Reproducibility as a goal for all scientists

Faical Yannick Palingwende Congo

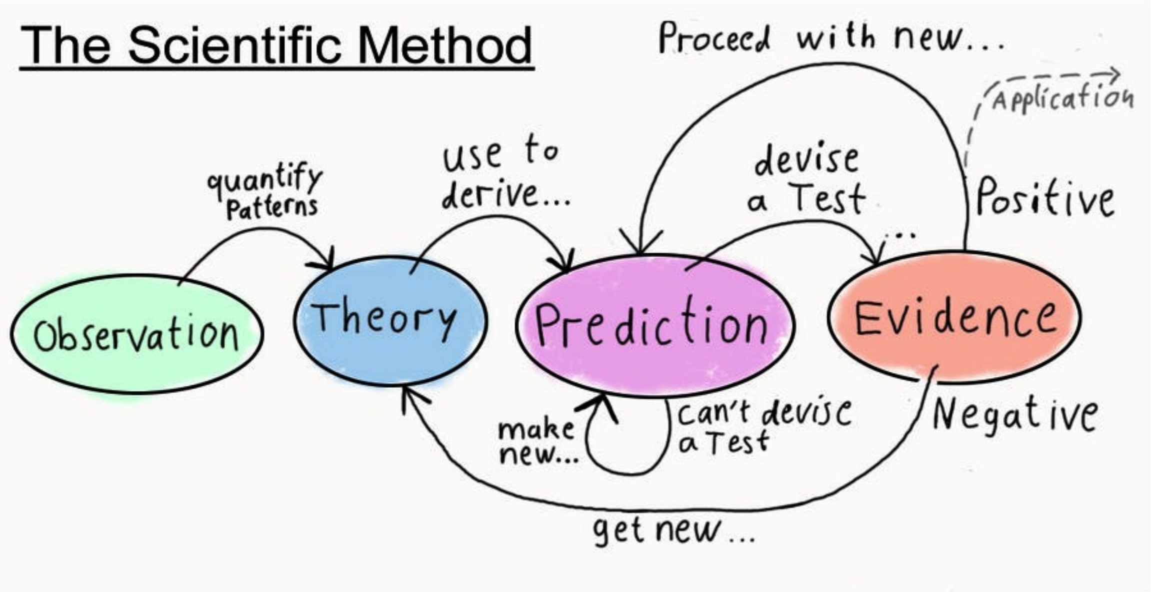
11-23-2015

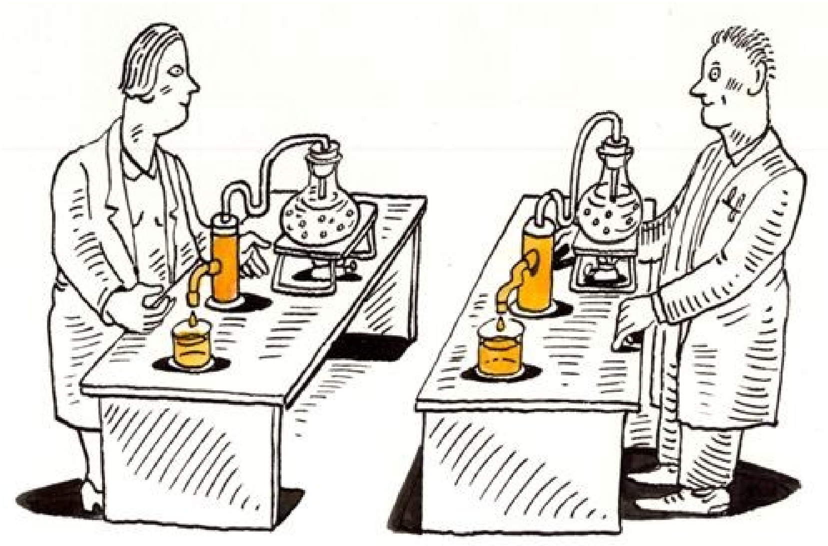
Simple Terminology by Example

- **repeatable**
  - Scientist O ran an investigation O.
  - The investigation O was performed with machine O.
  - The investigation O used an input O.
  - The execution deliver result O.
  - If Scientist O run the investigation O again some time later with input O on machine O and get result O again
  - Investigation O can be tagged: Repeatable.

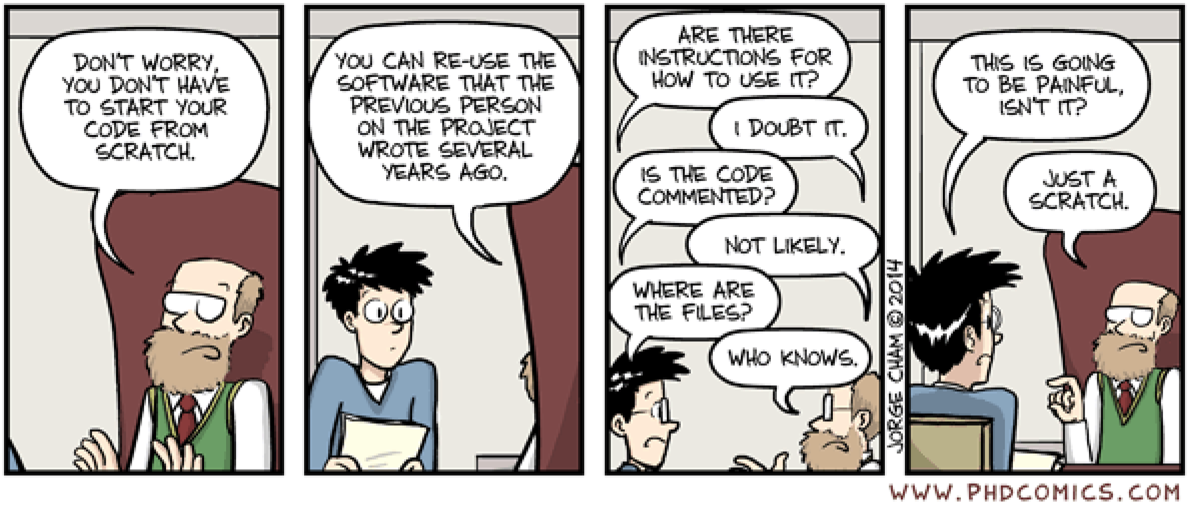
- **reproducible**
  - Scientist O ran an investigation O.
  - The investigation O was performed with machine O.
  - The investigation O used an input O.
  - The execution delivers result O.
  - Then Scientist U can manage to run investigation O/U with an input O/U on a machine O/U and result U in agreement with result O.
  - Investigation O can be tagged Reproducible.

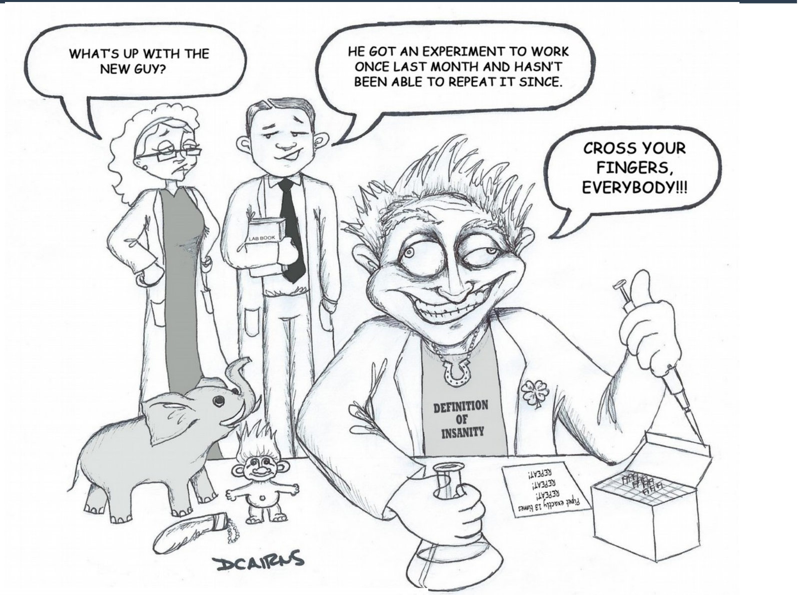
- **replicable**
  - Scientist O ran an investigation O.
  - The investigation O was performed with machine O.
  - The execution delivers result O.
  - Then scientist A can manage to get a copy of machine O, and input O and investigation O.
  - If scientist A can manage to get result O, Investigation O can be tagged: Replicable.
  - Other configurations are possible with same scientist O doing like A.

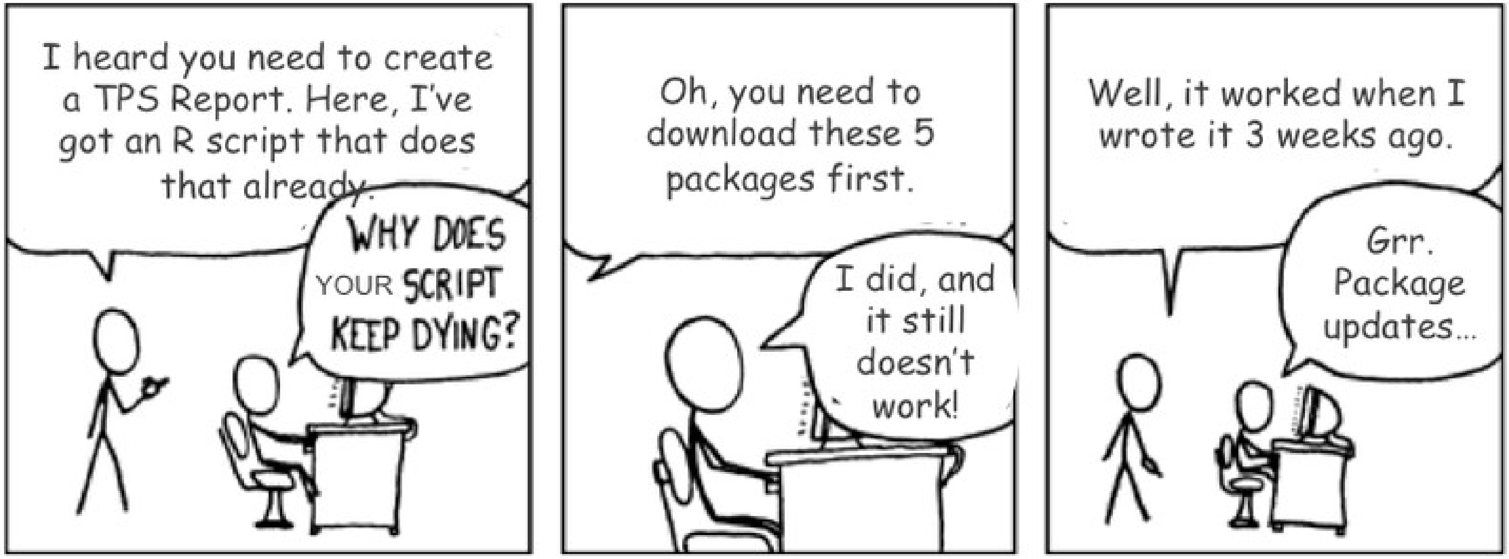




The Problems


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
JORGE CHAM © 2014
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
WWW.PHDCOMICS.COM

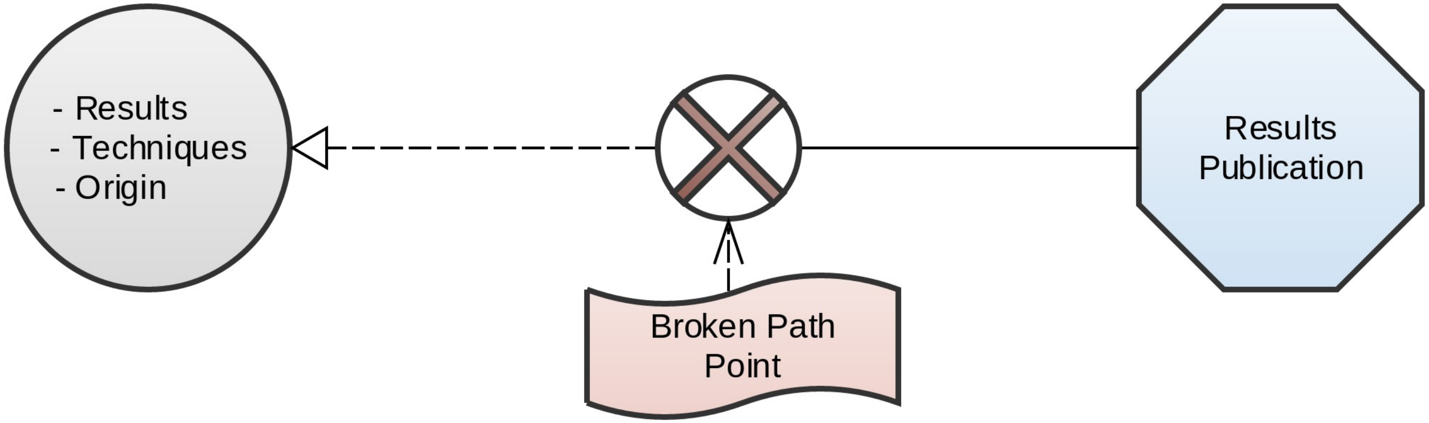

WHAT'S UP WITH THE NEW GUY?
HE GOT AN EXPERIMENT TO WORK ONCE LAST MONTH AND HASN'T BEEN ABLE TO REPEAT IT SINCE.
CROSS YOUR FINGERS, EVERYBODY!!!
LAB BOOK
DEFINITION OF INSANITY
DCAIRNS


I heard you need to create a TPS Report. Here, I've got an R script that does that already.
WHY DOES YOUR SCRIPT KEEP DYING?
Oh, you need to download these 5 packages first.
I did, and it still doesn't work!
Well, it worked when I wrote it 3 weeks ago.
Grr. Package updates...

Unidentified elements

- **Computational**
  - Machine requirements
  - Operating system variations
  - Dependencies variations
  - Execution parameters
  - Input files

- **Experimental**
  - Machine variation
  - Machine calibration
  - Experimental specifications
  - Experimental setup
  - Experimental input

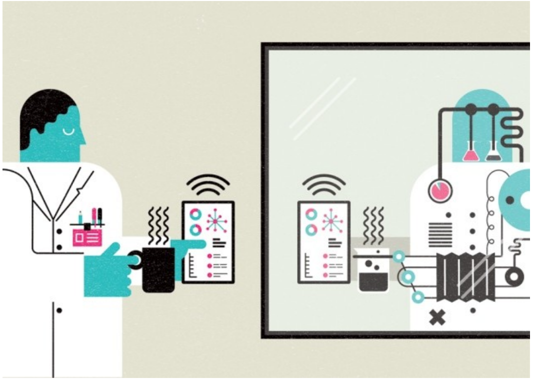


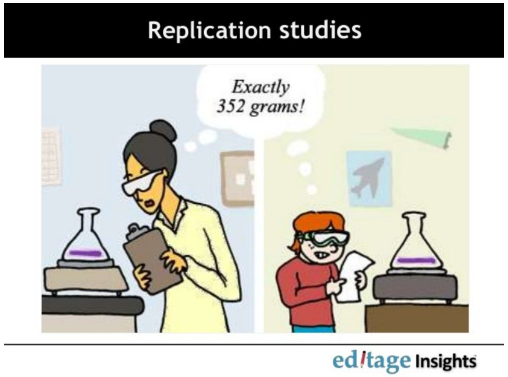
Approach

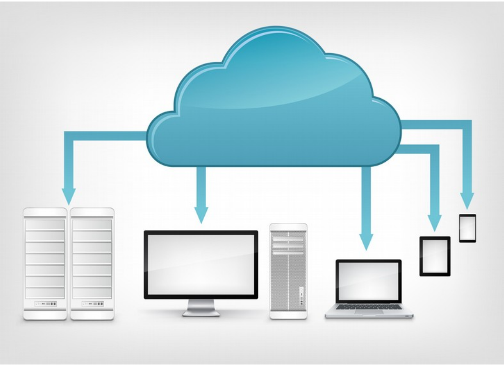
- ## Record
  - Important elements
  - Reproducible record
  - Generic structure
  - Flexible use
  - Result: Standardized

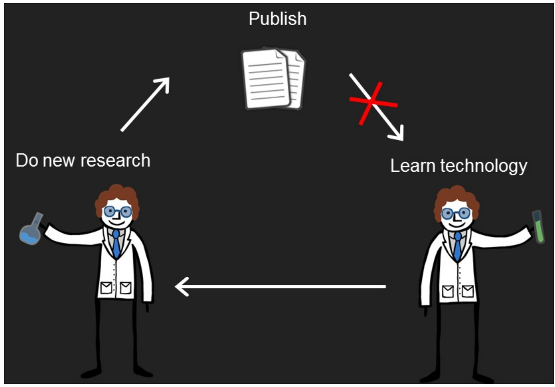
- ## Disseminate
  - Cloud
  - Central platform
  - Easy to share
  - Easy to reference
  - Provenance id

- ## Collaborate
  - Reproducibility tags
  - Interactions on records
  - Education from records
  - Results grouping
  - Rationals about records







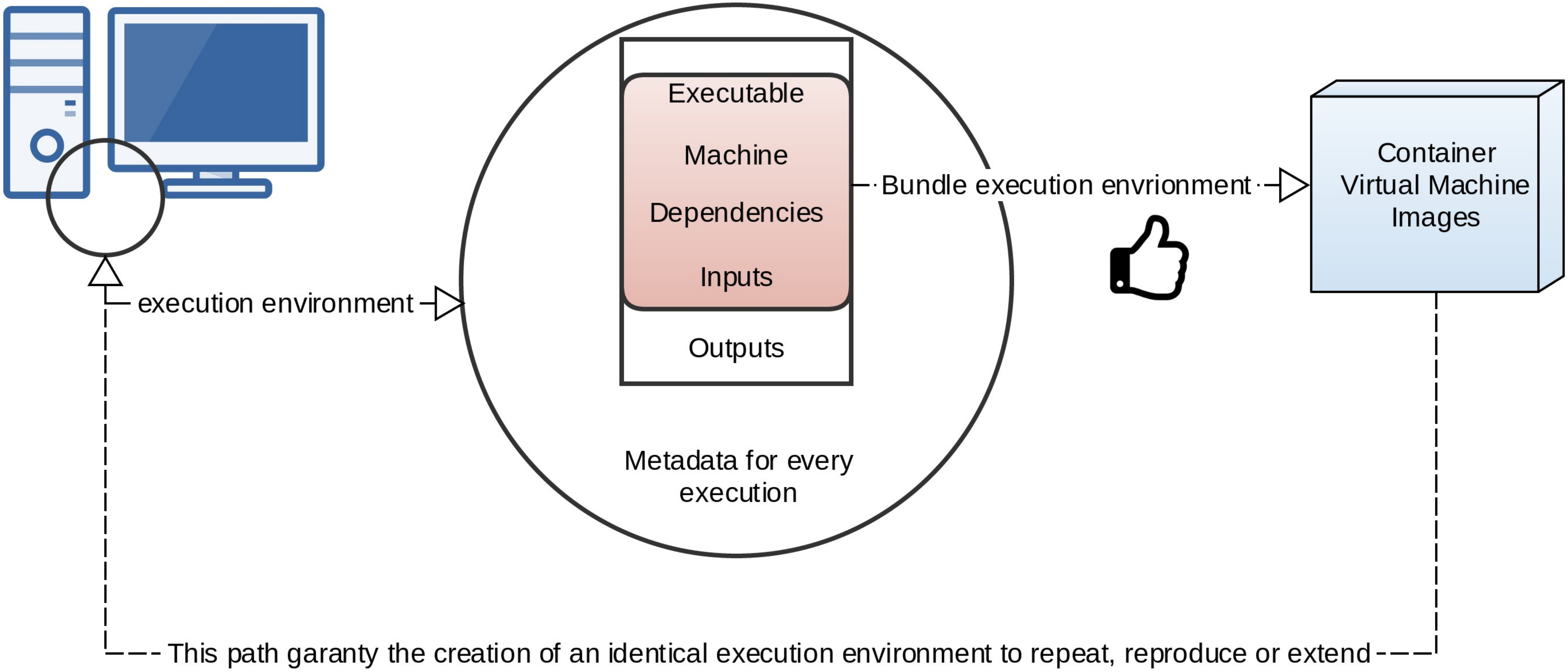


- ## Automate
  - Environment setup
  - Investigation setup
  - Execution
  - Avoid manual configs
  - Standard representation

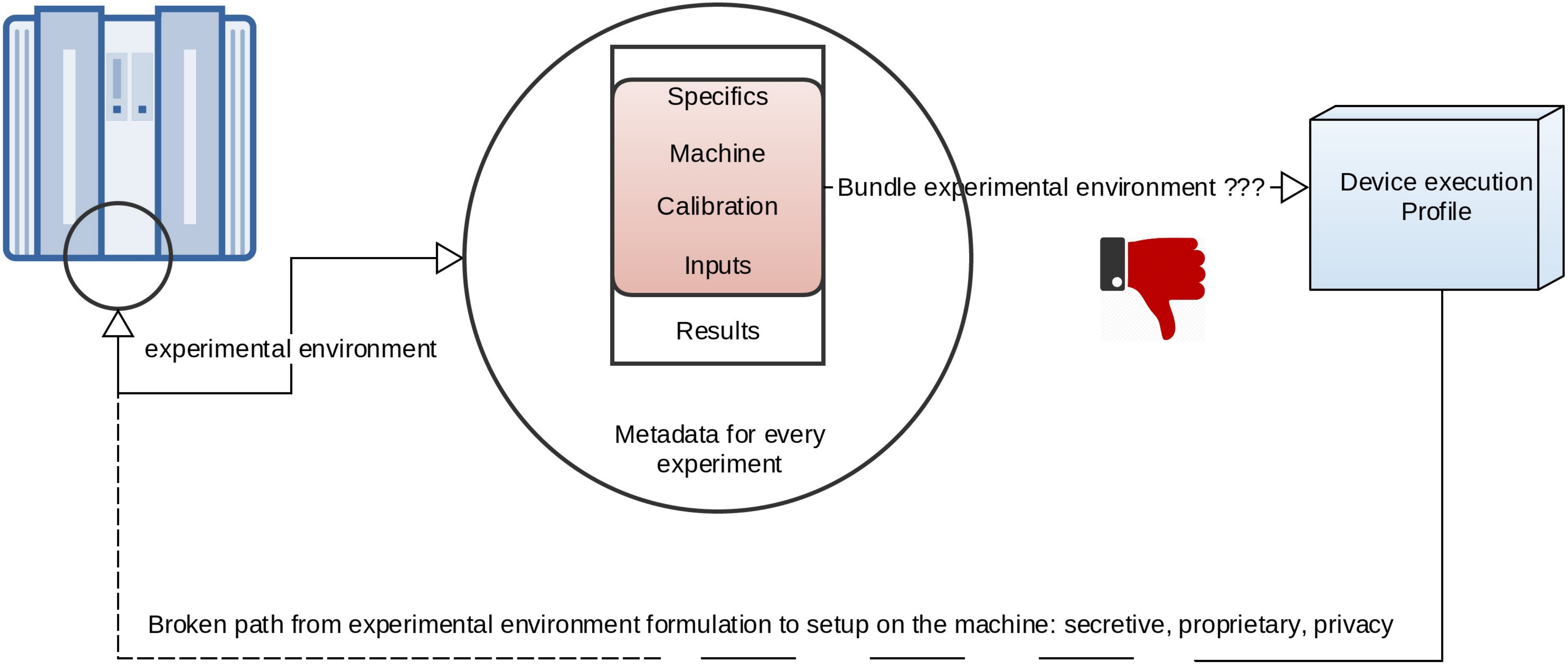
- ## Inter-operate
  - Record equivalence
  - Machine Interface
  - Machine to Machine
  - Machine to Cloud
  - Cloud to Machine

- ## Consensus
  - Openness ground
  - Advancement sacrifices
  - Standards
  - Automated world
  - Cloud future

The computational solution


execution environment
Executable
Machine
Dependencies
Inputs
Outputs
Metadata for every execution
Bundle execution envrionment
Container
Virtual Machine
Images
This path garanty the creation of an identical execution environment to repeat, reproduce or extend

The experimental solution?


Specifics
Machine
Calibration
Inputs
Results
Metadata for every experiment
experimental environment
Bundle experimental environment ???
Device execution Profile
Broken path from experimental environment formulation to setup on the machine: secretive, proprietary, privacy

Recording Computational Research

- **Workflow tools:**
  - Black box model: Input & Output Description
  - Recording of the pipeline involved
  - Taverna, Galaxy, Madgascar, VisTrails
- **Event based control tools:**
  - Event based model: Track all interactions with the OS as a process parent.
  - Records: System infos, Inputs, Outputs, Executables, Dependencies.
  - Sumatra, ReproZip
- **Libraries:**
  - Integration: Provide alternative objects and track interactions from with the code.
  - Documentations generation, inputs, outputs, dependencies
  - Dexy, Sumatra,

Recording Experimental Research

- **Api for developers:**
  - Open machine: interactions and events captured
  - Everything can be virtually captured
  - Rare configuration: Generaly for open hardware/source
- **Computer Files/Projects watchdogs:**
  - Project storage and Data storage can be tacked
  - Outputs and some parameters/inputs can be tracked
  - Common for non proprietary software/Always possible except warranty
- **External tracking devices:**
  - Sensors places at key positions
  - Capture: Inputs, Outputs and critical events
  - A bit of a hacking but still possible too.

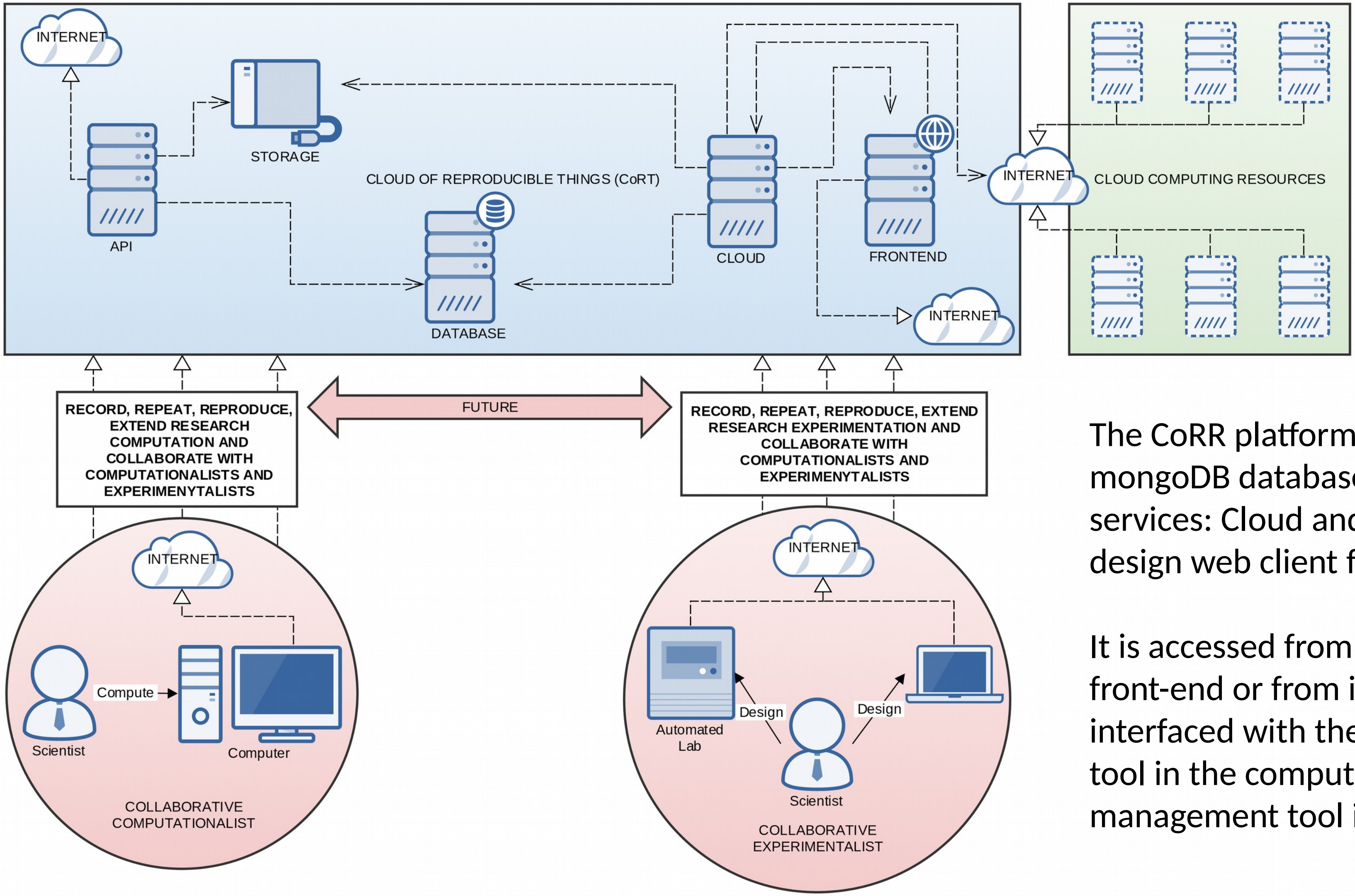
Building the bridge to a Standard

Minimal Machine Programming Interface

- **Run:**
  - Machine execute provided command.
  - Machine execute experimental design file.
- **Status:**
  - Machine provide current experiment status.
  - Running/Failed/Finished/Terminated/Lost.
- **Calibrate**
  - Machine can load a calibration file if provided.
  - Machine can give its calibration state and file.
- **Result:**
  - Machine return result of the last experiment ran.
- **Info:**
  - Machine informations and specifications.
- **Cloud:**
  - Configure and check the cloud configuration.

Cloud of Reproducible Records (CoRR)

A unified General Purpose approach to reproducibility


INTERNET
STORAGE
API
CLOUD OF REPRODUCIBLE THINGS (CoRT)
DATABASE
CLOUD
FRONTEND
INTERNET
INTERNET
CLOUD COMPUTING RESOURCES
RECORD, REPEAT, REPRODUCE, EXTEND RESEARCH COMPUTATION AND COLLABORATE WITH COMPUTATIONALISTS AND EXPERIMENYTALISTS
FUTURE
RECORD, REPEAT, REPRODUCE, EXTEND RESEARCH EXPERIMENTATION AND COLLABORATE WITH COMPUTATIONALISTS AND EXPERIMENYTALISTS
INTERNET
Compute
Scientist
Computer
COLLABORATIVE COMPUTATIONALIST
INTERNET
Automated Lab
Design
Design
Scientist
COLLABORATIVE EXPERIMENTALIST

The CoRR platform is composed of: A mongoDB database server; two flask services: Cloud and API; and a material design web client front-end.

It is accessed from the web through the front-end or from its API which is interfaced with the Software management tool in the computer or experiment management tool in the machine.

Collaborations

- **FACT Lab**
  - Laura Espinal
  - Uncertainty Quantification
  - Adsorption/Desorption
  - 3 machines
  - Record and 0.5 interoperability

- **MDCS**
  - Team
  - Data Curation
  - Representation standards
  - Input and Output storage
  - Meta-data storage

- **DFT Benchmarking**
  - Francesca Tavazza
  - Standardized data structure
  - Unified filtering access
  - MDCS integration
  - Software runs recorded

- **Sample Reference**
  - Zach Trautt
  - QR code tag to sample
  - Link to experiment
  - Link all results to sample
  - Query all experiments from QR code

- **Materials Framework**
  - ShengYen Lee
  - Machine learning
  - Material Science
  - Record framework runs
  - Analyze results

- **PyMKS**
  - Surya Kalidindi (Georgia Tech)
  - Material Knowledge System
  - Record experiments
  - Work-flow management
  - Reproducibility in Science

Why interests in BioScience?

- **Science high presence in both worlds**
  - Experimental presence
  - Computational presence
  - Most likely experimental/computational combination
- **Heterogeneous environment**
  - Multiple devices usage
  - Manual transitions
  - Need for automation
- **Reproducibility challenges**
  - Most experiments are hard to replicate.
  - Machine interoperability is a challenge.
  - So much potential still there to unleash.

Thank you

faical.congo@nist.gov