Linear Score Normalization Revisited

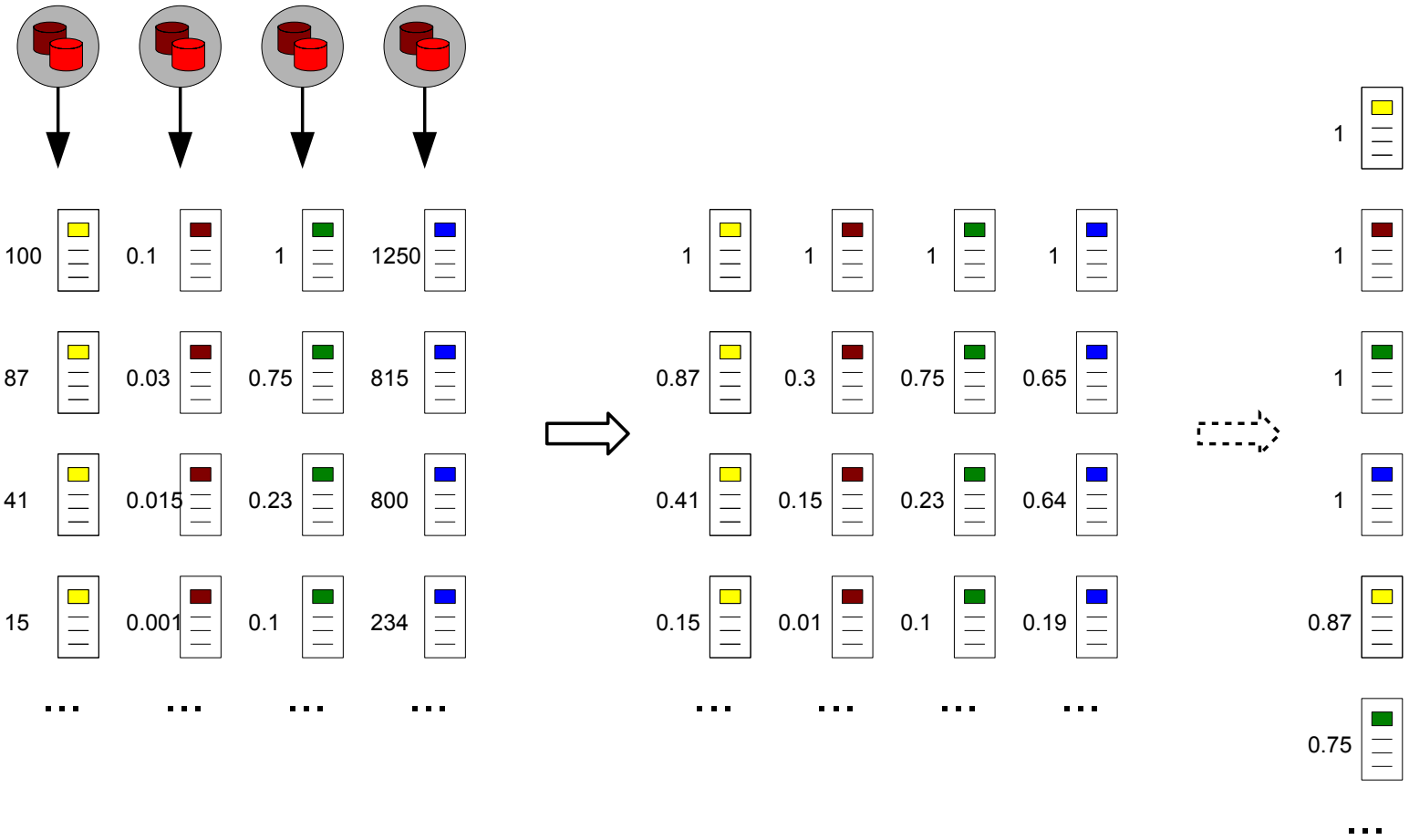
Ilya Markov

ilya.markov@usi.ch

University of Lugano

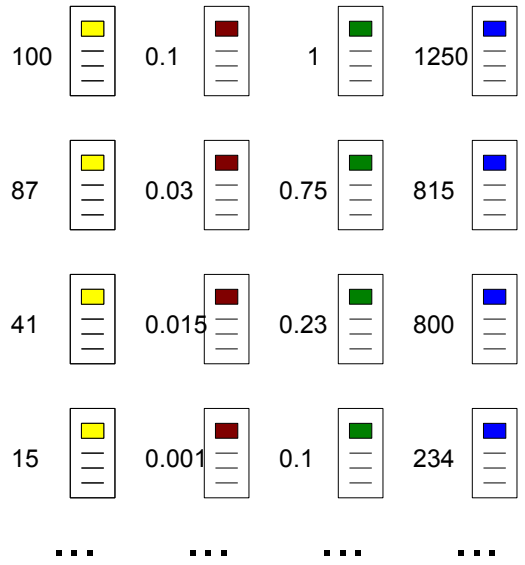
Outline

Score Normalization

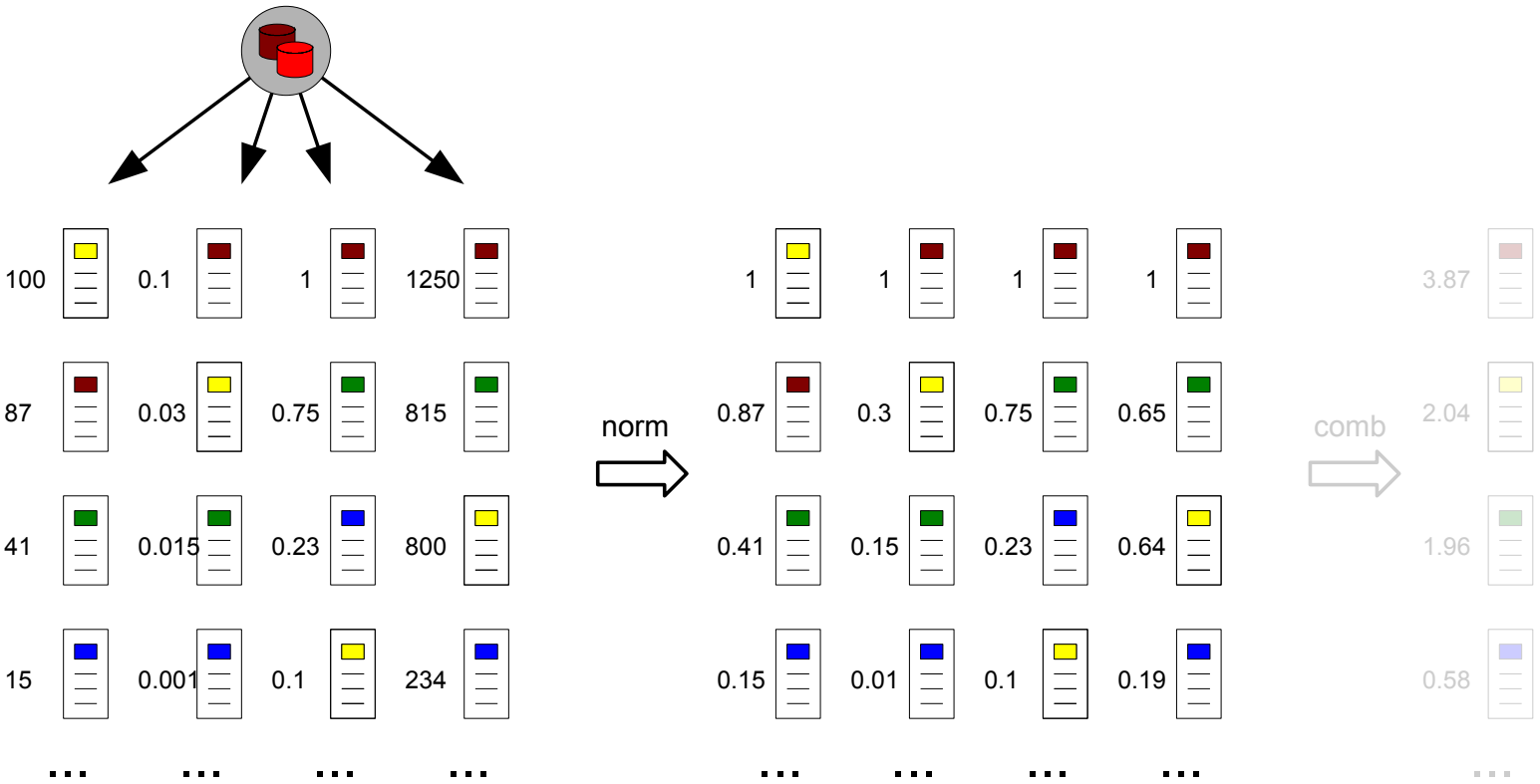


Assumptions

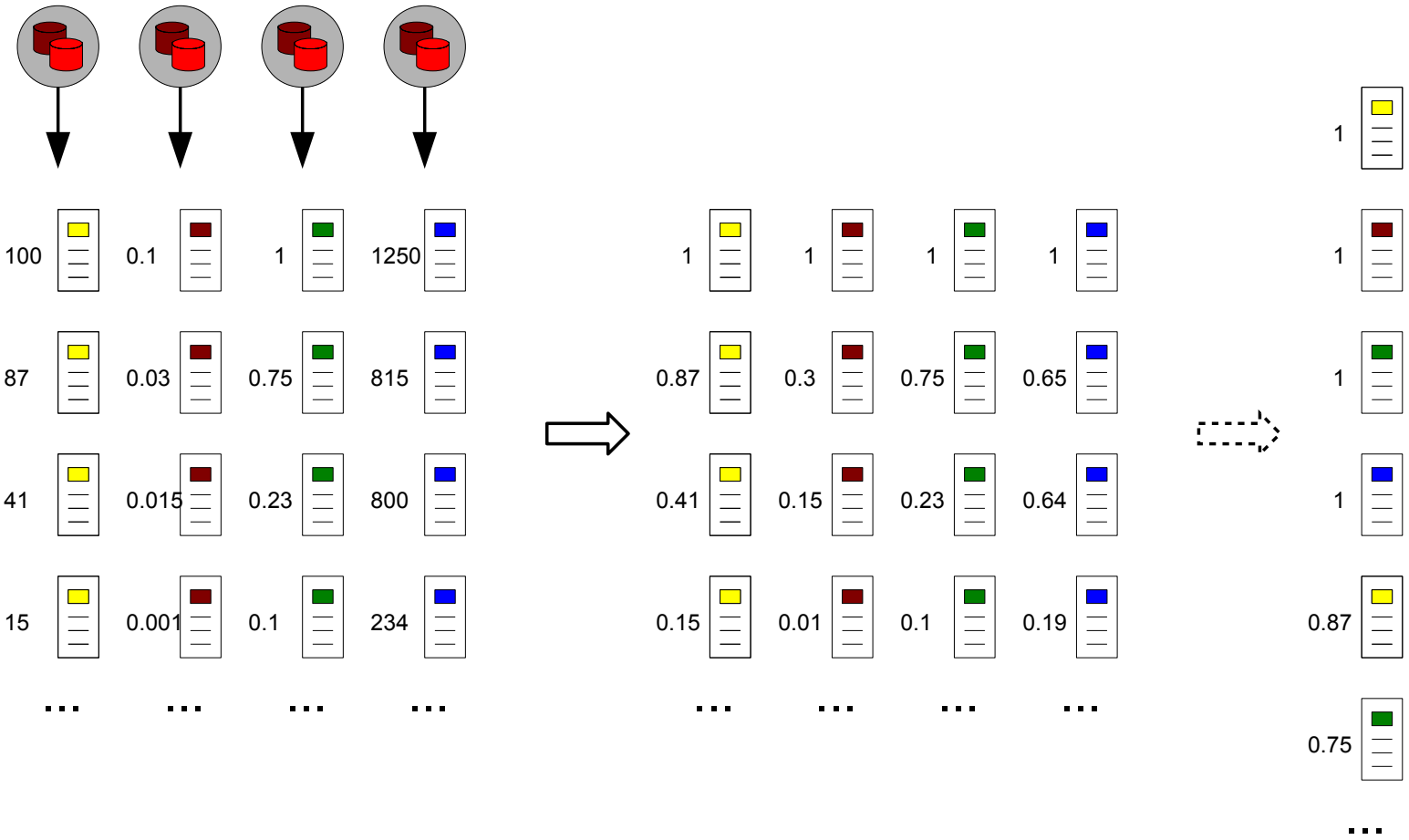
1. Document scores are provided
2. Document lists are disjoint



Data Fusion


100
0.1
1
1250
87
0.03
0.75
815
41
0.015
0.23
800
15
0.001
0.1
234
norm
1
1
1
1
0.87
0.3
0.75
0.65
0.41
0.15
0.23
0.64
0.15
0.01
0.1
0.19
comb
3.87
2.04
1.96
0.58

Score Normalization



Outline

Linear Score Normalization

- MinMax

$$s_{MinMax} = \frac{s - s_{min}}{s_{max} - s_{min}}$$

- Z-Score

$$s_{ZScore} = \frac{s - \mu}{\sigma}$$

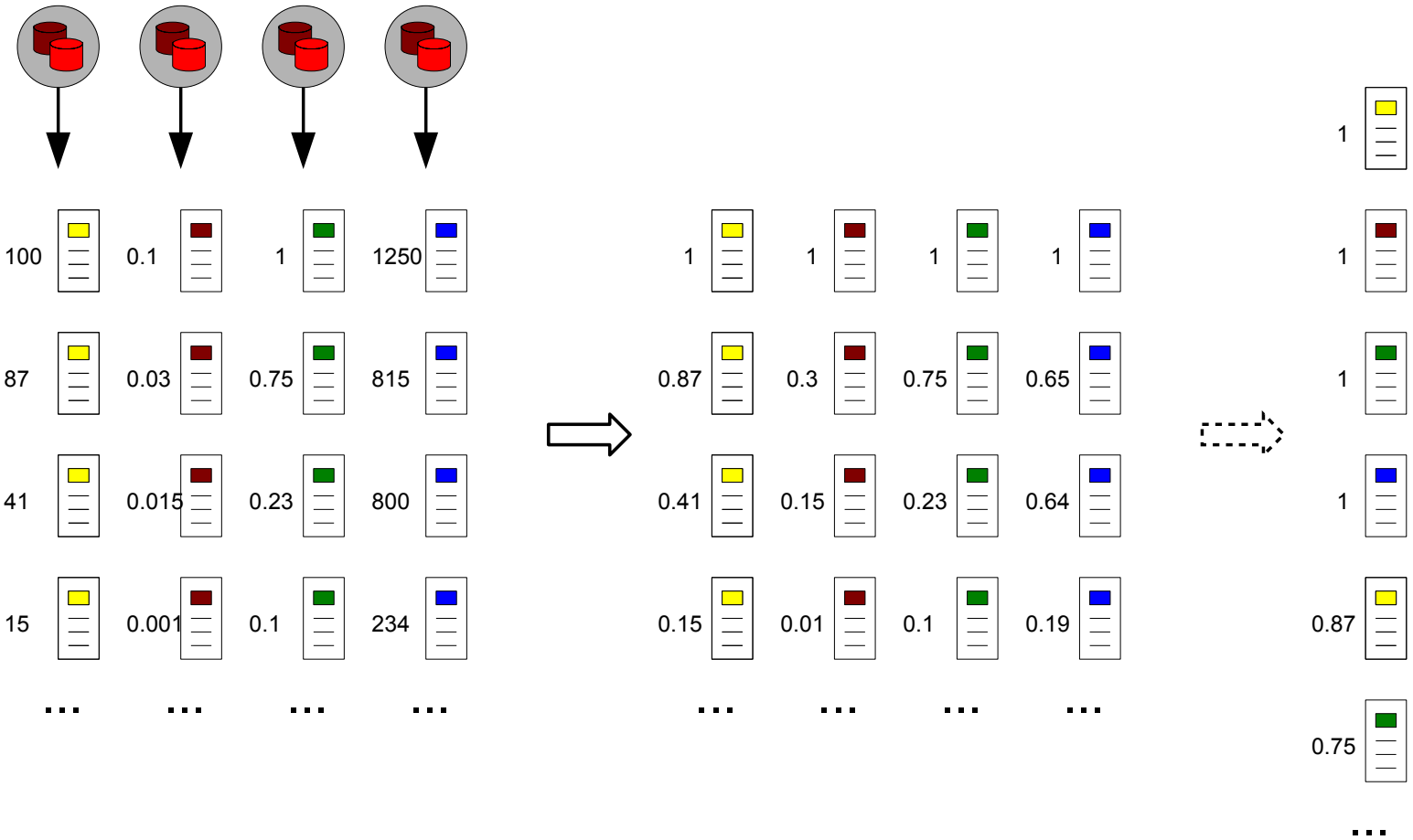
- Sum

$$s' = s - s_{min}, \; s_{Sum} = \frac{s'}{\sum_i s'_i}$$

MinMax

$$s_{MinMax} = \frac{s - s_{min}}{s_{max} - s_{min}}$$

MinMax


100
0.1
1
1250
87
0.03
0.75
815
41
0.015
0.23
800
15
0.001
0.1
234
1
1
1
1
0.87
0.3
0.75
0.65
0.41
0.15
0.23
0.64
0.15
0.01
0.1
0.19
1
1
1
1
0.87
0.75

MinMax Assumptions

A1 Each collection contains at least 1 relevant document

A2 This document is most likely to be ranked 1st

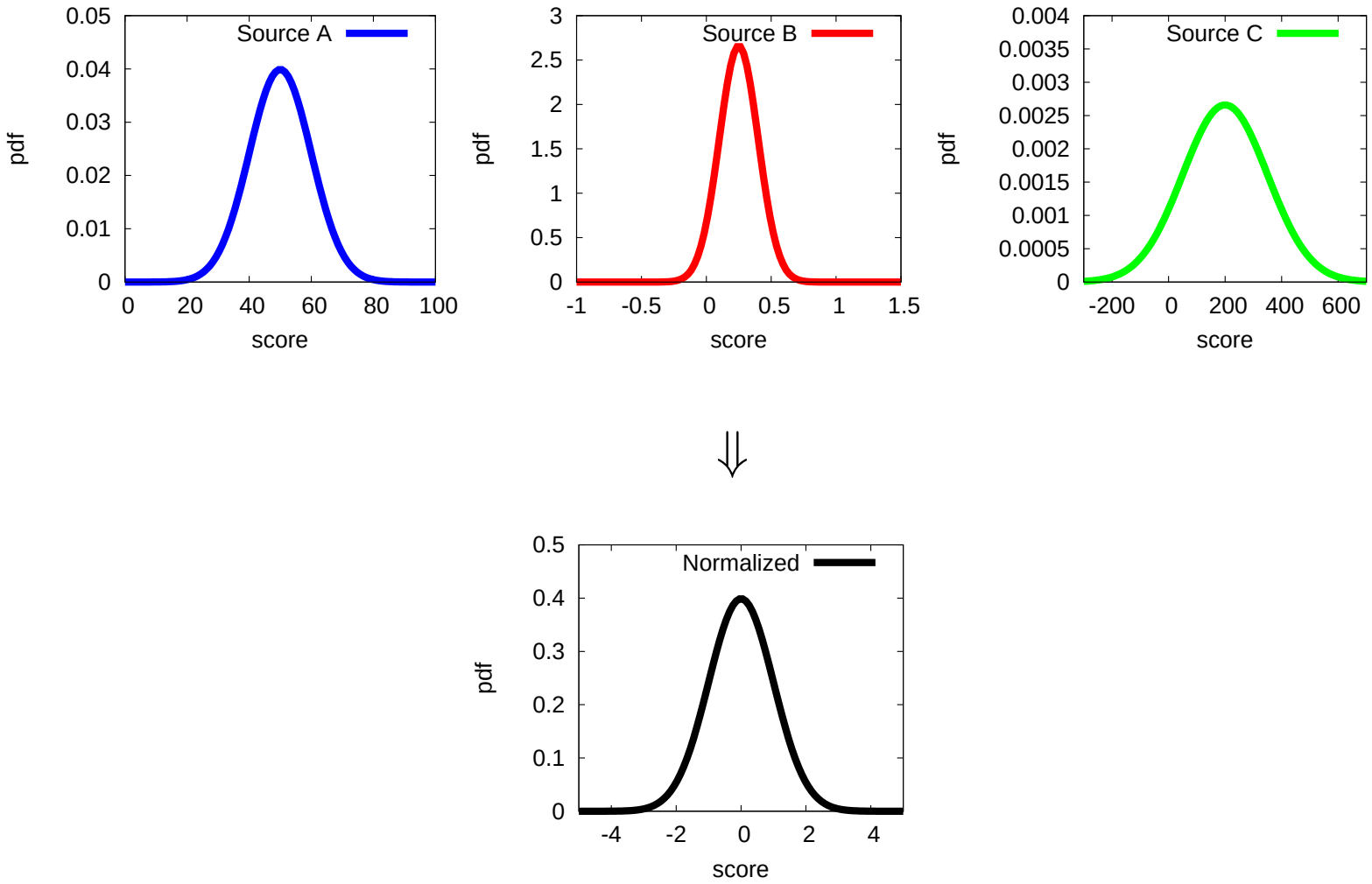
MinMax Results

# sources	top-10		top-100		top-1000	
	MAP	P@10	MAP	P@10	MAP	P@10
30	0.063	0.213	0.110	0.213	0.111	0.213
1000	0.015	0.037	0.017	0.037	0.016	0.037

Z-Score

$$s_{ZScore} = \frac{s - \mu}{\sigma}$$

Z-Score


Source A
pdf
score
0.05
0.04
0.03
0.02
0.01
0
0 20 40 60 80 100
Source B
pdf
score
3
2.5
2
1.5
1
0.5
0
-1 -0.5 0 0.5 1 1.5
Source C
pdf
score
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
-200 0 200 400 600
Normalized
pdf
score
0.5
0.4
0.3
0.2
0.1
0
-4 -2 0 2 4

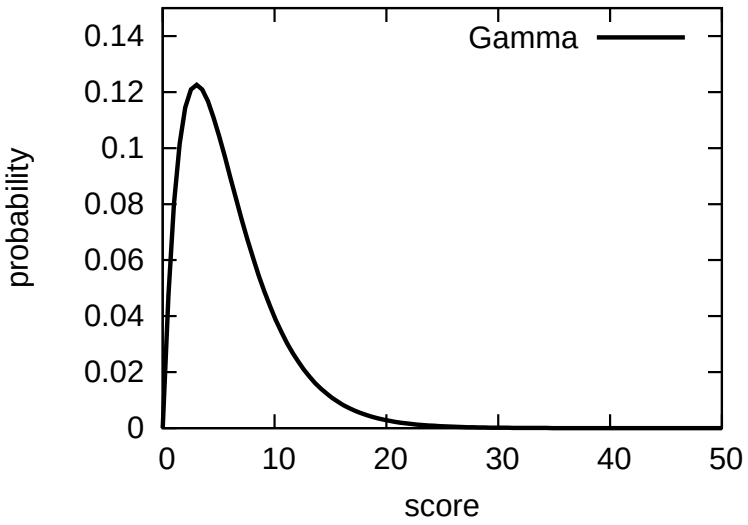
Z-Score Assumptions

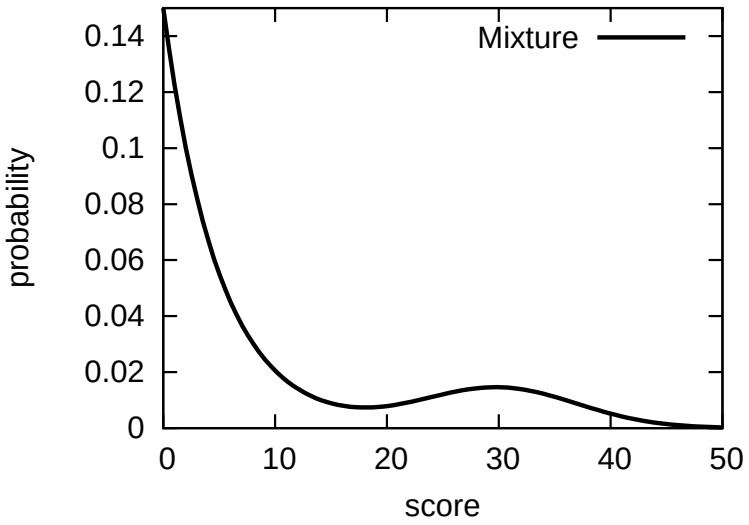
A1 Each collection contains relevant documents

A2 These documents are likely to be ranked high

A3 Document score distributions have the bell shape

Document Score Distributions





Z-Score Results

# sources	top-10		top-100		top-1000	
	MAP	P@10	MAP	P@10	MAP	P@10
30	0.057	0.168	0.059	0.072	0.066	0.133
1000	0.016	0.047	0.010	0.034	0.011	0.035

Sum

$$s' = s - s_{min}, \quad s_{Sum} = \frac{s'}{\sum_i s'_i}$$

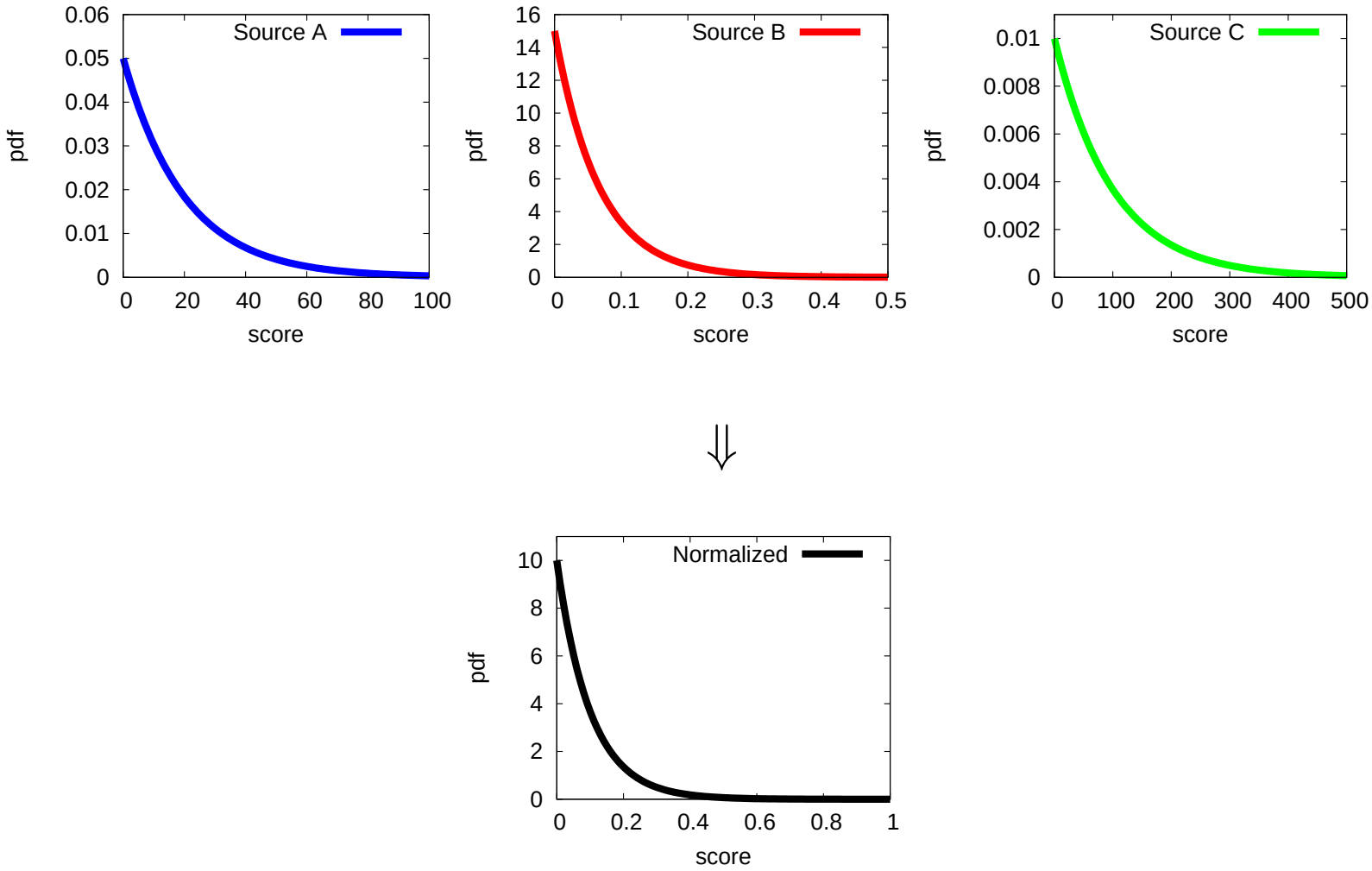
Sum

$$s \sim \mathcal{E}(s;\lambda),\ \lambda = \frac{n}{\sum_i s_i}$$

$$s_{Sum} \sim \mathcal{E}(s_{Sum};\lambda_{Sum})$$

$$\lambda_{Sum} = \frac{n}{\sum_i s_{Sum}} = n$$

Sum


Source A
pdf
score
0.06
0.05
0.04
0.03
0.02
0.01
0
0 20 40 60 80 100
Source B
pdf
score
16
14
12
10
8
6
4
2
0
0 0.1 0.2 0.3 0.4 0.5
Source C
pdf
score
0.01
0.008
0.006
0.004
0.002
0
0 100 200 300 400 500
Normalized
pdf
score
10
8
6
4
2
0
0 0.2 0.4 0.6 0.8 1

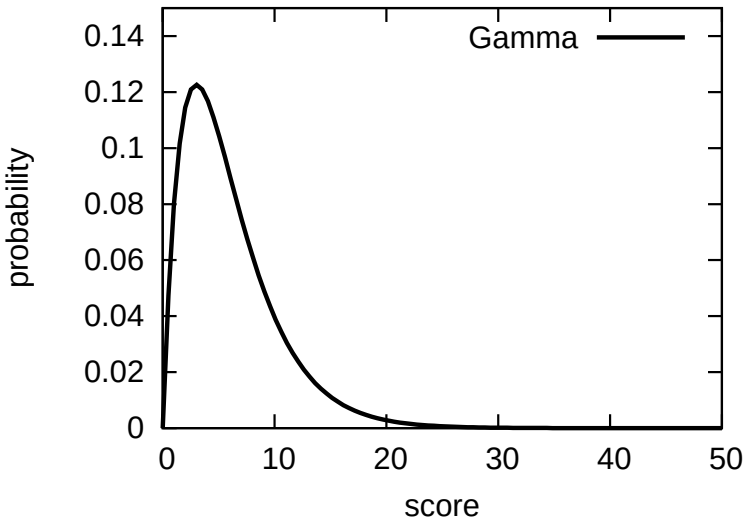
Sum Assumptions

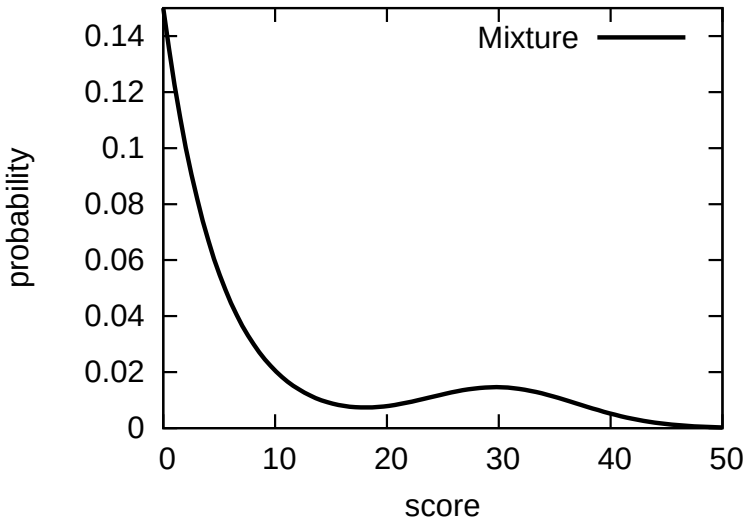
A1 Each collection contains relevant documents

A2 These documents are likely to be ranked high

A3 Document scores are distributed exponentially

Document Score Distributions





Sum Results

# sources	top-10		top-100		top-1000	
	MAP	P@10	MAP	P@10	MAP	P@10
30	0.056	0.158	0.052	0.054	0.035	0.041
1000	0.015	0.039	0.006	0.024	0.004	0.019

Linear Score Normalization Assumptions

A1 Each collection contains relevant documents

A2 These documents are likely to be ranked high

A3 [Z-Score and Sum] Document scores are distributed in a certain way

Linear Score Normalization Results

	top-10		top-100		top-1000	
	MAP	P@10	MAP	P@10	MAP	P@10
MinMax	$0.063^{\triangle}$	$0.213^{\triangle}$	$0.110^{\triangle}$	$0.213^{\triangle}$	$0.111^{\triangle}$	$0.213^{\triangle}$
Z-Score	0.057	0.168	0.059	0.072	0.066	0.133
Sum	$0.056^{\triangledown}$	$0.158^{\triangledown}$	$0.052^{\triangledown}$	$0.054^{\triangledown}$	$0.035^{\triangledown}$	$0.041^{\triangledown}$

Table: 30 sources

	top-10		top-100		top-1000	
	MAP	P@10	MAP	P@10	MAP	P@10
MinMax	0.015	0.037	$0.017^{\triangle}$	0.037	$0.016^{\triangle}$	0.037
Z-Score	0.016	0.047	0.010	0.034	0.011	0.035
Sum	0.015	0.039	$0.006^{\triangledown}$	$0.024^{\blacktriangledown}$	$0.004^{\triangledown}$	$0.019^{\triangledown}$

Table: 1000 sources

$$s_{MinMax} = \frac{s - s_{min}}{s_{max} - s_{min}}$$

Outline

Linear Score Normalization Assumptions

A1 Each collection contains relevant documents

A2 These documents are likely to be ranked high

A3 [Z-Score and Sum] Document scores are distributed in a certain way

Dealing with A1 Assumption

1. Force data to satisfy A1
2. *Change methods to relax A1*

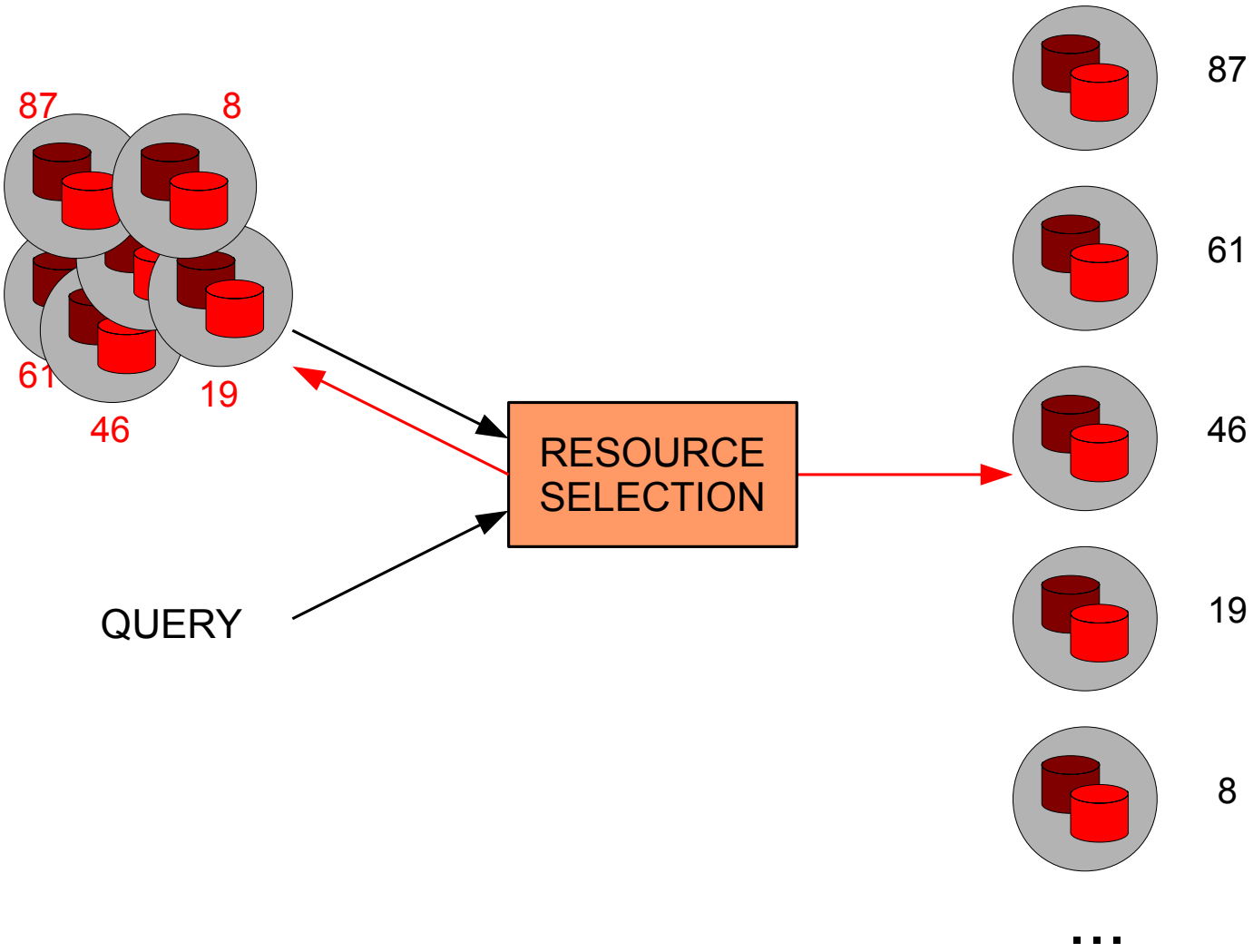
Weighted Linear Score Normalization

$$s_{weighted} = w(C|q) \cdot s_{linear}$$

Calculating Weights

1. Based on a ranked list itself
2. *Based on external evidence*

Resource Selection



CORI Results Merging

$$s_{CORI} = \frac{1 + 0.4 \cdot s_{MinMax}(C|q)}{1.4} \cdot s_{MinMax}$$

Parameterized CORI Results Merging

$$s_{CORI} = \frac{1 + \lambda \cdot s_{MinMax}(C|q)}{1 + \lambda} \cdot s_{MinMax}$$

CORI Results

	top-10		top-100		top-1000	
	MAP	P@10	MAP	P@10	MAP	P@10
CORI(cv)	0.092$^{\triangle}$	0.376$^{\blacktriangle}$	0.154$^{\triangle}$	0.353	0.157	0.361
CORI	0.073	0.349	0.129	0.369$^{\blacktriangle}$	0.162	0.371
SAFE	0.068	0.273	0.087	0.241	0.085	0.234
MinMax	0.063	0.213	0.110	0.213	0.111	0.213

Table: 30 sources

	top-10		top-100		top-1000	
	MAP	P@10	MAP	P@10	MAP	P@10
CORI(cv)	0.049$^{\triangle}$	0.281	0.065$^{\triangle}$	0.289	0.065	0.282
CORI	0.028	0.269	0.042	0.287	0.056	0.292$^{\blacktriangle}$
SAFE	0.031	0.116	0.030	0.114	0.026	0.105
MinMax	0.015	0.039	0.006	0.024	0.004	0.019

Table: 1000 sources

$$s_{MinMax} = \frac{s - s_{min}}{s_{max} - s_{min}}$$

$$s_{MinMax} = w(C|q) \cdot s_{MinMax}$$

$$s_{CORI} = \frac{1 + \lambda \cdot s_{MinMax}(C|q)}{1 + \lambda} \cdot s_{MinMax}$$

Answers