

Variable Bit Quantisation for LSH

Sean Moran

School of Informatics
University of Edinburgh



TU Graz, August 2013

Variable Bit Quantisation for LSH

Fast search in large-scale datasets

Locality Sensitive Hashing (LSH)

Variable Bit Quantisation (VBQ)

Evaluation

Summary

Variable Bit Quantisation for LSH

Fast search in large-scale datasets

Locality Sensitive Hashing (LSH)

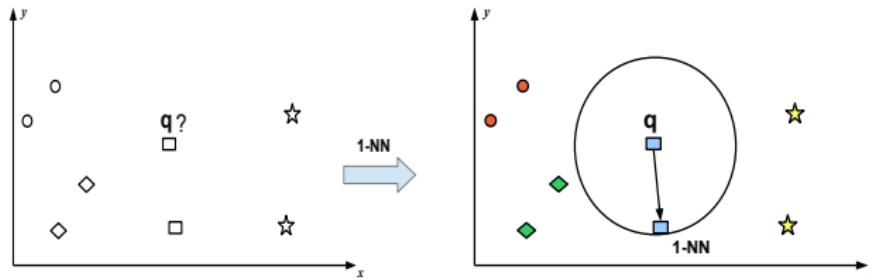
Variable Bit Quantisation (VBQ)

Evaluation

Summary

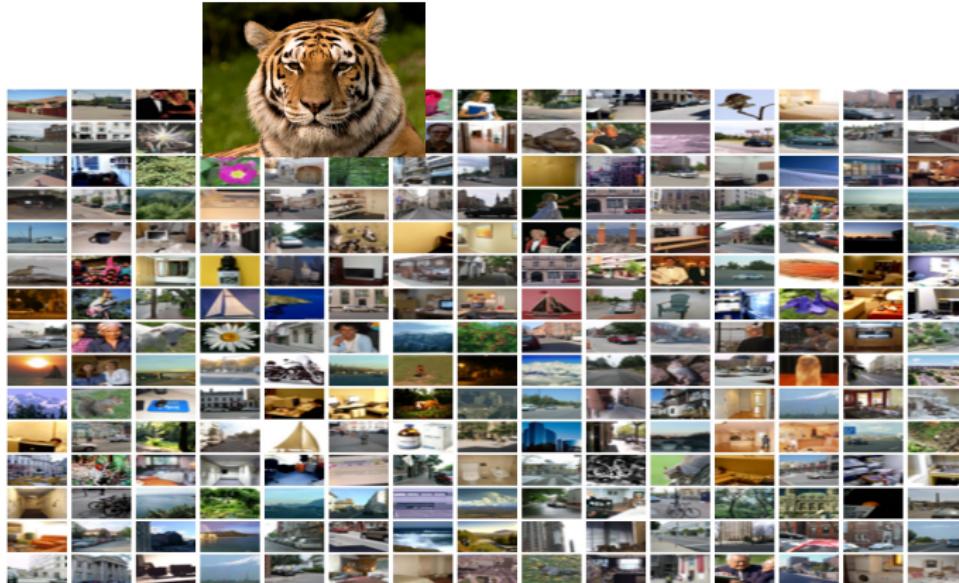
Fast search in large-scale datasets

- ▶ Problem: retrieve nearest neighbour(s) to a given query item



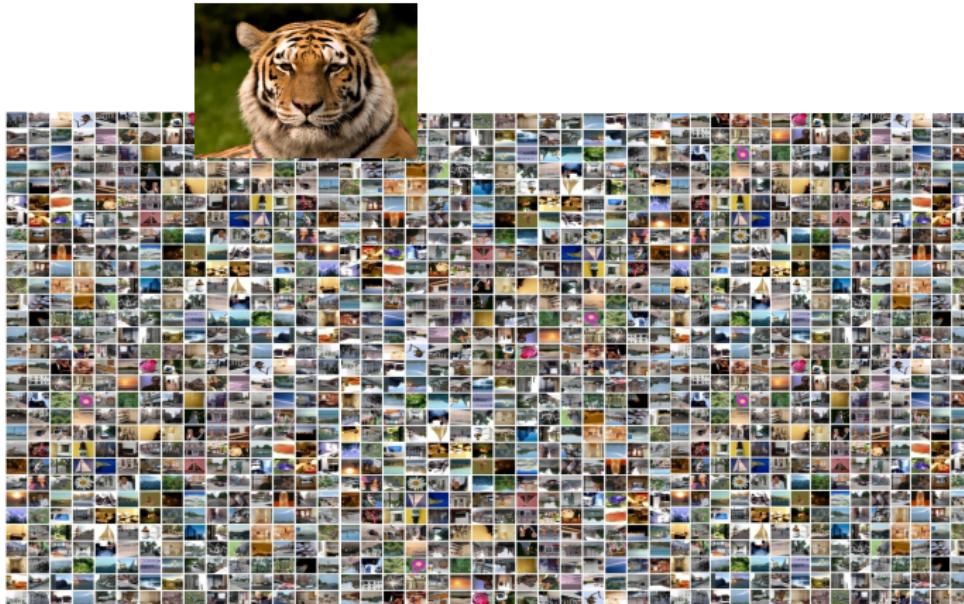
- ▶ Naïve approach: compare query to all N database items
 - ▶ Scales linearly $\mathcal{O}(N)$ with the size of the database
 - ▶ Impractical for all but the smallest of databases

Fast search in large-scale datasets



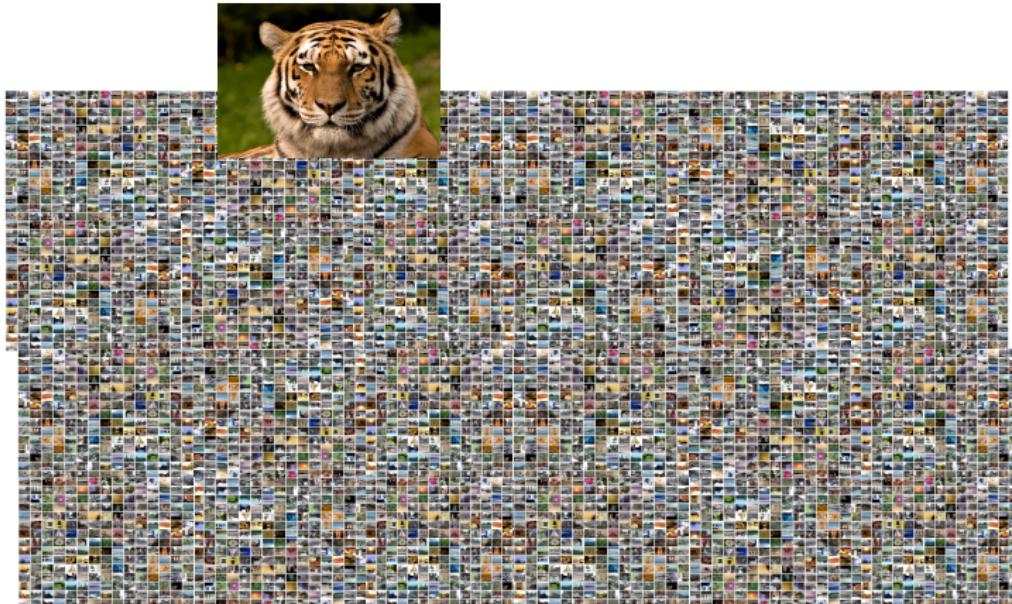
[1] M. Norouzi and D. Fleet. *Minimal Loss Hashing*. ICML '11.

Fast search in large-scale datasets



[1] M. Norouzi and D. Fleet. *Minimal Loss Hashing*. ICML '11.

Fast search in large-scale datasets



[1] M. Norouzi and D. Fleet. *Minimal Loss Hashing*. ICML '11.

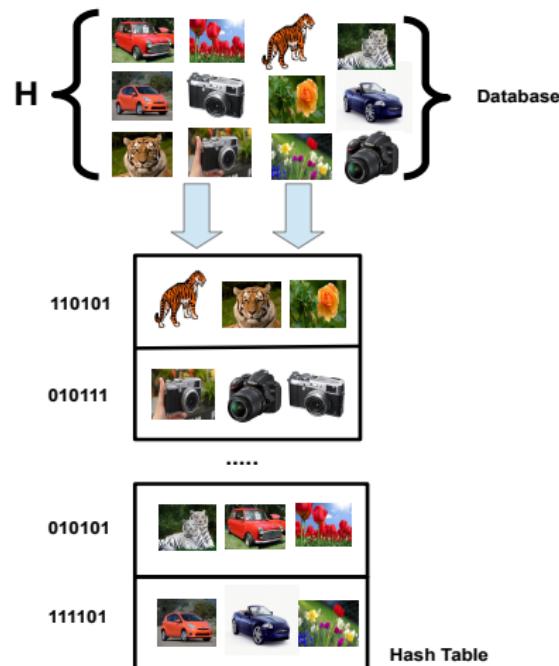
Approximate Nearest Neighbour

- ▶ An approximate nearest neighbour may almost be as good as the exact neighbour
- ▶ Do not guarantee to return the true nearest neighbour in every case, in return for improved speed or memory savings
- ▶ We can solve the nearest neighbour problem by enumerating all approximate nearest neighbours and returning the closest point.

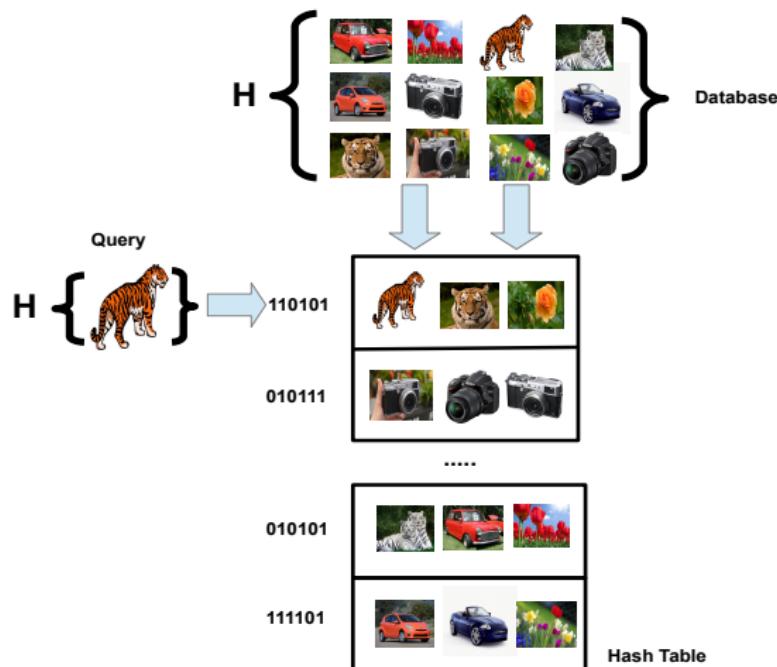
Hashing-based search using binary codes



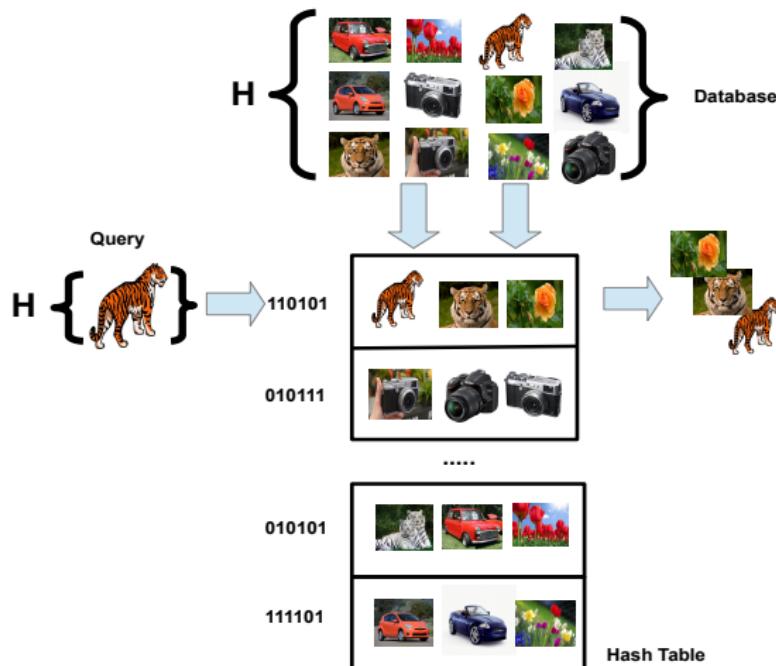
Hashing-based search using binary codes



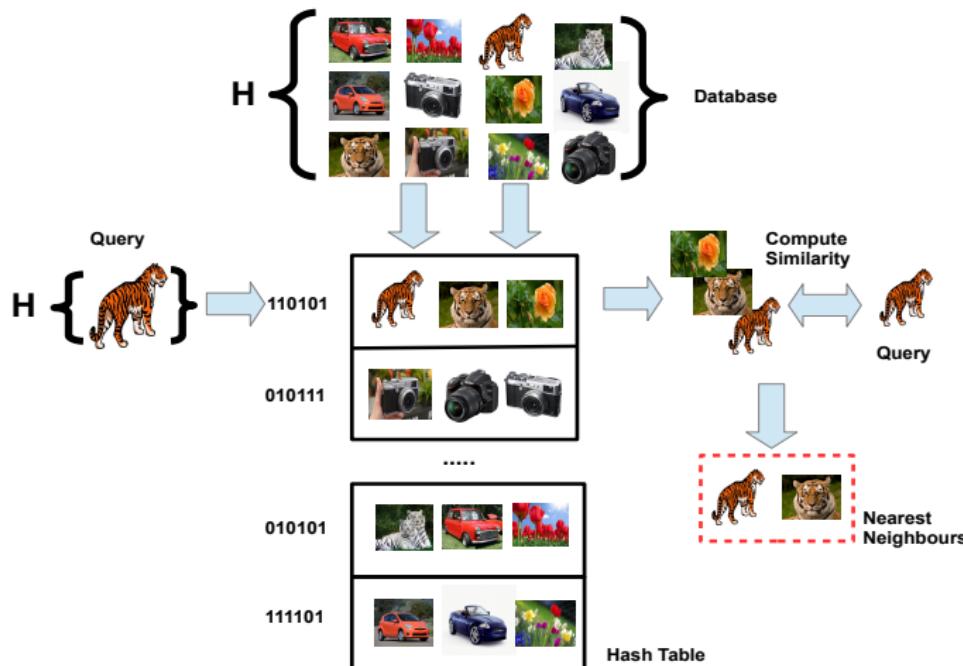
Hashing-based search using binary codes



Hashing-based search using binary codes



Hashing-based search using binary codes



Why transform our data to binary codes?

- ▶ Constant $O(1)$ query time with respect to the dataset size.
- ▶ Binary code comparison requires few machine instructions (XOR followed by popcount).
- ▶ Binary codes are extremely compact e.g. 16Mb to store 1 million, 128 bit encoded data points.

Applications

More Items to Consider

You viewed



The Appeal
John Grisham
Paperback
\$14.00 \$11.20

imense
picturesearch

castle
 Advanced options
 Model released Rights managed
 Property released Royalty free

lightbox

To and plot
Content based IR

Recommendation



The Innocent Man
John Grisham
Mass Market Paperback
\$7.99



The Associate: A Novel
John Grisham
Mass Market Paperback

amazon.com



Ford County: Stories
John Grisham
Paperback
\$15.00 \$8.19



Location Recognition

computer science

mechanical engineering
electrical engineering
chemical engineering
Civil Engineering
ECONOMICS
ENGINEERING

TSUNAMI

tidal wave
LANDSLIDE
EARTH
volcanic
HAILSTORM
Typhoon

Louis Vuitton

PRADA
Fendi
BURBERRY
GUCCI

Noun Clustering

SFGate.com

Obama Takes on Question of Faith

By MICHAEL KITKIN, Associated Press Writer
Monday, January 21, 2008

(AP) —
Barack Obama is stepping up his effort to revert the misconception that he's a Muslim now that the presidential campaign has hit the Bible belt.

At a rally to kick off a weekend campaign swing for the South Carolina primary, Obama tried to set the record straight by attacking church doctrine while on the Unites States Air Force One flight to

- 1. Tijuana Police arrest man found alone in San Diego
- 2. Out of love or death: Driven by love, young man kills his wife
- 3. I could see Phoenix could be a great place to live, so I moved here
- 4. The most popular place to buy a house in the U.S.
- 5. Barack Obama is running for the Senate Committee's
- 6. Obama wants to develop land in AT&T
- 7. More than 100,000 people have applied for stimulus aid
- 8. Obama takes on question of faith
- 9. Obama takes on question of faith
- 10. Obama takes on question of faith

Near Duplicate Detection

The New York Times U.S.

POLITICS / DOMESTIC / ECONOMY / TECHNOLOGY / SCIENCE / HEALTH / OPINION / ARTS / ENTERTAINMENT / BOOKS / METRO

Obama Takes on Question of Faith

By THE ASSOCIATED PRESS
COLUMBIA, S.C. (AP) — Barack Obama is stepping up his effort to revert the misconception that he's a Muslim now that the presidential campaign has hit the Bible belt.

As a rally to kick off a weekend campaign swing for the South Carolina primary, Obama tried to set the record straight by attacking church doctrine while on the Unites States Air Force One flight to

- 1. Michael S. Knight (History), Harvard, Experience
- 2. Paul Krugman (Economics), Princeton, Experience
- 3. Michael Lewis (Finance), Princeton, Experience
- 4. Michael David Red, White and Blue Pig Tales
- 5. Roger Cohen (U.S. Politics and Foreign Affairs), The New York Times
- 6. Steve Kroft (Worldview), Host of 60 Minutes
- 7. Bob Woodward (Politics), Washington Post
- 8. Carl Bernstein (Politics), Washington Post

Variable Bit Quantisation for LSH

Fast search in large-scale datasets

Locality Sensitive Hashing (LSH)

Variable Bit Quantisation (VBQ)

Evaluation

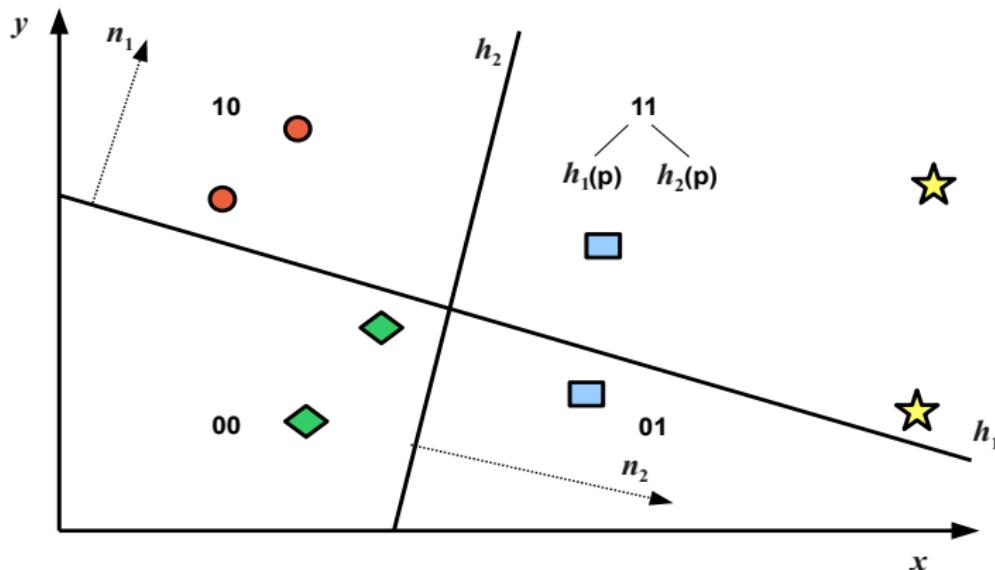
Summary

Computing similarity preserving binary codes

- ▶ Locality Sensitive Hashing (LSH) [1]:
 - ▶ Randomised algorithm for approximate nearest neighbour search
 - ▶ Generates similarity preserving binary codes (fingerprints) for a dataset: data points that are close together are likely to have similar fingerprints.
 - ▶ Compare similarity between data points based upon the fingerprints.
 - ▶ Preserved similarity depends on the selected hash family

[1] P. Indyk and R. Motwani. *Approximate nearest neighbors: removing the curse of dimensionality*. In STOC '98.

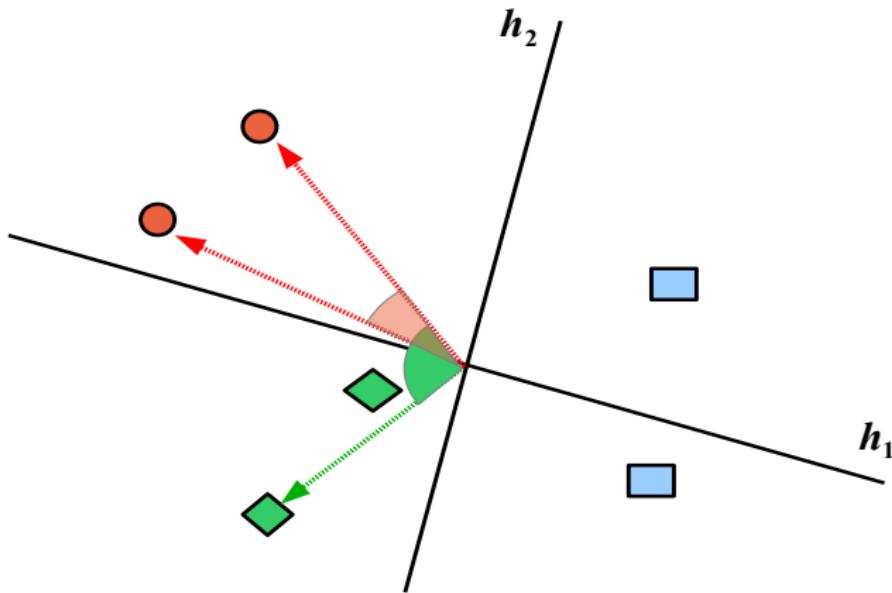
Locality Sensitive Hashing (Cosine Similarity)



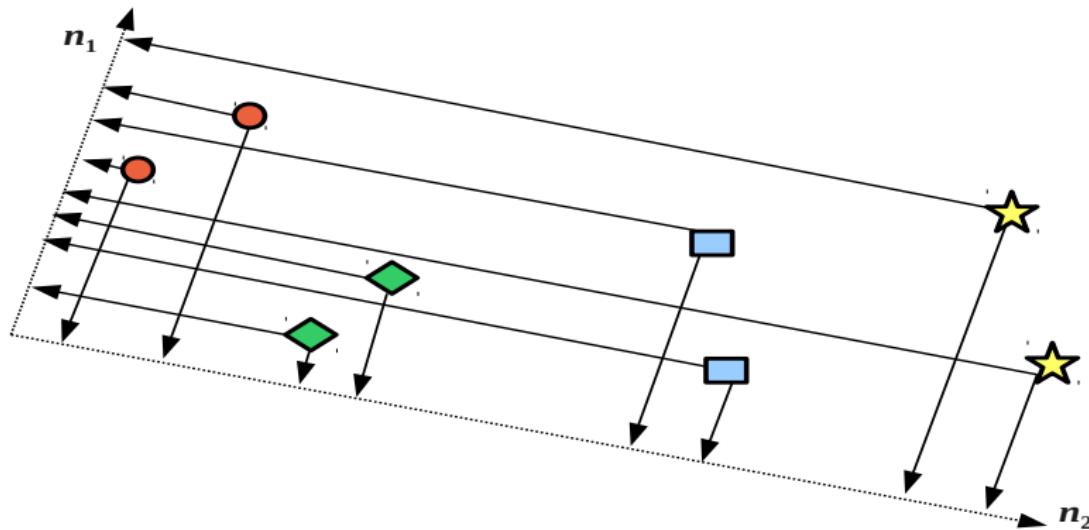
- ▶ Vanilla LSH: Each hyperplane gives 1 bit to the bit encoding for a data point

Preserving cosine similarity in the binary representation

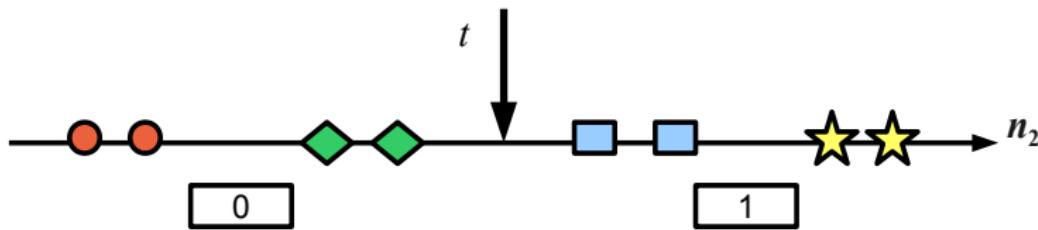
- Two vectors' bits match with probability proportional to the cosine of the angle between them:



Step 1: Low Dimensional Projection



Step 2: Single Bit Quantisation (SBQ)



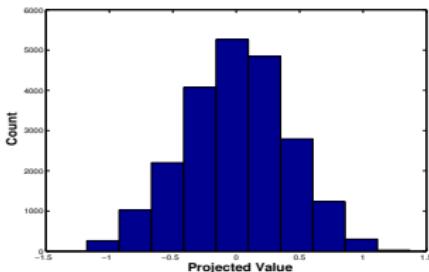
Plus many more...

- ▶ Kernel methods [1]
- ▶ Spectral methods [2] [3]
- ▶ Neural networks [4]
- ▶ Loss based methods [5]
- ▶ **All use single bit quantisation (SBQ)...**

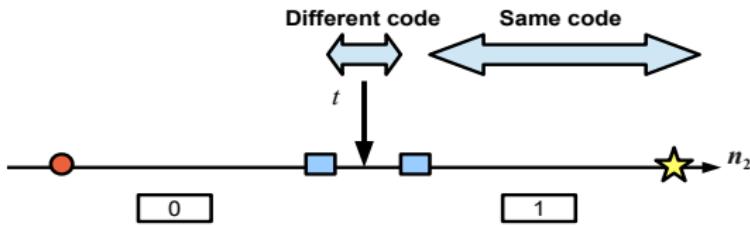
- [1] M. Raginsky and S. Lazebnik. *Locality-sensitive binary codes from shift-invariant kernels*. In NIPS '09.
- [2] Y. Weiss and A. Torralba and R. Fergus. *Spectral Hashing*. NIPS '08.
- [3] J. Wang and S. Kumar and SF. Chang. *Semi-supervised hashing for large-scale search*. PAMI '12.
- [4] R. Salakhutdinov and G. Hinton. *Semantic Hashing*. NIPS '08.
- [5] B. Kulis and T. Darrell. *Learning to Hash with Binary Reconstructive Embeddings*. NIPS '09.

Problem 1: SBQ leads to high quantisation errors

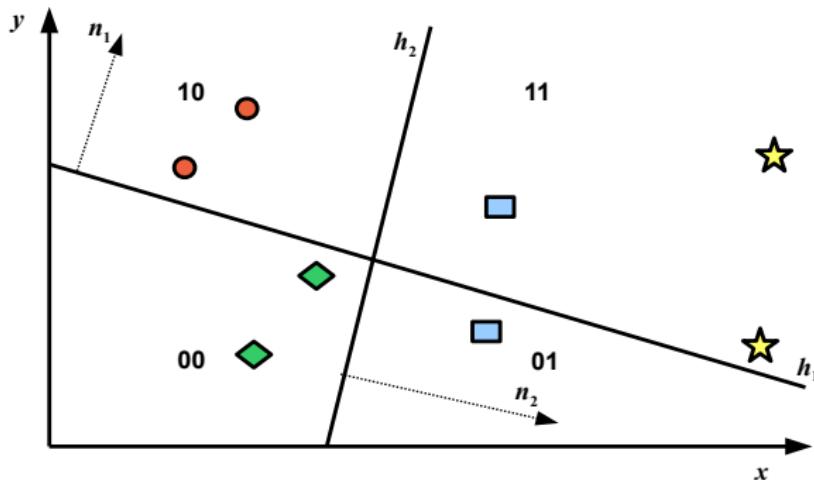
- Highest point density typically occurs around zero:



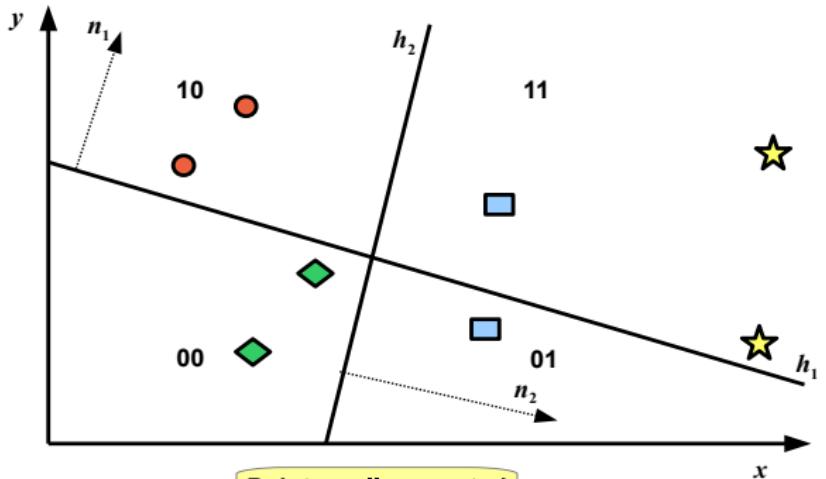
- Closer points can have a greater Hamming distance than distant points:



Problem 2: some hyperplanes are better than others



Problem 2: some hyperplanes are better than others

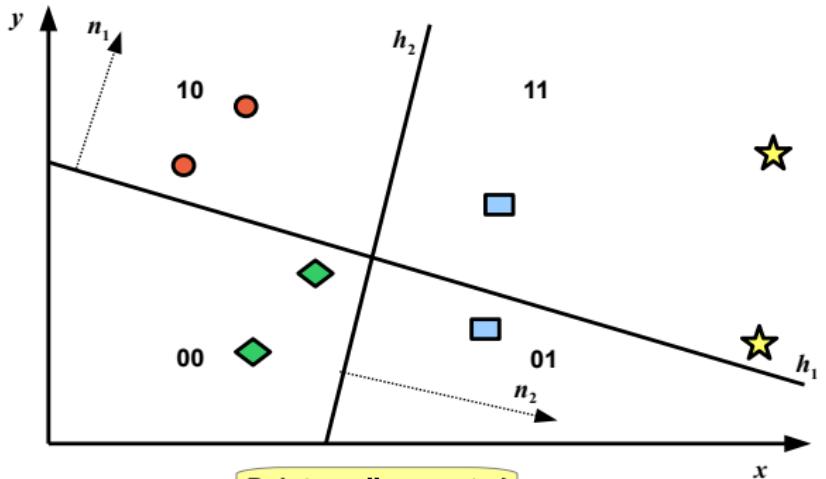


Projected Dimension 2 n_2

Points very mixed

Projected Dimension 1 n_1

Problem 2: some hyperplanes are better than others



Variable Bit Quantisation for LSH

Fast search in large-scale datasets

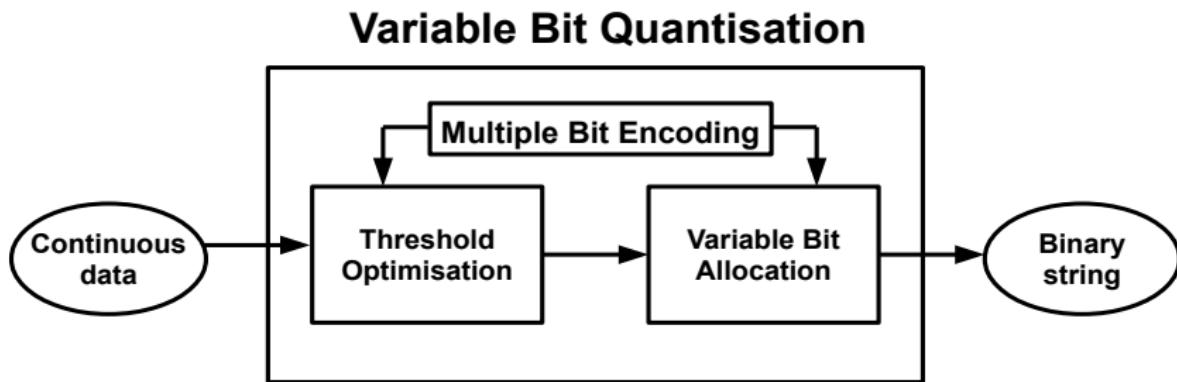
Locality Sensitive Hashing (LSH)

Variable Bit Quantisation (VBQ)

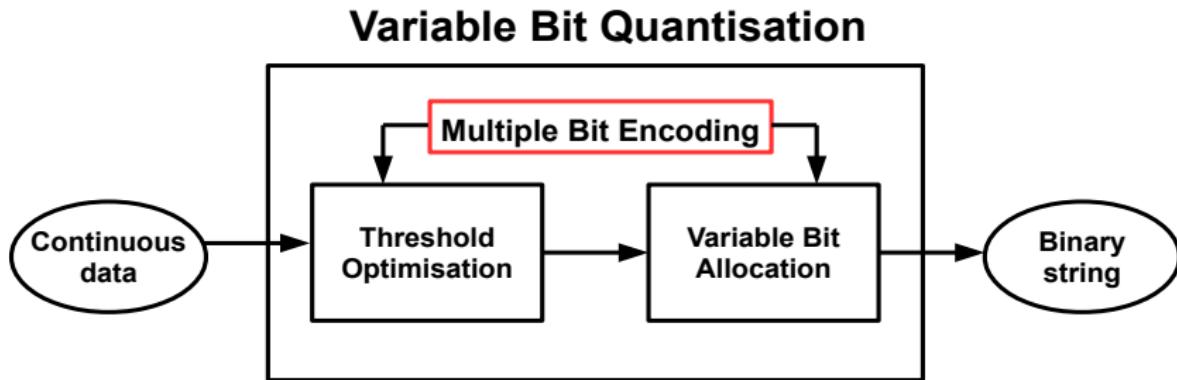
Evaluation

Summary

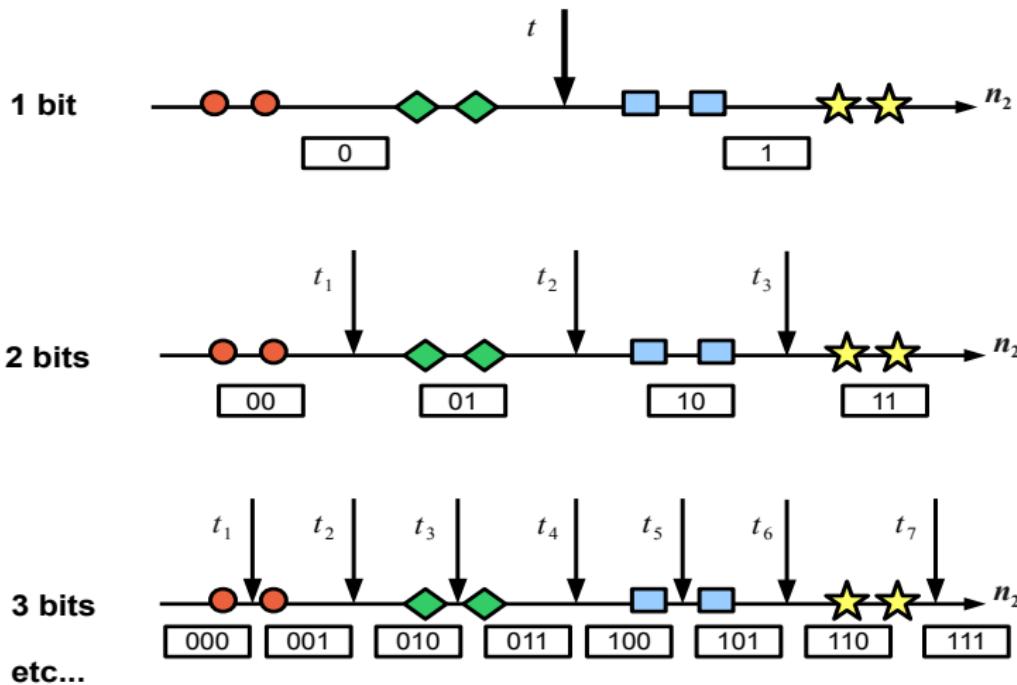
Our Solution: Variable Bit Quantisation



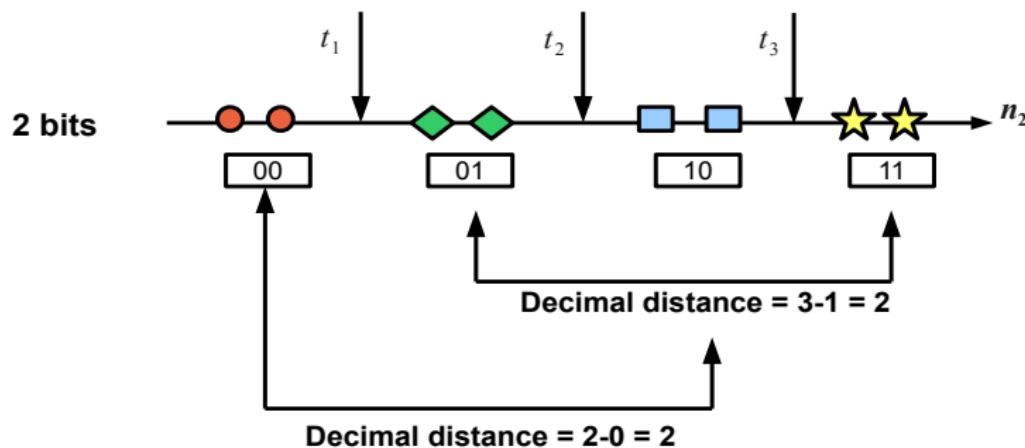
Our Solution: Variable Bit Quantisation



Multiple Bit Encoding: Natural Binary Code

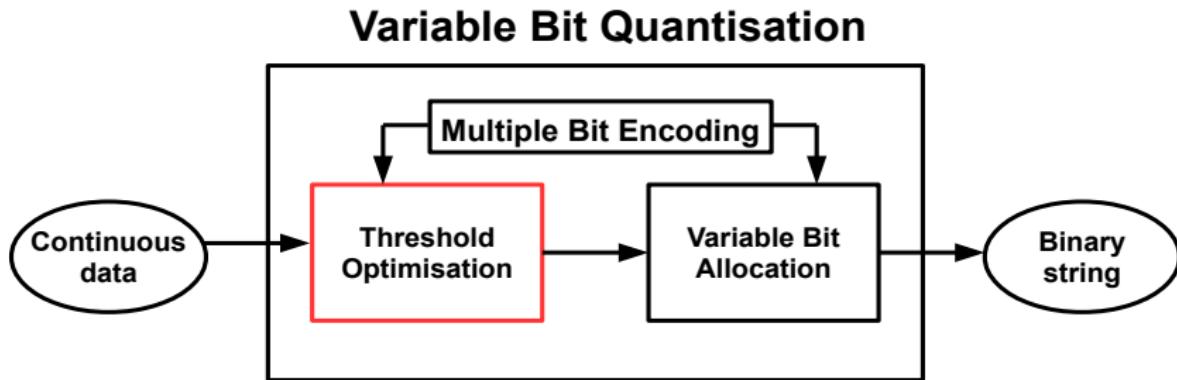


Multiple Bit Encoding [1]



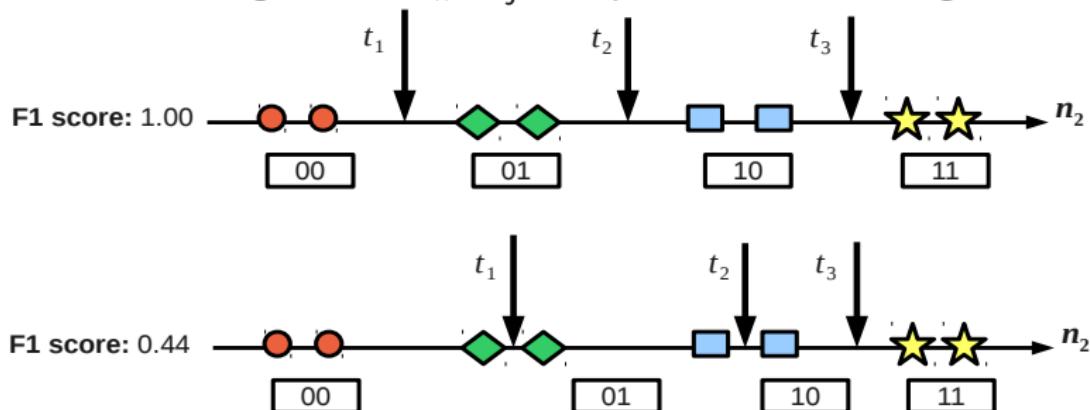
[1] W. Kong and W. Li and M. Guo. *Manhattan hashing for large-scale image retrieval*. SIGIR '12.

Our Solution: Variable Bit Quantisation

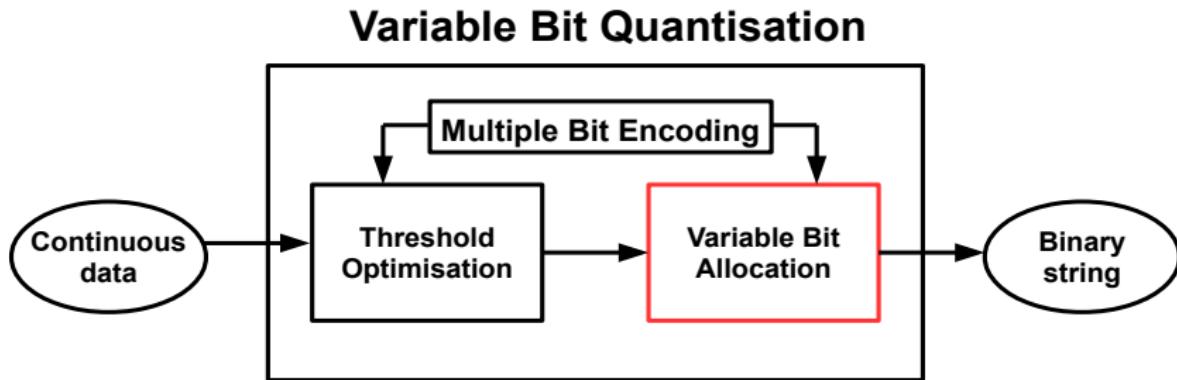


Threshold Optimisation

- ▶ Multiple bits per hyperplane requires multiple thresholds.
- ▶ F_β -based optimisation using pairwise constraints matrix S :
TP: # $S_{ij} = 1$ pairs in the same region. FP: # $S_{ij} = 0$ pairs in the same region. FN: # $S_{ij} = 1$ pairs in different regions.

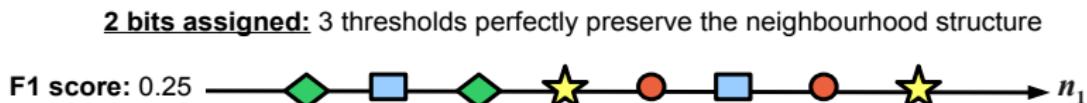
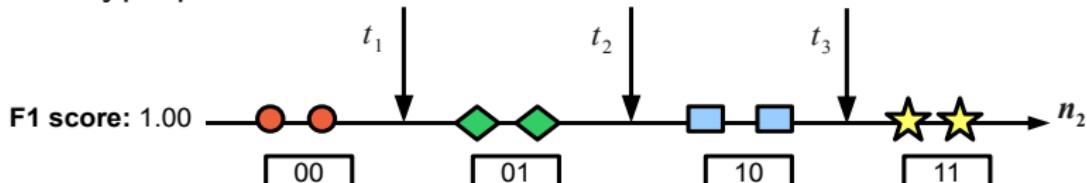


Our Solution: Variable Bit Quantisation



Variable Bit Allocation

- ▶ F_β score is a measure of the *neighbourhood preserving* quality of a hyperplane:



- ▶ Compute bit allocation that maximises the *cumulative F_1* score across all hyperplanes subject to a bit budget B .
- ▶ Bit allocation solved as a binary integer linear program (BILP).

Variable Bit Allocation

$$\begin{aligned} & \max \quad \|\mathbf{F} \circ \mathbf{Z}\| \\ \text{subject to } & \|\mathbf{Z}_h\| = 1 \quad h \in \{1 \dots B\} \\ & \|\mathbf{Z} \circ \mathbf{D}\| \leq B \\ & \mathbf{Z} \text{ is binary} \end{aligned}$$

- ▶ \mathbf{F} contains the F_β scores per hyperplane, per bit count
- ▶ \mathbf{Z} is an indicator matrix specifying the bit allocation
- ▶ \mathbf{D} is a constraint matrix
- ▶ B is the bit budget
- ▶ $\|\cdot\|$ denotes the Frobenius L_1 norm
- ▶ \circ the Hadamard product

Variable Bit Allocation

$$\begin{aligned}
 & \max \quad \| \mathbf{F} \circ \mathbf{Z} \| \\
 \text{subject to } & \|\mathbf{Z}_h\| = 1 \quad h \in \{1 \dots B\} \\
 & \|\mathbf{Z} \circ \mathbf{D}\| \leq B \\
 & \mathbf{Z} \text{ is binary}
 \end{aligned}$$

$$\begin{array}{ccccc}
 \mathbf{F} & h_1 & h_2 & \mathbf{D} & \mathbf{Z} \\
 b_0 & \begin{pmatrix} 0.25 & 0.25 \end{pmatrix} & & \begin{pmatrix} 0 & 0 \end{pmatrix} & \begin{pmatrix} 1 & 0 \end{pmatrix} \\
 b_1 & \begin{pmatrix} 0.35 & 0.50 \end{pmatrix} & & \begin{pmatrix} 1 & 1 \end{pmatrix} & \begin{pmatrix} 0 & 0 \end{pmatrix} \\
 b_2 & \begin{pmatrix} 0.40 & 1.00 \end{pmatrix} & & \begin{pmatrix} 2 & 2 \end{pmatrix} & \begin{pmatrix} 0 & 1 \end{pmatrix}
 \end{array}$$

- ▶ Sparse solution possible: lower quality hyperplanes can be discarded.

Variable Bit Quantisation for LSH

Fast search in large-scale datasets

Locality Sensitive Hashing (LSH)

Variable Bit Quantisation (VBQ)

Evaluation

Summary

Evaluation Objectives

1. Quantify performance of the threshold optimisation procedure:
 - ▶ How does assigning a constant 2 bits per hyperplane based upon our adaptive thresholding scheme affect retrieval performance?

2. Quantify performance of the variable bit assignment routine:
 - ▶ How does assigning more bits to higher quality hyperplanes affect retrieval performance?

Evaluation Protocol

- ▶ **Task:** Text and image retrieval on three standard datasets: *CIFAR-10*, *100k TinyImages* and *Reuters-21578*.
- ▶ **Projections:** LSH [1], Shift-invariant kernel hashing (SIKH) [2], Spectral Hashing (SH) [3] and PCA-Hashing (PCAH) [4].
- ▶ **Baselines:** Single Bit Quantisation (SBQ), Manhattan Hashing (MQ)[5], Double-Bit quantisation (DBQ) [6].
- ▶ **Hamming Ranking:** how well do we retrieve the ϵ -NN of query points?

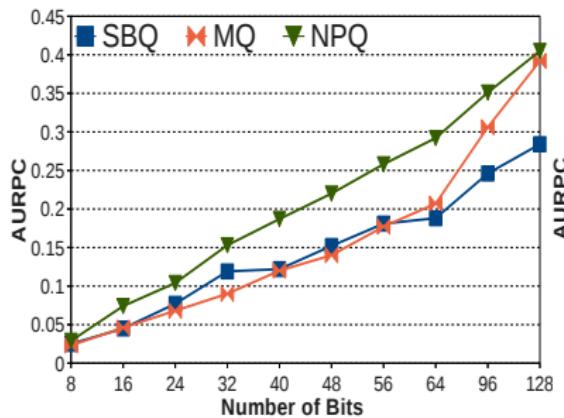
- [1] P. Indyk and R. Motwani. *Approximate nearest neighbors: removing the curse of dimensionality*. In STOC '98.
- [2] M. Raginsky and S. Lazebnik. *Locality-sensitive binary codes from shift-invariant kernels*. In NIPS '09.
- [3] Y. Weiss and A. Torralba and R. Fergus. *Spectral Hashing*. NIPS '08.
- [4] J. Wang and S. Kumar and SF. Chang. *Semi-supervised hashing for large-scale search*. PAMI '12.
- [5] W. Kong and W. Li and M. Guo. *Manhattan hashing for large-scale image retrieval*. SIGIR '12.
- [6] W. Kong and W. Li. *Double Bit Quantisation for Hashing*. AAAI '12.

AUPRC across different projections (constant # bits)

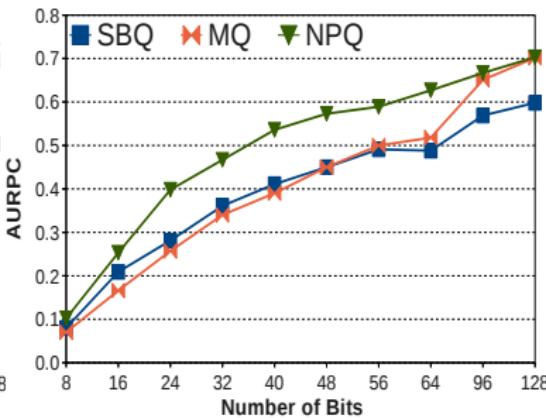
Dataset	CIFAR-10 (32 bits)				Tiny Images (32 bits)			
	SBQ	MQ	DBQ	NPQ	SBQ	MQ	DBQ	NPQ
SIKH	0.042	0.063	0.047	0.090	0.135	0.221	0.182	0.365
LSH	0.119	0.093	0.066	0.153	0.361	0.340	0.285	0.464
SH	0.051	0.135	0.111	0.167	0.117	0.237	0.136	0.356
PCAH	0.036	0.137	0.107	0.153	0.046	0.257	0.295	0.312

- ▶ Assigning multiple bits via our adaptive thresholding scheme yields substantial gains.
- ▶ NPQ and a cheap projection (e.g. LSH) can outperform SBQ and an expensive projection (e.g. PCA).

AUPRC for LSH across a broad bit range



(a) CIFAR-10



(b) 100k TinyImages

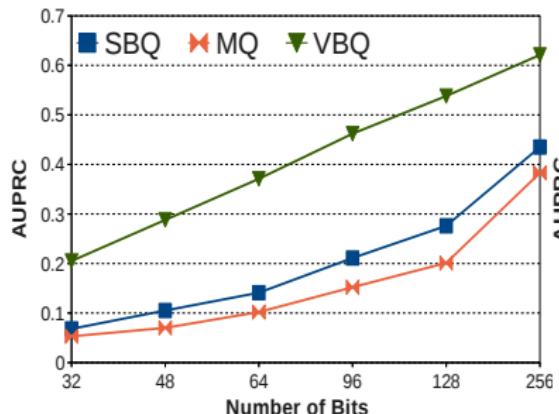
- ▶ NPQ is effective across a wide range of bits.

AUPRC across different projections (variable # bits)

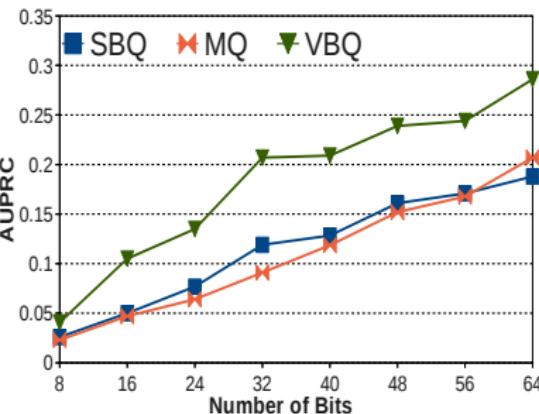
Dataset	CIFAR-10 (32 bits)				Reuters-21578 (128 bits)			
	SBQ	MQ	DBQ	VBQ	SBQ	MQ	DBQ	VBQ
SIKH	0.042	0.046	0.047	0.161	0.102	0.112	0.087	0.389
LSH	0.119	0.091	0.066	0.207	0.276	0.201	0.175	0.538
SH	0.051	0.144	0.111	0.202	0.033	0.028	0.030	0.154
PCAH	0.036	0.132	0.107	0.219	0.095	0.034	0.027	0.154

- ▶ Varying the number of bits based on hyperplane quality gives additional gains in retrieval accuracy.
- ▶ VBQ is an effective quantisation scheme for both image and text datasets.

AUPRC for LSH across a broad bit range



(a) Reuters-21578



(b) CIFAR-10

- VBQ is also effective across a wide range of bits.

Variable Bit Quantisation for LSH

Fast search in large-scale datasets

Locality Sensitive Hashing (LSH)

Variable Bit Quantisation (VBQ)

Evaluation

Summary

Summary

- ▶ Proposed a general data-driven technique to adaptively assign variable bits per LSH hyperplane
- ▶ Hyperplanes better preserving the neighbourhood structure are afforded more bits from the bit budget
- ▶ VBQ substantially increased retrieval performance across standard text and image datasets
- ▶ Future work: evaluate with an LSH system that uses hash tables for fast retrieval

Publications

- ▶ **Neighbourhood Preserving Quantisation for LSH.** Moran, S., Lavrenko, V., Osborne, M. (2013), *Special Interest Group on Information Retrieval (SIGIR)*, Dublin, Ireland.

- ▶ **Variable Bit Quantisation for LSH.** Moran, S., Lavrenko, V., Osborne, M. (2013), *Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Thank you for your attention

**Sean Moran
University of Edinburgh**

sean.moran@ed.ac.uk
www.seanjmoran.com