Data Science in the Pharmaceutical Industry

Driving Innovation through Insights

Dr James Black,
Director, Insights Engineering
Pharma Development Data Sciences

German Data Science Days | 2023-03-09

Table of Contents

Data science across the Roche Group

The diversity of data science across drug development lifecycles

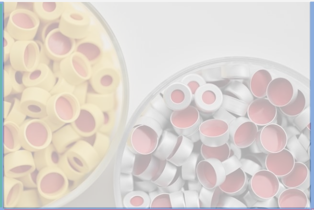
Examples of data science in late stage

A snapshot of data science in our team in 2023

Roche in numbers

1896
Founded in Basel and family still holds majority stake

>63 billion
Sales 2022
(CHF)

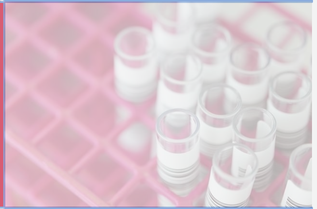


>14 billion
R&D investment
(CHF)

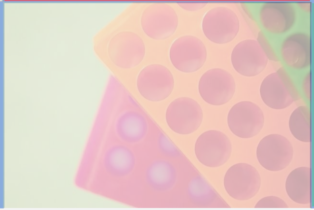
#1
R&D investor in healthcare

14 million
Patients treated

29 billion
Diagnostic tests conducted



87
Drugs under development



103,613
Employees

27
Drug approvals in 2022

Data science across Roche divisions

Drug development

Roche

Genentech
A Member of the Roche Group

Spark THERAPEUTICS

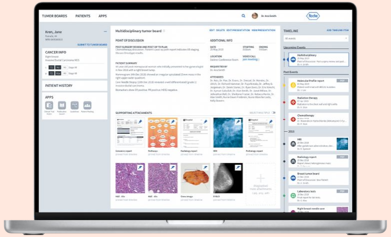
CHUGAI

Diagnostics

FOUNDATION MEDICINE®

Insights business

flatiron.

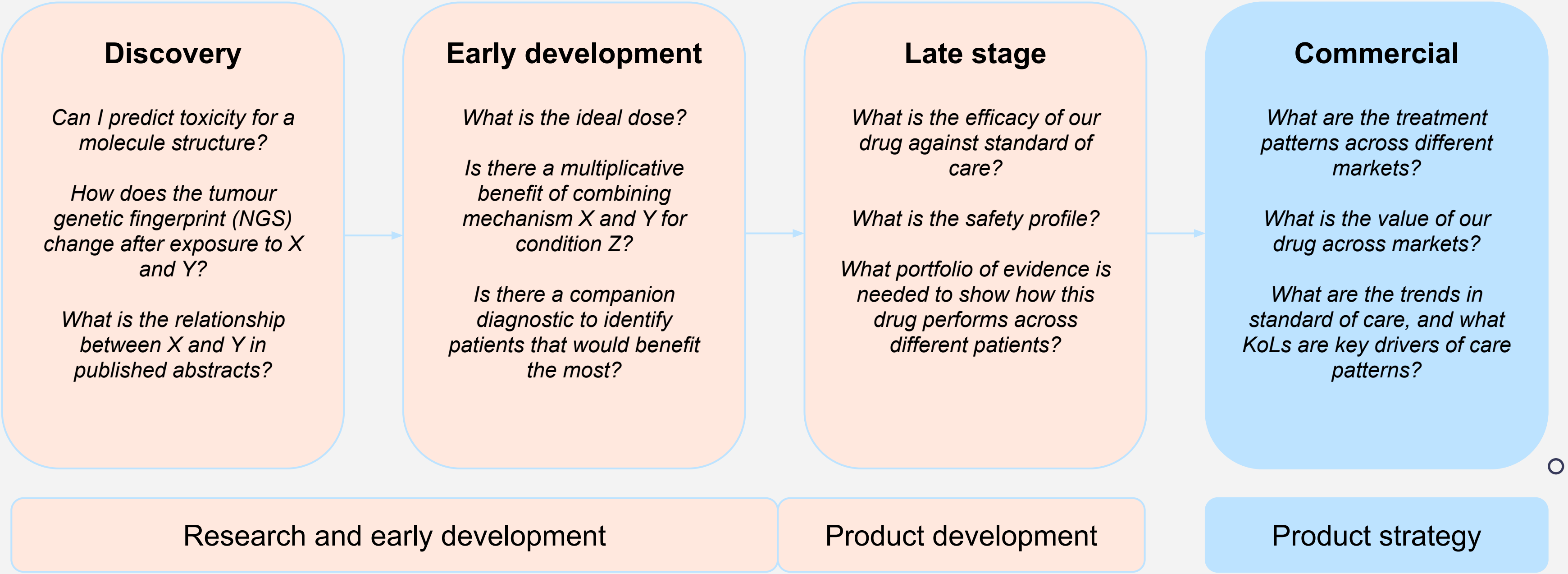


Medicines

Diagnose & SaMD

Aid treatment

The diversity of data science across drug development lifecycles



Discovery

Can I predict toxicity for a molecule structure?

How does the tumour genetic fingerprint (NGS) change after exposure to X and Y?

What is the relationship between X and Y in published abstracts?

Early development

What is the ideal dose?

Is there a multiplicative benefit of combining mechanism X and Y for condition Z?

Is there a companion diagnostic to identify patients that would benefit the most?

Late stage

What is the efficacy of our drug against standard of care?

What is the safety profile?

What portfolio of evidence is needed to show how this drug performs across different patients?

Commercial

What are the treatment patterns across different markets?

What is the value of our drug across markets?

What are the trends in standard of care, and what KoLs are key drivers of care patterns?

Research and early development

Product development

Product strategy

Ex.1: External controls

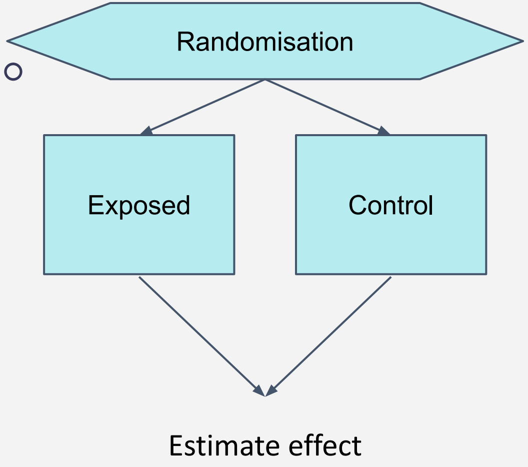
Randomised clinical trials are the gold standard for causal inference, ***but they are slow, expensive and tightly scoped to a specific hypothesis***

Can we leverage the routinely collected data from EHR and claims, across millions of patients, to improve our evidence portfolio?

External controls 101

Randomised clinical trial

Gold standard for treatment effect.



Randomisation

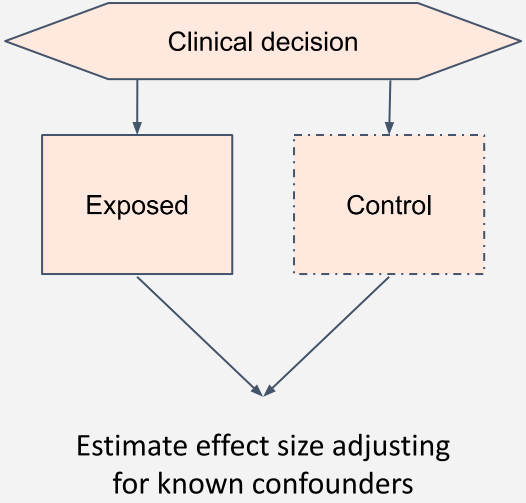
Exposed

Control

Estimate effect

Adjustment

Adjust and estimate effect in one step, e.g. a retrospective cohort study.



Clinical decision

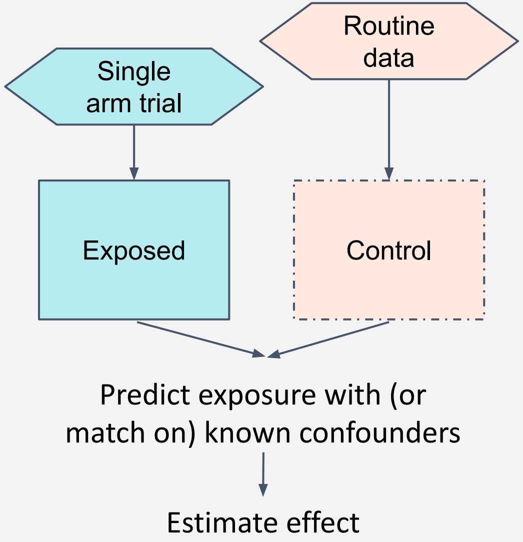
Exposed

Control

Estimate effect size adjusting for known confounders

Matching and weighting

Includes propensity scores.



Single arm trial

Routine data

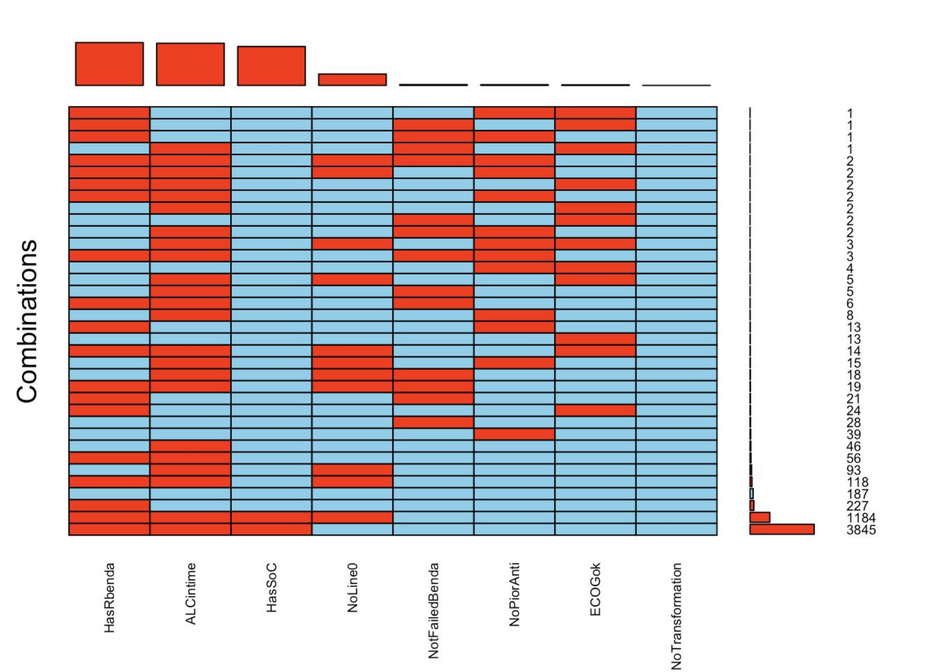
Exposed

Control

Predict exposure with (or match on) known confounders

Estimate effect

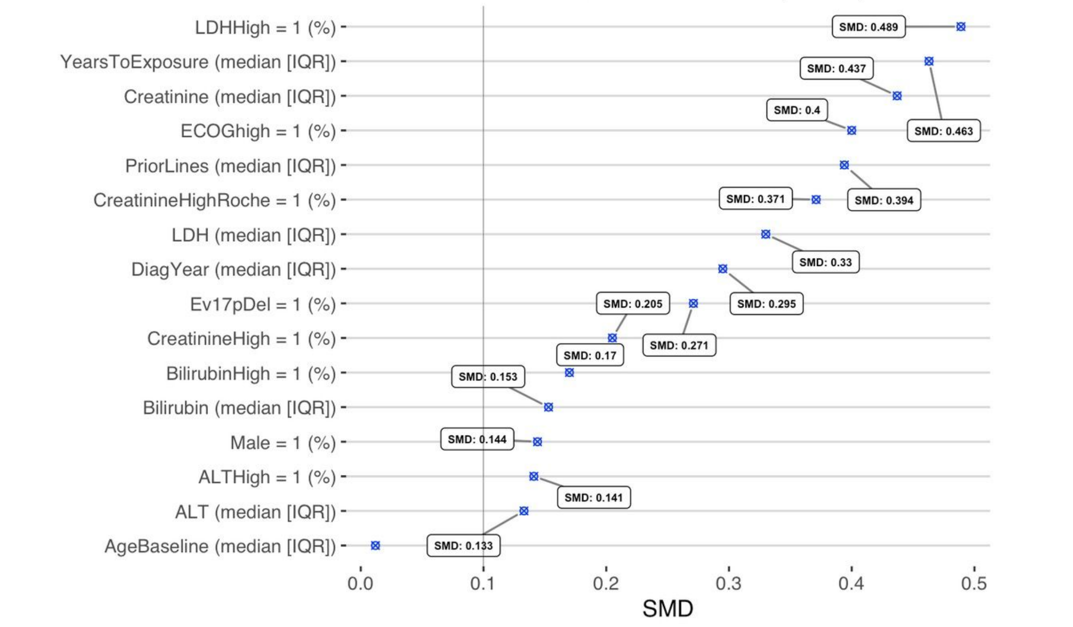
Applying inclusion/exclusion criteria



An example inc/exc attrition for a 2 LoT+ analysis

Starting with 6,012 patients in our real world database with the same disease and >1 LoT, we drop to 187 that meet real world *approximations* of the inclusion and exclusion criteria

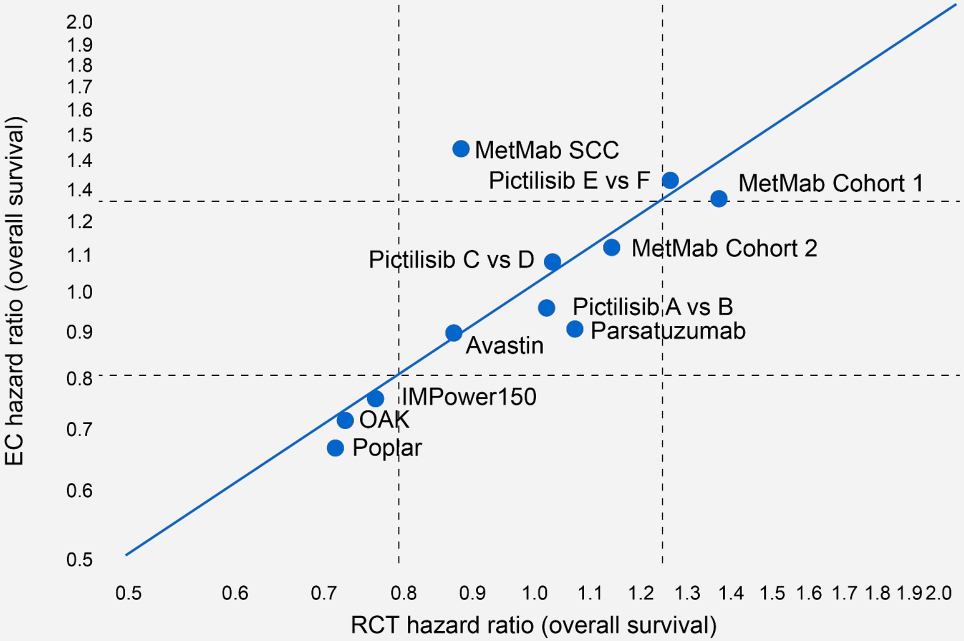
Imbalance pre-analysis



Standardised mean difference of baseline characteristics comparing real world control to single arm trial

Analysing our populations as-is would be likely to lead to bias.

External controls:
an area of active research



Ex.2: The RAAD Challenges

Roche Advanced Analytics Data (RAAD) Challenges

RAAD Challenge 1.0

Predict the probability that a patient will be alive at 1 year after 1L treatment initiation, using all the patient data available up to the start of 1L treatment.

Data:
9,500 patients' across train and test from across seven different cancer types were used to train the models.

Participation:
500+ Roche employees, 132 teams, 28 sites

RAAD Challenge 2.0

Predict response in a new drug and indication combination using training data containing either that drug in a different indication, or that indication with other drugs

Data:
6,000 patients across train and test from historical trials

Participation:
500+ Roche employees, 141 teams, 38 sites

RAAD Challenge 3.0 judging was last Friday!

Data-Centricity: clinico-genomic feature library

Hackathon tasks

With a fixed prediction model, enrich the training data and create 10 features that improve prediction on the test data.

Benefit to Roche & our Patients

This challenge focuses our attention not on the model, but on the preparation of data for modeling, and will help to develop a rich catalog of features enabling future researchers to develop more impactful insights and products (e.g .to support clinical decision support at point of care, enhanced patient segmentation for clinical trials etc).

Data-Centricity: clinico-genomic feature library

NGS test

Vitals

Labs

Demographics

Feature generation

Model

Prediction target is time till death

96,686 patients

118 variables

58,066 events

744 MBs of data

93,673,017 data points

Train - 60%

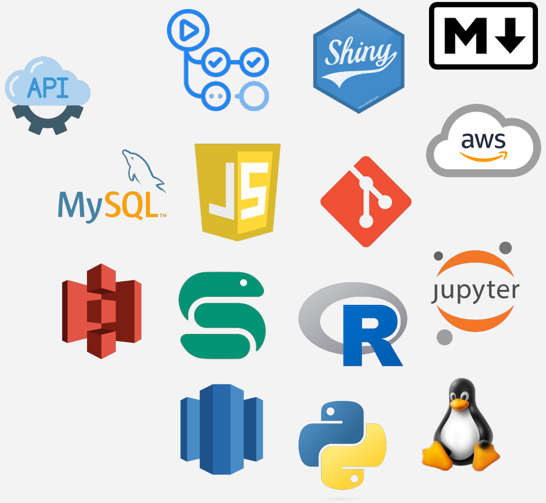
Leaderboard - 20%

Test - 20%

Data science in our team in 2023

Being multilingual is the new-norm

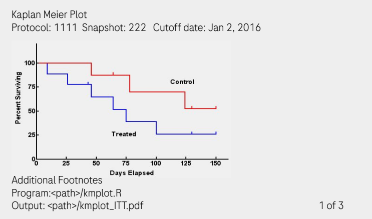
5 years ago



Today

We focus on R-backbone for analytics, and python for tooling, but there is a continued growth in other languages and technologies that we leverage in the pursuit of constantly doing things ‘better’

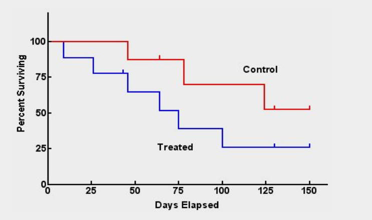
Data products, in addition to static insights, is now the new normal



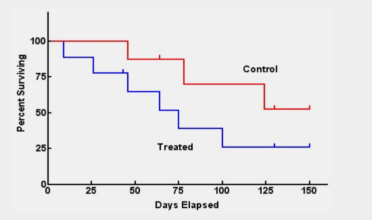
Kaplan Meier Plot
Protocol: 1111 Snapshot: 222 Cutoff date: Jan 2, 2016

Additional Footnotes
Program:<path>/kmplot.R
Output: <path>/kmplot_ITT.pdf
1 of 3

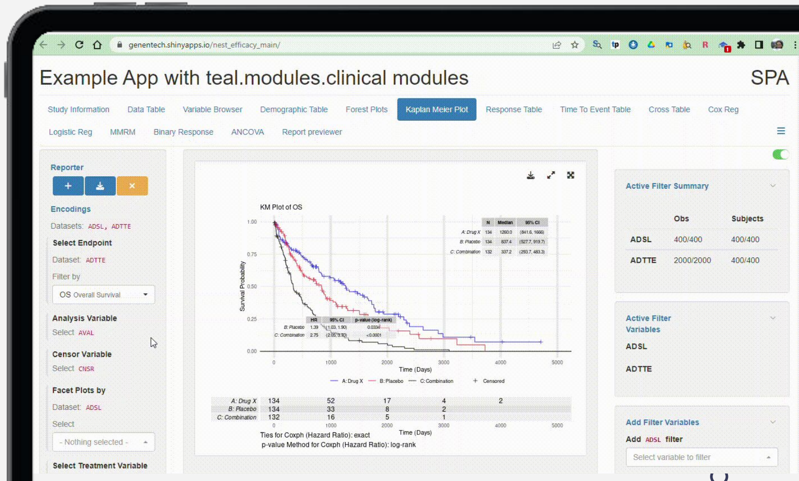
Per-SAP static output on **OS**



Ad-hoc analysis on **PFS**



Ad-hoc analysis on **PFS** in **subpopulation**



genentech.shinyapps.io/nest_efficacy_main/

Example App with teal.modules.clinical modules — SPA

Datasets: ADSL, ADTTE

Select Endpoint
Dataset: ADTTE
Filter by
OS Overall Survival

Analysis Variable
Select AVAL

Censor Variable
Select CNSR

Facet Plots by
Dataset: ADSL
Select
- Nothing selected -

Select Treatment Variable

KM Plot of OS

	N	Median	95% CI
A: Drug X	134	1260.0	(841.6, 1666)
B: Placebo	134	837.4	(527.7, 919.7)
C: Combination	132	337.2	(293.7, 483.3)

	HR	95% CI	p-value (log-rank)
B: Placebo	1.39	(1.03, 1.90)	0.0334
C: Combination	2.75	(2.05, 3.70)	<0.0001

	0	1000	2000	3000	4000	5000
A: Drug X	134	52	17	4	2	
B: Placebo	134	33	8	2		
C: Combination	132	16	5	1		

Time (Days)

Ties for Coxph (Hazard Ratio): exact
p-value Method for Coxph (Hazard Ratio): log-rank

Active Filter Summary

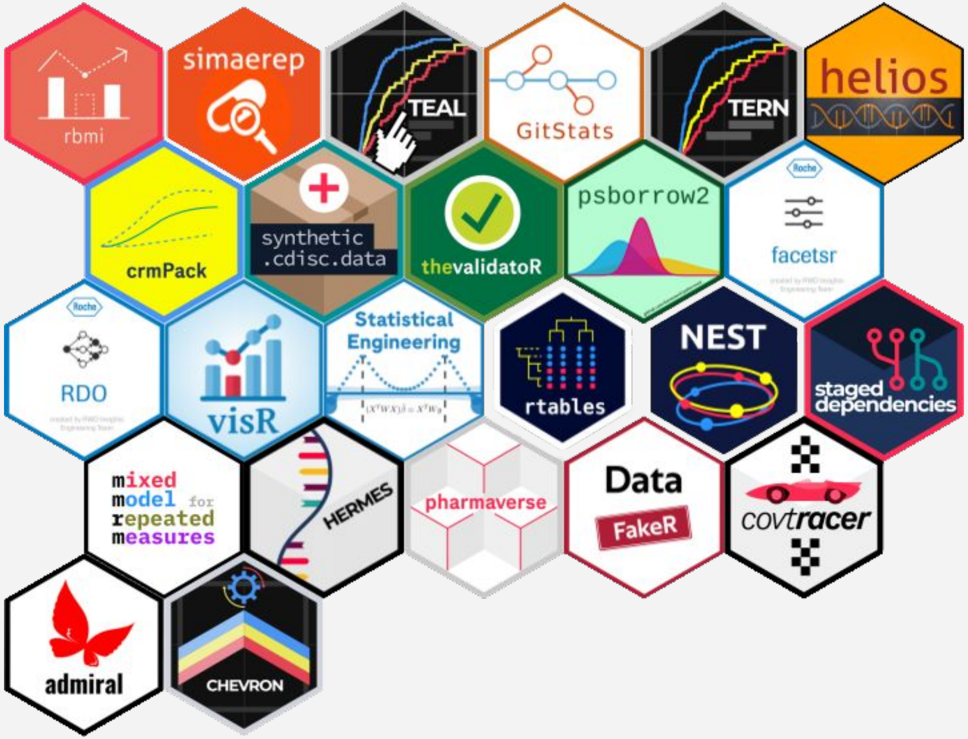
	Obs	Subjects
ADSL	400/400	400/400
ADTTE	2000/2000	400/400

Active Filter Variables
ADSL
ADTTE

Add Filter Variables
Add ADSL filter
Select variable to filter

Self-service data products to enable rapid exploration

We have more of a focus on re-use, and open source



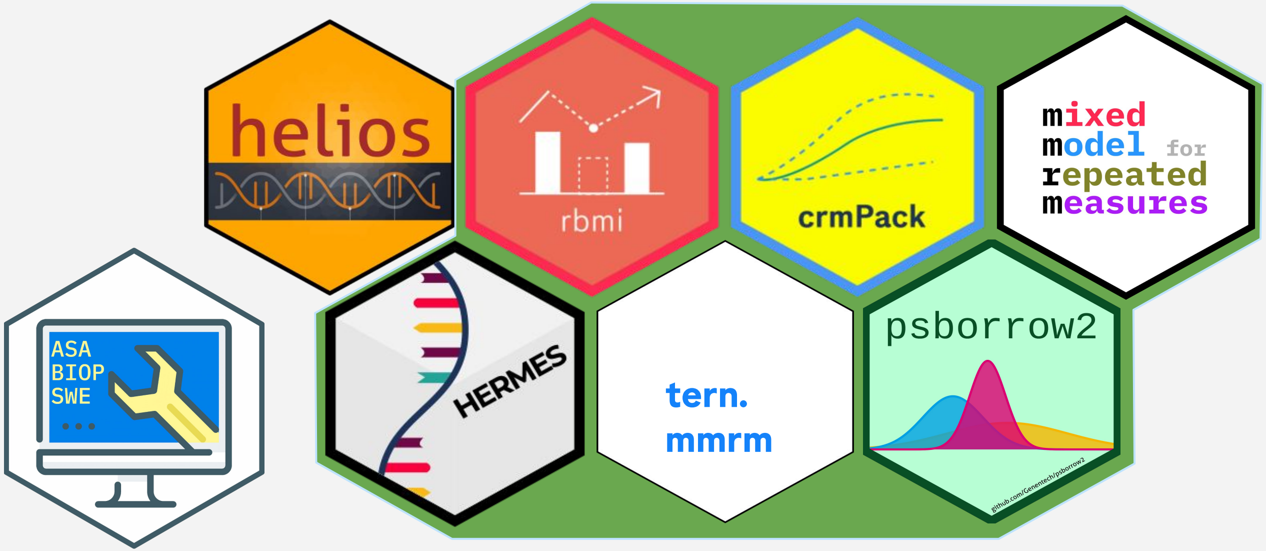
Example R packages our team work on.

In Feb 2023, our team of **~50 data scientists** contributed at least one commit to **100 open source repositories**, and **117 closed source repositories**.

An area of recent growth: Statistical Engineering

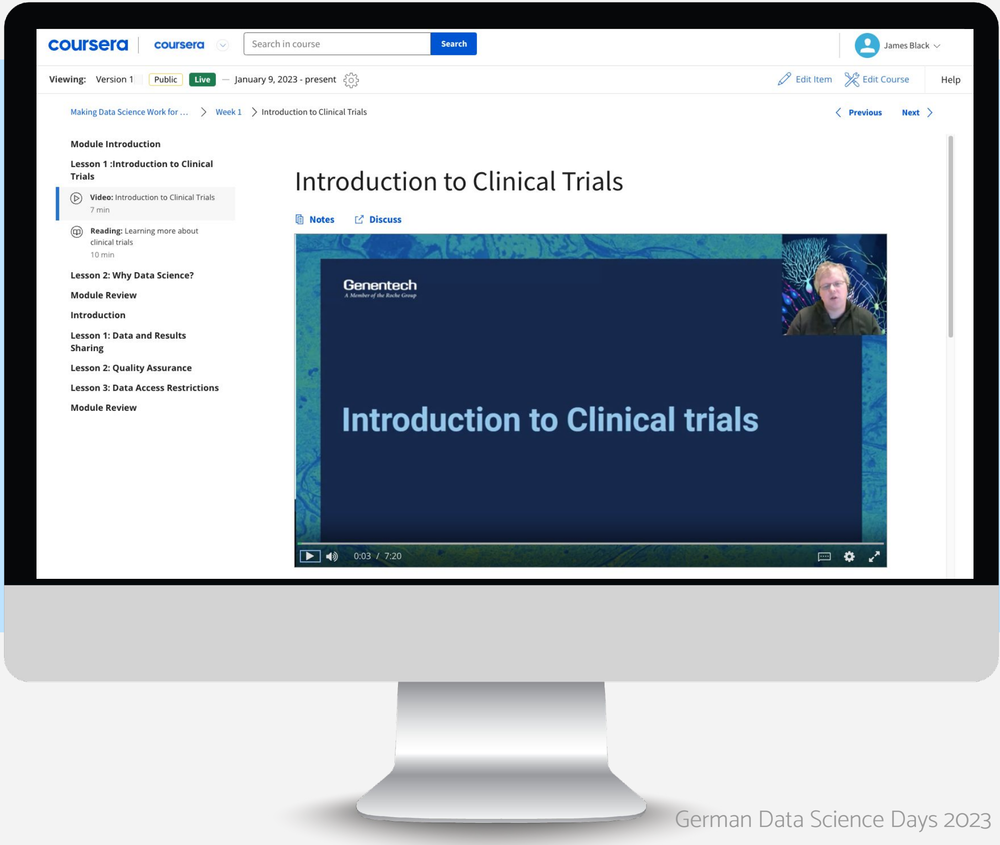
rconsortium.github.io/asa-biop-swe-wg

Open Source


helios
rbmi
crmPack
mixed model for repeated measures
ASA
BIOP
SWE
...
HERMES
tern.
mmrm
psborrow2
github.com/Genentech/psborrow2

Our first coursera course is now live!

- Bridging general DS into clinical reporting DS
- Statistical Programming, Statistical engineering & RWE courses coming
- go.roche.com/course



coursera | coursera

Search in course | Search

James Black

Viewing: Version 1 | Public | Live — January 9, 2023 - present

Edit Item | Edit Course | Help

Making Data Science Work for … > Week 1 > Introduction to Clinical Trials

< Previous | Next >

Module Introduction

Lesson 1 :Introduction to Clinical Trials

Video: Introduction to Clinical Trials
7 min

Reading: Learning more about clinical trials
10 min

Lesson 2: Why Data Science?

Module Review

Introduction

Lesson 1: Data and Results Sharing

Lesson 2: Quality Assurance

Lesson 3: Data Access Restrictions

Module Review

Introduction to Clinical Trials

Notes | Discuss

Contact: james.black.jb2@roche.com

I’m currently hiring!
Research Software Engineer
positions open across Basel,
San Francisco, Welwyn, Mississauga
go.roche.com/dsjobs

Data science careers:
code4life.roche.com