# Data Quality

## Banco de Dados

Aluno:
**Ivan Valentim Santos**

# Roteiro

- Conceito
- Coleta dos dados
- Estágios de sofisticação
- Critérios e suas Categorias
- Problemas dos dados “sujos”
- Arquitetura de processos
- Conclusão


DATA
EXAMPLE
ACTION
RISK
DATA
associated
REPORTING
example
REQUIREMENTS
ENSURE
data
data
quality
relevant
EXTERNAL
governance
associated
ensure
training
commitment
processes
BODY
processes
RISK
clearly
risk
MANAGEMENT
business
REPORTING
commitment
AUDIT
USED
POLICIES
responsability
training
PLACE
level
data
quality
body
information
used
TOOLS
DATA QUALITY
charged
reviewed
security
CLEANSING
TIMELY
data
people
RELEVANT
relevant
level
COLLECTION
risk
action
MANAGEMENT
reporting
data
ensure

# Conceito

1960 1990

# Conceito

"O estado de completude, validade, consistência, atualidade e precisão que torna um dado apropriado para um uso específico."

*(Government of British Columbia)*

# Qualidade de dados

- Conceito multidimensional;
- Composição dos dados;
- Planejamento estratégico;
- Manipulação de grandes volumes de dados;
- Atualização frequente da informação;
- Diversas fontes.

# Sem Qualidade dos dados

A má qualidade de dados significa a possibilidade da existência de informação imprecisa, incompleta, redundante e até mesmo fictícia.

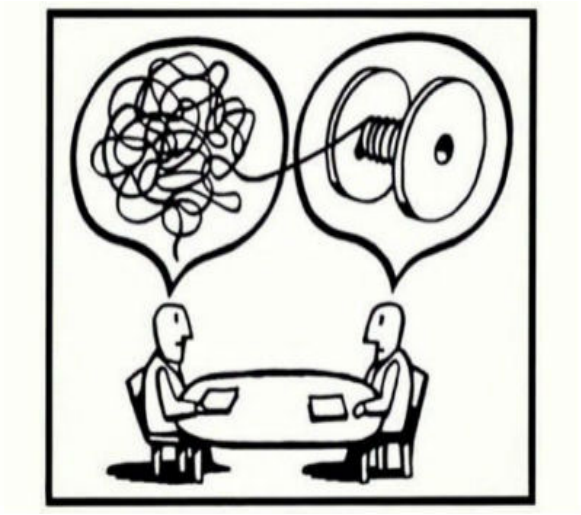
*(Drescher, 2004)*

# Sem Qualidade dos dados

- Diminuição da confiança do cliente;
- Perda de oportunidade de negócio;
- Tomadas de decisão equivocadas ocasionadas pela imprecisão e a falta de completude dos dados;
- Risco de Imagem.

# Proveniência dos dados

"É a descrição das origens de um dado e do processo pelo qual ele foi produzido, auxiliando na avaliação da qualidade, da validade e de quão recente é a informação."



(BUNEMAN *et al.*, 2001)

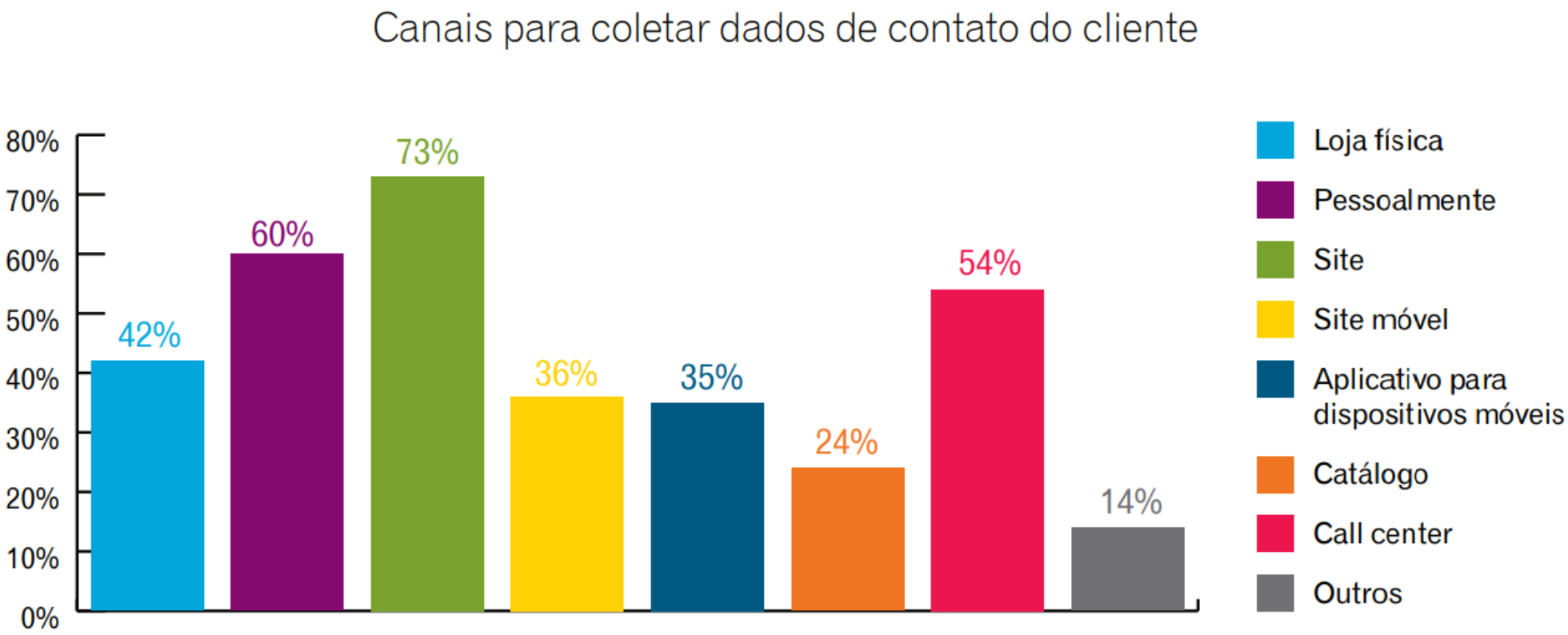
# Proveniência dos dados

- Qual a origem dos dados?
  - Cópias de cópias?
  - Edições de conteúdo?
- O quão confiáveis e atuais eles são?
  - Fontes não confiáveis?
  - Dados desatualizados?

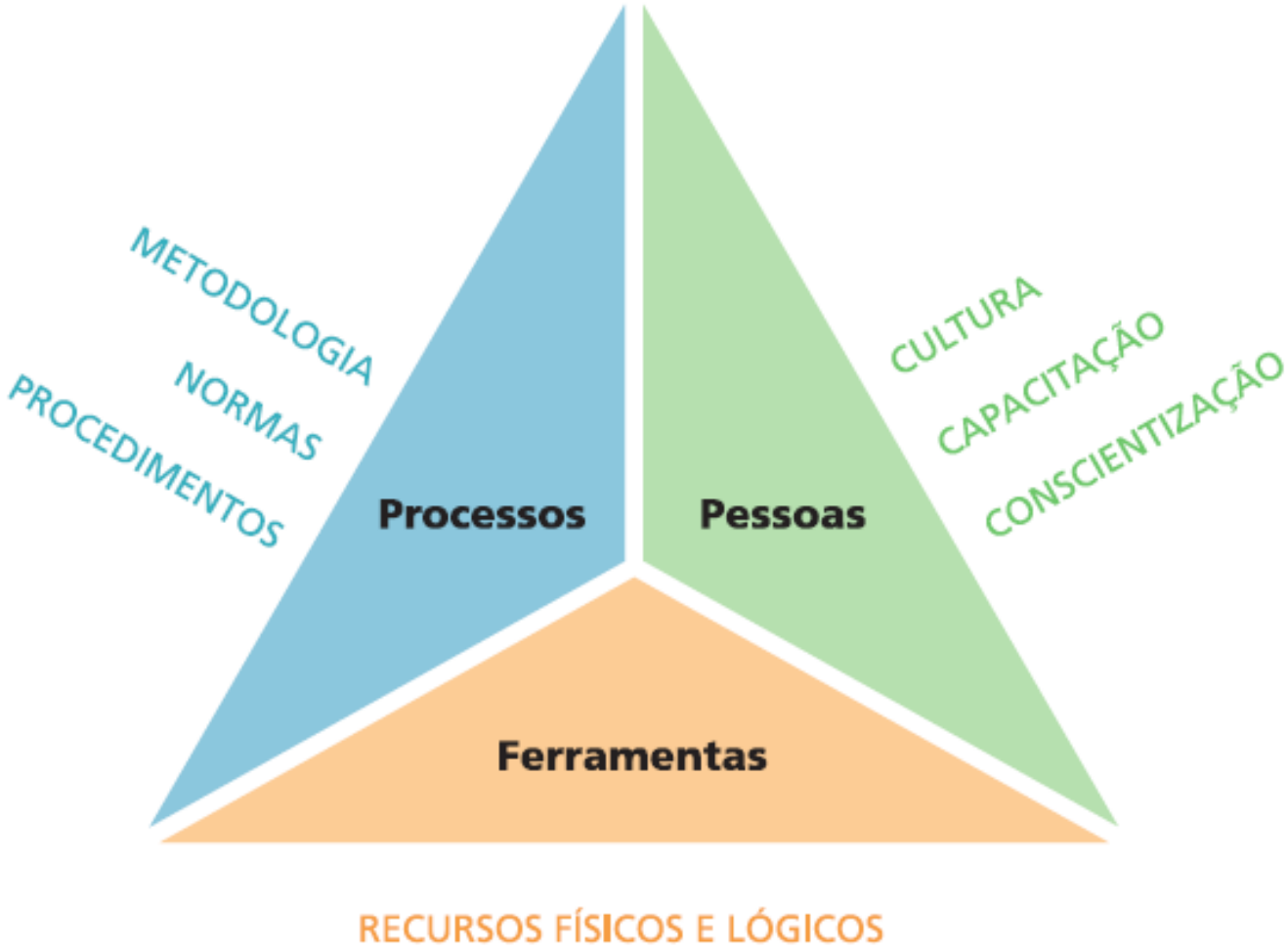
# Proveniência dos dados

- A proveniência permite avaliar a qualidade dos dados para uma aplicação;
- Erros introduzidos por defeitos nos dados tendem a inflar quando propagados;
- O nível de detalhamento da proveniência determina que grau a qualidade dos dados pode ser estimada;
- Com um certificado do dado, é possível avaliá-lo baseado em métricas de qualidade.

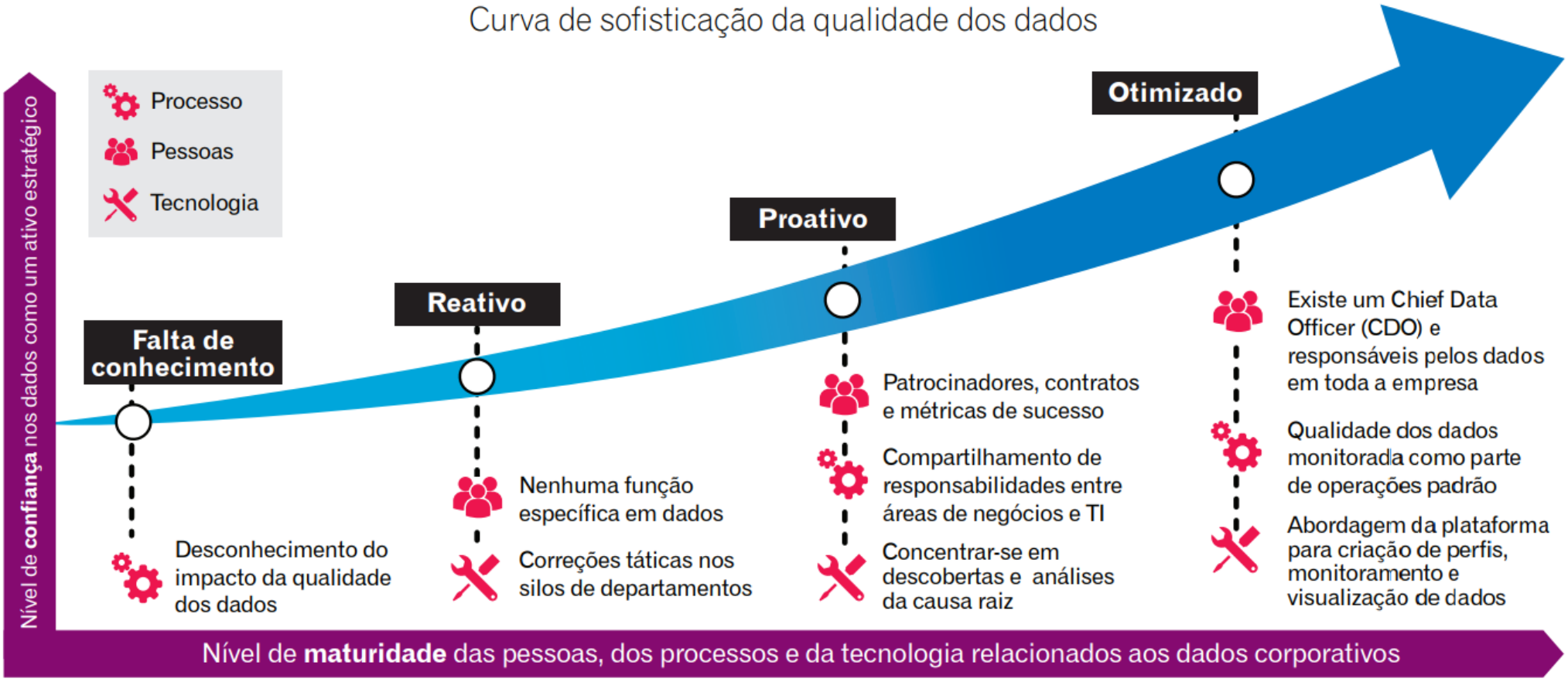
# Coleta dos dados



# Data quality tríade



# Estágios de sofisticação de Data Quality




*(Serasa Experian)*

# Dados com qualidade

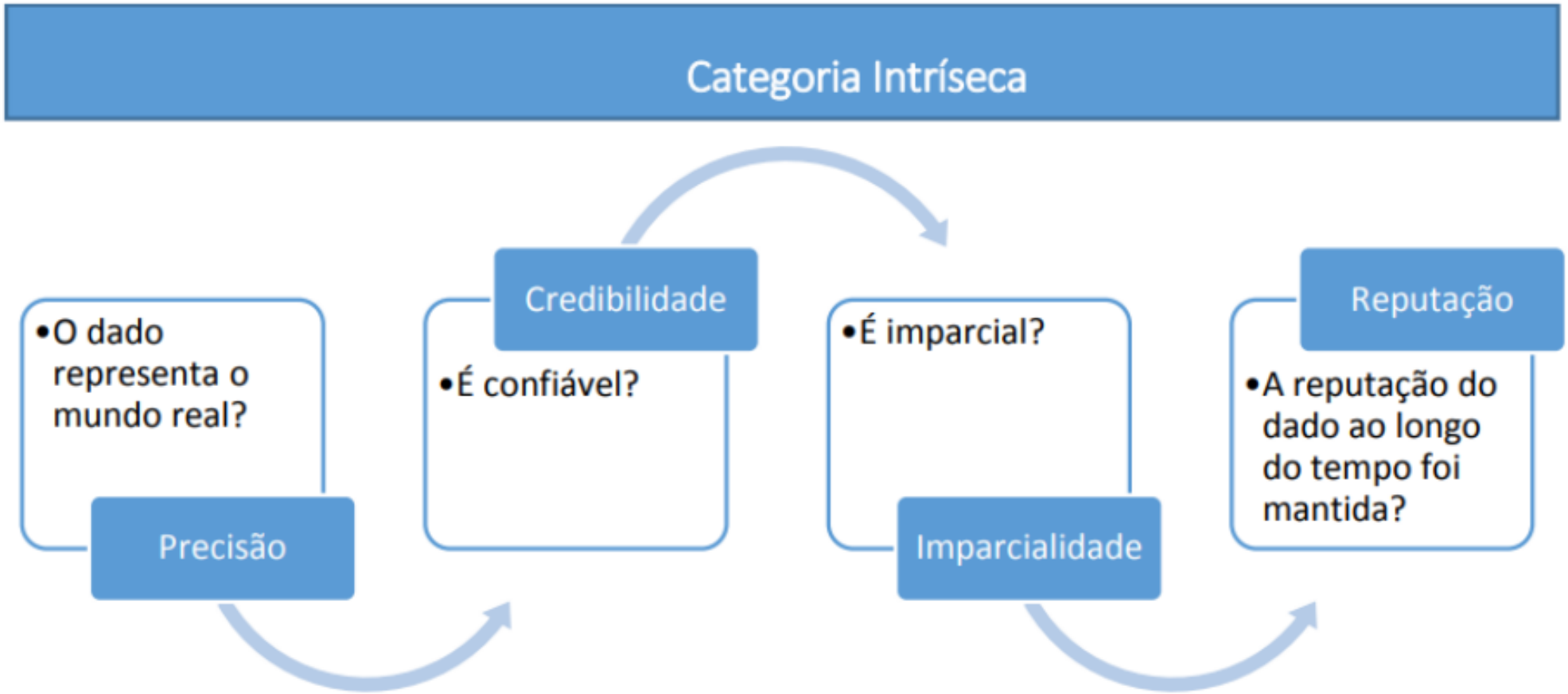
- Precisão;
- Veracidade;
- Relevância;
- Pontualidade.

(Total Data Quality Management - MIT)

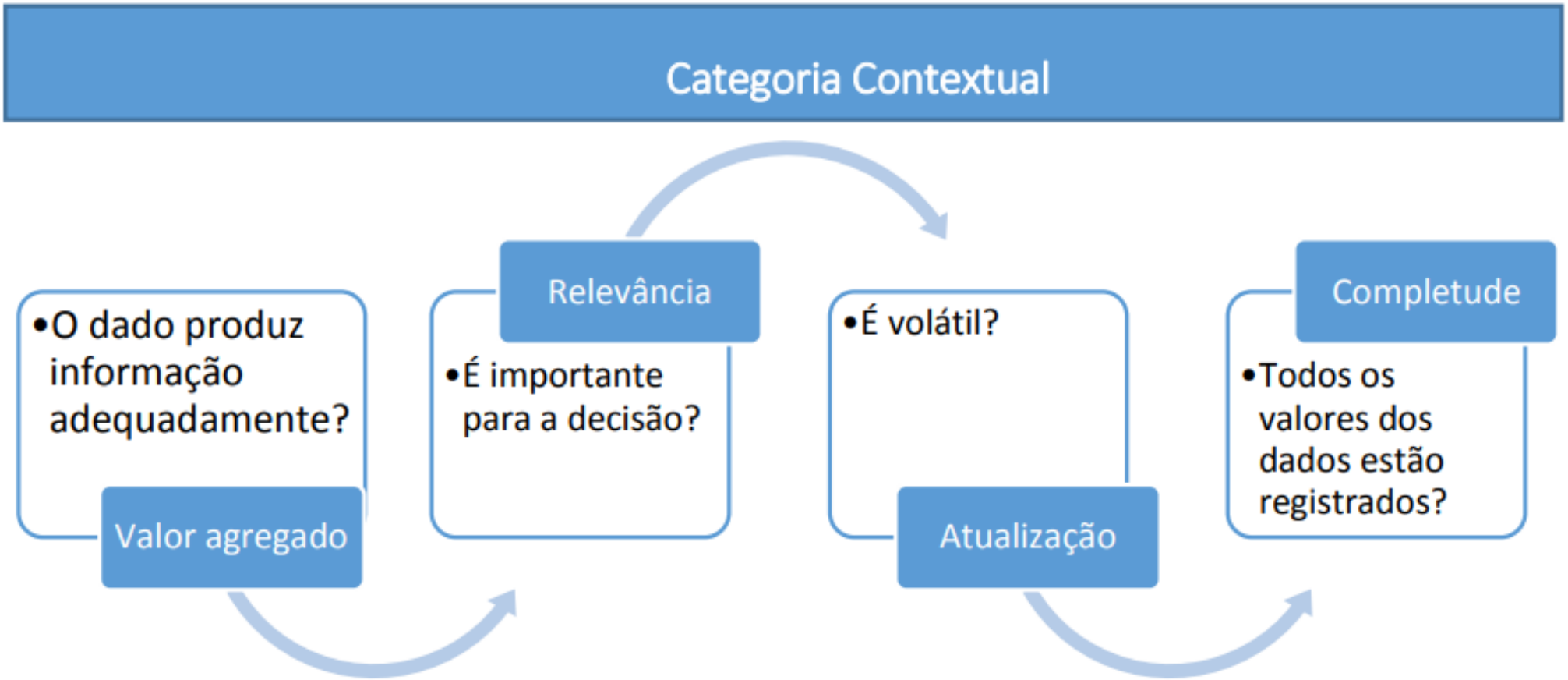
# Categorias da verificação de qualidade

- Intrínseca
- Contextual
- Representacional
- Acessibilidade

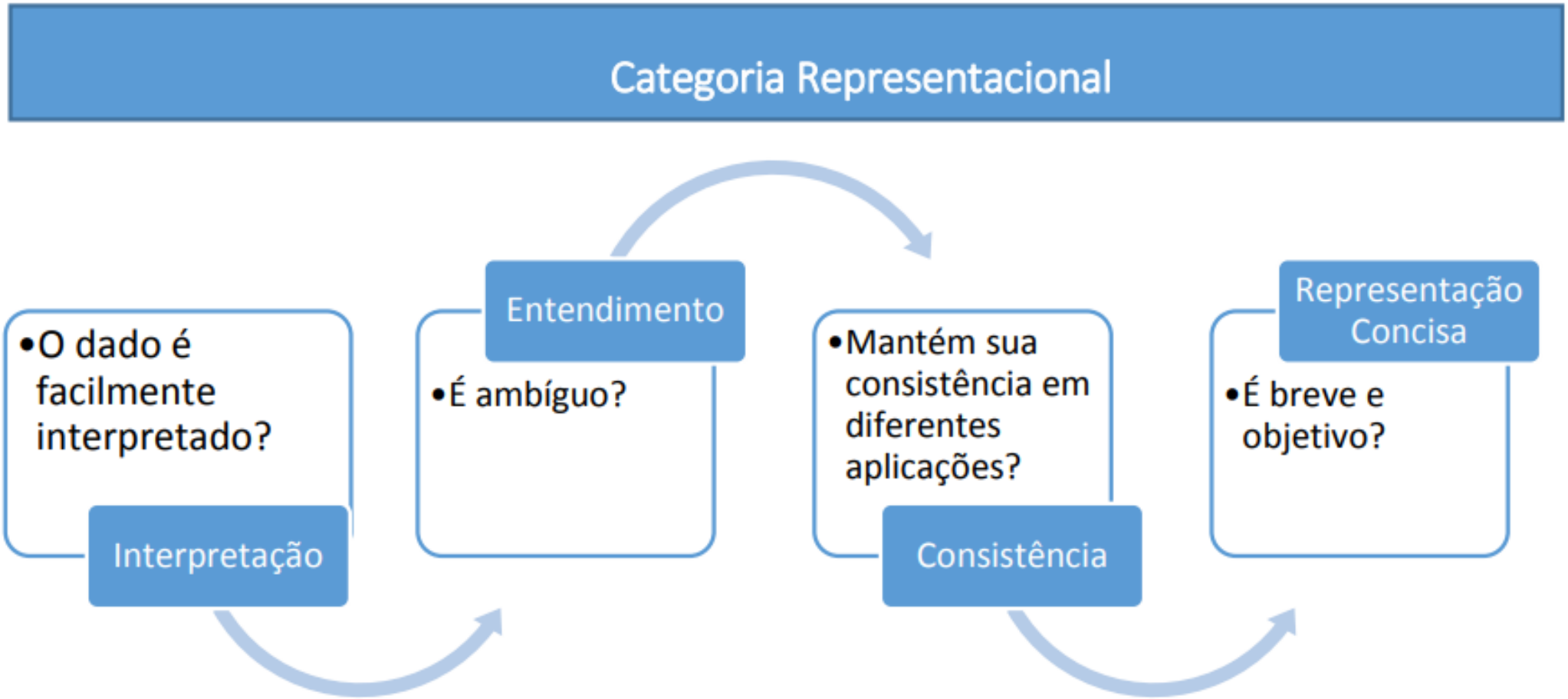
# Intrínseca


Categoria Intríseca
•O dado representa o mundo real?
Precisão
Credibilidade
•É confiável?
•É imparcial?
Imparcialidade
Reputação
•A reputação do dado ao longo do tempo foi mantida?

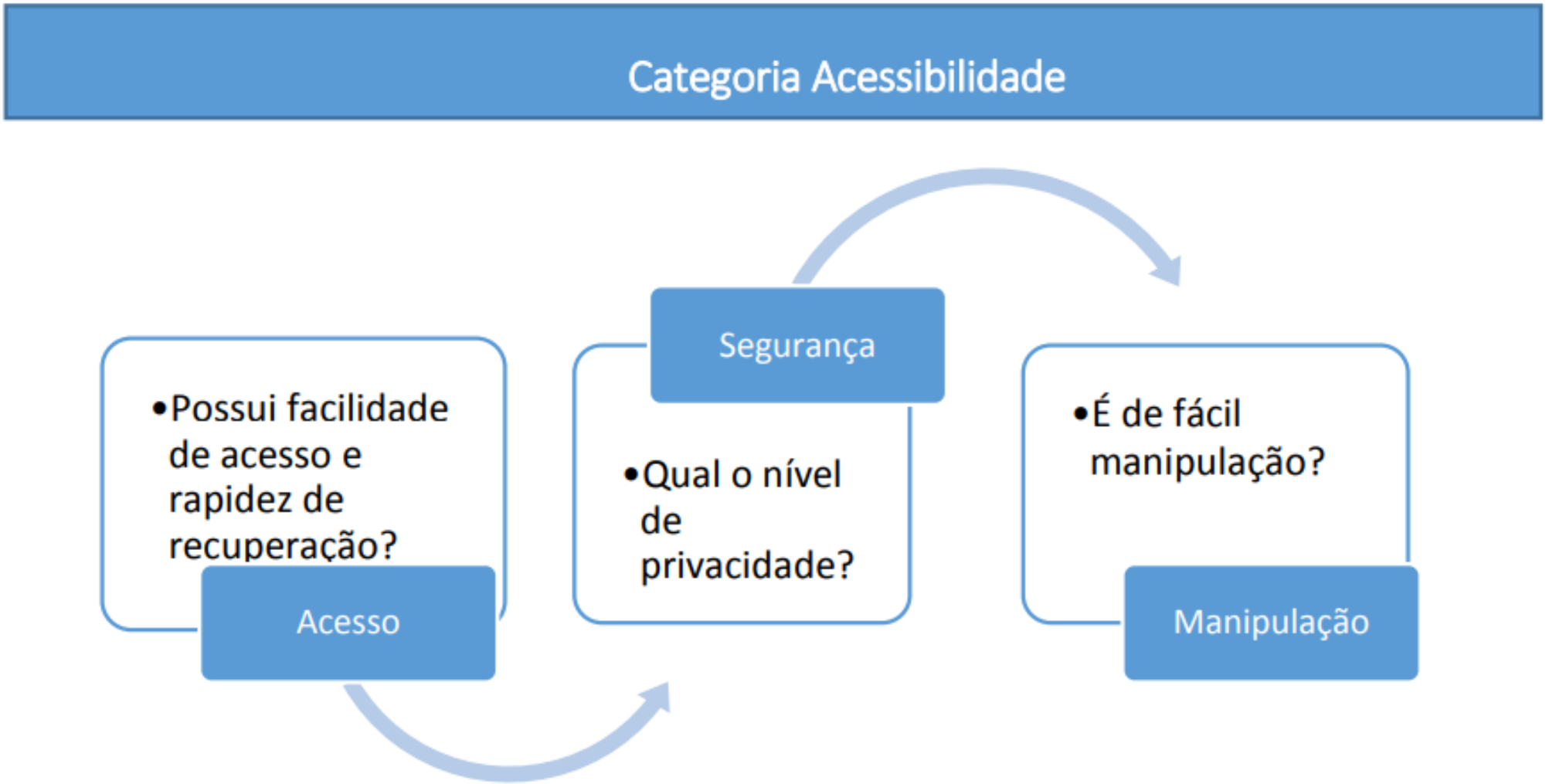
# Contextual


Categoria Contextual
•O dado produz informação adequadamente?
Valor agregado
Relevância
•É importante para a decisão?
•É volátil?
Atualização
Completude
•Todos os valores dos dados estão registrados?

# Representacional



# Acessibilidade



# Categorias e critérios

| Qualidade de dados | | | |
|---|---|---|---|
| Categoria Intrínseca | Categoria Contextual | Categoria Representacional | Categoria de Acessibilidade |
| • Precisão<br>• Credibilidade<br>• Imparcialidade<br>• Reputação | • Valor agregado<br>• Relevância<br>• Atualização<br>• Completude<br>• Quantidade | • Interpretação<br>• Entendimento<br>• Consistência<br>• Representação concisa | • Acesso<br>• Segurança<br>• Manipulação |

# Origem dos problemas de qualidade de dados

- Os dados foram inseridos incorretamente
- Os dados foram gerados sem cuidados

Os dados não estão corretos

- O método de geração dos dados não foi rápido o suficiente para suprir as necessidades

Os dados não foram gerados no momento correto

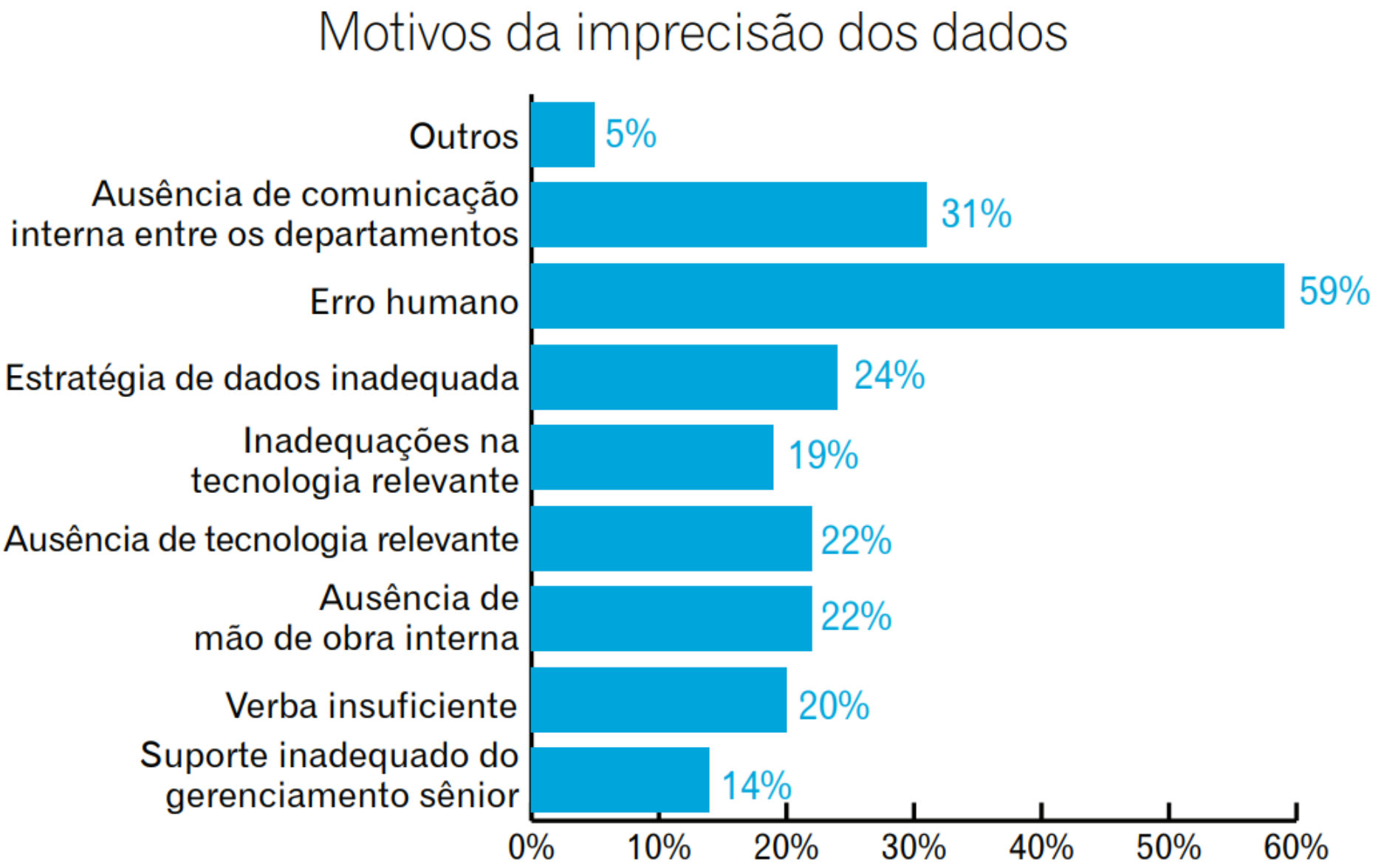
- O dados brutos foram coletados sob uma lógica ou periodicidade não adequada

Os dados não foram medidos ou classificados corretamente

- Não houve armazenamento
- Os dados nunca existiram

Os dados necessários não existem

# Imprecisão dos dados



# O que seriam dados “sujos”?

**Valores Default DUMMY**:

Quando encontramos Defaults para os valores de colunas ou campos obrigatórios.

*Exemplo:*

CPF com 999.999.999-9

# Dados “sujos”

**Valores Default “INTELIGENTES”**:

Quando os Defaults possuem significado.

*Exemplo:*

Se a coluna IDADE contiver 000 o cliente é corporativo!

# Dados “sujos”

**Valores contraditórios**: Quando os valores de uma coluna ou campo são inconsistentes com os valores de outra coluna ou campo relacionado.

*Exemplo:*

Na Tabela de **CLIENTES** determinada linha possui os seguintes valores:

Coluna **CEP**: 031085-020

Coluna **Endereço**: Rua Amazonas;

Todavia, na Tabela de **CEPs** este **CEP** não é da rua Amazonas!

# Dados “sujos”

**Valores em desacordo com o domínio**:

Quando os valores de uma coluna ou campo não obedecem ao domínio estabelecido.

*Exemplo*:

Na Tabela de **FUNCIONARIO** a coluna **SITUACAO-DO-FUNCIONARIO** deve conter os seguintes valores: ‘**ATIVO**’, ‘**INATIVO**’,’**DEMITIDO**’. Todavia, encontramos em uma linha da Tabela que possui uma situação igual a ‘**AFASTADO’**!

# Outros tipos de problemas

- **Dados incompletos**:

  Ex: *nome* = “ ”.

- **Dados ruidosos**:

  Ex: *salário* = “-54”.

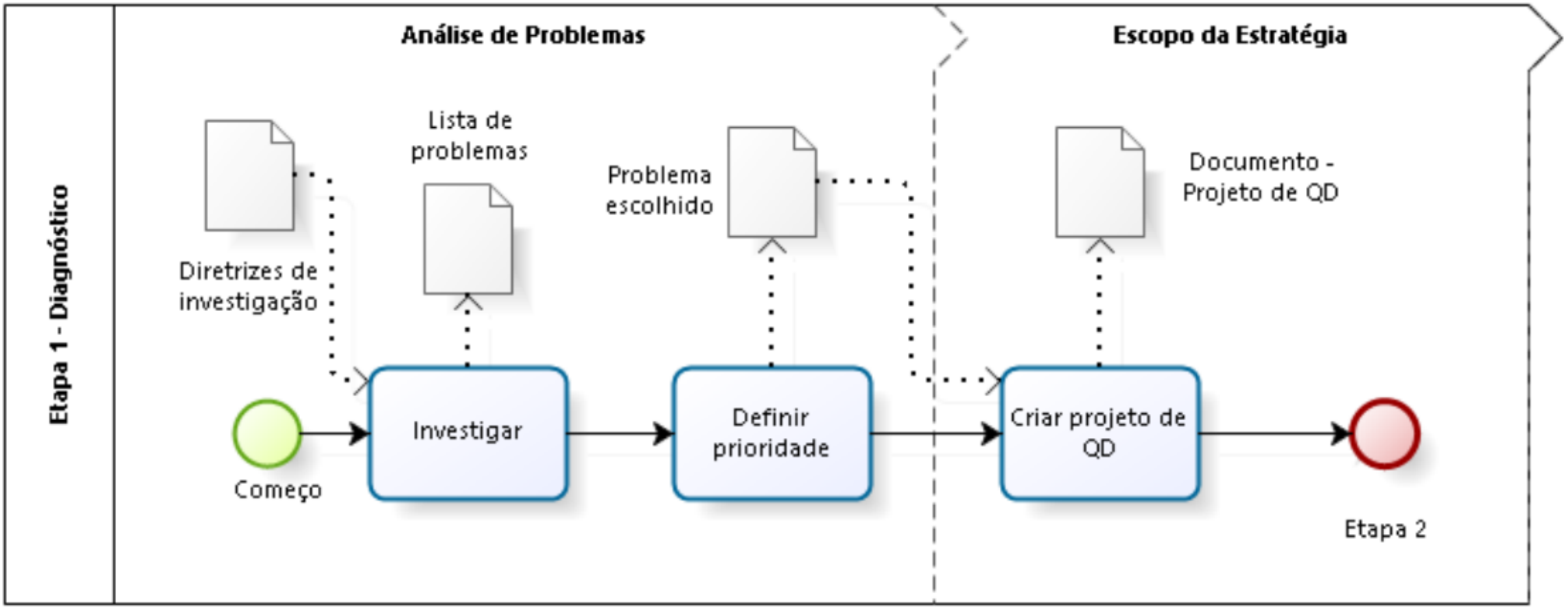
- **Dados inconsistentes**:

  Ex: *data de nascimento* = “29/05/1990”,

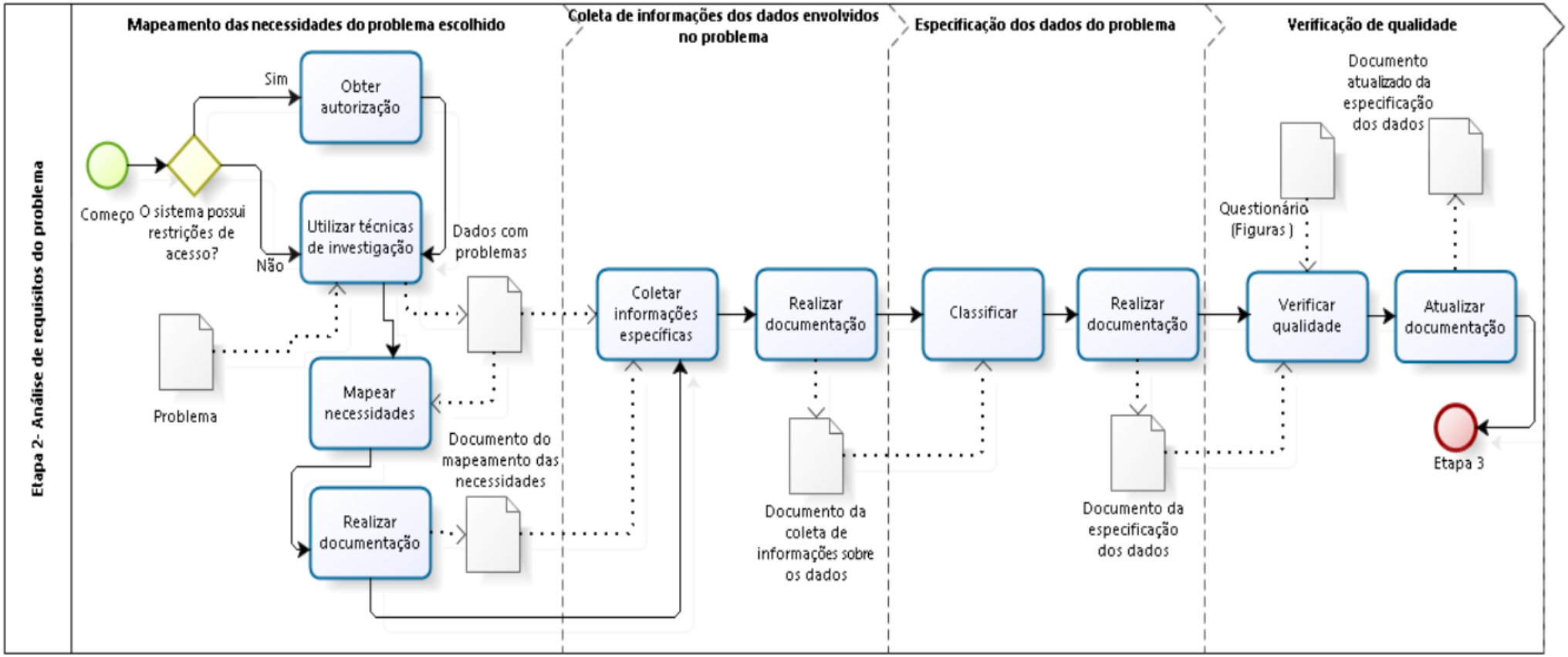
  *idade* = “38”.

# Arquitetura de Processos de Data Quality

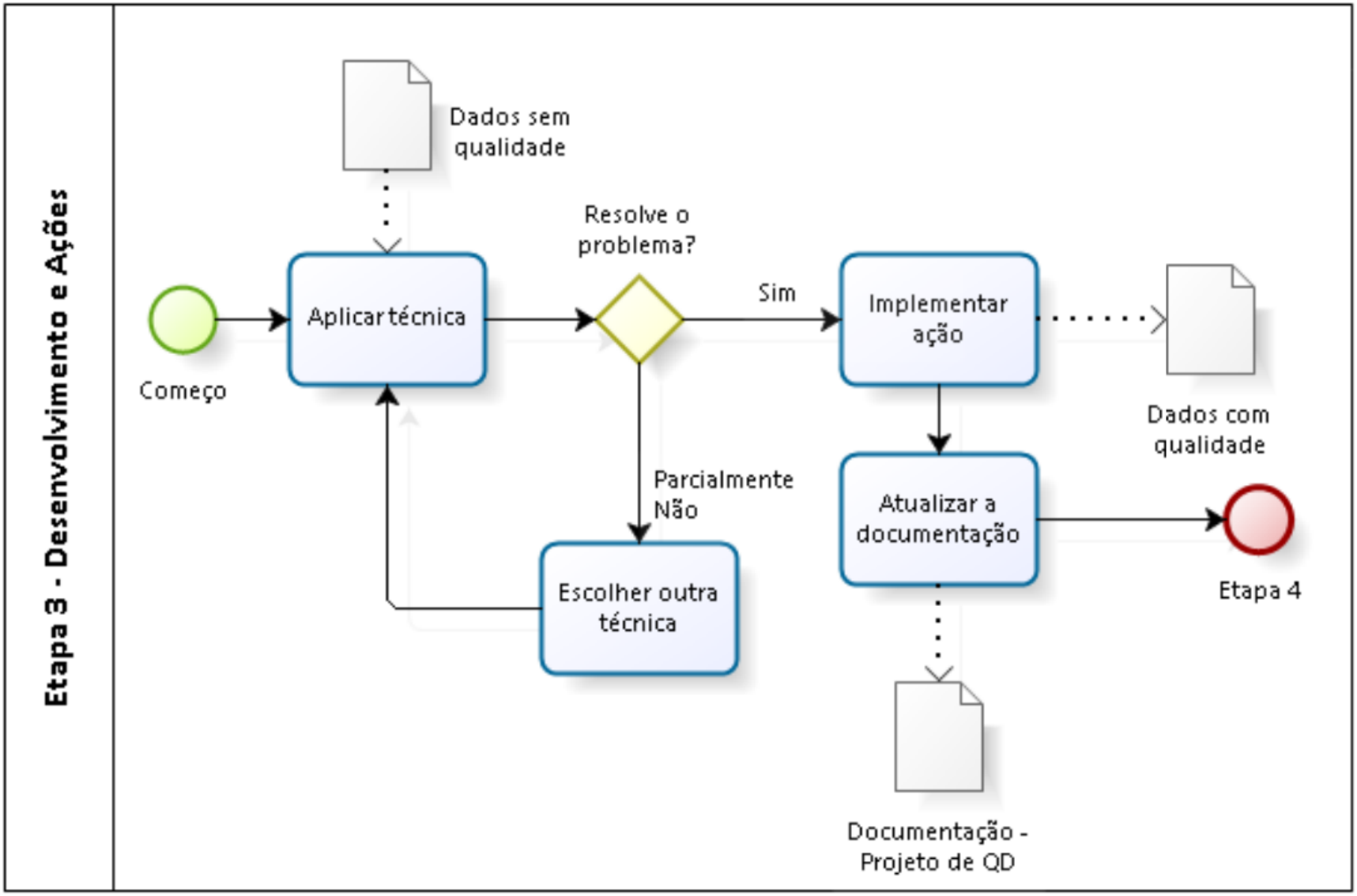
# Diagnóstico (Profiling)



# Análise de requisitos do problema (Audit)



# Desenvolvimento de ações (Cleansing)



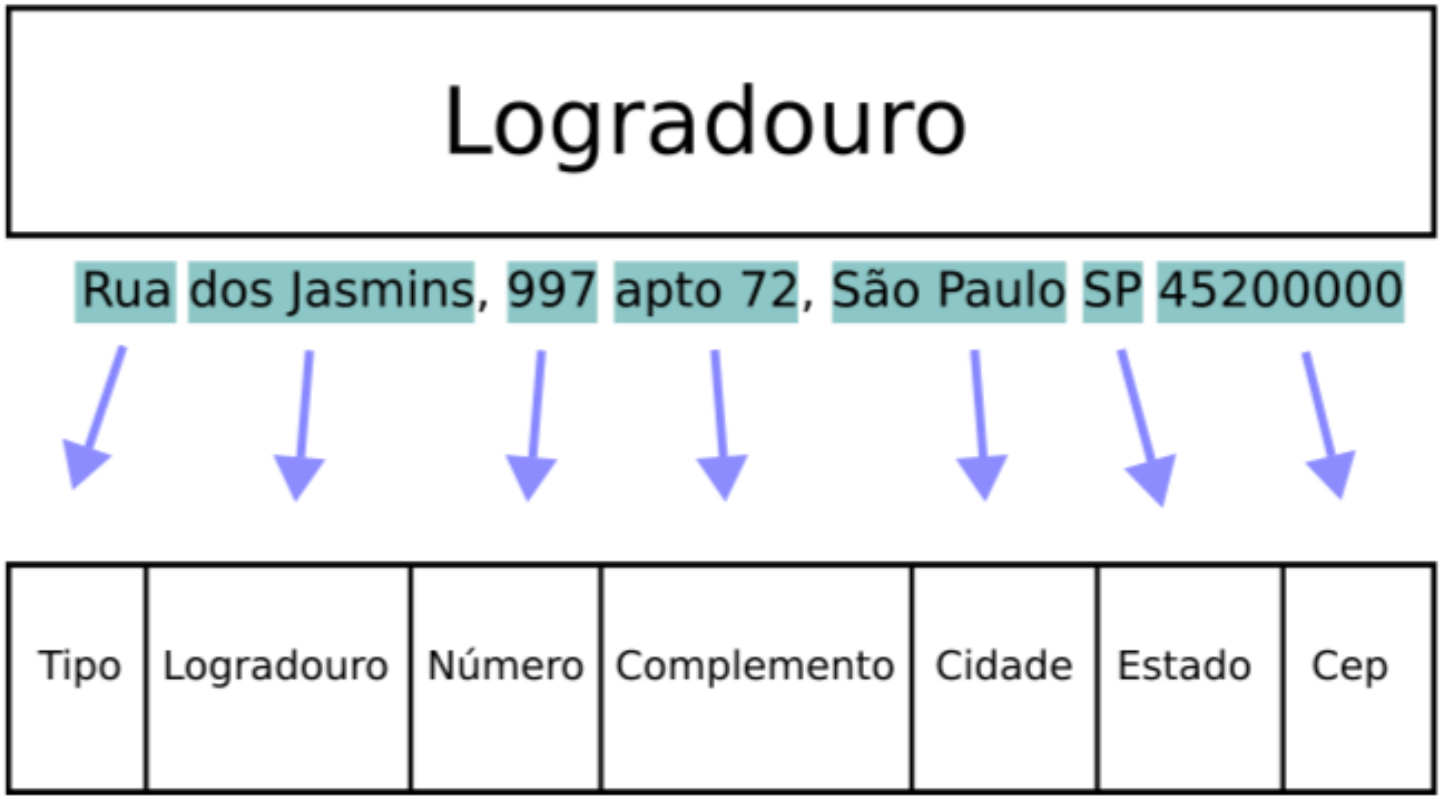
# Algumas soluções de mineração em Data Quality

- Resolver Redundâncias
- Resolver dados faltantes
- Resolver dados ruidosos
- Previsão
- Agrupamento

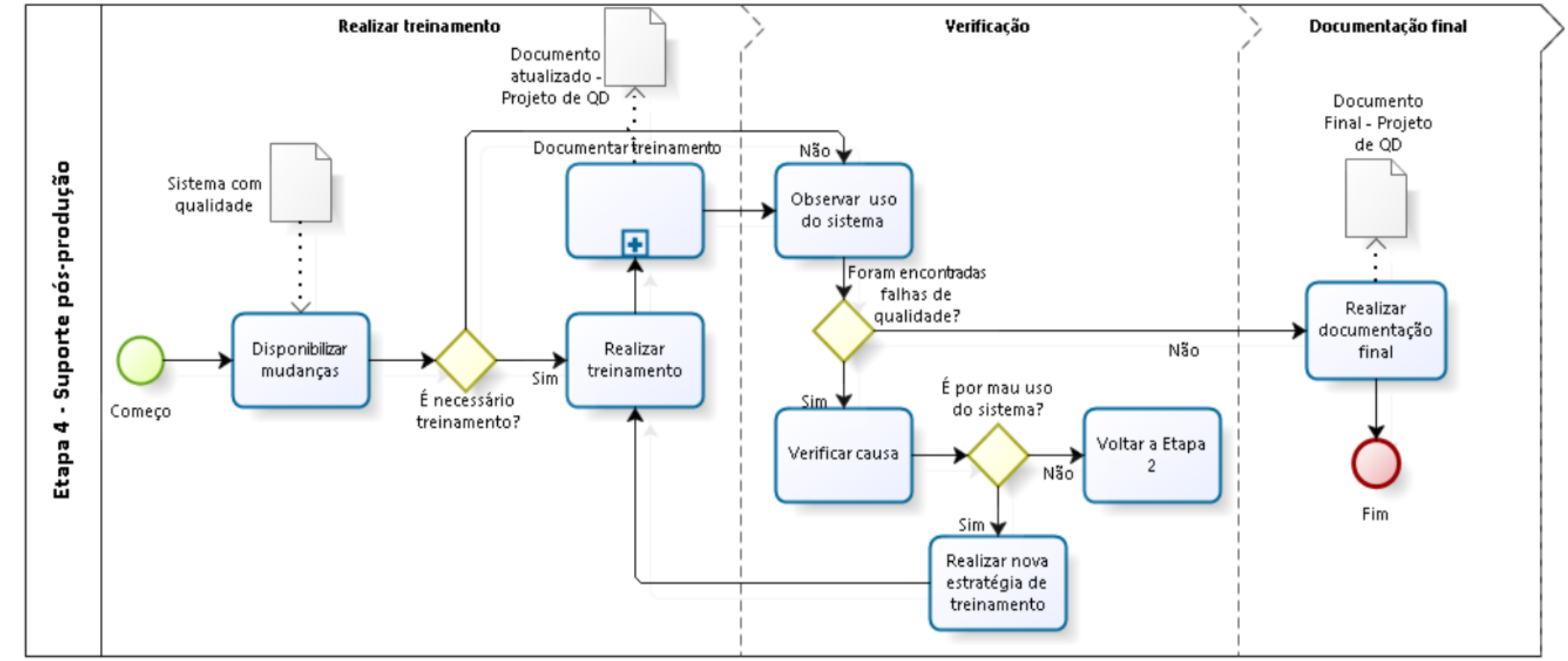
# Algumas regras de negócio

- **RN1** - Cada evento deve estar associado a apenas a uma tabela.
- **RN2 -** Todos os tipos de conta deverão seguir a nomenclatura CN_. Exemplo CN_Corrente, CN_Investimento.
- **RN3 -** O sistema deverá aceitar apenas as abreviações previamente definidas.
- **RN4 -** Números não serão aceitos nos campos de tipo nome.

# Exemplo de parsing em uma variável *endereço*



# Suporte pós-produção (Monitoring)



# Fatores que afetam a qualidade dos dados

- Processos de negócios deficientes;
- Cultura empresarial;
- Falta de padronização entre sistemas;
- Negligência humana.

# Técnicas

- Armazenamento e integração de dados
- Engenharia de requisitos
- Mineração de dados
- Governança de dados.

# Conclusão

A qualidade não envolve somente a confiabilidade do dado, além disso, é necessário que as informações estejam disponíveis na hora, lugar certo e para as pessoas certas.


*(McGilvray 2008)*

# Perguntas?

# Referências


- http://www.devmedia.com.br/uma-visao-sobre-a-qualidade-dos-dados/6973
- https://www.binapratica.com.br/data-quality
- https://www.cetax.com.br/qualidade-de-dados/
- https://marketing.serasaexperian.com.br/artigos/como-testar-a-qualidade-de-dados-da-sua-empresa/
- https://pt.slideshare.net/cecamoraes/banco-de-dados-integrao-e-qualidade-de-dados
- http://qualidadededados.blogspot.com.br/
- https://www.tiespecialistas.com.br/2012/05/data-quality-management-dados-sao-a-causa-e-consequencia-de-um-bom-negocio/

# Referências


- http://blog.in1.com.br/data-quality-o-que-e-quais-beneficios-e-como-adotar
- http://www.triscal.com.br/web/pt/servicos/?bi&integracao-e-qualidade-de-dados
- https://smartbridge.com/data-done-right-6-dimensions-of-data-quality-part-1/
- http://svetlana.dbsdataprojects.com/2015/07/18/data-quality-the-core-of-analytics/
- http://www.devmedia.com.br/data-quality-como-esta-a-qualidade-dos-nossos-dados/3021
- http://blog.mjv.com.br/ideias/data-quality-processos-pessoas-e-ferramentas
- https://marketing.serasaexperian.com.br/wp-content/uploads/2016/04/Whitepaper-Guia-para-compra-de-ferramentas-de-Data-Quality.pdf
- **http://tcc.ecomp.poli.br/20142/monografia-Gizia_Dielle.pdf**
- http://us.trinity-data.com/?FID=9&CID=87
- http://slideplayer.com.br/slide/1393891/


Gracie
tack
THANK YOU
thanks
Gracie
thank you
Obrigado
thank you
merci
OBRIGADO
thank you
Gracie
Obrigado
TACK
DIOLCH
Gracie
diolch
BEDANKI
OBRIGADO
GRAZIE
TACK
thank you
TACK
thanks
Gracie
THANK YOU
thank you
DIOLCH
bedanki
TACK
hvala
Gracie
danke
grazie
BEDANKI
merci
Gracie
diolch
Obrigado
tack
Obrigado
merci
bedanki
thanks
merci
Gracie
tack
tack
Obrigado
THANK YOU
TACK
hvala
tack
HVALA
merci
thanks