"Benchmarking via Challenges & Competitions"

… the RoCKIn legacy

… how I stopped worrying about RoboCup and learned to love benchmarcking.

Matteo Matteucci, matteo.matteucci@polimi.it

Dept. of Electronics, Information and Bioengineering

Politecnico di Milano

Where did it all start?

RoboCup European Championship

Amsterdam-2000

- General Chair: Frans Groen, (groen@wins.uva.nl)
- Organizing Co-Chairs:
  - Enrico Pagello, (epv@dei.unipd.it)
  - Gerhard Kraetzschmar Ulm, (gkk@acm.org)
- Simulator League Chair: Huosheng Hu, (hhu@essex.ac.uk)
- F180 (Small Size) League Co-Chair: Andeas Birk, (cyrano@arti.vub.ac.be)
- F2000 (Middle Size) League Co-Chair: Pedro Lima, (pal@isr.ist.utl.pt)
- Workshop Co-chair: Giovanni Adorni,(adorni@ce.unipr.it)
- Jarkko Kemppainen, (jarkko.kemppainen@nokia.com)
- Pierre Blazevic,(blazevic@robot.uvsq.fr)
- Andreas Zell, (zell@Informatik.Uni-Tuebingen.de)
- Raul Rojas,(rojas@inf.fu-berlin.de)
- Leo Stankevich,(stankevitch@phtf.stu.neva.ru)
- Josep De la Rosa,(peplluis@eia.udg.es)
- Silvia Coradeschi (silvia.coradeschi@tech.oru.se)
-
- **Advisory Board**:
  - Manuela Veloso,(mmv@school.coral.cs.cmu.edu)
  - Peter Stone, (pstone@research.att.com)
  - Hiroaki Kitano,(kitano@symbio.jst.go.jp)
  - Minoru Asada,(asada@er.ams.eng.osaka-u.ac.jp)

Amsterdam 2000

Where did it all start?

Lisbon 2004

Bremen 2006

Are we really benchmarking AI?

Competition as Experiments

"Challenge and competition events in robotics

provide an excellent vehicle for advancing the state of the art

and evaluating new algorithms and techniques

in the context of a common problem domain.

[...] treat competitions and challenges as repeatable experiments."

Monica Anderson, Odest Chadwicke Jenkins, and Sarah Osentoski

"Recasting Robotics Challenges as Experiments"

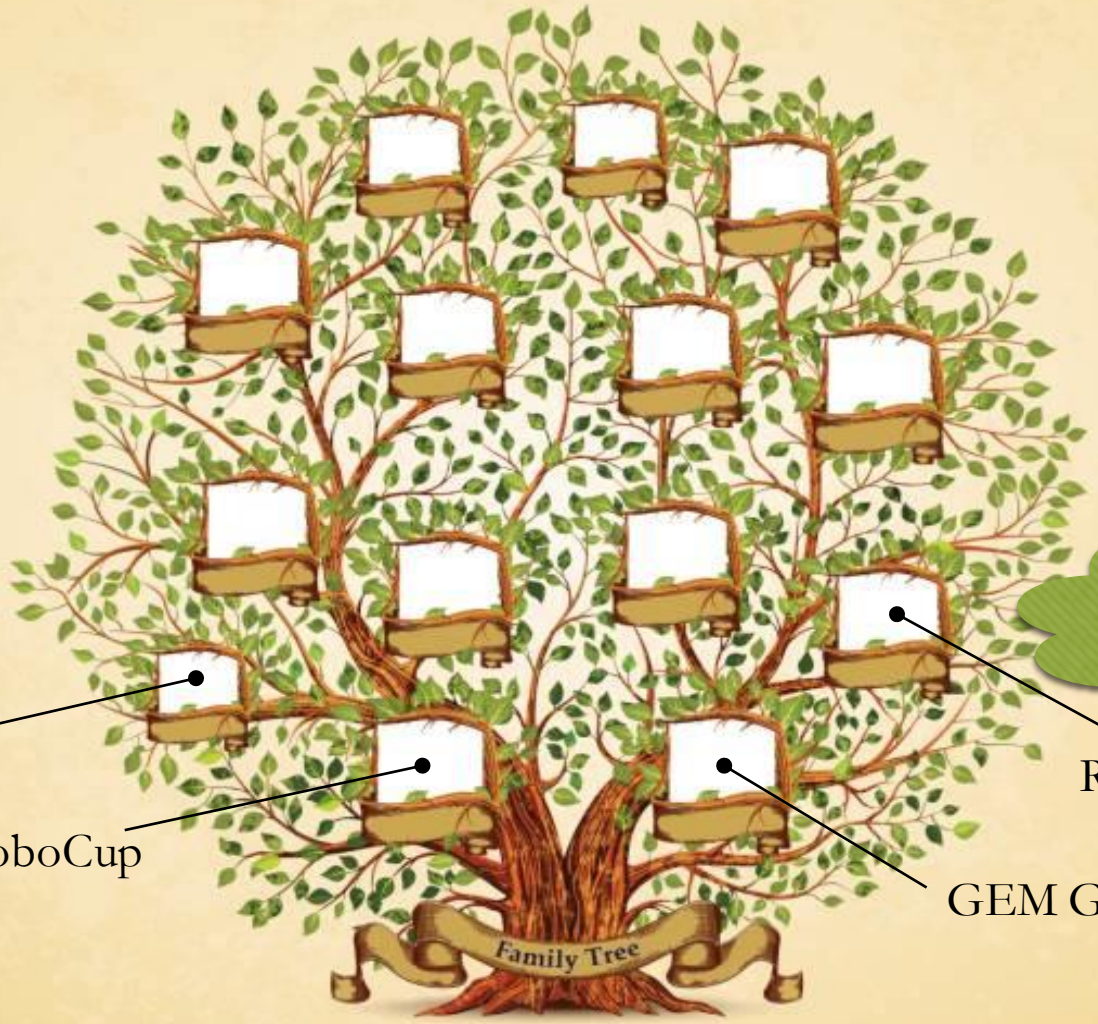
Competition as Experiments

"One-time demonstrations of robot performance
(e.g., grand challenges or other competitions) in robotics
are one way of comparing the performance of robots,
but they do not necessarily prove that one's robotics research
is consistently better or worse than another lab's."

Leila Takayama (Google[x], formerly at Willow Garage)
"Towards a Science of Robotics: Goals and Standards for Experimental Research"

Good Experimental Methodologies

- *"General Guidelines for Robotics Papers using Experiments"*
  - Is it an experimental paper?
  - Are the system assumptions/hypotheses clear?
  - Are the performance criteria spelled out explicitly?
  - What is being measured and how?
  - Do the methods and measurements match the criteria?
  - Is there enough information to reproduce the work?
  - Do the results obtained give a fair and realistic picture of the system being studied?
  - Are the drawn conclusions precise and valid



RoCKIn

RoboCup

GEM Guidelines

RAWSEEDS

I talked about this at the CHIST-ERA Workshop in Warsaw (2017)

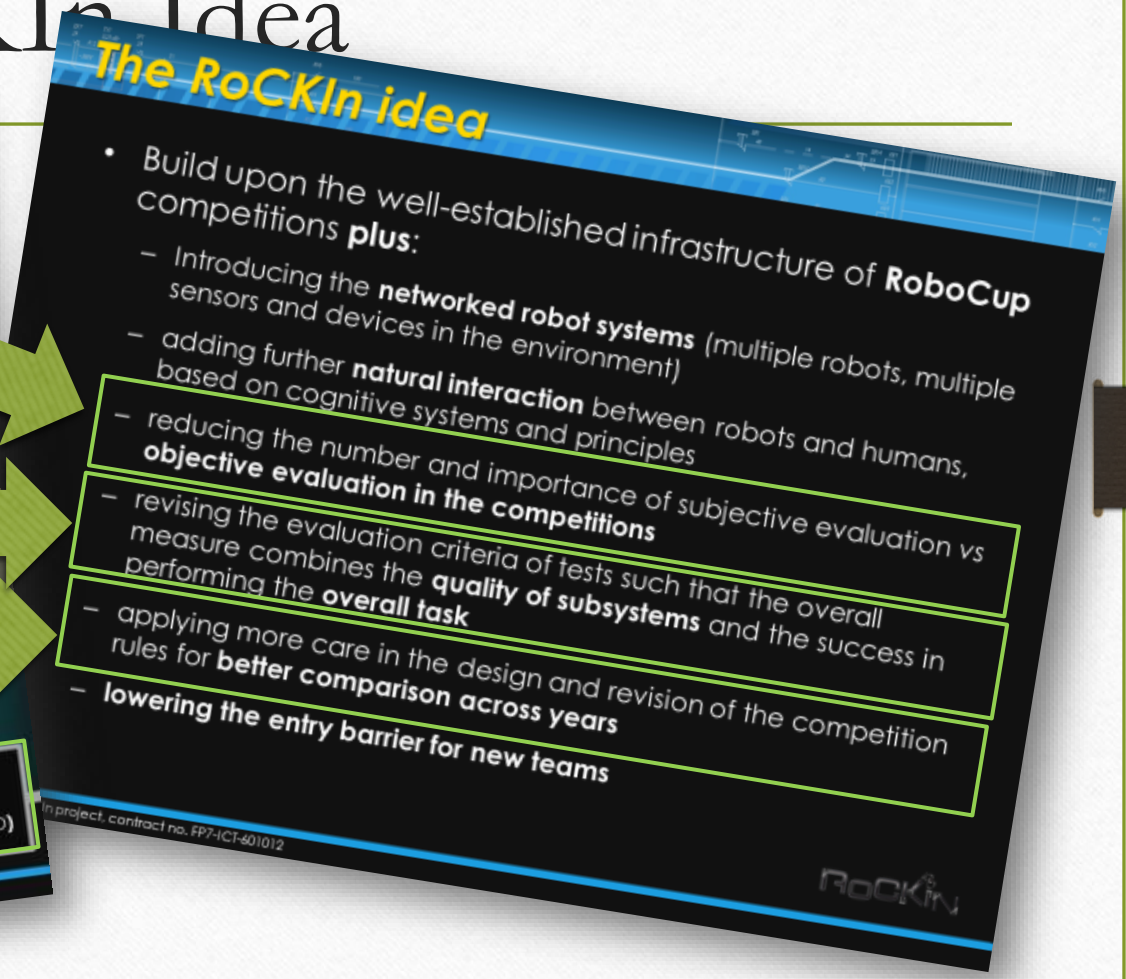
The RoCKIn Idea

“Robots are doomed to close the loop”

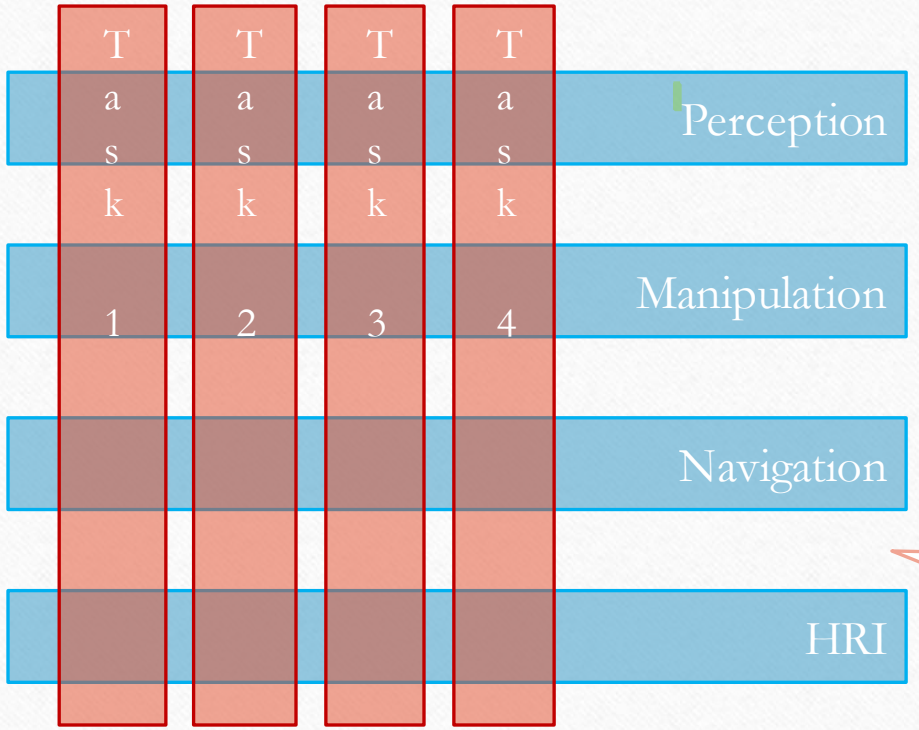
- Robots experiments can be done at two different levels
  - Task Level: evaluation of whole systems on a specific task (e.g., the “bring me the glasses” tasks)
  - Functionality Level: evaluation of modules implementing, in a general manner, functionalities required by the competition tasks (e.g., vision, planning, manipulation)
- RoCKIn competitions should allow independent evaluation at both levels

Intelligence comes from the effective interaction of different capabilities

The RoCKIn Idea



“Robots are doomed to close the loop”

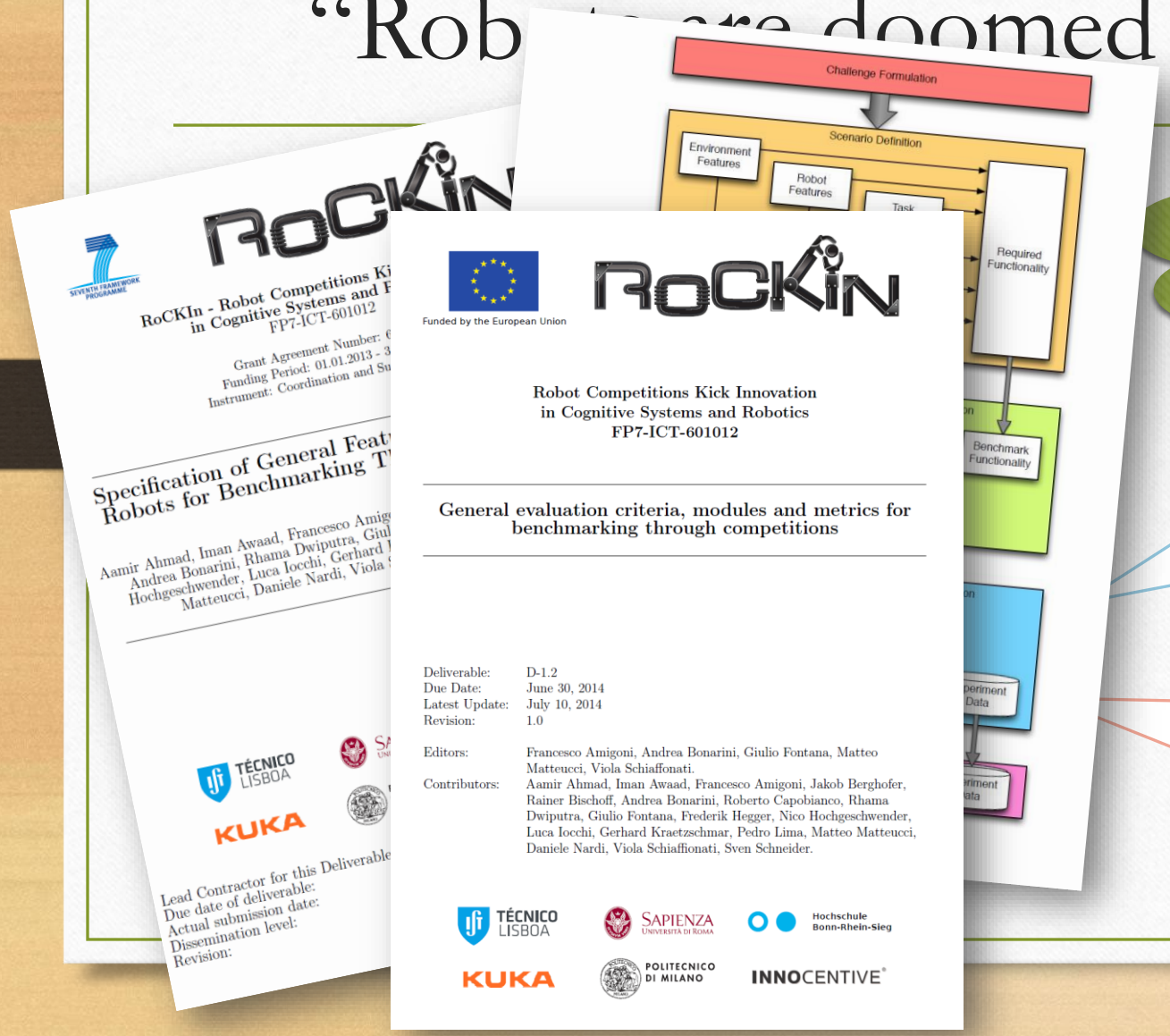


Most of what we call today performance evaluation in intelligent systems is somehow limited to functionality benchmarks!

Functionality Benchmarks (FBM) evaluate performance of modules for specific functionalities

Task Benchmarks (TBM) assess performance of integrated robot systems facing complex tasks

"Rob[illegible] doomed to close the loop"

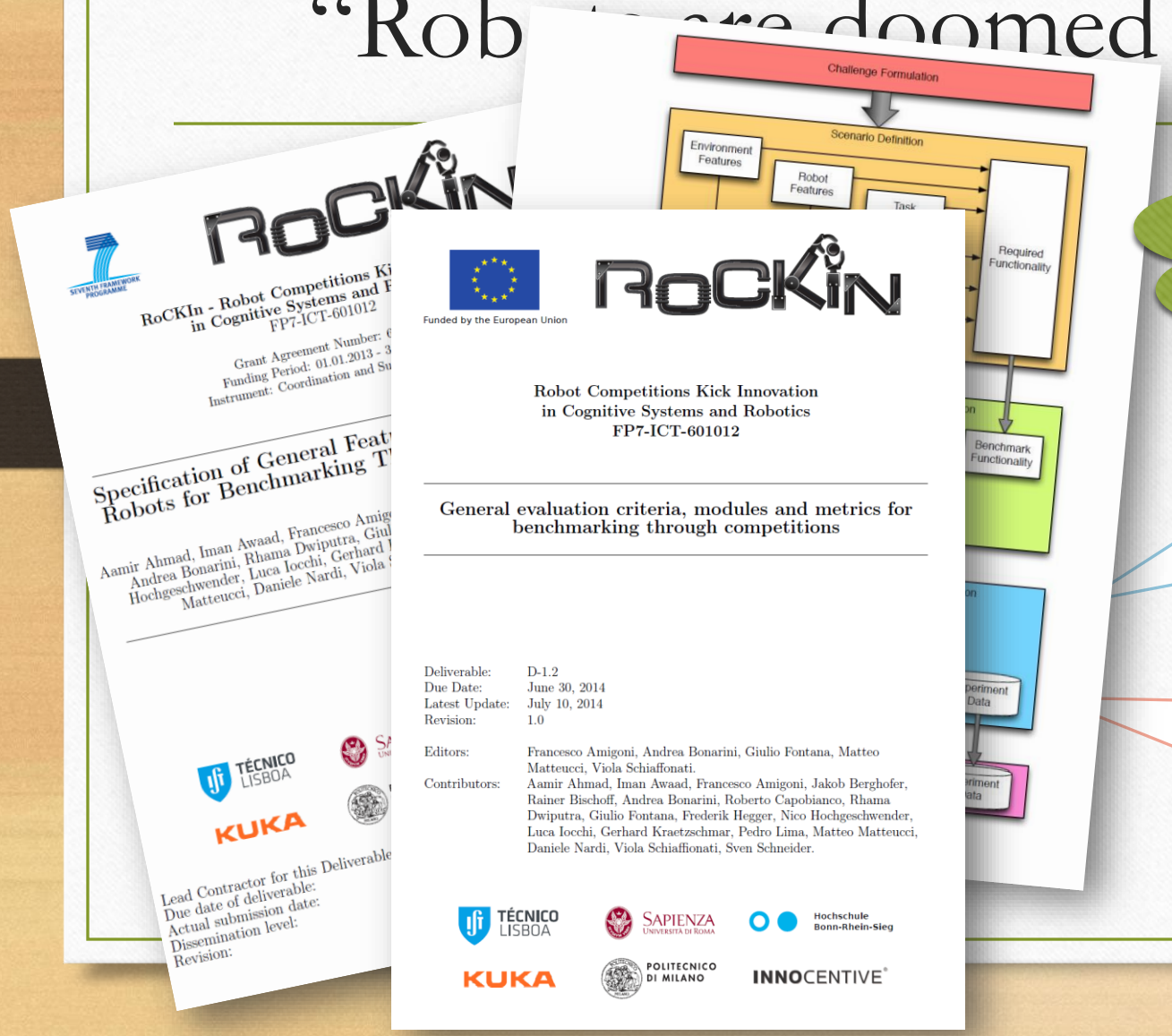


Most of what we call today competitions/challenges are task benchmarks!

Functionality Benchmarks (FBM) evaluate performance of modules for specific functionalities

Task Benchmarks (TBM) assess performance of integrated robot systems facing complex tasks

“Rob … doomed to close the loop”

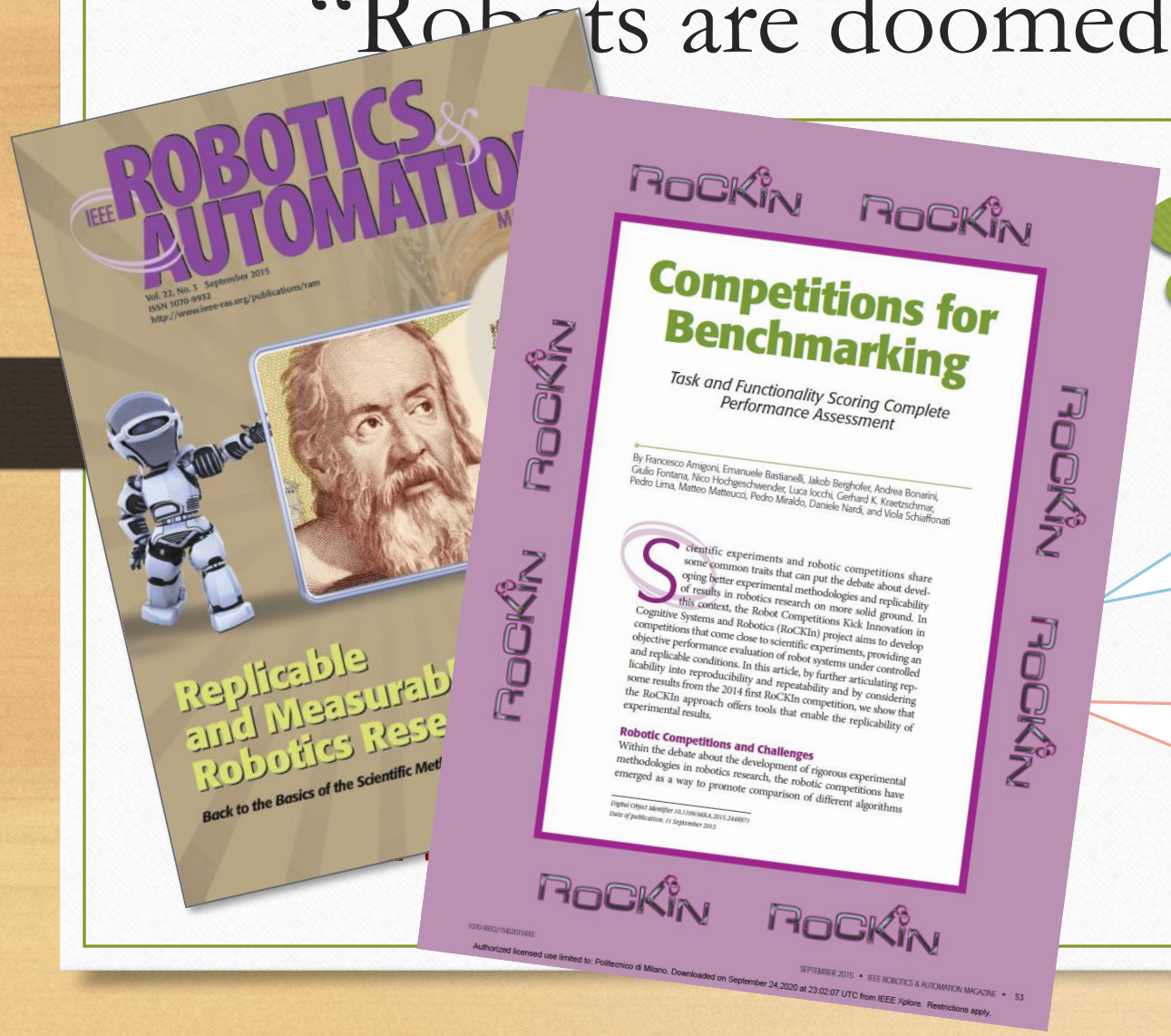


This integrated approach is what the RoCKIn PO dubbed as «RoCKIn legacy» !!!

Functionality Benchmarks (FBM) evaluate performance of modules for specific functionalities

Task Benchmarks (TBM) assess performance of integrated robot systems facing complex tasks

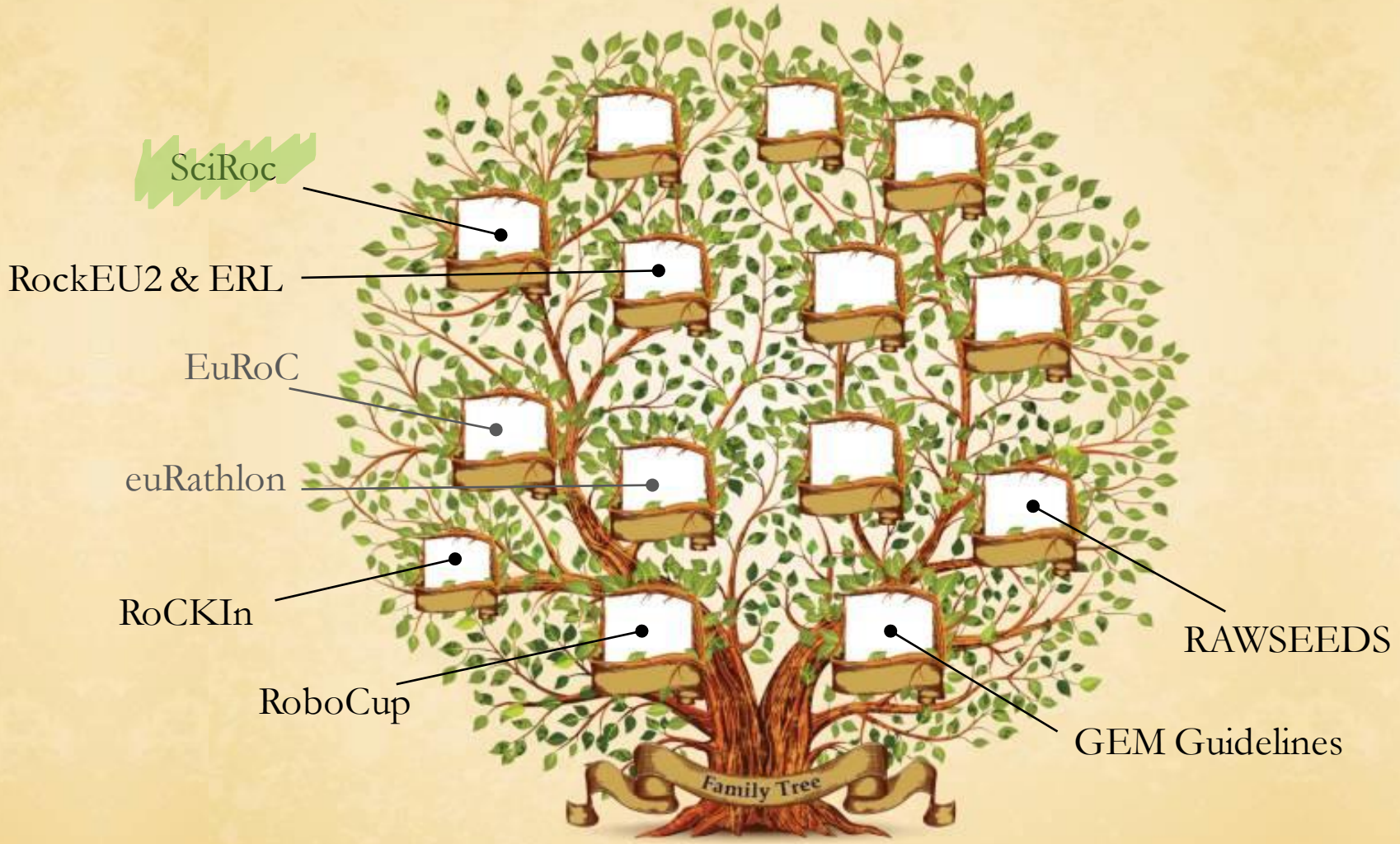
"Robots are doomed to close the loop"



This integrated approach is what the RoCKIn PO dubbed as «RoCKIn legacy» !!!

Functionality Benchmarks (FBM) evaluate performance of modules for specific functionalities

Task Benchmarks (TBM) assess performance of integrated robot systems facing complex tasks

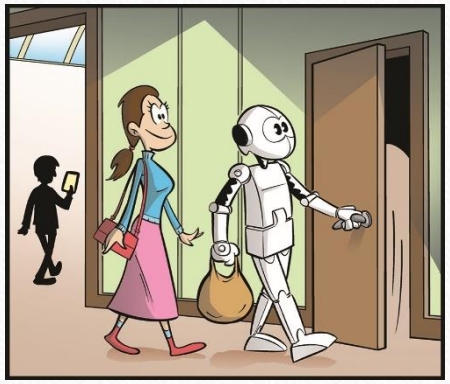

SciRoc
RockEU2 & ERL
EuRoC
euRathlon
RoCKIn
RoboCup
RAWSEEDS
GEM Guidelines
Family Tree

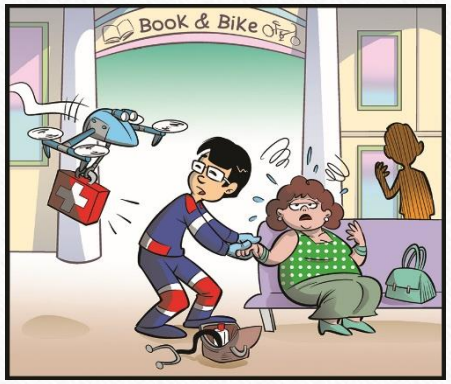
The Sciroc Idea

Sciroc Episodes

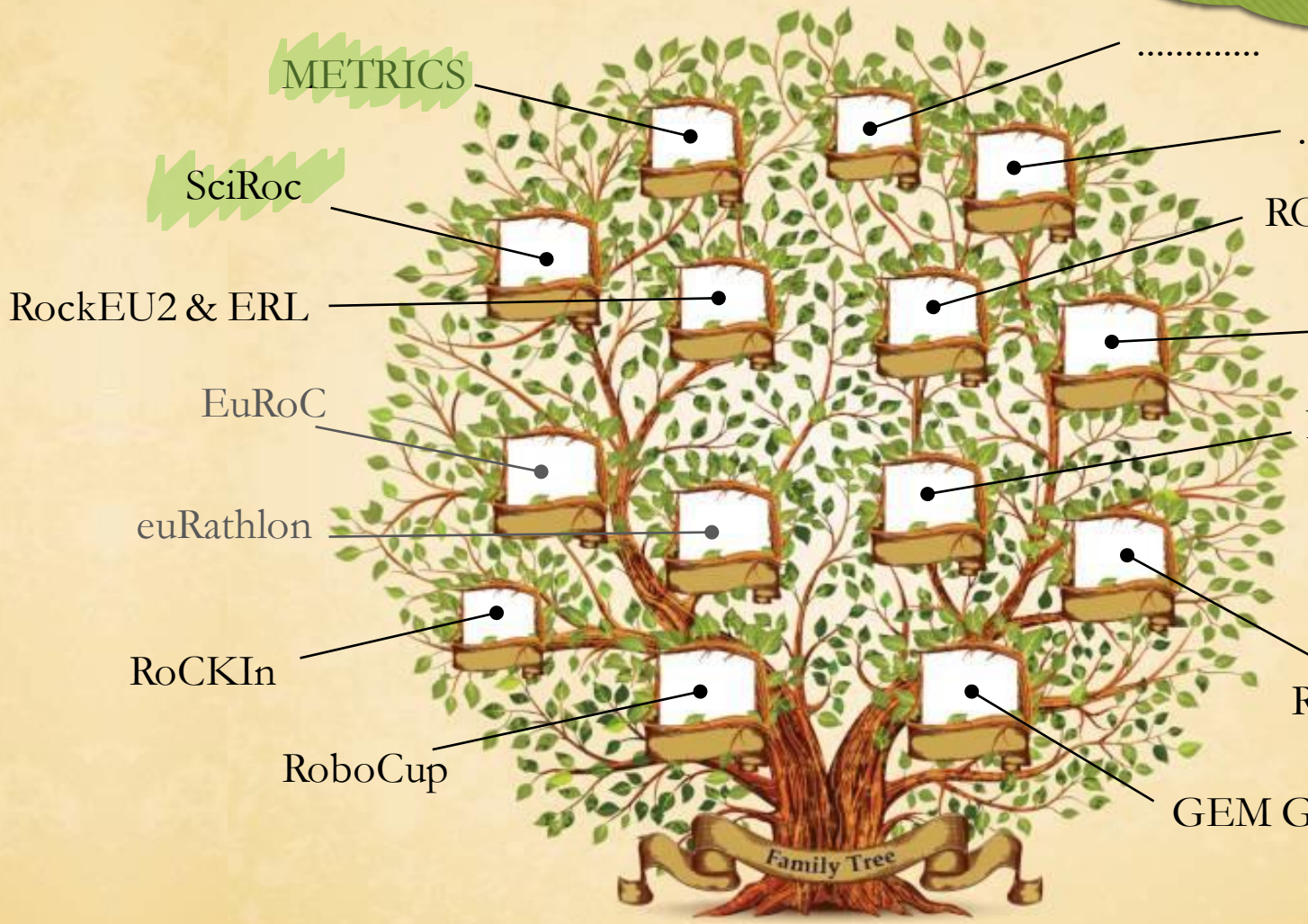
Human-robot interaction in social environments as well as social impact are the focus

- Deliver coffee shop orders
- Talk the elevator
- Shopping pick and pack
- Open the door
- Fast delivery of emergency pills





Still lot to do from a methodological/metrological perspective …



METRICS

SciRoc

RockEU2 & ERL

EuRoC

euRathlon

RoCKIn

RoboCup

.....

.....

ROSE Challenge

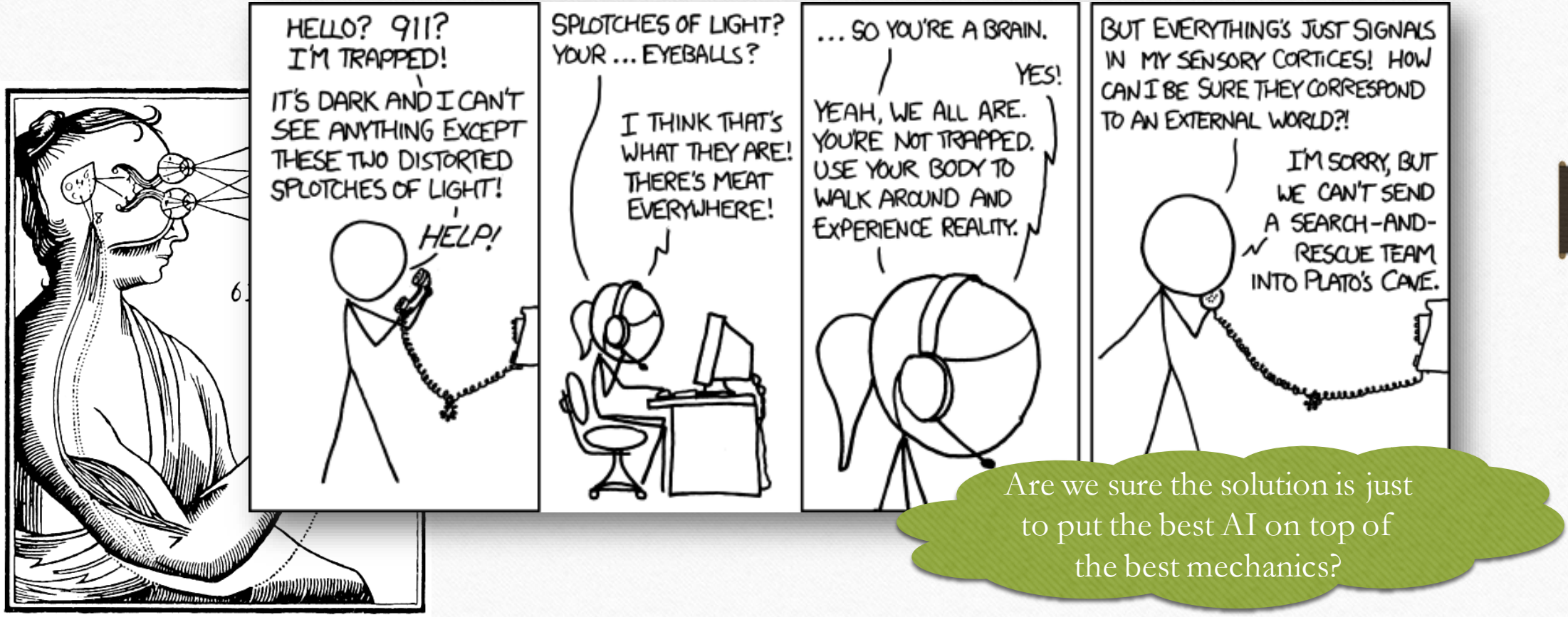
EUROBENCH

Plug & Bench (RobMoSys)

RAWSEEDS

GEM Guidelines

Back to the Body and Mind Problem …



Are we sure the solution is just to put the best AI on top of the best mechanics?

“Benchmarking via Challenges & Competitions”

… the RoCKIn legacy

Thank you!

Matteo Matteucci, matteo.matteucci@polimi.it

Dept. of Electronics, Information and Bioengineering

Politecnico di Milano