CSE Framework: A UIMA-based Distributed System for Configuration Space Exploration

Elmer Garduno[2], Zi Yang[1], Yan Fang[3], Avner Maiberg[1], Collin McCormack[4], Eric Nyberg[1]

[1)] Carnegie Mellon University {ziy, amaiberg, ehn}@cs.cmu.edu
[2)] Sinnia elmerg@sinnia.com
[3)] Oracle Corporation yan.fang@oracle.com
[4)] The Boeing Company collin.w.mccormack@boeing.com

Carnegie Mellon University
School of Computer Science

Motivation

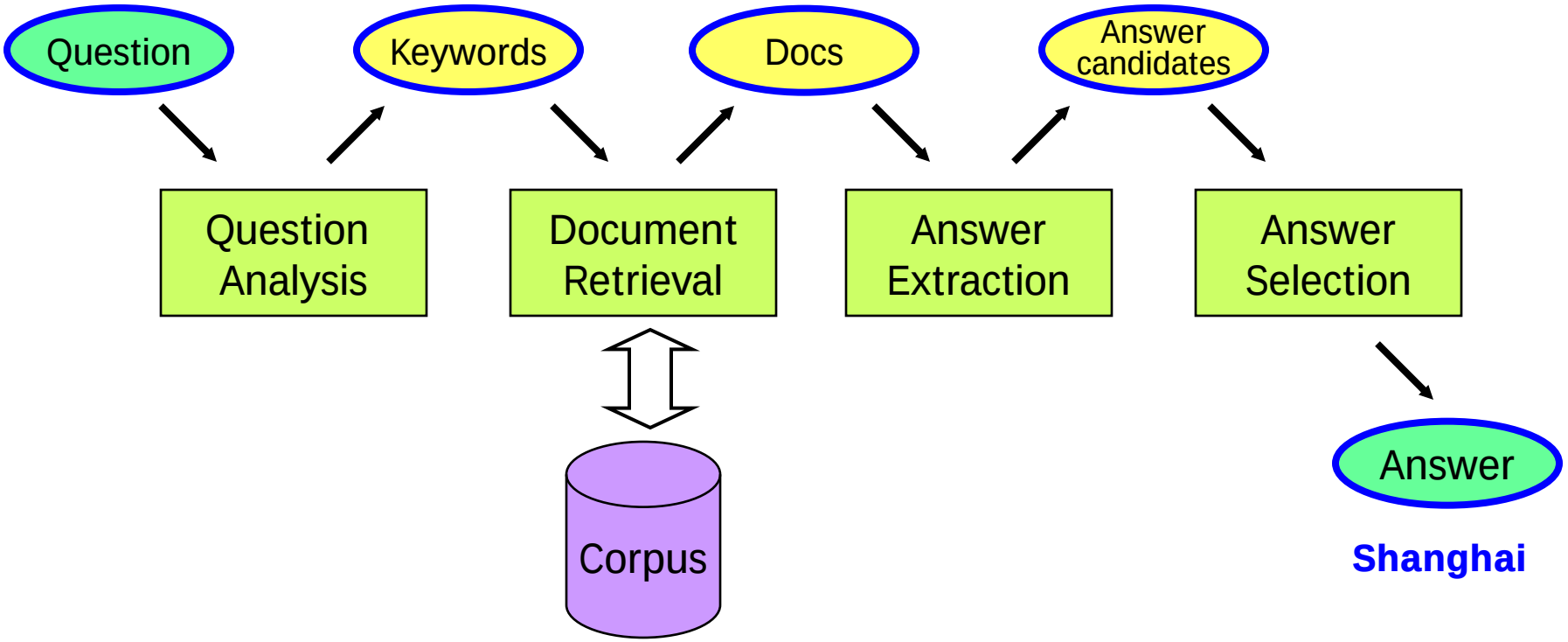
Typical QA Pipeline

Document ID	Rank
FBIS3-58 (relevant**)**	1
AP880603-0268	2
WSJ920110-0013	3
FBIS3-45320 (relevant**)**	4
FT942-2016	5

Answer candidates	Score	Document extracted
Beijing	0.7	AP880603-0268
Hong Kong	0.65	WSJ920110-0013
Shanghai	0.64	FBIS3-58
Taiwan	0.5	FT942-2016
Shanghai	0.4	FBIS3-45320

Which city in China has the largest number of foreign financial companies?

Keywords: China largest foreign financial company
Answer type: location (city)

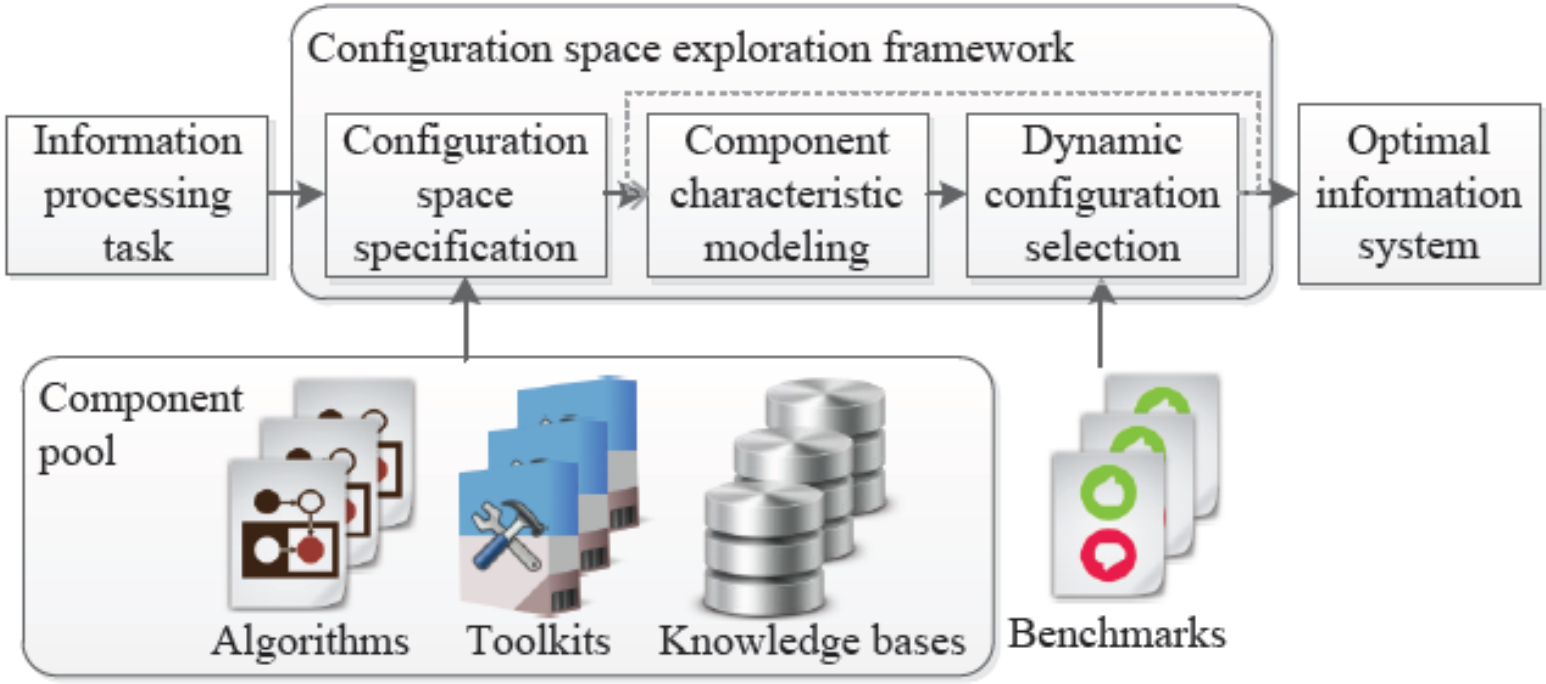


CURRENT RESEARCH IN QA

What did we learn from Watson?

- QA systems can be fast enough, accurate enough, and confident enough to perform in the real world
- Key factors:
  - Scalable, parallel architecture
  - Agile, open advancement process
- Next big challenge: *rapid domain adaptation*

Automatic Optimization of QA for TREC Genomics Questions



Results of Automatic Optimization

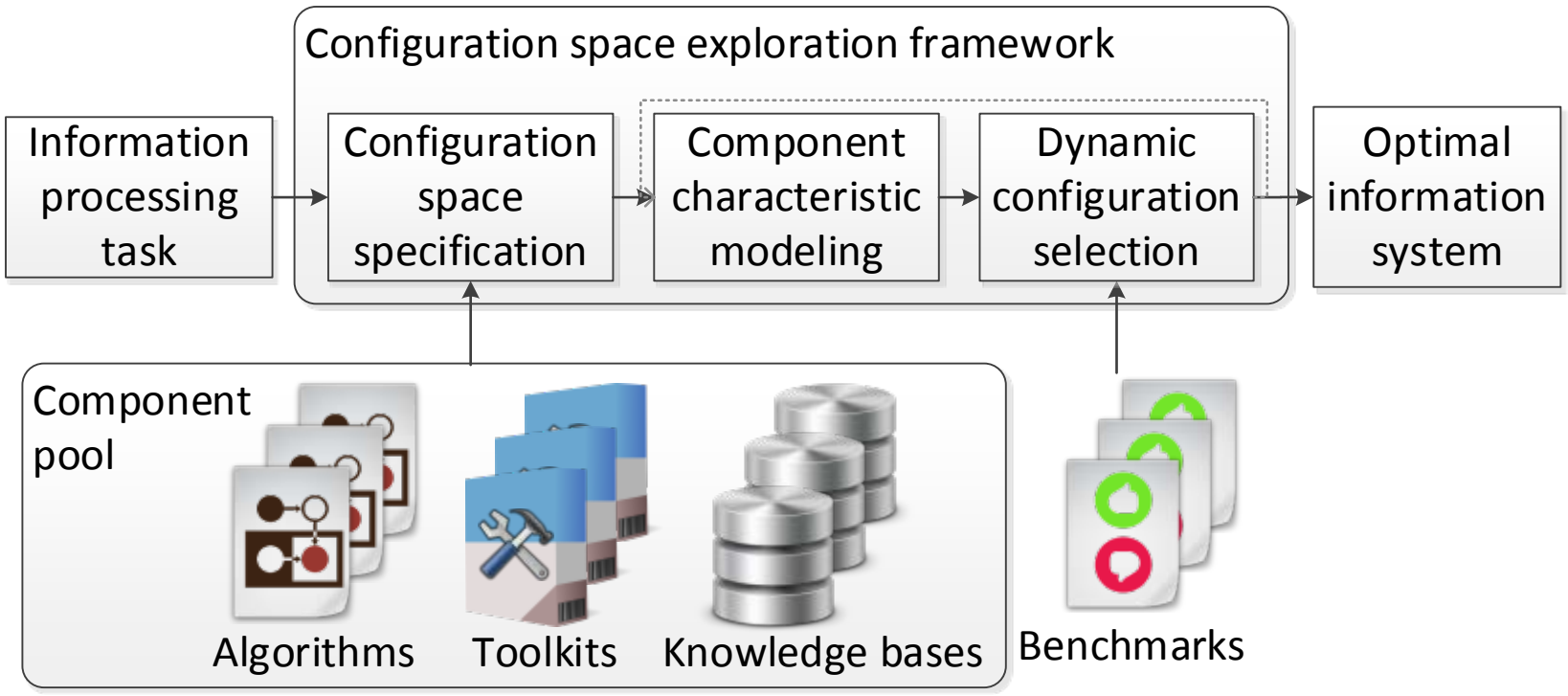
		Participants	CSE	Scaled CSE
# Component		~1,000	12	12
# Configuration		~1,000	32	2,946
# Trace		92	2,700	$\sim 1.426 \times 10^{12}$
# Execution		~1,000	190,680	$\sim 6.050 \times 10^{13}$
Capacity (hours)		N/A	24	24
DocMAP	Max	.5439	.5648	.5072
	Median	.3083	.4770	.3509
	Min	.0198	.1087	.2679
PsgMAP	Max	.1486[a]	.1773	.1181
	Median	.0345	.1603	.0713
	Min	.0007	.0311	.0164

[Yang, Z., Garduno, E., Fang, Y., Maiberg, A., McCormack, C. and Nyberg, E. (2013). "Building Optimal Information Systems Automatically: Configuration Space Exploration for Biomedical Information Systems", *Proceedings of the ACM Conference on Information and Knowledge Management*]

Automatically Building an Information System by Another Meta-System?

Building an Information System Automatically

- CSE framework

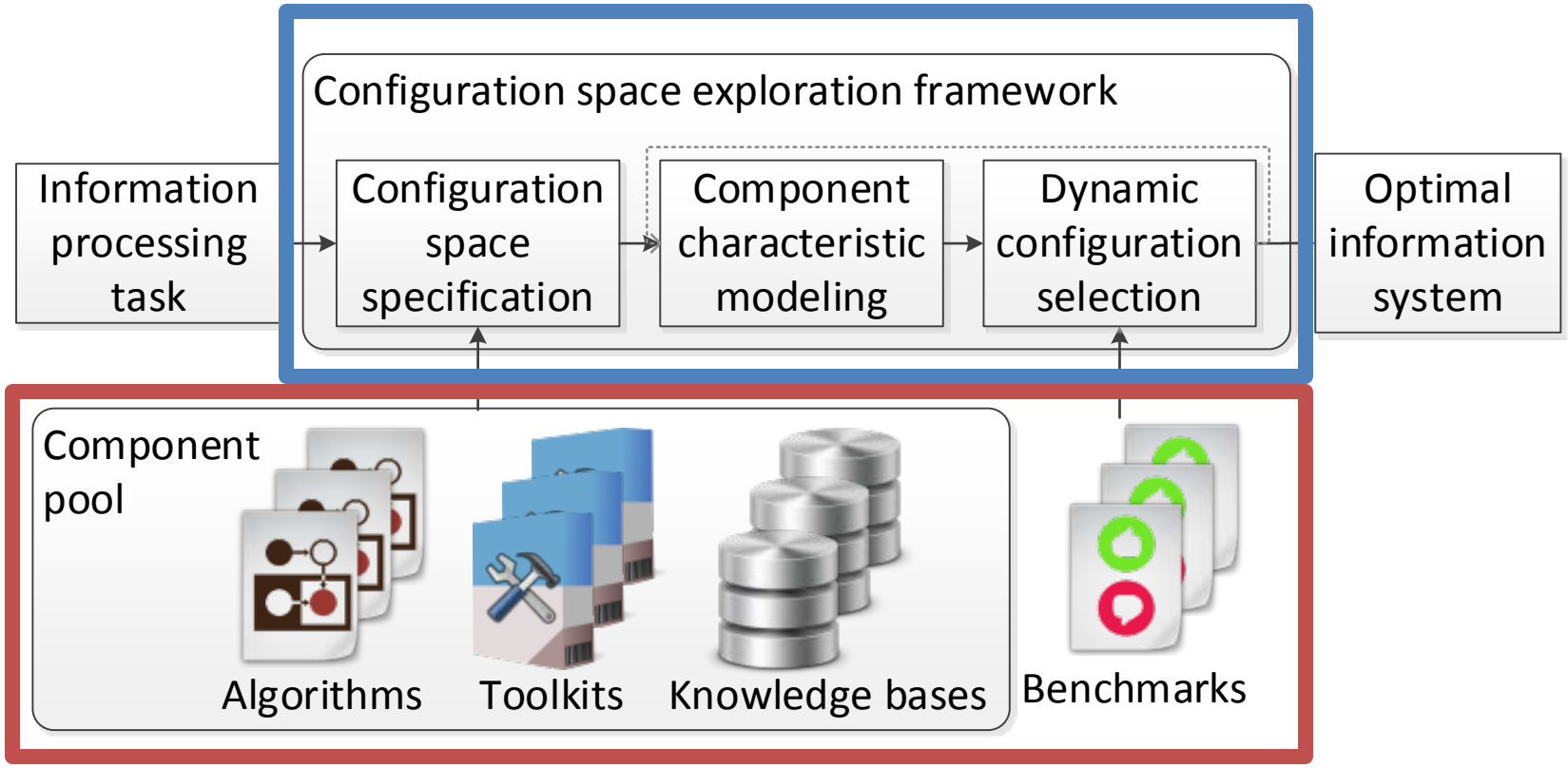


The *benefit* of CSE framework

- Accelerate the system development cycle by automating the component selection and tuning!
- Save *cost*!

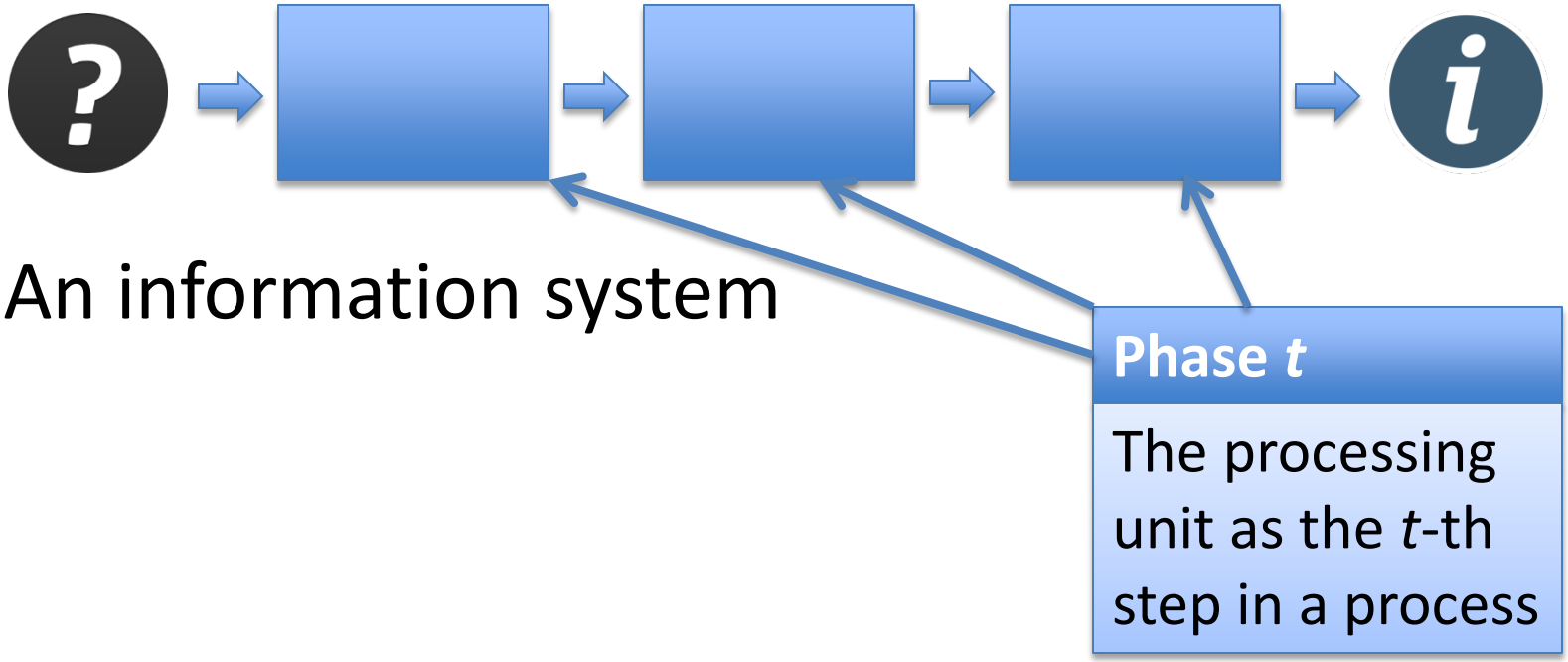
It requires

- Identify the tool, knowledge base, task algorithm candidates
- Provide information needs with known outcomes, e.g. answers to questions in the domain.



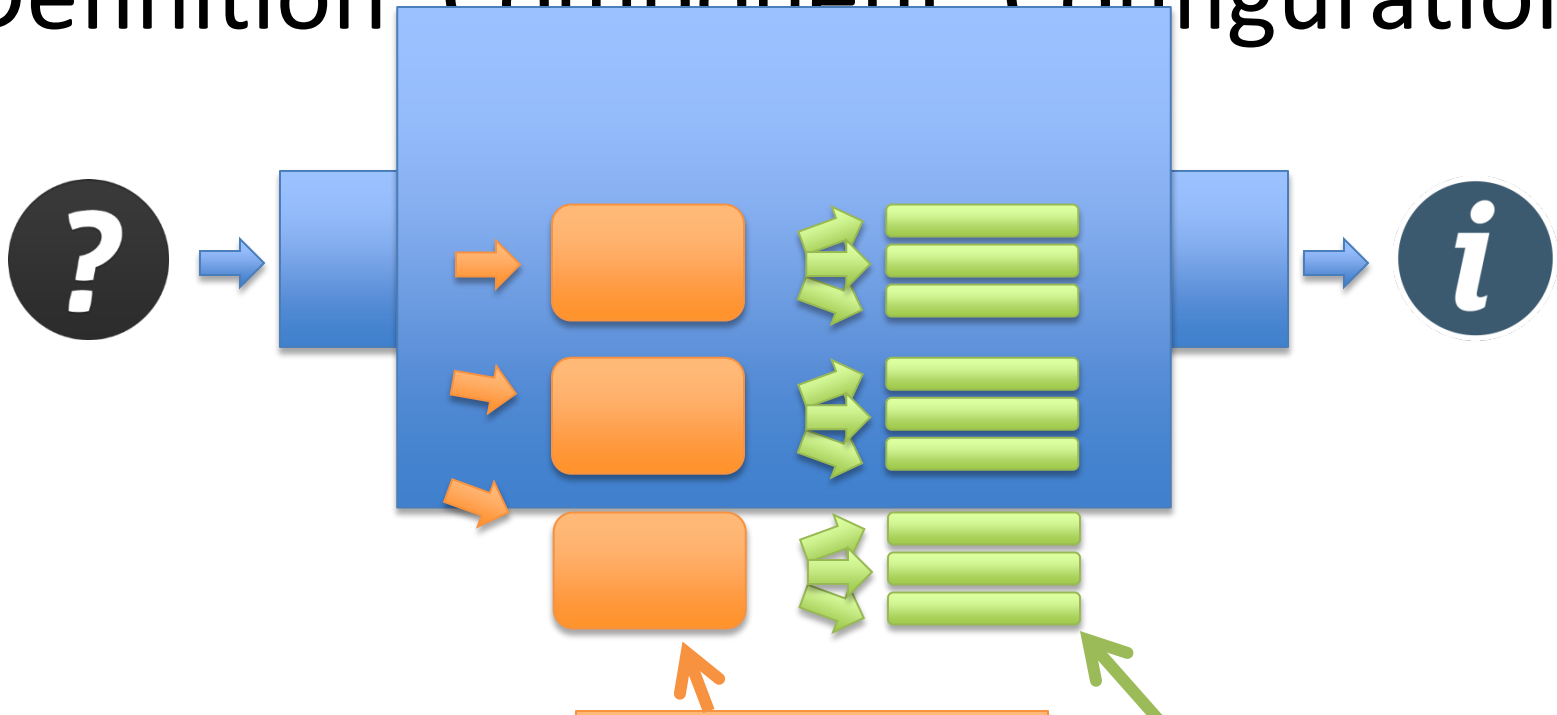
CSE - FRAMEWORK

Definition: Phase



- An information system

Definition: Component, Configuration



- Inside phase t

Component f_t^c

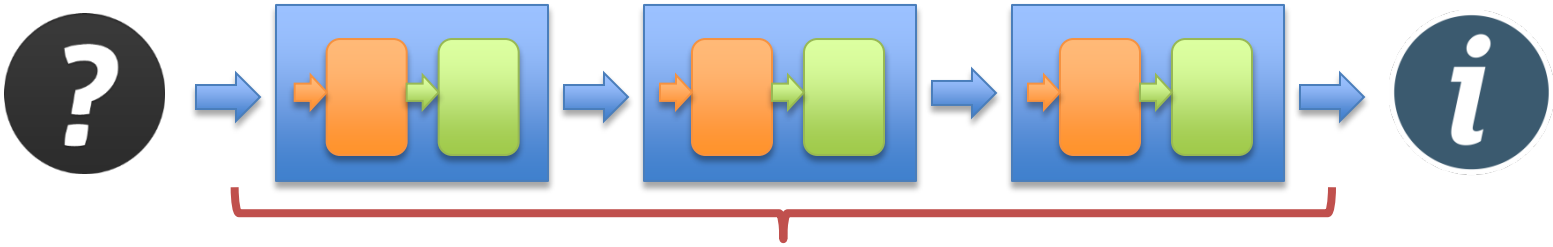
An instantiated processing unit in phase t

Configuration ω_t^c

Consists of a set of **parameters** $\{\omega_t^{c,p}\}_p$

Configured component $f_t^c|\omega_t^c$

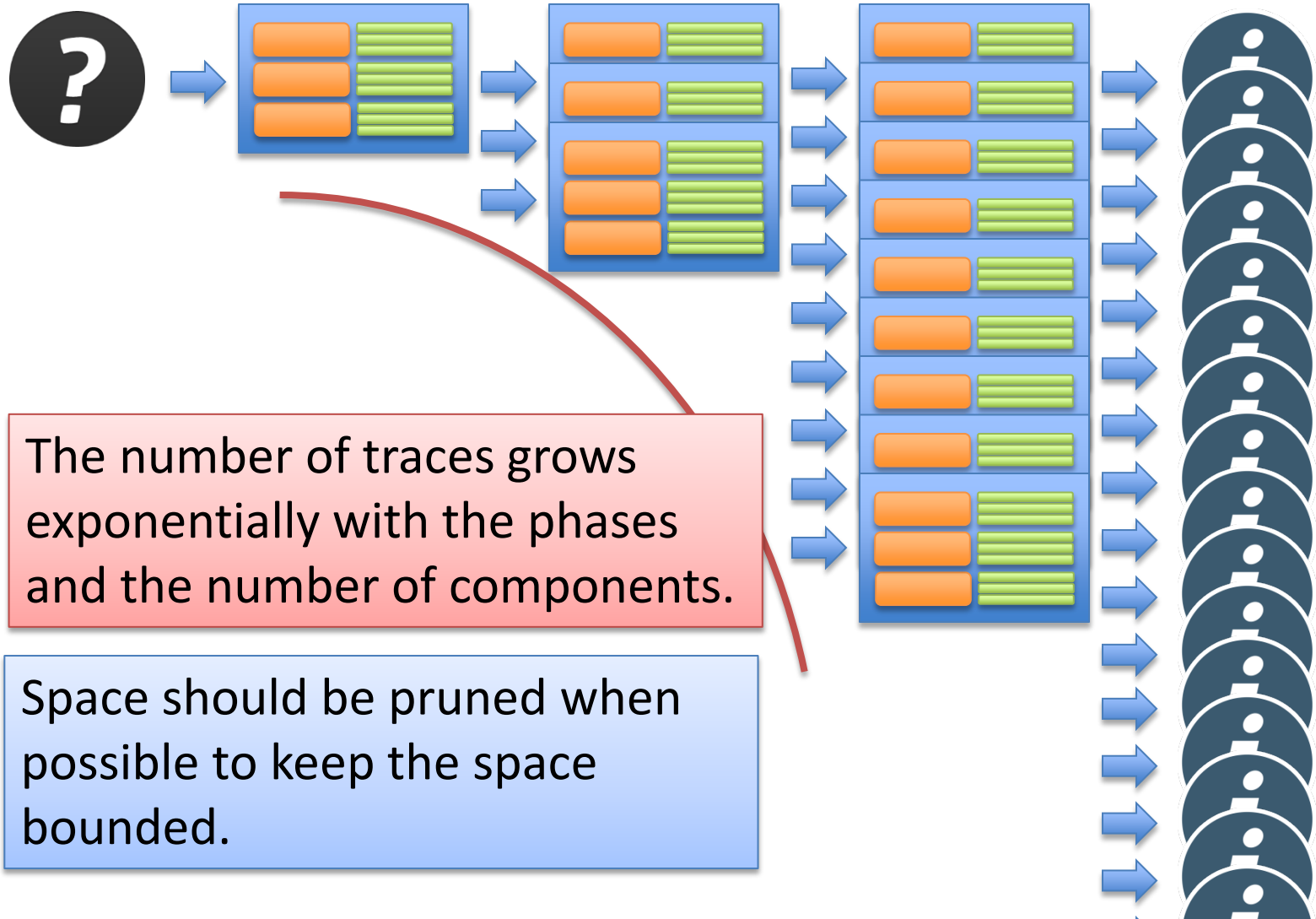
Definition: Trace



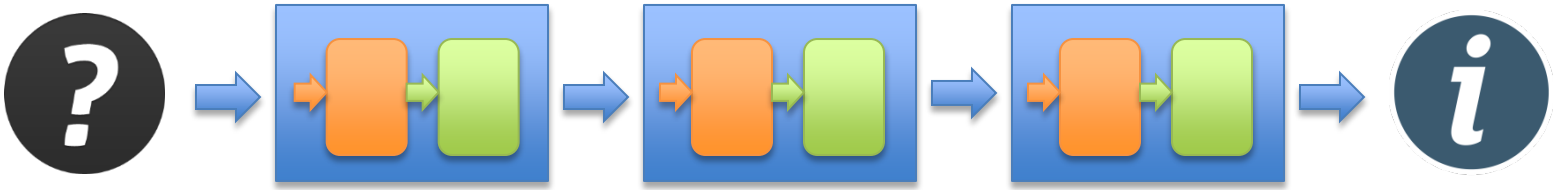
Trace $f^c|\omega^c = \left(f_1^{c_1}|\omega_1^{c_1}, f_2^{c_2}|\omega_2^{c_2}, \ldots, f_n^{c_n}|\omega_n^{c_n}\right)$

An execution path that involves a single configured component for each phase

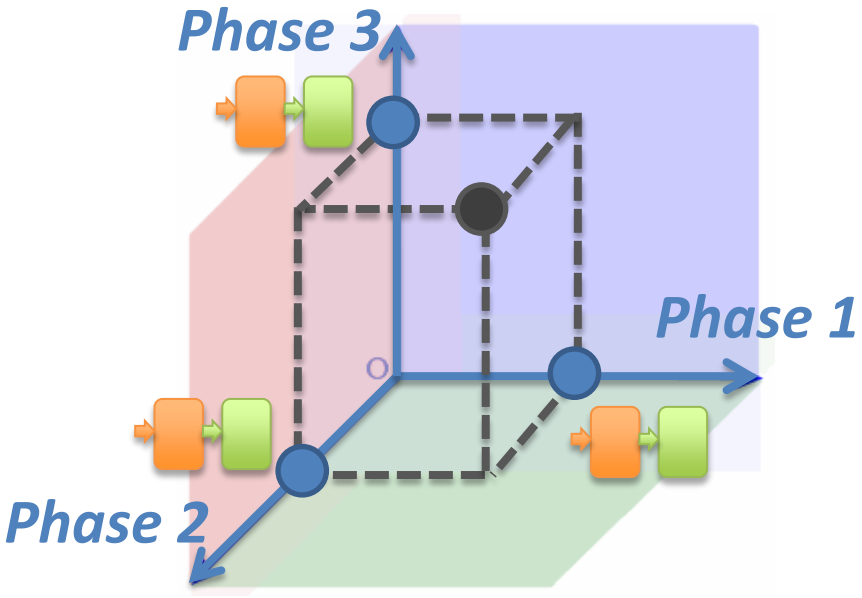
Exponential problem



Definition: Configuration space



- Pipeline



Configuration space
$\mathcal{F}|\Omega = \{f^c|\omega^c\}_c$

Set of all configured components

UIMA - EXTENDED CONFIGURATION DESCRIPTOR

Extended Configuration Descriptor

- YAML format
- A simple yet complete configuration descriptor

```
configuration:
  name: testqa-ziy-test
  author: ziy

persistence-provider:
  inherit: jdbc.db.persistence-provider

collection-reader:
  inherit: jdbc.db.collection-reader
  dataset: BIO-COMBINED
  sequence-start: 160
  sequence-end: 187
```

Extended Configuration Descriptor

- Phases and components *inherit* configuration properties or are declared as *class*es.

```
pipeline:
  - inherit: jdbc.cse.phase
    name: keyterm-extractor
    options: |
      - inherit: default.keyterm.default
      - inherit: default.keyterm.faster
  - inherit: jdbc.cse.phase
    name: retrieval-stategist
    options: |
      - inherit: default.retrieval.default
      - inherit: default.retrieval.better
  - inherit: jdbc.cse.phase
    name: passage-extractor
    options: |
      - class: cmu.edu.default.ie.Default
```

Component configuration

```
class:
edu.cmu.lti.oaqa.ecd.example.FirstPhaseAnnotatorA1
extract: true
cross-opts
  param-a: [value100,value200]
  param-b: [value300,value400]
```

This evaluates to the following `Object[]` param lists.

```
[extract: true, param-a: value100, param-b: value300]
[extract: true, param-a: value200, param-b: value300]
[extract: true, param-a: value100, param-b: value400]
[extract: true, param-a: value200, param-b: value400]
```

Extended Configuration Descriptor

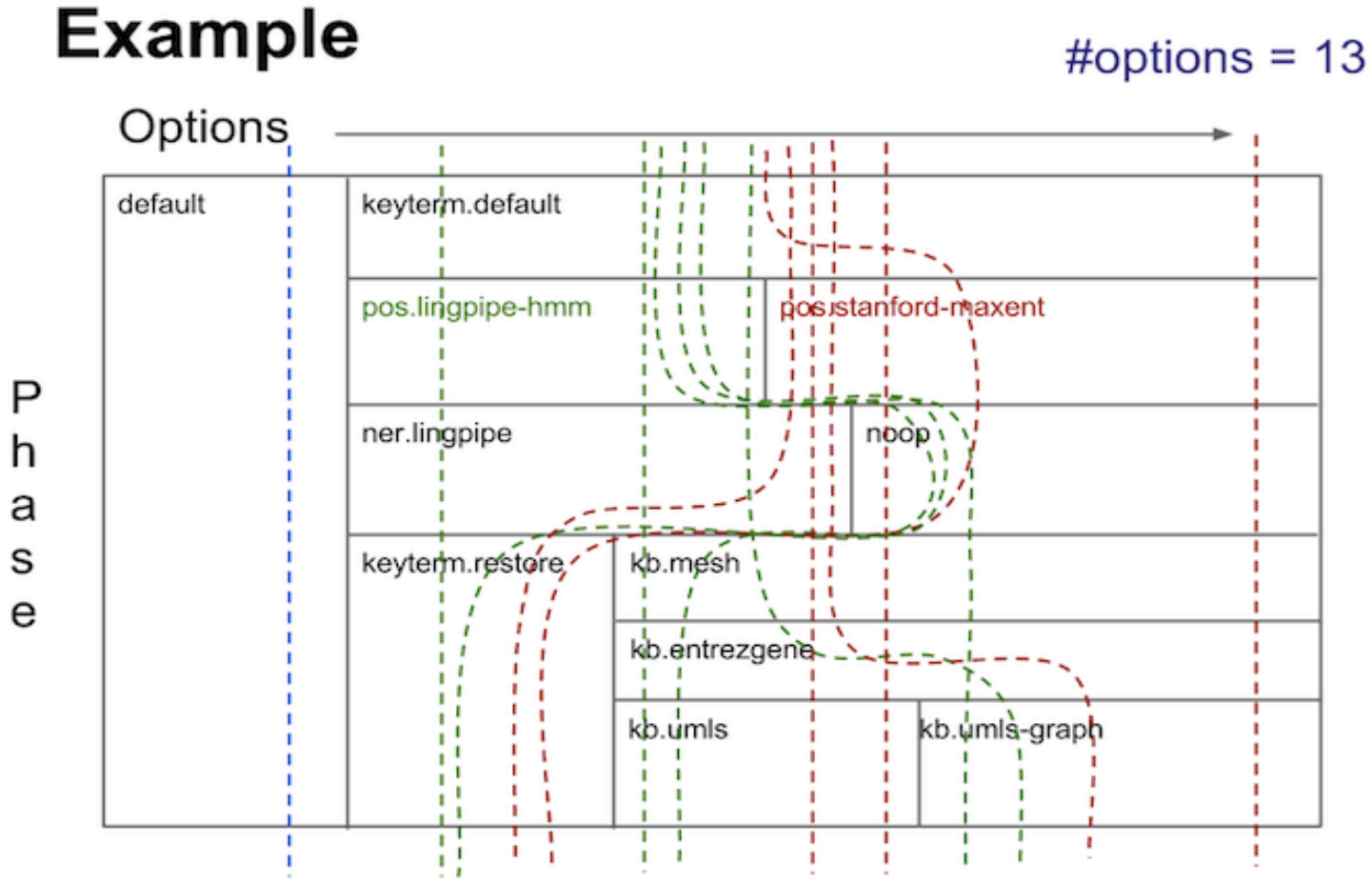
- Evaluation metrics are pluggable, and can be specified at the local or global level.

```
  - inherit: jdbc.eval.cse-retrieval-aggregator-consumer
  - inherit: bioqa.eval.cse-passage-map-aggregator-consumer

post-process:
  - inherit: jdbc.eval.cse-retrieval-evaluator-consumer
  - inherit: report.csv-report-generator
    builders: |
      - inherit: jdbc.report.f-measure-report-component

  - inherit: bioqa.eval.cse-passage-map-evaluator-consumer
  - inherit: report.csv-report-generator
    builders: |
      - inherit: bioqa.report.map-report-component
```

In-phase pipelines

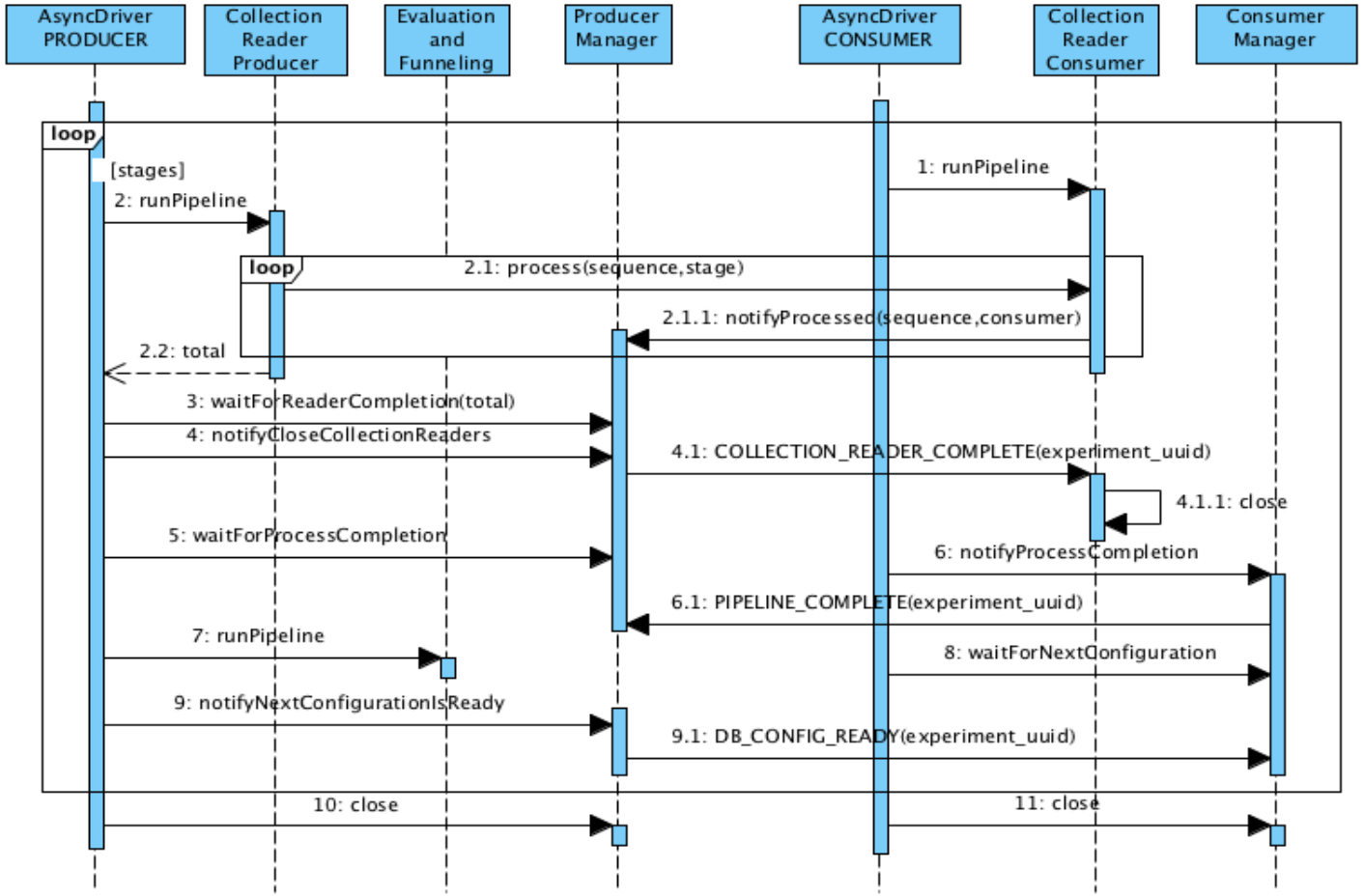


IMPLEMENTATION

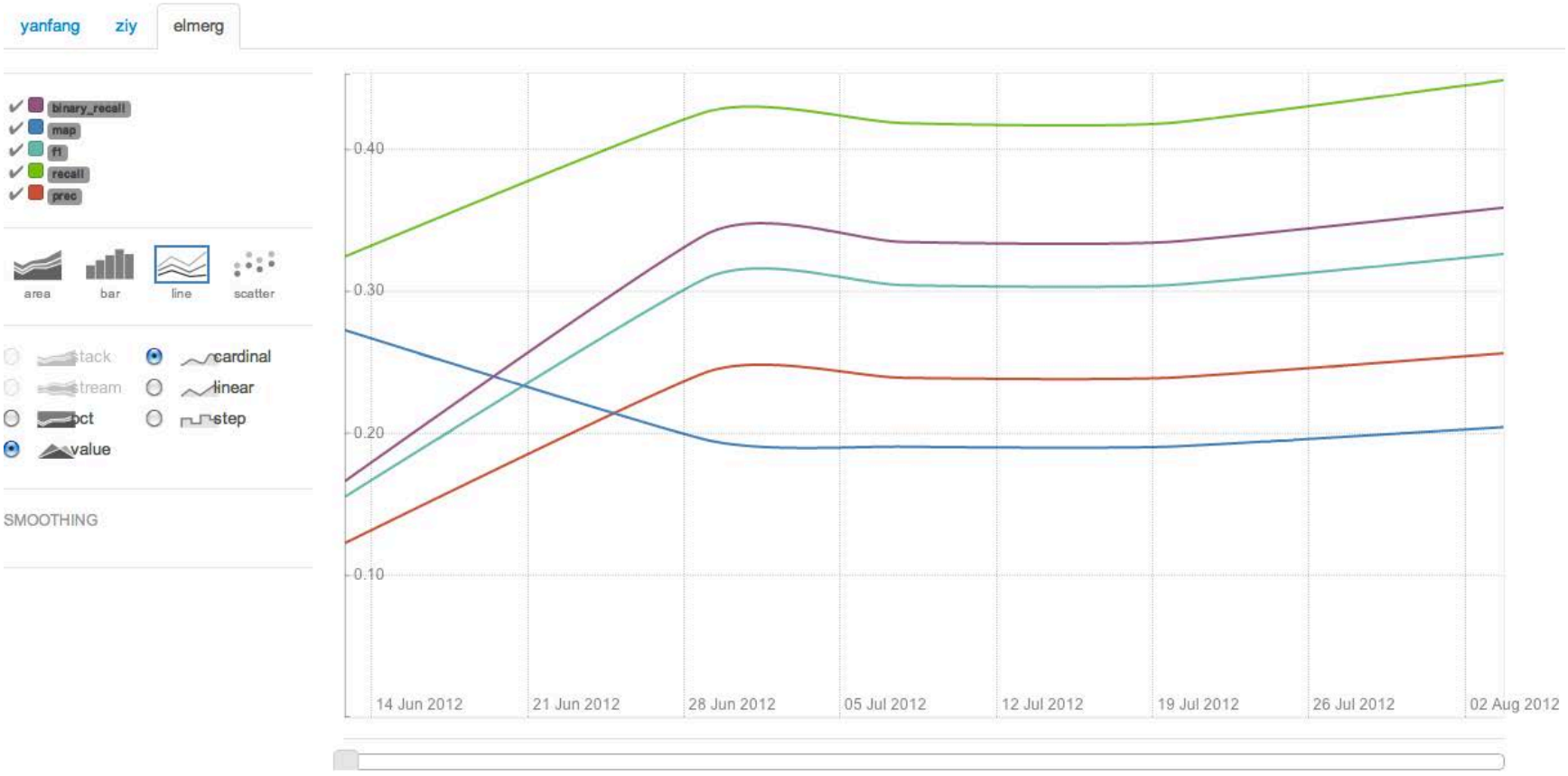
Implementation details

- Built on top of uimaFIT
- Combinatorial features are implemented using CAS Multiplier.
- CASes are persisted as compressed XMI
  - Once per trace at each phase.
  - Experiments can be restarted at any arbitrary point.
- Experimentation specific Type System
- Use UIMA-AS for external resources.

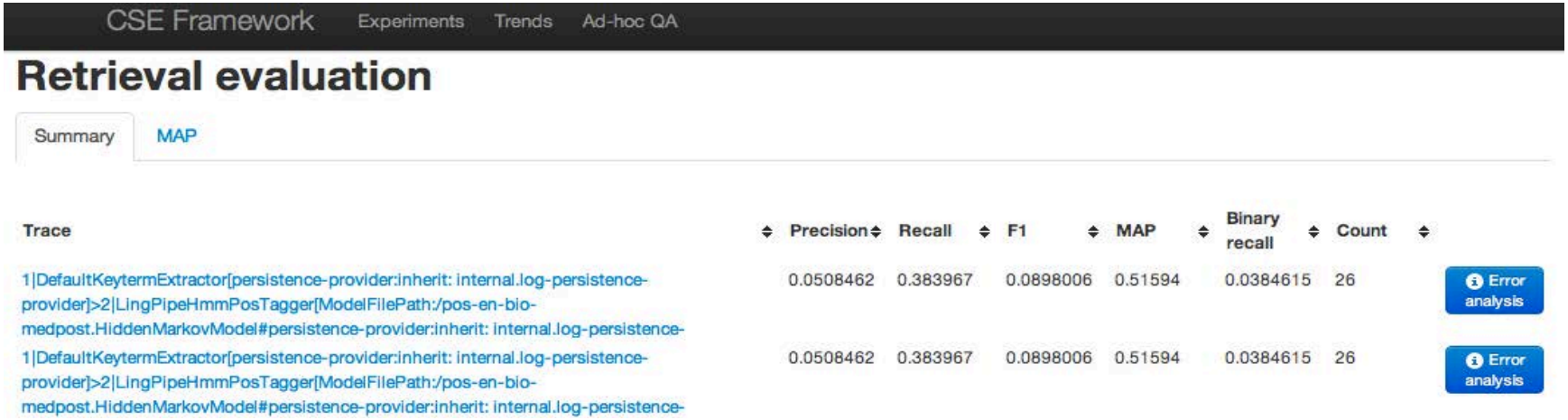
Distributed execution



Incremental improvement



Per trace visibility



Error analysis

CSE Framework Experiments Trends Ad-hoc QA

Error analysis

Id	Question	Retrieved	Relevant	Average precision
187	How do mutations in familial hemiplegic migraine type 1 (FHM1) gene affect calcium ion influx in hippocampal neurons?	1000	3	1.00000
168	How does BARD1 regulate BRCA1 activity?	1000	243	0.87095
175	How does L2 interact with L1 to form HPV11 viral capsids?	1000	33	0.84186
160	What is the role of PrnP in mad cow disease?	1000	525	0.83522
181	How do mutations in the Huntingtin gene affect Huntington's disease?	1000	589	0.78279
170	How does COP2 contribute to CFTR export from the endoplasmic reticulum?	1000	36	0.76515
186	How do mutations in the Presenilin-1 gene affect Alzheimer's disease?	1000	388	0.71872
167	How does nucleoside diphosphate kinase (NM23) contribute to tumor progression?	1000	208	0.69712
164	What is the role of Nurr-77 in Parkinson's disease?	1000	7	0.66017
163	What is the role of APC (adenomatous polyposis coli) in colon cancer?	1000	262	0.62560
184	How do mutations in the Pes gene affect cell growth?	1000	5	0.55556
165	How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease?	1000	17	0.54619
161	What is the role of IDE in Alzheimer's disease	1000	68	0.48033
183	How do mutations in the NM23 gene affect tracheal development?	1000	19	0.47449
185	How do mutations in the hypocretin receptor 2 gene affect narcolepsy?	1000	25	0.46111
172	How does p53 affect apoptosis?	1000	587	0.46068
166	What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?	1000	34	0.42790

Other domains:QA4MRE

- Question Answering for Machine Reading
- Configuration space:
  - 12 UIMA components were first developed
  - Replace UIMA descriptors with ECD
- CSE
  - 46 configurations
  - 1,040 combinations
  - 1,322 executions

The best trace identified by CSE achieved 59.6% performance gain over the original pipeline.

[Building Optimal Question Answering System Automatically using Configuration Space Exploration (CSE) for QA4MRE 2013 Tasks Alkesh Patel, Zi Yang, Eric Nyberg and Teruko Mitamura]

FUTURE WORK AND COLLABORATION

Future work

- Advanced *Configuration Space* exploration and pruning (Bagpipes Framework).
- Run arbitrary UIMA pipelines on top of industry grade distributed systems (Spark, Mesos, HDFS).
- Further investigation on space, time, resources constraining.
- Use differential CAS storage.

Collaboration

http://oaqa.github.io

Thanks!