

FIP 606

Análise e Visualização de dados

Unidade 1 - Introdução à Análise de Dados

Emerson M. Del Ponte

Departamento de Fitopatologia - UFV

delponte@ufv.br

Ao final, o estudante deverá:

1. Conhecer **princípios e técnicas de experimentação** aplicada à Fitopatologia (e áreas correlatas)
2. Dominar o **ambiente computacional** R (IDE RStudio), para:
 - Importar, transformar e visualizar dados
 - Definir e ajustar modelos estatísticos
 - Extrair e interpretar as estatísticas de interesse
 - Definir e preparar gráficos mais apropriados para publicar
 - Preparar um compêndio da pesquisa (dados e códigos) e disponibilizar em repositório público

Metodologia de ensino



Informes, chat e envio
de arquivos

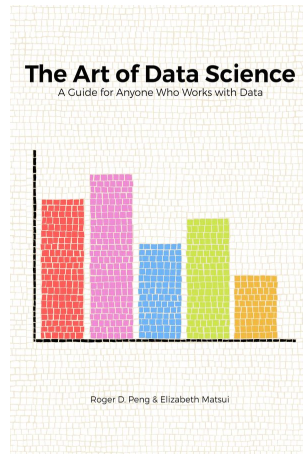
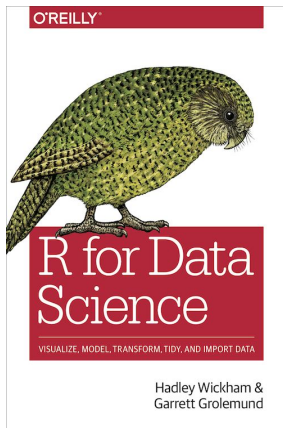


Aulas expositivas -
fundamentos e teoria



Criação de scripts R
usando R + RStudio

Bibliografia utilizada



Definindo Análise de Dados

Processo de inspeção, limpeza, transformação e modelagem de dados com o objetivo de descobrir informação útil, sugerir conclusões e apoiar a tomada de decisão (*Wikipedia*)

Processo de encontrar os dados corretos para responder a uma pergunta, entender os processos subjacentes aos dados, descobrir padrões importantes nos dados e comunicar os resultados visando o maior impacto possível (*Coursera course on data analysis*)

Definindo Estatística

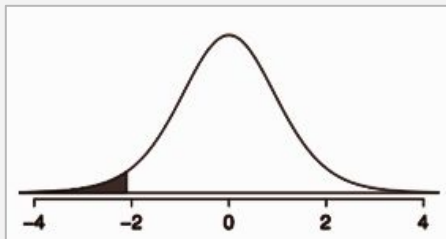
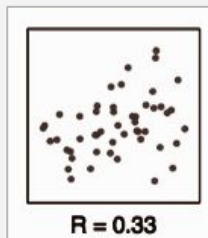
Estudo da coleta, análise, interpretação, apresentação e organização de dados (Dodge, 2006)



Estatística / Análise de Dados

Dois tipos básicos

- Estatística descritiva
- Estatística inferencial



Especializações

Bioestatística, Biometria, Sociometria, Psicometria, Demografia, Atuária, Epidemiologia, Quimiometria, Econometria, etc.



Tipologia da análise de dados



Tipos de análises

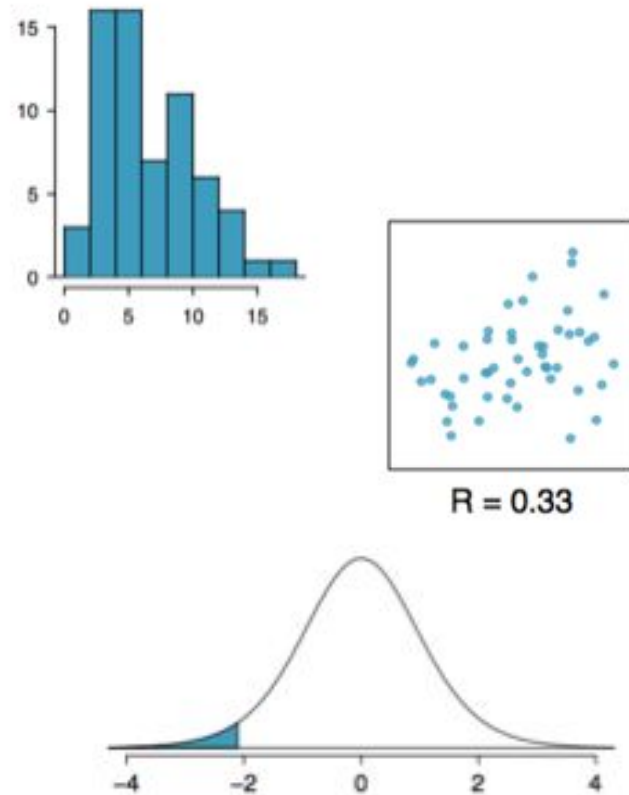
Descritiva

Exploratória

Inferencial

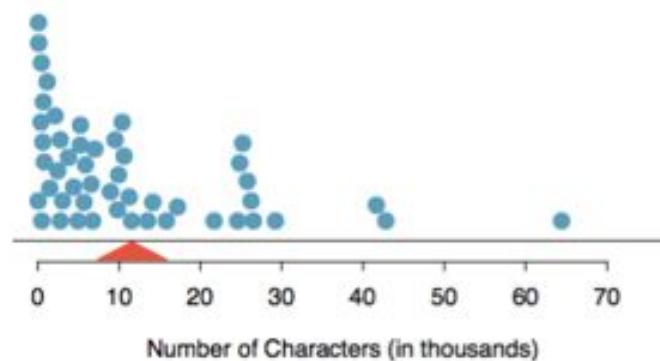
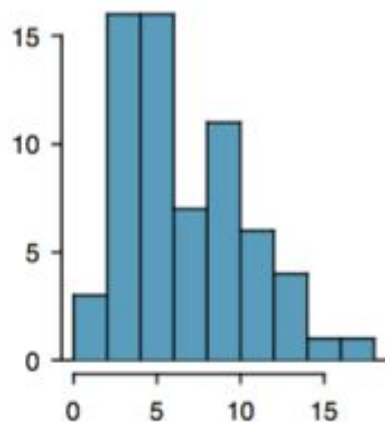
Preditiva

Experimental/causal



Análise descritiva

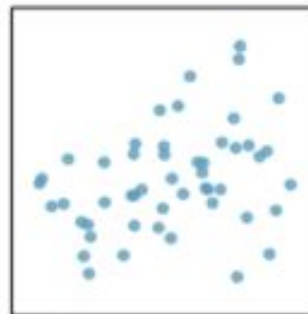
- Primeiro passo da análise a preparação dos dados
- Descreve (por estatísticas) o conjunto de dados



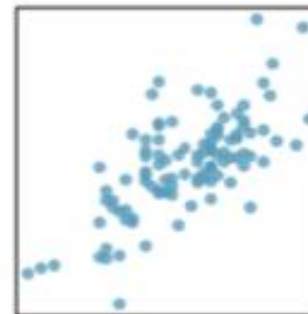
$$\begin{aligned}x_1 - \bar{x} &= 21.7 - 11.6 = 10.1 \\x_2 - \bar{x} &= 7.0 - 11.6 = -4.6 \\x_3 - \bar{x} &= 0.6 - 11.6 = -11.0 \\&\vdots \\x_{50} - \bar{x} &= 15.8 - 11.6 = 4.2\end{aligned}$$

Análise exploratória

- Encontra relações entre variáveis
- Relações não bem conhecidas
- Descobrir novas conexões ou associações
- *Correlation is not causation*



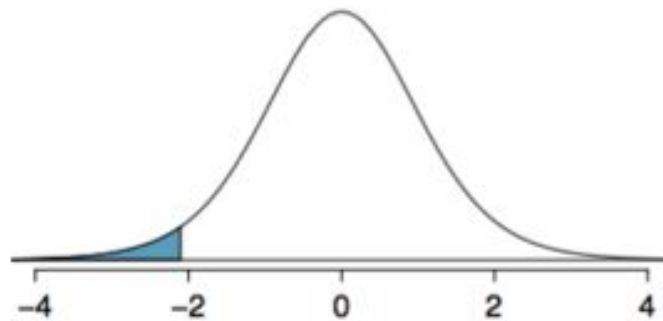
R = 0.33



R = 0.69

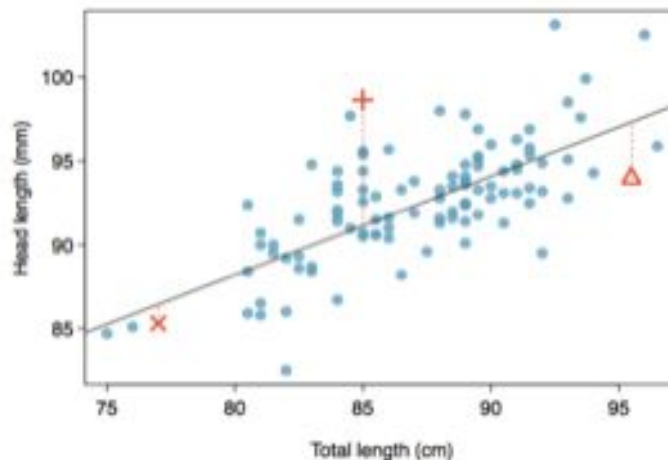
Análise Inferencial

- Amostra pequena -> infere sobre a população
- Modelos estatísticos inferencial
- Envolve a estimativa da incerteza
- Depende da população e do esquema de amostragem




Análise Preditiva

- Valores de uma variável predizem outra variável
- Se x prediz y, não significa que x causa y
- Predição acurada depende da medição correta




Análise Experimental

Avalia o que acontece com uma variável (dependente) quando se altera outra variável (independente) que é o **tratamento** -> aleatorizado e repetido



A Rotina do Pesquisador que Trabalha com Dados



Rotina do pesquisador

Aprende



Uma Idéia -> hipótese

Desenha o estudo

Prepara os materiais

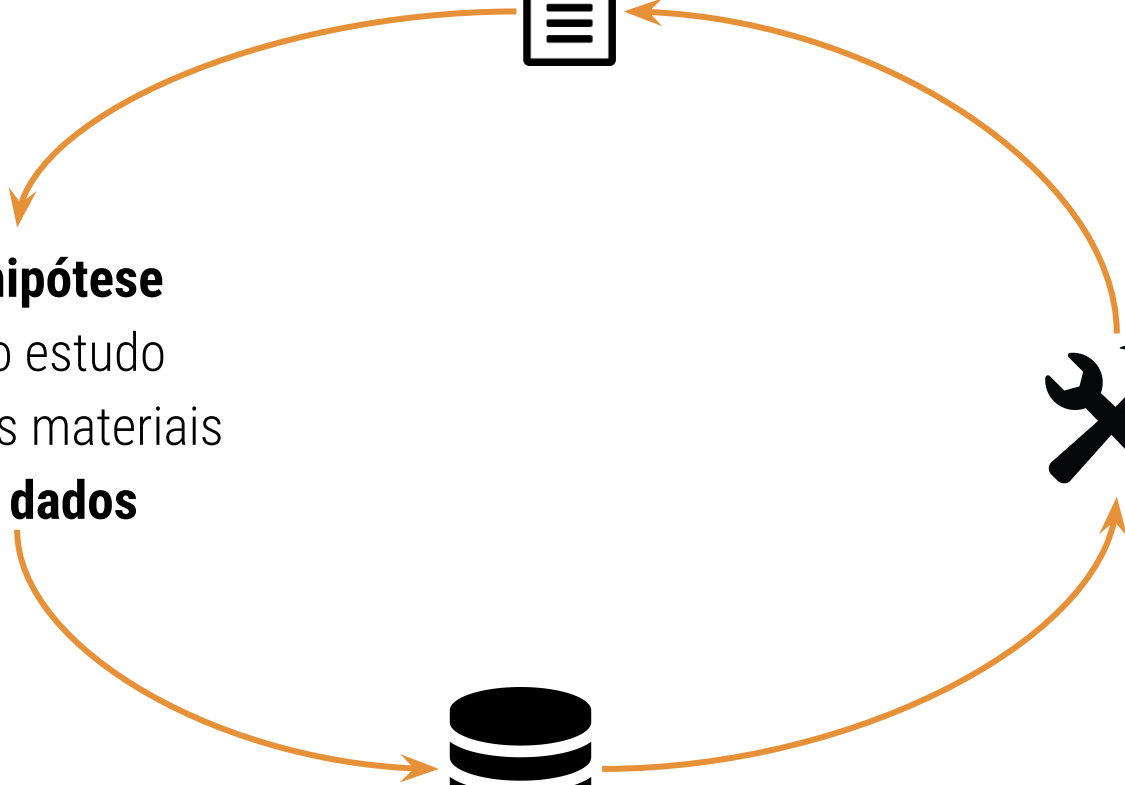
Obtém os dados



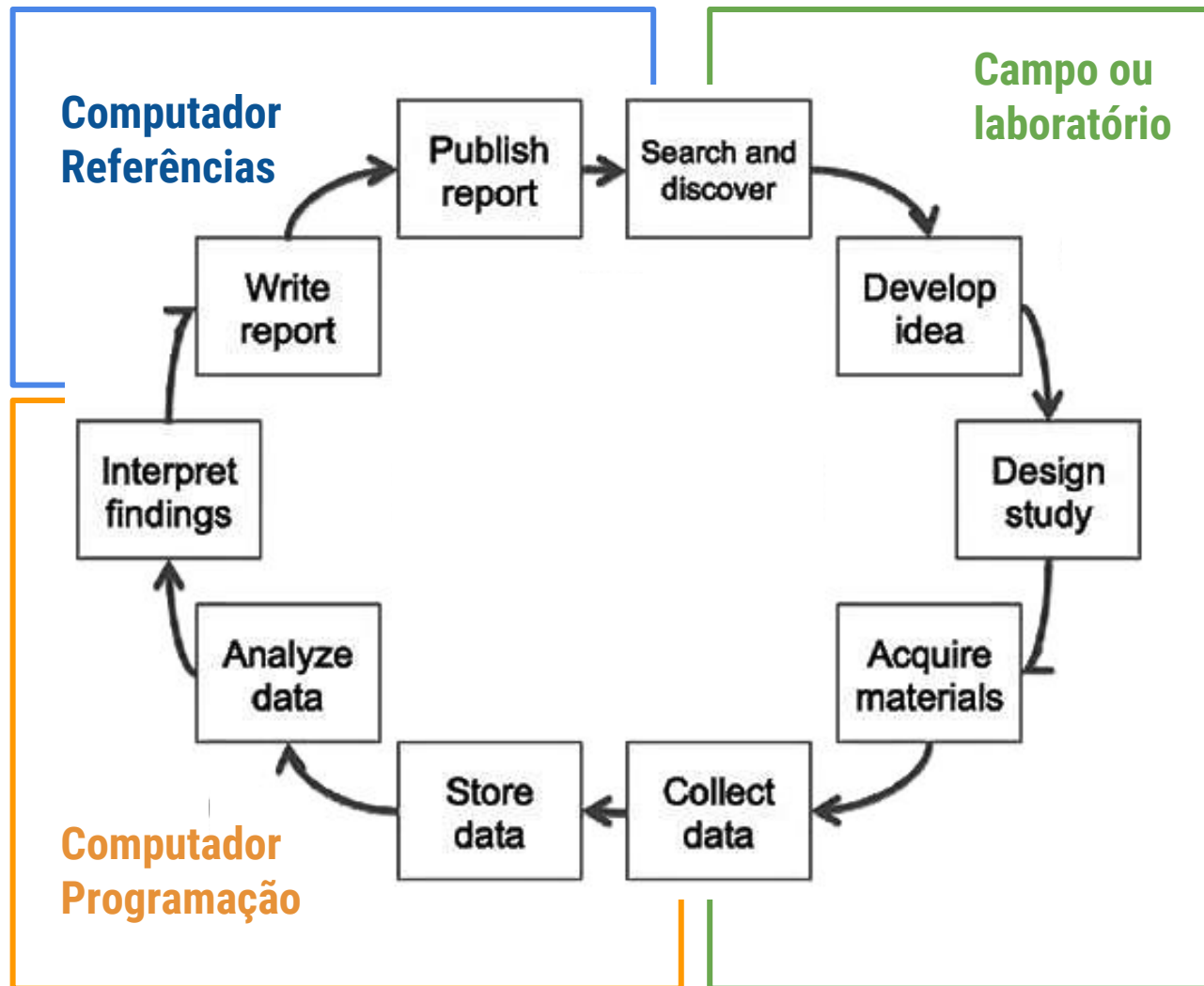
Analisa



Prepara



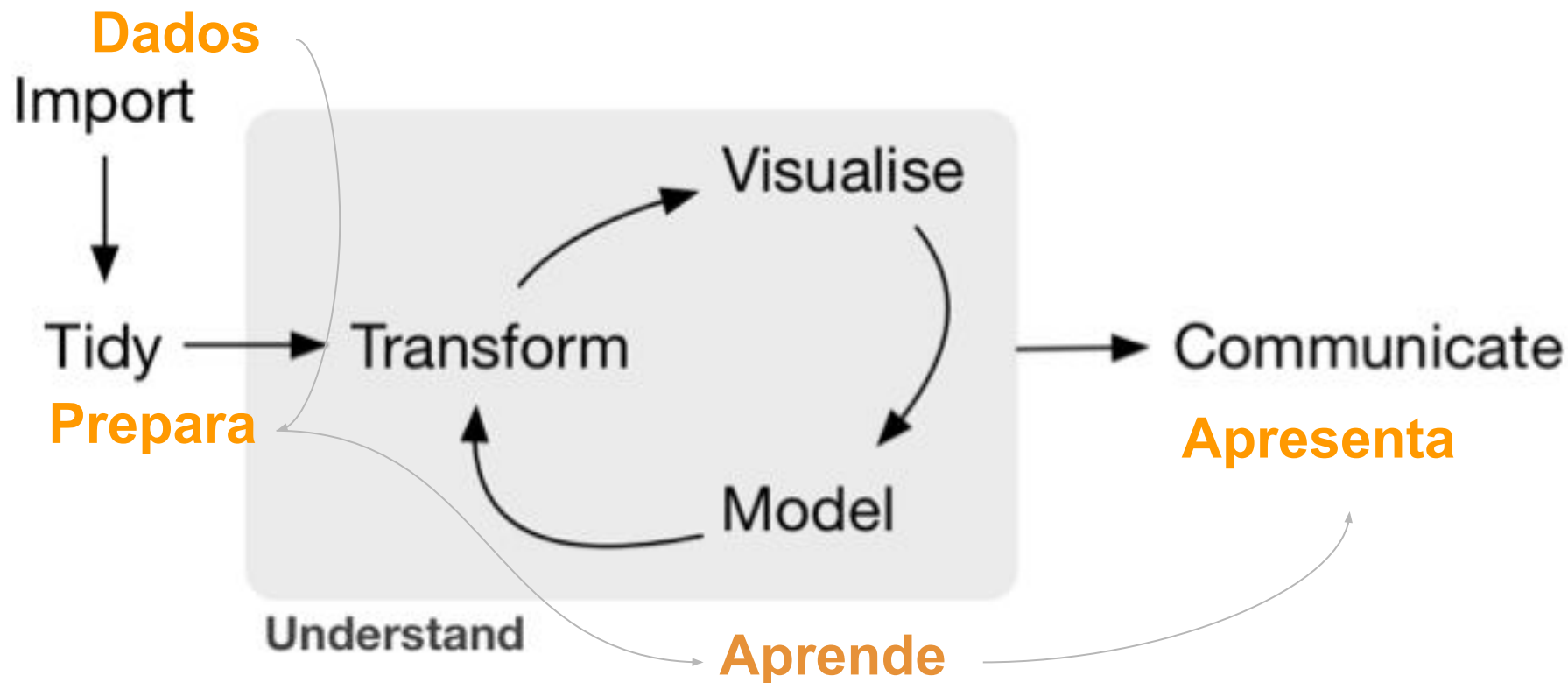
Ciclo de vida da pesquisa científica



Open Science Framework

<https://osf.io/>

Fluxo de uma análise (no computador)



Adaptado de

Fonte: Grolemund & Wickham (2017) R for data science

Reconhece esse fluxo?

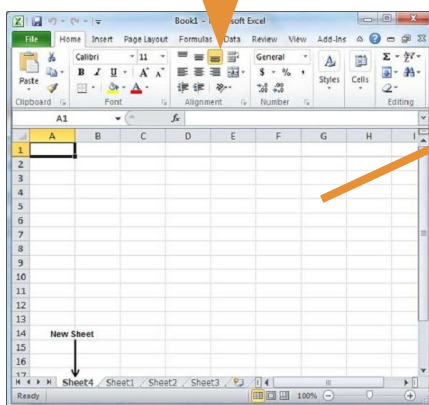
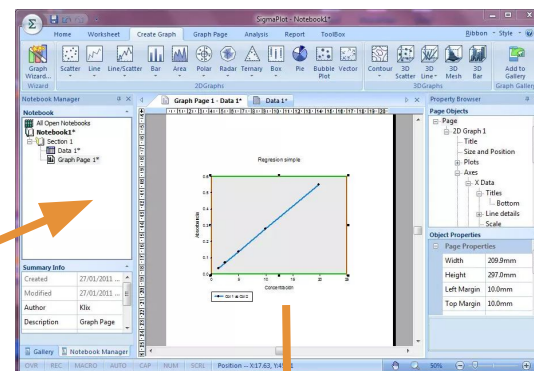
Coleta manual de dados em papel

[illegible]

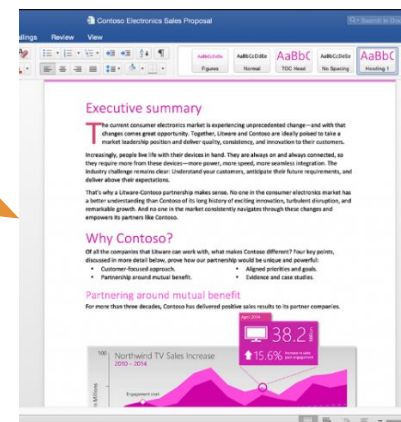
Programa de estatística



Programa para gerar gráficos



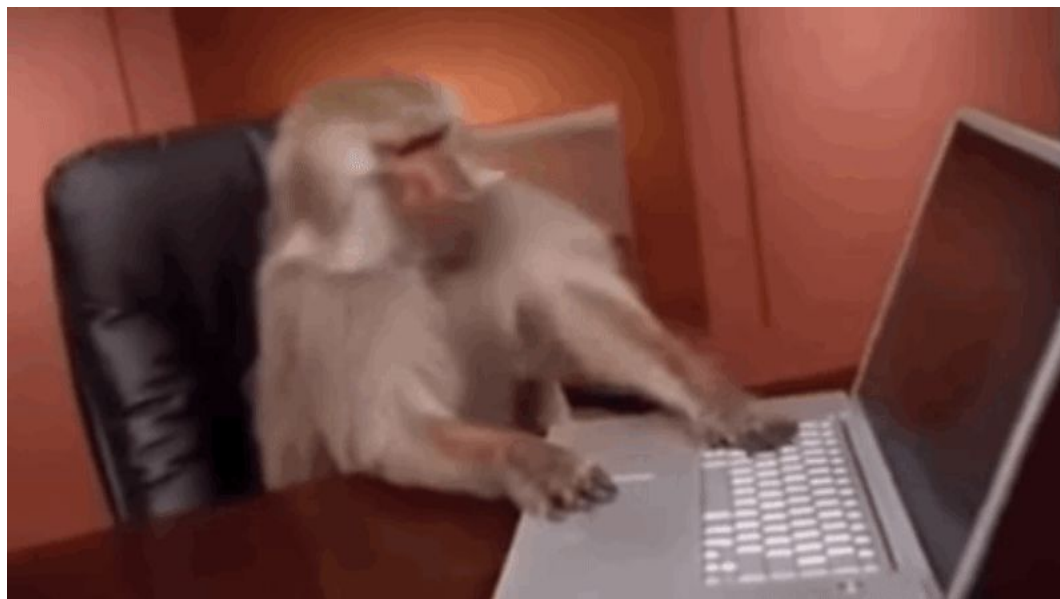
planilha manual dos dados



Processador de texto

No dia a dia da pesquisa

Pesquisadores utilizam softwares em computadores mas não são formalmente treinados a preparar e organizar os arquivos (dados, códigos, gráficos, tabelas, figuras, relatórios e manuscrito).



Então, muitos problemas...

- Transposição excessiva de dados -> erros
- Vários arquivos criados (desnecessariamente)
- Inconsistências e redundâncias na análise
- Documentação é difícil ou descentralizada
- Análises em programas "point & click"
- Gráficos devem ser refeitos a cada reanálise
- Documento de texto é formatado (infinitamente!)

No dia a dia da pesquisa

Boas práticas de **Pesquisa Reprodutível** levará a:

- Maior **eficiência** e **produtividade** no trabalho
- Preparo e análise **colaborativa** (orientador!) de dados e códigos
- **Compartilhamento** de dados e da "pipeline" da análise
- Maior **visibilidade** do trabalho científico como um todo

Afinal, o que produzimos?

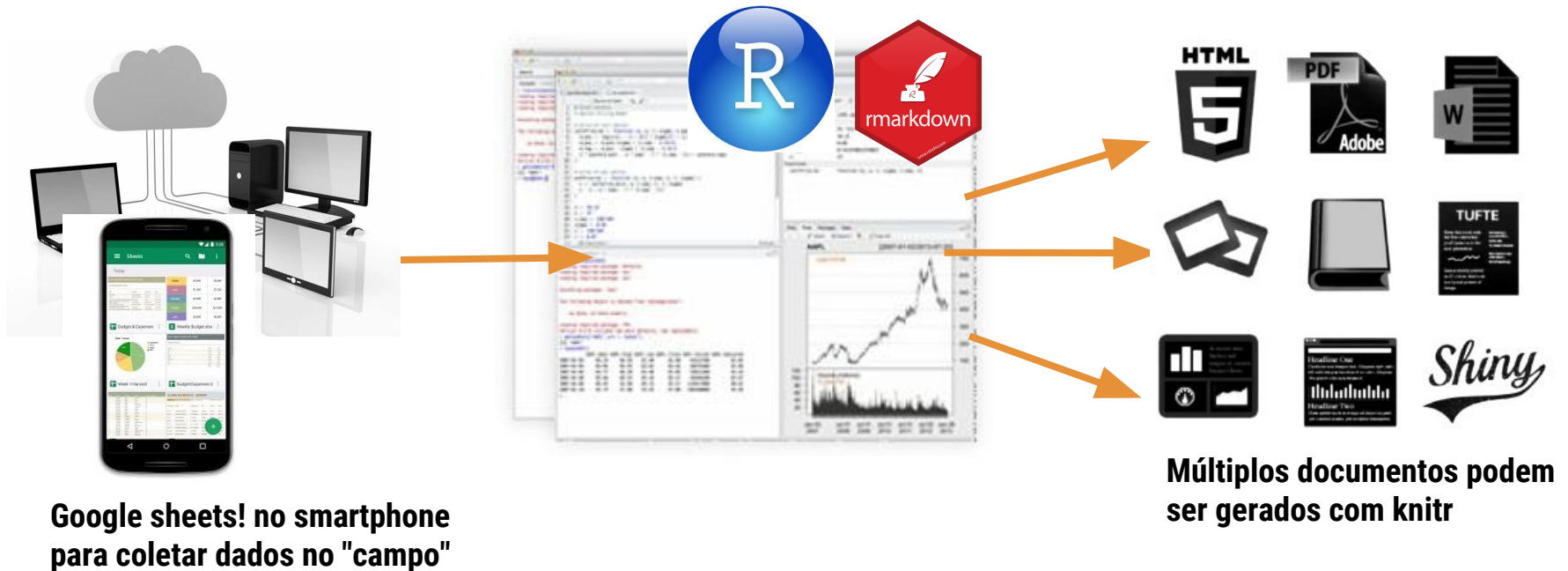
"The final product of **research** is not only the paper itself, but also the full computation environment used to produce the results in the paper such as the **code** and **data** necessary for **reproduction** of the results and building upon the research."



<https://yihui.name/>
(2014)

A much better WORLD!

R + RStudio + RMarkdown



Um pesquisador "reproduzível"

- Diligente e regrado
- Dedica tempo, sem pressa de ter resultados!
- É organizado e sistemático
- Aprende novas ferramentas (R + RMarkdown)
- Documenta todas as etapas do trabalho



Compêndio de pesquisa

Uma **coletânea** de **arquivos** digitais organizados segundo uma **padronização** e convenção que facilita a inspeção, compreensão e reprodução de resultados de análise (Marwick et al. 2017).

com·pen·di·um  (kəm-pĕn'dē-əm)

n. pl. com·pen·di·ums or com·pen·di·a (-dē-ə)

1. A short but complete summary of something.
2. A list or collection of various items.

<https://www.thefreedictionary.com/compendium>

Exemplos de compêndios

Arquivos R + dados + descrição

<https://zenodo.org/record/17804#.Wgh2xRNSwxg>

Preview

BroodParasiteDescription-v1.0.zip

duffymeg-BroodParasiteDescription-8986418

- .gitignore 29 Bytes
- 2014ParasiteSurveyJustBrood.csv 32.4 kB
- BroodParasiteDescription.Rproj 205 Bytes
- CedarBPLifeTable2014.csv 909 Bytes
- CodeforBPpaper.R 13.3 kB
- Metadata.txt 1.7 kB
- NorthBPLifeTable2013.txt 1.3 kB
- NorthBPLifeTable2014.csv 1.6 kB
- README.md 83 Bytes

Dados + códigos R + manuscrito em PDF

<https://github.com/USEPA/LakeTrophicModelling>

Branch: master New pull request Find file Clone or download

jholist Merge pull request #28 from jsta/master Latest commit 4ccb031 on Dec 1, 2016

R	proofs	2 years ago
data	Bryan's final edits	2 years ago
inst	some updates	2 years ago
man	working on final figs	2 years ago
vignettes	proofs	2 years ago
.Rbuildignore	added some text to methods	3 years ago
.gitignore	updated abstract	2 years ago
.travis.yml	added some text to methods	3 years ago
DESCRIPTION	updates plus first shot of cdf CI. Need to work on prob_cdf to get wo...	2 years ago
LakeTrophicModelling.Rproj	updated confusion matrix tables	2 years ago
NAMESPACE	working on final figs	2 years ago
README.md	typo in installation instructions	a year ago
test.txt	testing connection	2 years ago



This repository

Search

Pull requests

Issues

Marketplace

Explore



emdelponete / paper-FGSC-fitness

Unwatch

2

★ Star

2

🍴 Fork

1

<> Code

! Issues 0

🔗 Pull requests 0

📁 Projects 0

📖 Wiki

📊 Insights

⚙ Settings

A research compendium with data, R codes, figures and versions of the manuscript for a research aimed to assess and compare the fitness of strains of the *F. graminearum* species complex causing Fusarium head blight in Brazil

Edit

<http://emdelponete.github.io/paper-FGS...>

Add topics

🕒 17 commits

🌿 1 branch

📦 0 releases

👤 1 contributor

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download



emdelponete Check spell and text edit

Latest commit 7af95d5 on Jan 21

📁 .Rproj.user	Small changes	3 months ago
📁 data	Check spell and text edit	2 months ago
📁 docs	Check spell and text edit	2 months ago
📁 figs	Check spell and text edit	2 months ago
📁 site_libs (1)/jqueryui-1.11.4	first commit	3 months ago
📁 supp	Several changes including .doc as output	3 months ago
📄 .gitignore	created gitignore	3 months ago
📄 CONDUCT.md	first commit	3 months ago
📄 CONTRIBUTING.md	first commit	3 months ago
📄 README.md	Change in readme	3 months ago
📄 site.vml	new changes	3 months ago

Introduction

Perithecia production

Import

Transform

Visualize

Model

Figures

Mycelial growth

Import

Visualize

Model

Sporulation & Germination

Import

Visualize

Model

Figure 2

Pathogenicity

Import

Transform

Visualize temporal progress

Visualize AUDPC

Mixed model

Fungicide sensitivity

Import

Summarize

```
y = "MGR (mm/day)", color = "Species"
) +
ylim(0.2, 1.8)
```

Model

Let's fit a mixed model for the `mgr` data. We first test the effect of the interaction.

```
mgr <- mgr %>%
  unite(species_genotype, species, genotype, sep = "_", remove = F)
lmer_mgr <- lmer(mgr ~ species_genotype * temperature + (1 | species_genotype / isolate), data = mgr, REML = FALSE)
```

```
## fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
```

```
Anova(lmer_mgr)
```

	Chisq <dbl>	Df <dbl>	Pr(>Chisq) <dbl>
species_genotype	8.660965	4	7.015551e-02
temperature	410.223975	1	3.276219e-91
species_genotype:temperature	11.018529	3	1.162610e-02
3 rows			

The interaction was significant. We now create different data sets for each temperature and test the

Vantagens do Compêndio

- Compartilhar para ficar conhecido - e citado!
- Maior eficiência do trabalho com as rotinas de códigos automatizadas
- Maior transparência do processo (e decisões) de análise
- Aproveita a estrutura para outros projetos

Compêndios como "R packages"

<https://github.com/cboettig/nonparametric-bayes>

DOI 10.5281/zenodo.13794

- Authors:
 - Carl Boettiger
 - Marc Mangel
 - Steve Munch

doi: 10.5281/zenodo.12669

This repository contains the research compendium of our work in nonparametric Bayesian inference for improving ecosystem management under deep structural uncertainty. The compendium contains all data, code, and text associated with the publication and has been permanently archived at the DOI indicated by the above badge.

R package

build passing

This repository is organized as an R package, providing functions to integrate the **stochastic dynamic programming** and **Gaussian process inference** methods explored here. Nevertheless, this package has been written explicitly for this project and may not yet be suitable for more general purpose use.

Por que um "R package"? preferência

<https://github.com/cboettig/template>

build **passing** coverage **38%**

This repository provides the current template I use for new research projects.

Why an R package structure?

Academic research isn't software development, and there are many other templates for how to organize a research project. So why follow an R package layout? Simply put, this is because the layout of an R package is familiar to a larger audience and allows me to leverage a rich array of tools that don't exist for more custom approaches.

"But", you say, "a paper doesn't have unit tests, or documented functions! Surely that's a lot of needless overhead in doing this!?"

Exactly...

While there is certainly no need to use all the elements of a package in every research project, or even to have a package that can pass `devtools::check()` or even `devtools::install()` for it to be useful. Most generic layout advice starts sounding like an R package pretty quickly: have a directory for `data/`, a separate one for `R/` scripts, another one for the manuscript files, and so forth. Temple Lang and Gentleman (2007) advance the proposal for using the R package structure as a "Research Compendium," an idea that has since caught on with many others.

"R package" para manuscrito

<https://github.com/jhollist/manuscriptPackage>

build passing

manuscriptPackage


This repository contains materials for using an R package to distribute a reproducible manuscript. Since it is structured as an R package all scripts can be saved as functions in `/R`, data stored in `/data`, and the manuscript stored as an R Markdown file in `/vignettes`.

The current template is a generic manuscript and could be used for draft versions of a manuscript or possible for pre-prints. In the future having additional `.latex` files for other journals (e.g. plos, JSS, etc.) that parse the YAML would be useful. Don't know if I will get around to that.


Install the Package

To install the package and gain access to the materials do the following:

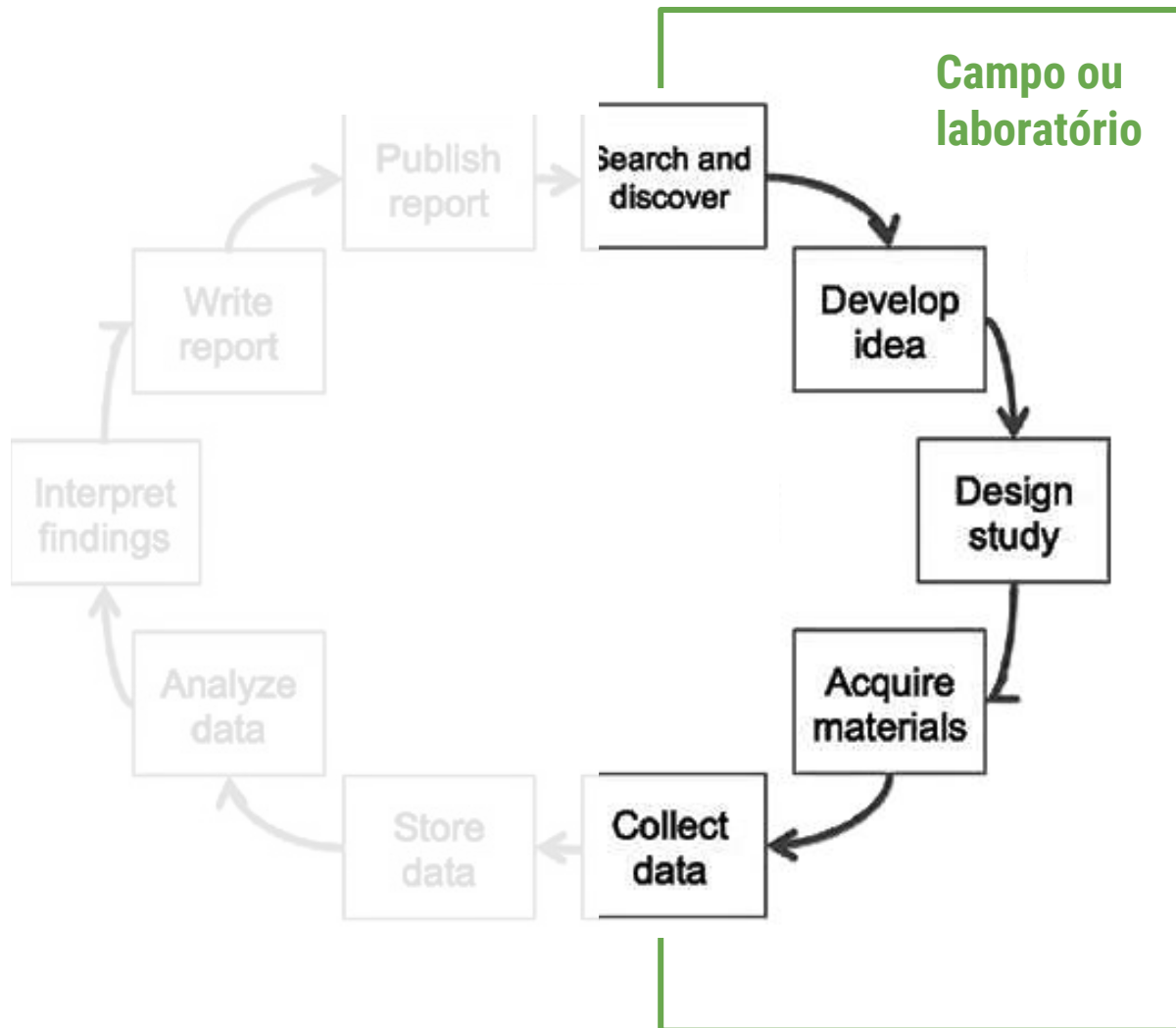
```
install.packages("devtools")
library("devtools")
install_github("jhollist/manuscriptPackage", build_vignettes=TRUE)
library("manuscriptPackage")
```



Planejando o estudo: o que e como
coletar e como org



Ciclo de vida da pesquisa científica



Open Science Framework

<https://osf.io/>

Tipos de Pesquisa ou Estudos

Pesquisa observacional
(Levantamento de campo)

Censitário
(População)

Descritivos
(Descrevem a
situação)

Amostral
(Parte da população)

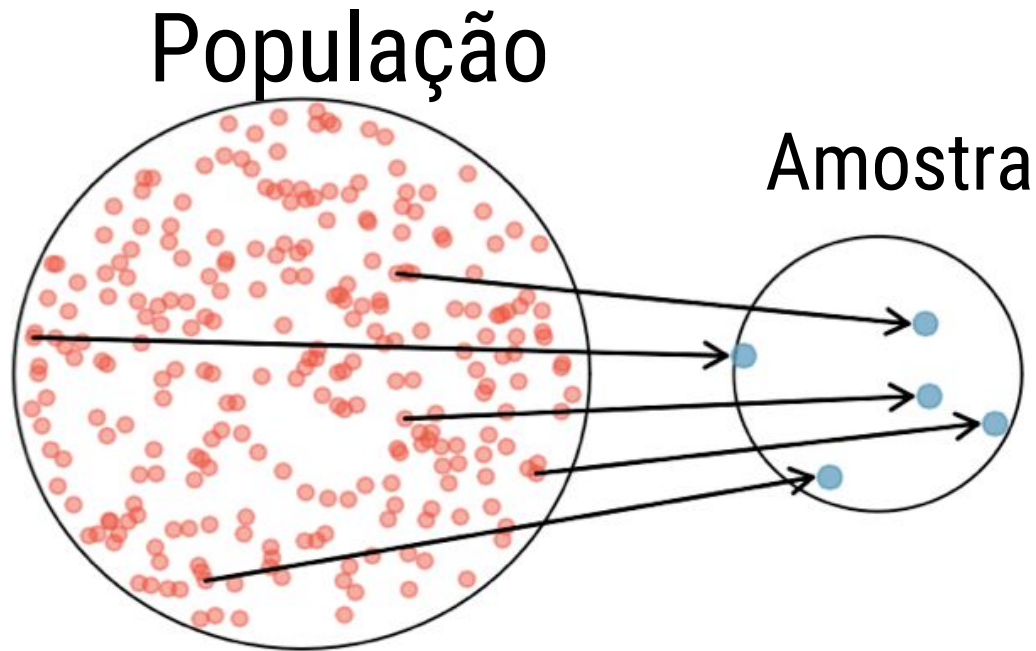
Analíticos
(Tentam explicar a
situação)

Pesquisa experimental

Exposição a um **fator**
(controlado) por intervenção do
pesquisador

Seguem os **princípios** de
casualização, repetição e
(às vezes) controle local

Pesquisa Observacional



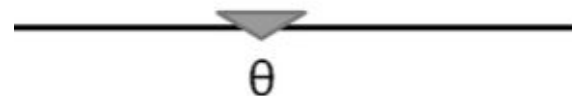
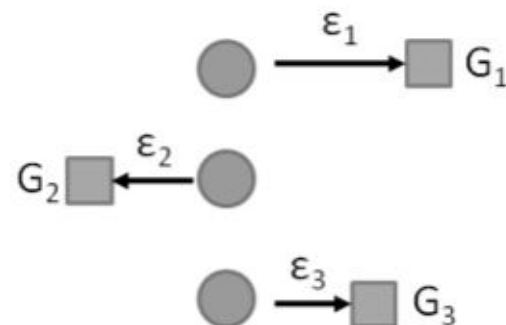
Exemplos de população e amostras?

Pesquisa Experimental

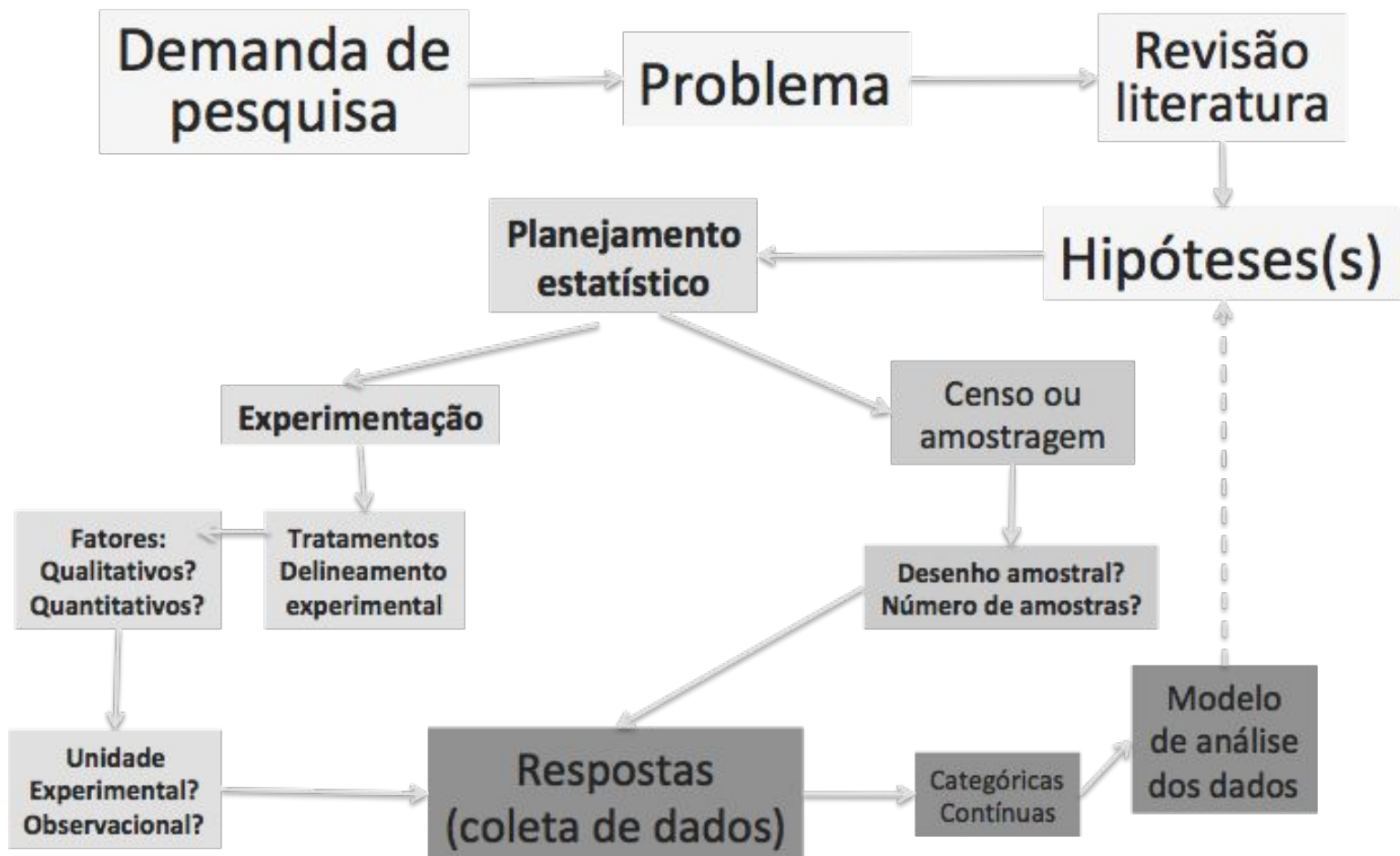
O fenômeno observado tem as causas de sua variação conhecidas

Tratamentos

Erro aleatório



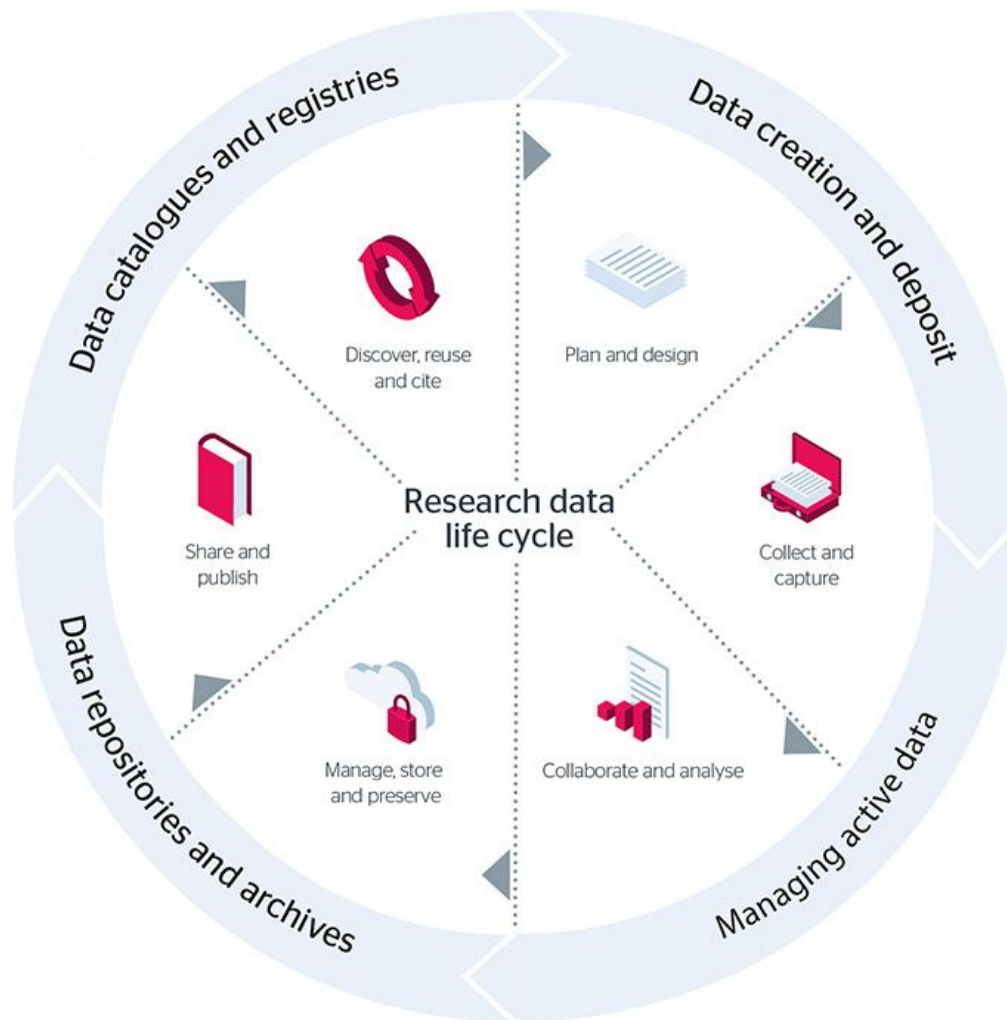
Passos e decisões



Plano, execução e análise experimental

1. Define o problema (científico)
2. Define observações (variáveis ou características)
3. Define as unidades experimentais e observacionais
4. Define os tratamentos
5. Define o delineamento experimental
6. Coleta os dados: variáveis principais e suplementares
7. Confere, organiza, documenta e prepara os dados
8. Conduz análise exploratória (tabelas e gráficos)
9. Ajusta o modelo estatístico (inferencial)
10. Interpreta os resultados
11. Prepara o relatório da análise

"Life cycle" dos dados



How and why you should manage your research data: a guide for researchers

<https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data>

Os Dados

FAIR

Findable, **A**ccessible, **I**nteroperable and **R**eusable

(Wilkinson et al., 2016)

Por que compartilhar os dados?

- Permitir que novas análises sejam feitas com os dados, ou mesmo combinando com outros conjuntos de dados (metanálise).
- Agências de fomento estarão em breve solicitando!

Os dados

Boas práticas

- **Documentação:** dados mais fáceis de entender
- **Formatação:** uso em programas de computador
- **Distribuição:** repositórios e licença aberta

Por que muitos não compartilham?

- Receio de perder a "corrida" pela publicação
- Desconhecimento de como compartilhar
- Postura contrária ao compartilhamento
- Percepção de que é tecnicamente difícil..
- .. e de que toma tempo



Descrição dos dados

- O que são, quando, como e por quem foram coletados
- Como encontrar e acessar os dados (links para repositórios)
- Comentário sobre a utilidade dos dados
- Alertas sobre possíveis problemas ou inconsistências
- Descrição de limitações nos dados ou problemas específicos
- Informação de como conferir se os dados foram importados OK
- Número total de linhas e colunas
- Nome das variáveis, unidades, etc.

Como organizar os dados?

Dados brutos

Os dados podem ser modificados de sua forma original de como foram coletados e anotados. Valores são sumarizados, unidades convertidas, nomes alterados, ou índices calculados. É importante que os dados originais sejam mantidos para se saber o quanto o conjunto foi alterado

Há um formato ideal?

- Dados tabulares de formato retangular são preferidos
- Formato **"Tidy"**: cada linha representa uma observação/sujeito (registro) e cada coluna representa uma variável
- Não duplique colunas com a mesma informação (há exceções)
- Não coloque fórmulas nas células
- Células não são coloridas ou mescladas
- Nomeie as variáveis de maneira simples (trat, rep, peso, altura)
- Datas no formato YYYY-MM-DD
- Não deixe espaços vazios nas células
- Dado faltante: célula vazia ou adicione "NA"
- Exportar para arquivo texto (.csv é o mais comum)
- Escolha um nome curto e descritivo (sem espaços!)