An introduction to Evaluation of Knowledge Processing Technologies

Edouard Geoffrois
French National Defence Project Agency (DGA)
& French National Research Agency (ANR)

CLARA Course
Paris, Nov 26th, 2012

Covered domains and sub-domains

- Natural language processing
  - Topic detection, Named Entity detection, Question answering, dialogue, summarization, translation
- Speech processing
  - Language recognition, speaker recognition, transcription
- Image processing
  - Detection and recognition of persons, objects, movements, attitudes, situations
- Scanned document processing
  - Language recognition, writer recognition, handwriting recognition
- Audio-visual document processing, information fusion
- Etc...
  - Behaviour analysis, inconsistency detection...

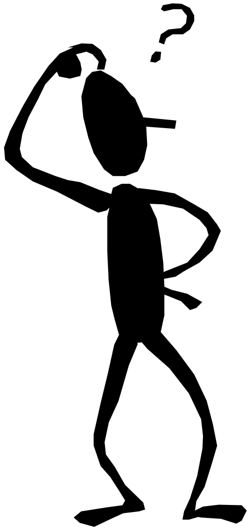
Does it work?

☺ “It works, I've seen a product in a shop.”

☺ “I've read that a start-up has solved the problem.”

“It has been 30 years that it is expected for next year” ☹

“This is just science-fiction” ☹



How can we really know?

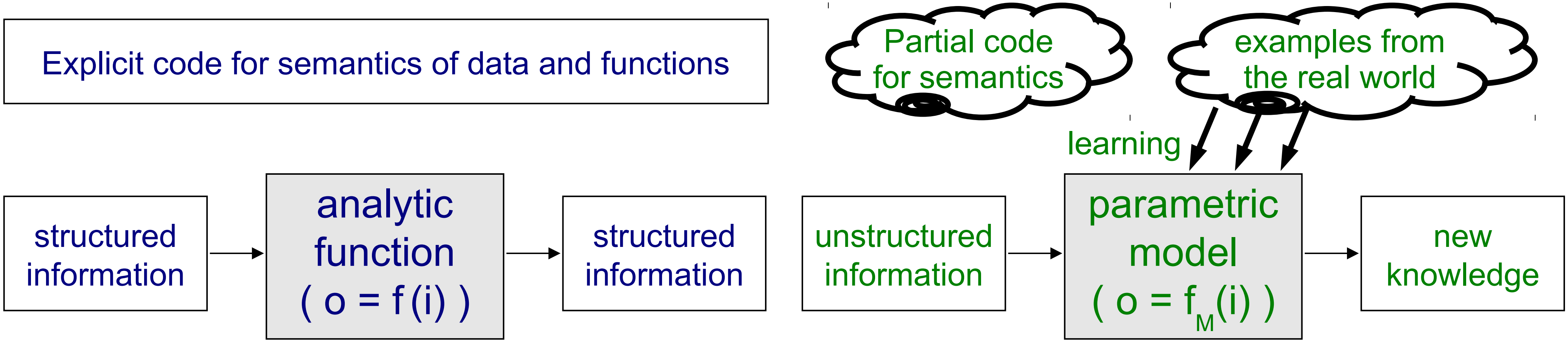
Questions

- How to evaluate knowledge processing technologies?
- How useful is evaluation?
- How much does it cost?
- Who should care?

Induced questions

- How to evaluate knowledge processing technologies?
  - What are the different types of evaluation?
  - Why is a specific organization needed?
  - What is specific to the domain of knowledge processing?

Structured vs. unstructured information


Explicit code for semantics of data and functions
Partial code for semantics
examples from the real world
learning
structured information
analytic function (o = f (i))
structured information
unstructured information
parametric model (o = f_M(i))
new knowledge

The data express the semantics through an *explicit* code	The data is not enough to derive the semantics, which are partially *implicit*
The data are *transformed* using an explicit mathematical function (rules, etc.)	The data are *interpreted* using a mathematical model of the world (probabilities, etc...)
Theoretical approach (model is the mathematical proof)	*Experimental* approach (model is natural science)

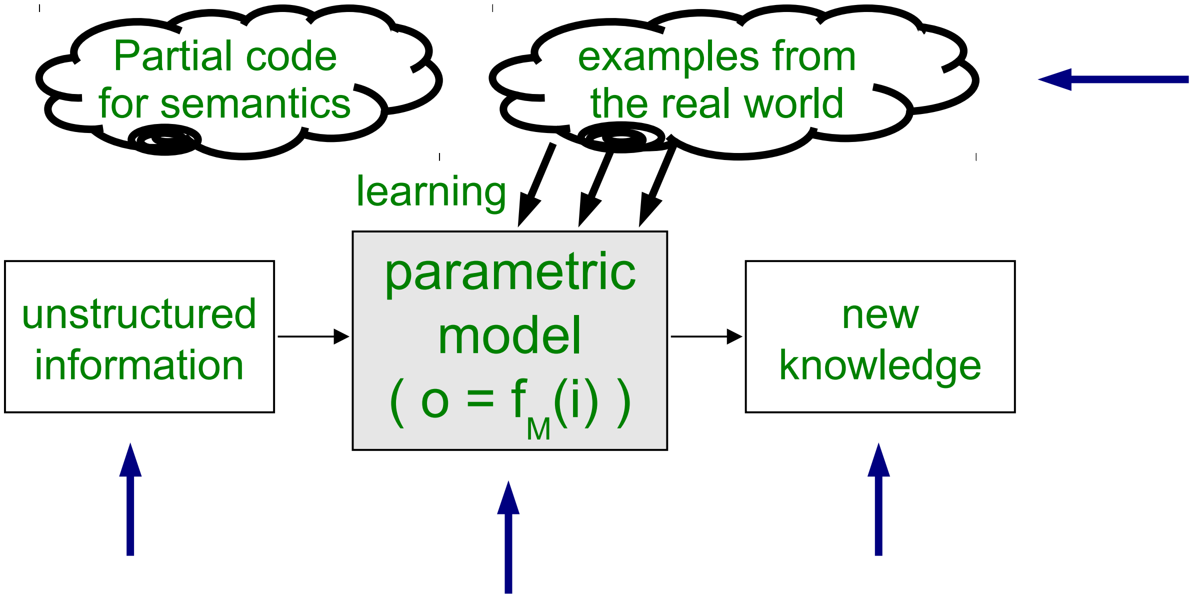
Trigger keywords: *data* processing, *computing*

Trigger keywords: *intelligent / semantic* processing of digital / multimedia *content / knowledge*

Examples of domains: *formal languages*, traditional *signal processing*

Examples of domains: *natural language and speech* processing, *scanned documents, image and video* processing, information *fusion*

Need n°1: Manually annotated data


Partial code for semantics
examples from the real world
learning
unstructured information
parametric model (o = f$_M$(i))
new knowledge

A task is defined by a representative sample data set

A good model should agree well with the observed data

Data is also important for training models

Example of metric
(for speech transcription)

“I *would like to go to London tomorrow morning hum*”

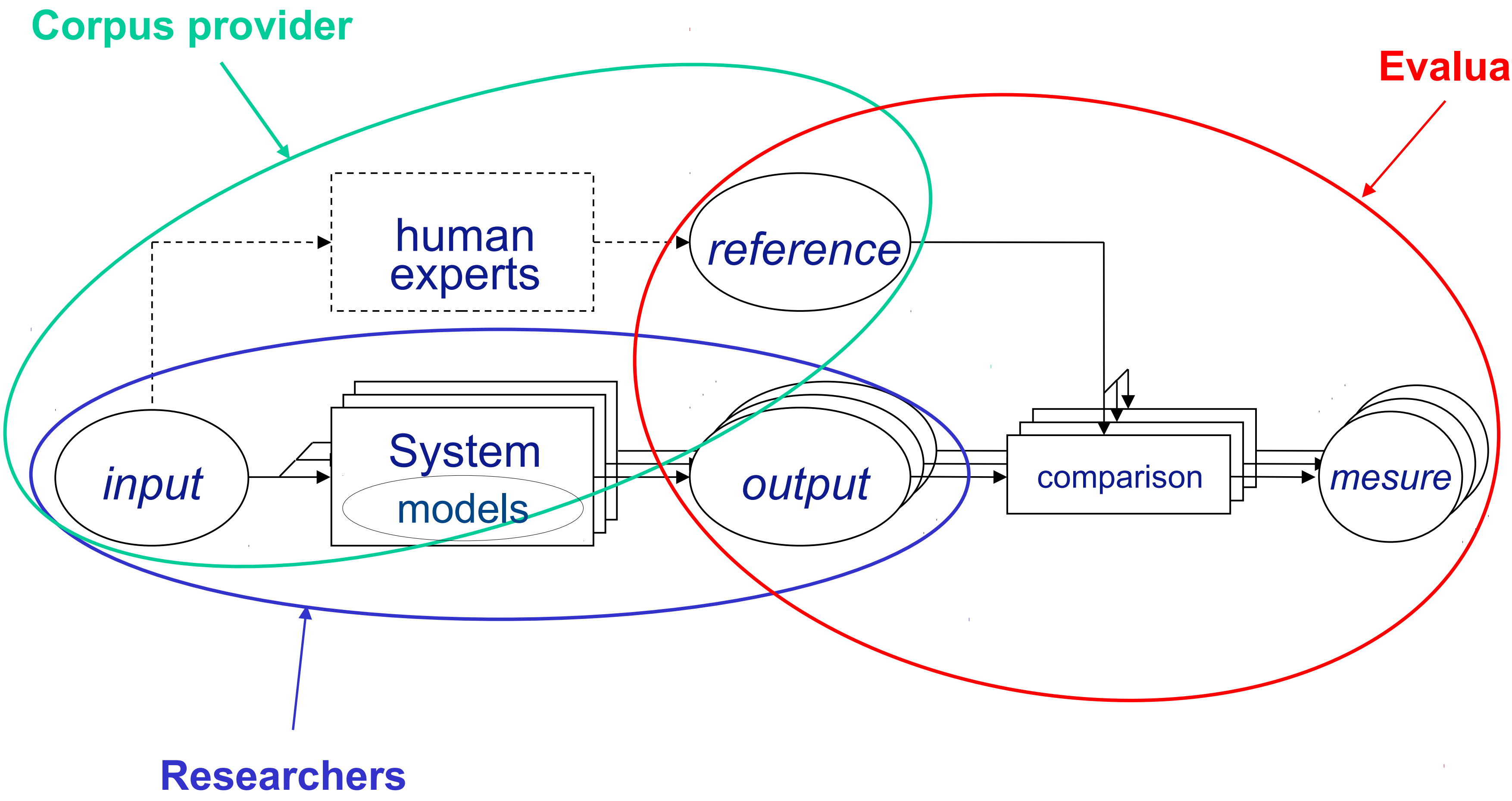
```
I will like to go to lone done tomorrow morning
```

Error rate = (2+1+1)/10 = 40%

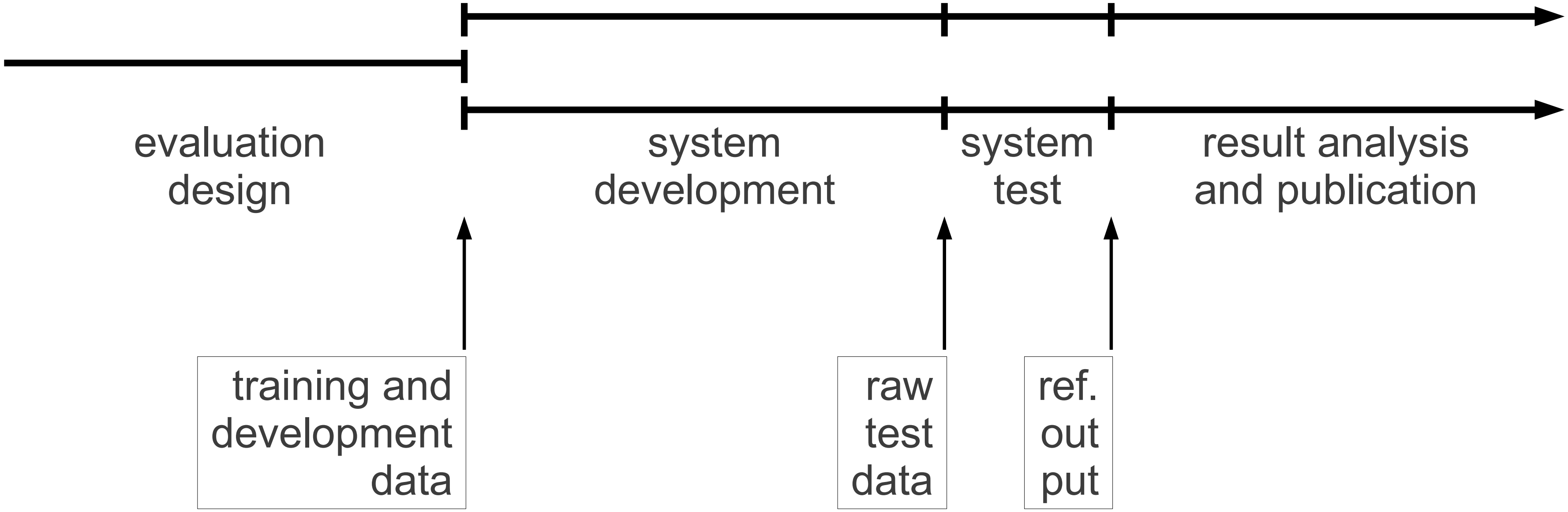
... or ... (2+1)/10 = 30%

Error rate = edit distance between an hypothesis and a reference or a set of references

Evaluation data flow



Need n°2: Synchronized evaluations



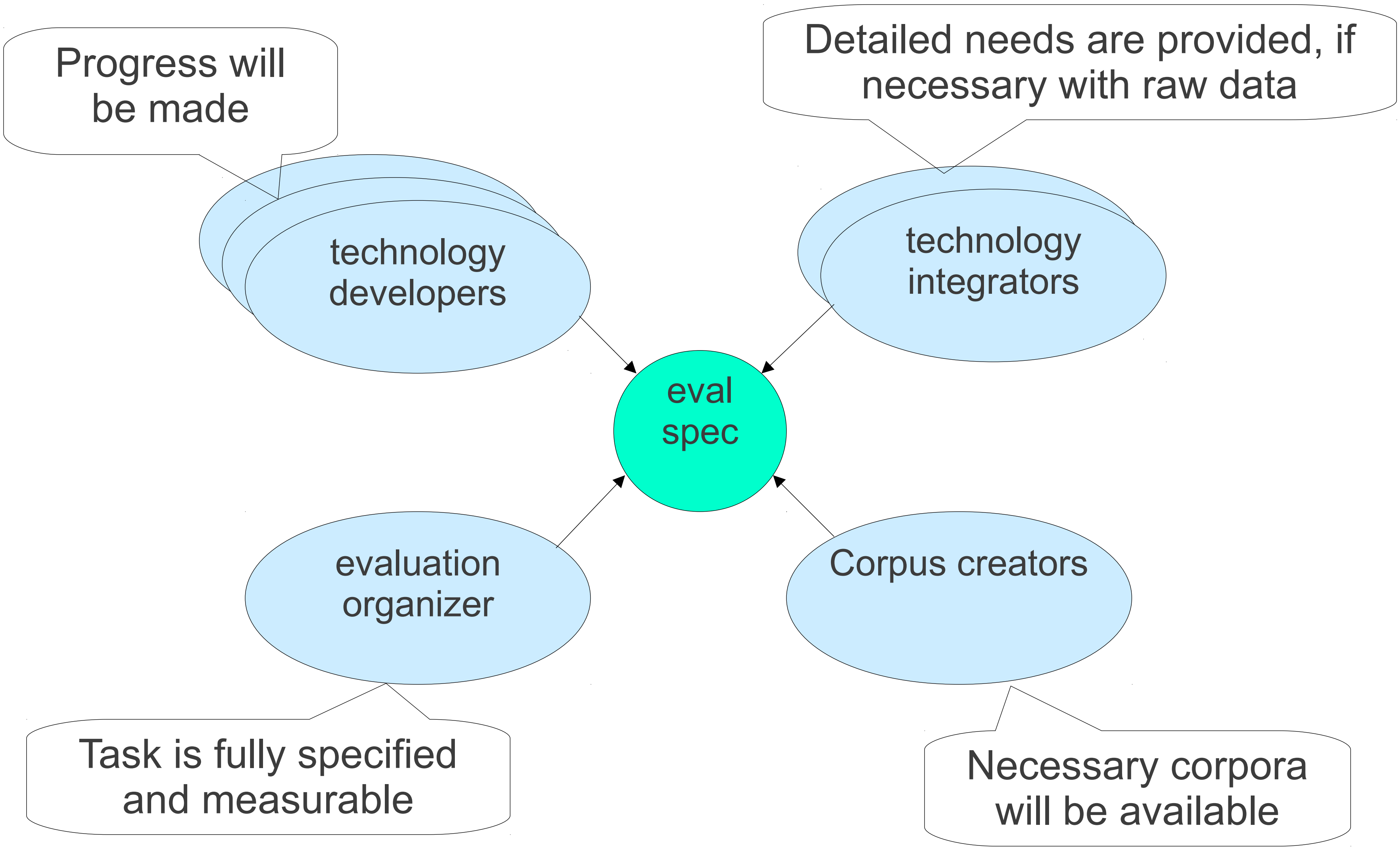
Data should be shared for the sake of reproducibility

Tests should occur almost simultaneously to avoid bias

Evaluation design should serve the community

→ Evaluation campaigns

Coordination of technology development



Progress will be made

Detailed needs are provided, if necessary with raw data

technology developers

technology integrators

eval spec

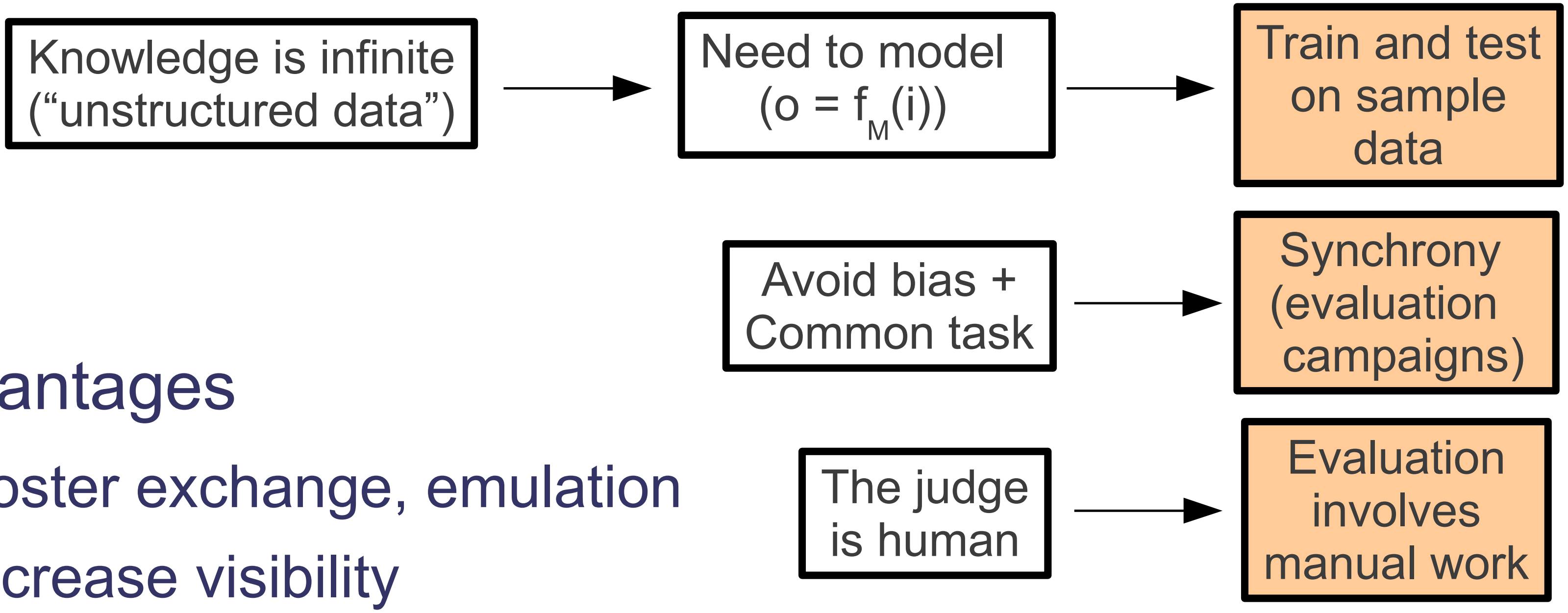
evaluation organizer

Corpus creators

Task is fully specified and measurable

Necessary corpora will be available

Specificities of evaluation for content processing technologies



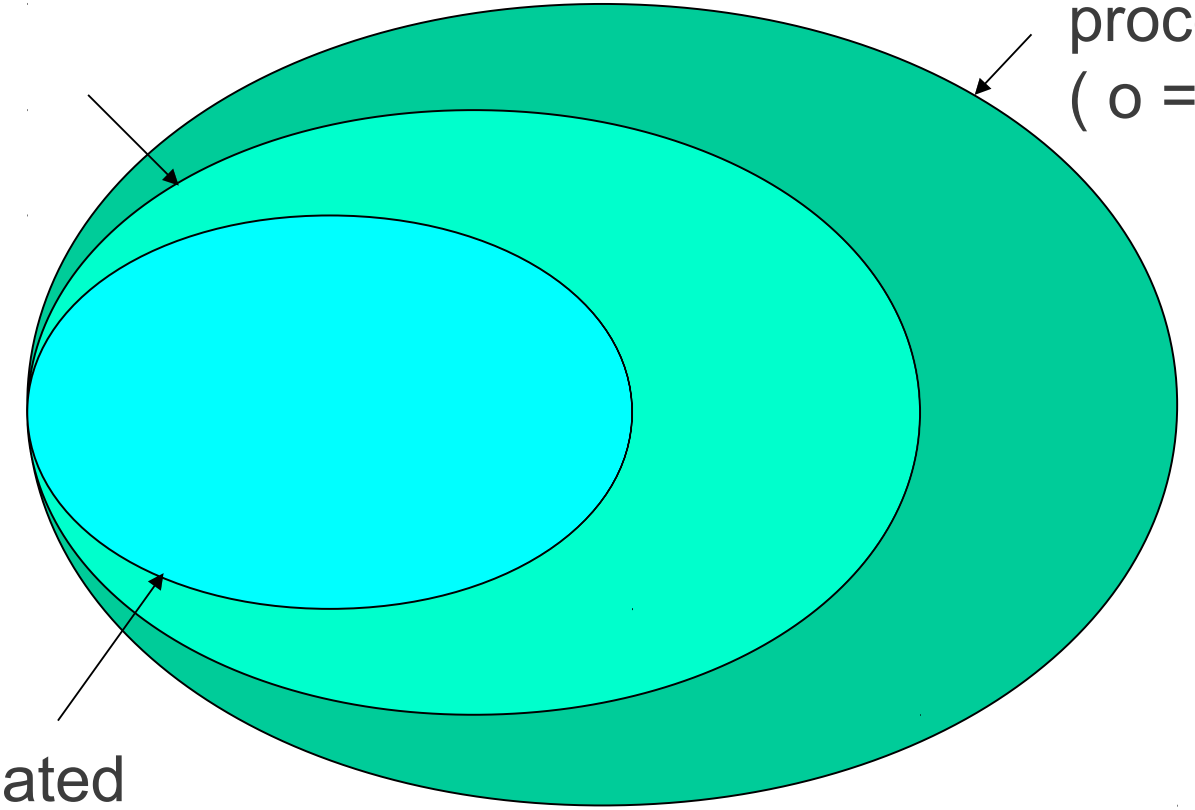
Knowledge is infinite (“unstructured data”) → Need to model $(o = f_M(i))$ → Train and test on sample data

Avoid bias + Common task → Synchrony (evaluation campaigns)

The judge is human → Evaluation involves manual work

- Advantages
  - Foster exchange, emulation
  - Increase visibility
- Inconvenients
  - Research constrained by deadlines
  - Research focused on topics of common interest

Perimeter



Unstructured information processing
(o = f_M(i))

Information processing
(o = f(i))

Actually evaluated unstructured information processing

Benefits of evaluation

1. Explicit problems
2. Validate new ideas
3. Identify missing science
4. **Compare approaches and systems**
5. Determine maturity for a given application
6. Facilitate technology transfer
7. Incite innovation
8. **Organise the community**
9. Support competitiveness
10. Assess public funding efficiency

The power of evaluation

Before

After

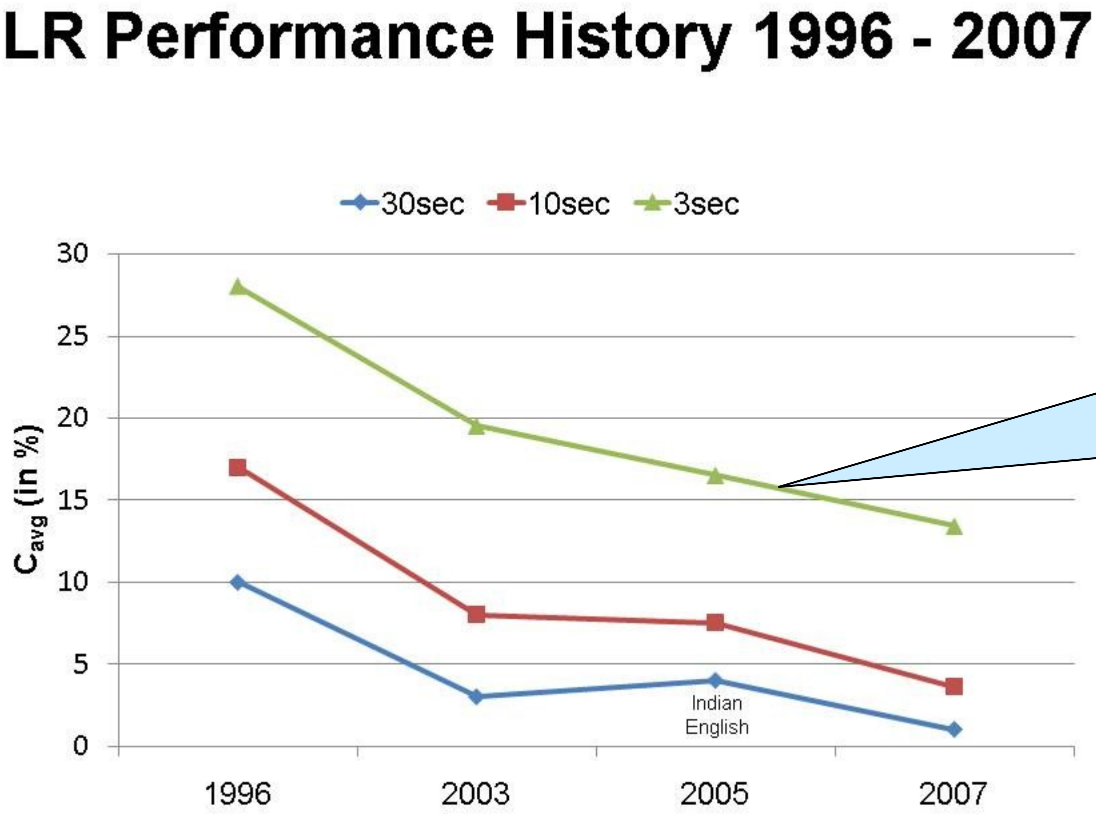
History

Late 70's	**NATO** Research Study Group on Automatic **Speech Recognition** (ASR) produces a common benchmark database in several languages
Mid 80's	After failure of earlier programs, the **US** (DARPA ans NIST) introduce systematic objective performance measurement in ASR programs
Early 90's	DARPA and NIST extend evaluation to automatic **Textual information processing** (TIPSTER program, then TREC, MUC, DUC, …) and opens its evaluation campaings to non-US participants
Mid 90's	First **European** program including evaluation (SQALE program on ASR)
Late 90's	First **French** evaluation program on speech and language processing, followed by a larger one in the early 2000's (Technolangue) First **Japanese** evaluation on information retrieval (NTCIR)
2001	DARPA and NIST extend evaluation to **Machine Translation**
2003	The major European programs on language processing (TC-STAR, CHIL) include evaluation
Mid 2000's	Evaluation methodology gradually extends to **Image processing** (TRECVid, US-EU CLEAR evaluations, French Techno-Vision program, ...)

Examples of evaluation campaigns today

Funding	Organisers	Name	Topic
DARPA, DoC	NIST	Rich Transcription	Speech transcription
DARPA, DoC	NIST	Text REtrieval Conference	Documents retrieval
DARPA, DoC	NIST	OpenMT	Translation
DoC, ...	NIST, ...	TRECVid	Video analysis
DoC, IARPA, FBI	NIST	SRE, LRE	Speaker and language recognition
DoD	NIST	Text Analysis Conference	Natural language
NII, NICT, U. Tokyo	NII, NICT, U. Tokyo	NTCIR	Information retrieval
EU	U. Pisa, Delft, ...	CLEF, MultiMediaEval	Crosslingual, ...
OSEO	DGA, LNE, IRIT, UJF, LIPN, GREYC	Quaero	Multimedia document processing
DGA	DGA	RIMES, ICDAR	Handwriting recognition
DGA	LNE	REPERE	Multimodal person reco
Trento	CELCT, ...	Evalita	Natural language

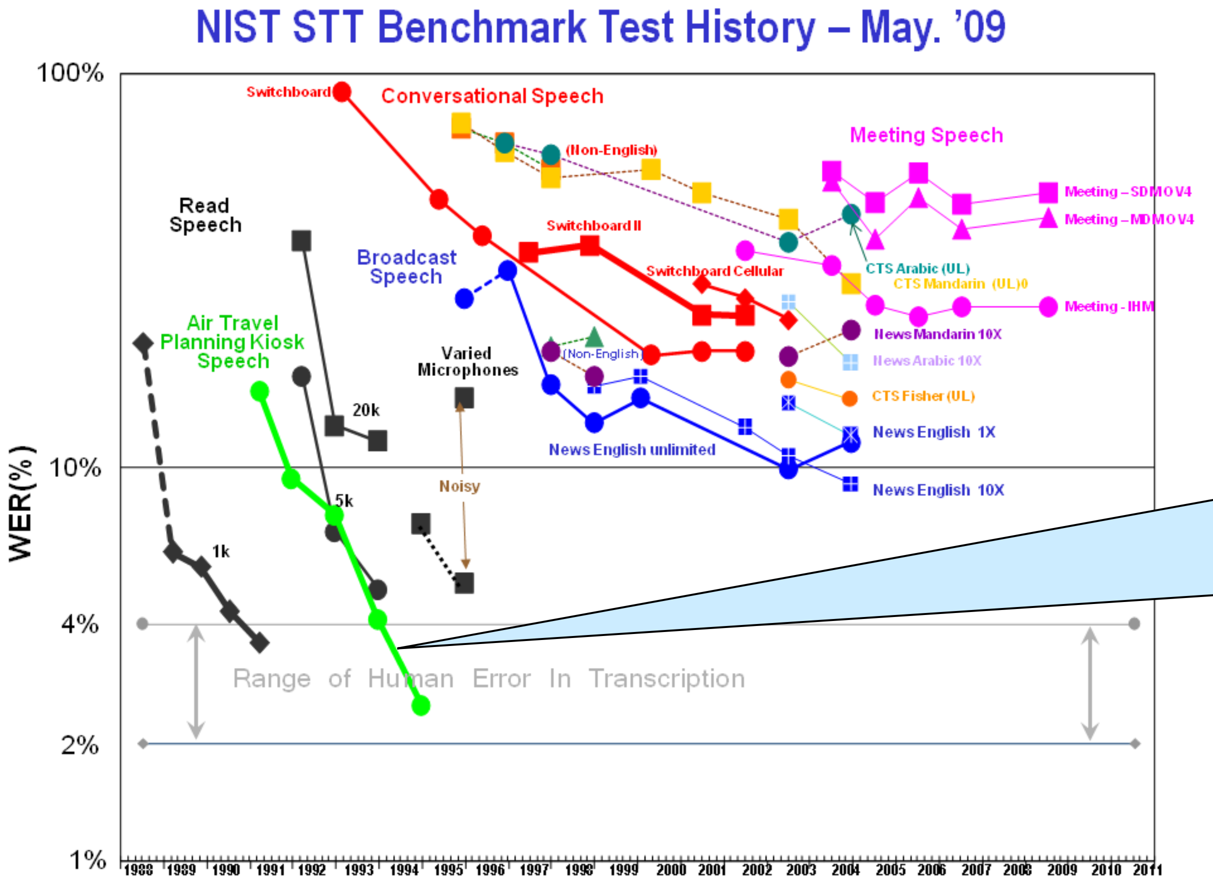
Impact on the evolution of performances (example of spoken language recognition)


LR Performance History 1996 - 2007

Evolution of the error rate of the best system over the years

Source : NIST

Impact on the evolution of performances (example of speech transcription)



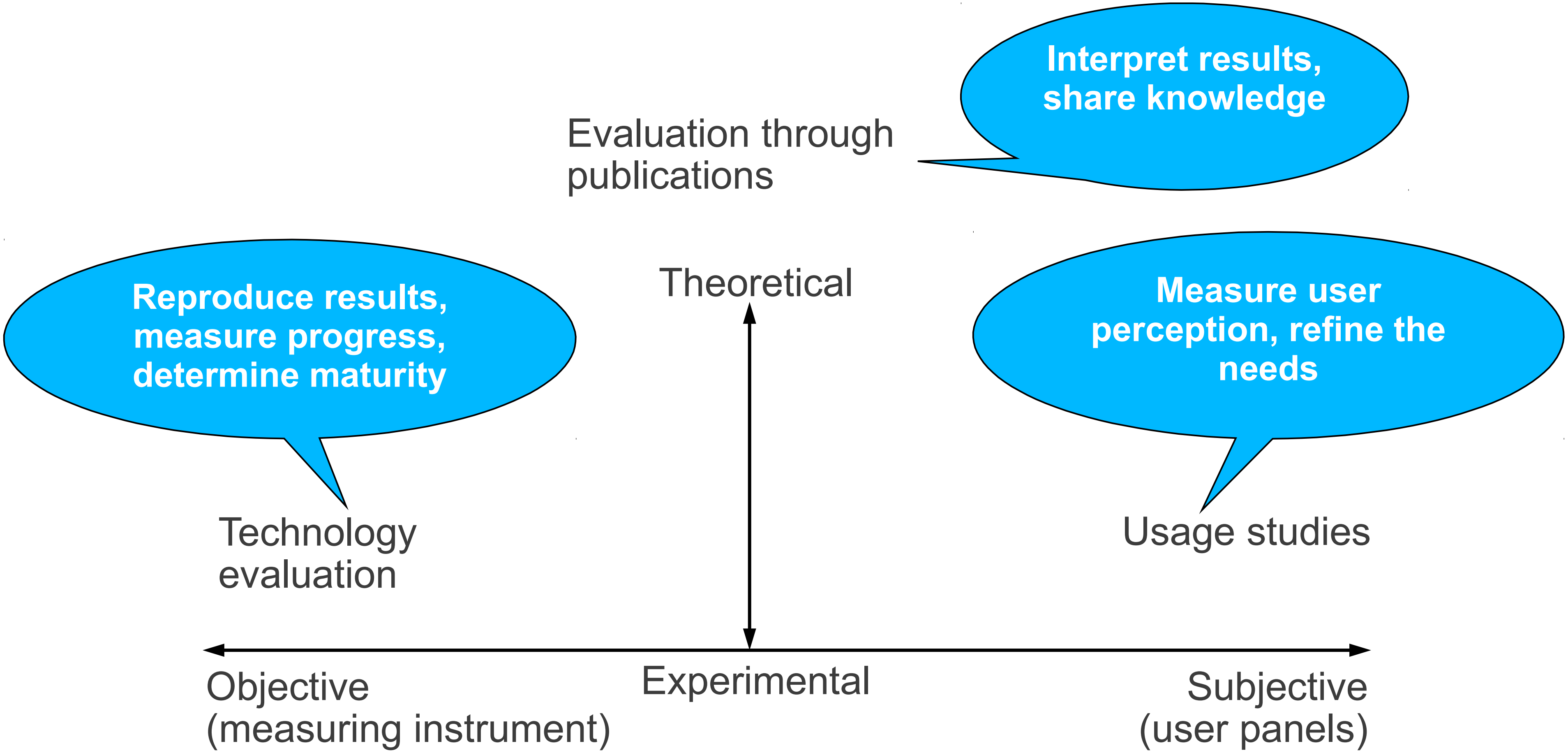
When a problem (one colored curve) is considered as solved, move on to a more difficult one

Source : NIST

Issues

- Why evaluate?
  - *“We did without it until now. Why change?”*
  - *“It is not a research activity. Why bother?”*
  - *“It creates additional constraints...”*
- How to evaluate?
  - *“It works on the examples shown in the demonstration.”*
  - *“The algorithm is mathematically proven. Isn't that enough?”*
  - *“We conducted user tests. Isn't that enough?”*
  - *“There are publications. Isn't that enough?”*
- Why so much debate?
  - A relatively young science with an even younger metrology
  - A relatively unknown economic model

Technology evaluation vs. usage studies



Evaluation through publications

Interpret results, share knowledge

Theoretical

Reproduce results, measure progress, determine maturity

Measure user perception, refine the needs

Technology evaluation

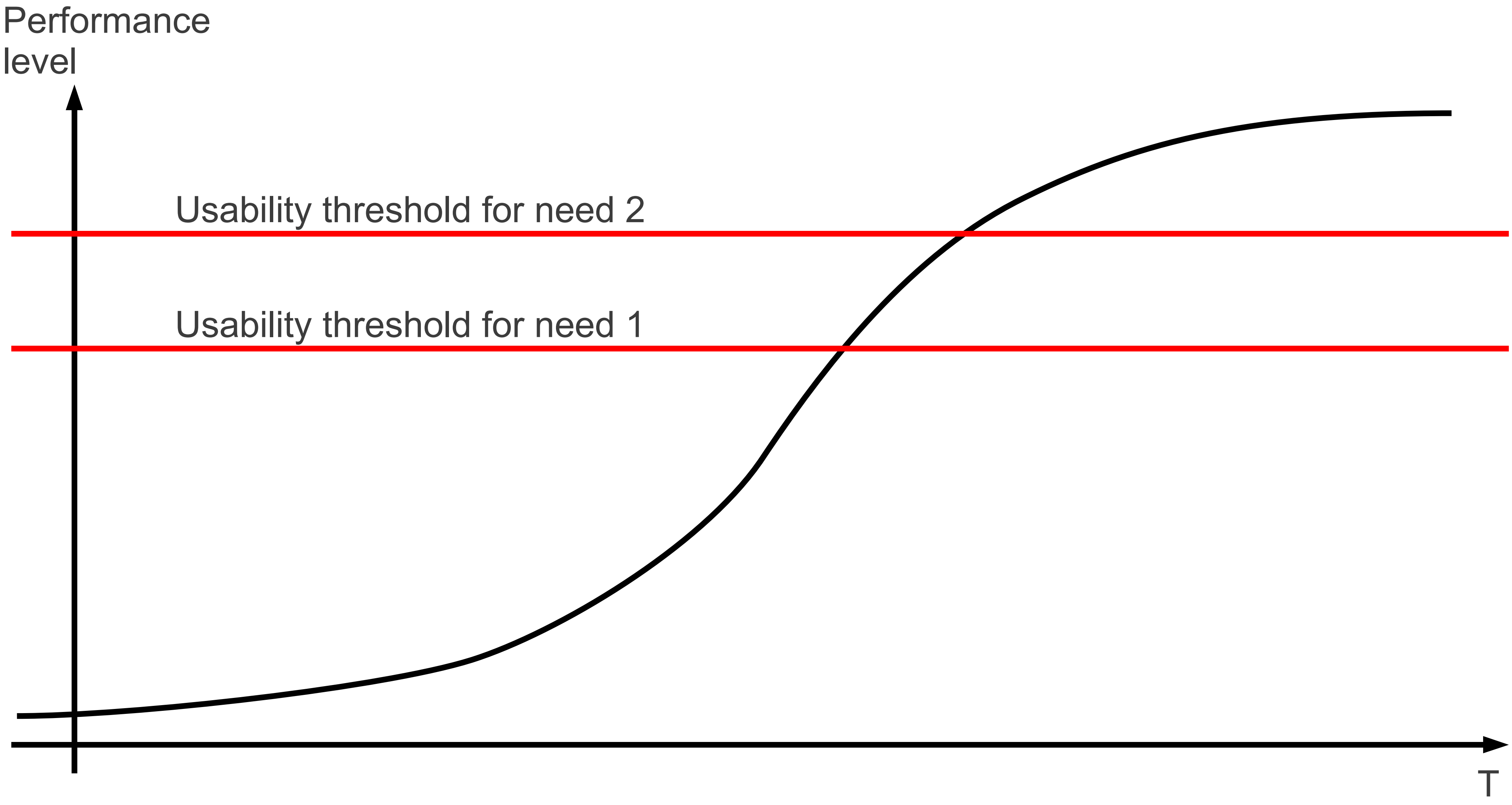
Usage studies

Objective (measuring instrument)

Experimental

Subjective (user panels)

Technology performance vs. satisfaction of user need



Need for a strong incentive

- A critical component...
  - It represents only a few % of the investments
  - It dramatically increases the return on these investments
- … which must be funded by those who want to see the field make progress as a whole...
  - Campaigns must be organized regularly to measure progress
  - Most of the costs are fixed ones
  - The infrastructure must be open to all to support scientific progress
  - There is no direct return on investment for the party doing the measurements
- … and must be prepared early in project design
  - Data, evaluation and R&D activities are tightly linked and should be jointly designed in integrated projects

Private vs. public goods

non-excludable	Common goods (e.g., fish stocks, timber, coal)	Public goods (e.g., free-to-air television, air, national defense) Corpus paid by public funding and distributed without a fee
excludable	Private goods (e.g., food, clothing, car, personal electronics) Corpus paid by a company for it own purpose and not distributed	Club goods (e.g., cinema, private parks, satellite television) Corpus sold for a fee
	rivalrous	non-rivalrous

Conclusions

- A relatively large but homogeneous domain
  - characterised by the interpretation of data using a model of the world to create new knowledge,
- with a need for manually annotated data
  - representative of the task under study
- and for synchronised evaluations
  - in the form of evaluation campaigns,
- both deserving special attention
  - to really happen and serve the research needs

Thank you for you attention!