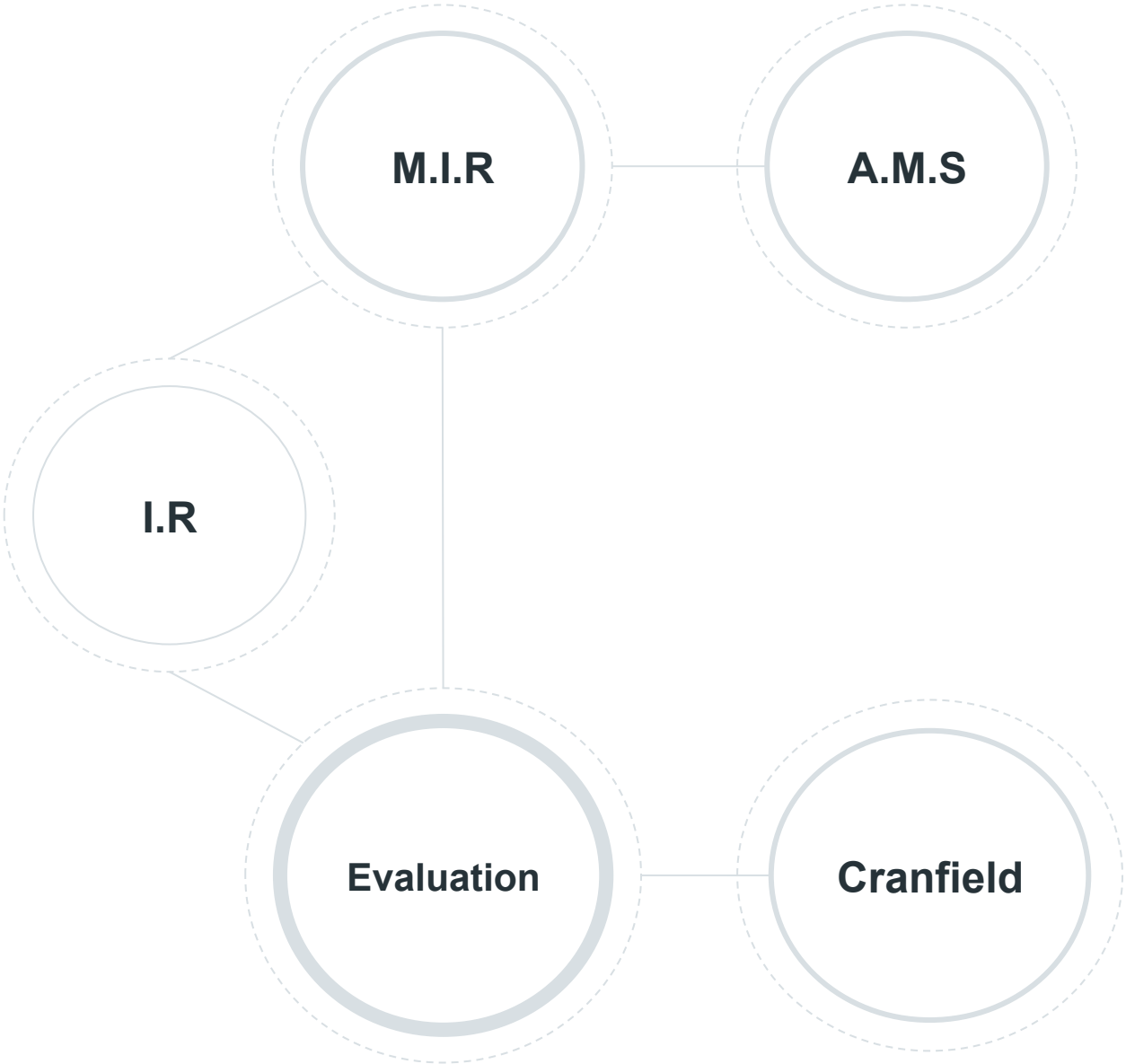
Universitat Pompeu Fabra Barcelona

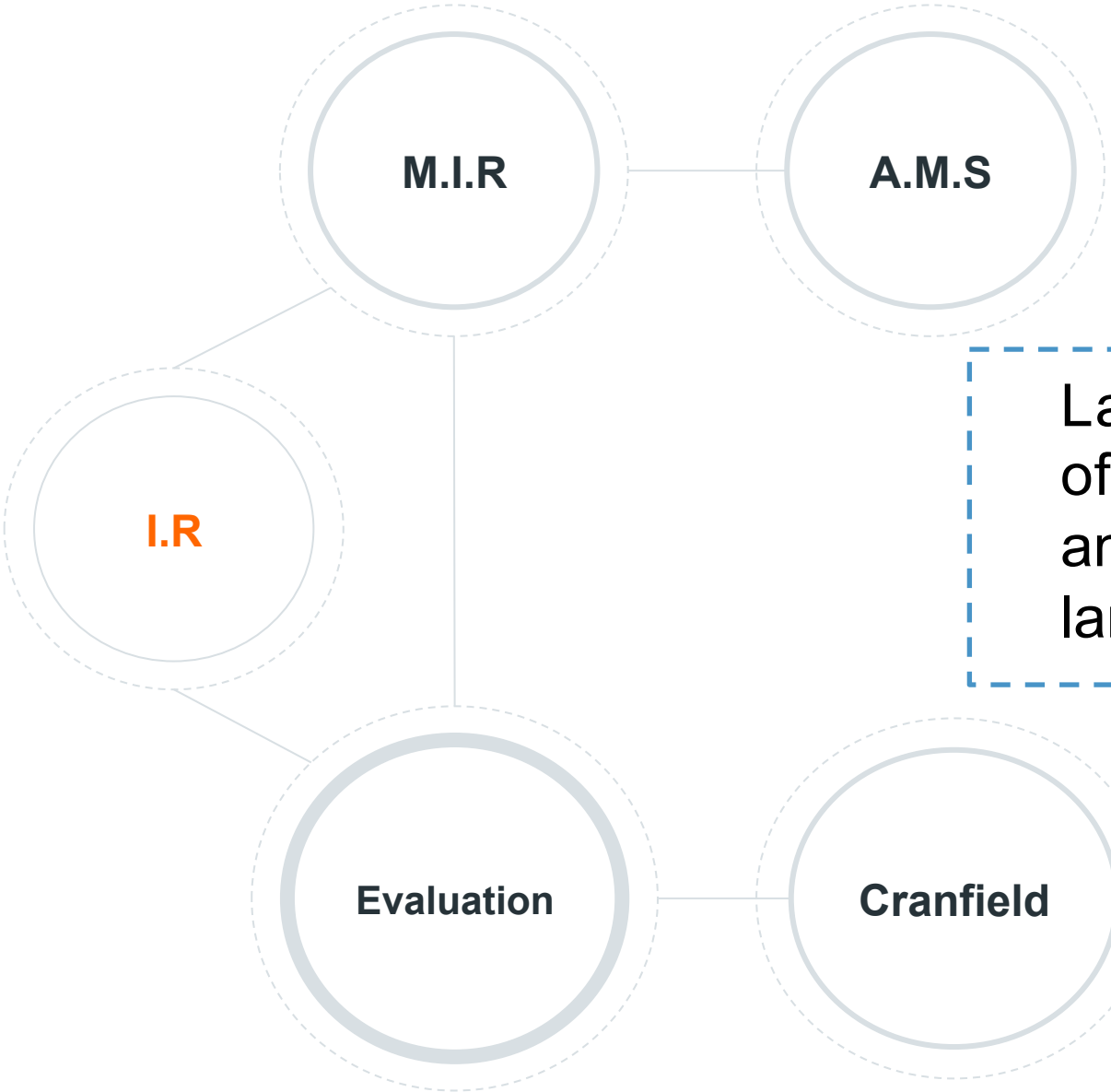
Reducing bias in the probabilistic evaluation of Audio Music Similarity

Adriana M. Suárez I
Supervisor: Julián Urbano

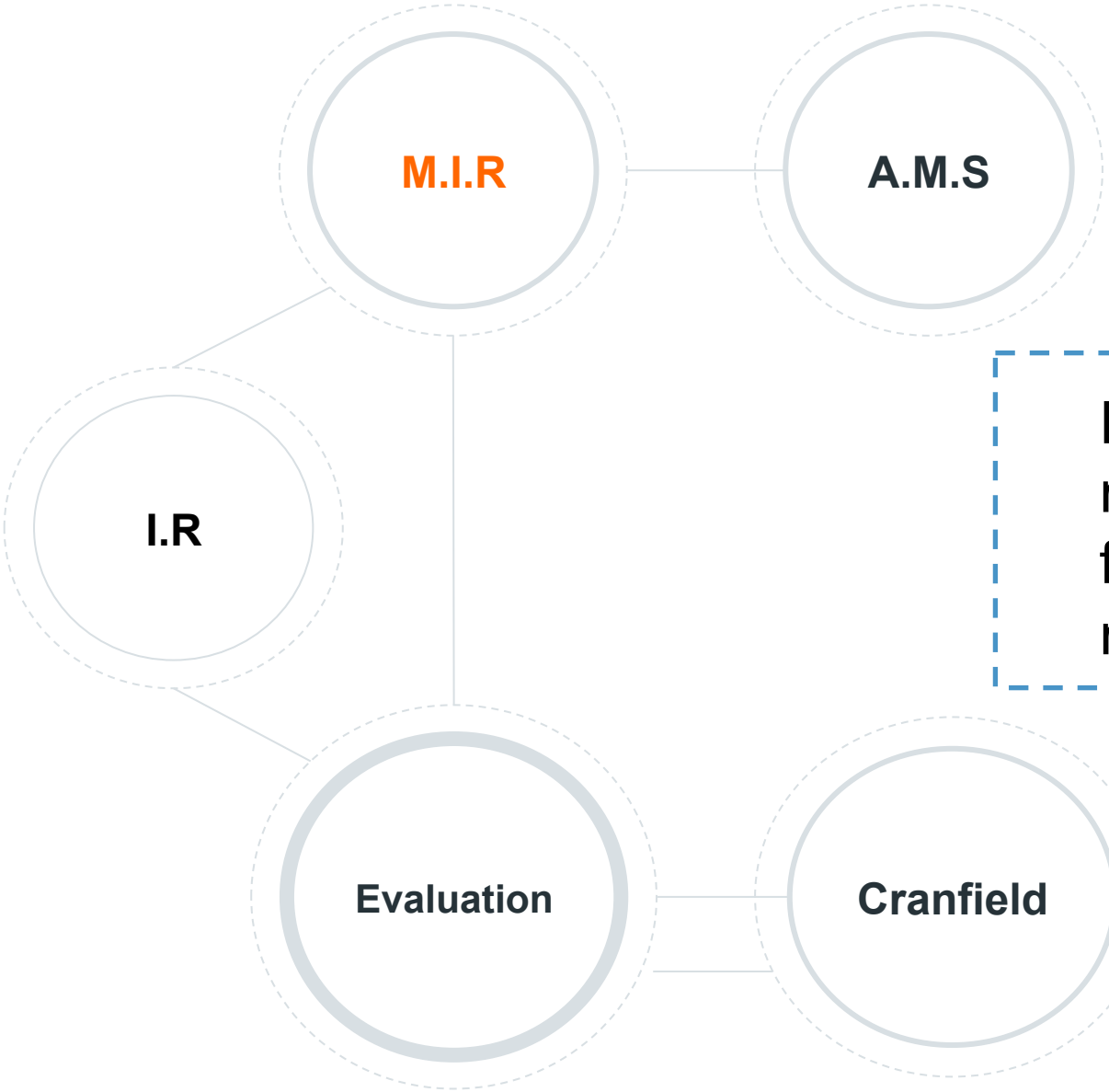
Content

- Introduction
- Motivation
- State of the Art
- Methodology
- Results
- Short-term planning

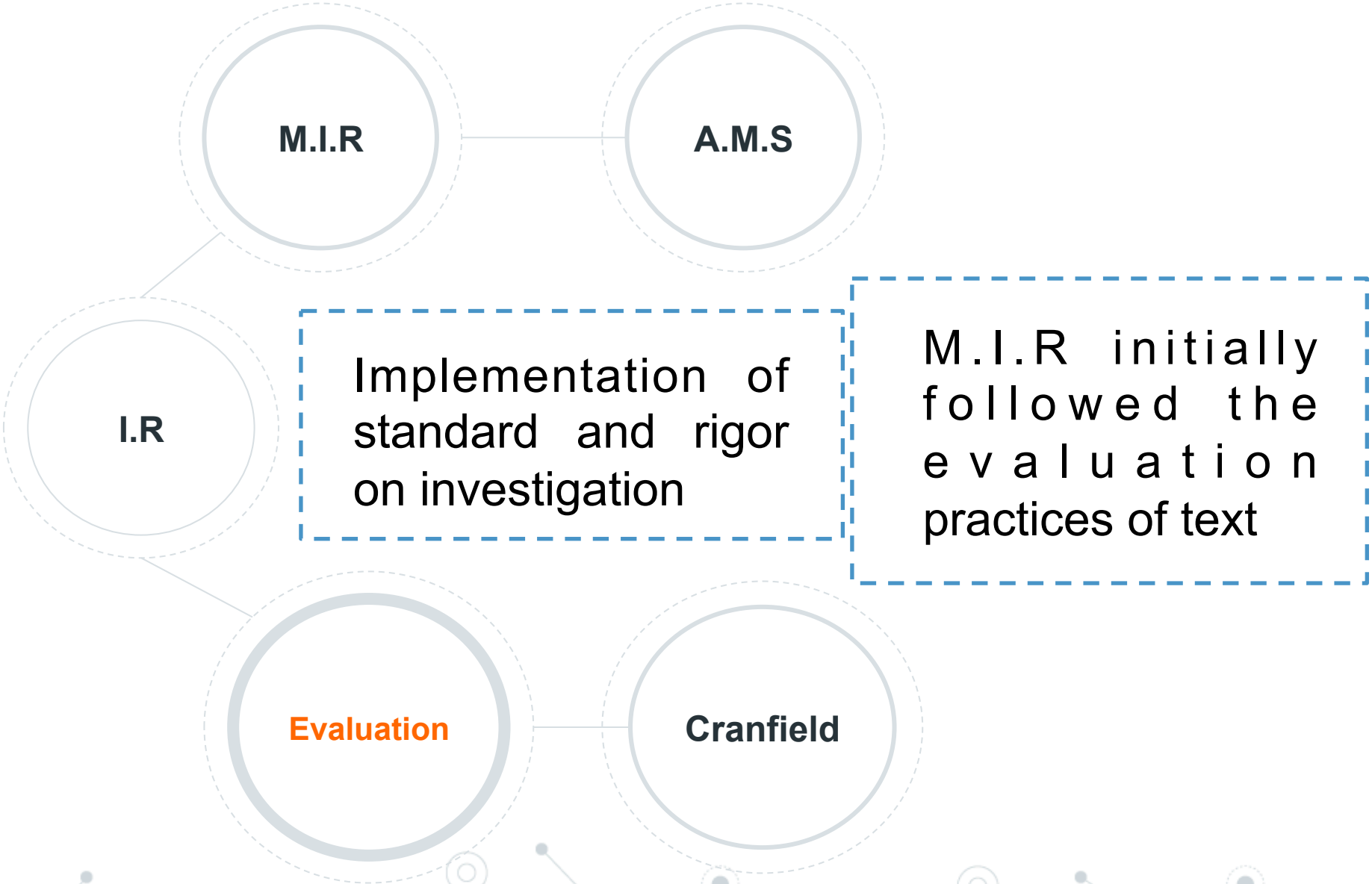

M.I.R
A.M.S
I.R
Evaluation
Cranfield

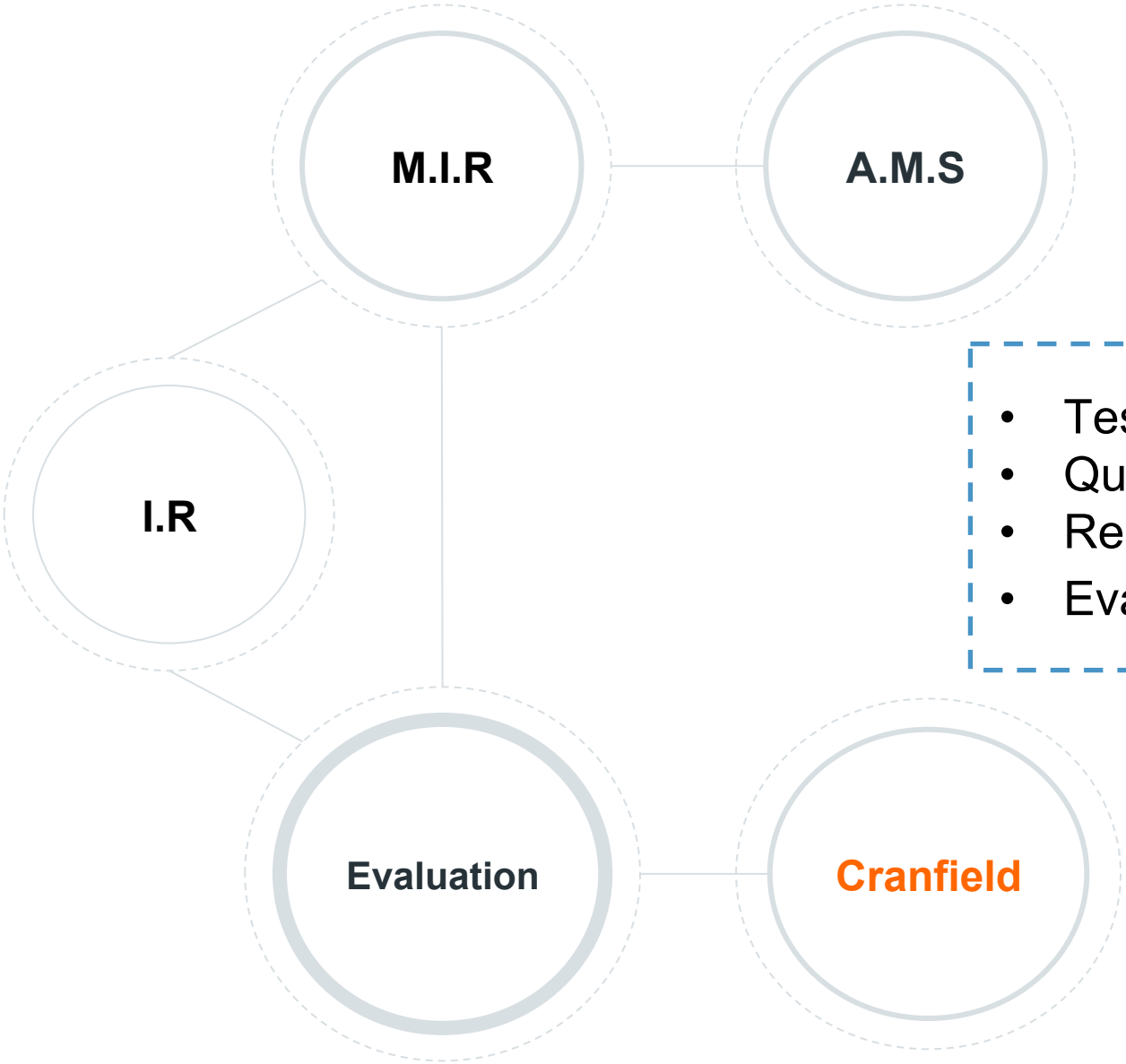


Large collections of electronic text and other human-language data.

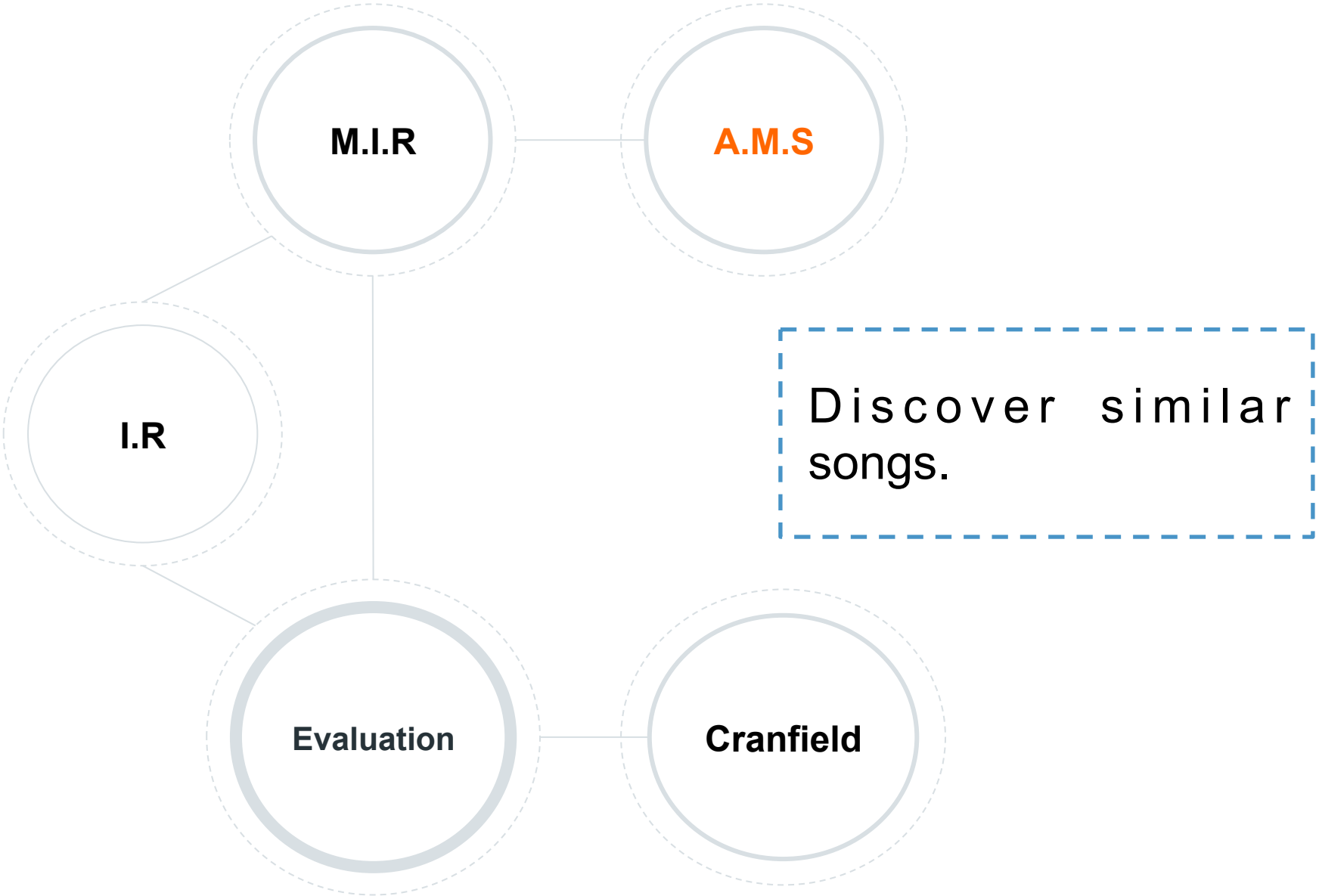


Extraction of meaningful features of music.


M.I.R
A.M.S
I.R
Implementation of standard and rigor on investigation
M.I.R initially followed the evaluation practices of text
Evaluation
Cranfield



- Test Collection
- Queries
- Relevance Judgments
- Evaluation Metrics


M.I.R
A.M.S
I.R
Evaluation
Cranfield
Discover similar songs.


M.I.R
A.M.S
I.R
Evaluation
Cranfield
Focus of this thesis

A.M.S in MIREX

- Annual evaluation of MIR algorithms
- Evaluation in A.M.S:

Audio clip (the query), a system returns a list of songs from a corpus (candidate songs), sorted by their musical similarity to the query.

Annotations

- Information contrasted using human assessors
- ***Two scales***: Broad-Fine

Annotations

- Time consuming
- Expensive
- Boring

Solution:

Low cost evaluation methodologies:

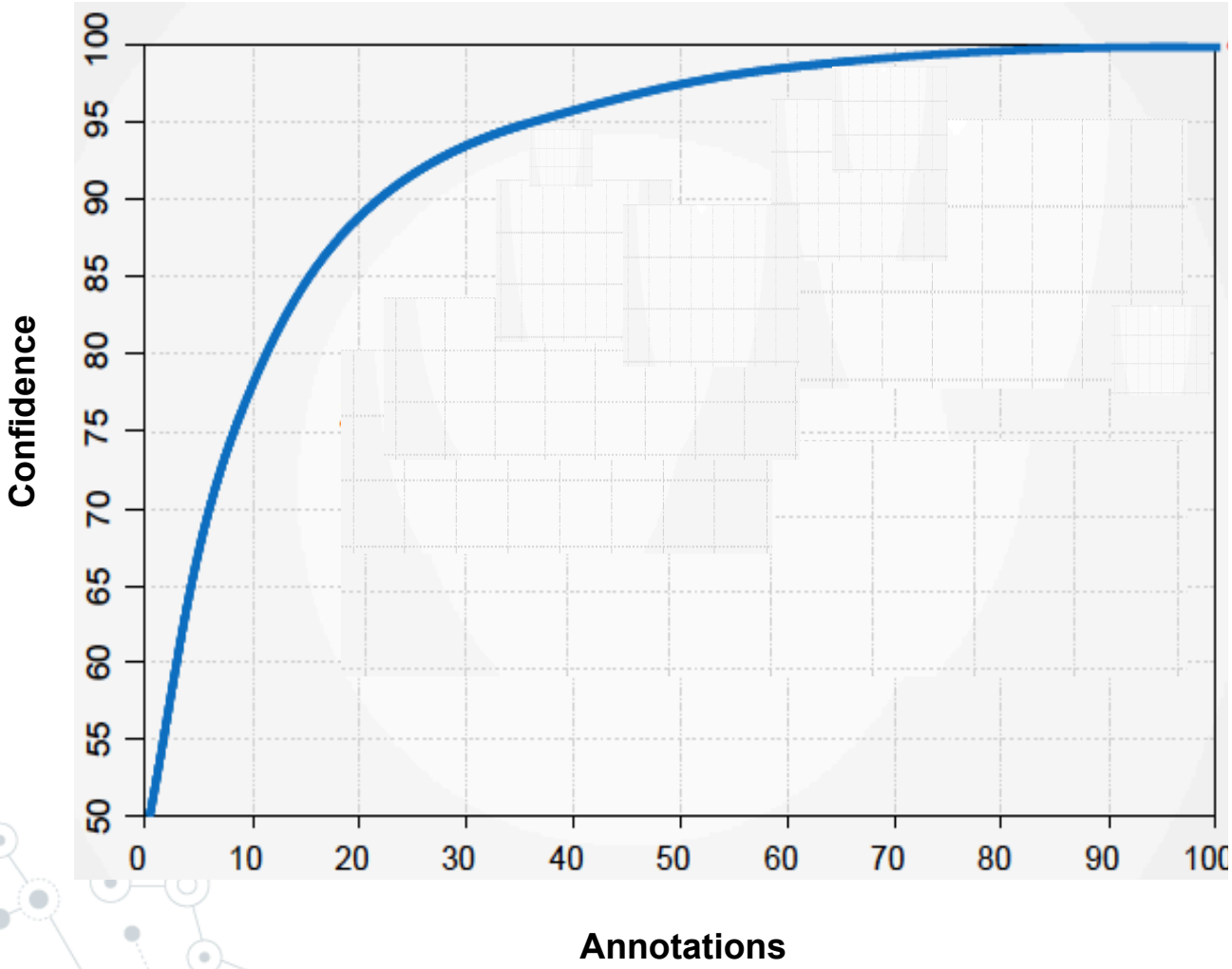
Probabilistic

Motivation

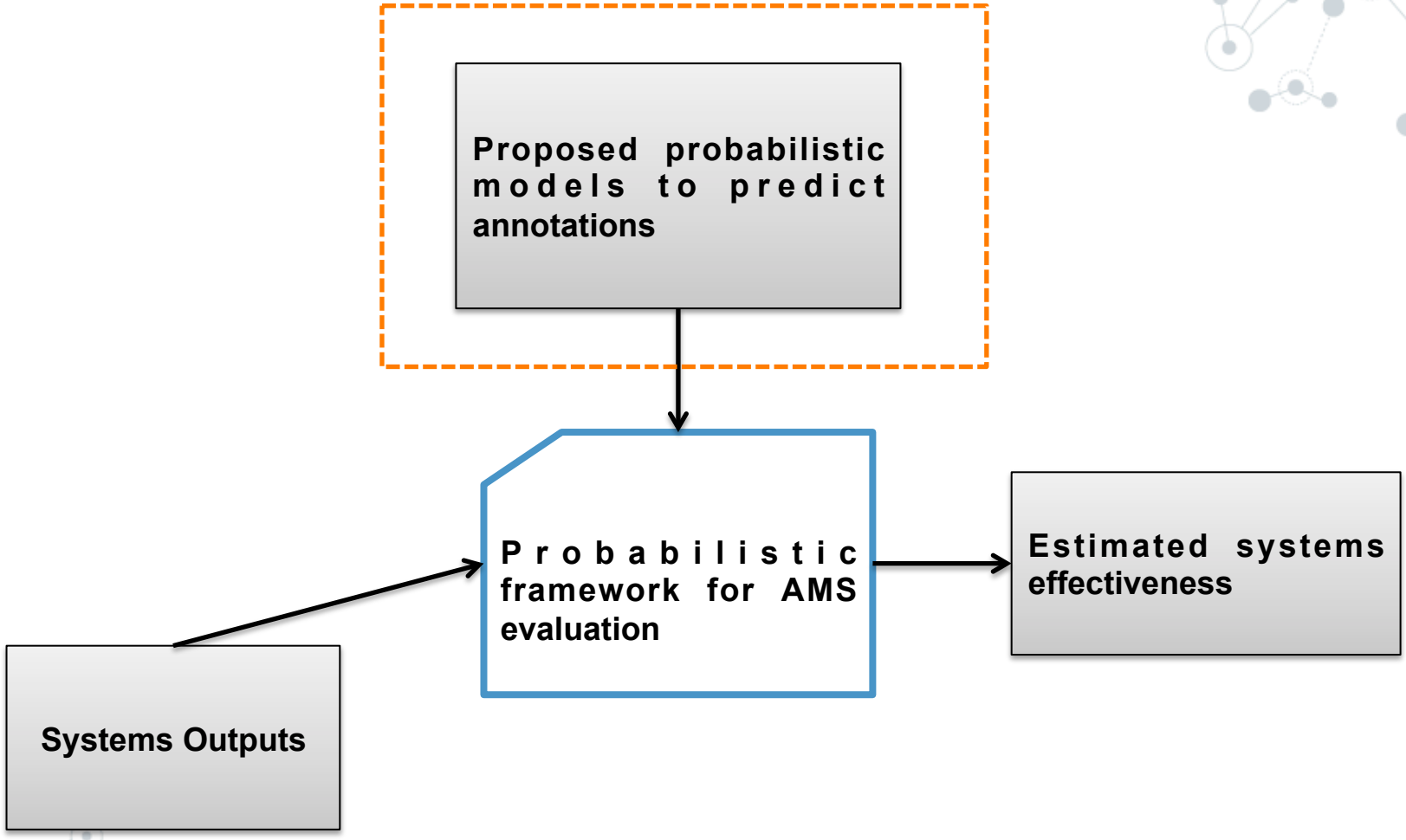
Improve efficiency in evaluation of A.M.S task in MIREX

reducing the number of needed relevance judgments

Probabilistic Evaluation



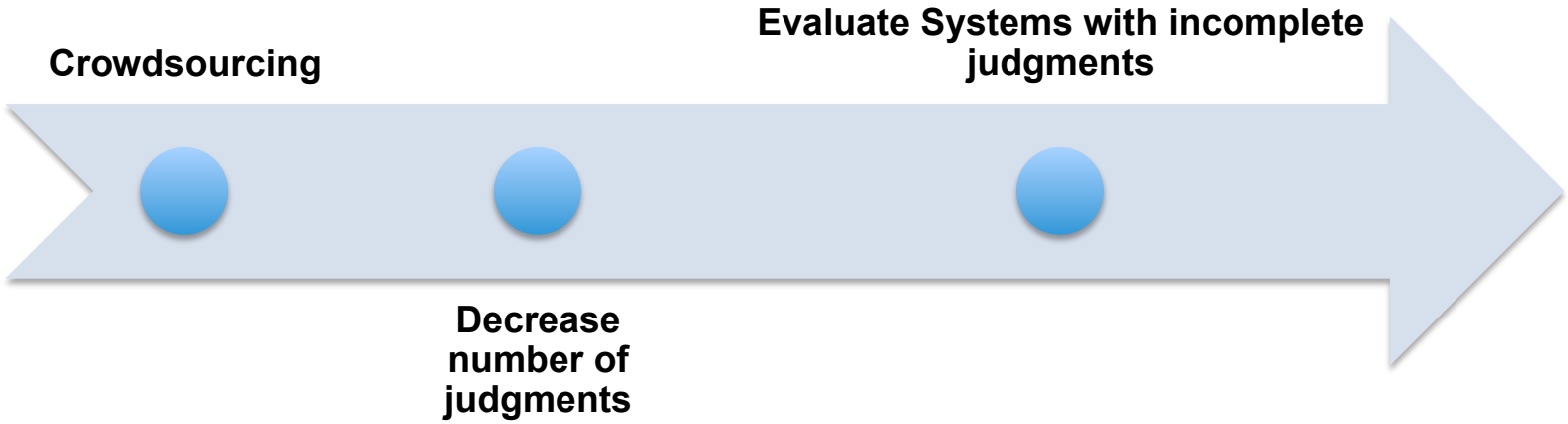
Scope


Proposed probabilistic models to predict annotations
Probabilistic framework for AMS evaluation
Estimated systems effectiveness
Systems Outputs

Content

- Motivation
- State of the Art

Low cost evaluation methodologies



Incomplete Judgments

◎Model probabilistically the relevance judgments using random variables

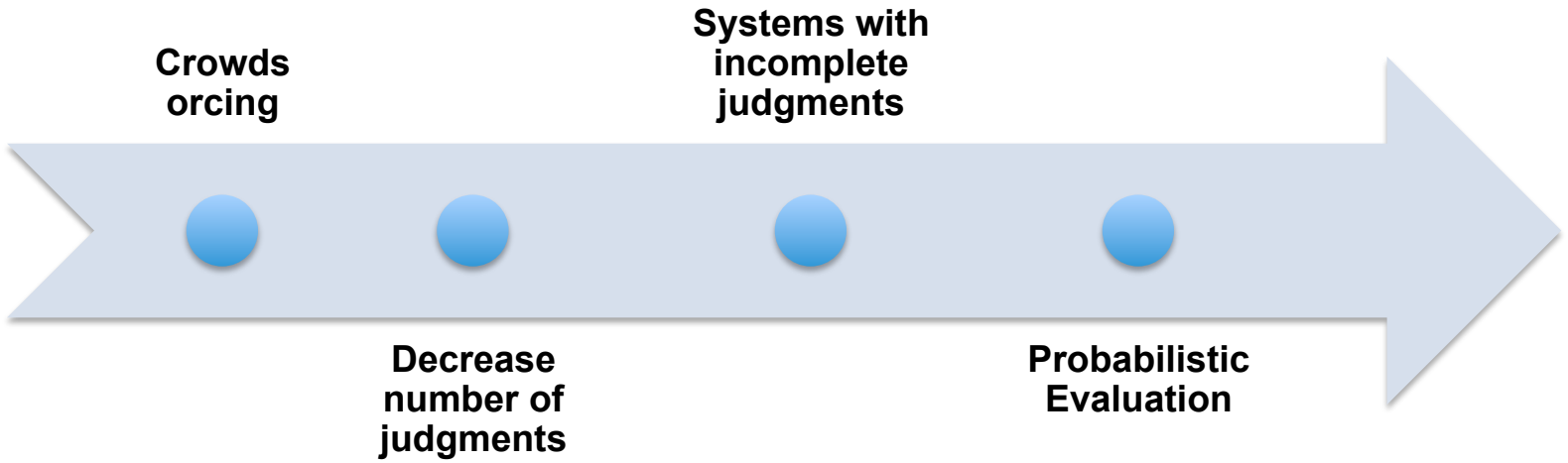
◎Let G_i being a Random Variable representing the relevance level assigned to document *d. Expectation: and Variance:*

$$E[G_i] = \sum_{l \in \mathcal{L}} P(G_i = l) \cdot l$$

$$Var[G_i] = \sum_{l \in L} P(G_i = l) \cdot l^2 - E[G_i]$$

$$\mathcal{L}_{BROAD} = \{0, 1, 2\} \qquad \mathcal{L}_{FINE} = \{0, 1, \ldots, 100\}$$

Low cost evaluation methodologies



Probabilistic Evaluation

◎To estimate the relevance of a document : to know: $P(G_i=l)$ for each relevance level L: the distribution of G_i must be calculated

◎Two models were fitted to estimate these relevance judgments

◎Use this relevance judgments to evaluate systems using metrics

Estimation of Relevance Judgments

- Framework: Ordinal Logistic Regression

$$\log \frac{P(R_d \geq \ell | \boldsymbol{\theta}_d)}{P(R_d < \ell | \boldsymbol{\theta}_d)} = \alpha_\ell + \sum_{k=1}^{|\boldsymbol{\theta}_d|} \beta_k \cdot \boldsymbol{\theta}_{d,k}$$

$$P(R_d = \ell | \boldsymbol{\theta}_d) = P(R_d \geq \ell | \boldsymbol{\theta}_d) - P(R_d \geq \ell + 1 | \boldsymbol{\theta}_d)$$

Probabilistic Evaluation

◎To estimate the relevance of a document : to know: $P(G_i=l)$ for each relevance level L: the distribution of G_i must be calculated

◎Two models were fitted to estimate these relevance judgments

◎Use this relevance judgments to evaluate systems using metrics

Output-based features

- Used when there are not judgments: M_{out}

Feature	Description
fSYS	% of systems that retrieved *d* for *q*
OV	Degree of overlap between systems
sGEN	Whether the musical genre of *d* is the same as *q*
fGEN	% of al documents retrieved for *q* that belong to the same musical genre than *d* does
fART	% of documents retrieved for *q* that belong to the same artist as *d* does

Judgment -based features

- Used with known results: M_{jud}

Feature	Description
aSYS	Average relevance of documents retrieved by the system
aGEN	Average relevance of all the documents retrieved for *q* that belongs to the same genre as *d* does
aART	Average relevance of all the documents retrieved for *q* performed by the same artist as *d* .

Performance of Models

- Used with known results: M_{jud}

Model	R2 (Coefficient of Determination)
Mjud	0.9
Mout	0.35

Probabilistic Evaluation

◎To estimate the relevance of a document : to know: $P(G_i=l)$ for each relevance level L: the distribution of G_i must be calculated

◎Two models were fitted to estimate these relevance judgments

◎Use this relevance judgments to evaluate systems using metrics

Estimation of Effectiveness

First Scenario: *not relevance judgments available*.

Order of systems estimated M_{out} average accuracy of 92%

Average confidence in the rankings of 94%

Estimation of Effectiveness

Second Scenario: *systems' performance differences.*

Using 2% of judgments, differences are estimated in

93%

Estimation of Effectiveness

Third Scenario: *the estimation of absolute effectiveness scores.*

Using 25% of relevance judgments they can estimate with an error of +-0.05

Effectiveness in the ranking of systems is highly overestimated

Estimation of Effectiveness

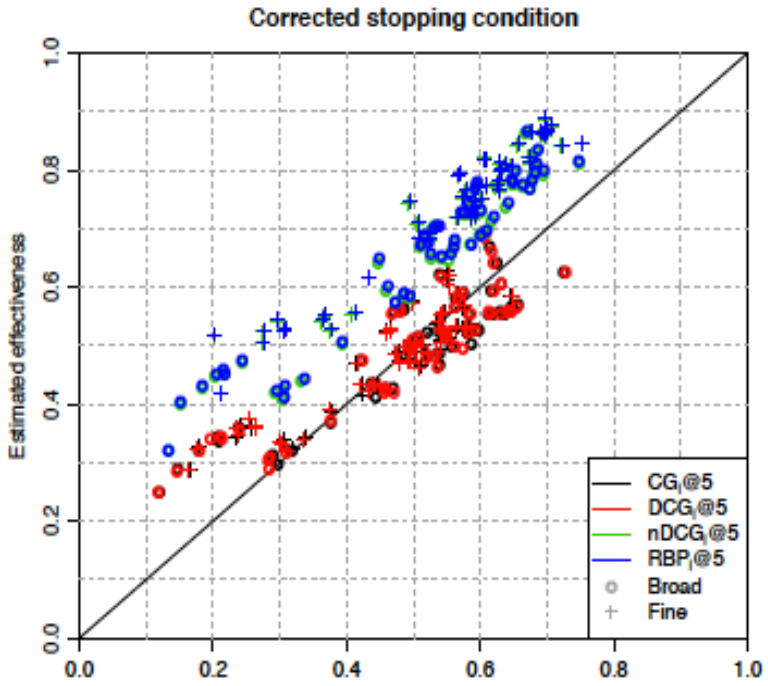


Fig 2. Estimated vs. actual absolute effectiveness scores in MIREX 2007, 2009, 2010 and 2011 when judging documents until expected error is +-0.05 with an uncorrected (left) or corrected (right) stopping condition. Adapted from (Urbano, 2013)

Content

- Motivation
- State of the Art
- Methodology

Methodology

1. Implementing others frameworks.

2. Implementing new attributes for models.

3. Using several models according to the amount of judgments

1. Using other frameworks

- Before we had Ordinal values 0,1,2…
- Logit and Probit Regression
- Estimated variable between [0-1]

Framework	RMSE	Variance
Logit	-4,5%	-7%
Probit	-2,8%	-6%

1. Using other frameworks

- Regression + trunking

- Estimated variable, real number, truncated in the interval [0-1]

Framework	R^2
Regression	-7%

Methodology

1. Implementing others frameworks.

2. Implementing new attributes for models.

3. Using several models according to the amount of judgments

2. New attributes

- Backward elimination

MODELS TRIALS		
Model	**Predictors**	**Deviance**
1	fSYS * OV + fSYS * sGEN + fSYS * fGEN + fSYS * fART + OV * sGEN + OV * fGEN + OV * fART + sGEN * fGEN + sGEN * fART + fGEN * fART	
2	fSYS + OV + sGEN + fGEN + fART	
3b	fSYS+OV	
3c	fSYS+sGEN	
3d	fSYS+fGEN	
3e	fSYS+fART	
3j	OV+sGEN	
3k	OV+fGEN	
3l	OV+fART	
3m	sGEN+fGEN	
3n	sGEN+fART	
3o	fGEN+fART	
5	fSYS*fGEN + OV*fGEN + sGEN*fGEN + fGEN*fART	**36.401**
	fSYS*OV + fART + sGEN*fGEN	**36.837**

2. New attributes

- Clustering genres by subjective similarity

Genres	Cluster
Baroque-Classical-Romantic	Cluster 1: ***Classical***
RapHiphop – Edance	Cluster 2: ***Electronic***
Blues-Rockandroll-Country	Cluster 3: ***Romantic***
Jazz	Cluster 4: ***Jazz***
Metal	Cluster 5: ***Metal***

2. New attributes

similarity-distances between
y - document

- Implementing similarity-distances between genres of query - document

GENRE q	GENRE d	Distances
JAZZ	JAZZ	65,2
METAL	METAL	63,3
CLASSICAL	CLASSICAL	59,3
ELECTRONIC	ELECTRONIC	58
ROMANTIC	ROMANTIC	49
ROMANTIC	METAL	39,2
ROMANTIC	JAZZ	37,5
METAL	ELECTRONIC	28,2
ELECTRONIC	ROMANTIC	20,4
CLASSICAL	JAZZ	17,5
ELECTRONIC	JAZZ	16,5
METAL	JAZZ	16,1
JAZZ	ELECTRONIC	13
ROMANTIC	ELECTRONIC	12,9
CLASSICAL	ROMANTIC	12,4
METAL	CLASSICAL	9,8
ROMANTIC	CLASSICAL	8,9
JAZZ	METAL	8,6
CLASSICAL	ELECTRONIC	6,4
CLASSICAL	METAL	4,8
ELECTRONIC	CLASSICAL	4,7

2. Improved M_{out}

RMSE

5,28%

R^2

Broad: 21%

Fine: 23%

Var.

7,05%

Content

- Motivation
- State of the Art
- Methodology
- Results
- Short-term planning

Short-term planning

Task	Description	Possible Date
New features for artist	Ex: Metadata from Internet Musical Databases, machine learning for clustering of genres.	*June 26th – July 15 th*
Training models to predict using different amount of available information	Training models using less data in order to consider the cases in which not much (incomplete) information will be available.	*June 26th – July 15 th*

Short-term planning

Task	Description	Possible Date
Coding the model	Coding in JAVA or C++ to interact with the other system	*July 16 th to 31 th*
Finishing the monograph	Consigning the final results or the research.	*August 1 th -20 th*
Final Revision by Advisor		*August 21 th - 31 th*

Thanks!

Gracias