Academic Research to Open Source

Open Source AI Workshop

Andy Hind
April 5, 2019

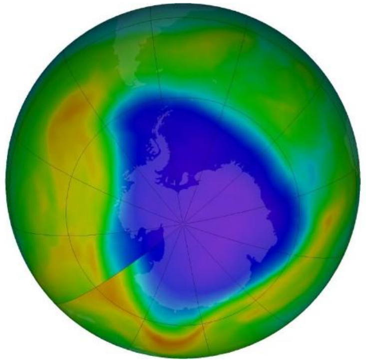
Adaptive Intelligent Apps

Presented by

ORACLE

Who am I?

- Andy Hind - Reformed academic?
  - Oracle
  - Alfresco
  - Campden BRI
  - University of Edinburgh – Chemical Engineering
  - British Antarctic Survey



Agenda

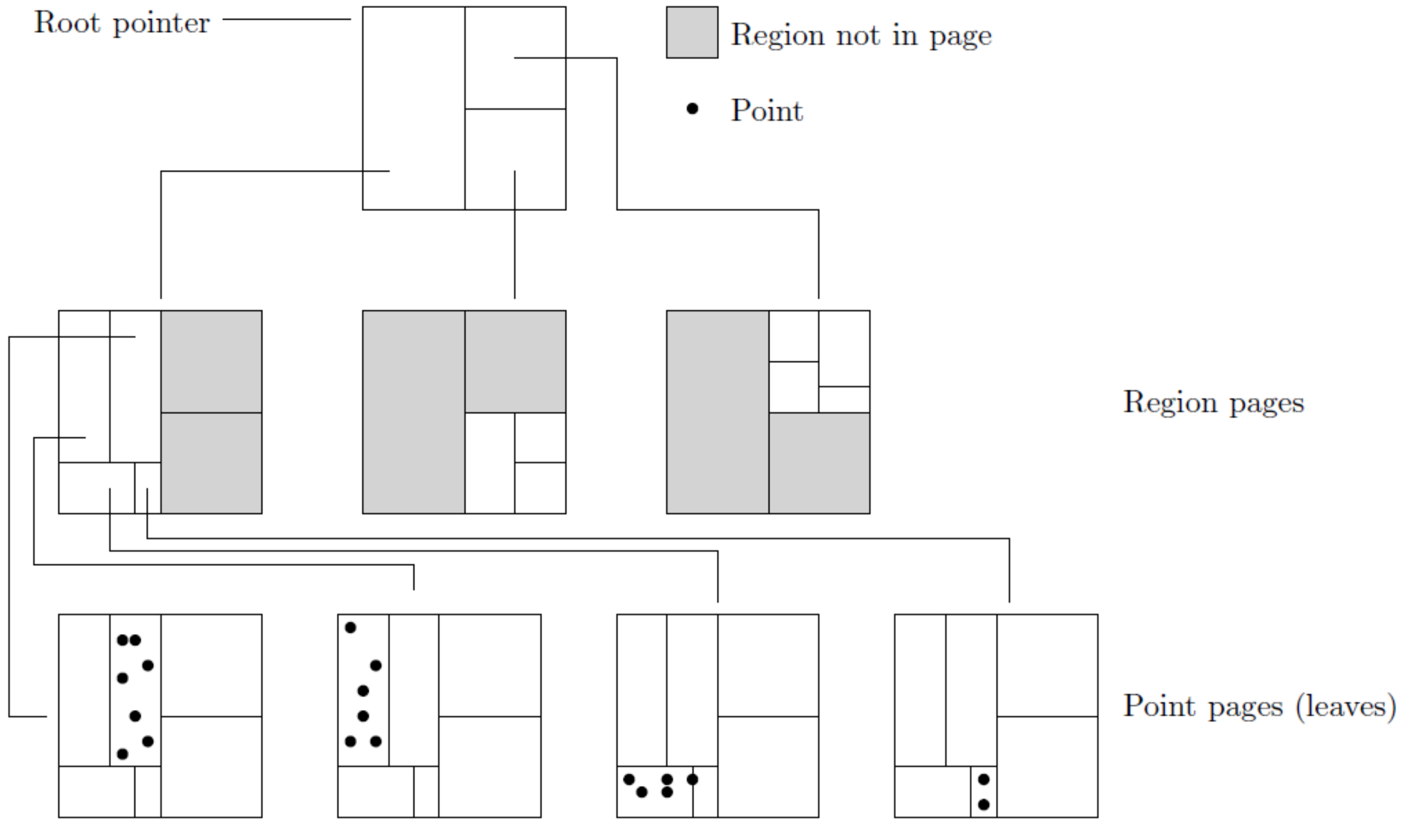
1. **Introduction**
2. Document Fingerprints
3. Getting it into Lucene and SOLR
4. Vectors are interesting …
5. The journey

New ideas appearing in Lucene/SOLR

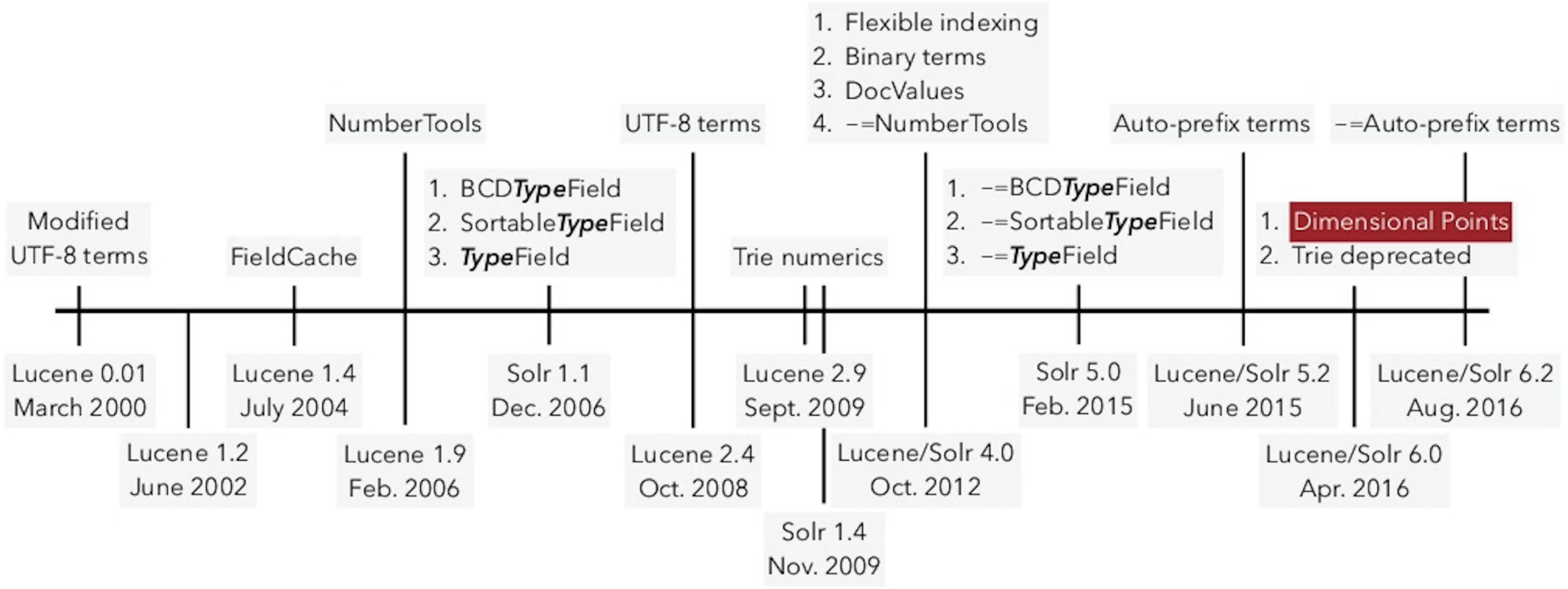
- Learning to rank
  - RankNet 2005/LambdaMART 2010
  - SOLR 2015 (rerank 2014) - Elastic 2017

New ideas appearing in Lucene/SOLR

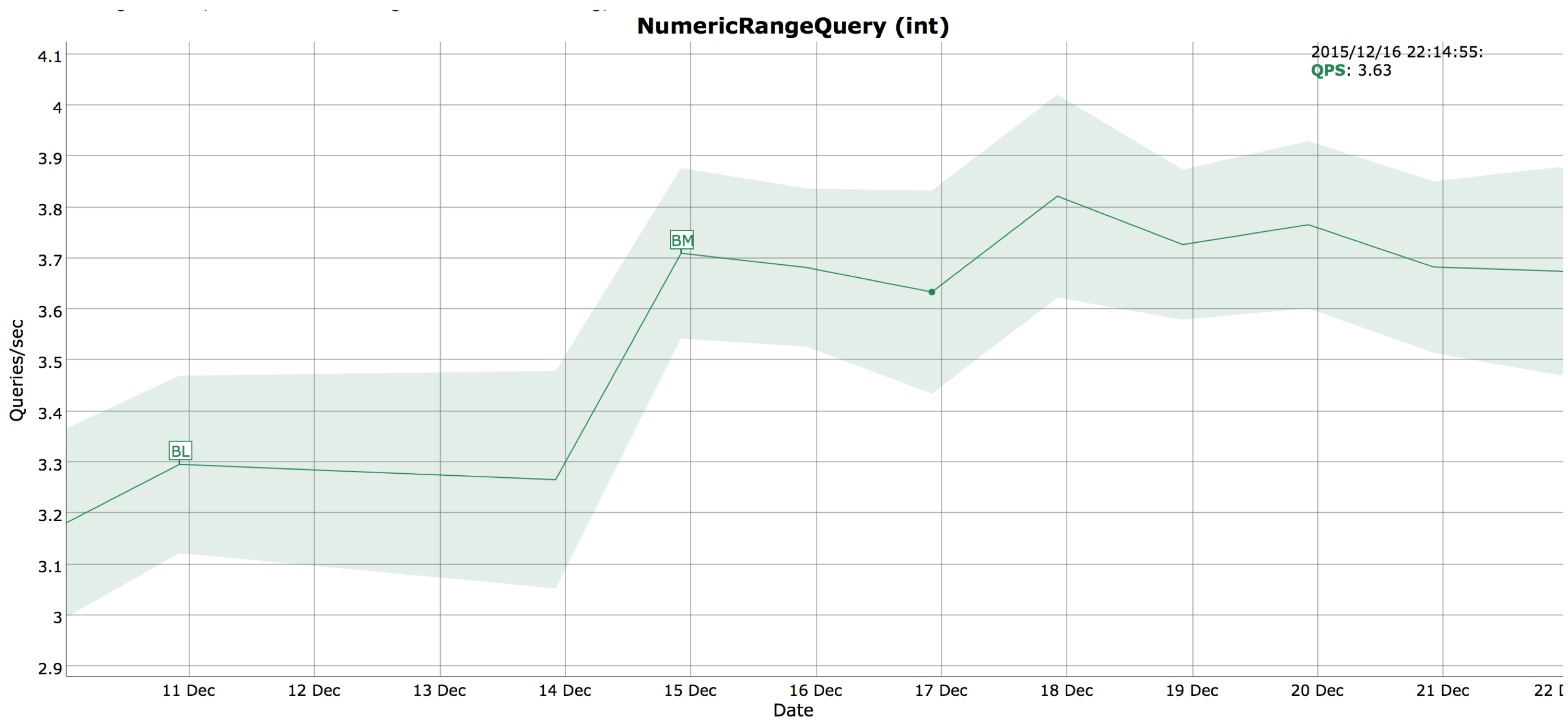
- (b)kd – trees
  - Paper 2003
  - Lucene 2015



Numeric Types in Lucene



Numeric Types in Lucene



Encoded Strings

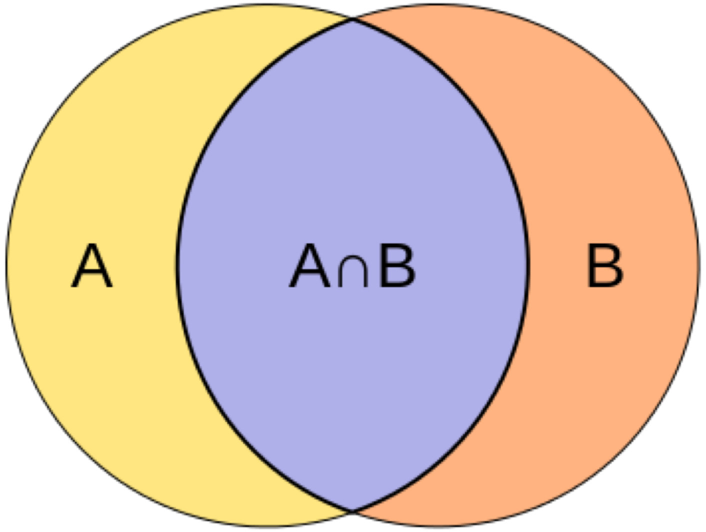
- Encode information in tokens

- Multi-lingual indexing
  - Encode locale/analysis chain … {en}woof

- Many fields
  - Encode field id ….   woof:1
  - Salesforce – Activate 2018

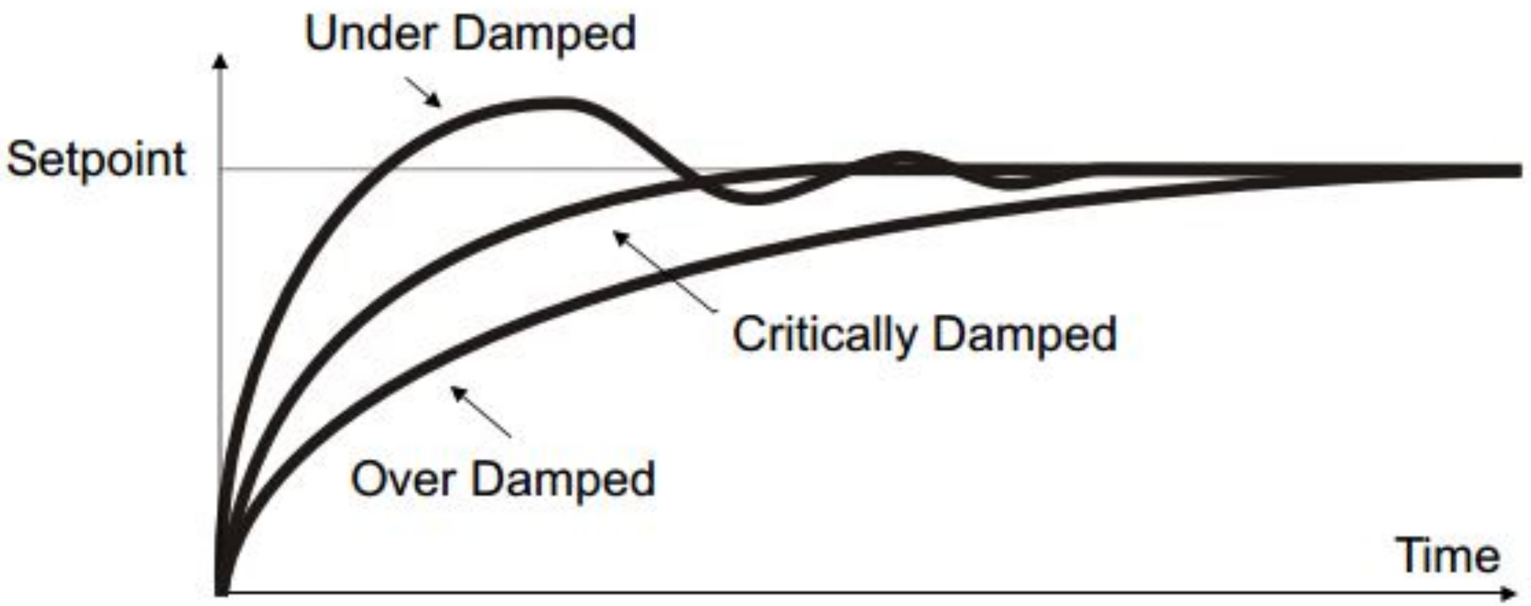
New ideas appearing in Lucene/SOLR

- Locality Sensitive Hashing & Minhash
  - AltaVista - 1997
  - Lucene 2016/SOLR 2018



New ideas appearing in Lucene/SOLR

- PID control from the 1920s …..



Academia

Labs

IP

Industry

Pre-competitive

Production

Practice

Community

Commodity

Open Source

Agenda

1. Introduction
2. **Document Fingerprints**
3. Getting it into Lucene and SOLR
4. Vectors are interesting …
5. The journey

Document Fingerprints

- Document similarity
- “More like this”
  - SOLR term vectors
    - index is 7.8 x larger (http://blog.mikemccandless.com/2012/)
- ???
  - (Near) duplicates
  - Inclusion
  - Query expansion (recall)
  - Feature for LTR (precision)
  - Smaller

Document Fingerprints – LSH – Minhash

- Mining of Massive Datasets - http://www.mmds.org
  - Chapter 3 “Finding Similar Items”
  - Jaccard similarity of documents - BOW
  - Similarity does not have to be high to be significant
    - Character N-grams
    - Word Shingles
  - Minhash
  - Locality Sensitive Hashing – approximate nearest neighbour search
    - Data dependent or independent

Document Fingerprints – LSH – Minhash - Timeline

- 1997 – Andrei Broder – AltaVista - **On the resemblance and containment of documents - https://ieeexplore.ieee.org/document/666900**
- 2012 - Mining of Massive Datasets - http://www.mmds.org
- 2014 - Densifying One Permutation Hashing via Rotation for Fast Near Neighbor Search - http://proceedings.mlr.press/v32/shrivastava14.pdf
- 2014 - Review - Locality Sensitive Hashing – approximate nearest neighbor search. https://arxiv.org/abs/1408.2927
- 2016 – https://issues.apache.org/jira/browse/LUCENE-6968

ORACLE

Document Fingerprints – LSH – Minhash

- Mining of Massive Datasets - http://www.mmds.org
  - Chapter 3 “Finding Similar Items”
  - Jaccard similarity of documents - BOW
  - Similarity does not have to be high to be significant
    - Character N-grams
    - Word Shingles
  - Minhash
  - Locality Sensitive Hashing – approximate nearest neighbour search
  - https://arxiv.org/abs/1408.2927

5 word shingle

Document Fingerprints – LSH – Minhash

- Mining of Massive Datasets - http://www.mmds.org
  - Chapter 3 “Finding Similar Items”
  - Jaccard similarity of documents - BOW
  - Similarity does not have to be high to be significant
    - Character N-grams
    - Word Shingles
  - Minhash
  - Locality Sensitive Hashing – approximate nearest neighbour search
  - https://arxiv.org/abs/1408.2927

5 word shingle

ORACLE

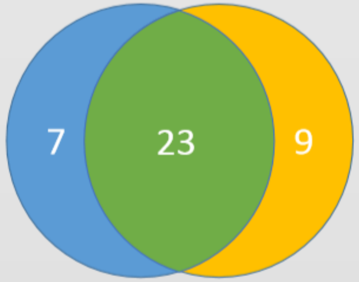
Document Fingerprints - Example

A **CMIS 1.0** **5 word n-grams**

The Content Management Interoperability Services (CMIS) standard defines a domain model and Web Services **and** Restful AtomPub bindings that can be used by applications to work with one or more Content Management repositories/systems.

B **CMIS 1.1** **5 word n-grams**

The Content Management Interoperability Services (CMIS) standard defines a domain model and Web Services, Restful AtomPub **and browser (JSON)** bindings that can be used by applications to work with one or more Content Management repositories/systems.



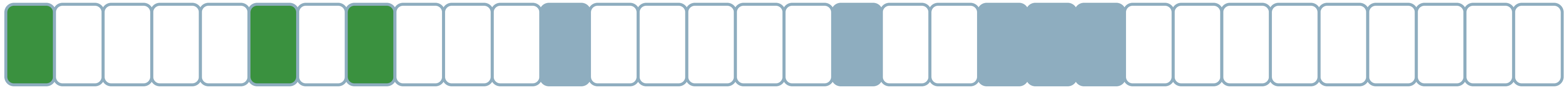
$$C(A,B) = \frac{23}{30} \approx 77\%$$

$$C(B,A) = \frac{23}{32} \approx 72\%$$

$$J(A,B) = \frac{23}{39} \approx 59\%$$

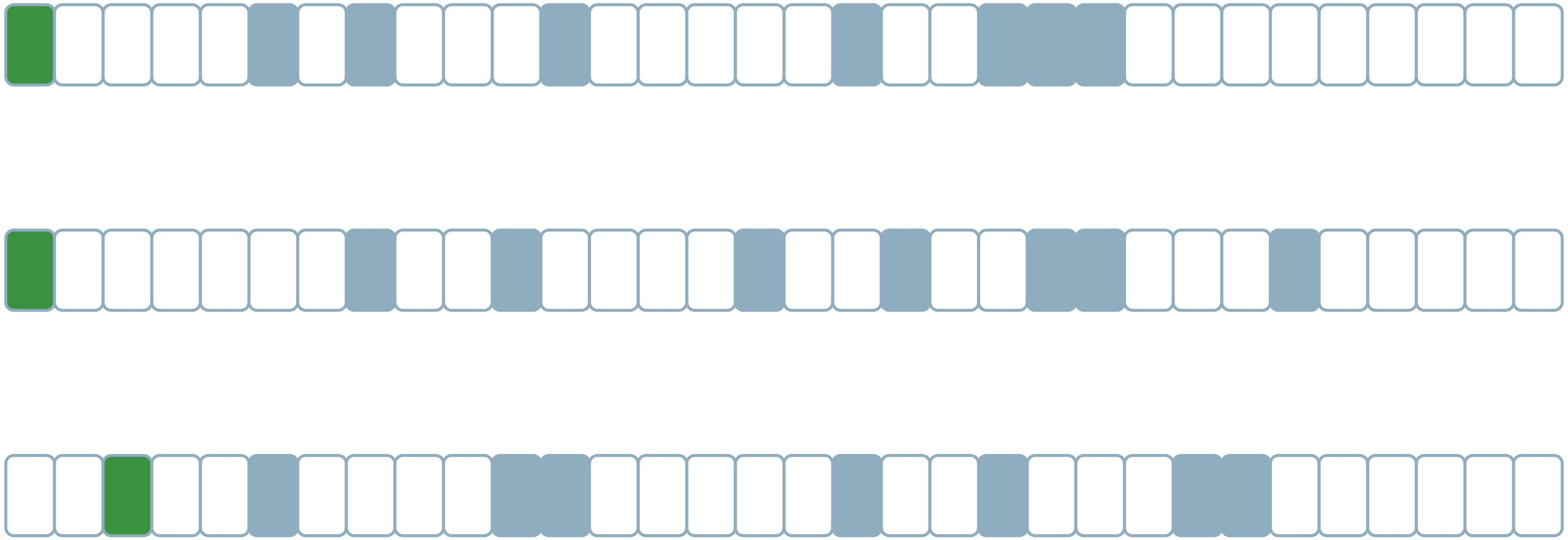
ORACLE

Min Hash – set

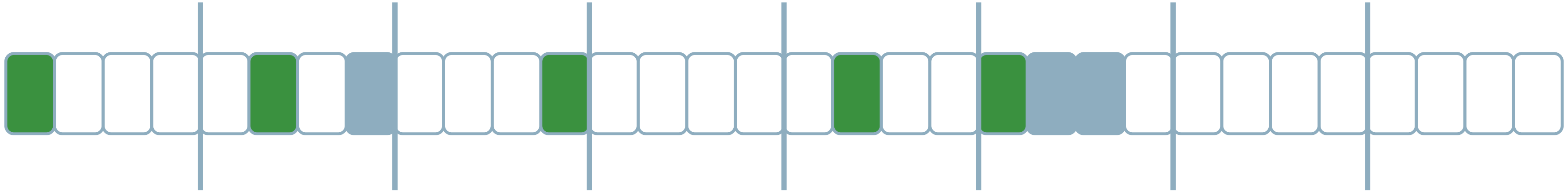


$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

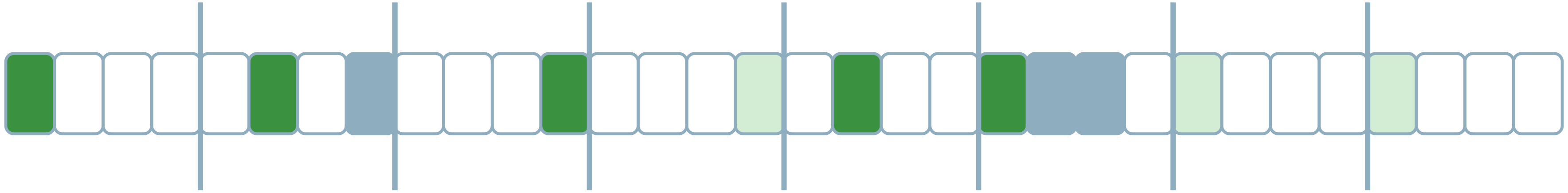
Min Hash – many hash functions



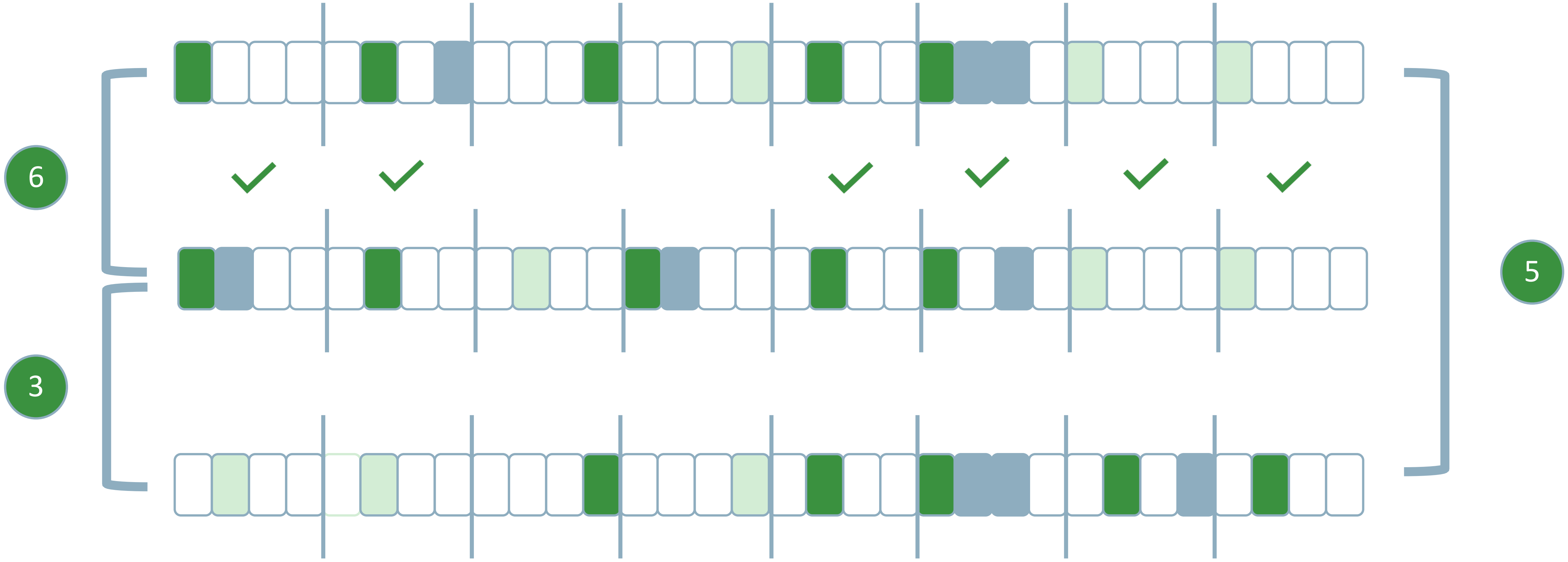
Min Hash – one hash with buckets



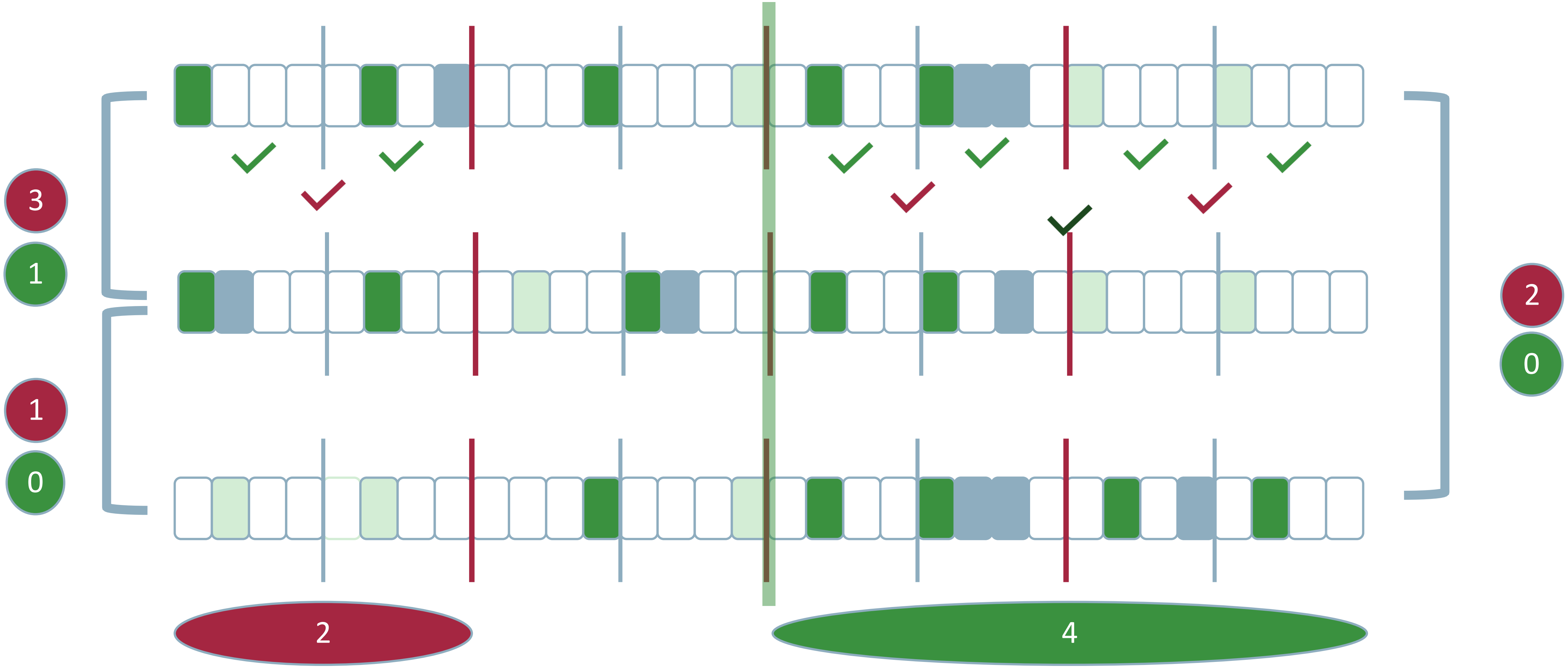
Min Hash – one hash with buckets + rotation



Min Hash – comparing hashes



Min Hash – comparing hashes – with banding



Agenda

Similar Documents

- Lucene
  - MinHashFilter
  - https://issues.apache.org/jira/browse/LUCENE-6968
  - 6 months

- SOLR
  - min_hash MinHashQParser
  - https://issues.apache.org/jira/browse/SOLR-12879
  - 3 days + a month to catch up on documentation

Similar Documents

- Analysed   vs   pre-analysed and stored
- Analysis chain
  - n-grams vs shingles etc
- Hashes, buckets, minimum set, rotation

- Similarity

Examples

- Wikipedia articles
- 5 - word shingles
- Pre-analysed and stored

- Aside
  - State in the index
  - Event sourcing/CQRS

Oracle Corporation

Page	Score	Normalised
Oracle Corporation	512	1.000
Oracle Cloud	9	0.018
Oracle Cloud Platform	5	0.010
Michelle K. Lee	5	0.010
Paul Grewal	4	0.008
Ultratech	4	0.008

Oracle Cloud

Page	Score	Normalised
Oracle Cloud	512	1.000
Oracle Cloud Platform	148	0.289
Oracle Corporation	17	0.033
Microsoft Azure	10	0.020
Recovery as a service	9	0.018
SHI International Corp	8	0.016
Cloud28+	8	0.016
Content as a service	6	0.012

ORACLE

Brexit

Page	Score	Normalised
Brexit	512	1.000
Brexit negotiations	30	0.059
Brexit in popular culture	22	0.043
History of European Union–United Kingdom relations	19	0.037
Economic effects of Brexit	11	0.021
European Parliament election, 2019	8	0.016
Aftermath of the United Kingdom European Union membership referendum, 2016	7	0.014
United Kingdom invocation of Article 50 of the Treaty on European Union	7	0.014

Scott Joplin

Page	Score	Normalised
Scott Joplin	512	1.000
Treemonisha	38	0.074
List of compositions by Scott Joplin	15	0.030
The Entertainer (rag)	15	0.030
Scott Joplin House State Historic Site	13	0.025
Scott Joplin: Piano Rags	13	0.025
Joshua Rifkin	7	0.014
Bethena	5	0.010

ORACLE

Stuff still to do …

- New BKD type
- Normalised score
- Positions/Highlighting
  - Where was it similar?
- Hash size and collisions
- Rotation bug …

Agenda

1. Introduction
2. Document Fingerprints
3. Getting it into Lucene and SOLR
4. **Vectors are interesting …**
5. The journey

ORACLE

Vectors are interesting …

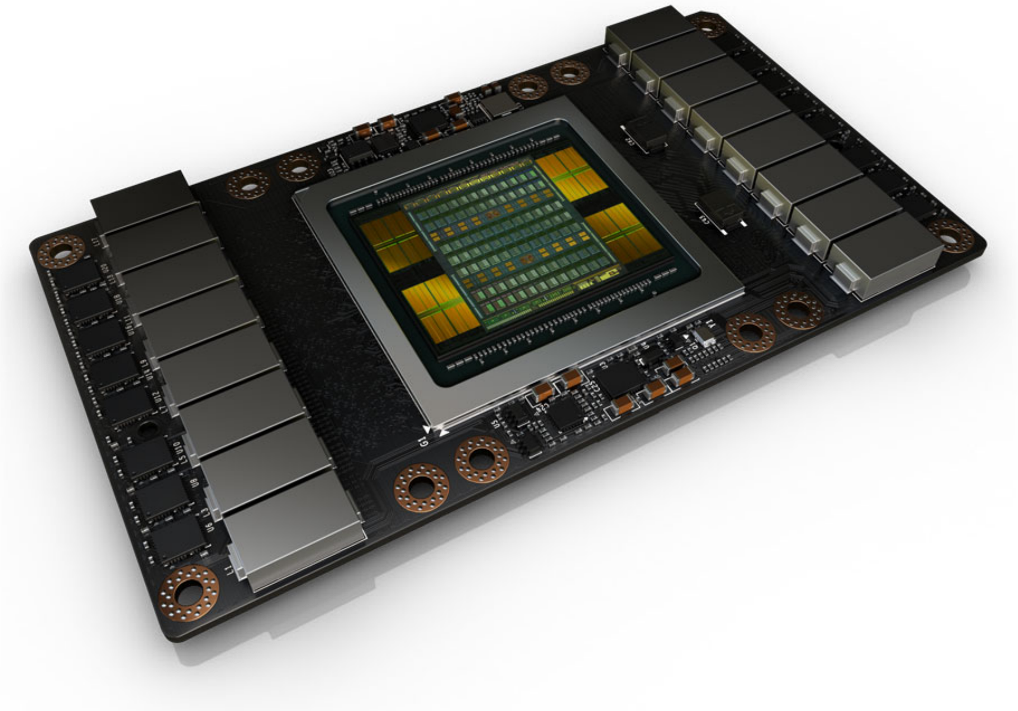
- Dense Vectors
  - Embeddings
  - Post processing

- Approximate nearest neighbours
  - Locality Sensitive Hashing (LSH – SimHash, spectral hashing, …)
  - K-Means tree
  - Randomized KD forest
  - Vector to text encoding
  - Brute force

ORACLE

NLP

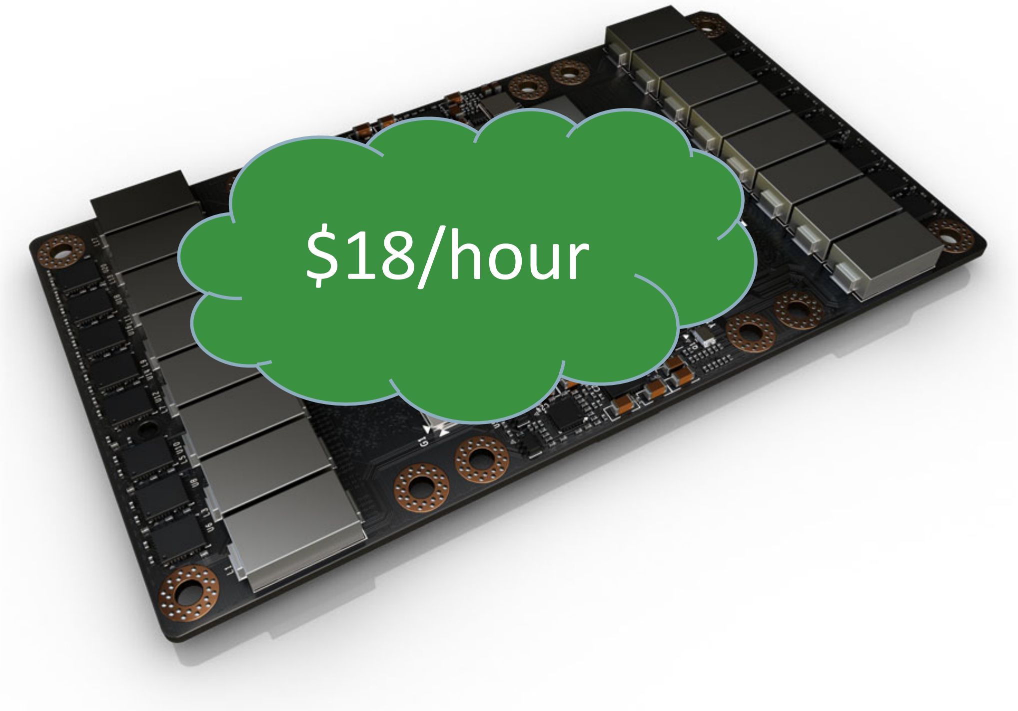
- **NVIDIA DGX-1 Deep Learning System with 8x 32GB Tesla V100 Volta GPUs, 12nm, HBM2, 1 petaFLOP FP16 Performance**



ORACLE

NLP

- **NVIDIA DGX-1 Deep Learning System with 8x 32GB Tesla V100 Volta GPUs, 12nm, HBM2, 1 petaFLOP FP16 Performance**



Vectors are interesting …

- NLP and transfer learning
  - Sentence representations
  - NLP’s image net moment?
  - Transfer learning needs less data ….

- Text vs images

Agenda

1. Introduction
2. Document Fingerprints
3. Getting it into Lucene and SOLR
4. Vectors are interesting …
5. **The journey**

ORACLE

The Journey

- Papers …
- Math vs application
- Implementation
- Scalability

Papers …

- Performance Comparison of Learning to Rank Algorithms for Information Retrieval

TABLE I. PERFORMANCE COMPARISON BETWEEN BASE ALGORITHMS AND LEARNING TO RANK ALGHORITHMS

Algorithm	NDCG@10
TF*IDF	0.7051
BM25	0.7800
RankSVM	0.8087
LambdaMART	0.8092
AdditiveGroves	**0.8165**

- https://pdfs.semanticscholar.org/cd12/e191d2c2790e5ed60e5186462e6f8027db1f.pdf

ORACLE

Papers …

Deep Rank: A new deep architecture for Relevance Ranking in Information Retrieval (https://arxiv.org/abs/1710.05649)

MQ2008

Model	NDCG@1	NDCG@3	NDCG@5	NDCG@10	P@1	P@3	P@5	P@10	MAP
BM25-TITLE	0.344^-	0.420^-	0.461^-	0.220^-	0.408^-	0.381^-	0.337^-	0.245^-	0.465^-
RANKSVM	0.375^-	0.431^-	0.479^-	0.229	0.441^-	0.390^-	0.348^-	0.249	0.478^-
RANKBOOST	0.381	0.436^-	0.477^-	0.231	0.455	0.392^-	0.347^-	0.248	0.481^-
ADARANK	0.360^-	0.422^-	0.462^-	0.222	0.430^-	0.384^-	0.339^-	0.247^-	0.468^-
LAMBDAMART	0.378	0.437^-	0.477^-	0.231	0.446^-	0.398	0.348^-	0.251	0.478^-
DSSM	0.286^-	0.336^-	0.378^-	0.178^-	0.341^-	0.307^-	0.284^-	0.221^-	0.391^-
CDSSM	0.283^-	0.331^-	0.376^-	0.175^-	0.335^-	0.302^-	0.279^-	0.222^-	0.395^-
ARC-I	0.295^-	0.363^-	0.413^-	0.187^-	0.361^-	0.336^-	0.311^-	0.229^-	0.424^-
SQA-NOFEAT	0.291^-	0.350^-	0.401^-	0.184^-	0.366^-	0.332^-	0.309^-	0.231^-	0.416^-
DRMM	0.368^-	0.427^-	0.468^-	0.220^-	0.437^-	0.392^-	0.344^-	0.245^-	0.473^-
ARC-II	0.299^-	0.340^-	0.394^-	0.181^-	0.366^-	0.326^-	0.305^-	0.229^-	0.413^-
MATCHPYRAMID	0.351^-	0.401^-	0.442^-	0.211^-	0.408^-	0.365^-	0.329^-	0.239^-	0.449^-
MATCH-SRNN	0.369^-	0.426^-	0.465^-	0.223^-	0.432^-	0.383^-	0.335^-	0.239^-	0.466^-
DEEPRANK-2DGRU	0.391	0.436	0.480	0.236	0.462	0.395	0.354	0.252	0.489
DEEPRANK-CNN	**0.406**	**0.460**	**0.496**	**0.240**	**0.482**	**0.412**	**0.359**	**0.252**	**0.498**
SQA	0.402	0.454	0.493	0.236	0.485	0.411	0.362	0.254	0.496
DEEPRANK-CNN-FEAT	0.418	0.475	0.507	0.248	0.497	0.422	0.366	0.255	0.508

ORACLE

Papers …

- A Dual Embedding Space Model for Document Ranking

	Explicitly Judged Test Set			Implicit Feedback based Test Set		
	NDCG@1	NDCG@3	NDCG@10	NDCG@1	NDCG@3	NDCG@10
BM25	21.44	26.09	37.53	11.68	22.14	33.19
LSA	04.61*	04.63*	04.83*	01.97*	03.24*	04.54*
DESM (IN-IN, trained on body text)	06.69*	06.80*	07.39*	03.39*	05.09*	07.13*
DESM (IN-IN, trained on queries)	05.56*	05.59*	06.03*	02.62*	04.06*	05.92*
DESM (IN-OUT, trained on body text)	01.01*	01.16*	01.58*	00.78*	01.12*	02.07*
DESM (IN-OUT, trained on queries)	00.62*	00.58*	00.81*	00.29*	00.39*	01.36*
BM25 + DESM (IN-IN, trained on body text)	21.53	26.16	37.48	11.96	22.58*	33.70*
BM25 + DESM (IN-IN, trained on queries)	**21.58**	26.20	37.62	11.91	22.47*	33.72*
BM25 + DESM (IN-OUT, trained on body text)	21.47	26.18	37.55	11.83	22.42*	33.60*
BM25 + DESM (IN-OUT, trained on queries)	21.54	**26.42***	**37.86***	**12.22***	**22.96***	**34.11***

- https://arxiv.org/abs/1602.01137

ORACLE

Papers …

- Learning a Deep Listwise Context Model for Ranking Refinement

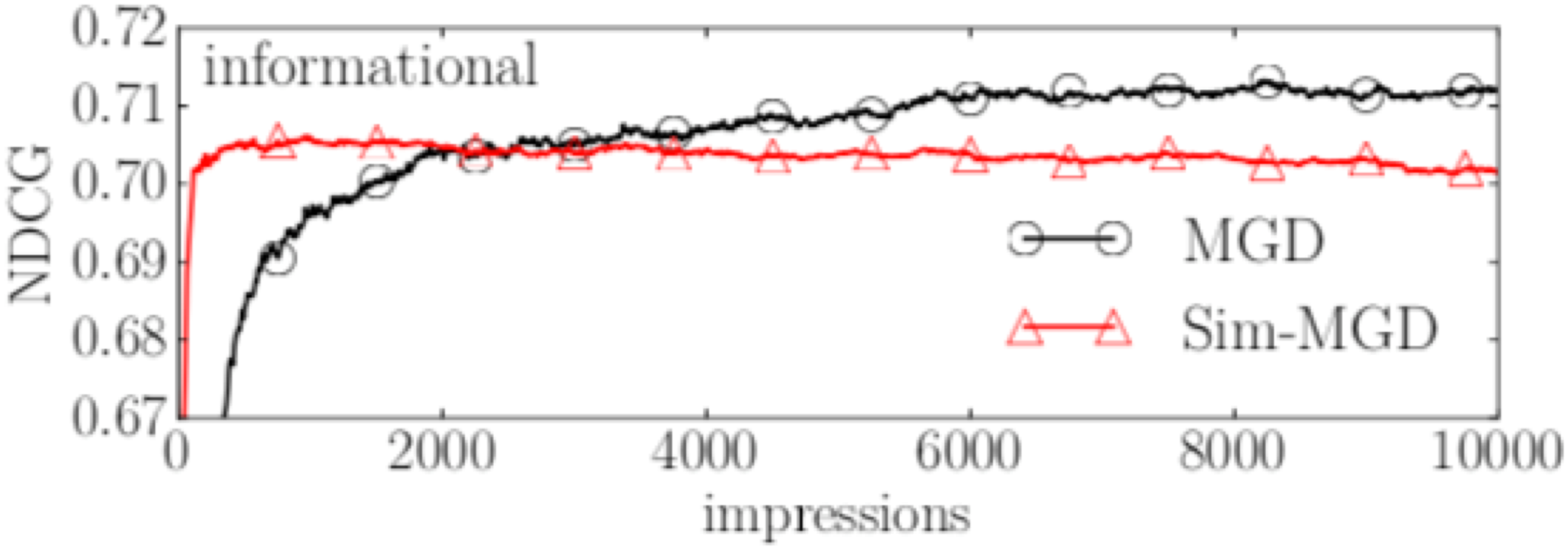
			Microsoft Letor Dataset 30K							
Initial List	Model	Loss Function	nDCG@1	ERR@1	nDCG@3	ERR@3	nDCG@5	ERR@5	nDCG@10	ERR@10
LambdaMART			0.457^{+}	0.235^{+}	0.442^{+}	0.314^{+}	0.445^{+}	0.336^{+}	0.464^{+}	0.355^{+}
LambdaMART	DNN	ListMLE	0.372	0.174	0.378	0.254	0.386	0.278	0.409	0.299
		SoftRank	0.384	0.209	0.373	0.281	0.378	0.302	0.397	0.321
		AttRank	0.388	0.199	0.386	0.274	0.393	0.297	0.416	0.317
	LIDNN	ListMLE	0.427^{+}	0.219^{+}	0.427^{+}	0.301^{+}	0.435^{+}	0.325^{+}	0.455^{+}	0.344^{+}
		SoftRank	0.457^{+}	0.234^{+}	0.442^{+}	0.314^{+}	0.445^{+}	0.336^{+}	0.464^{+}	0.355^{+}
		AttRank	0.455^{+}	0.237^{+}	0.432^{+}	0.312^{+}	0.436^{+}	0.334^{+}	0.458^{+}	0.354^{+}
	DLCM	ListMLE	0.457^{+}	0.235^{+}	0.442^{+}	0.314^{+}	0.445^{+}	0.336^{+}	0.464^{+}	0.355^{+}
		SoftRank	$\mathbf{0.463}^{*+\ddagger}$	$0.243^{*+\ddagger}$	$0.444^{*+\ddagger}$	$0.320^{*+\ddagger}$	$0.447^{*+\ddagger}$	$0.342^{*+\ddagger}$	$0.465^{*+\ddagger}$	$0.360^{*+\ddagger}$
		AttRank	$\mathbf{0.463}^{*+\ddagger}$	$\mathbf{0.246}^{*+\ddagger}$	$\mathbf{0.445}^{*+\ddagger}$	$\mathbf{0.322}^{*+\ddagger}$	$\mathbf{0.450}^{*+\ddagger}$	$\mathbf{0.344}^{*+\ddagger}$	$\mathbf{0.469}^{*+\ddagger}$	$\mathbf{0.362}^{*+\ddagger}$

https://arxiv.org/pdf/1804.05936.pdf

ORACLE

Papers …

- Balancing Speed and Quality in Online Learning to Rank for Information Retrieval https://arxiv.org/pdf/1711.09446.pdf




Adaptive
Intelligent
Apps
Presented by
ORACLE
Thank you
We’re hiring!
Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |