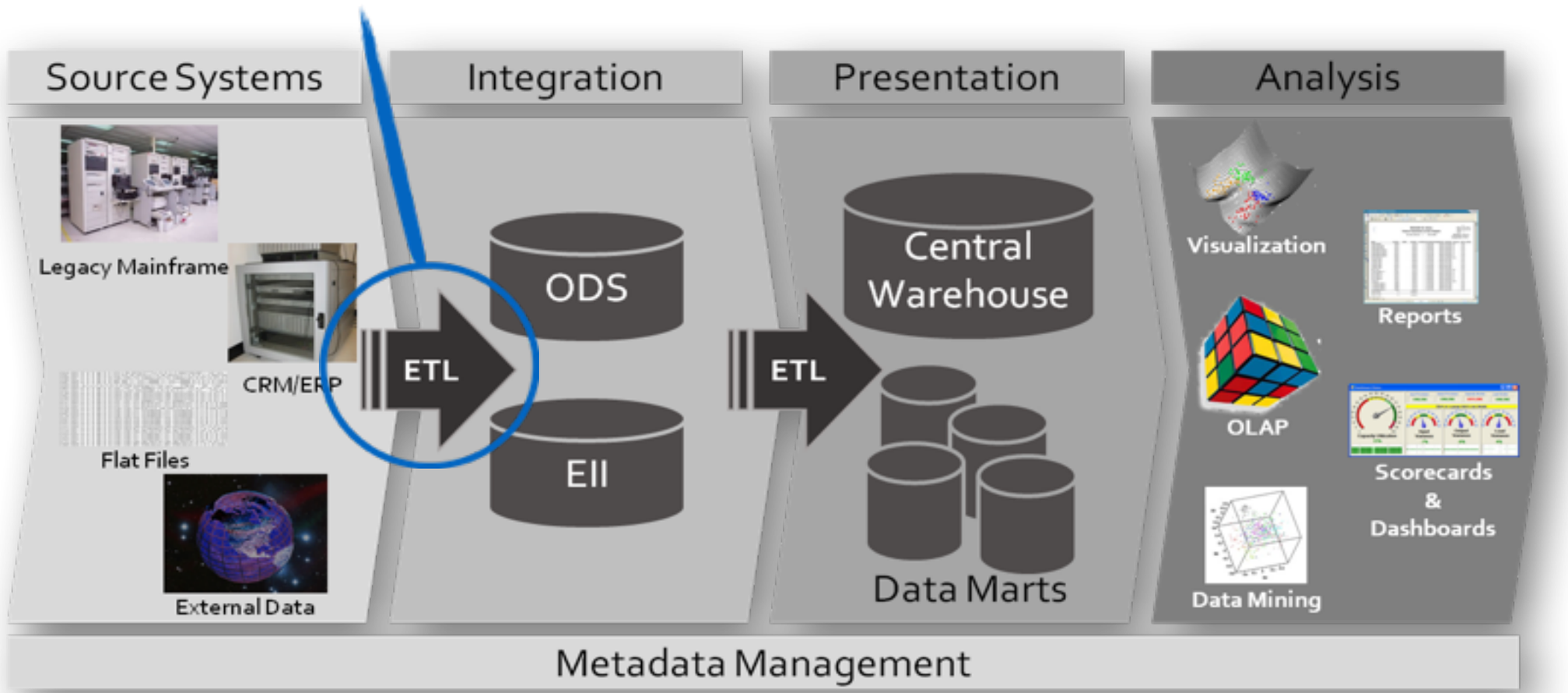


# ARQUITETURAS TRADICIONAIS DE BI & DW



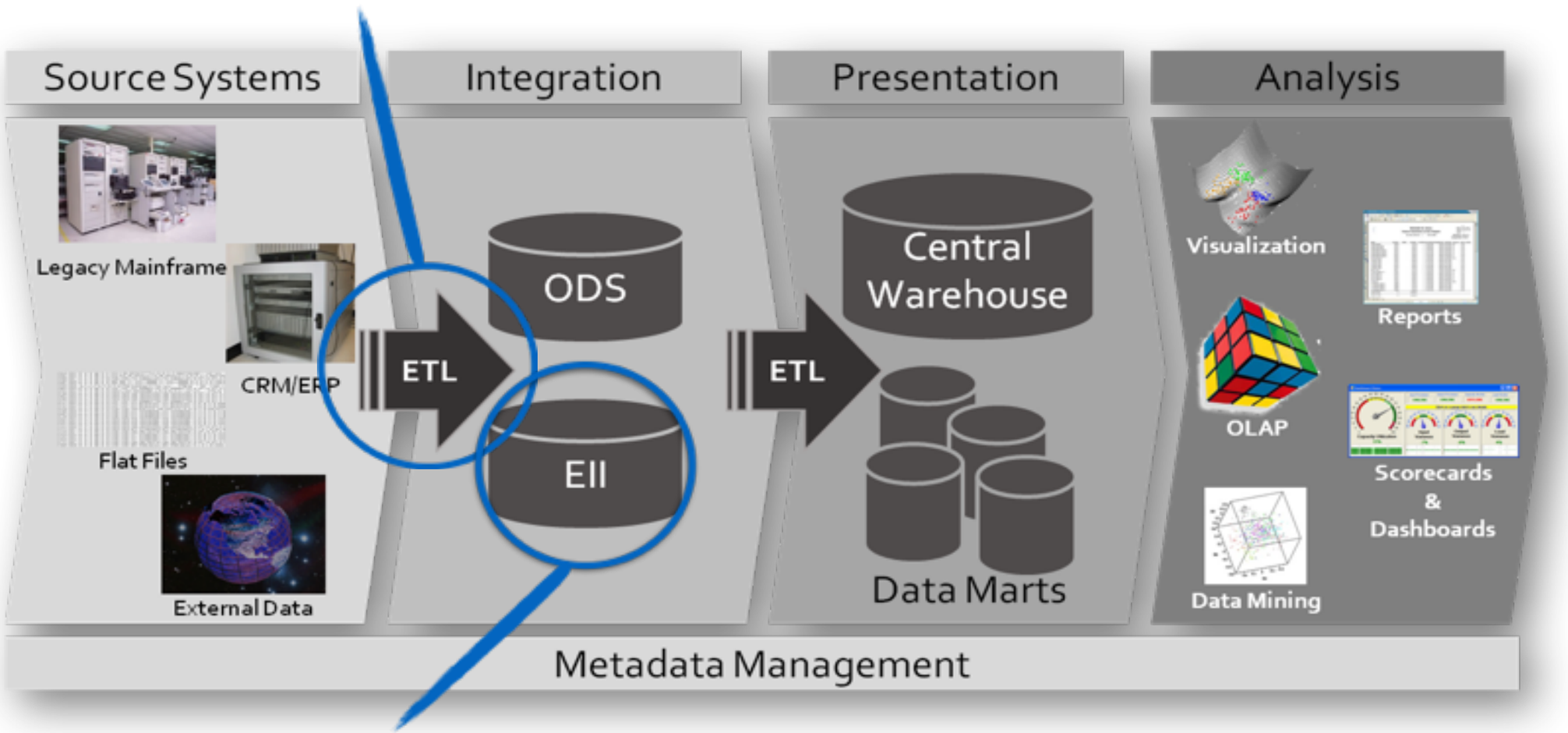
# ARQUITETURAS TRADICIONAIS DE BI & DW

*Extração ativa de dados em lote*



# ARQUITETURAS TRADICIONAIS DE BI & DW

*Extração ativa de dados em lote*

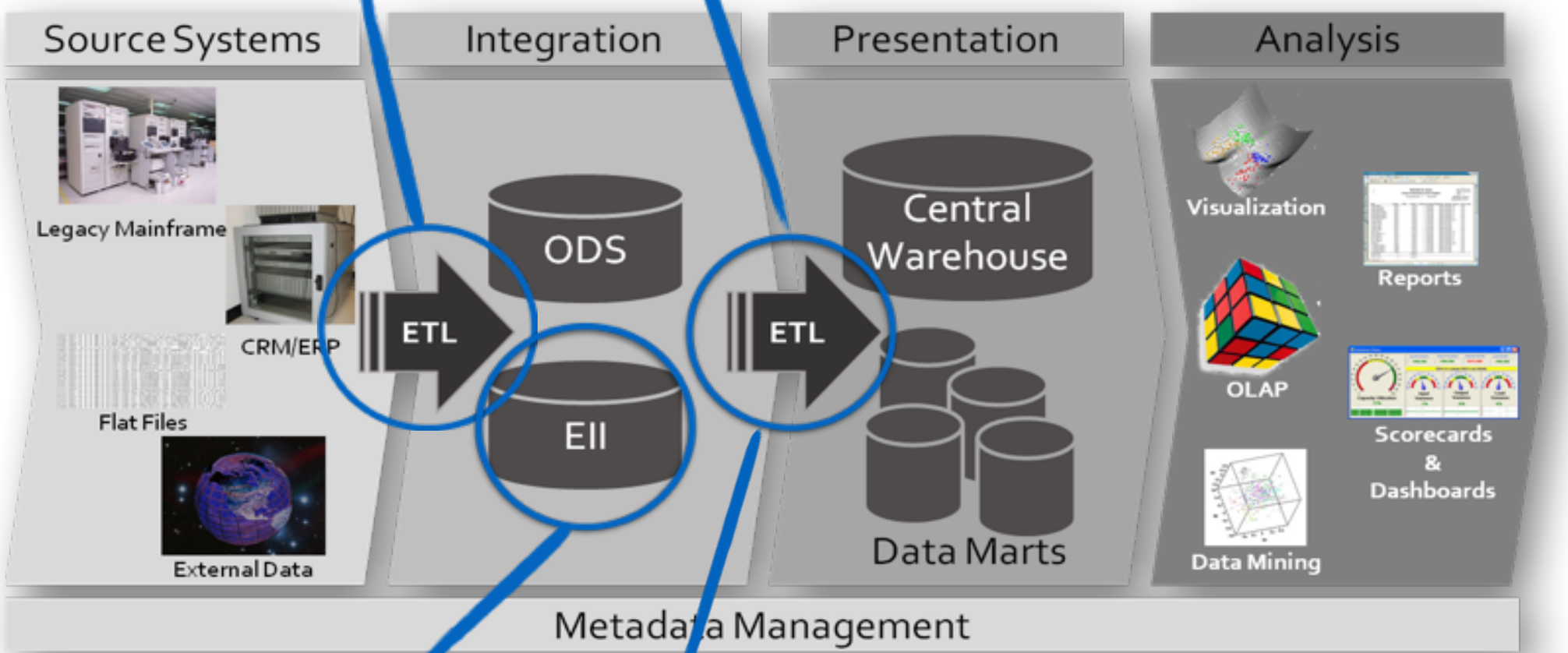



*Modelagem de dados corporativos*

# ARQUITETURAS TRADICIONAIS DE BI & DW

*Extração ativa de dados em lote*

***Lógica de ETL complexa***




*Modelagem de dados corporativos*

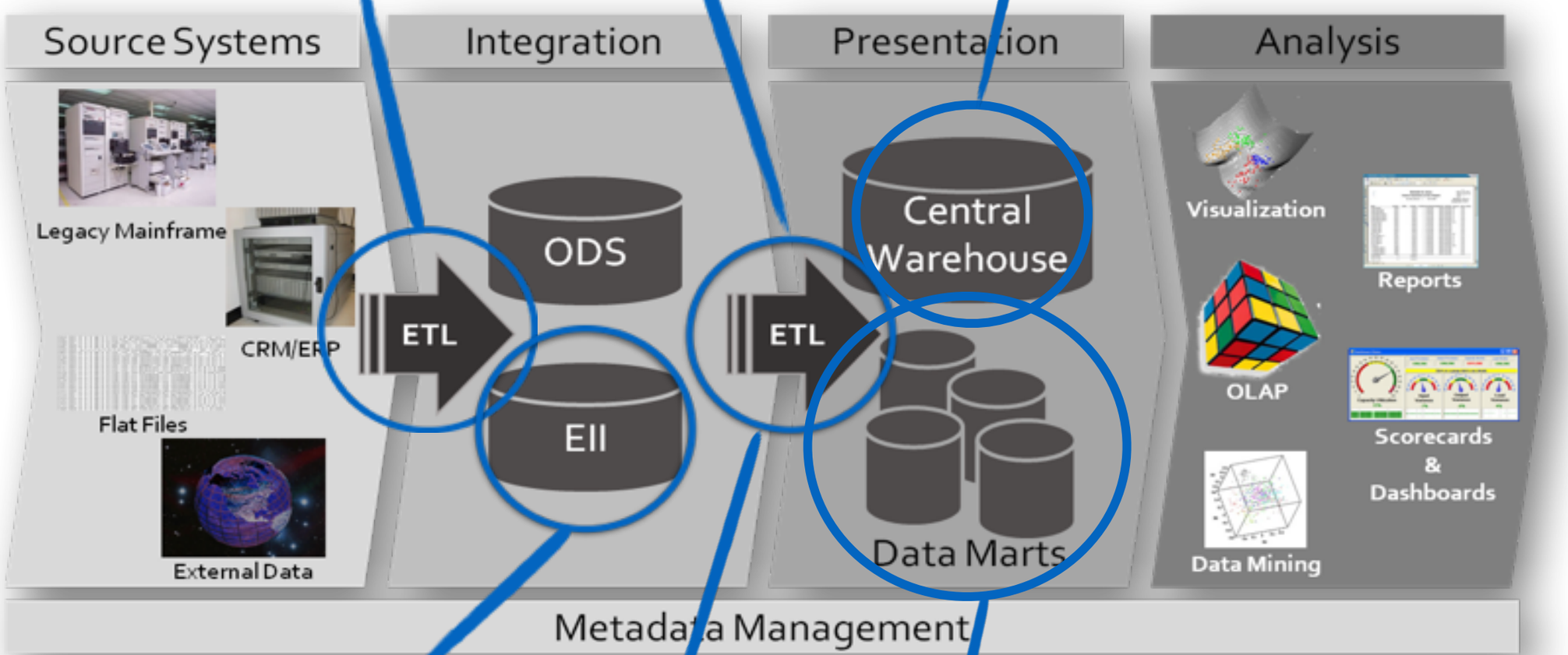
***Jobs ETL de longa duração***

# ARQUITETURAS TRADICIONAIS DE BI & DW

*Extração ativa de dados em lote*

*Lógica de ETL complexa*

***Design emergente é de difícil implementação***




*Modelagem de dados corporativos*

*Jobs ETL de longa duração*

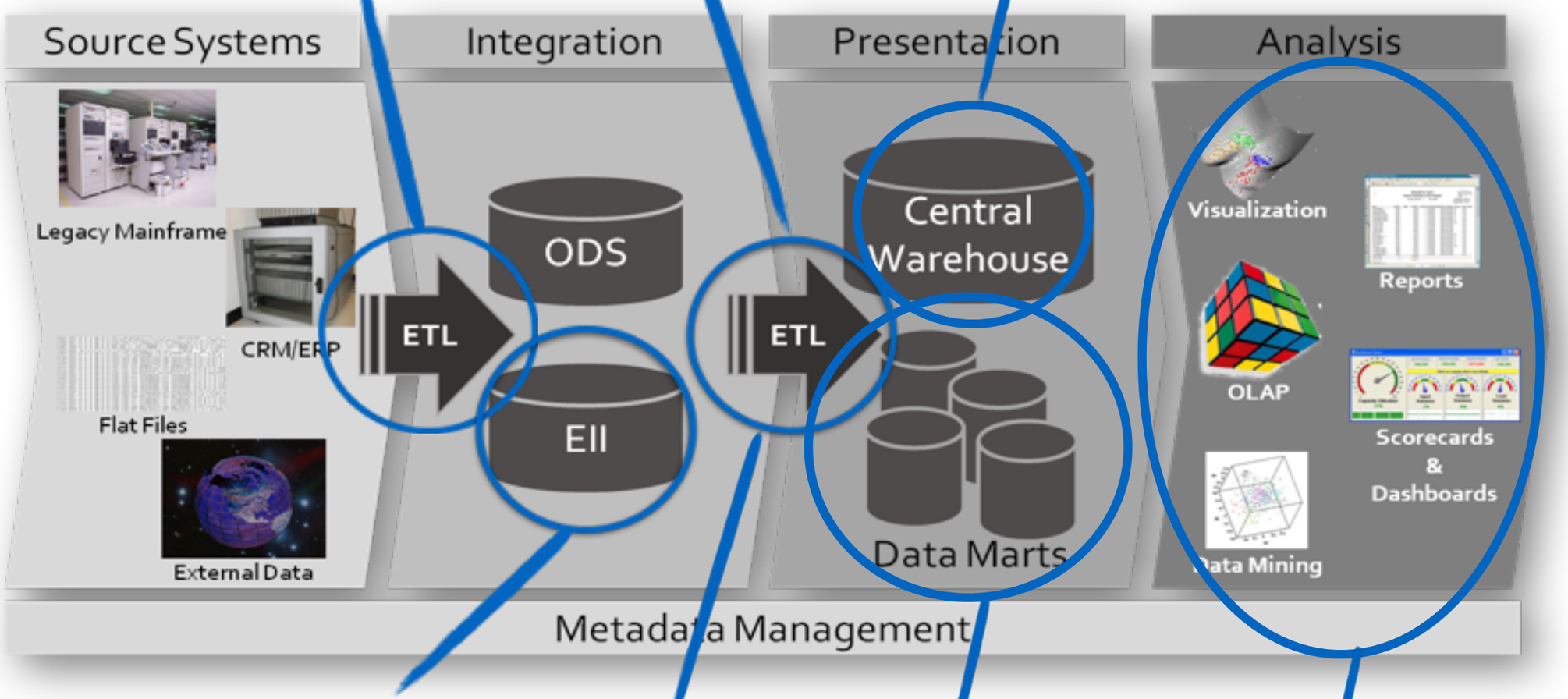
***Não adequado à dados não-relacionais***

# ARQUITETURAS TRADICIONAIS DE BI & DW

*Extração ativa de dados em lote*

*Lógica de ETL complexa*

*Design emergente é de difícil implementação*




*Modelagem de dados corporativos*

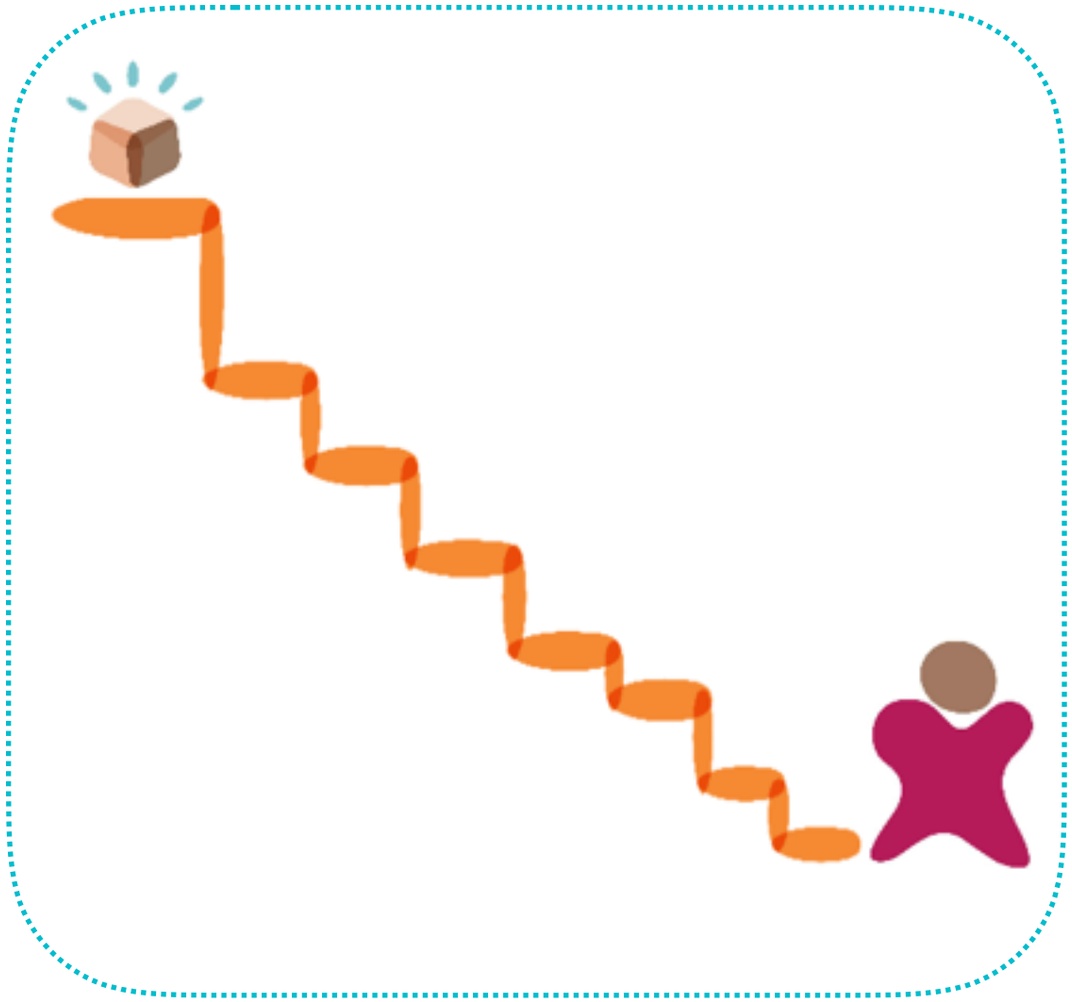
*Jobs ETL de longa duração*

*Não adequado à dados não-relacionais*

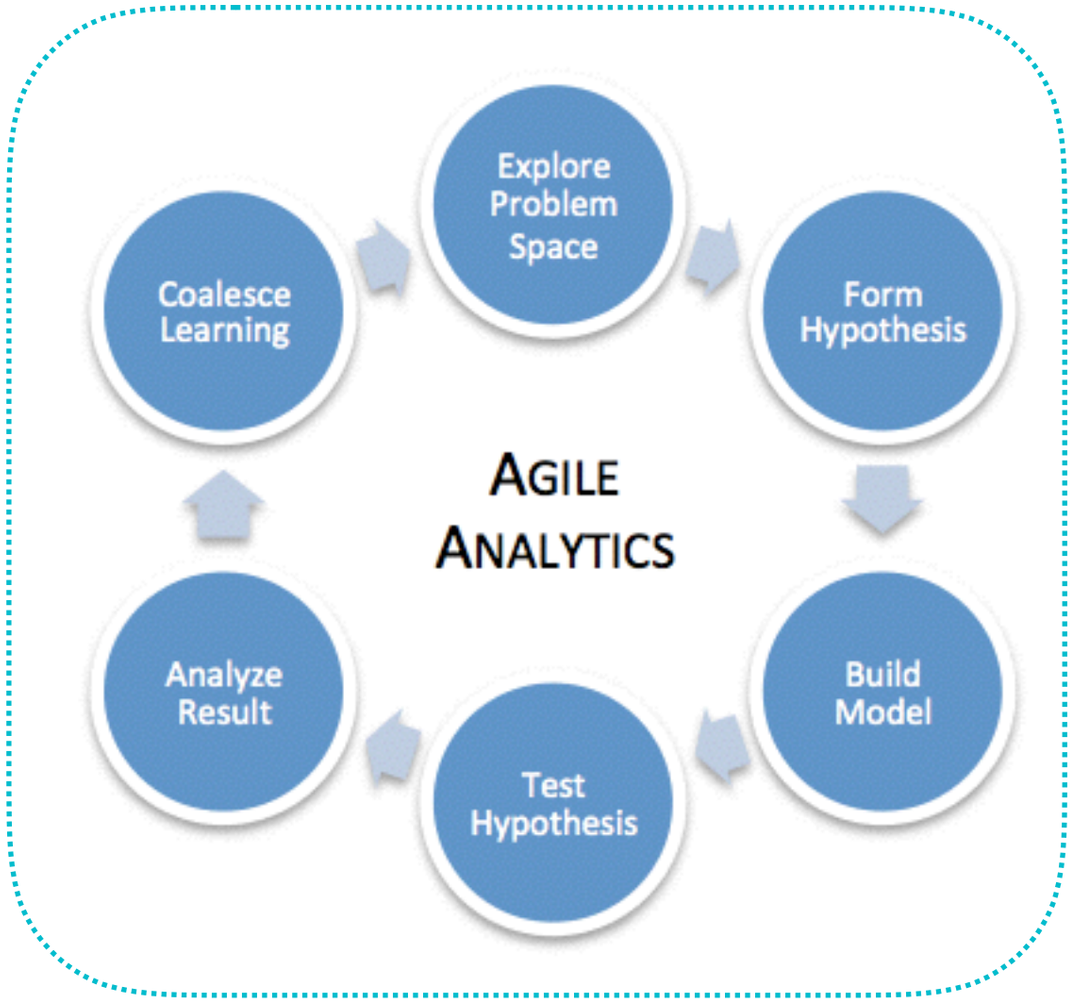
***Relatórios e Diagnósticos***

# MARÍLIA E A ARQUITETURA ATUAL

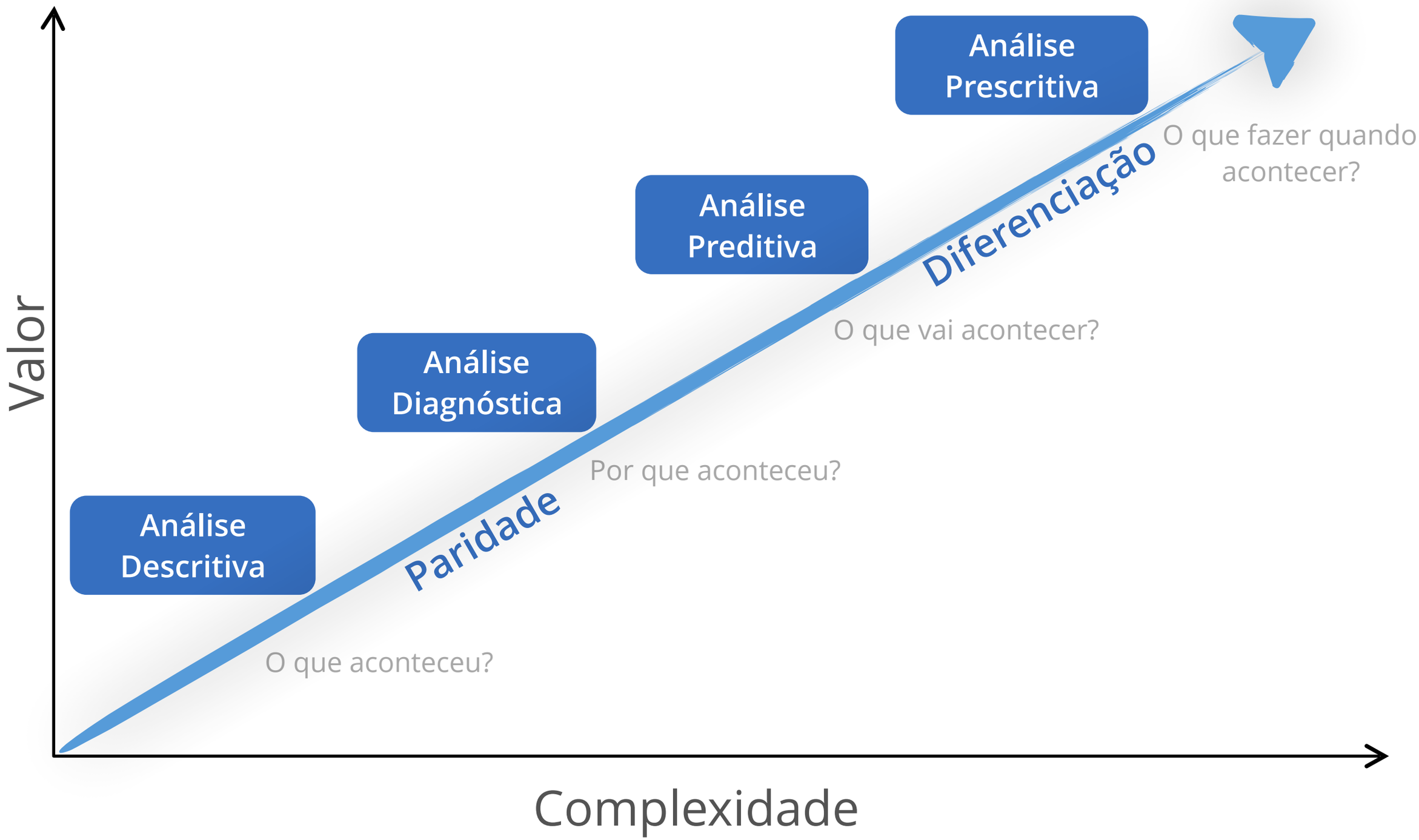
# NOVA ABORDAGEM



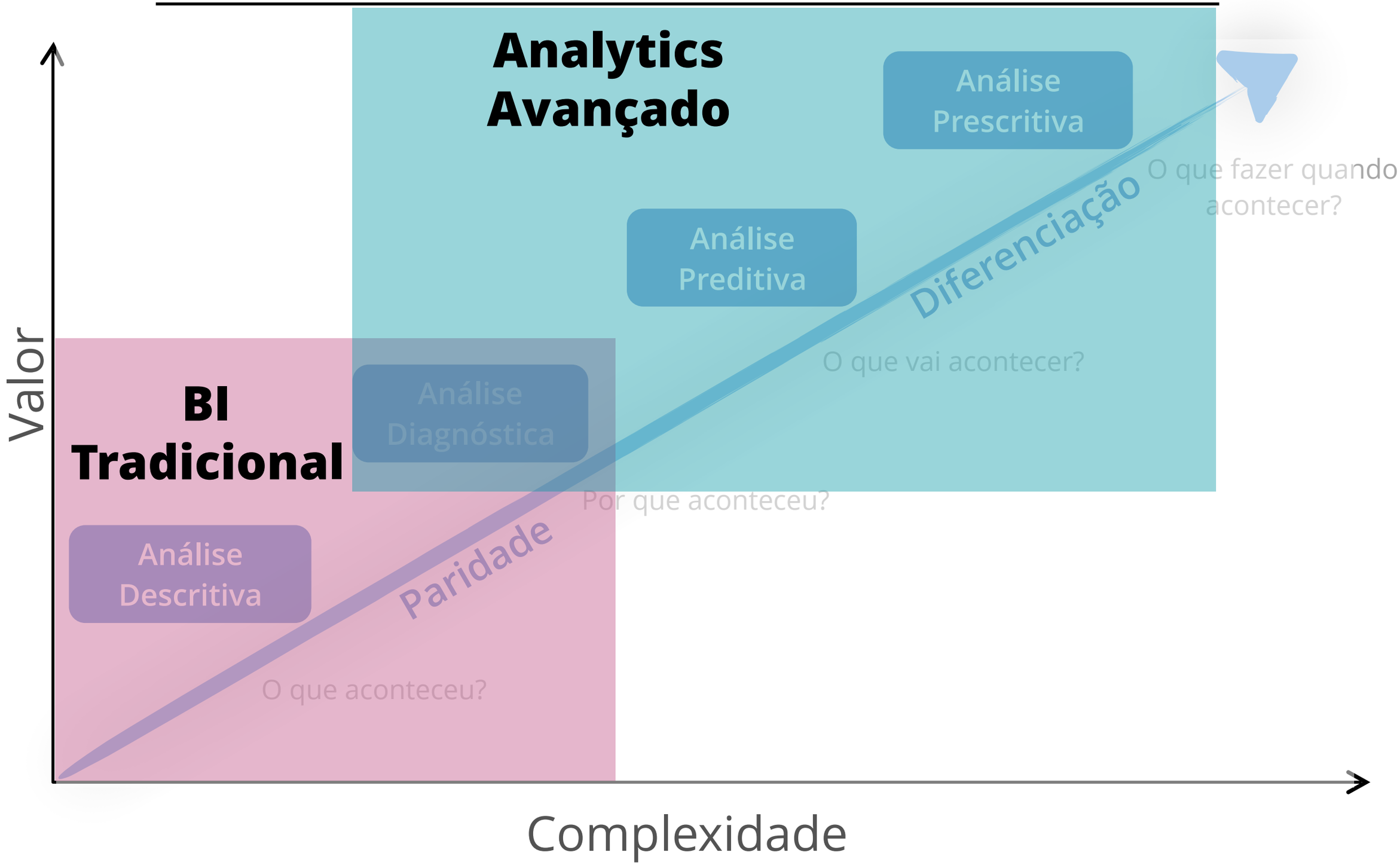
*VS*



# NÍVEIS DE MATURIDADE EM ANALYTICS

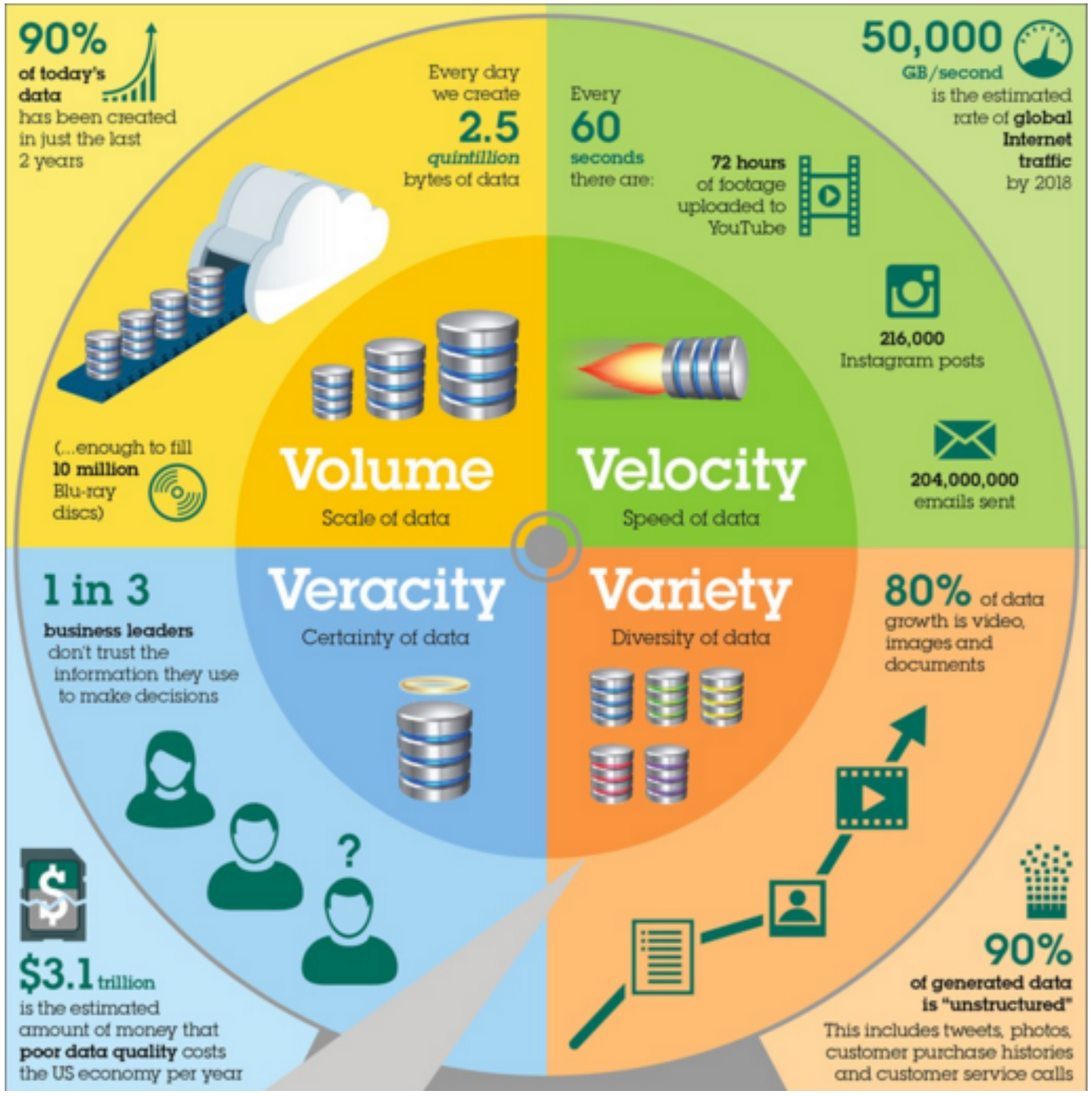


# NÍVEIS DE MATURIDADE EM ANALYTICS



# CENÁRIO ATUAL

- Volume, Velocidade e Variedade → Analytics Avançado



# CENÁRIO ATUAL

- **V**olume, **V**elocidade e **V**ariedade → Analytics Avançado
- **Oportunidade de Inovação**

# CENÁRIO ATUAL

- **V**olume, **V**elocidade e **V**ariedade → Analytics Avançado
- Oportunidade de Inovação
- **Desafios:**
  - **Processamento**
  - **Armazenamento**
  - **Análise**

# CENÁRIO ATUAL

- **V**olume, **V**elocidade e **V**ariedade → Analytics Avançado
- Oportunidade de Inovação
- Desafios:
  - Processamento
  - Armazenamento
  - Análise
- **BI Tradicional não lida com desafios**

# CENÁRIO ATUAL

- **V**olume, **V**elocidade e **V**ariedade → Analytics Avançado
- Oportunidade de Inovação
- Desafios:
  - Processamento
  - Armazenamento
  - Análise
- BI Tradicional não lida com desafios
- **Ferramentas proprietárias**
  - **Dificuldade para integração contínua**

# ANALYTICS AVANÇADO?

### ARQUITETURA E ENGENHARIA DE DADOS

Definição de estruturas e modelos de dados, escolha e implementação das tecnologias adequadas

### CIÊNCIA DE DADOS

Criação de modelos que permitam dar o passo de uma abordagem descritiva para uma preditiva.

### VISUALIZAÇÃO DE DADOS

Design colaborativo, prototipação, testes e avaliação de feedbacks para visualização efetiva dos dados

# É, NÃO É, FAZ, NÃO FAZ

| Analytics Avançado | BI Tradicional |
| --- | --- |
| Baseada em Ciência Estatística | Baseada em sistemas de suporte a decisão (CRM, ERP, ...) |

# É, NÃO É, FAZ, NÃO FAZ

| Analytics Avançado | BI Tradicional |
|---|---|
| Baseada em Ciência Estatística | Baseada em sistemas de suporte a decisão (CRM, ERP, ...) |
| **Computação em Nuvem, massivamente paralela, bancos de dados não relacionais** | **Bancos de dados relacionais** |

# É, NÃO É, FAZ, NÃO FAZ

| Analytics Avançado | BI Tradicional |
|---|---|
| Baseada em Ciência Estatística | Baseada em sistemas de suporte a decisão (CRM, ERP, ...) |
| Computação em Nuvem, massivamente paralela, bancos de dados não relacionais | Bancos de dados relacionais |
| **Data Lake** | **Data Warehouse** |

# É, NÃO É, FAZ, NÃO FAZ

| Analytics Avançado | BI Tradicional |
|---|---|
| Baseada em Ciência Estatística | Baseada em sistemas de suporte a decisão (CRM, ERP, ...) |
| Computação em Nuvem, massivamente paralela, bancos de dados não relacionais | Bancos de dados relacionais |
| Data Lake | Data Warehouse |
| **Foco em estatística e experimentação** | **Foco em relatórios e Dashboards** |

# É, NÃO É, FAZ, NÃO FAZ

| Analytics Avançado | BI Tradicional |
|---|---|
| Baseada em Ciência Estatística | Baseada em sistemas de suporte a decisão (CRM, ERP, ...) |
| Computação em Nuvem, massivamente paralela, bancos de dados não relacionais | Bancos de dados relacionais |
| Data Lake | Data Warehouse |
| Foco em estatística e experimentação | Foco em relatórios e Dashboards |
| **Aprendizado de Máquina** | **OLAP** |

# É, NÃO É, FAZ, NÃO FAZ

| Analytics Avançado | BI Tradicional |
|---|---|
| Baseada em Ciência Estatística | Baseada em sistemas de suporte a decisão (CRM, ERP, ...) |
| Computação em Nuvem, massivamente paralela, bancos de dados não relacionais | Bancos de dados relacionais |
| Data Lake | Data Warehouse |
| Foco em estatística e experimentação | Foco em relatórios e Dashboards |
| Aprendizado de Máquina | OLAP |
| **Ferramentas Open Source (poucas proprietárias)** | **Ferramentas proprietárias (poucas Open Source)** |

# É, NÃO É, FAZ, NÃO FAZ

| Analytics Avançado | BI Tradicional |
|---|---|
| Baseada em Ciência Estatística | Baseada em sistemas de suporte a decisão (CRM, ERP, ...) |
| Computação em Nuvem, massivamente paralela, bancos de dados não relacionais | Bancos de dados relacionais |
| Data Lake | Data Warehouse |
| Foco em estatística e experimentação | Foco em relatórios e Dashboards |
| Aprendizado de Máquina | OLAP |
| Ferramentas Open Source (poucas proprietárias) | Ferramentas proprietárias (poucas Open Source) |
| **Suporta à ausência de dados** | **Requer dados completos** |

# É, NÃO É, FAZ, NÃO FAZ

| Analytics Avançado | BI Tradicional |
|---|---|
| Baseada em Ciência Estatística | Baseada em sistemas de suporte a decisão (CRM, ERP, ...) |
| Computação em Nuvem, massivamente paralela, bancos de dados não relacionais | Bancos de dados relacionais |
| Data Lake | Data Warehouse |
| Foco em estatística e experimentação | Foco em relatórios e Dashboards |
| Aprendizado de Máquina | OLAP |
| Ferramentas Open Source (poucas proprietárias) | Ferramentas proprietárias (poucas Open Source) |
| Suporta à ausência de dados | Requer dados completos |
| **Dados (des)estruturados** | **Dados estruturados** |

# PONTOS CHAVE

- Analytics Avançado **não é** BI ou geração de **relatórios**.

- Analytics é uma forma de alavancar dados e algoritmos de modo a produzir **soluções novas e inovadoras**.

- **Agile** Analytics é modelada para reduzir custos

**Starwood buscava entender quais eventos possuíam o maior impacto nas ações da empresa e encontrou na ThoughtWorks a parceira ideal para ajudar na identificação desses eventos.**

Os resultados foram:

- Melhoria na medição da performance dos hotéis com o desenvolvimento de um **índice de sazonalidade ajustável**
- Identificadas as **tendências** potencialmente **prejudiciais à propriedade**
- Principais indicadores econômicos utilizados pela Starwood **não eram mais relevantes** para prever a performance futura do negócio

starwood
Hotels and Resorts

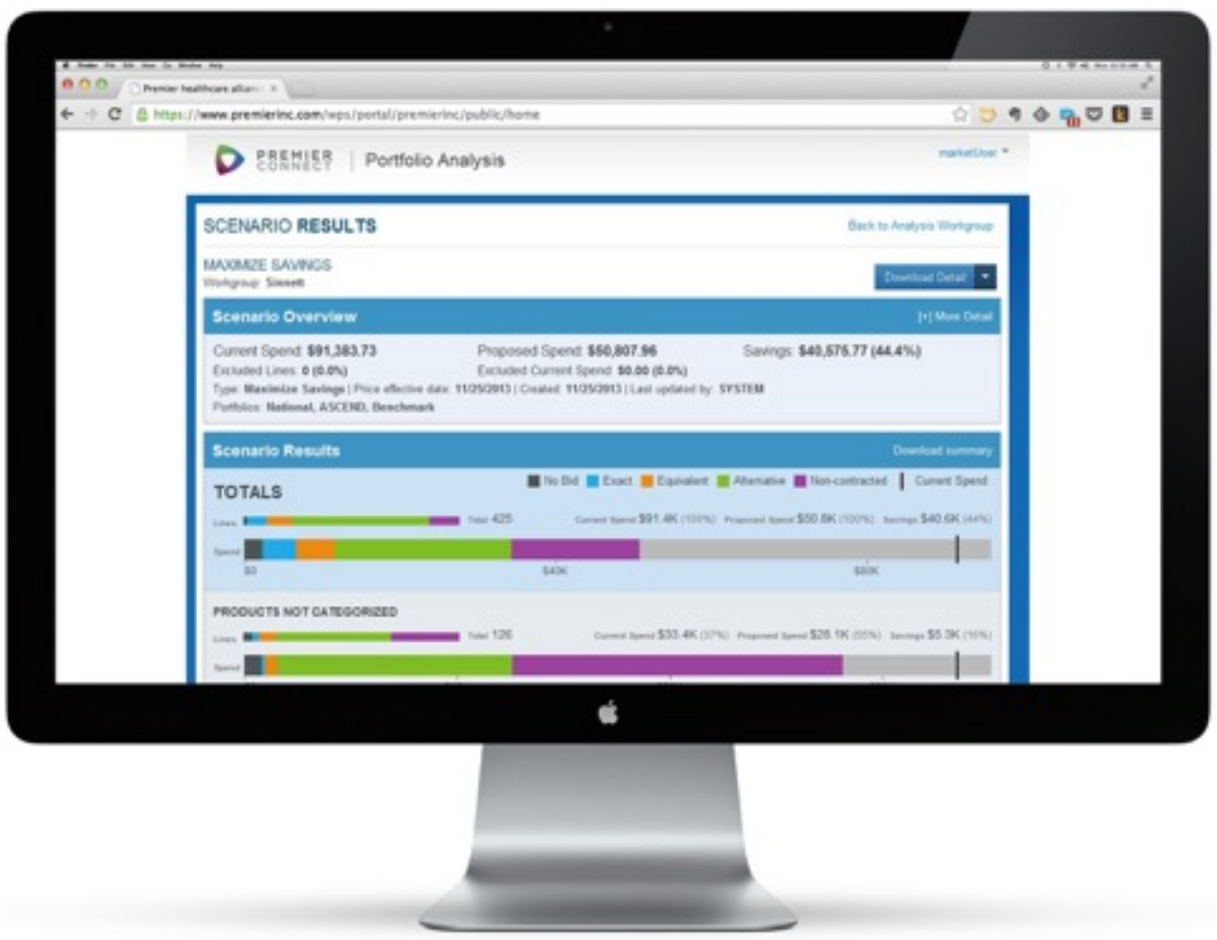
**Rackspace, em parceria com a ThoughtWorks, criou uma plataforma de Cloud Analytics para que o Grupo de Produtos pudesse extrair valor a partir dos dados de seus clientes.**

Os resultados foram:

- **Compreensão** de quais usuários aderem a um plano pago, porque eles aderem e sob quais circunstâncias.
- **Geração contínua de insights** possibilitando novas estratégias de produto.

# PREMIER

**A Premier buscou a ThoughtWorks com a necessidade de simplificar e melhorar a identificação de itens de listas de compras para negociação de contratos.**




Os resultados foram:

- Reimplementação do motor de identificação de itens de contrato para possibilitar cenários mais complexos e conjuntos de dados maiores
- Aumento do percentual de registros identificados de 80% para 95%
- Redução do tempo de processamento de horas para minutos
- App para visualização substituiu o Excel

# REACH LOCAL

**Construção de um sistema de gerenciamento de anúncios para que os usuários finais pudessem gerenciar campanhas online com foco na otimização de performance.**

Geração de recomendações ”inteligentes” com base em algoritmos de machine learning treinados com dados históricos de mercado.

# GRANDE VAREJISTA NOS EUA

## Migração da plataforma de dados utilizando tecnologias de Big Data

### PROBLEMA

- Os sistemas legados utilizados para armazenar e processar dados foram considerados inadequados
- Os algoritmos de previsão de demanda precisavam ser implementados em um ambiente de Big Data

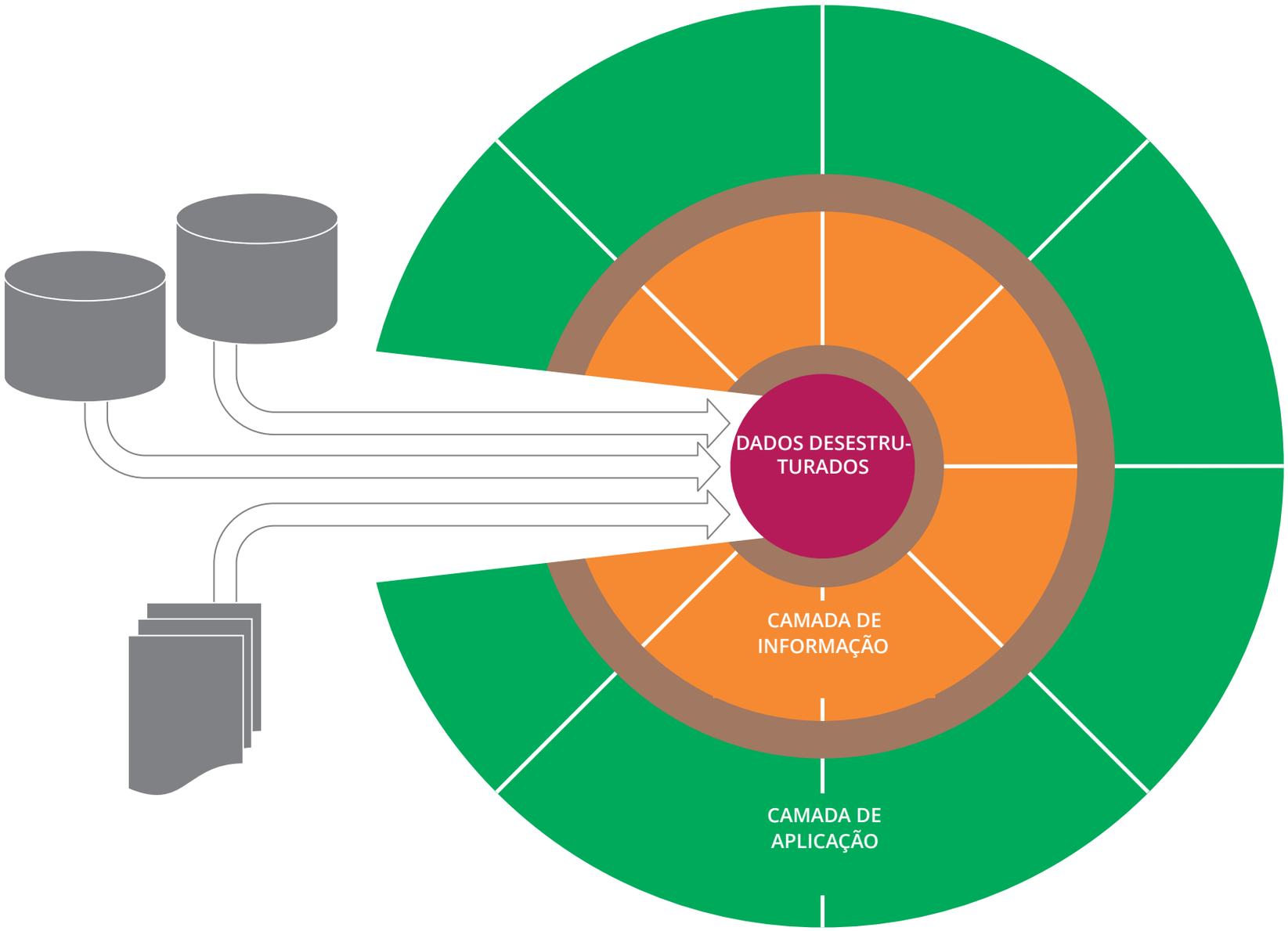
### SOLUÇÃO

- Os dados foram migrados para um cluster Hadoop
- Foram implementados workflows de dados totalmente automatizados para preparação de dados, processamento e publicação
- Foram implementados algoritmos de forecasting utilizando Hive e R (no cluster)
- O treinamento de modelos e previsões foi paralelizado através de operações MapReduce a partir do código em R

# TECNOLOGIAS

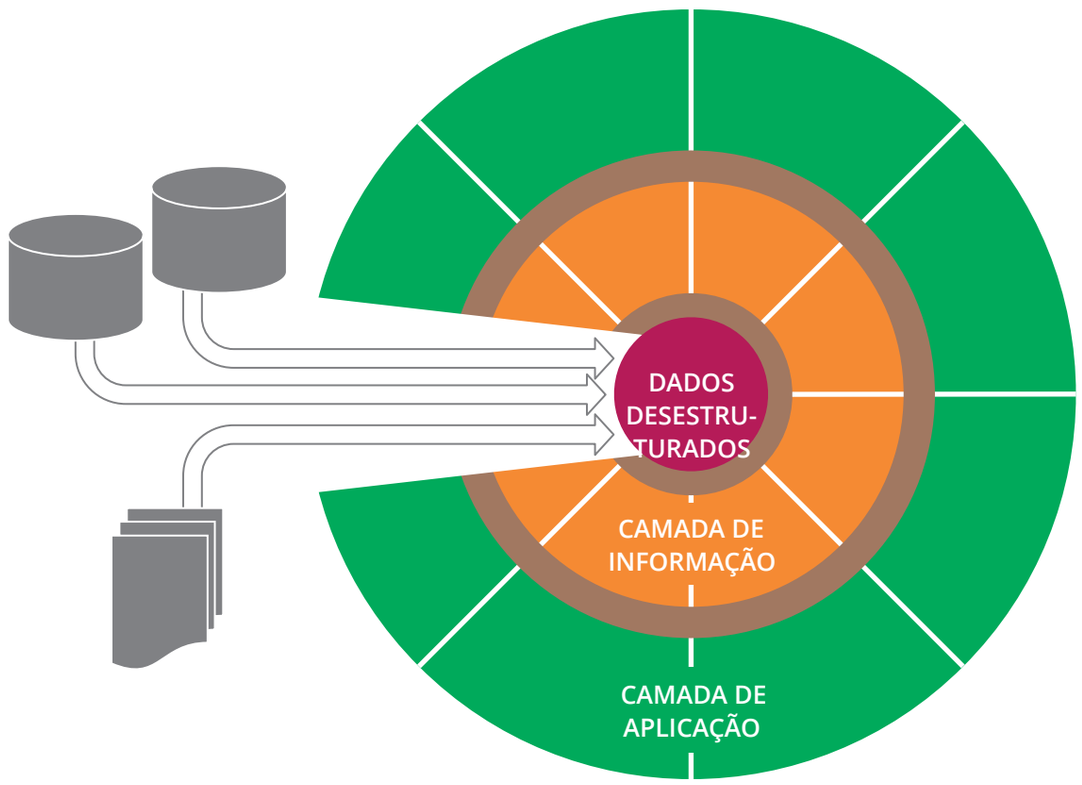
# AGILE ANALYTICS

# ESTRUTURA


DADOS DESESTRU-
TURADOS
CAMADA DE
INFORMAÇÃO
CAMADA DE
APLICAÇÃO

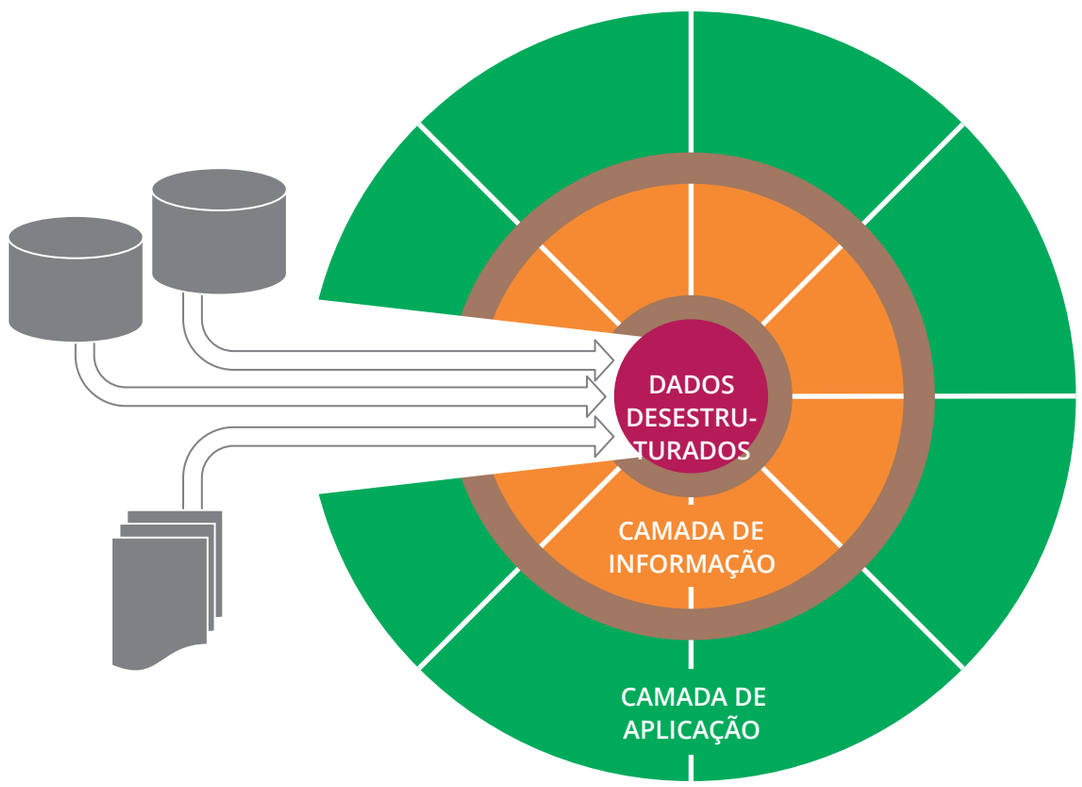
# ESTRUTURA

- Foco nos objetivos de negócio

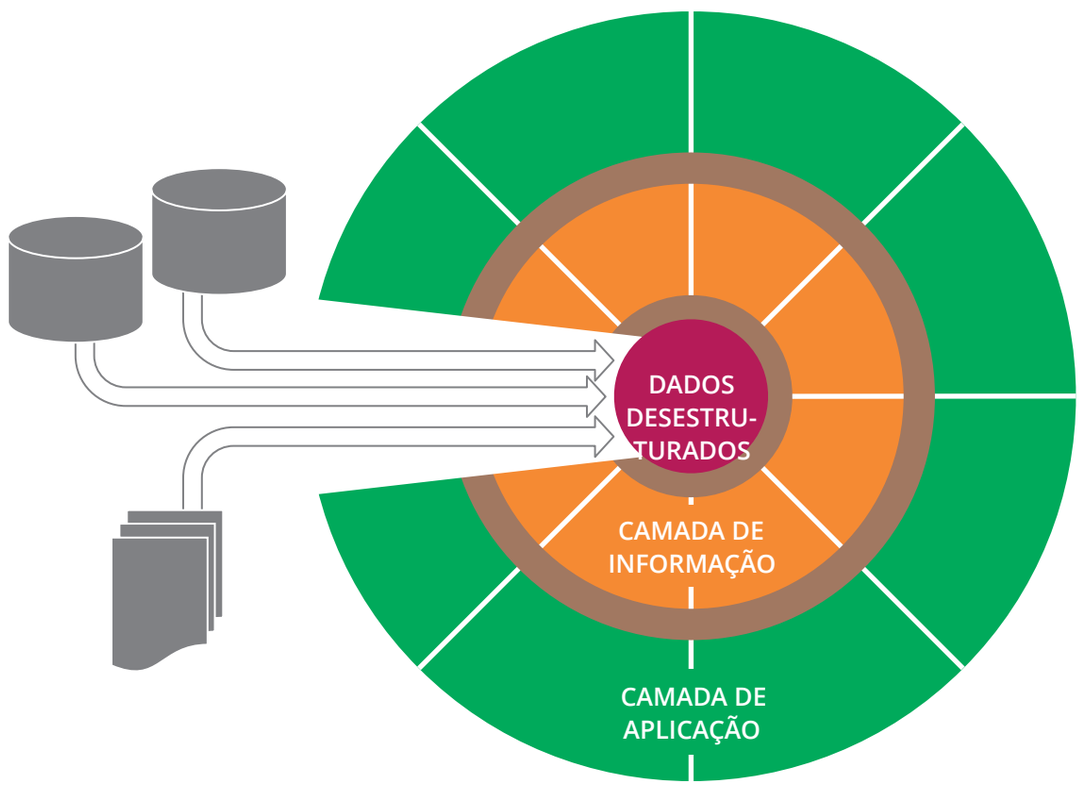


# ESTRUTURA

- Foco nos objetivos de negócio
- **Engajamento constante dos *stakeholders***

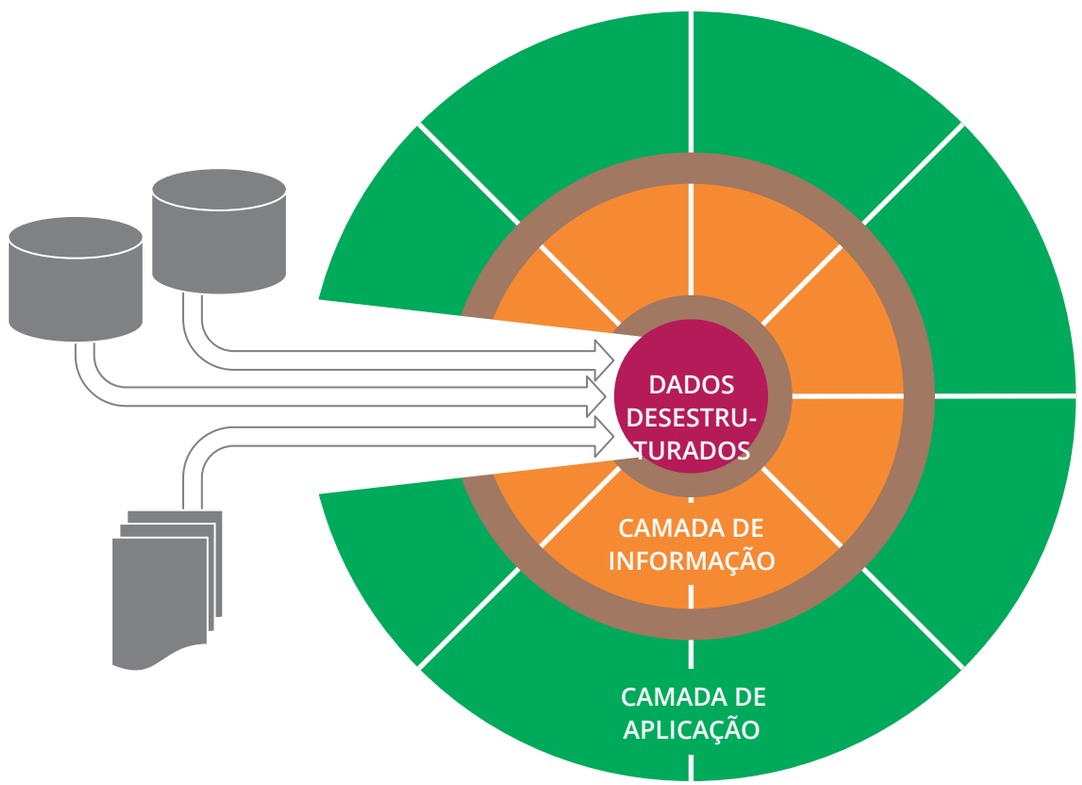


# ESTRUTURA



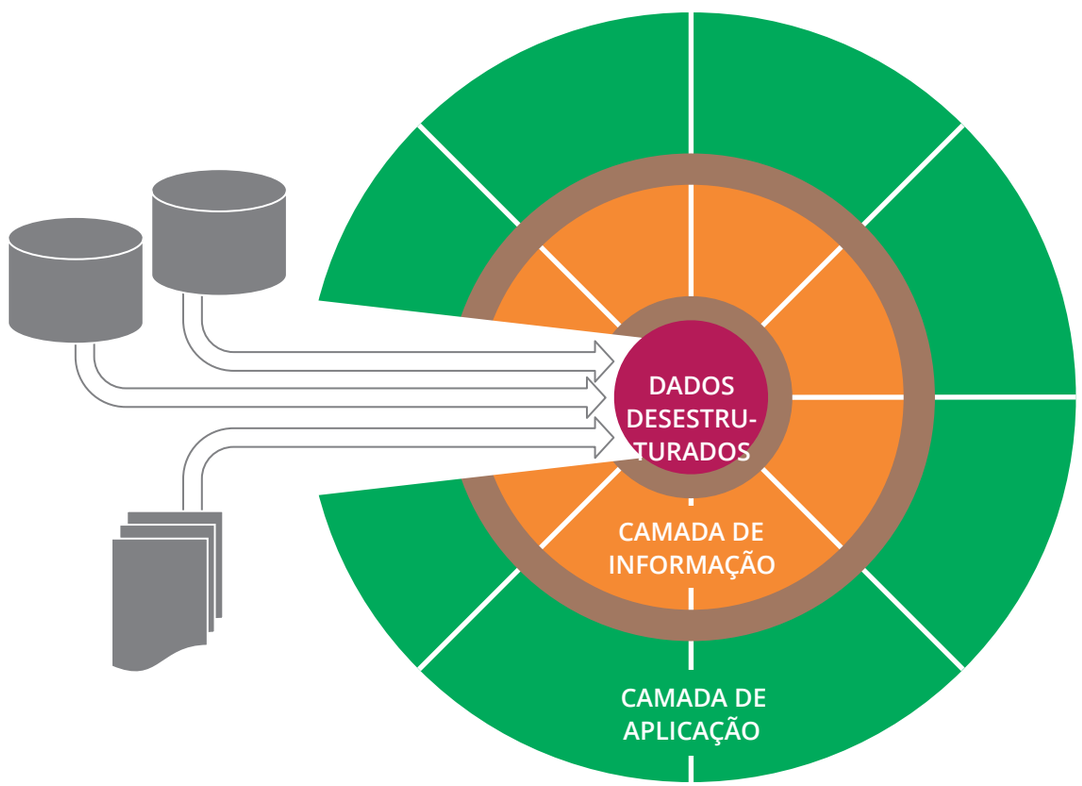

- Foco nos objetivos de negócio
- Engajamento constante dos *stakeholders*
- **Implementação em fatias verticais, da extração de dados às aplicações**

# ESTRUTURA



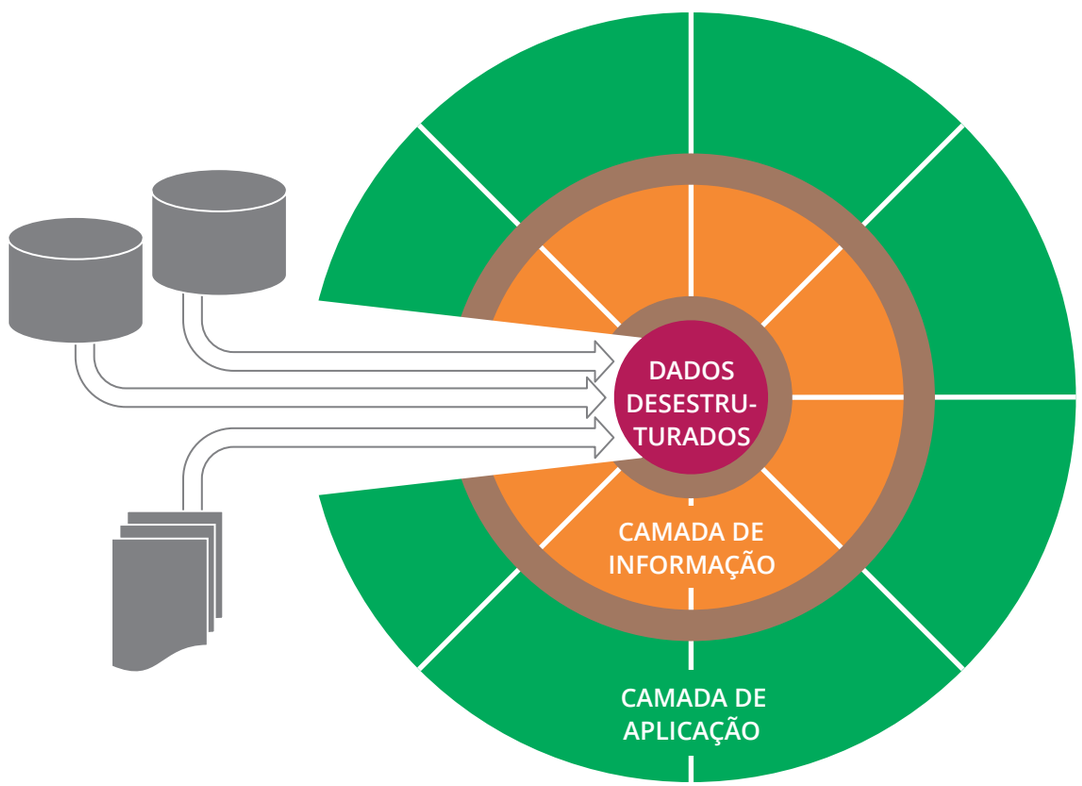

- Foco nos objetivos de negócio
- Engajamento constante dos *stakeholders*
- Implementação em fatias verticais, da extração de dados às aplicações
- **Ingestão incremental de dados, com capacidade *real time***

# ESTRUTURA



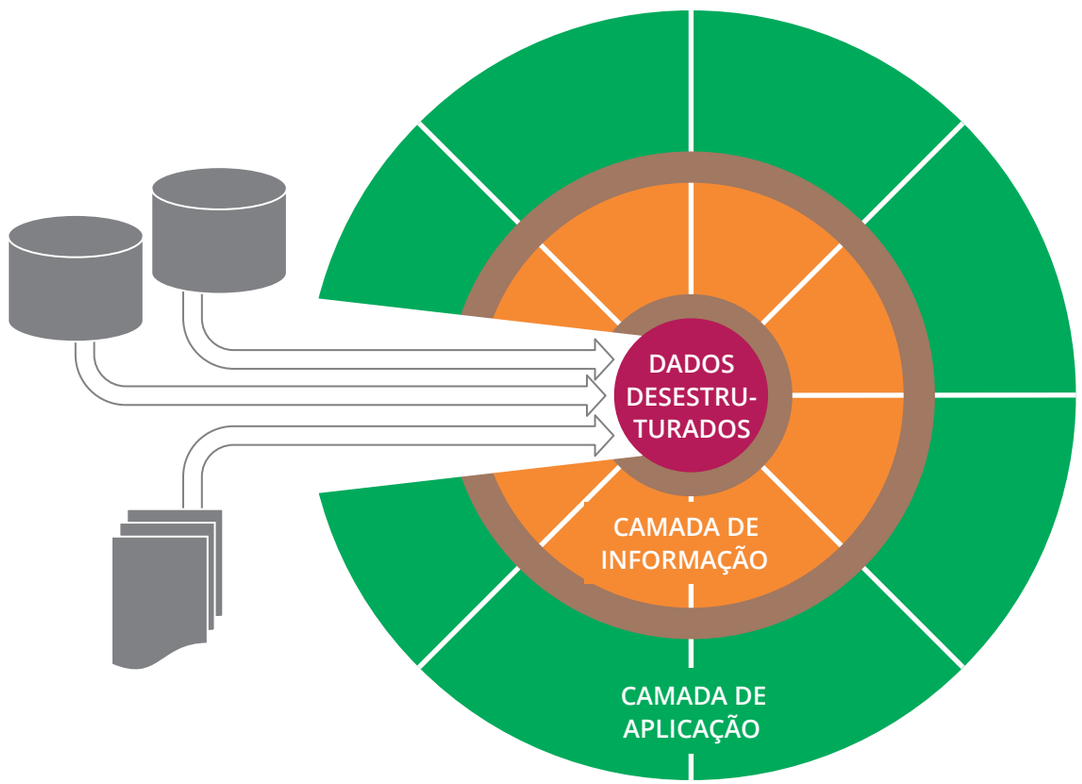

- Foco nos objetivos de negócio
- Engajamento constante dos *stakeholders*
- Implementação em fatias verticais, da extração de dados às aplicações
- Ingestão incremental de dados, com capacidade *real time*
- **Estruturas de dados de cada fatia adequadas à aplicação**

# ESTRUTURA




- Foco nos objetivos de negócio
- Engajamento constante dos *stakeholders*
- Implementação em fatias verticais, da extração de dados às aplicações
- Ingestão incremental de dados, com capacidade *real time*
- Estruturas de dados de cada fatia adequadas à aplicação
- **Lógica perto da aplicação**

# ESTRUTURA

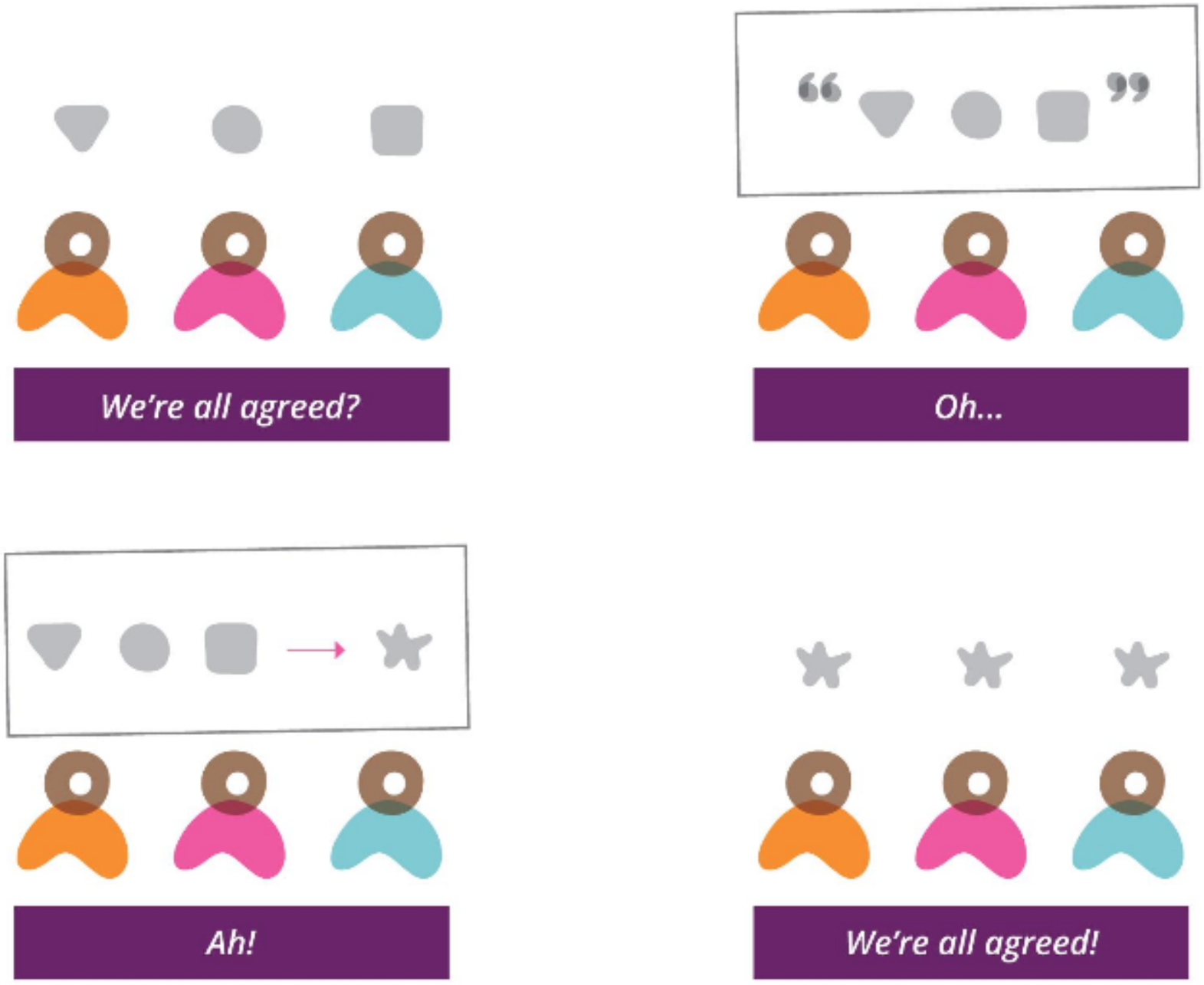



- Foco nos objetivos de negócio
- Engajamento constante dos *stakeholders*
- Implementação em fatias verticais, da extração de dados às aplicações
- Ingestão incremental de dados, com capacidade *real time*
- Estruturas de dados de cada fatia adequadas à aplicação
- Lógica perto da aplicação
- **Design emergente / *"just enough architecture"***

# NA PRÁTICA...

*Por onde eu começo?*

# ALINHAMENTO VIA *INCEPTION*



# ALINHAMENTO VIA *INCEPTION*

# QUAIS SÃO OS ELEMENTOS?

### FACILITAÇÃO E ANÁLISE DE NEGÓCIO

Facilitação de workshops, identificação de cenários e hipóteses, discussão de premissas e soluções

### NEGÓCIO E FONTES DE DADOS

Conhecimento do negócio e fontes de dados

### DESENVOLVIMENTO ÁGIL

Análise, desenvolvimento e qualidade de software com métodos ágeis

### ARQUITETURA E ENGENHARIA DE DADOS

Definição de estruturas e modelos de dados, escolha e implementação das tecnologias adequadas

### CIÊNCIA DE DADOS

Criação de modelos que permitam dar o passo de uma abordagem descritiva para uma preditiva.

### VISUALIZAÇÃO DE DADOS

Design colaborativo, prototipação, testes e avaliação de feedbacks para visualização efetiva dos dados

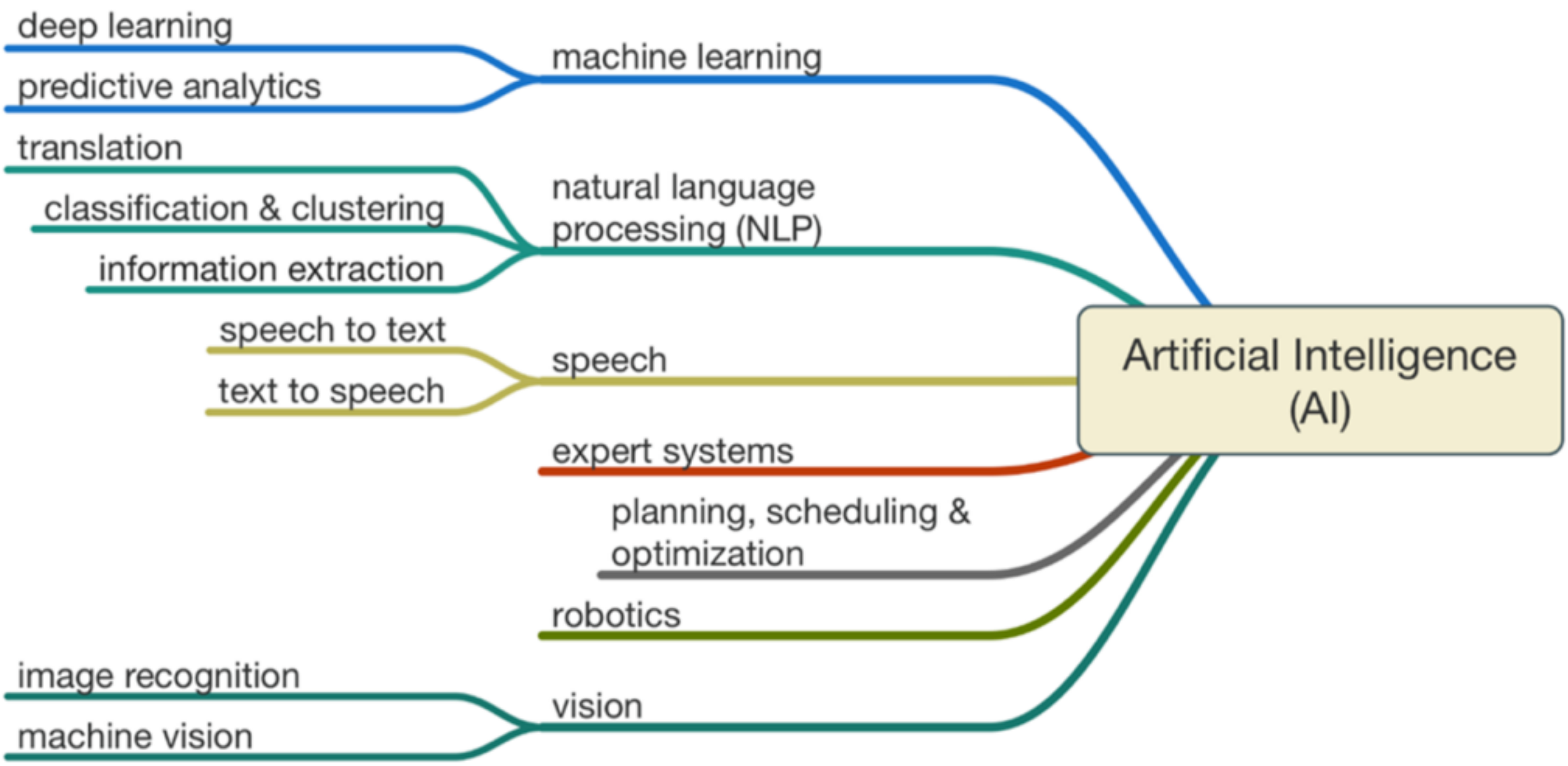
# POR ONDE COMEÇAR?

## CIÊNCIA DE DADOS

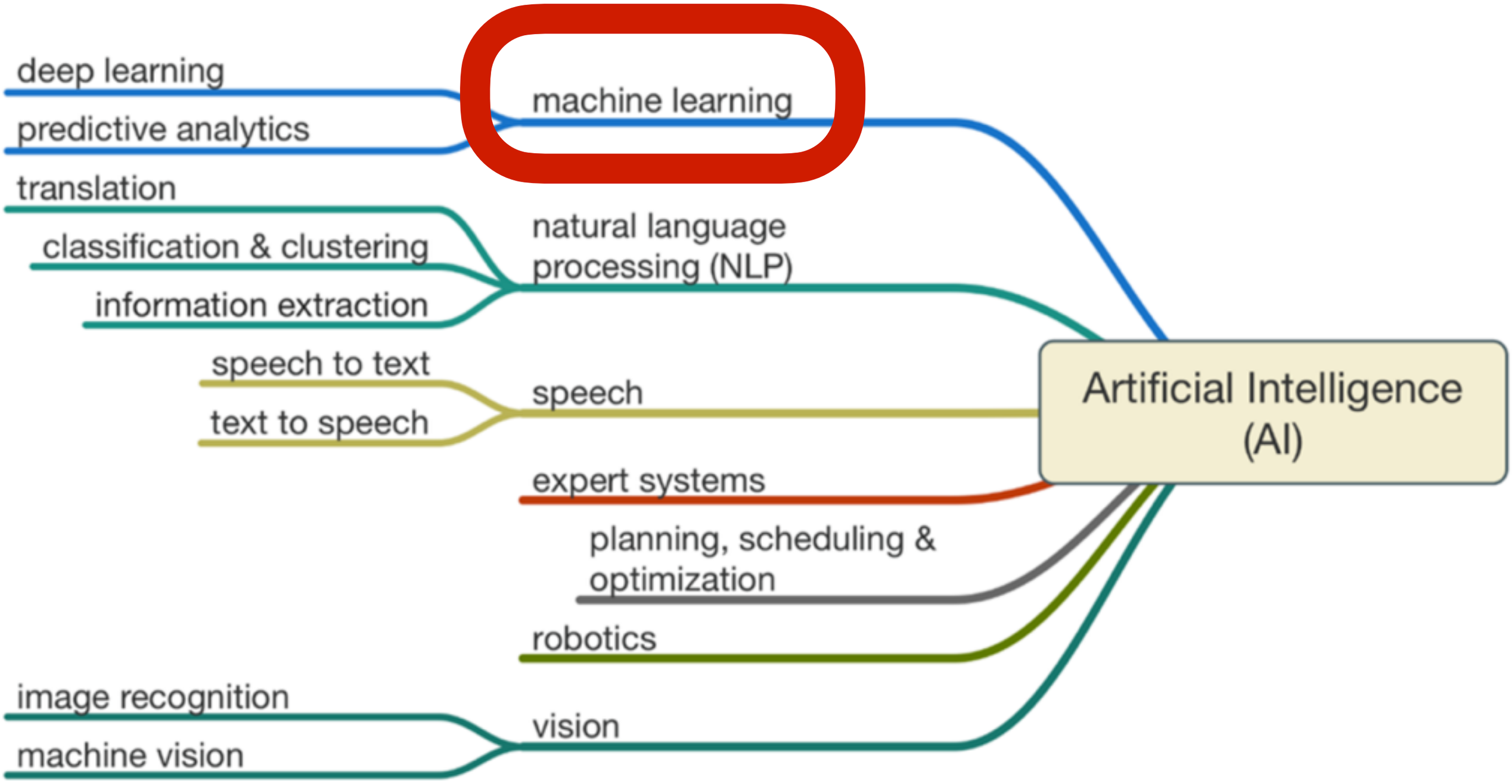
Criação de modelos que permitam dar o passo de uma abordagem descritiva para uma preditiva.

# OPEN SOURCE

# CIÊNCIA DE DADOS



# CIÊNCIA DE DADOS



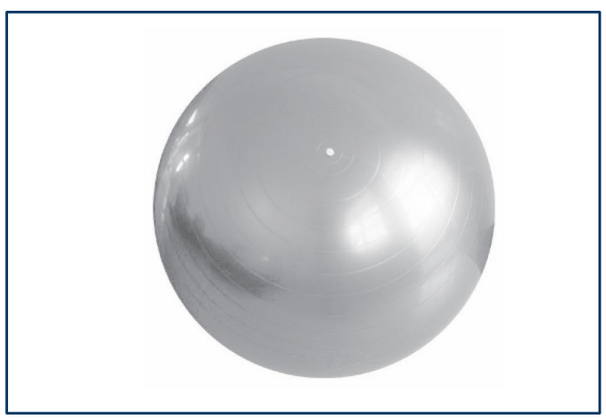
# O QUE É MACHINE LEARNING?

# ABORDAGEM DE REGRAS

- Cores?
- Tipo de costura?
- Tamanho?
- ...

# ABORDAGEM DE ANALYTICS AVANÇADO

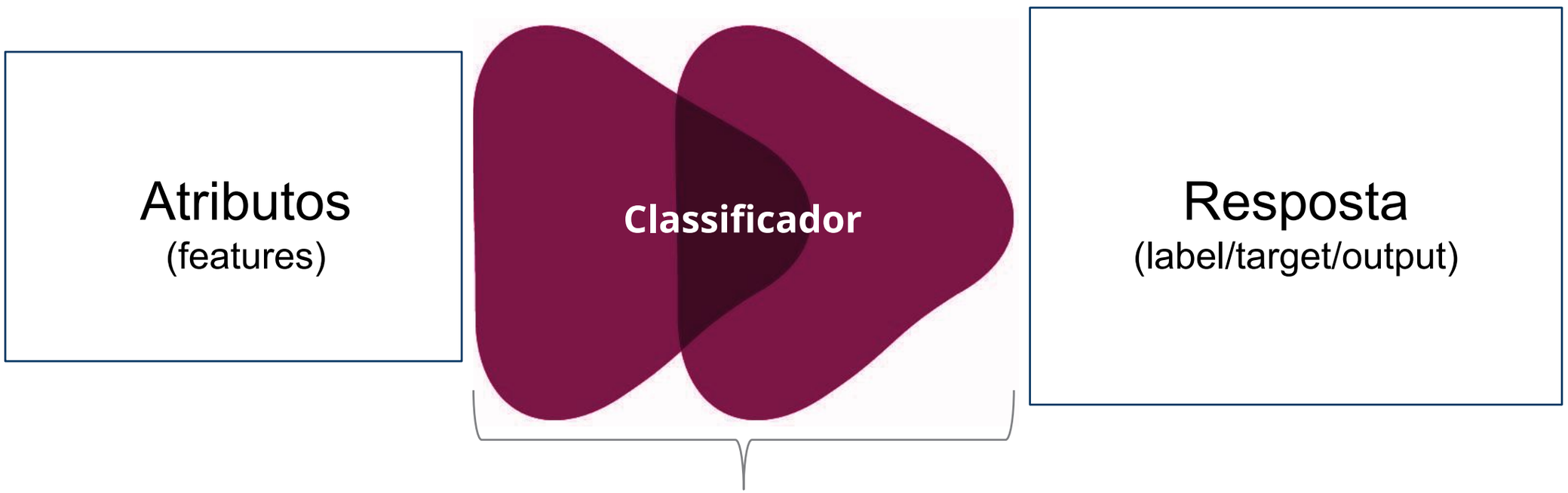
Atributos
(features)



Resposta
(label/target/output)

Treinado a partir de
conjunto conhecido
(supervisionado)

# ABORDAGEM DE ANALYTICS AVANÇADO




**Algoritmos** encontram padrões nas amostras conhecidas

# ABORDAGEM DE ANALYTICS AVANÇADO

Atributos
(features)

Resposta
(label/target/output)

- **árvore de decisão???**
- rede neural
- regressão linear
- ensemble
- ...

Marília joga tênis

# MARÍLIA JOGA TÊNIS?

- Tempo
- Umidade
- Vento
- ...

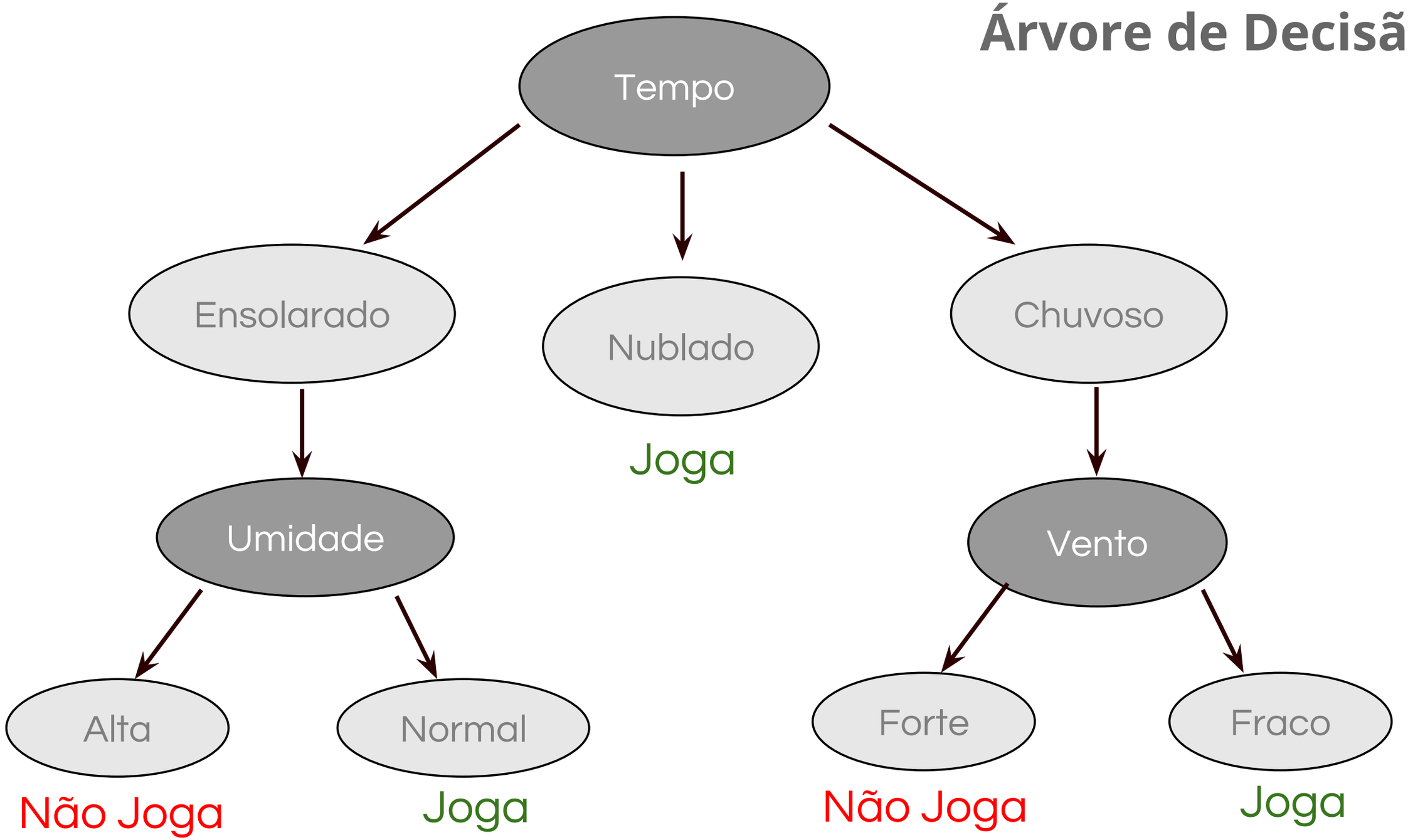
**Árvore de Decisão**

- Não joga
- Joga

| # Dia | Tempo | Umidade | Vento | Joga? |
|---|---|---|---|---|
| 1 | Ensolarado | Alta | Fraco | **Não** |
| 2 | Ensolarado | Alta | Fraco | **Não** |
| 3 | Nublado | Alta | Fraco | **Sim** |
| 4 | Chuvoso | Alta | Forte | **Sim** |
| 5 | Chuvoso | Normal | Fraco | **Sim** |
| 6 | Chuvoso | Normal | Forte | **Sim** |
| 7 | Nublado | Normal | Forte | **Sim** |
| 8 | Ensolarado | Alta | Fraco | **Não** |
| 9 | Ensolarado | Normal | Fraco | **Sim** |
| 10 | Chuvoso | Normal | Fraco | **Sim** |
| 11 | Ensolarado | Normal | Forte | **Sim** |
| 12 | Nublado | Alta | Forte | **Sim** |
| 13 | Nublado | Normal | Fraco | **Sim** |
| 14 | Chuvoso | Alta | Forte | **Não** |

2 semanas de Janeiro

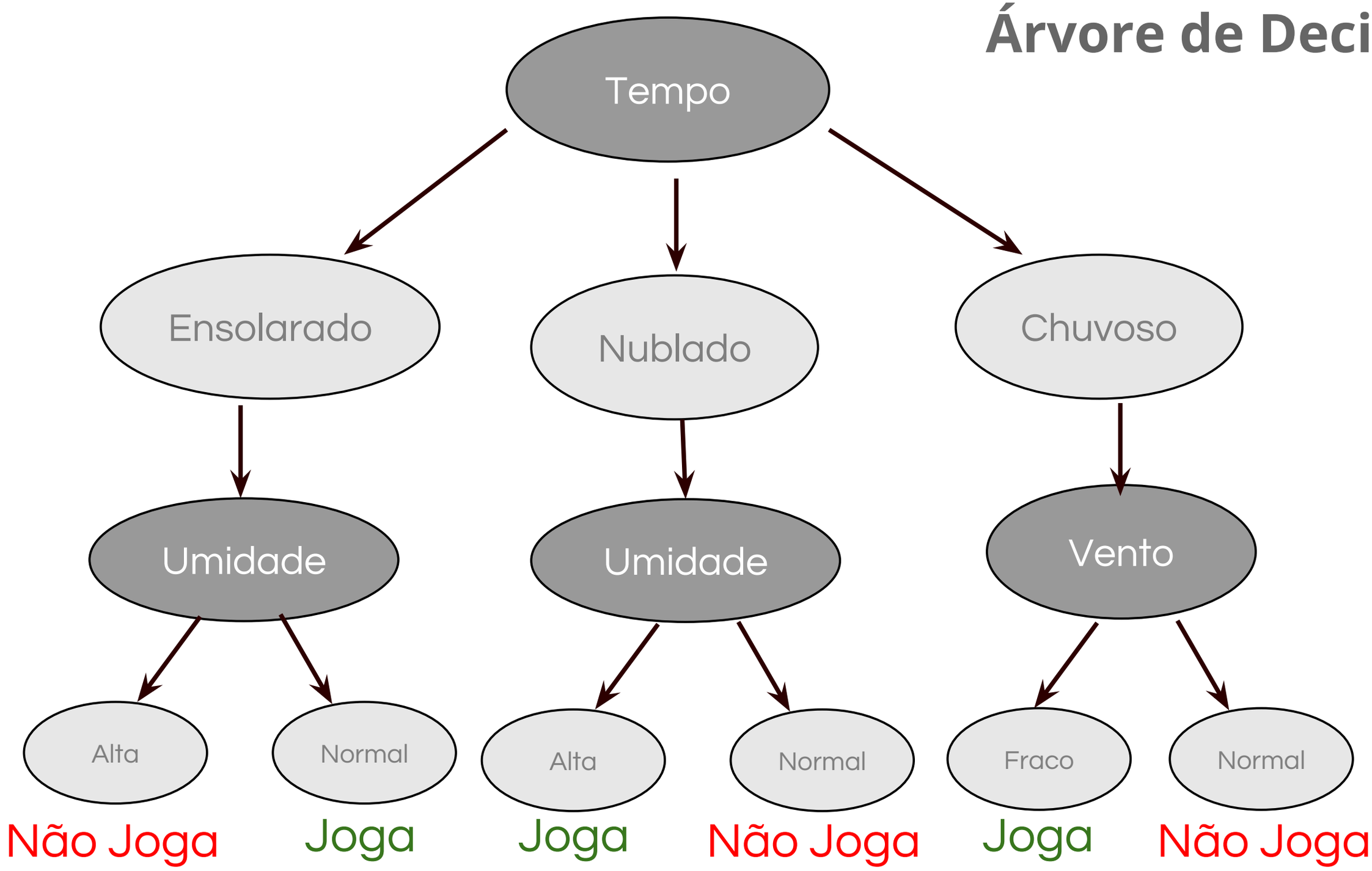
# Árvore de Decisão



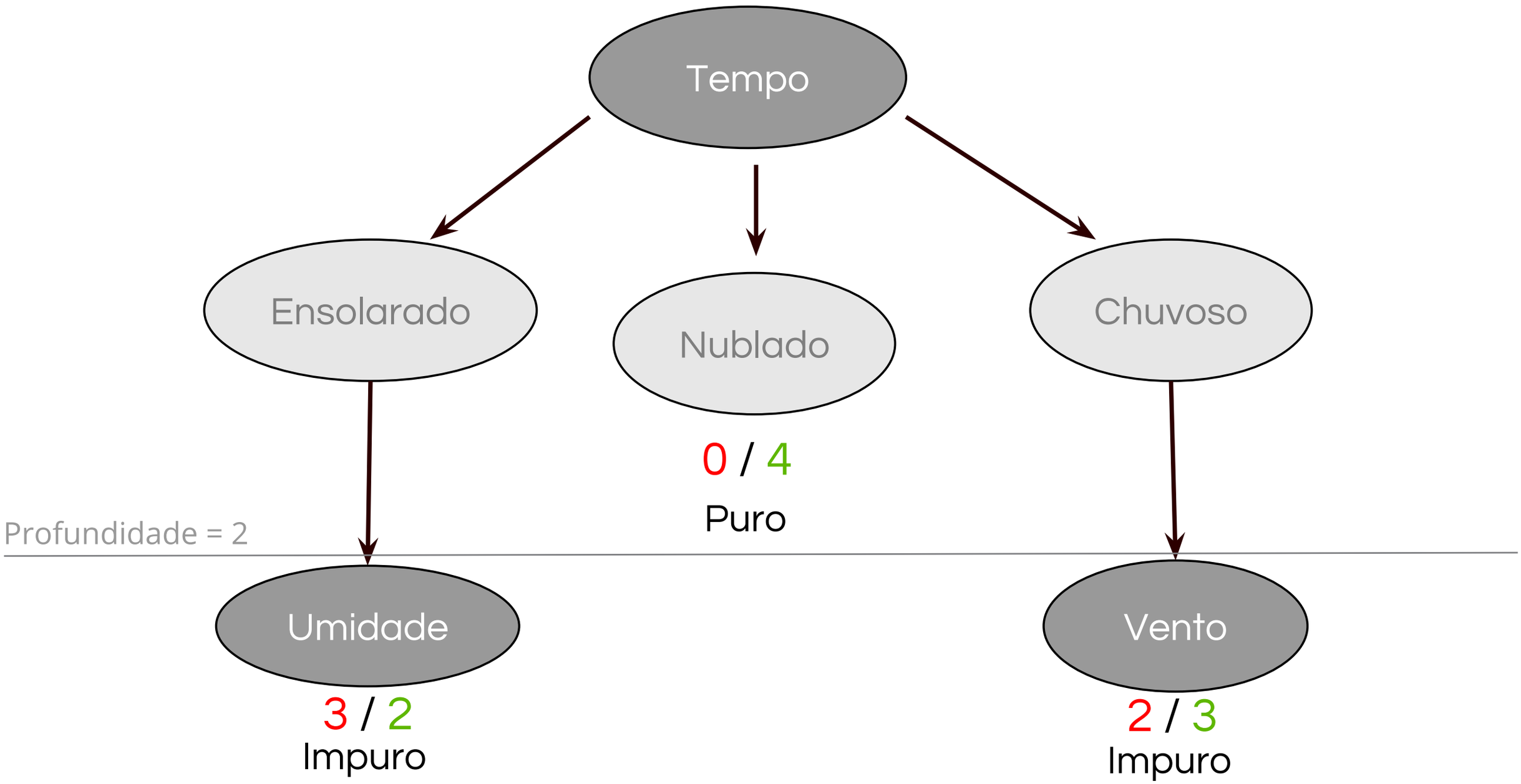
| # Dia | Tempo | Umidade | Vento | Joga? |
|---|---|---|---|---|
| 1 | Nublado | Alta | Fraco | **Sim** |
| 2 | Ensolarado | Alta | Forte | **Não** |
| 3 | Ensolarado | Alta | Fraco | **Não** |
| 4 | Chuvoso | Normal | Fraco | **Sim** |
| 5 | Chuvoso | Alta | Fraco | **Sim** |
| 6 | Chuvoso | Normal | Forte | **Não** |
| 7 | Nublado | Alta | Forte | **Sim** |
| 8 | Chuvoso | Alta | Forte | **Não** |
| 9 | Ensolarado | Normal | Fraco | **Sim** |
| 10 | Chuvoso | Alta | Fraco | **Sim** |
| 11 | Nublado | Normal | Fraco | **Não** |
| 12 | Nublado | Alta | Forte | **Sim** |
| 13 | Nublado | Alta | Fraco | **Sim** |
| 14 | Ensolarado | Alta | Fraco | **Não** |

2 semanas de Julho

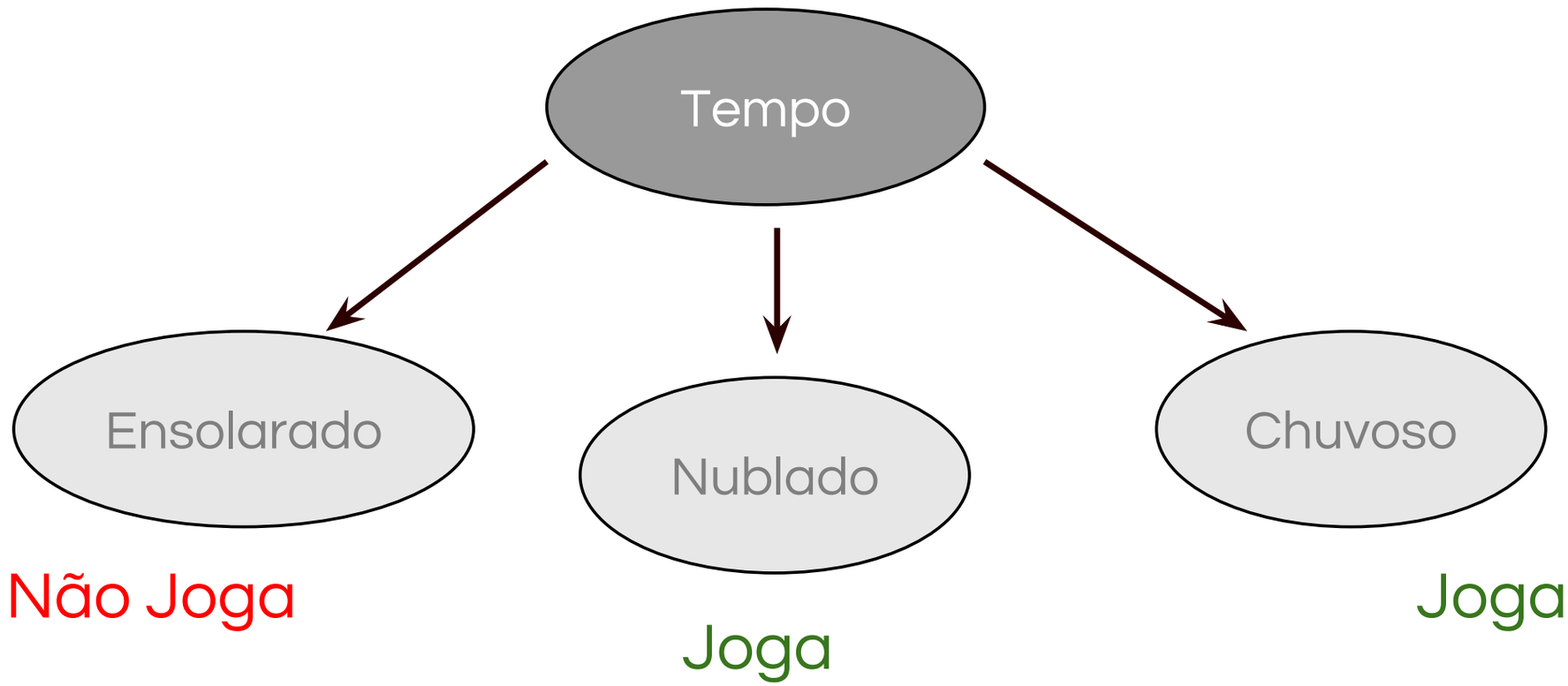
# Árvore de Decisão



# Sub-conjuntos (im)puros



# Sub-conjuntos (im)puros

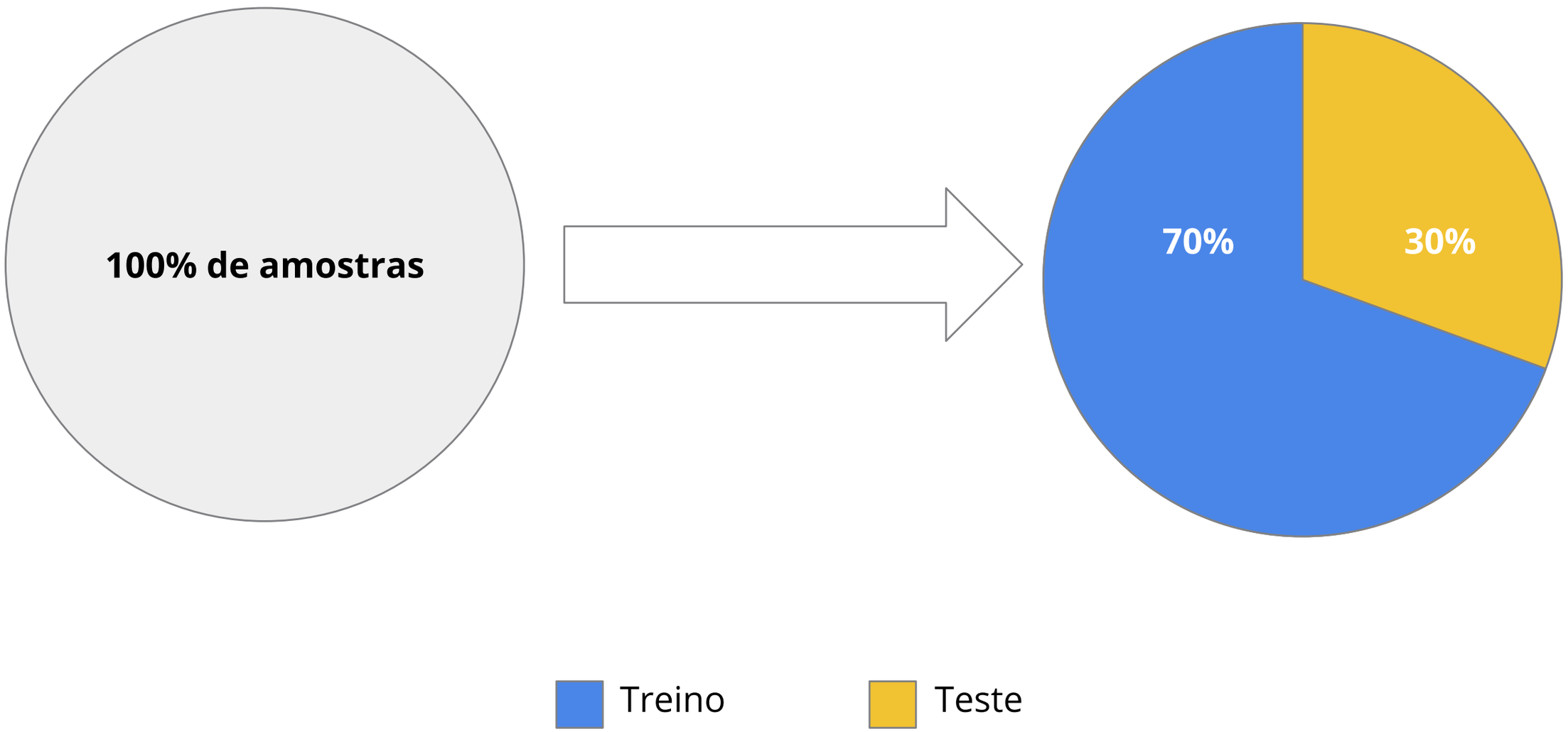



Profundidade = 2

# TREINAMENTO SUPERVISIONADO

Conjunto de amostras **rotulado**

# TREINAMENTO SUPERVISIONADO



# COMO COMPARAR DESEMPENHO?

Matriz de Confusão

| | | Palpite do Classificador | |
|---|---|---|---|
| | | Joga | Não Joga |
| Classificação Real | Joga | Verdadeiro Positivo (TP) | False Negativo (FN) |
| | Não Joga | False Positivo (FP) | Verdadeiro Negativo (TN) |

# COMO COMPARAR DESEMPENHO?

| | | Palpite do Classificador | |
|---|---|---|---|
| | | Joga | Não Joga |
| Classificação Real | Joga | Verdadeiro Positivo (TP) | False Negativo (FN) |
| | Não Joga | False Positivo (FP) | Verdadeiro Negativo (TN) |

$$\textbf{Precisão} = \frac{TP}{TP + FP}$$

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

$$\textbf{Acurácia} = \frac{TP + TN}{Total}$$

$$\textbf{F1 Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}}$$

# Código, mostre-me código!

# EVOLUÇÃO: MODELOS ENSEMBLE

*Em qual atleta eu posso apostar na final do US Open?*

# OPINIÃO 1

- Conhece bem o esporte
- ... mas tendencioso
- Não é imparcial

No passado, ele teve êxito em **68%** de seus palpites

# OPINIÃO 2

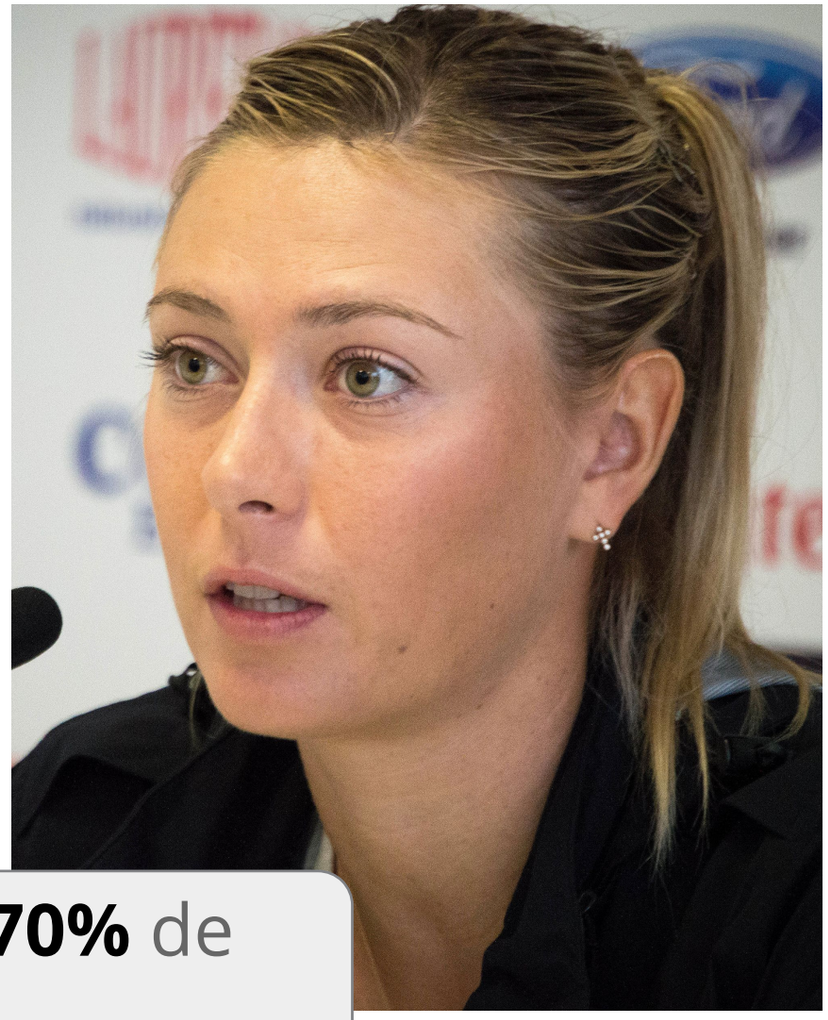
Blabous®

- Acompanha apostas há 10 anos
- Estuda atletas individualmente
- Excelência em apostas esportivas

No passado, ela teve êxito em **73%** de seus palpites

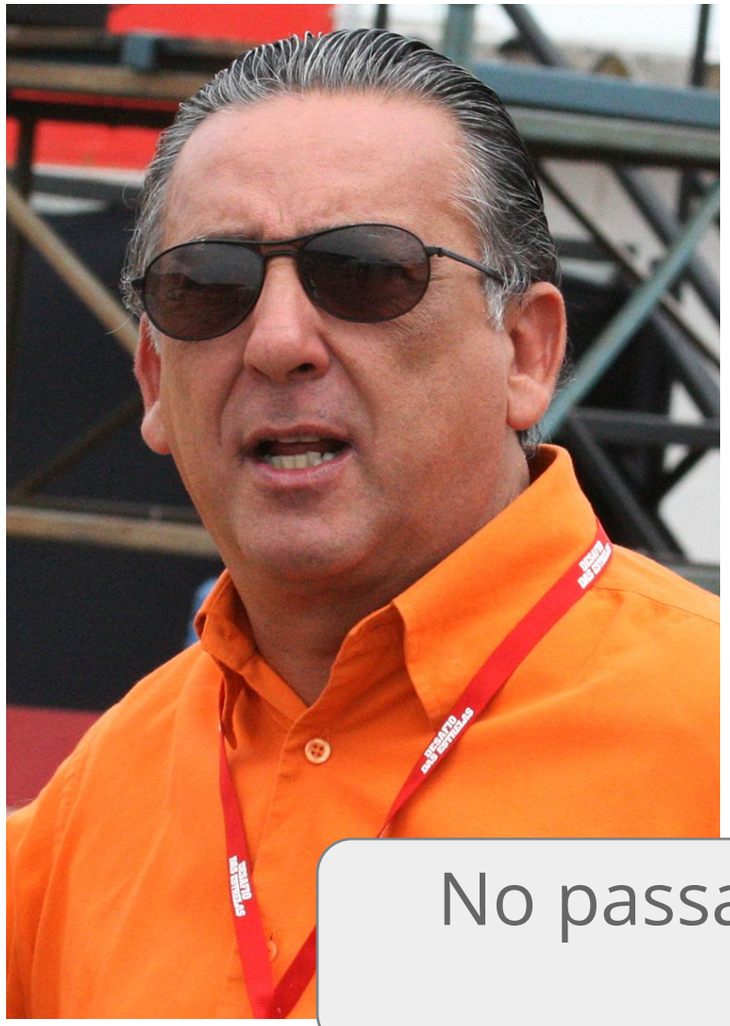
# OPINIÃO 3

- Mais afinidade com atletas mulheres
- Desconhece atletas novos
- Não vem acompanhando o esporte



No passado, ela teve êxito em **70%** de seus palpites

# OPINIÃO 4



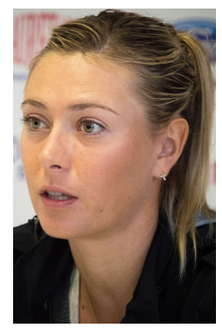
...

No passado, ela teve êxito em **62%** de seus palpites

# COMBINAÇÃO DOS PALPITES

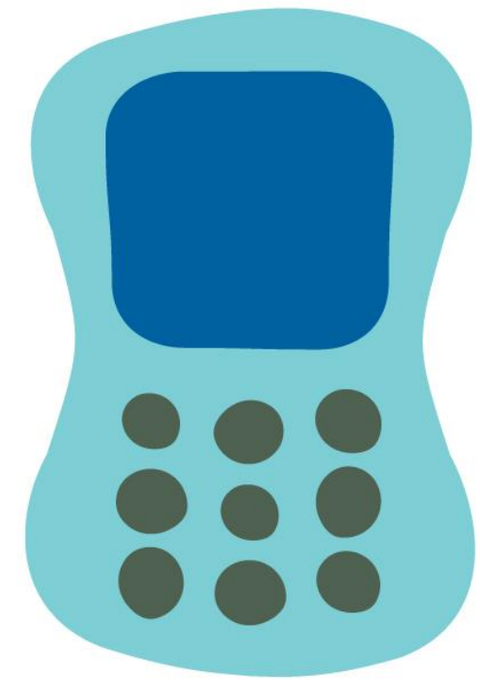
Guga

Empresa



Sharapova

Galvão



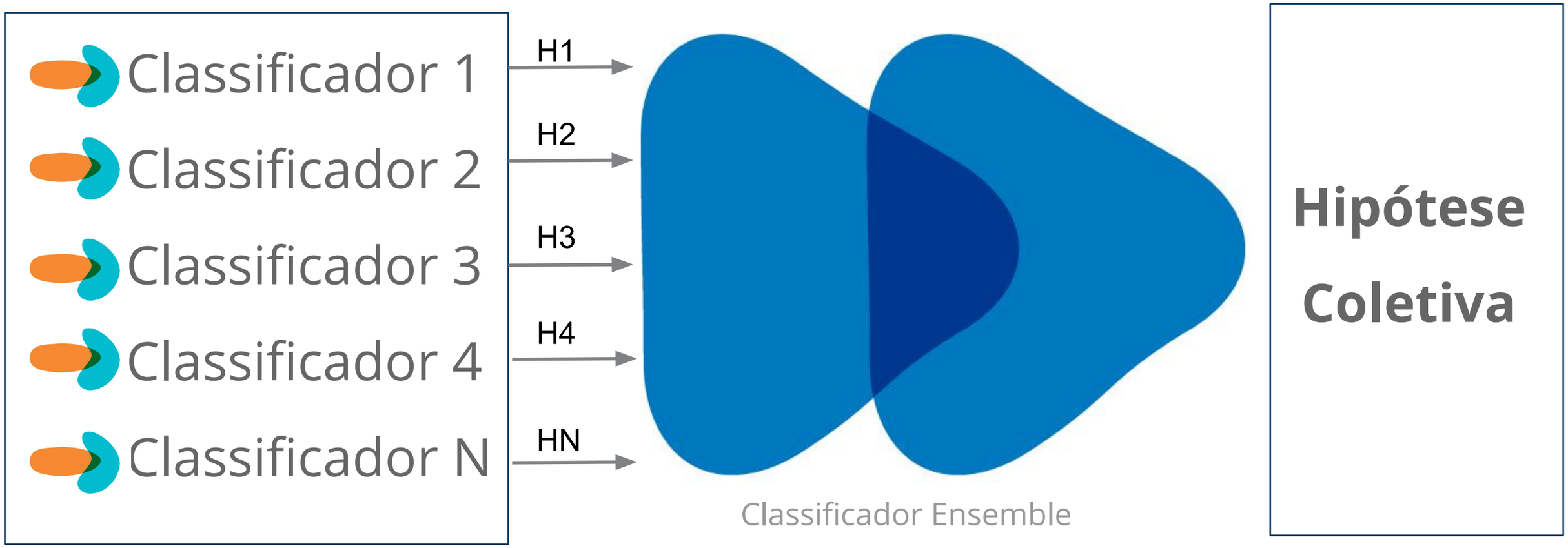
= 1 - (68% x 73% x 70% x 62%)

**= 79% de acurácia**

Correlacionados

Acurácia

# MODELO ENSEMBLE



# MODELO ENSEMBLE

Guga

Galvão

**Blabous**

Sharapova

Classificador Ensemble

**Hipótese Coletiva**

# CLASSIFICADORES

Algoritmos da mesma natureza

- Árvore de Decisão 1
- Árvore de Decisão 2
- Árvore de Decisão 3
- Árvore de Decisão 4
- Árvore de Decisão N

Diversidade de Algoritmos

- **SVM**
- Árvore de Decisão
- **Rede Neural**
- Deep Learner
- Regressor Linear

# CLASSIFICADORES

Hipótese Coletiva

Negativo Positivo

| | |
|---|---|
| Árvore de Decisão 1 | +1 |
| Árvore de Decisão 2 | -1 |
| Árvore de Decisão 3 | +1 |
| Árvore de Decisão 4 | -1 |
| Árvore de Decisão N | -1 |

| | |
|---|---|
| **SVM** | +2 |
| Árvore de Decisão | -1 |
| **Rede Neural** | +3 |
| Deep Learner | -1 |
| Regressor Linear | -1 |

# CLASSIFICADORES

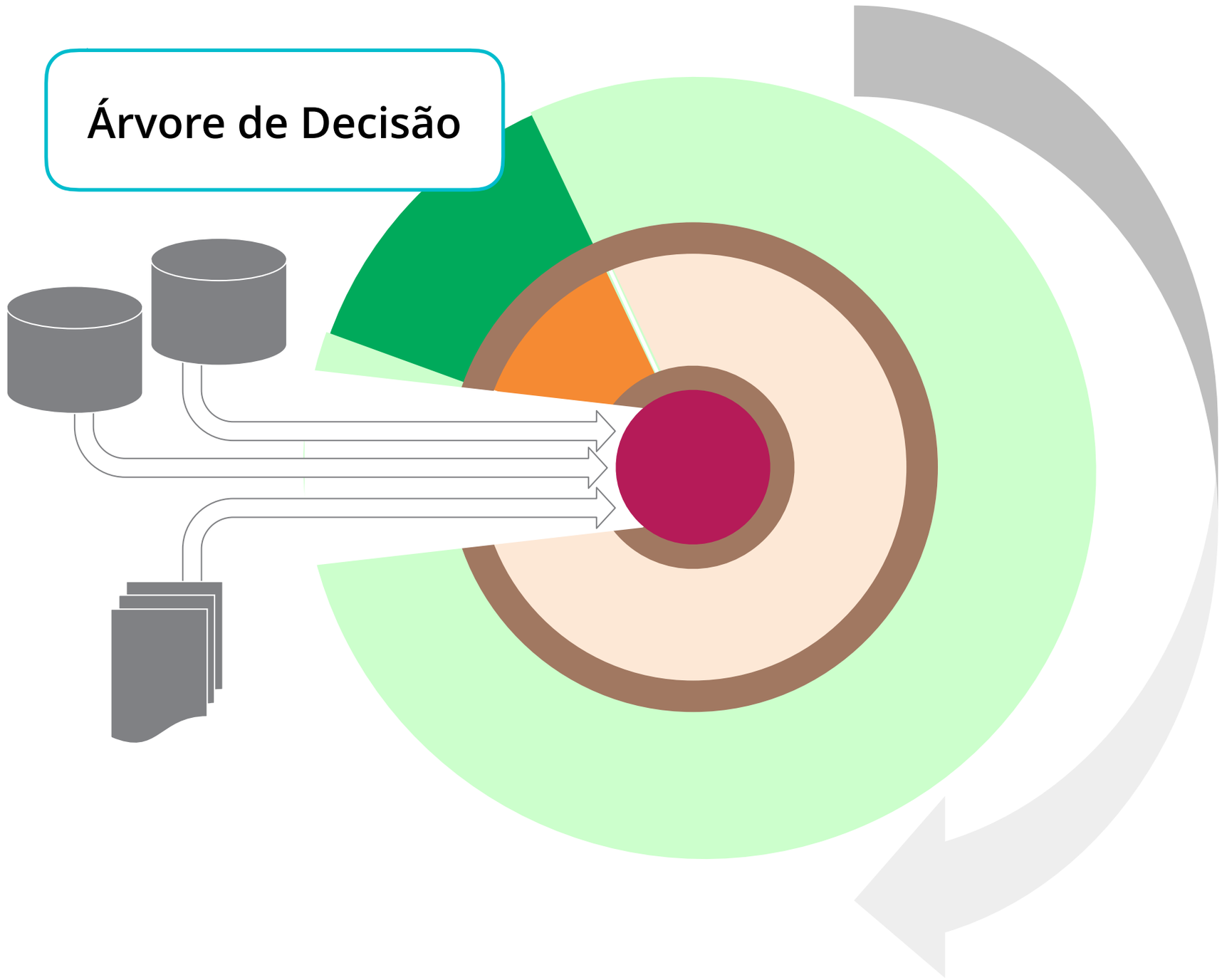
| | |
|---|---|
| Árvore de Decisão 1 | +1 |
| Árvore de Decisão 2 | -1 |
| Árvore de Decisão 3 | +1 |
| Árvore de Decisão 4 | -1 |
| Árvore de Decisão N | -1 |

**Hipótese Coletiva: -1**

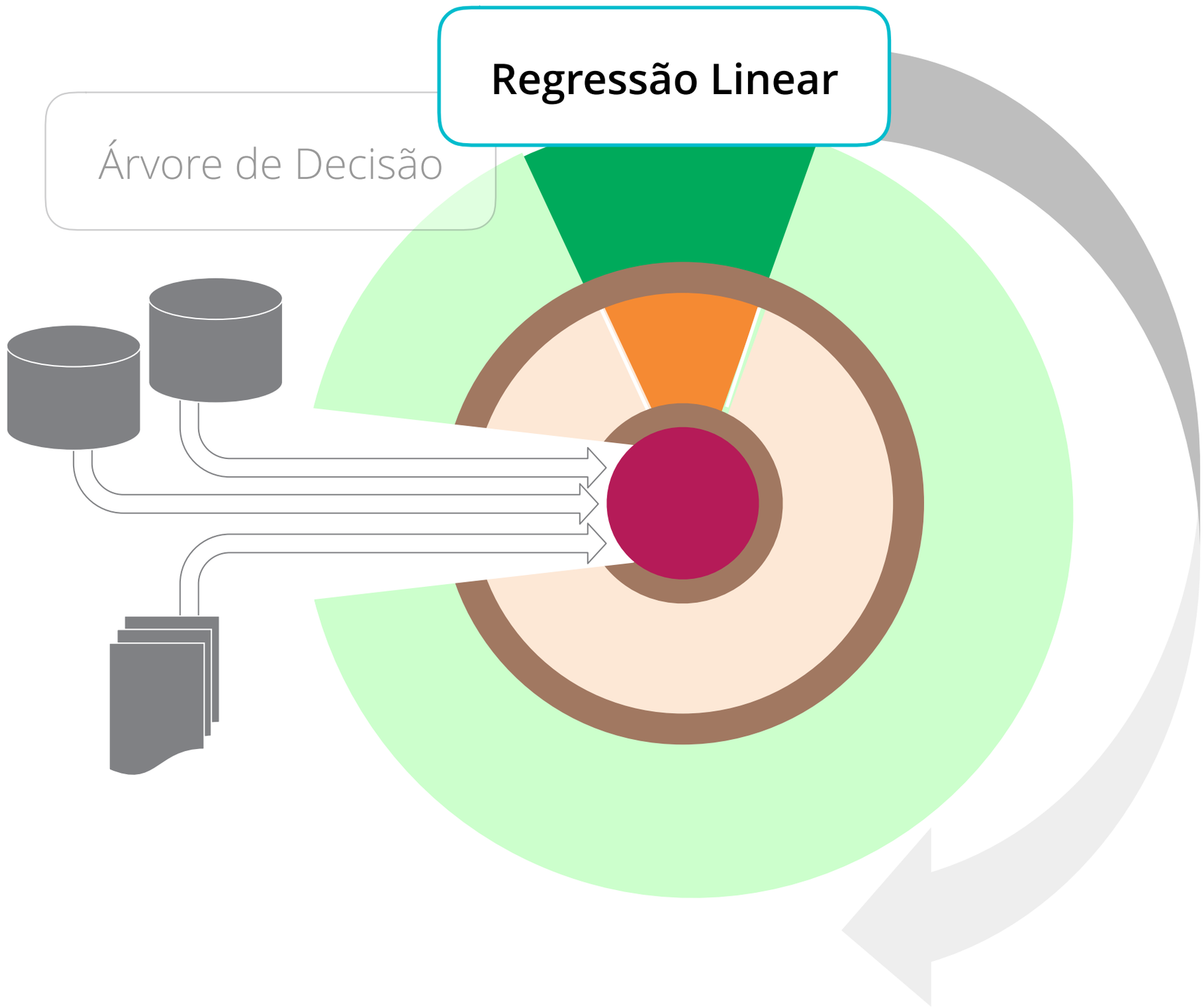
| | |
|---|---|
| **SVM** | +2 |
| Árvore de Decisão | -1 |
| **Rede Neural** | +3 |
| Deep Learner | -1 |
| Regressor Linear | -1 |

**Hipótese Coletiva: +2**

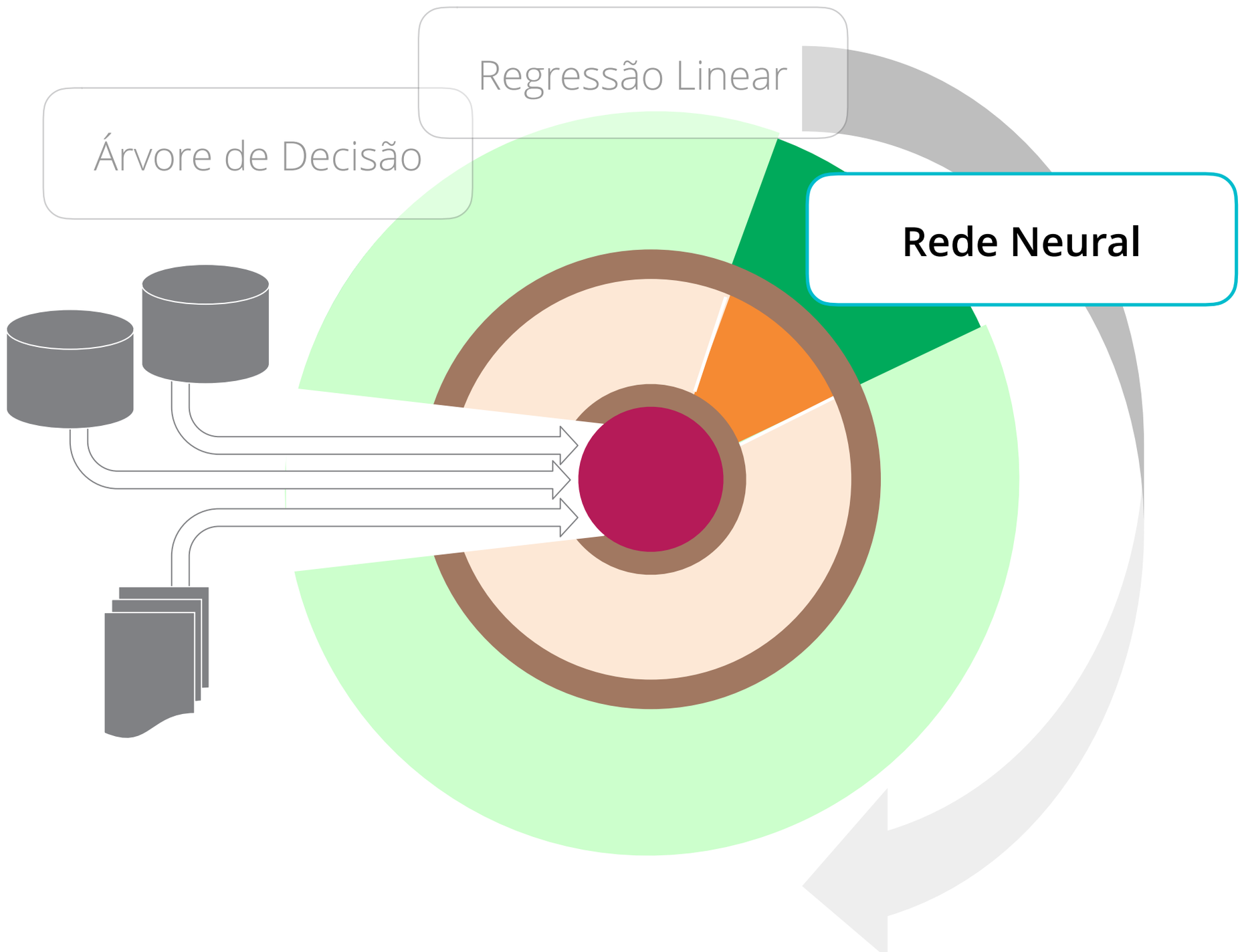
# AGILE ANALYTICS



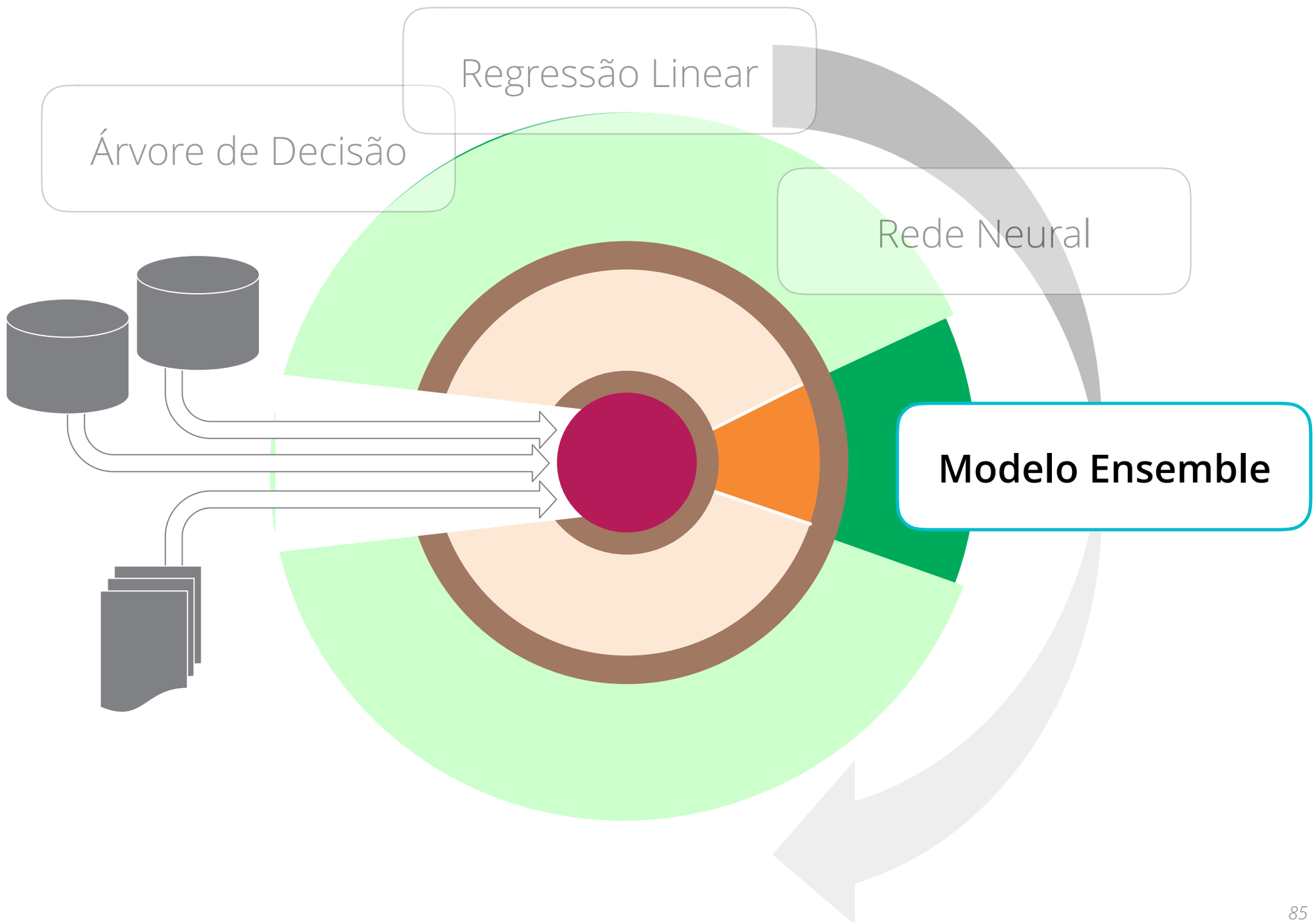
# AGILE ANALYTICS



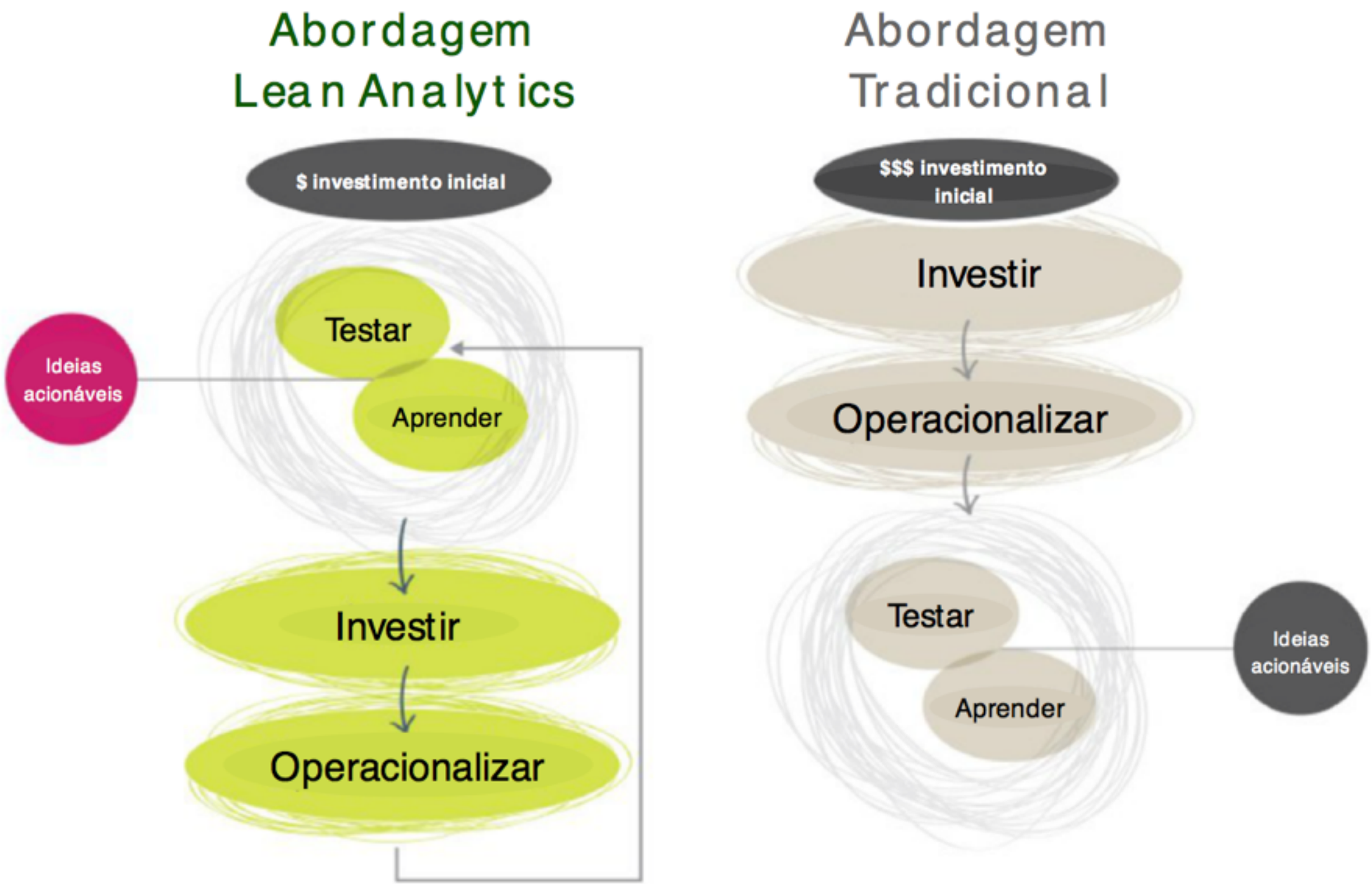
# AGILE ANALYTICS



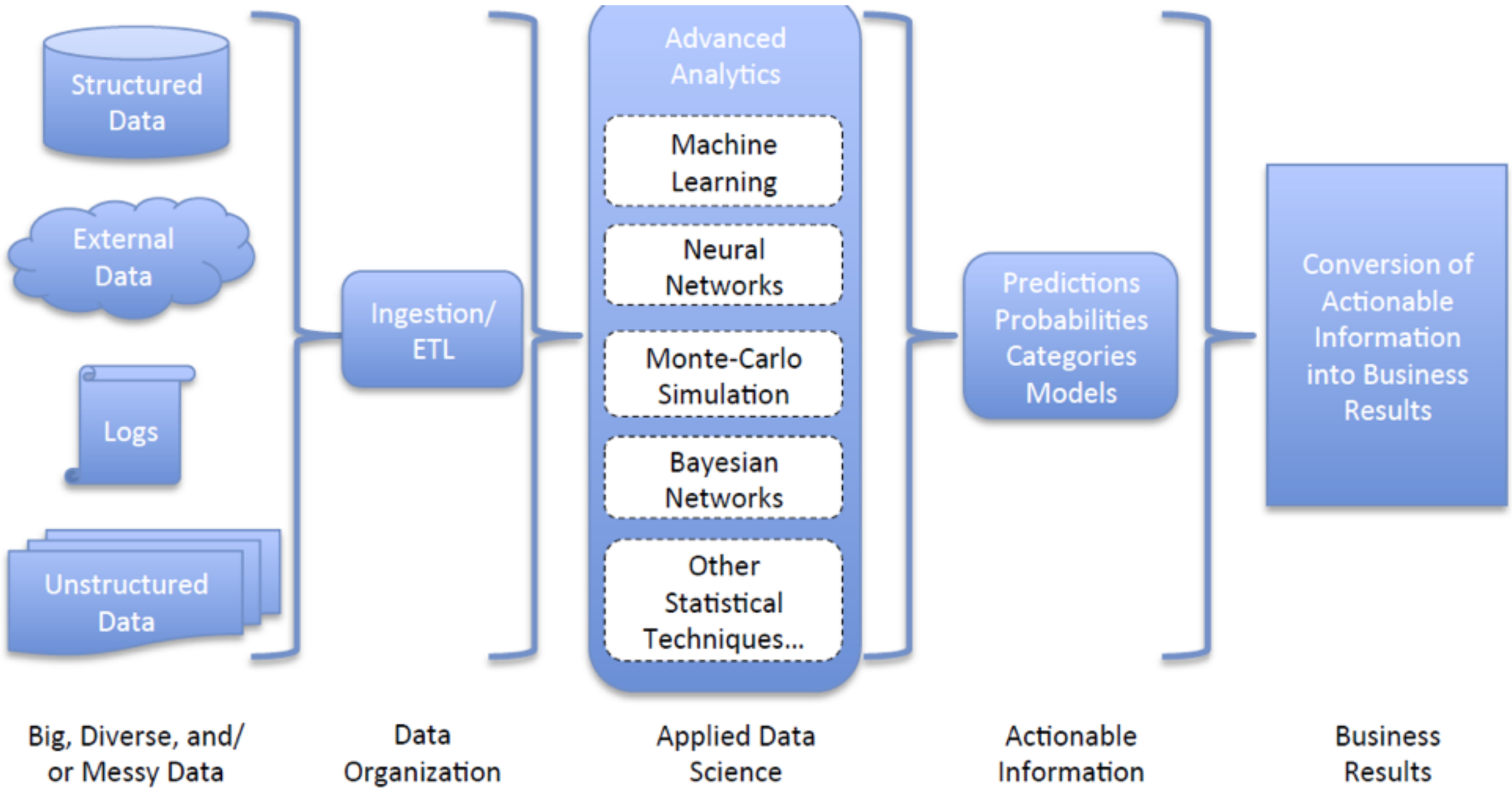
# AGILE ANALYTICS



# ABORDAGEM


Abordagem
Lean Analytics
$ investimento inicial
Ideias acionáveis
Testar
Aprender
Investir
Operacionalizar
Abordagem
Tradicional
$$$ investimento inicial
Investir
Operacionalizar
Testar
Aprender
Ideias acionáveis

# ABORDAGEM


Structured Data
External Data
Logs
Unstructured Data
Ingestion/ ETL
Advanced Analytics
Machine Learning
Neural Networks
Monte-Carlo Simulation
Bayesian Networks
Other Statistical Techniques...
Predictions Probabilities Categories Models
Conversion of Actionable Information into Business Results
Big, Diverse, and/ or Messy Data
Data Organization
Applied Data Science
Actionable Information
Business Results

# METODOLOGIAS

# CANVAS

| Personas & Plataformas | Visão do MVP | Resultado esperado |
|---|---|---|
| Para quem é este MVP?<br>Em que plataforma estará disponível? | Qual é a visão para este MVP? | *Acreditamos que este MVP irá alcançar...* |
| | **Funcionalidades**<br>O que vamos construir neste MVP?<br>Que ações serão simplificados/melhoradas nesta MVP? | Que aprendizado ou resultado estamos buscando neste MVP? |
| **Jornadas**<br>Quais jornadas são atendidas ou melhoradas com este MVP? | | **Métricas para validar as hipóteses de negócio**<br>*Acreditamos que este MVP irá alcançar (resultado esperado). Saberemos que isso aconteceu baseado em...*<br>Como podemos medir os resultados deste MVP? |
| | **Custo & Cronograma**<br>*Qual custo e a data prevista para entrega do MVP?* | |

# ESTÓRIAS DE USUÁRIOS

Como um **<papel>**

Eu quero **<objetivo>**

Para então **<valor>**

# PAREAMENTO

# TIMES AUTO-GERENCIÁVEIS

# OUTRAS PRÁTICAS

- Stand-ups
- Showcases
- Atividades de *Team Building*
- Retrospectivas Semanais

# RADAR TÉCNICO

## Techniques

**hold**
1. Hour/Point estimation
2. Normalized MongoDB schemas
3. Raw callbacks

**assess**
4. API UX
5. Promises (ES6)
6. Executable Documentation
7. Hypermedia (REST)

**trial**
8. Trunk-based development
9. Continuous Delivery
10. BFF
11. Specification by Example
12. No estimation
13. Demos as documentation

**adopt**
14. Node.js

## Tools

**assess**
15. Doctest.js
16. Mesosphere + Marathon
17. PostgreSQL 9.x

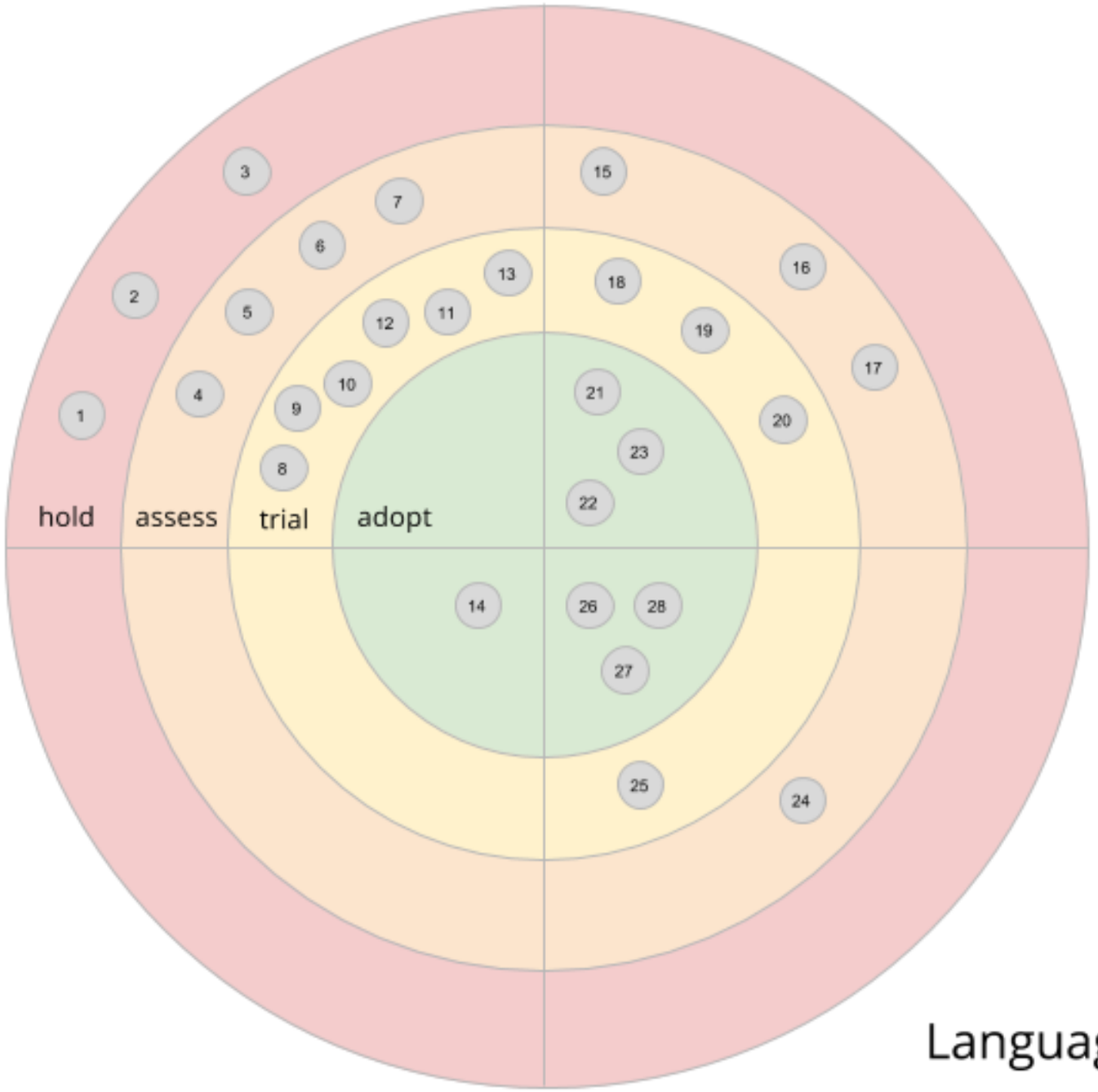
**trial**
18. Readme.io
19. ELK
20. EcmaScript 6

**adopt**
21. MongoDB
22. Docker
23. ElasticSearch

**assess**
24. Nock

**trial**
25. SinonJS

**adopt**
26. Mocha
27. Chai
28. SuperTest




## Platforms

## Languages & Frameworks

# *Andressa Sivolella*

*@asivolella*

*asivolella@thoughtworks.com*

ThoughtWorks®