The importance of the SCE in enabling our shift from proprietary programming to open-source data science

James Black, Data & Statistical Sciences, Roche

R/Pharma keynote | 2023

Talk contents

- My context
- RWE/PHC experiences
- What is an SCE?
- Bringing a modern SCE for data science into clinical reporting

My context

My journey with compute at Roche

~2015

Joined Roche as 'RWD analyst'

~2015-18

RWE diverges from clinical reporting with its own SCE

Move into management

~Q1 2018

RWE moves into 'PHC' - bringing imaging, omics, etc into the SCE mix

~2021/22

Shift from doing/managing RWE to a team that focuses on data science tooling

2023

New clinical reporting platform goes into production, using RWE/PHC platform for compute

My current role's connection to SCEs

- Insights Engineering Product Family Lead
  - Sponsor our pan-study 'insights' codebase (e.g. Nest)
- Business lead for compute for late-stage
  - RWD/AA compute (Apollo) business owner
  - Clinical trial (Ocean) compute product owner
- Roche representative on the PHUSE SCE Council

Data science - and compute needs are diverse within a Pharma company

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?

Development

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?

Access/ Commercial

What are the treatment patterns across different markets?

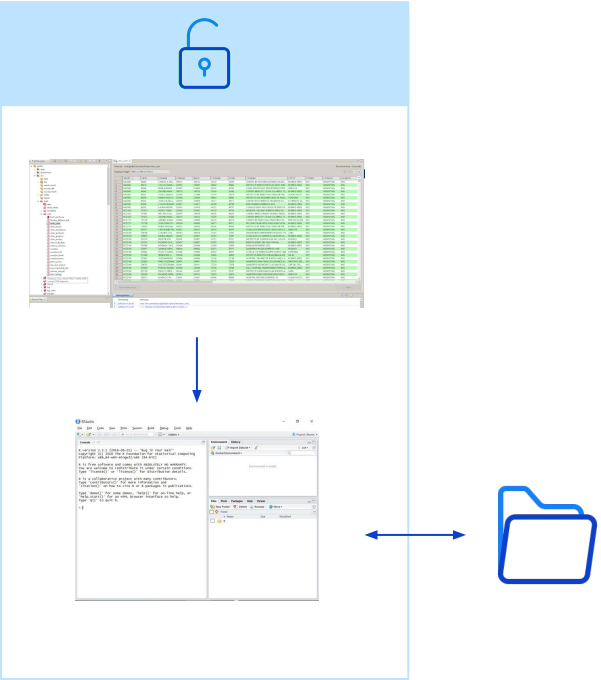
What is the value of our drug across markets?

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**

RWE/PHC experiences

RWE in 2015



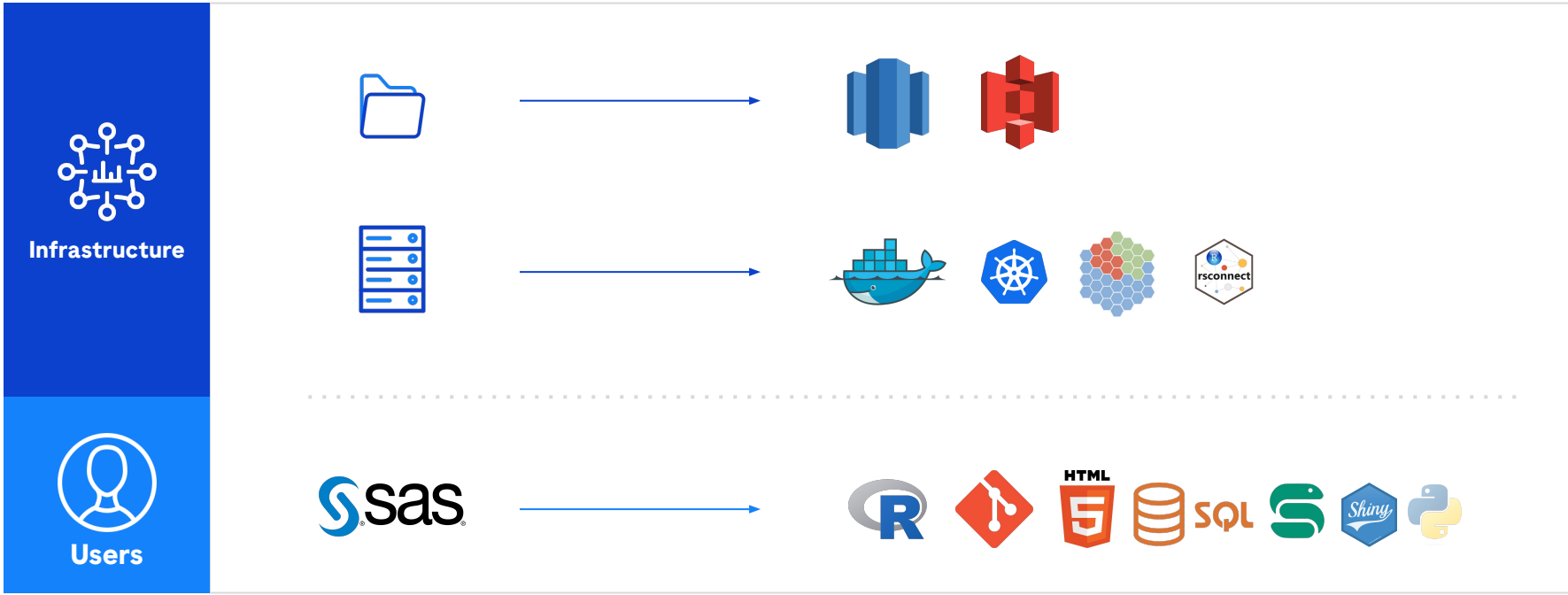
The environment

- Clinical trial data (and compute) in a SAS only ‘kitchen sink’ platform
- Data can be exported one-way into an R server
- Managed by IT with in-frequent releases
- RWD on a network drive (raw and derived)

The experience

- Reading data takes a long time, and data is duplicated as convenience copies
- Have to make tickets for system libraries
- Local laptops are a preferred place to work

RWE tooling before and after 2018

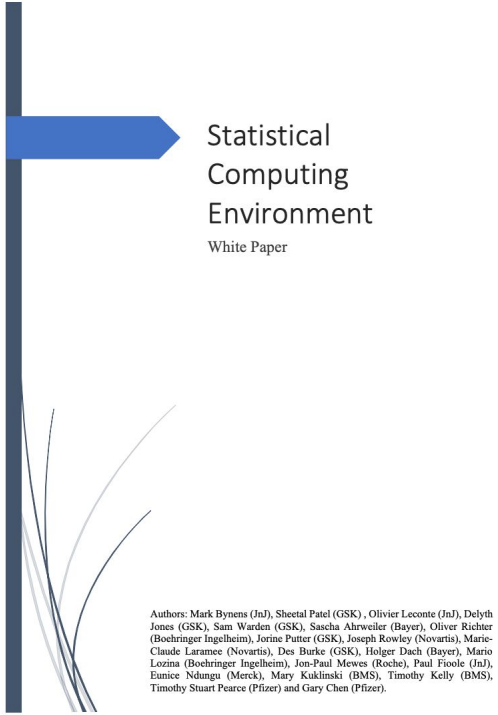


Processes and infrastructure catalysing data science adoption

- Git used for version control → Users familiar with github/gitlab and gitflows
- Snakemake for orchestration → Understanding of codebased workflow tooling
- Production runs off git repos → Familiar with CICD and ‘don’t trust interactively run code’ mindset
- Redshift/S3 for data → Understanding of secrets management, APIs, and databases
- SAS to R → Exposed to huge open source libraries, external collaboration and thinking through dependencies
- Managed servers to containers → Users can take ownership of their environment at the system level

What is an SCE?

What is an SCE?

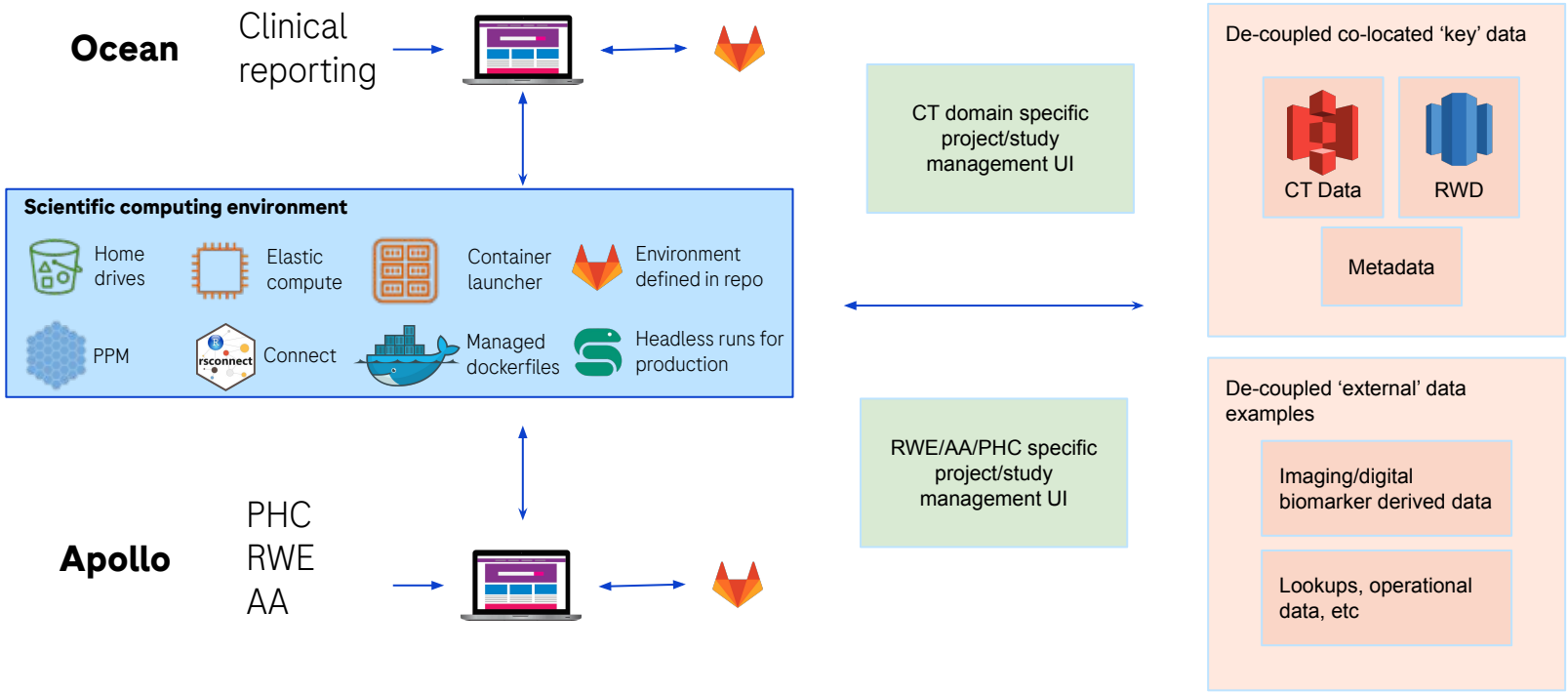

Statistical
Computing
Environment
White Paper

Authors: Mark Bynens (JnJ), Sheetal Patel (GSK) , Olivier Leconte (JnJ), Delyth Jones (GSK), Sam Warden (GSK), Sascha Ahrweiler (Bayer), Oliver Richter (Boehringer Ingelheim), Jorine Putter (GSK), Joseph Rowley (Novartis), Marie-Claude Laramee (Novartis), Des Burke (GSK), Holger Dach (Bayer), Mario Lozina (Boehringer Ingelheim), Jon-Paul Mewes (Roche), Paul Fioole (JnJ), Eunice Ndungu (Merck), Mary Kuklinski (BMS), Timothy Kelly (BMS), Timothy Stuart Pearce (Pfizer) and Gary Chen (Pfizer).

https://www.lexjansen.com/phuse-us/2021/hw/PAP_HoW04.pdf

- *Statistical* or *Scientific* Computing Environment?
- *"A modern SCE must allow for functional, meaningful and delightful experiences to the end users."*
- What is the scope?
  - CDISC derived submission data and TLGs for GxP?
  - Exploratory work?
  - Trial design?
  - RWE?
  - Bioinformatics, imaging, digital biomarkers?
  - SaMD development/MLops?
- **Is data storage (source and derived) part on an SCE?**

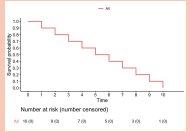
What an SCE looks like at Roche

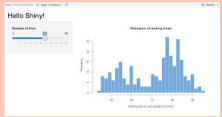


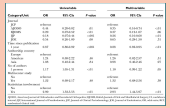
Validation shouldn’t kill innovation in your SCE

Roche’s approach to isolate validated insights from users normal workflows

Work in IDE

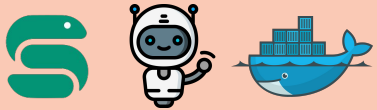


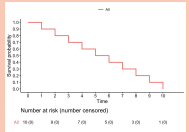




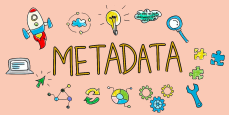
GitLab

Batch runs





+



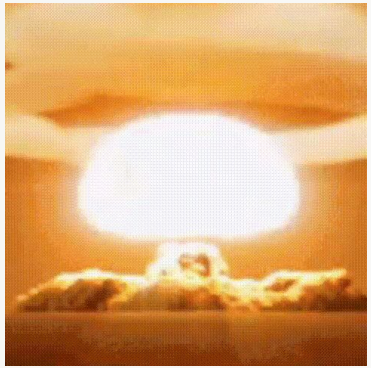
Bringing a modern SCE for data science into clinical reporting

Scaling from RWE to clinical reporting

RWE: New-ish field with ~40 people touching data (in 2018, 500+ in 2023)

Clinical Trials: ~800 Statistical Programmers + additional Statisticians. 20+ years of tools and experience with prior processes

~~Scaling from RWE to clinical reporting~~
One simple trick to make a new SCE easy.....



Define how people work from a blank slate

Support an evolution to a new SCE and ways of working

Setting the dialogue on what a new SCE means to change management

“Existing processes will be supported as-is”

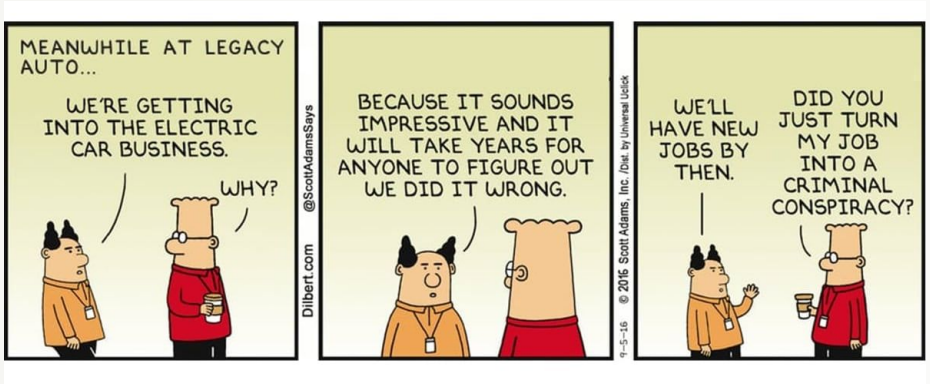
“A multilingual future”

Platform needs to support legacy in an **efficient** way for the foreseeable future

“Open source, with an R backbone”

Negotiated timelines to move to new ways of working

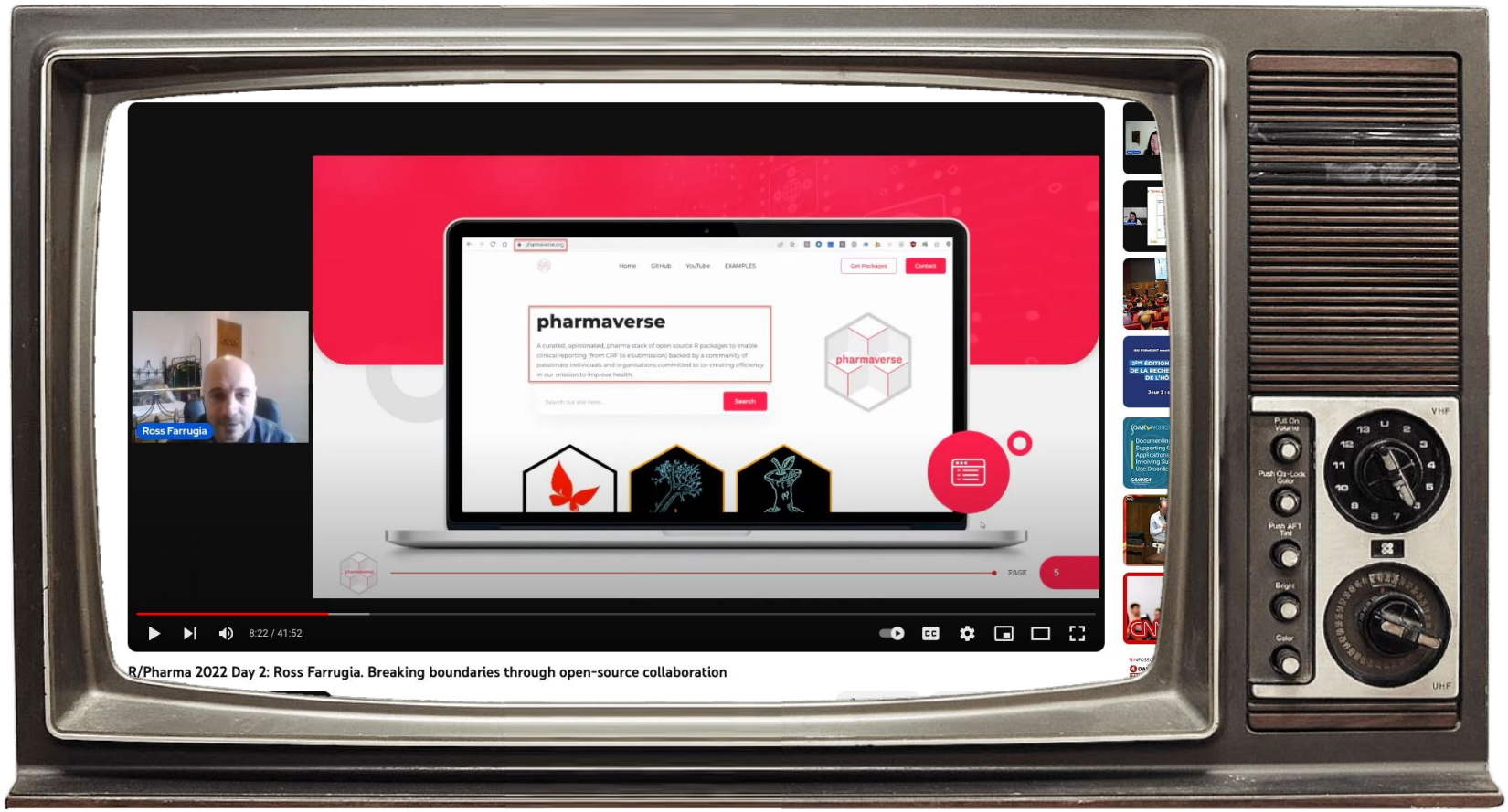
Why disrupt clinical reporting?



"We've spent decades with one language and vendor products that abstract away the technical aspects of data science into an end to end platform! Why throw that all away with a swamp of tools?"

"Git, codebased workflow management, sharing code via R packages - this all too technical and slows people down!"

"Statistical programming is a distinct, highly specialised and very process driven role - generic and flexible tools and platforms will only decrease efficiency"


Ross Farrugia
pharmaverse
8:22 / 41:52
R/Pharma 2022 Day 2: Ross Farrugia. Breaking boundaries through open-source collaboration

We need to refactor clinical reporting SCEs into more generalised data science SCEs

- Talent are graduating with experience in R, Python and Git, so why can't statistical programmers use the same language as the statisticians?
- Clinical trial design and data modalities used continues to grow in complexity
- The pharmaverse has shifted the codebase for clinical reporting to be a collaborative effort in R
  - Statistical programmers are now:
    - Spending more time in software development
    - Need to be able to handle new data modalities and emerging tools in other languages
    - Can be expected to have core data science competencies like git, environment management and keep up with data sciences evolution

Breaking clinical reporting out of it's domain specific SCE design is required to support this shift to statistical reporting becoming a specialty within data science, rather than a separate silo'd world

Notes from R/Pharma round tables

"What are the bounds of an SCE?"

"We have many legacy workflows to support"

"Supporting legacy workflows (e.g. filesystem based data and old macros) cripples innovation"

"Maintaining provenance/replicability gets harder moving away from one platform"

Scene from the round tables in Chicago this year

"How will a new SCE make the company money?"

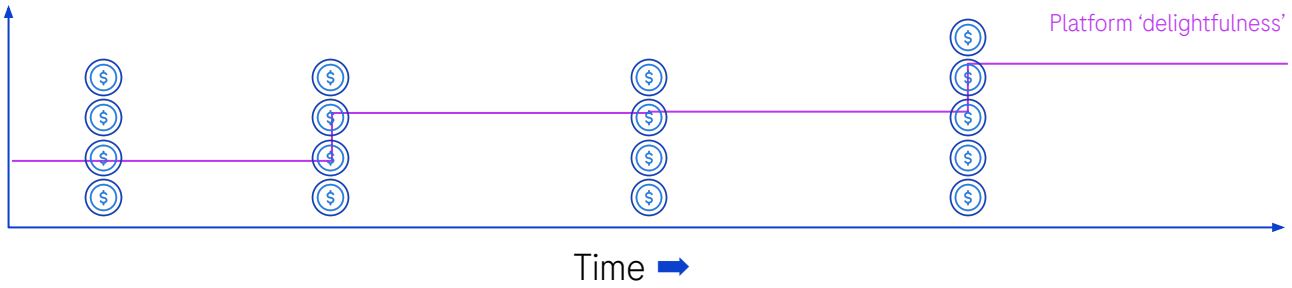
"Can GxP and exploratory co-exist?"

"Validation is painful and orientated towards vendor platforms"

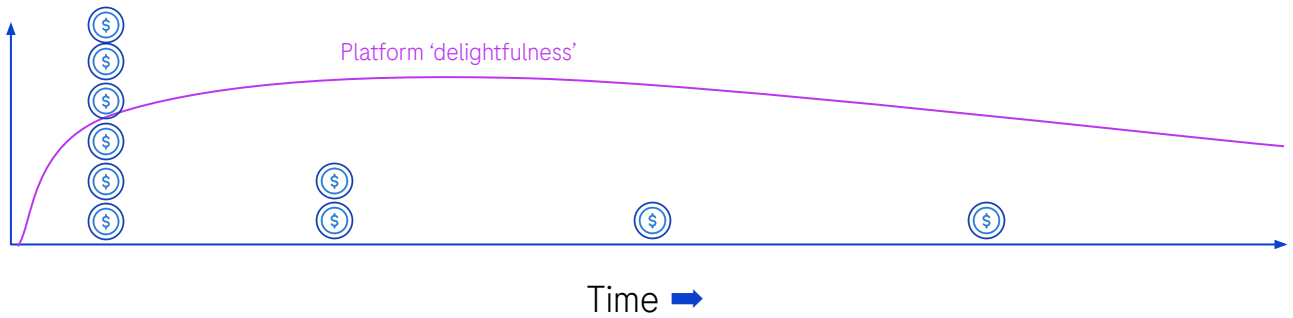
"Sustained funding for homegrown post 1.0 is a challenge"

The curse of internal funding models

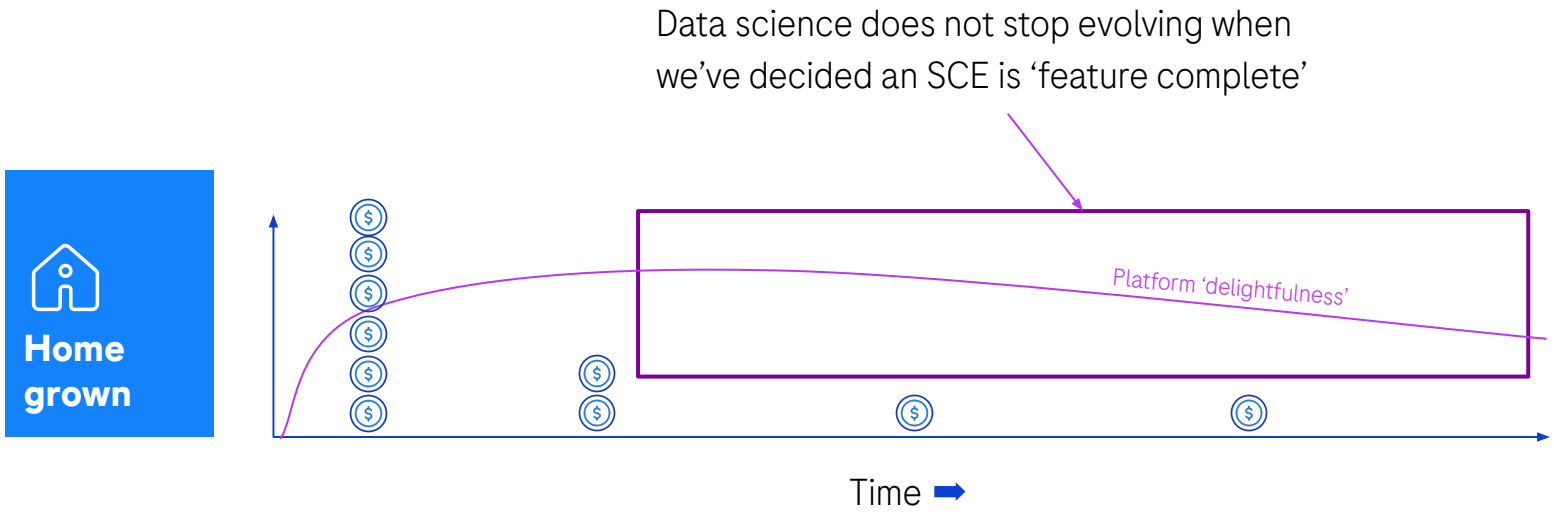
Vendor product



Home grown



The curse of internal funding models



Re-capping some key questions to tackle with a modern clinical reporting SCE

- Is the scope clinical reporting by statistical programmers, late stage evidence, or even broader?
- Be clear on what you expect from users
  - Should statisticians use git/batchless runs for trial design?
  - Do we want statistical programmers on a platform optimised for specific tasks with lots of abstraction, or data scientists on an adaptive platform?
  - Can we isolate validation to a specific subset of the platform? (e.g. batch runs)
  - For what do we need to capture provenance/metadata? (e.g. batch runs)
  - Will/can legacy/existing studies be migrated?
  - If homegrown - can we find a sustainable budget for innovation?

The discussion continues later today....

Please also join us and the end of day 2 (today) for a panel discussion about what is a next-generation SCE

Doing now what patients need next