A vida secreta dos dados

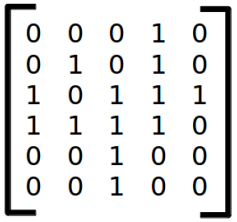
Manejo, Visualização e Compartilhamento de Dados

Nicholas A. C. Marino

nac.marino@gmail.com

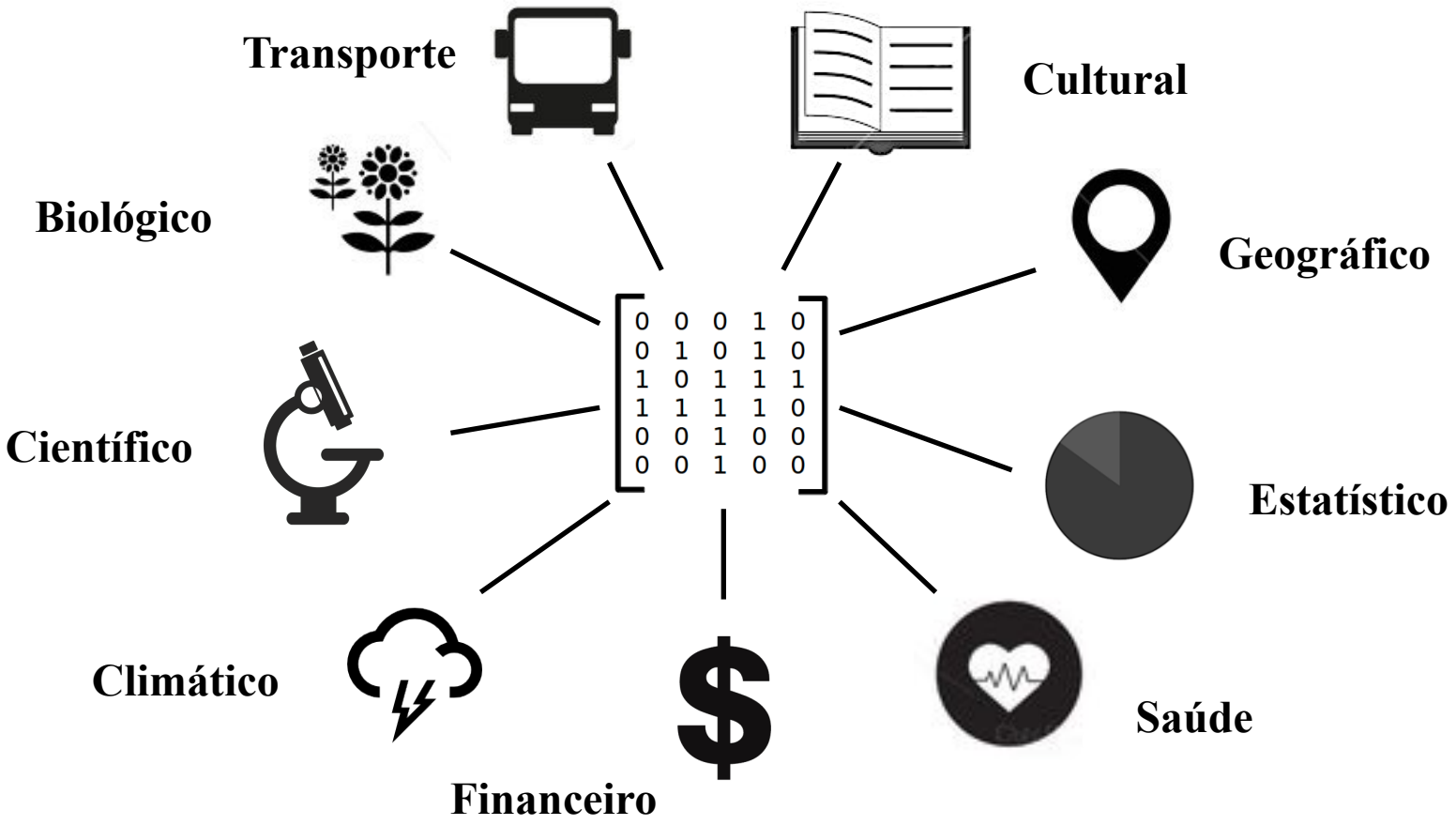
github.com/nacmarino/compartilhaR

O que são ‘*dados*’?

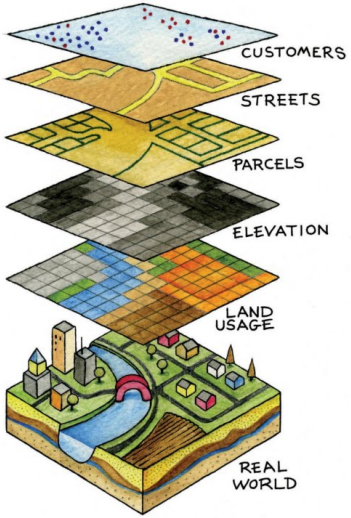


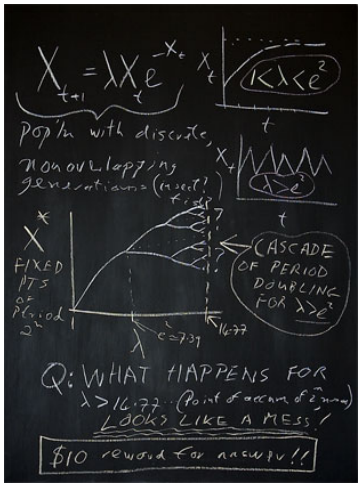
O que são '*dados*'?

- É um conjunto de valores qualitativos ou quantitativos que definem uma variável, processo ou fenômeno.

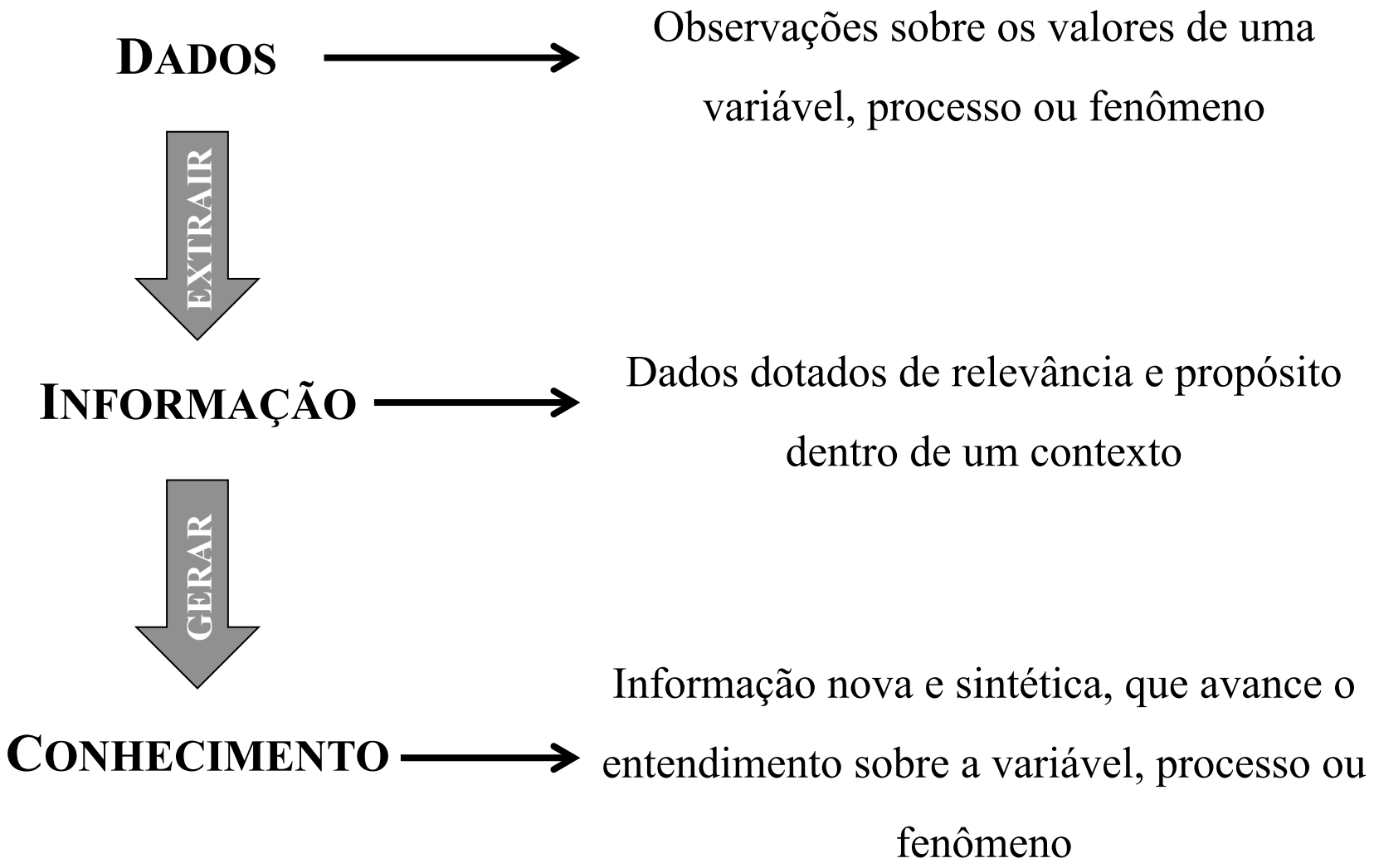


Os dados podem ter praticamente qualquer origem

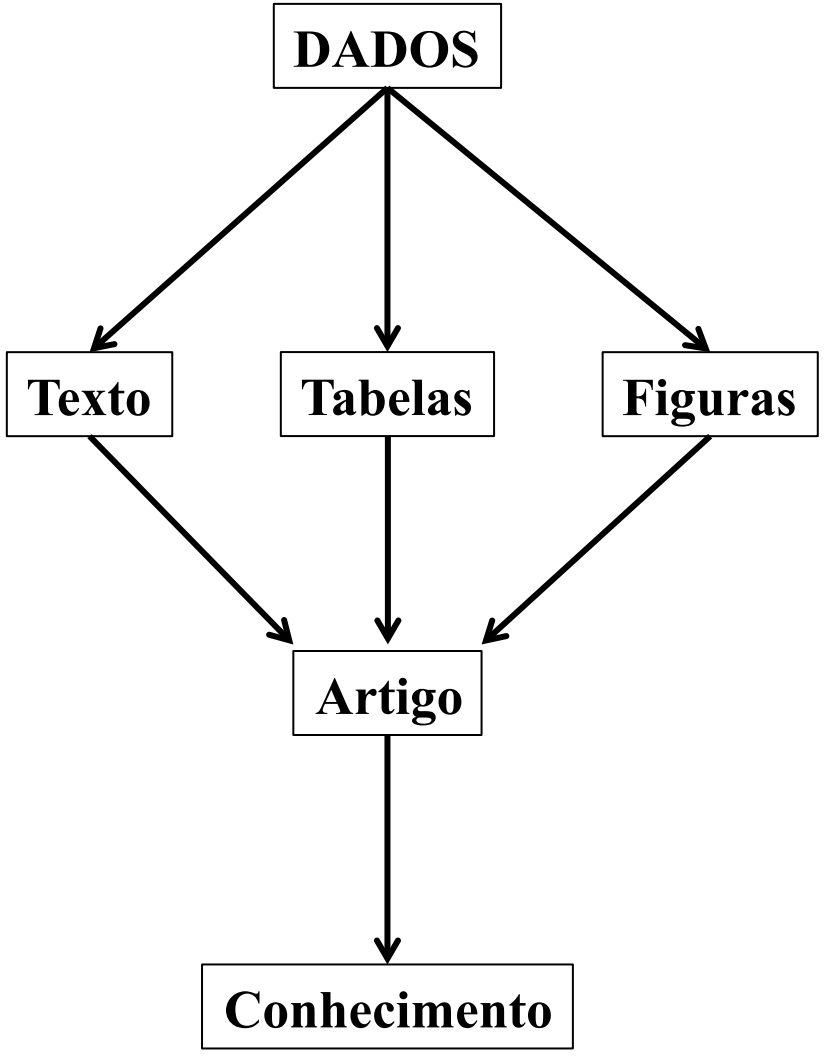




Para que usamos um conjunto de dados?



Os dados, da maneira como pensamos e usamos



Existe um *gap* enorme entre os dados que temos e o conhecimento que podemos gerar a partir deles

Existe um *gap* enorme entre os dados que temos e o conhecimento que podemos gerar a partir deles

Qual(is) dado(s) são importante(s) para responder a minha pergunta?

Existe um *gap* enorme entre os dados que temos e o conhecimento que podemos gerar a partir deles

Qual a procedência e qualidade do dado selecionado?

Existe um *gap* enorme entre os dados que temos e o conhecimento que podemos gerar a partir deles

Como manipular e integrar dados tão heterogêneos?

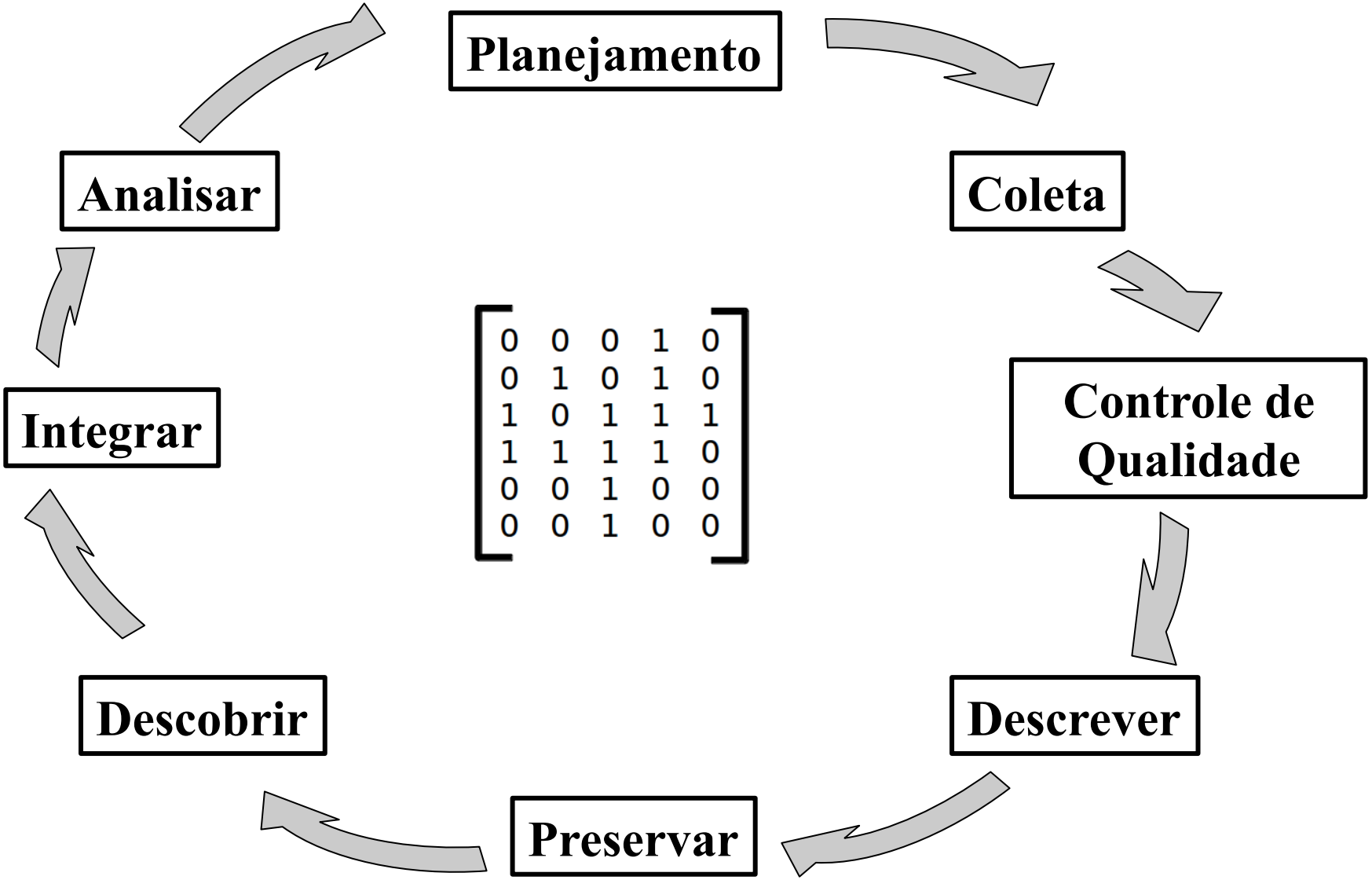
Existe um *gap* enorme entre os dados que temos e o conhecimento que podemos gerar a partir deles

Como interpretar e apresentar o que significado da informação gerada?

Existe um *gap* enorme entre os dados que temos e o conhecimento que podemos gerar a partir deles

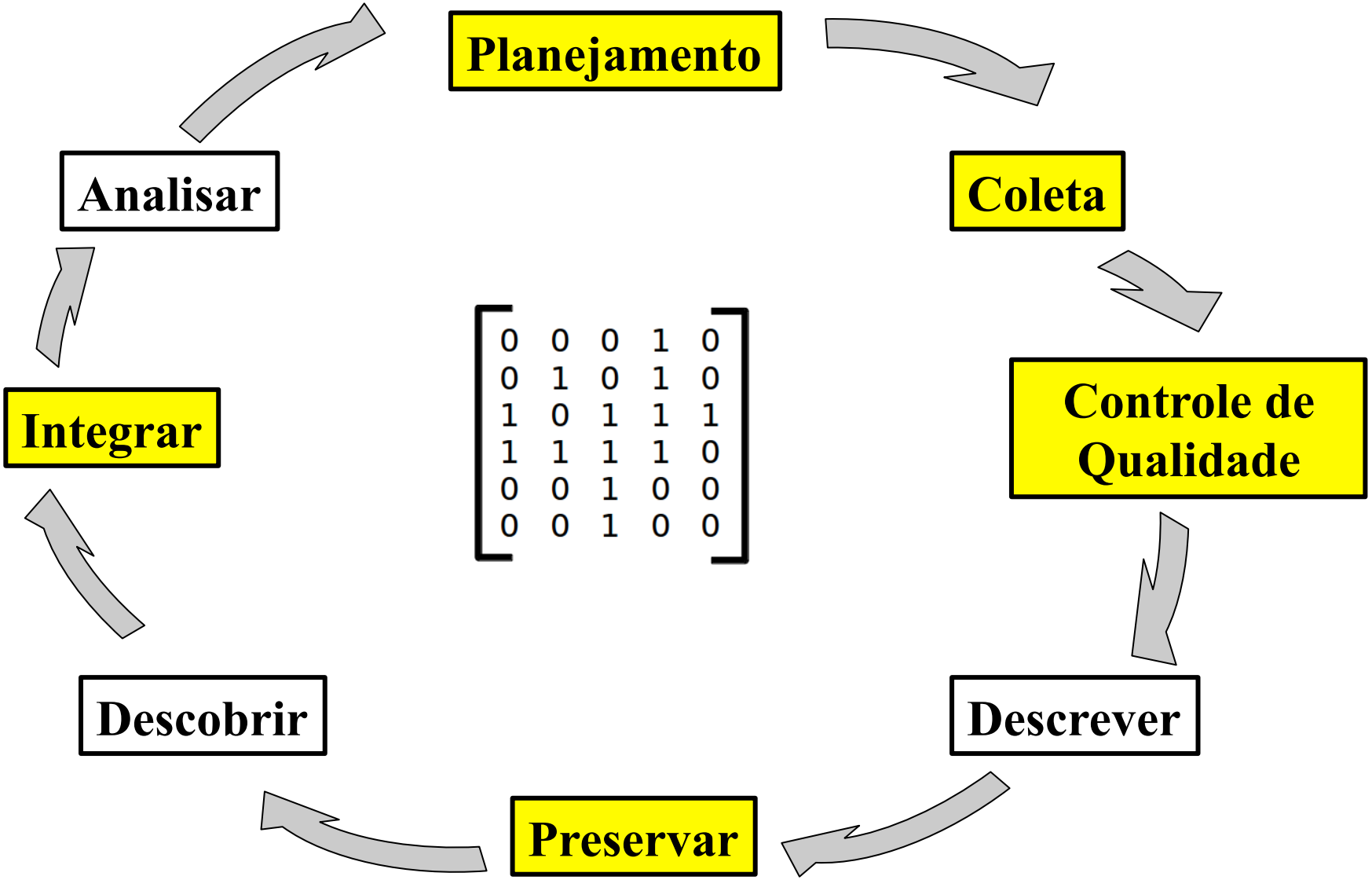
Como garantir que o meu eu do futuro e/ou outras pessoas consigam utilizar esses dados e chegar à mesma informação ou, mais importante, utilizar estes mesmos dados em outro contexto e para responder outras perguntas?

O ciclo de vida dos dados



(Strasser et al., 2011)

O ciclo de vida dos dados, neste curso



(Strasser et al., 2011)

Por que focar nessas etapas?

- Parecem, mas não são nem um pouco triviais;
- São as fases do trabalho que mais demandam tempo;
- Envolve a manipulação e processamento constante dos dados (*e.g.*, criação de informações derivadas, cálculos, etc);
- Exige atenção constante para evitar erros (*e.g.*, digitação), mas ainda assim, você vai errar e precisará repetir muitas vezes as mesmas operações e tarefas;
- É provável que você se perda no que já foi feito, especialmente se todo o processamento for feito de maneira manual.

Exercício – Planejamento, Coleta e CQ de dados

- A Lei de Acesso à Informação (Lei nº 12.527/2011) foi criada para atender o direito constitucional de qualquer cidadão poder solicitar e receber dos órgãos e entidades públicos as informações (públicas) por ele produzidas e custodiadas.

- Qualquer tipo de solicitação de informação pode ser feita, desde que dentro dos limites da razoabilidade e, também, não seja uma informação classificada.

- Solicitações de informação podem ser feitas através do sítio eletrônico https://esic.cgu.gov.br/sistema/site/index.aspx, mediante cadastro prévio.

Exercício – Planejamento, Coleta e CQ de dados

- A partir dos relatórios finais dos projetos, extrair e registrar informações referentes:

  a) à coordenação do projeto;

  b) aos bolsistas envolvidos;

  c) à produção científica, técnica, artística ou cultural;

  d) à rientações e supervisões;

  e) aos recursos financeiros destinados ao projeto.

O maior obstáculo ao uso de um conjunto de dados é a própria forma como nós os armazenamos

- Algumas possíveis dificuldades que vocês podem ter notado:
  - ✓ Dados dispersos e apresentados de forma inconsistente;
  - ✓ Ausência de informações importantes;
  - ✓ Falta de clareza sobre o que cada dado representa;
  - ✓ Mesmo dado apresentado de formas diferentes;
  - ✓ Um mesmo dado representa múltiplas entidades;
  - ✓ ...

Boas práticas para o armazenamento de dados

- **Regra de Ouro:** uma observação por linha, uma variável por coluna.
  - ✓ Todos os dados referentes a um evento observacional de uma dada entidade devem estar na mesma linha;
  - ✓ Cada coluna deve abrigar apenas um tipo de variável (somente números, ou só caracteres, ou só lógicos);
  - ✓ Dados compostos devem ficar em colunas diferentes (*e.g.*, lat/long).

	PONTO 1		PONTO 2		PONTO 3	
variavel	chuva	estiagem	chuva	estiagem	chuva	estiagem
v1	presente	presente	ausente	ausente	presente	ausente
v2	7.2	4.5	3.4	3.2	6.8	3.6
v3	F	V	V	V	F	V
v4	-72.4;-23.4	-72.4;-23.4	-73.2;-22.1	-73.2;-22.1	-71.5;-20.5	-71.5;-20.5

Boas práticas para o armazenamento de dados

- **Regra de Ouro #1:** uma observação por linha, uma variável por coluna.

site	estacao	v1	v2	v3	v4.long	v4.lat
Ponto 1	chuva	presente	7.2	F	-72.4	-23.4
Ponto 1	estiagem	presente	4.5	V	-72.4	-23.4
Ponto 2	chuva	ausente	3.4	V	-73.2	-22.1
Ponto 2	estiagem	ausente	3.2	V	-73.2	-22.1
Ponto 3	chuva	presente	6.8	F	-71.5	-20.5
Ponto 3	estiagem	ausente	3.6	V	-71.5	-20.5

Boas práticas para o armazenamento de dados

- **Regra de Ouro #2:** dados de latitude e longitude sempre em decimal;
- **Regra de Ouro #3:** datas são representadas pelo formato-padrão "ano-mês-dia";
- **Regra de Ouro #4:** dados não disponíveis para uma observação recebem o valor "NA".
- **Regra de Ouro #5:** atenção à maiúsculas e minúsculas – computadores são sensíveis à isso

site	v1	v2	v3
site1	2018-07-10	20	10
site2	2017-05-05	NA	5
site3	2016-04-29	15	5

Boas práticas para o armazenamento de dados

- Uma tabela de dados pode estar em dois tipos de formato:
  - ✓ **Largo:** dados de uma mesma natureza são adicionados às *colunas* – a tabela cresce para o lado;

Sites amostrados

Identidade das espécies

site	sp1	sp2	sp3	sp4	sp5
site1	0	20	0	5	30
site2	10	30	0	0	25
site3	0	15	0	0	5

Registros de ocorrência (ou abundância)

Boas práticas para o armazenamento de dados

- Uma tabela de dados pode estar em dois tipos de formato:
  - ✓ **Longo:** dados de uma mesma natureza são adicionadas às *linhas* – a tabela cresce para baixo.

Identidade das espécies

Sites amostrados

site	especie	abundancia
site1	sp2	20
site1	sp4	5
site1	sp5	30
site2	sp1	10
site2	sp2	30
site2	sp5	25
site3	sp2	15
site3	sp5	5

Registros de ocorrência (ou abundância), sem ausências

Boas práticas para o armazenamento de dados

- Alguns tipos de dados não podem ou não precisam ficar na mesma tabela.

Composição/site			Traits/espécie		Ambiente/site	
site	**especie**	**abund**	**trait 1**	**trait 2**	**altitude**	**vazão**
site1	sp2	20	A	2.1	560	3.4
site1	sp4	5	B	4.3	560	3.4
site1	sp5	30	C	1.0	560	3.4
site2	sp1	10	A	2.9	200	2.1
site2	sp2	30	A	1.8	200	2.1
site2	sp5	25	C	1.1	200	2.1
site3	sp2	15	A	2.3	940	3.5
site3	sp5	5	C	0.8	940	3.5

Boas práticas para o armazenamento de dados

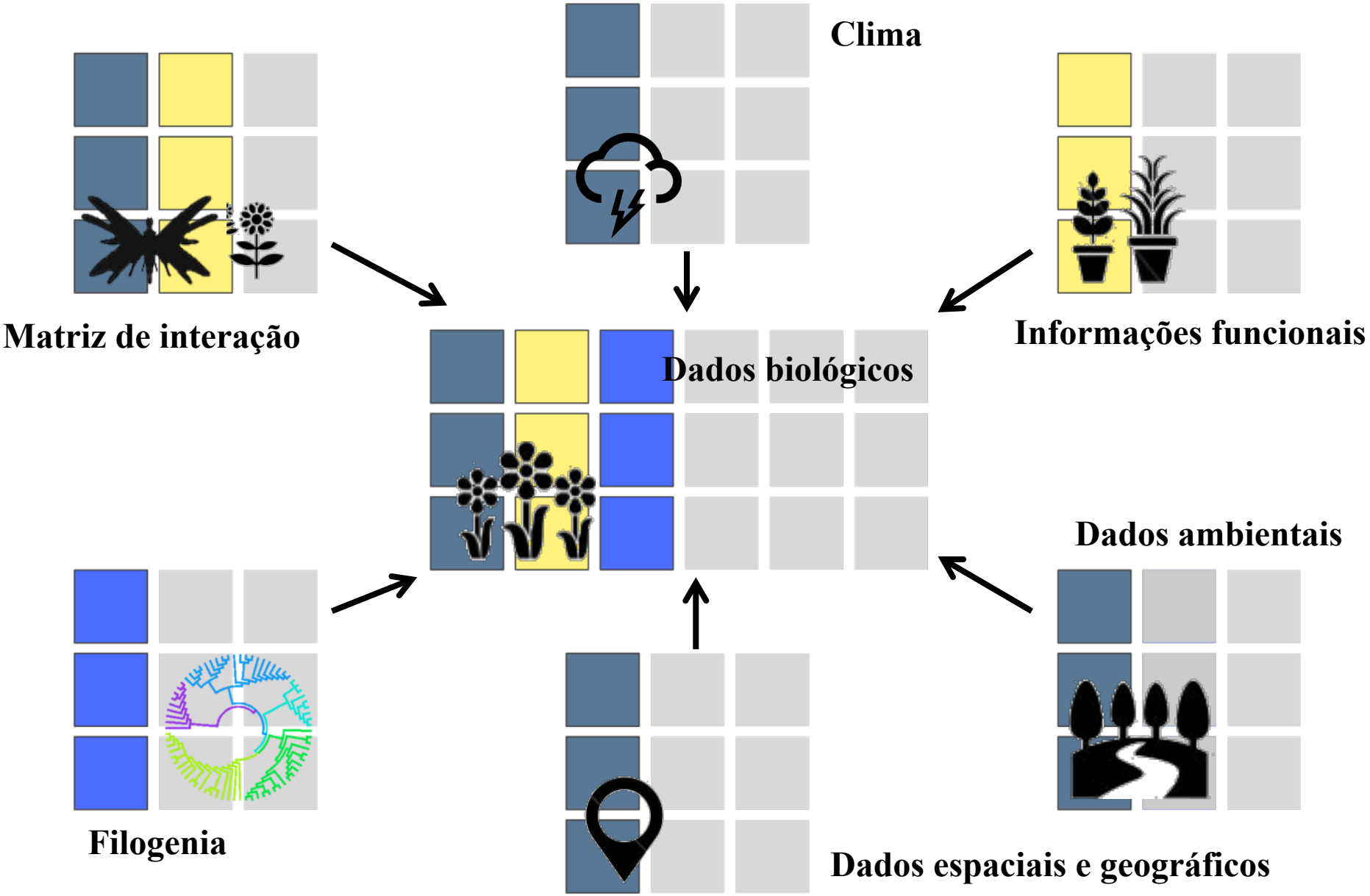
- **Dados relacionais:** conjunto de dados organizados em duas ou mais tabelas, cujas linhas possuem um identificador em comum, que pode ser usado para relacionar as observações entre as tabelas.

especie	trait 1	trait 2
sp2	A	2.1
sp4	B	4.3
sp5	C	1.0
sp1	A	2.9

site	altitude	vazão
site1	560	3.4
site2	200	2.1
site3	940	3.5

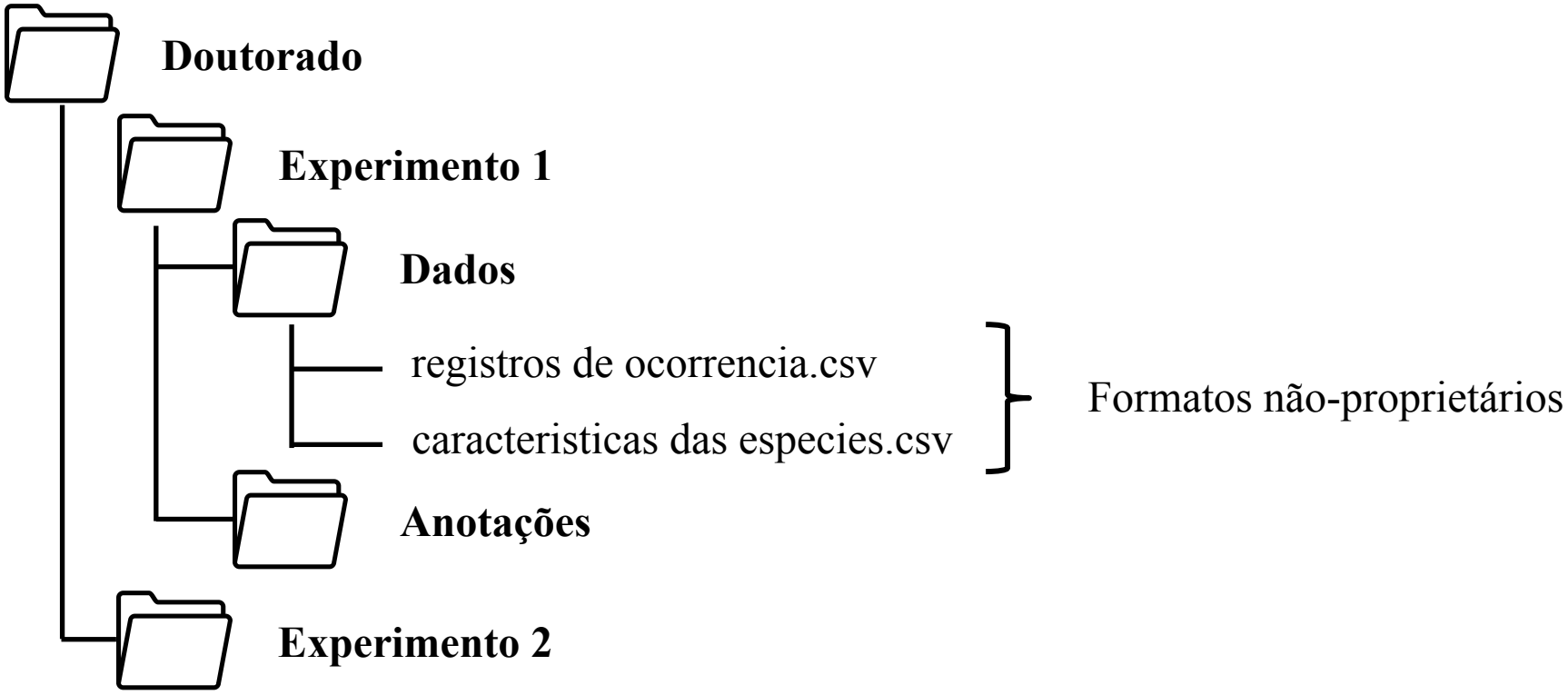
site	especie	abund
site1	sp2	20
site1	sp4	5
site1	sp5	30
site2	sp1	10
site2	sp2	30
site2	sp5	25
site3	sp2	15
site3	sp5	5

Boas práticas para o armazenamento de dados



Boas práticas para o armazenamento de dados

- Mantenha os arquivos de um mesmo projeto organizados em um mesmo diretório, separados em sub-diretórios se necessário, e com nomes auto-explicativos.



Resumindo

- Os dados compreendem todo e qualquer conjunto de valores que definem uma variável, fenômeno e/ou processo de interesse;
- O objetivo final de um conjunto de dados é extrair informações úteis para gerar um novo conhecimento;
- A maior dificuldade do uso de um conjunto de dados está na própria forma como ele é armazenado e processado;
- Existem algumas técnicas e práticas que, se empregadas, podem facilitar e potencializar o uso de um conjunto de dados.