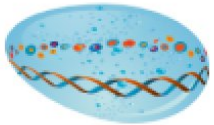
R WORKSHOP

Exploring the variability of the immune responses
using the Milieu Intérieur datasets

INTRODUCTION

Vincent Rouilly, TIL, Institut Pasteur Paris.



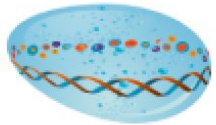
Milieu
Intérieur

INSTITUT
PASTEUR

June 2025, Tartu, Estonia.

R WORKSHOP TEAM

- **Darragh Duffy**, TIL lab PI, Institut Pasteur Paris.
- **Etienne Villain**, TIL lab, Institut Pasteur Paris.
- **Anthony Bertrand**, TIL lab, Institut Pasteur Paris.
- **Vincent Rouilly**, TIL Lab, Institut Pasteur Paris.

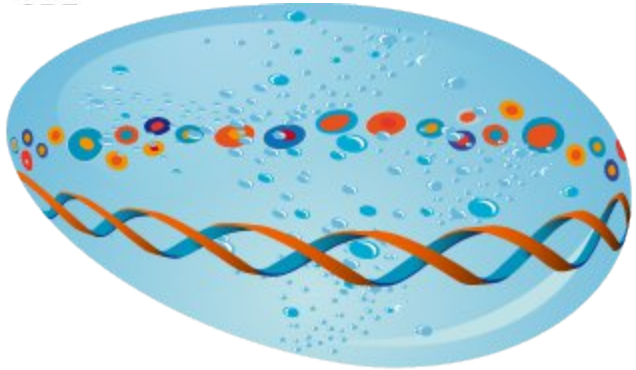


Milieu Intérieur

INSTITUT PASTEUR

The Milieu Intérieur Project

Hosted by Institute Pasteur on behalf of the Milieu Intérieur Consortium



Milieu Intérieur

Vers une médecine personnalisée

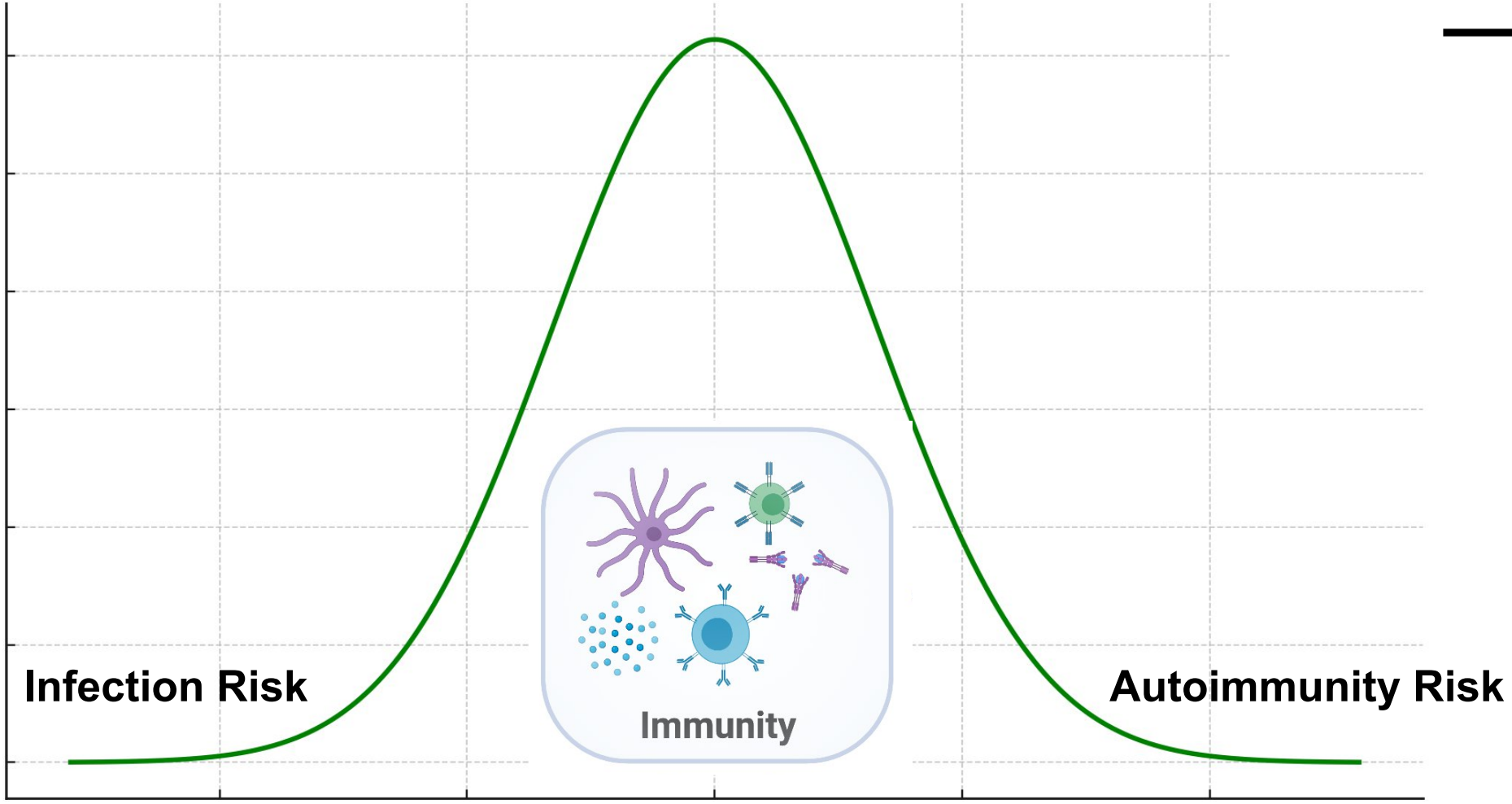
Milieu Intérieur explores the **human immune system** by examining the **genetic** and **environmental** factors contributing to the **variability of healthy immune responses**.

Agence Nationale de la Recherche ANR

INVESTIR L'AVENIR

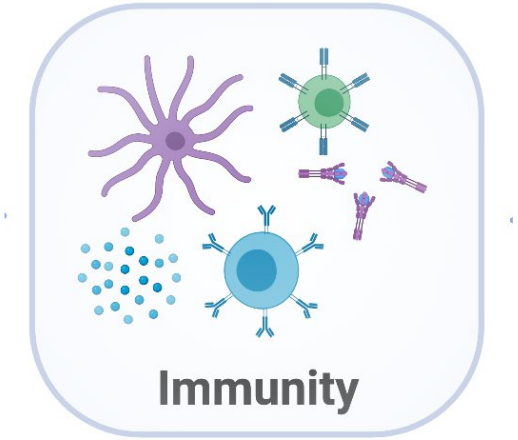
INSTITUT PASTEUR

WHY ARE IMMUNE RESPONSES DIFFERENT BETWEEN INDIVIDUALS?



Duffy, Curr Opin Immunol. 2020

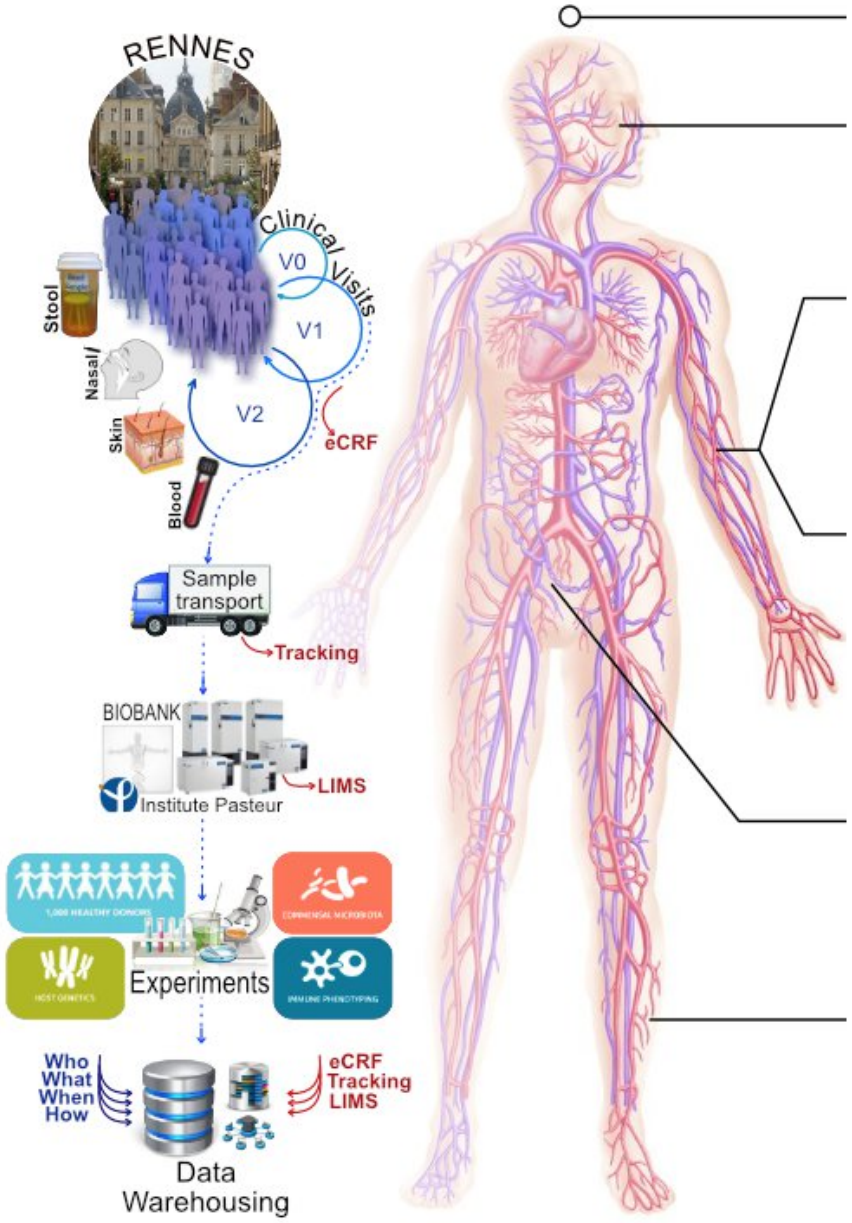
WHY ARE IMMUNE RESPONSES DIFFERENT BETWEEN INDIVIDUALS ?



Duffy, Curr Opin Immunol. 2020

The Milieu Intérieur Project

Hosted by Institute Pasteur on behalf of the Milieu Intérieur Consortium



Healthy cohort

1000 donors

50% Women – 50% Men
20yr < Age < 70yr

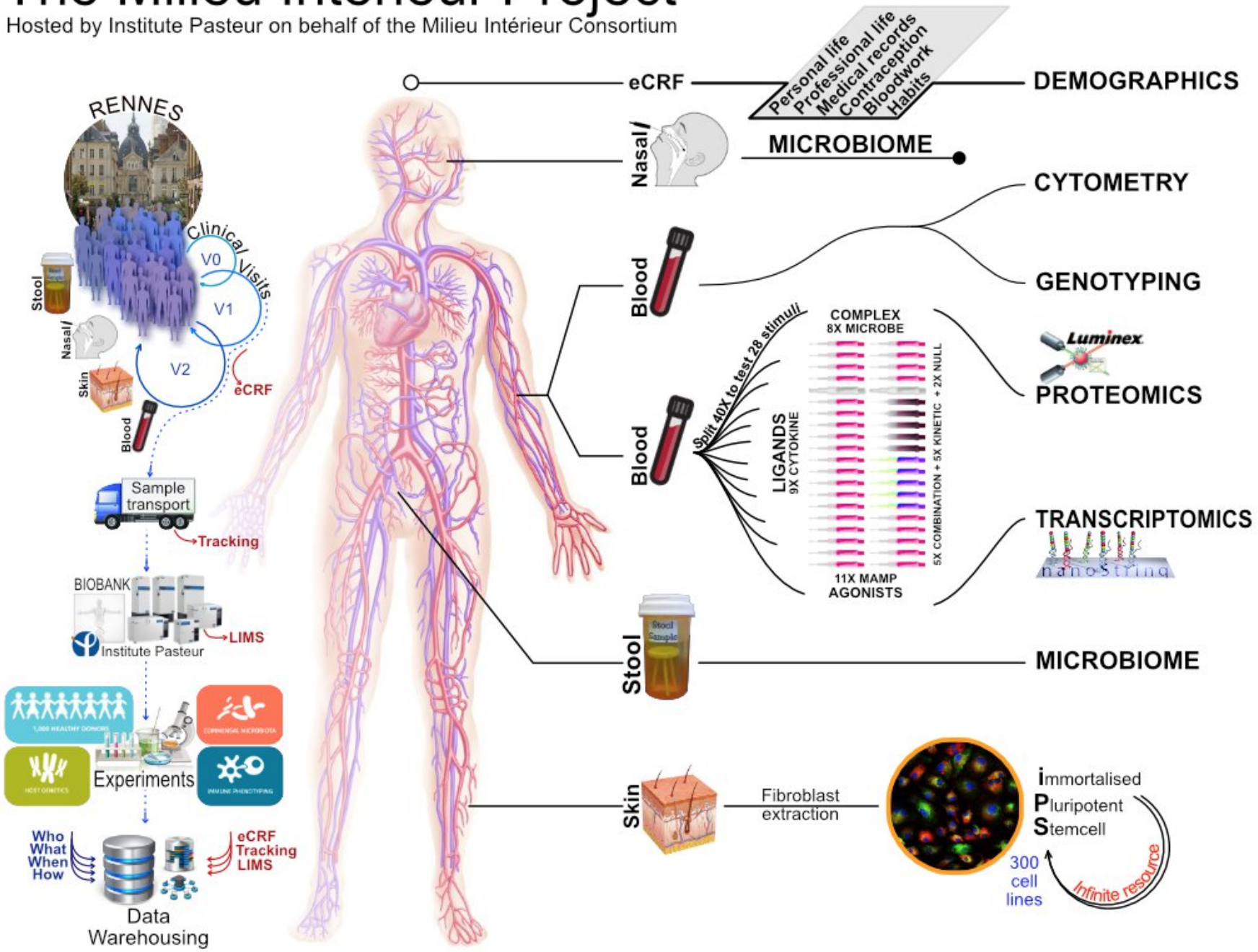
2012 – 2022 (3 visits)

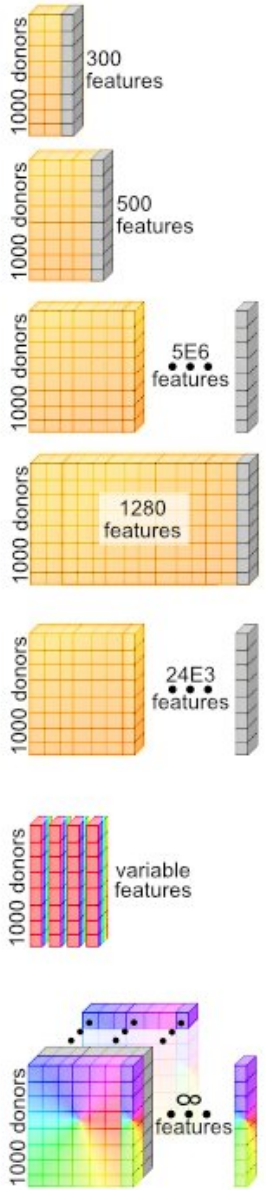
Objectives:

Study the determinants of the variability of a healthy immune response

The Milieu Intérieur Project

Hosted by Institute Pasteur on behalf of the Milieu Intérieur Consortium





Healthy cohort

1000 donors

50% Women – 50% Men
20yr < Age < 70yr

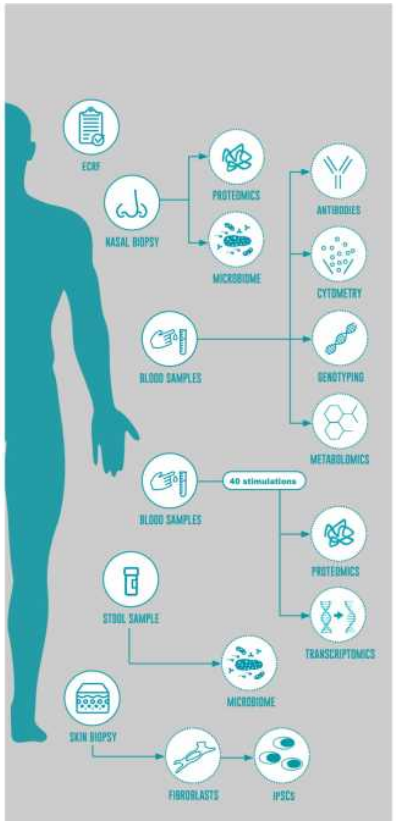
2012 – 2022 (3 visits)

Objectives:

Study the determinants of the variability of a healthy immune response

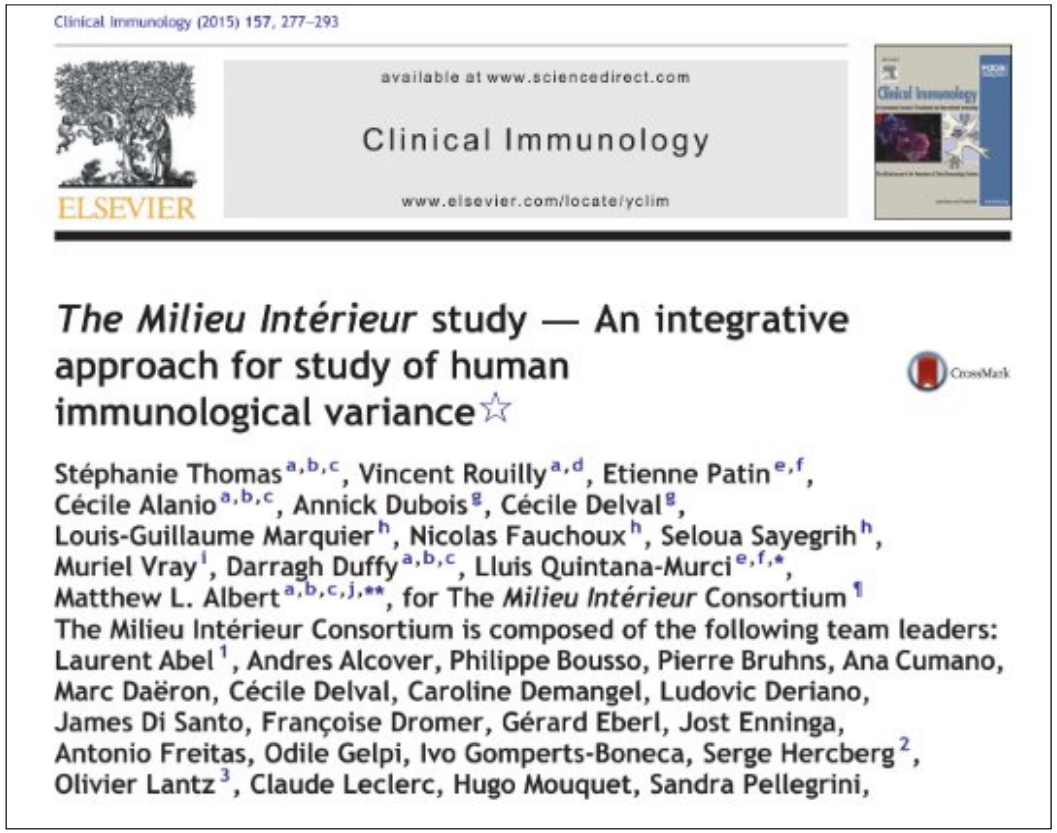
MILIEU INTÉRIEUR DATASETS FOR OUR WORKSHOP

Milieu Intérieur | INSTITUT PASTEUR



- eCRF variables
- TruCulture Luminex dataset
- TruCulture Nanostring dataset
- Flow Cytometry dataset

MI DATASETS: ECRF



Clinical Immunology (2015) 157, 277–293

available at www.sciencedirect.com

Clinical Immunology

www.elsevier.com/locate/yclim

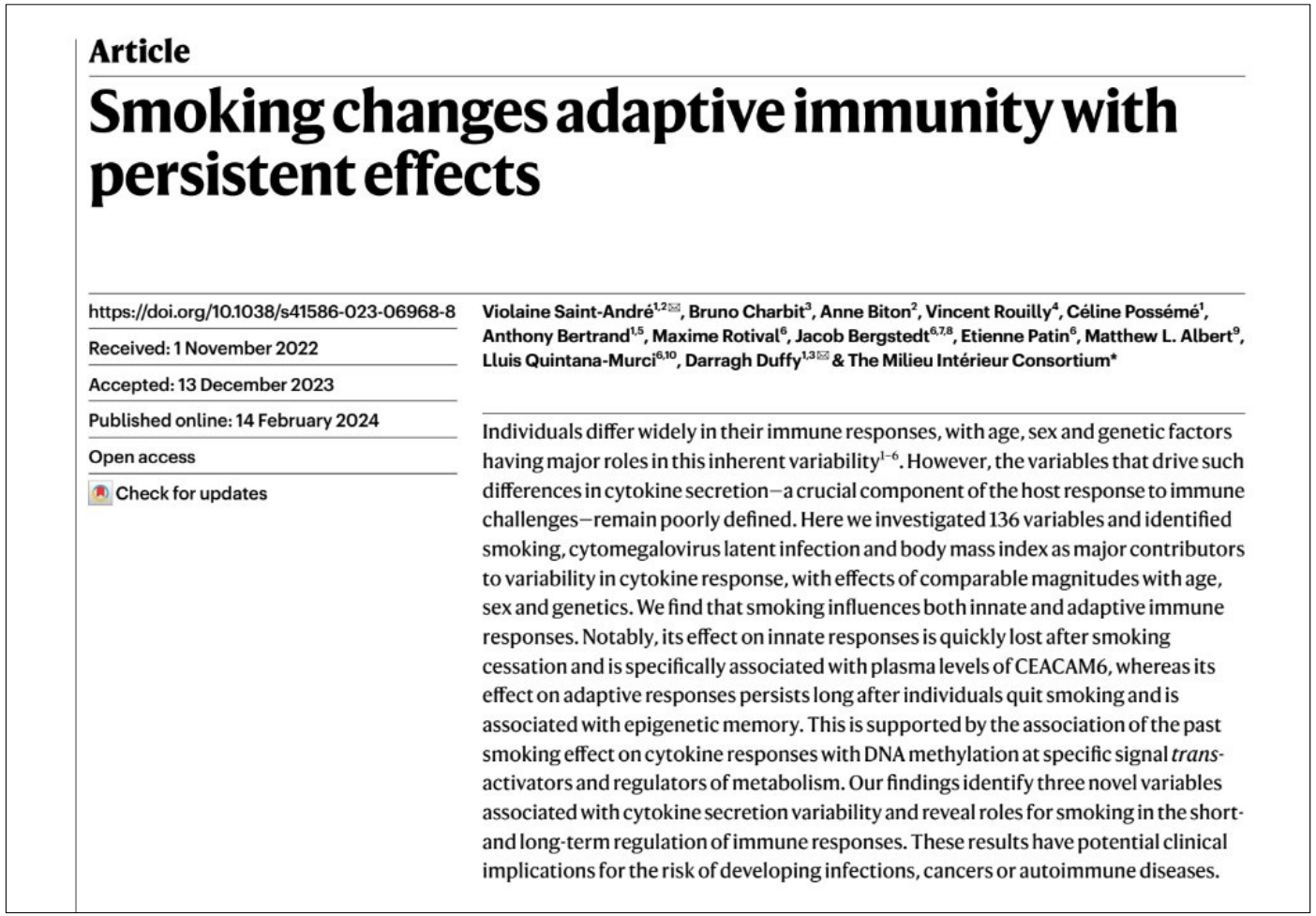
The Milieu Intérieur study — An integrative approach for study of human immunological variance ☆

Stéphanie Thomas[a,b,c], Vincent Rouilly[a,d], Etienne Patin[e,f], Cécile Alanio[a,b,c], Annick Dubois[g], Cécile Delval[g], Louis-Guillaume Marquier[h], Nicolas Fauchoux[h], Seloua Sayegrih[h], Muriel Vray[i], Darragh Duffy[a,b,c], Lluis Quintana-Murci[e,f,*], Matthew L. Albert[a,b,c,j,**], for The *Milieu Intérieur* Consortium[¶]

The Milieu Intérieur Consortium is composed of the following team leaders: Laurent Abel[1], Andres Alcover, Philippe Bousso, Pierre Bruhns, Ana Cumano, Marc Daëron, Cécile Delval, Caroline Demangel, Ludovic Deriano, James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga, Antonio Freitas, Odile Gelpi, Ivo Gomperts-Boneca, Serge Hercberg[2], Olivier Lantz[3], Claude Leclerc, Hugo Mouquet, Sandra Pellegrini,

- Age
- Sex
- Smoking
- CMV status
- Vaccination
- Physical Activity
- Sleep
- …

MI DATASETS: TRUCULTURE LUMINEX



Article

Smoking changes adaptive immunity with persistent effects

https://doi.org/10.1038/s41586-023-06968-8

Received: 1 November 2022

Accepted: 13 December 2023

Published online: 14 February 2024

Open access

Check for updates

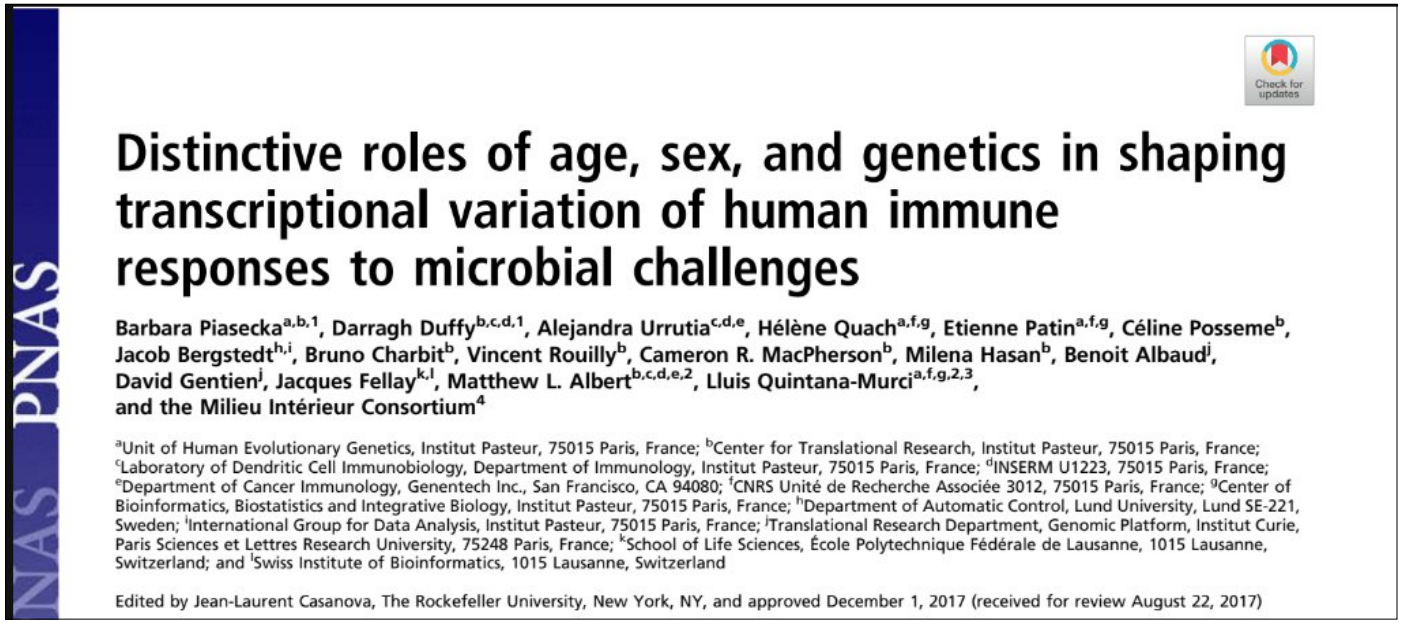
Violaine Saint-André[1,2]✉, Bruno Charbit[3], Anne Biton[2], Vincent Rouilly[4], Céline Possémé[1], Anthony Bertrand[1,5], Maxime Rotival[6], Jacob Bergstedt[6,7,8], Etienne Patin[6], Matthew L. Albert[9], Lluis Quintana-Murci[6,10], Darragh Duffy[1,3]✉ & The Milieu Intérieur Consortium*

Individuals differ widely in their immune responses, with age, sex and genetic factors having major roles in this inherent variability[1–6]. However, the variables that drive such differences in cytokine secretion—a crucial component of the host response to immune challenges—remain poorly defined. Here we investigated 136 variables and identified smoking, cytomegalovirus latent infection and body mass index as major contributors to variability in cytokine response, with effects of comparable magnitudes with age, sex and genetics. We find that smoking influences both innate and adaptive immune responses. Notably, its effect on innate responses is quickly lost after smoking cessation and is specifically associated with plasma levels of CEACAM6, whereas its effect on adaptive responses persists long after individuals quit smoking and is associated with epigenetic memory. This is supported by the association of the past smoking effect on cytokine responses with DNA methylation at specific signal *trans*-activators and regulators of metabolism. Our findings identify three novel variables associated with cytokine secretion variability and reveal roles for smoking in the short- and long-term regulation of immune responses. These results have potential clinical implications for the risk of developing infections, cancers or autoimmune diseases.

- 816 Donors
- 12 TruCulture Stimulations:
  - `Null, TNFa, IL1b, IFNg, BCG, PolyIC, C.albicans, E.coli, Influenza, CD3+CD28, LPS, SEB`
- 13-Plex:
  - `CXCL5, CSF2, IFNg, IL1b, IL2, IL6, IL8, IL10, IL12p70, IL13, IL17, IL23, TNFa`

Luminex®
xMAP® Technology

MI DATASETS: TRUCULTURE NANOSTRING



PNAS

Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges

Barbara Piasecka[a,b,1], Darragh Duffy[b,c,d,1], Alejandra Urrutia[c,d,e], Hélène Quach[a,f,g], Etienne Patin[a,f,g], Céline Posseme[b], Jacob Bergstedt[h,i], Bruno Charbit[b], Vincent Rouilly[b], Cameron R. MacPherson[b], Milena Hasan[b], Benoit Albaud[j], David Gentien[j], Jacques Fellay[k,l], Matthew L. Albert[b,c,d,e,2], Lluis Quintana-Murci[a,f,g,2,3], and the Milieu Intérieur Consortium[4]

[a]Unit of Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France; [b]Center for Translational Research, Institut Pasteur, 75015 Paris, France; [c]Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, 75015 Paris, France; [d]INSERM U1223, 75015 Paris, France; [e]Department of Cancer Immunology, Genentech Inc., San Francisco, CA 94080; [f]CNRS Unité de Recherche Associée 3012, 75015 Paris, France; [g]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, 75015 Paris, France; [h]Department of Automatic Control, Lund University, Lund SE-221, Sweden; [i]International Group for Data Analysis, Institut Pasteur, 75015 Paris, France; [j]Translational Research Department, Genomic Platform, Institut Curie, Paris Sciences et Lettres Research University, 75248 Paris, France; [k]School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; and [l]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Edited by Jean-Laurent Casanova, The Rockefeller University, New York, NY, and approved December 1, 2017 (received for review August 22, 2017)

- 638 Donors
- 7 TruCulture Stimulations:
  - `Null, BCG, C.albicans, E.coli, IAV, S.aureus, SEB`
- 560 genes

nanoString®

MI DATASETS: FLOW CYTOMETRY

nature immunology

RESOURCE

https://doi.org/10.1038/s41590-018-0049-7

Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors

Etienne Patin[1,2,3,26*], Milena Hasan[4,26], Jacob Bergstedt[5,6,26], Vincent Rouilly[3,4], Valentina Libri[4], Alejandra Urrutia[4,7,8,9], Cécile Alanio[4,7,8], Petar Scepanovic[10,11], Christian Hammer[10,11], Friederike Jönsson[12,13], Benoît Beitz[4], Hélène Quach[1,2,3], Yoong Wearn Lim[9], Julie Hunkapiller[14], Magge Zepeda[15], Cherie Green[16], Barbara Piasecka[1,2,3,4], Claire Leloup[14], Lars Rogge[4,17], François Huetz[18,19], Isabelle Peguillet[20,21], Olivier Lantz[20,21,22,23], Magnus Fontes[6,24], James P. Santo[4,8,25], Stéphanie Thomas[4,7,8], Jacques Fellay[9,10], Darragh Duffy[4,7,8], Lluís Quintana-Murci[1,2,3,27], Matthew L. Albert[4,7,8,9,27*] and The Milieu Intérieur Consortium[28]

- 816 Donors
- Immune cell populations

OUR LEARNING OBJECTIVES

- Explore the Variability of Healthy Immune Responses using the datasets from the Milieu Intérieur (MI) Project (milieuintérieur.fr).

- Practice your R Skills
  - R/RStudio Good Practices
  - Tidy Data Principles and Tidyverse
  - Data Wrangling and Data Visualization
  - Foundational statistical methods (PCA, NHST, Linear Models)

- Be equipped to continue your path in learning R and Statistics

R WORKSHOP PROGRAM

Times	Tuesday, June 17th.
9am	**Lecture**: Welcome & Workshop Introduction Vincent Rouilly
10am	Break
10.15am	RStudio set-up & Student introduction
12pm	Lunch
1pm	Data Wrangling
2.30pm	Break
2.45pm	Data Wrangling / Data Visualization
4pm	End of day

R WORKSHOP PROGRAM

Times	Tuesday, June 17th.	Wednesday, June 18th.
9am	**Lecture**: Welcome & Workshop Introduction Vincent Rouilly	**Lecture**: Milieu Intérieur Healthy Donor Studies Darragh Duffy
10am	Break	Break
10.15am	RStudio set-up & Student introduction	Data Visualization
12pm	Lunch	Lunch
1pm	Data Wrangling	PCA and Clustering
2.30pm	Break	Break
2.45pm	Data Wrangling / Data Visualization	PCA and Clustering
4pm	End of day	End of day

R WORKSHOP PROGRAM

Times	Tuesday, June 17th.	Wednesday, June 18th.	Thursday, June 19th
9am	**Lecture**: Welcome & Workshop Introduction Vincent Rouilly	**Lecture**: Milieu Intérieur Healthy Donor Studies Darragh Duffy	**Lecture**: Comparing Immune Phenotypes Across Cohorts Etienne Villain
10am	Break	Break	Break
10.15am	RStudio set-up & Student introduction	Data Visualization	PCA and Clustering
12pm	Lunch	Lunch	Lunch
1pm	Data Wrangling	PCA and Clustering	Statistical tests (NHST)
2.30pm	Break	Break	Break
2.45pm	Data Wrangling / Data Visualization	PCA and Clustering	Statistical tests (NHST)
4pm	End of day	End of day	End of day

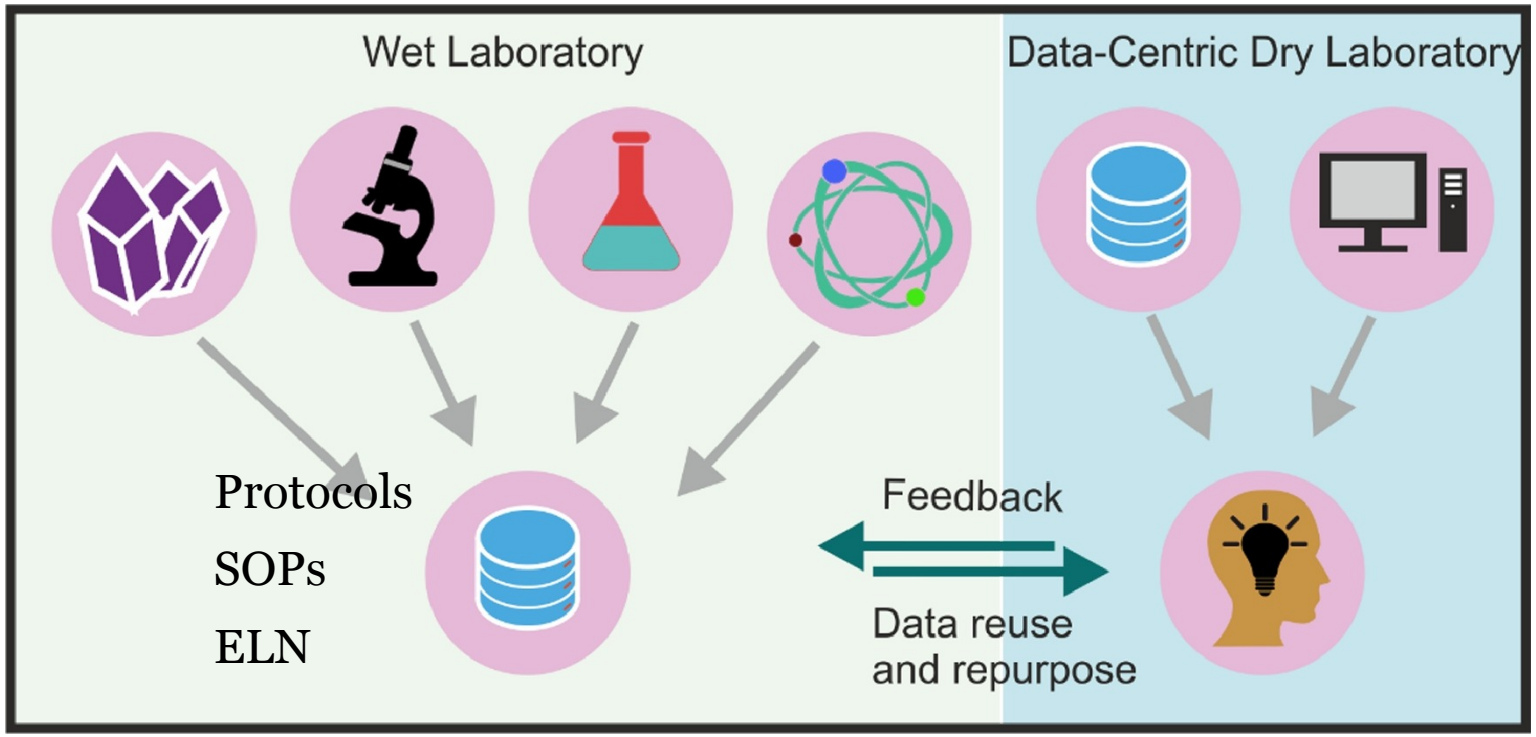
R WORKSHOP PROGRAM

Times	Tuesday, June 17th.	Wednesday, June 18th.	Thursday, June 19th	Friday, June 20th
9am	**Lecture**: Welcome & Workshop Introduction Vincent Rouilly	**Lecture**: Milieu Intérieur Healthy Donor Studies Darragh Duffy	**Lecture**: Comparing Immune Phenotypes Across Cohorts Etienne Villain	**Lecture**: Transcriptional gene regulation & SES effects Anthony Bertrand
10am	Break	Break	Break	Break
10.15am	RStudio set-up & Student introduction	Data Visualization	PCA and Clustering	Linear models
12pm	Lunch	Lunch	Lunch	Lunch
1pm	Data Wrangling	PCA and Clustering	Statistical tests (NHST)	Linear models
2.30pm	Break	Break	Break	Break
2.45pm	Data Wrangling / Data Visualization	PCA and Clustering	Statistical tests (NHST)	Linear models
4pm	End of day	End of day	End of day	End of workshop


Questions?
Comments?

TIDY ANALYSIS PRINCIPLES IN R

WET LAB AND DRY LAB ARE BOTH EXPERIMENTS

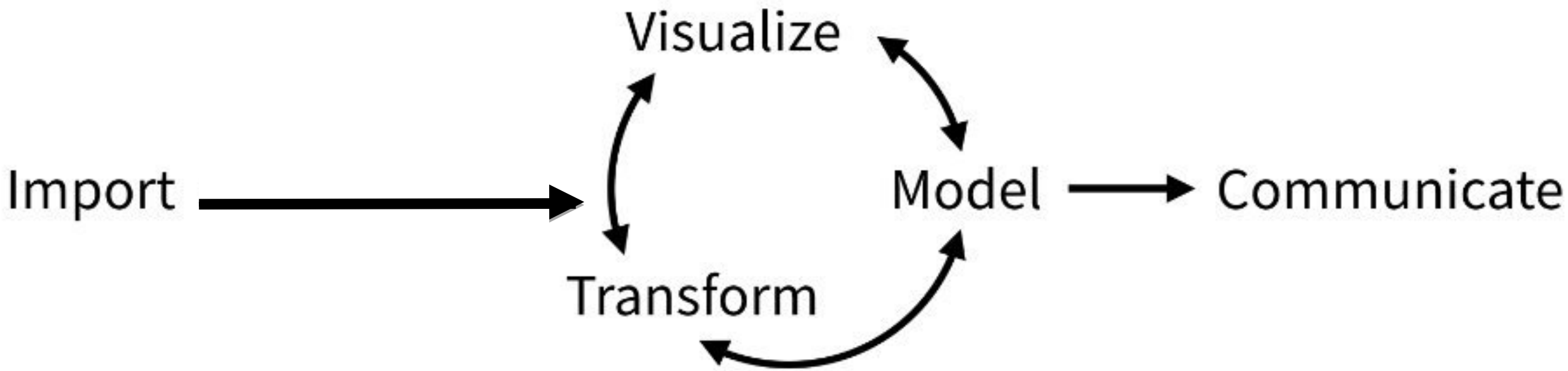


Many possible computational methods to try

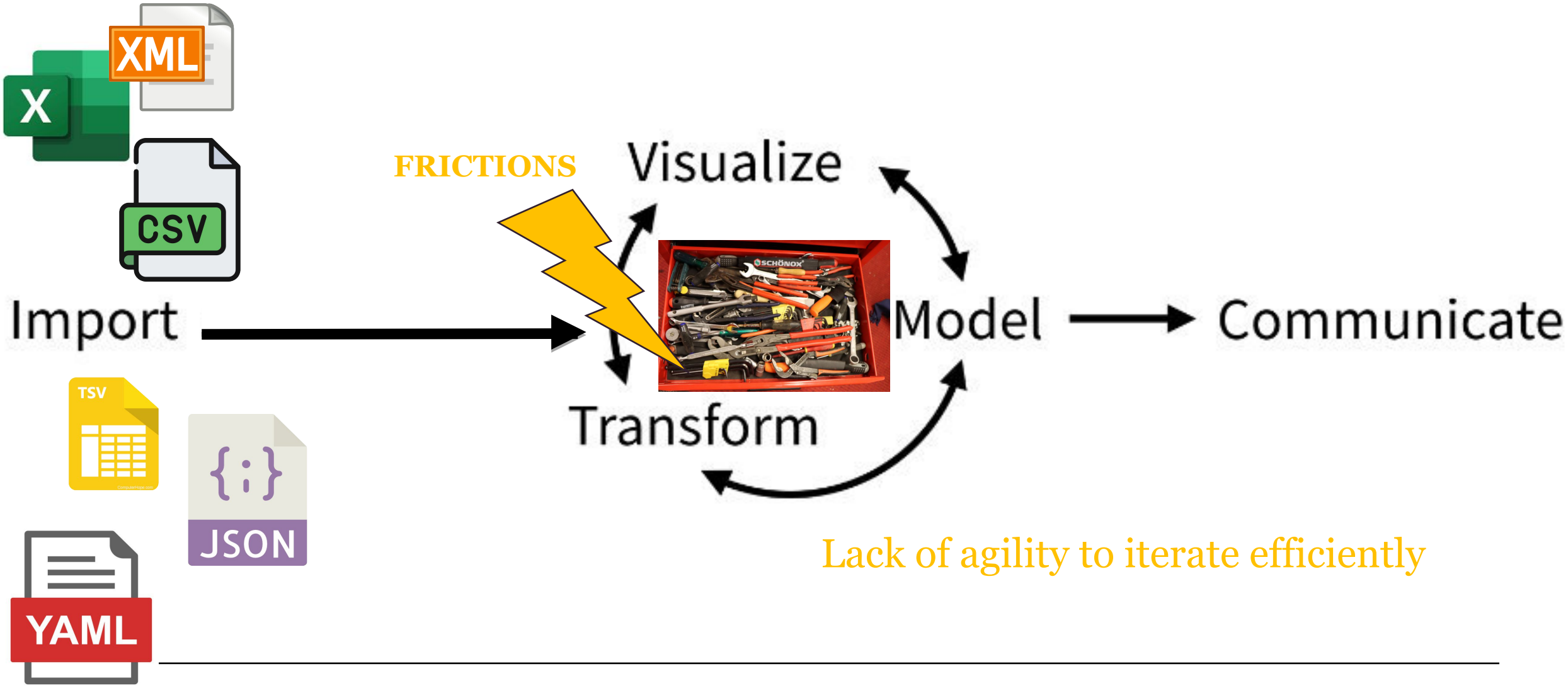
Both need to be **highly iterative**

Both need **Agility**, **Traceability** and **Reproducibility**

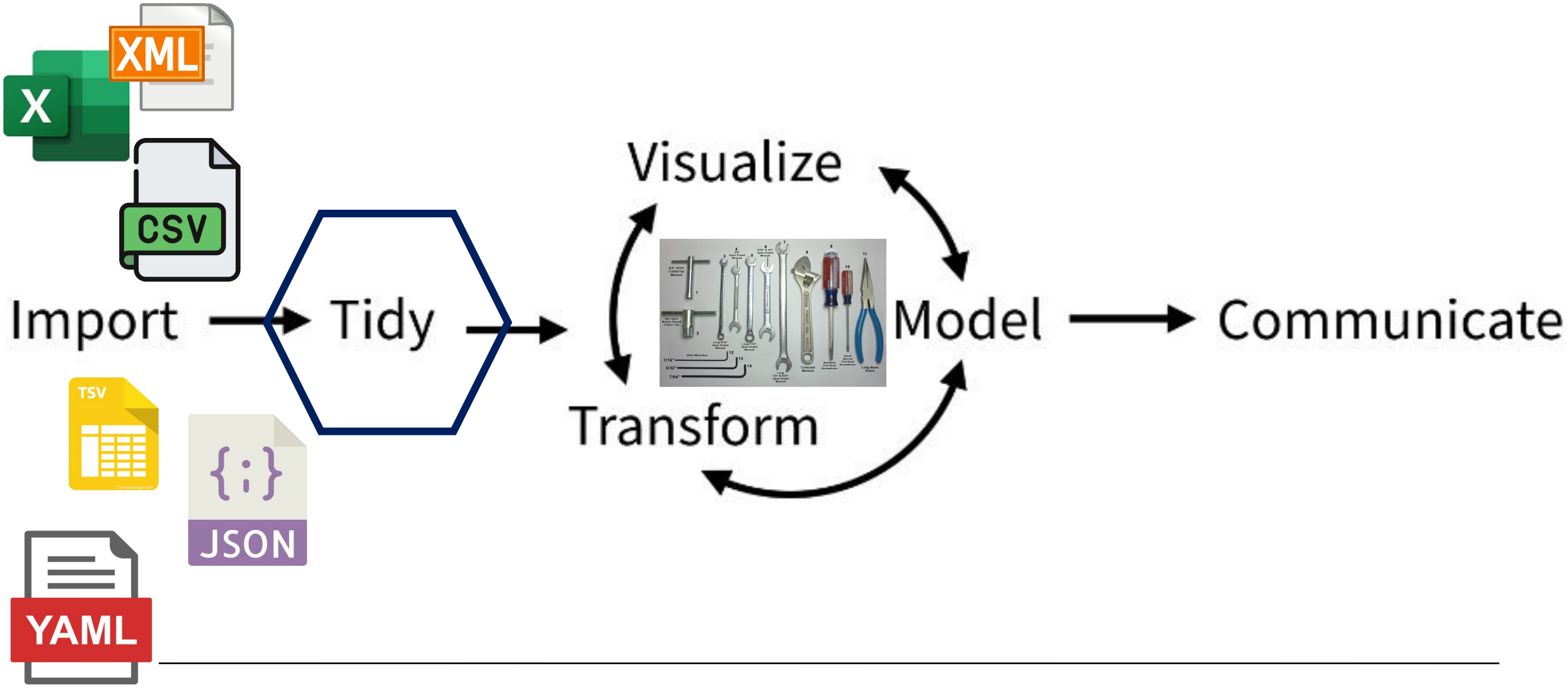
TYPICAL DATA ANALYSIS WORKFLOW



DIVERSE DATA REQUIRES DIVERSE CUSTOM TOOLS



TIDY DATA OPENS OPPORTUNITY FOR STANDARD TOOLS



TIDY PRINCIPLES

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

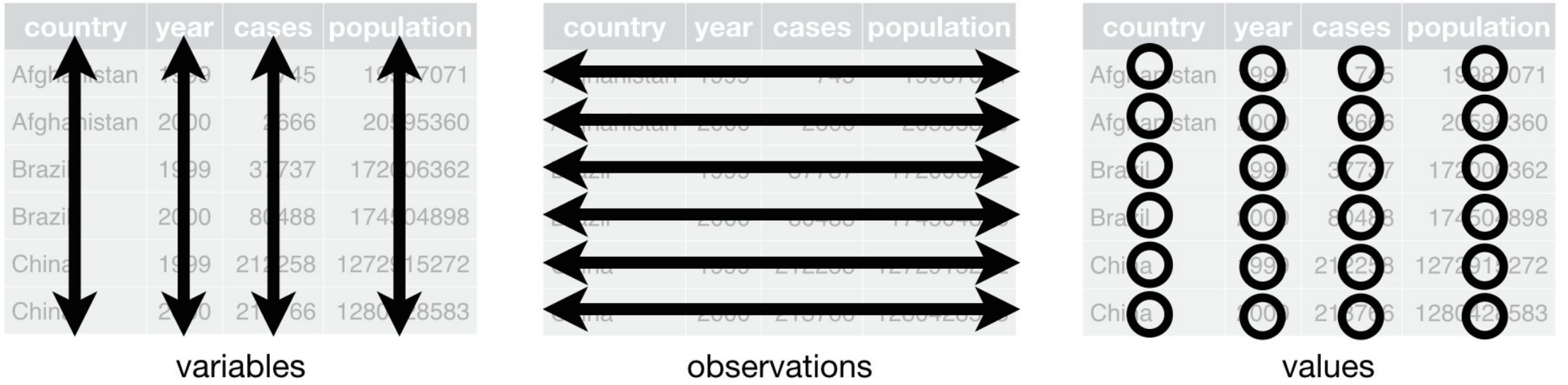
each column a variable

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation

Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

TIDY DATA


country
year
cases
population
variables
observations
values

DATA.FRAME OR TIBBLE AS A STANDARD CONTAINER

```
> flights
Source: local data frame [336,776 x 16]

    year month   day dep_time dep_delay arr_time arr_delay carrier tailnum
   <int> <int> <int>    <int>     <dbl>    <int>     <dbl>   <chr>   <chr>
1   2013     1     1      517         2      830        11      UA  N14228
2   2013     1     1      533         4      850        20      UA  N24211
3   2013     1     1      542         2      923        33      AA  N619AA
4   2013     1     1      54          4       4       -18      B6  N804JB
5   2013     1     1      55          2       2       -25      DL  N668DN
6   2013     1     1      55          0       0        12      UA  N39463
7   2013     1     1      55          3       3        19      B6  N516JB
8   2013     1     1      55          9       9       -14      EV  N829AS
9   2013     1     1      55          8       8        -8      B6  N593JB
10  2013     1     1      558        -2     753         8      AA  N3ALAA
..   ...   ...   ...      ...       ...     ...       ...     ...     ...
Variables not shown: flight <int>, origin <chr>, dest <chr>, air_time
  <dbl>, distance <dbl>, hour <dbl>, minute <dbl>.
```

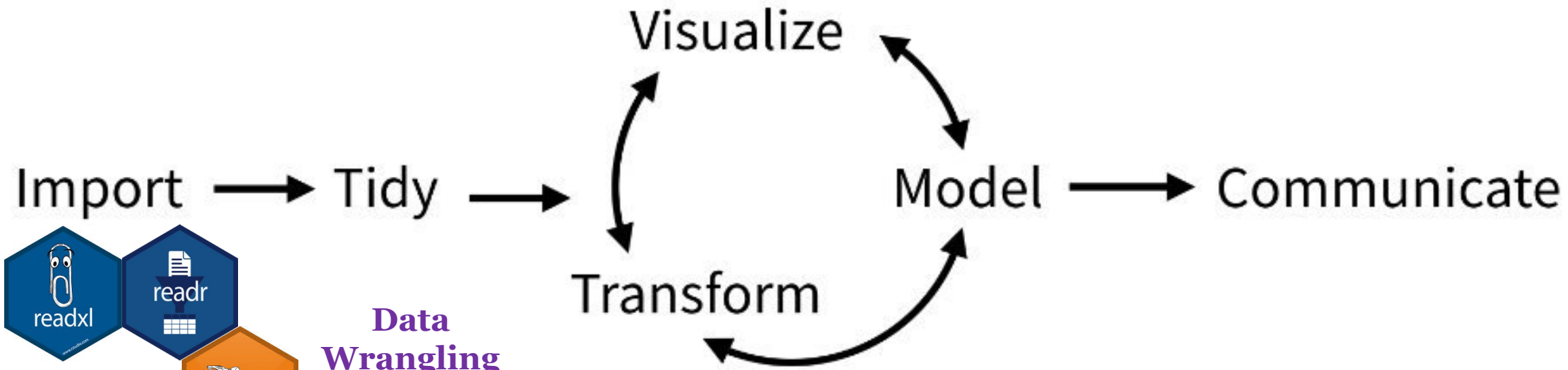
Column types

First 10 rows printed by default (Useful with large data sets)

Columns that don't fit to the screen are not shown

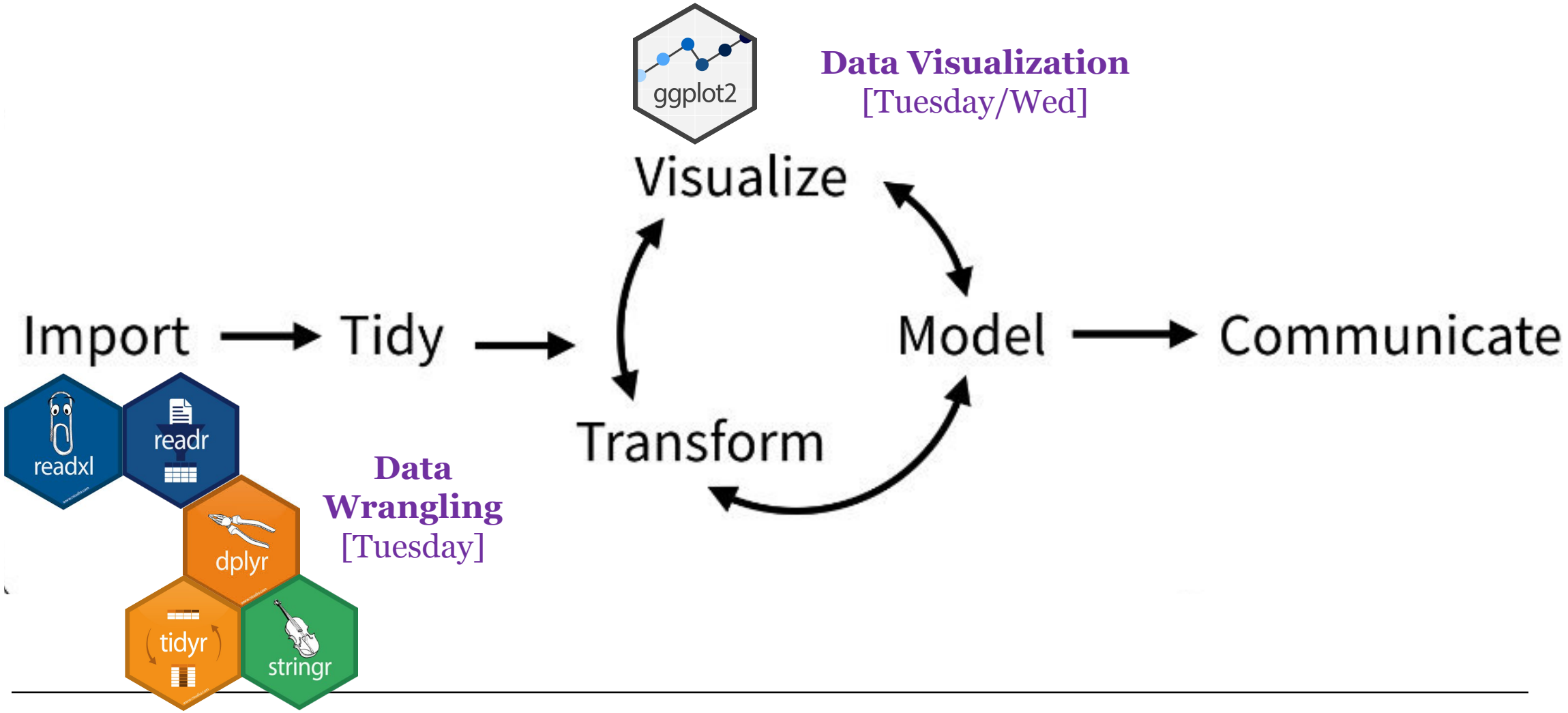
[Tibble vs Data.frame]

TIDY DATA ANALYSIS

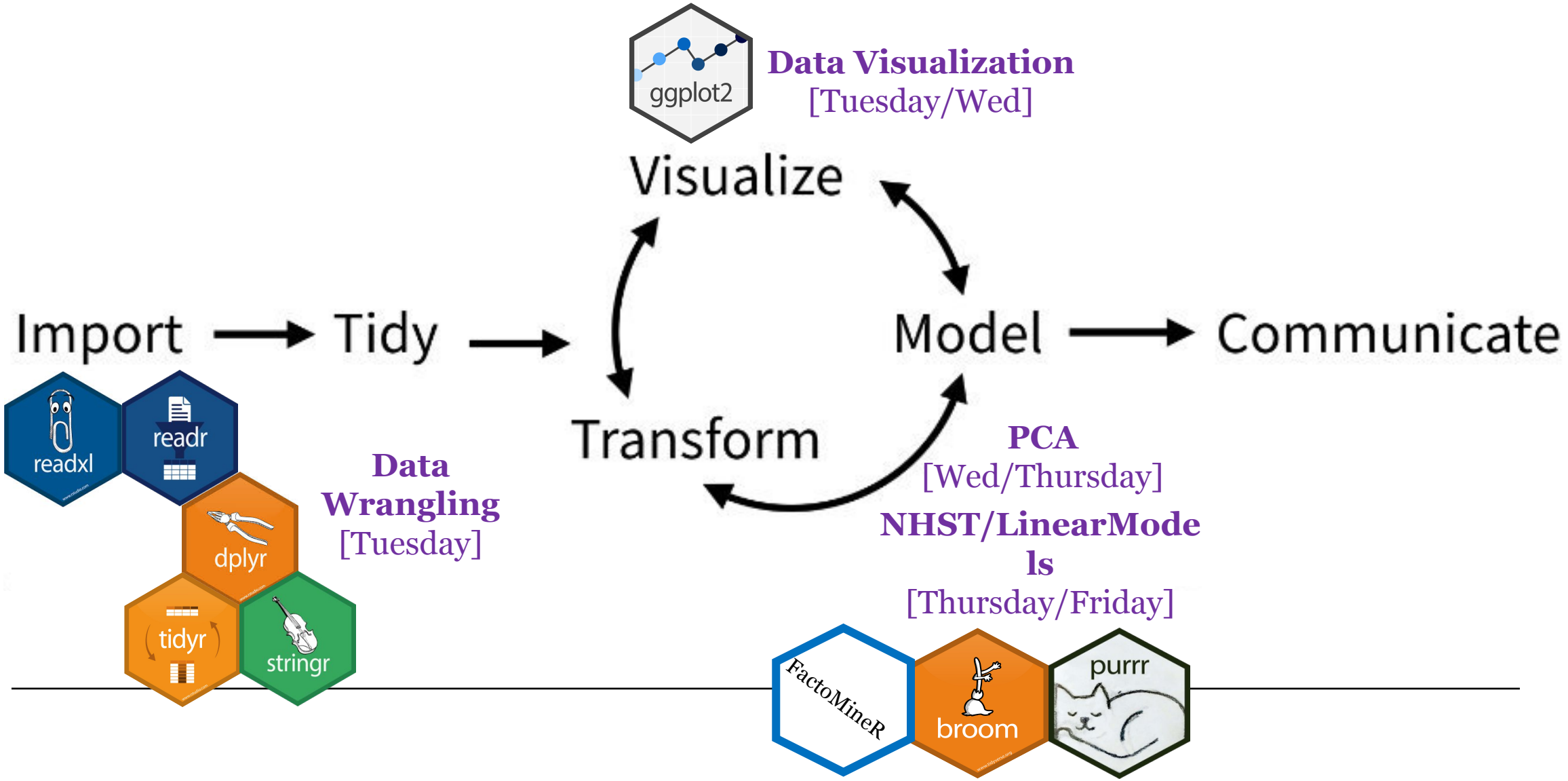


Data Wrangling
[Tuesday]

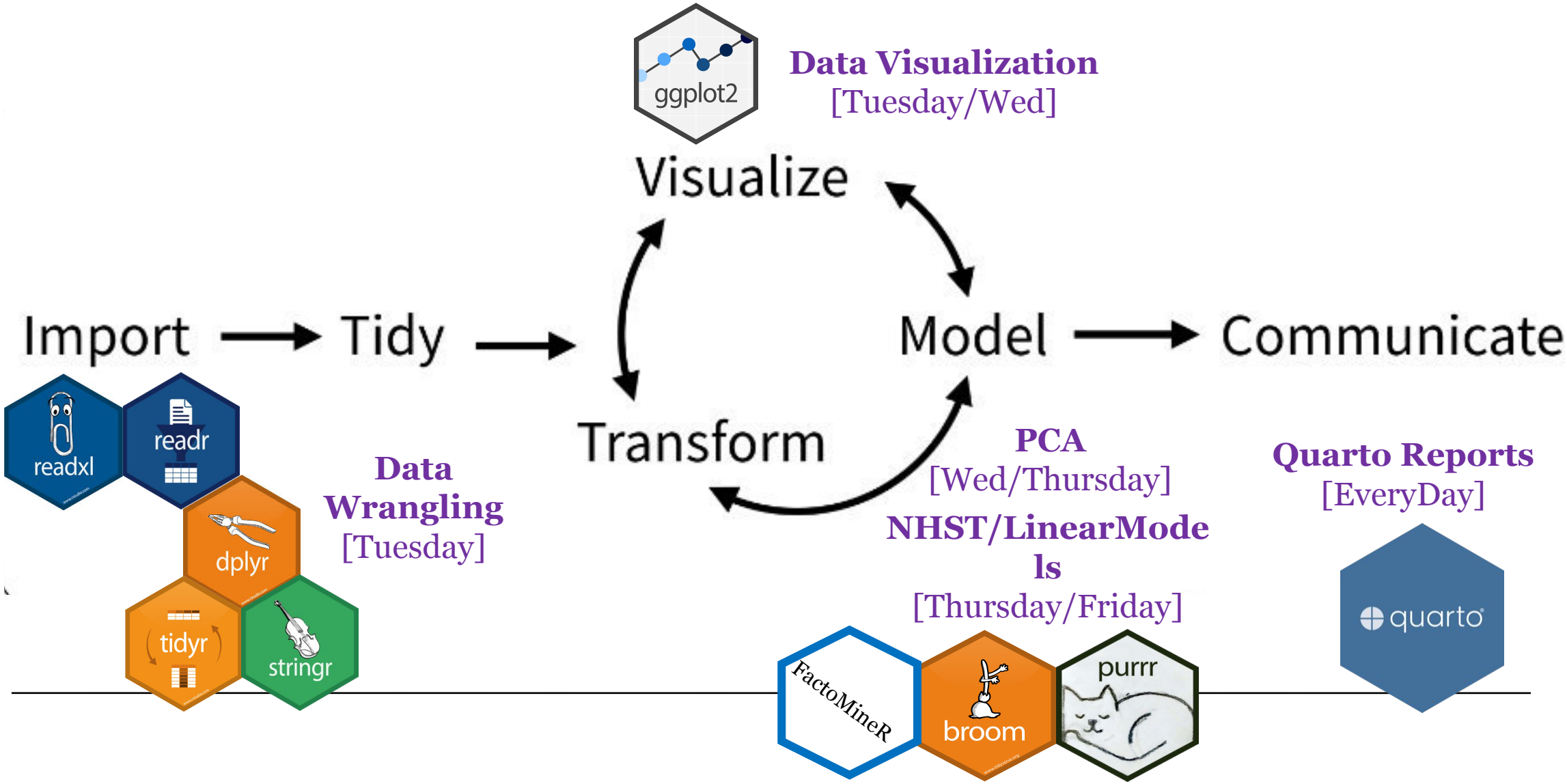
TIDY DATA ANALYSIS



TIDY DATA ANALYSIS



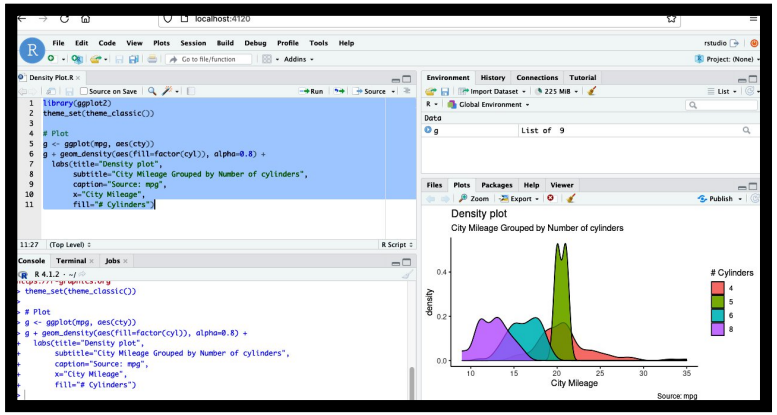
TIDY DATA ANALYSIS



GENERAL WORKSHOP PRINCIPLES

Topic Introduction
Overview

Cookbook
R Recipes
Synthetic Data
Use case



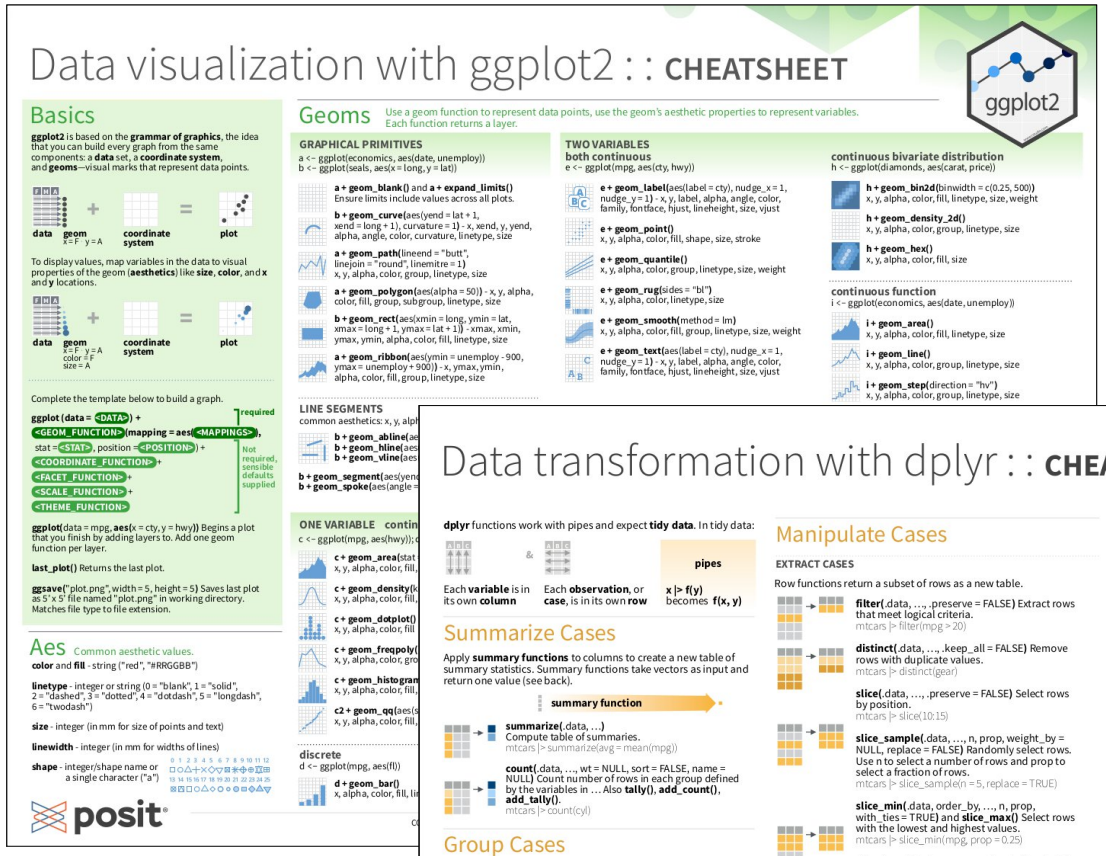
Application
to MI Datasets

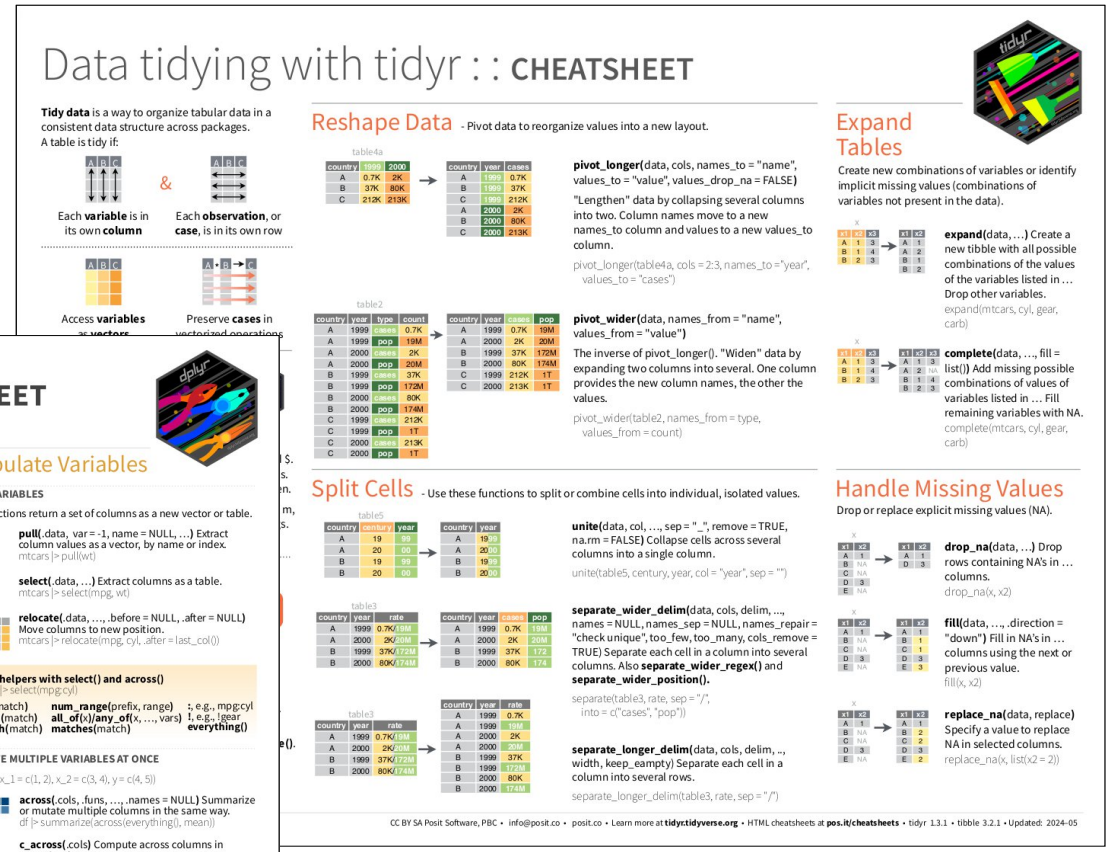
R CHOICES WE MADE FOR THIS WORKSHOP

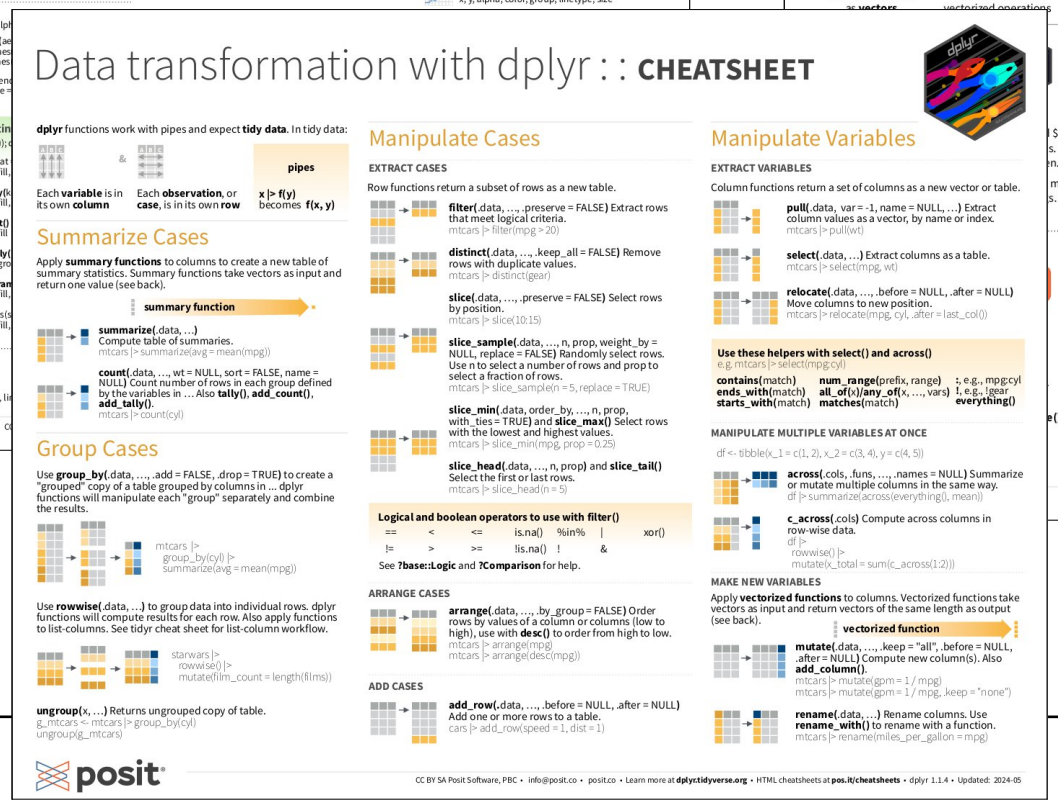
- R Coding-related:
  - Favor Tidyverse packages
  - Extensive use of R Pipes: |> or %>%
  - Package prefixing e.g. dplyr::filter(…)
- Use of Synthetic Data for R examples
- Use of Quarto instead of RMarkdown
- Use of RStudio Cloud instead of a local RStudio

ALWAYS FAVORS TYDIVERSE R PACKAGES

R CHEAT SHEETS



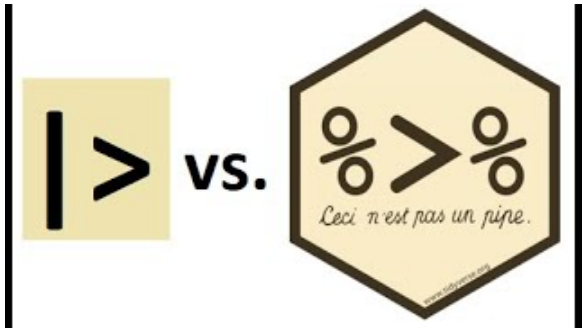




R PIPE APPROACH

Cognitive process:

1. Take the **ydat** dataset, ***then***
2. **filter()** for genes in the leucine biosynthesis pathway, ***then***
3. **group_by()** the limiting nutrient, ***then***
4. **summarize()** to correlate rate and expression, ***then***
5. **mutate()** to round *r* to two digits, ***then***
6. **arrange()** by rounded correlation coefficients



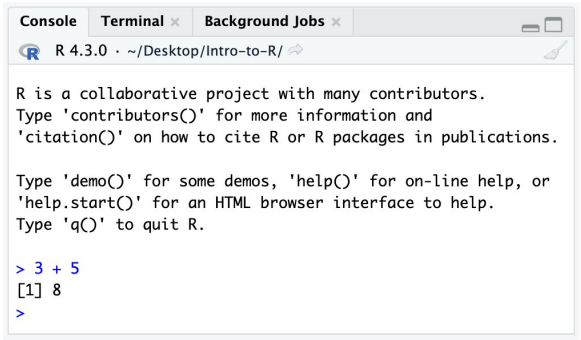
USE OF SYNTHETIC DATASETS

Great for testing

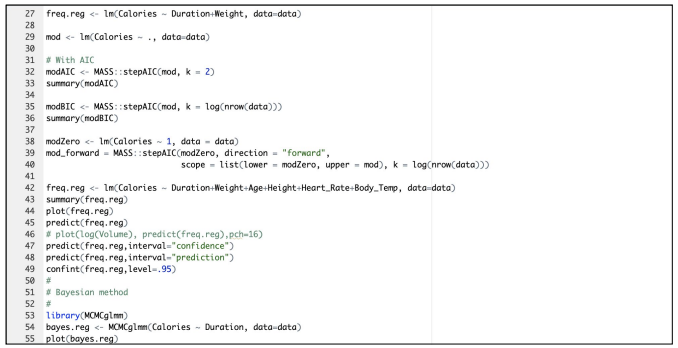
No confidentiality issue

Reproducible

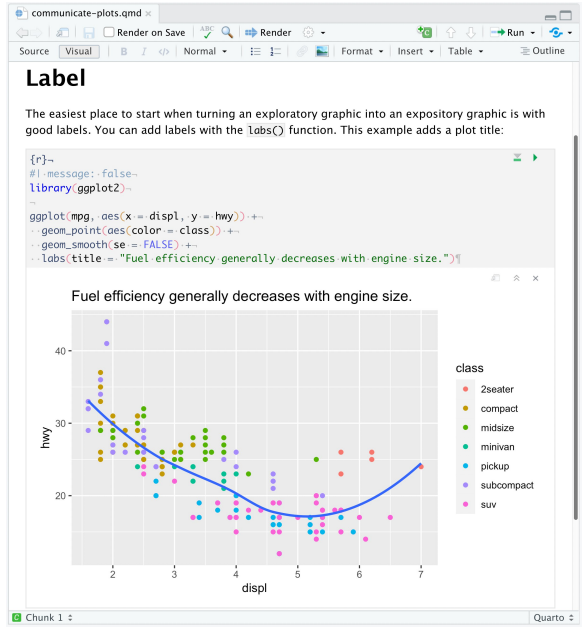
TRACEABILITY, REPRODUCIBILITY AND COMMUNICATION



R Console

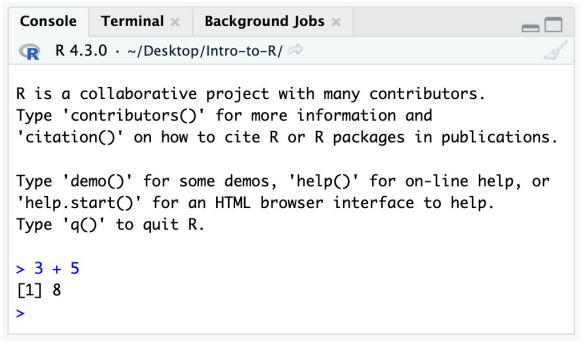


R Script

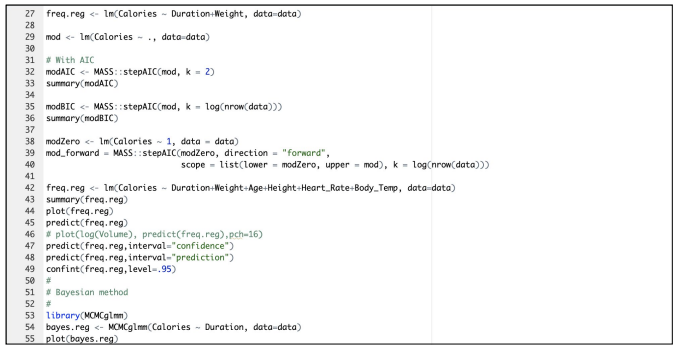


R Report

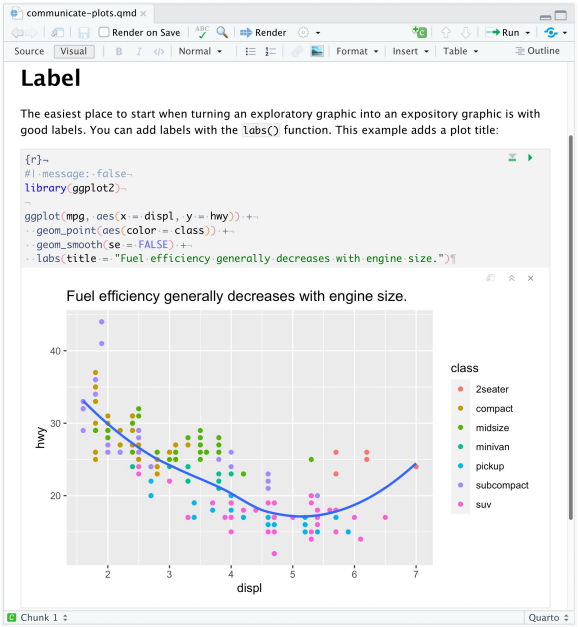
QUARTO REPORTS



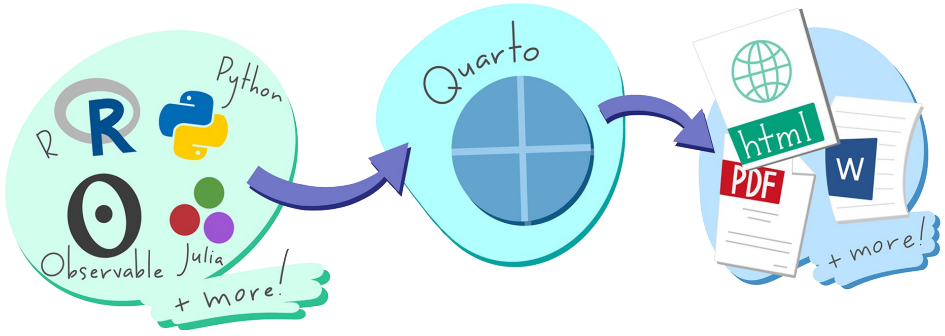
R Console



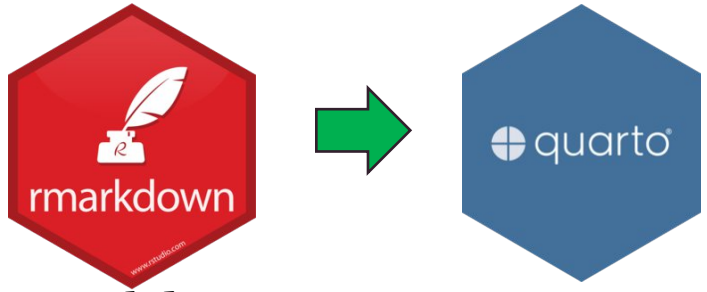
R Script



R Report



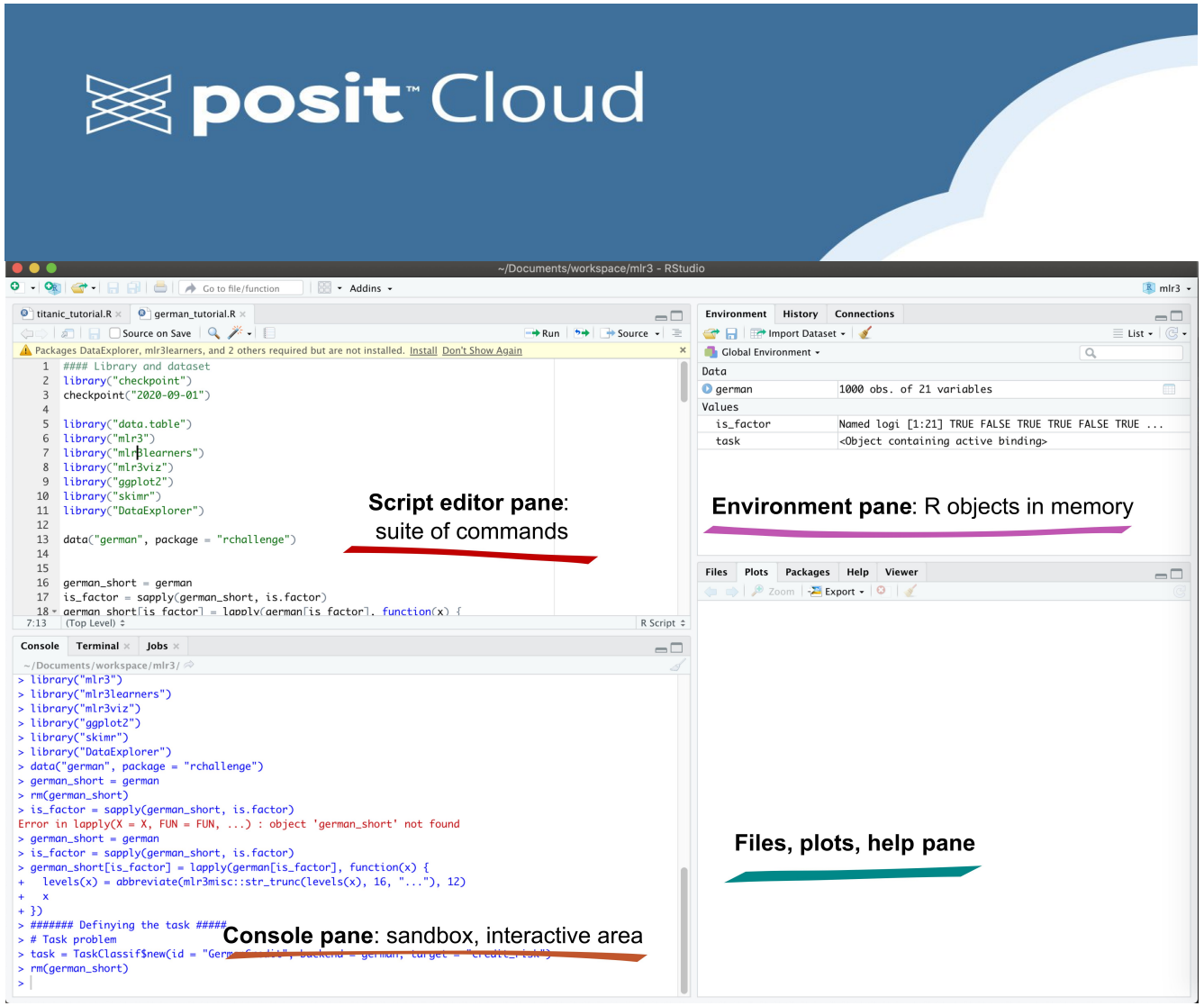
[Quarto Documentation]



RMarkdown

Quarto

R ENVIRONMENT FOR THE WORKSHOP



- Web application
- Uniform R environment for all
- Use of a shared RStudio project template
  - Pre-installed libraries
  - Pre-installed files for workshop
- But you can still download the RStudio project to run locally

R: WHERE TO FIND HELP

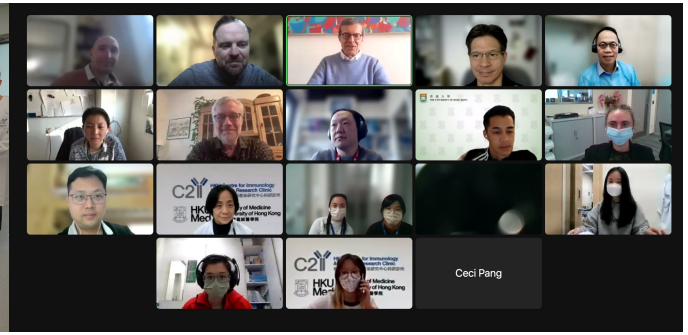
- Your Instructors for this week
- Within the group
- ?function in R console
- R Cheat Sheets
- StackOverflow / Google
- …
- And …

A NEW TOOL IN OUR TOOLBOX

- Have you already used it for R programming ?
  - To explain some code you found
  - To customize some existing code
  - To explain an error message and suggest a fix
  - … other case ?

- What is your favorite prompt for R questions ?

“Art is I, Science is We” Claude Bernard 1865



Translational Immunology Unit
Darragh Duffy
Chloe Albert
Anthony Bertrand
Vincent Bondet
Tom Dott
Florian Dubois
Elizabeth Maloney
Thomas Moreau
Farah Rahal
Marie Robert
Vincent Rouilly
Jamie Sugrue
Violaine Saint-Andre
Divya Unni
Etienne Villain
MC Vougny
Nader Yatim

Human Evolutionary Genetics Unit
Yann Aquino
Etienne Patin
Maxime Rotival
Lluis Quintana-Murci

Innate Immunity Unit
James Di Santo
Jean Marc Doisne
Pedro Goncalves

CBUTechS
Slobodan Culina
Milena Hasan

DARRI
Isabelle Buckle
Omar Choa
Camille Tribout
Lea Vanni

C2i, HK
Wilson Ng
Rex Hung
Rosanna Fong
Raven Cheng
Cynthia Wong
Hebie Chen
Minnie Lo
Max Hui
Jeffrey Yu
Cheska Li
Anna Lee
Hilda On
Garrick Yip
Simon Muller
Amandine Schneider
Peter Chung
Christy Wong
Grace Wu
Theodora Luk
Jimmy Lai
Rachel Telford
Elina Li
Kylie Chang
Raina Mok
Hebronson Yung
Atilla ERISIR
Sonia YOUNAS
Ashley Li

HKU-Pasteur Research Pole, HKU/C2i
Roberto Bruzzone
Malik Peiris
Leo Poon

The University of Hong Kong
Yu Lung Lau
Daniel Leung
Jaime Rosa Duque

FAMILY Cohort, HKU
Gabriel Leung
Michael Ni

Milieu Interieur Consortium
Lluis QUINTANA MURCI
Darragh DUFFY
Matthew AILBERT
Laurent ABEL (Necker)
Andres ALCOVER
Hughues ASCHARD
Philippe BOUSSO
Nollaig BOURKE (TCD)
Petter BRODIN (Karolinksa)
Pierre BRUHNS
Nadeine CERF-BENSUSSAN (Imagine)
Ana CUMANO
Caroline DEMANGEL
Christophe D’ENFERT
Ludovic DERIANO
Marie-Agnes DILLIES
James DISANTO
Gérard EBERL
Jost ENNINGA
Jacques FELLAY (EFL)
Ivo GOMPERTS-BONECA
Milena HASAN
Magnus FONTES (Roche)
Gunilla KARLSSON HEDESTAM (Karolinska)
Serge HERCBERG (Université Paris)
Molly INGERSOLL
Rose ANNE KENNY (TCD)
Olivier LANTZ (Curie)
Frederique MICHEL
Hugo MOUQUET
Cliona O’FARRELLY (TCD)
Eteinne PATIN
Antonio RAUSELL (Imagine)
Lars ROGGE
Anavaj SAKUNTABHAI
Olivier SCHWARTZ
Benno SCHWIKOWSKI
Spencer SHORTE
Fréderic TANGY
Antoine TOUBERT (St Louis)
Mathilde TOUVIER Université Paris)
Marie-Noelle UNGHEUER
Christophe ZIMMER

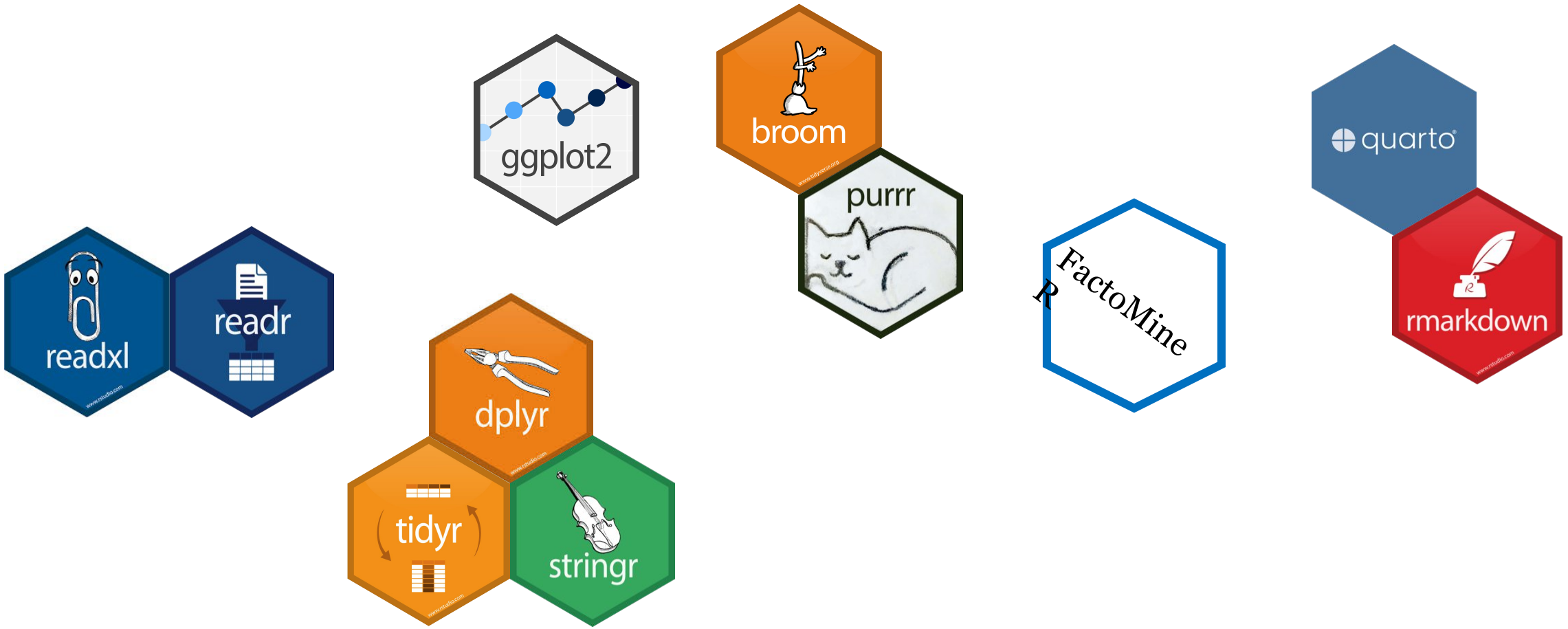
C2i Centre for Immunology & Infection 免疫與感染研究中心

THANK YOU

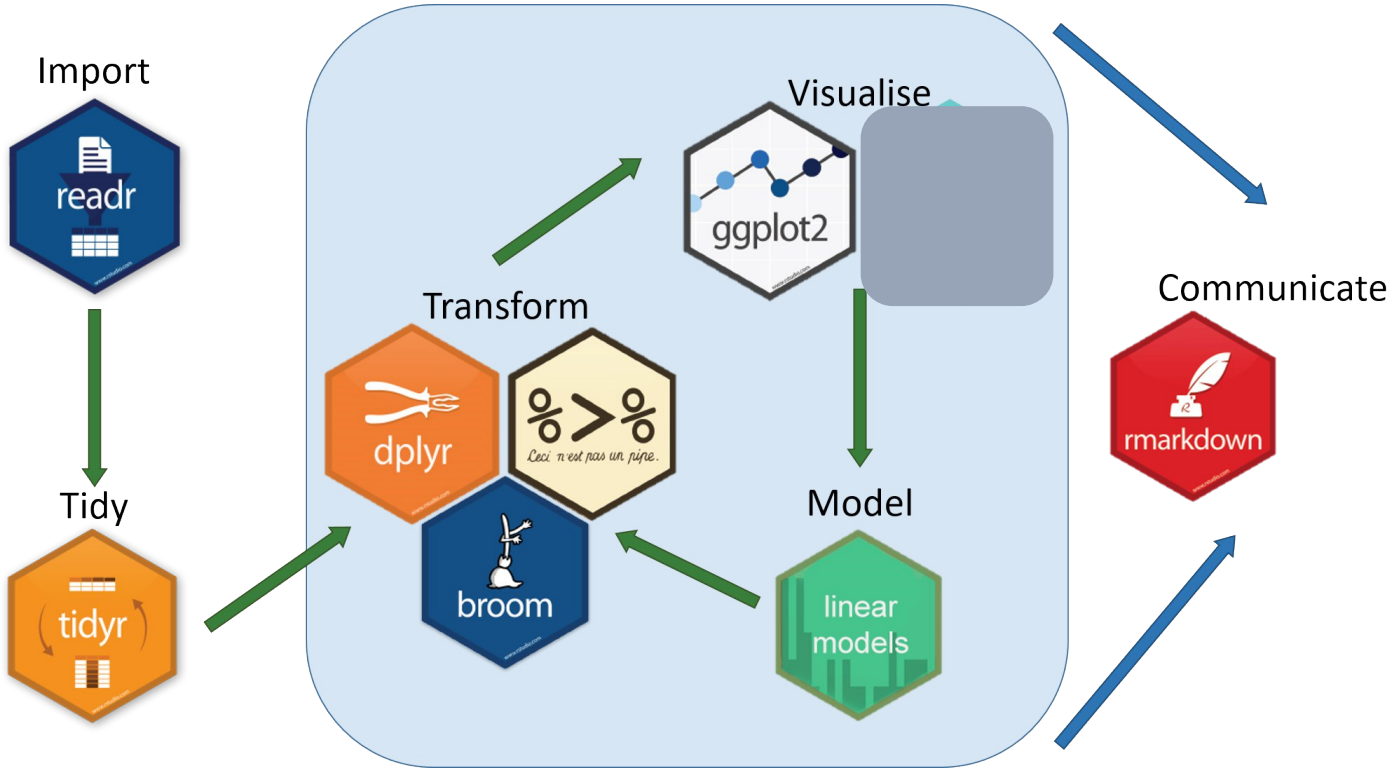
RESOURCES

- R Cheat Sheets
- R Books
  - R for Data Science
  - R Cookbook
  - R Graphics Cookbook
- R Tutorials
  - Carpentries Open Science

R PACKAGES USED IN THIS WORKSHOP



TIDYVERSE


Import
readr
Tidy
tidyr
Transform
dplyr
%>%
Leci n'est pas un pipe.
broom
Visualise
ggplot2
Model
linear models
Communicate
rmarkdown