Distilling Collective Intelligence from Twitter

Crowdsourcing and Human Computation
Lecture 17
Instructor: Chris Callison-Burch
TA: Ellie Pavlick

Website: crowdsourcing-class.org

Today's slides come courtesy of Miles Osborne and Benjamin Van Durme

Tapping into Collective Intelligence on Twitter

- What can we learn about the real world from Twitter?
- How can we use scalable machine learning algorithms to detect facts quickly?

Twitter

- Poor signal-to-noise ratio
- Multilingual
  - About 65% of Tweets are in English
- > 400 million posts / day
- > 500 million registered users (2012)
- At peak more than 140k tweets per second

Representative examples of non-events

- J’aime pas Bieber, 1D le rap et plein d’autres conneries. Vous pouvez m’amener 500 haters je changerai pas d’avis.
- This wine is going down a lil to smoothly. Here comes trouble.
- LIMA HARI BULAN LIMA ! KEK SEBESAR GUNUNG ! kena belajar buat
- kek ni, tinggal 2 bulan jea lagi -.-’
- RT ZorianRamone: Happy Bday to one of my Closest friends bra I love youe

Event Detection

- Find breaking news as quickly as possible:
  - Earthquake happens on Monday at 9am.
    - Report story as soon after 9am as possible
    - Don’t report follow-up mentions.
- Intensively studied as part of *Topic Detection and Tracking* (DARPA TIDES program, 1997 – 2004)

First Story Detection

- Typical FSD system:
  - Store stories (vectors) as they are seen
  - First Story Detection
    - Need to compare stories with each other using a distance metric
    - For some new story, find the story that is its nearest neighbor
    - If the new story is ‘far away’ from its nearest story, announce it as a new story

Need for Efficient Search

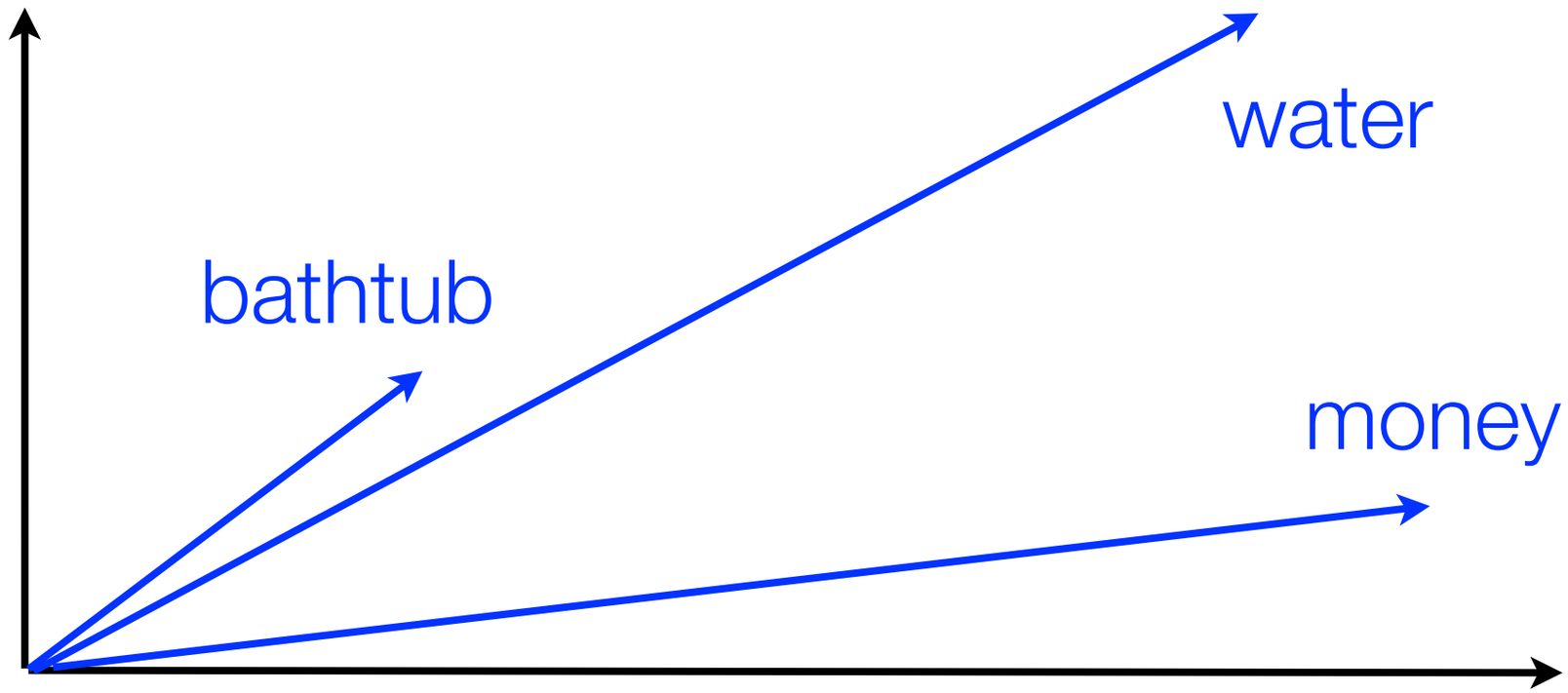
- Nearest Neighbors search implies comparing **all** stories against the current one
- If there are 400 million posts / day, how many comparisons do we need to do for some new post in order to decide whether it represents something new?

Vector Space Models of Word Similarity

- Represent a word through the contexts that it has been observed in

He found five fish swimming in an old bathtub.
He slipped down in the bathtub.

a	1
down	1
find	1
fish	1
five	1
he	2
in	2
slip	1
swim	1
the	1

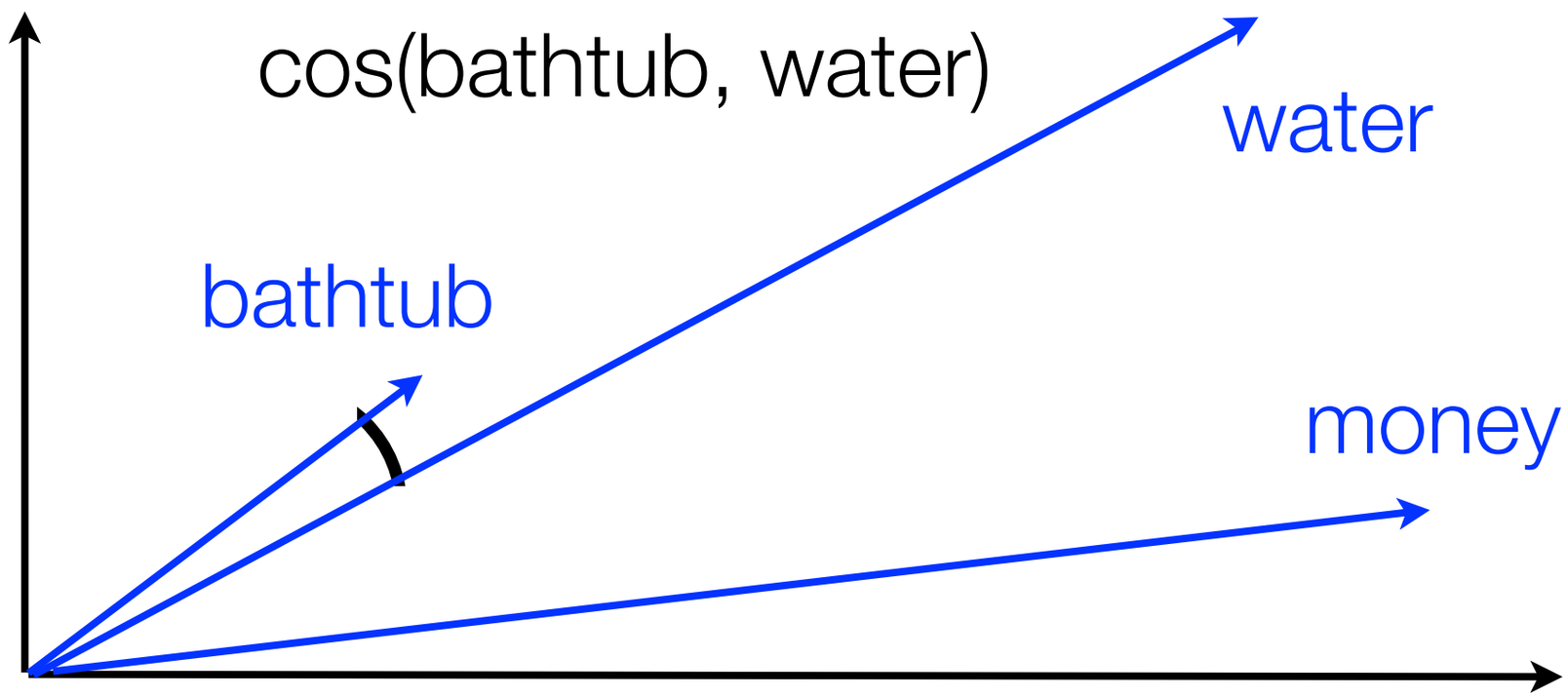


Vector Space Models of Word Similarity

- Represent a word through the contexts that it has been observed in

He found five fish swimming in an old bathtub.
He slipped down in the bathtub.

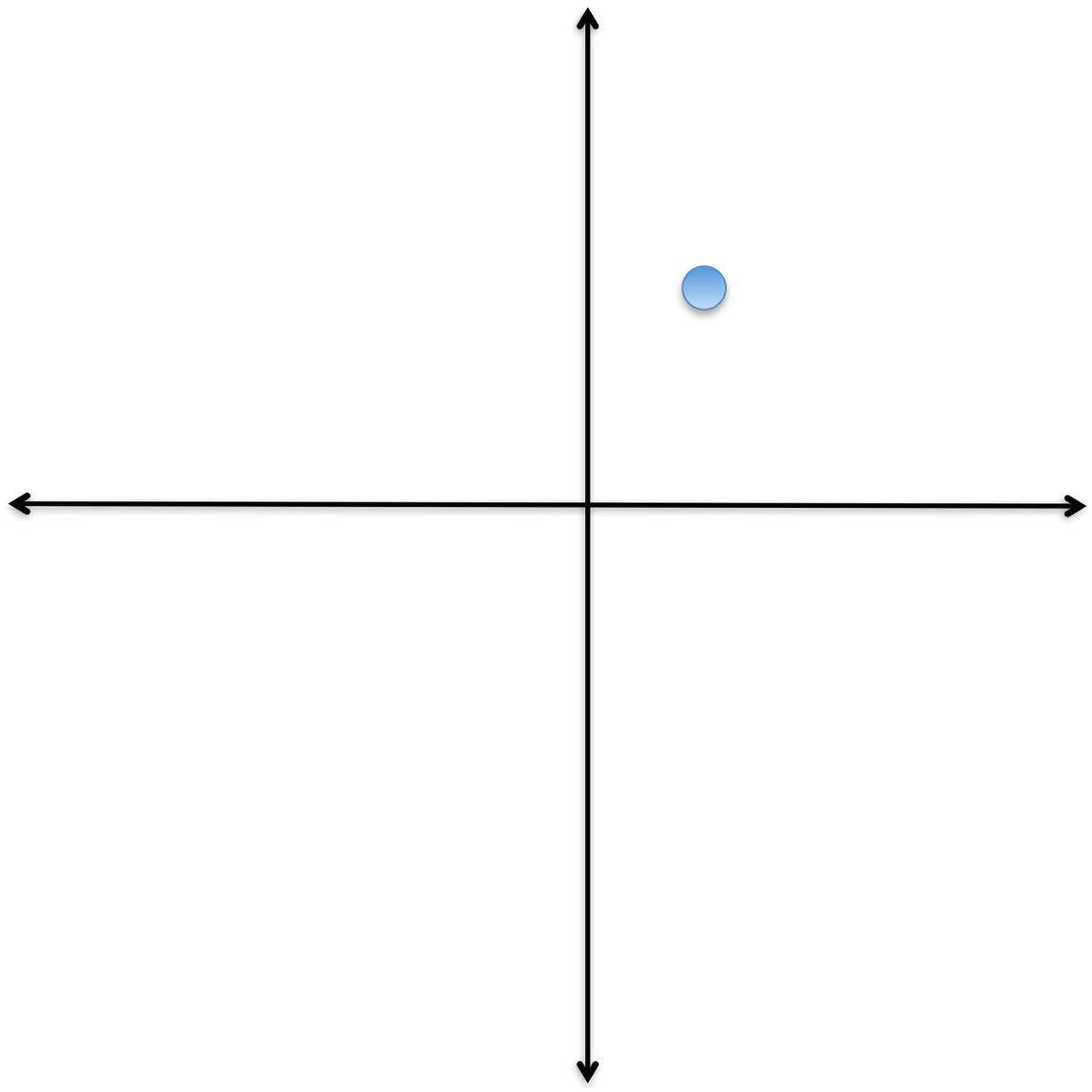
a	1
down	1
find	1
fish	1
five	1
he	2
in	2
slip	1
swim	1
the	1

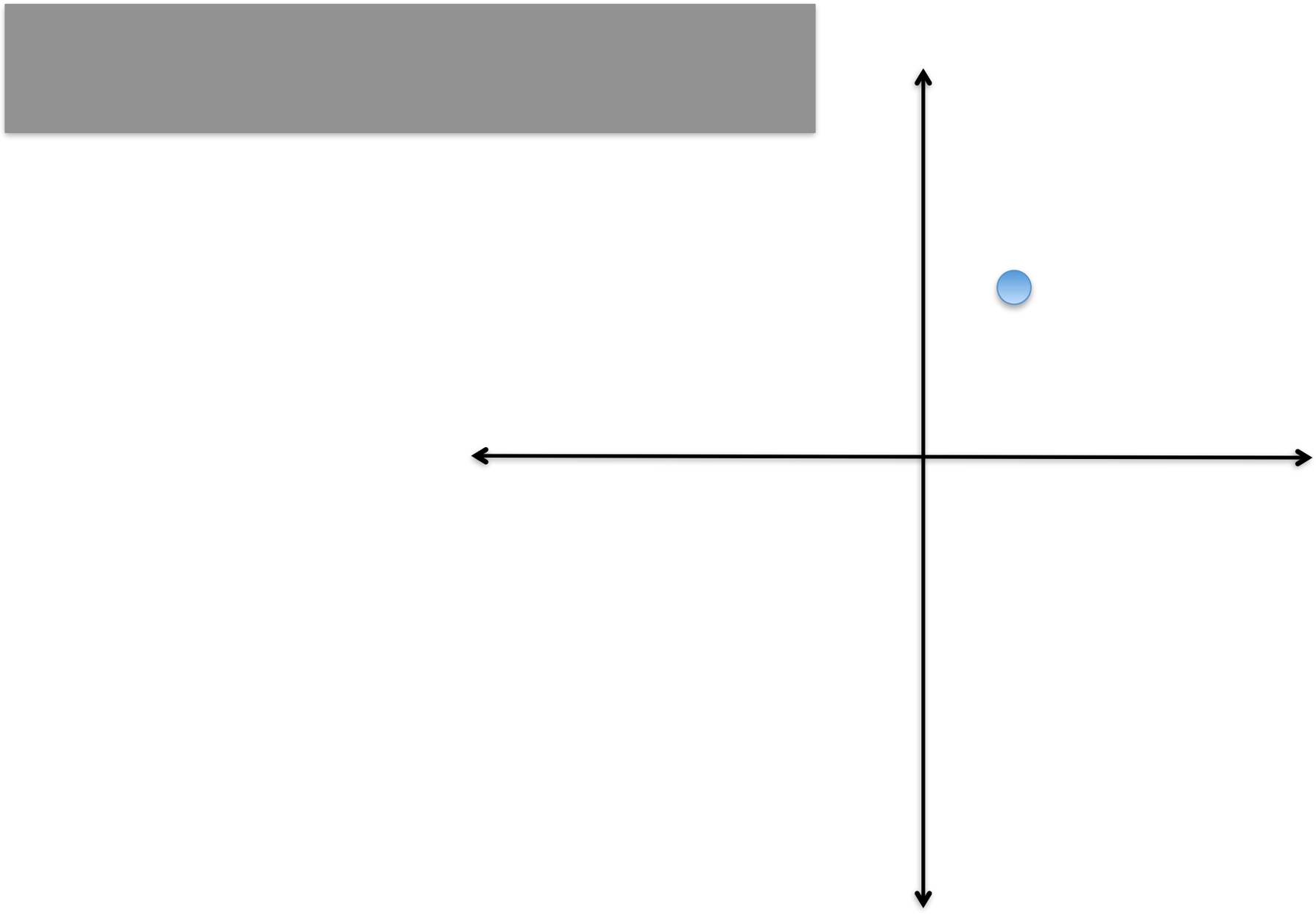


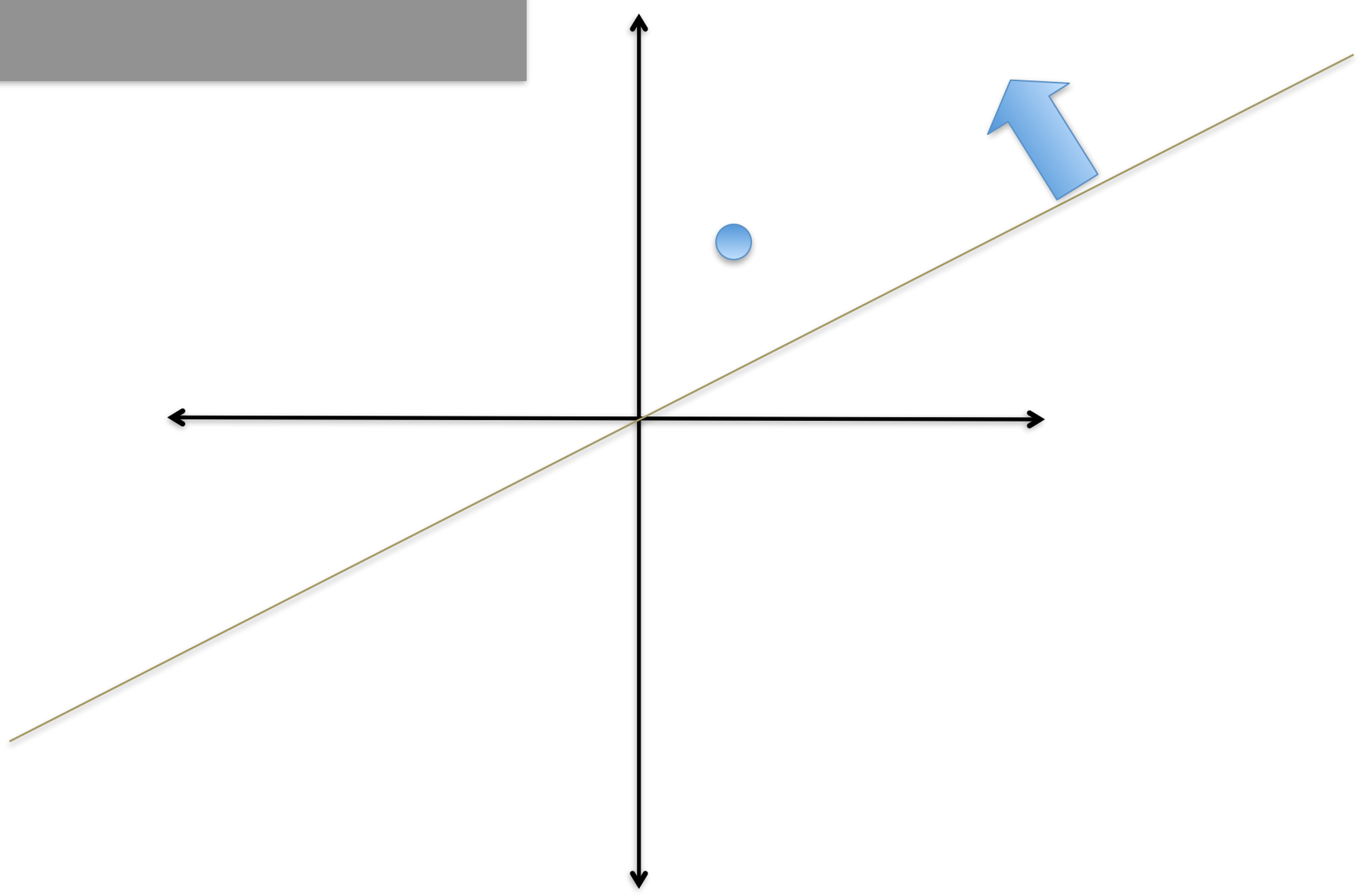
Locality Sensitive Hashing

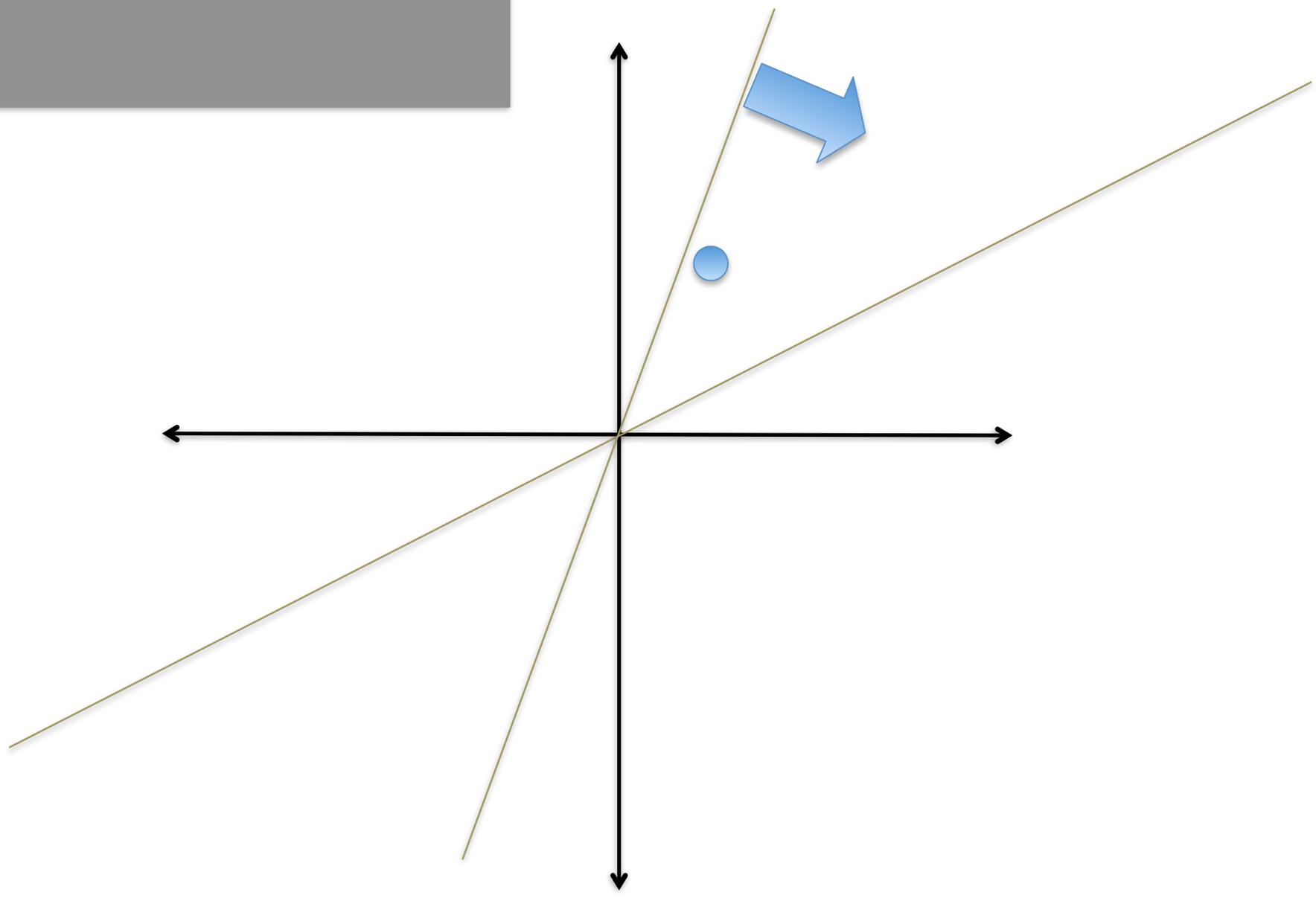
- Goal: fast comparison between points in very high dimensional space

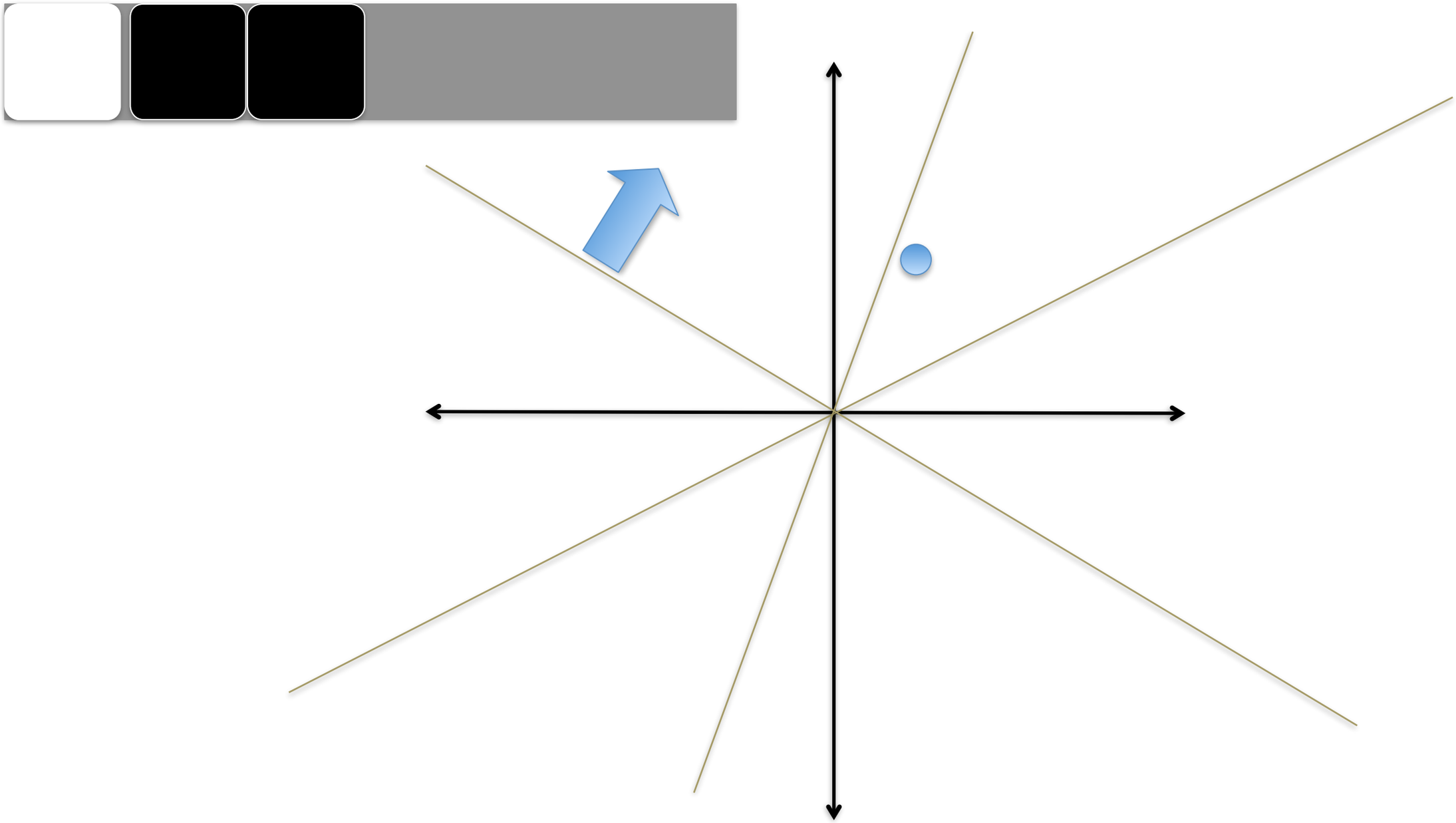
- Randomly project points to low dimensional *bit signatures* such that cosine distance is roughly preserved

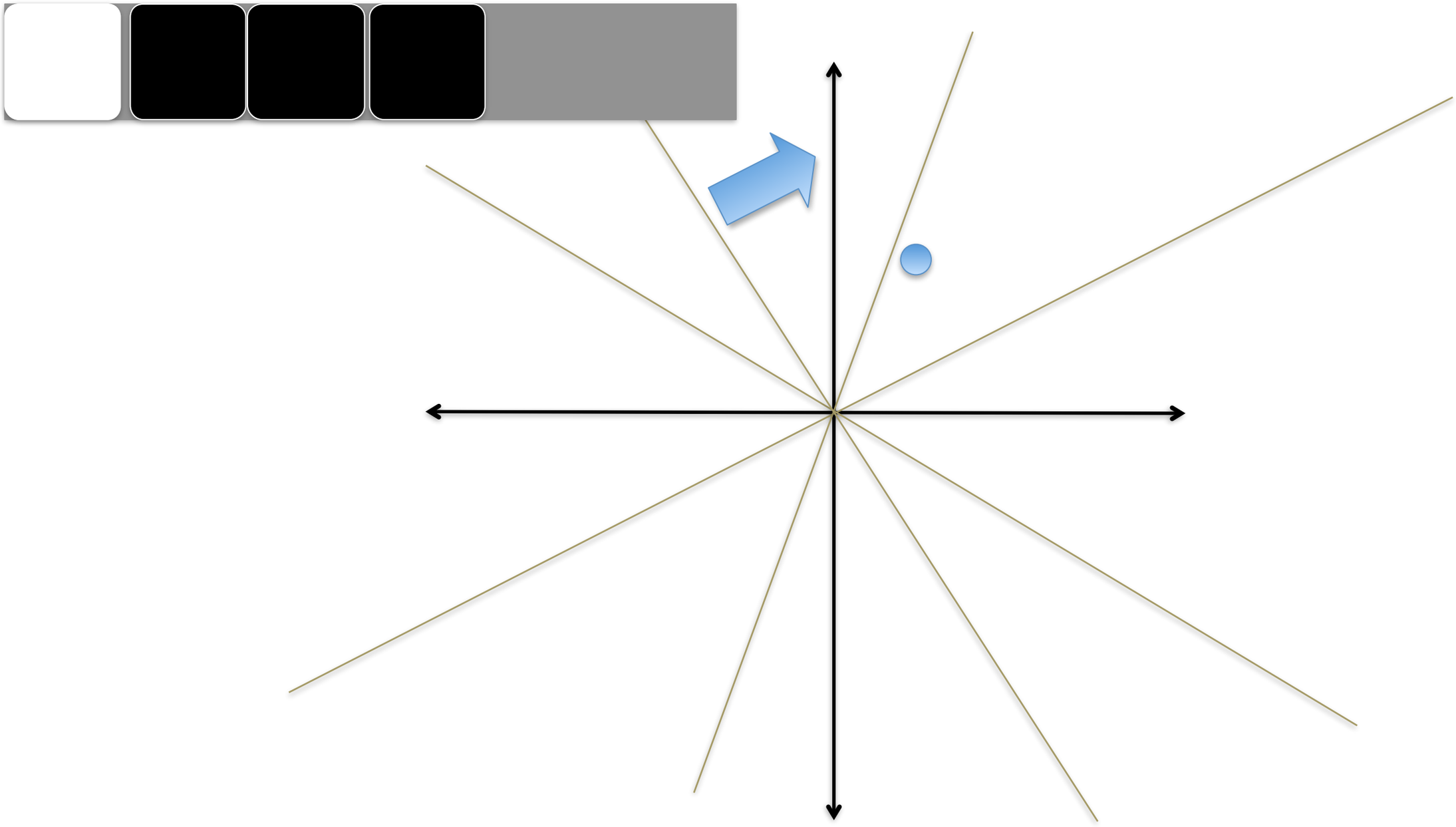


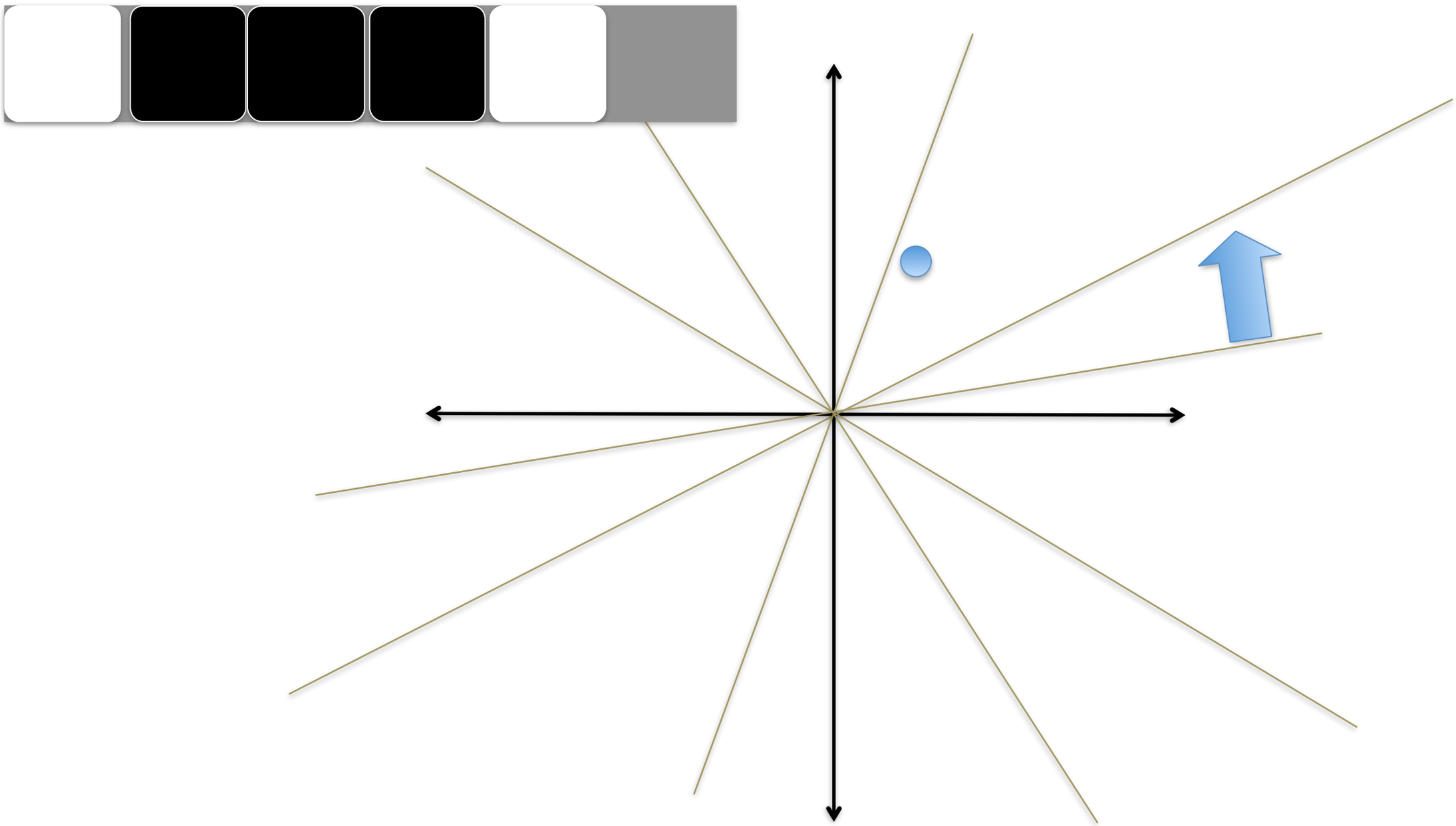


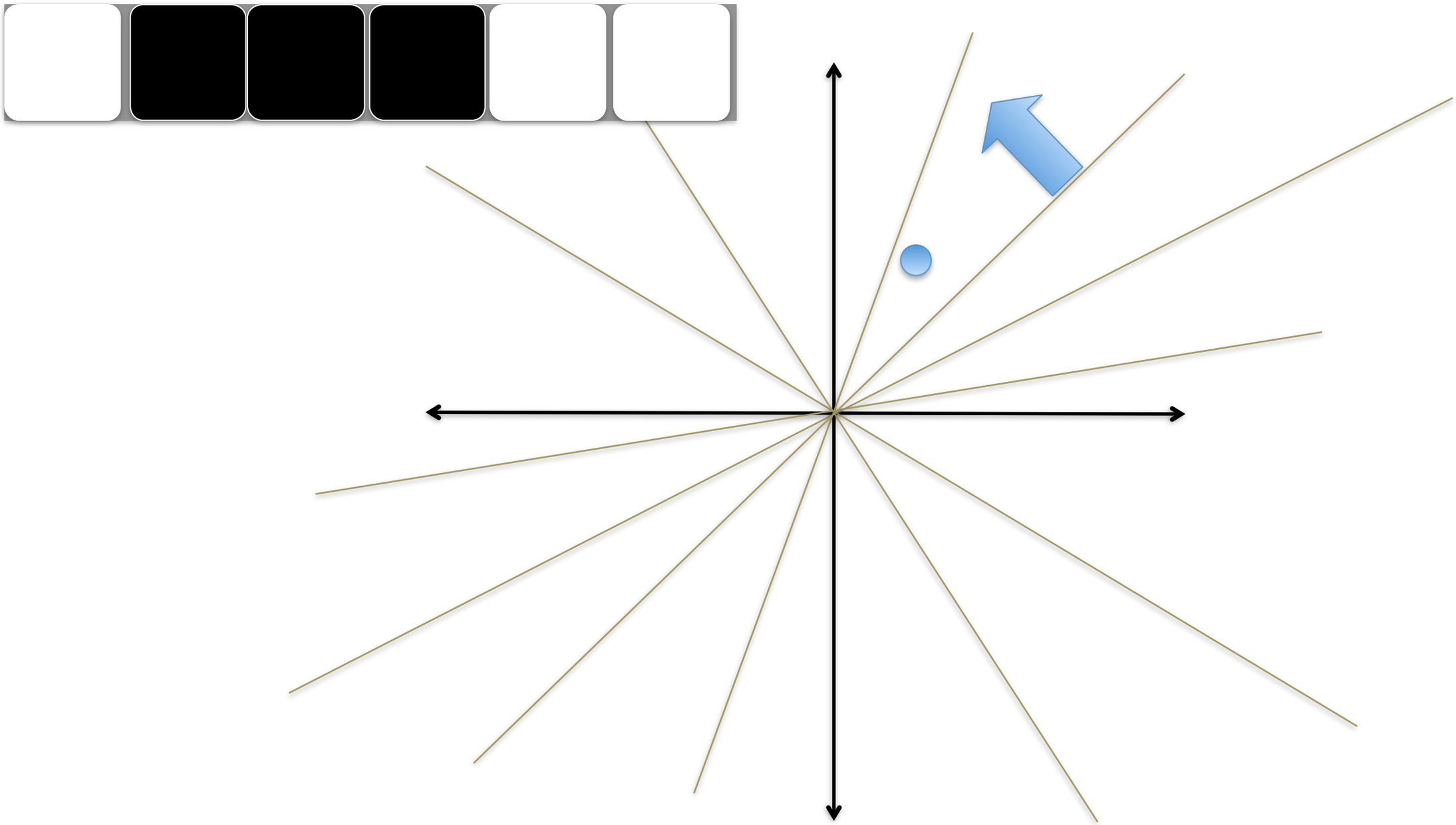


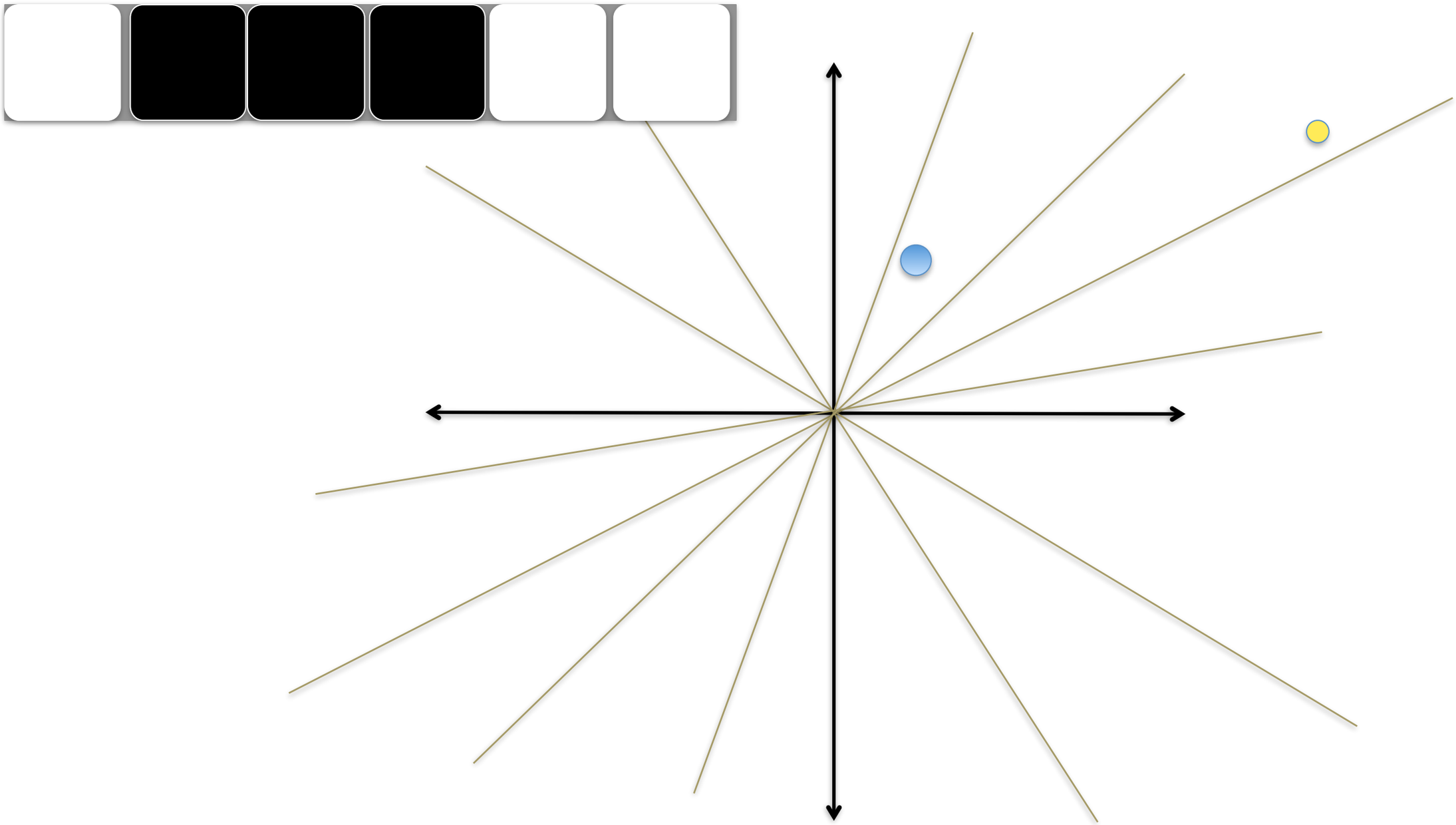


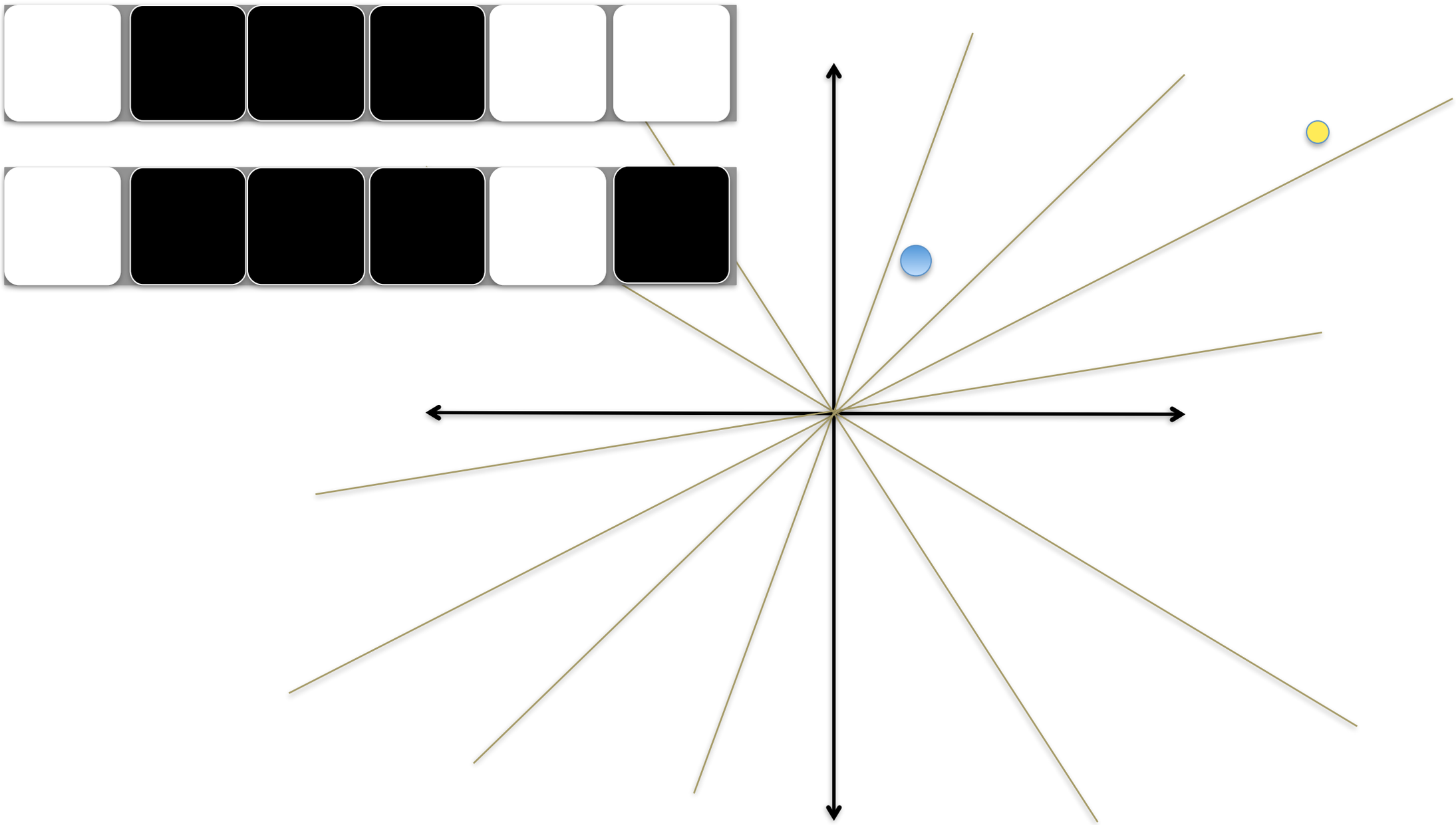


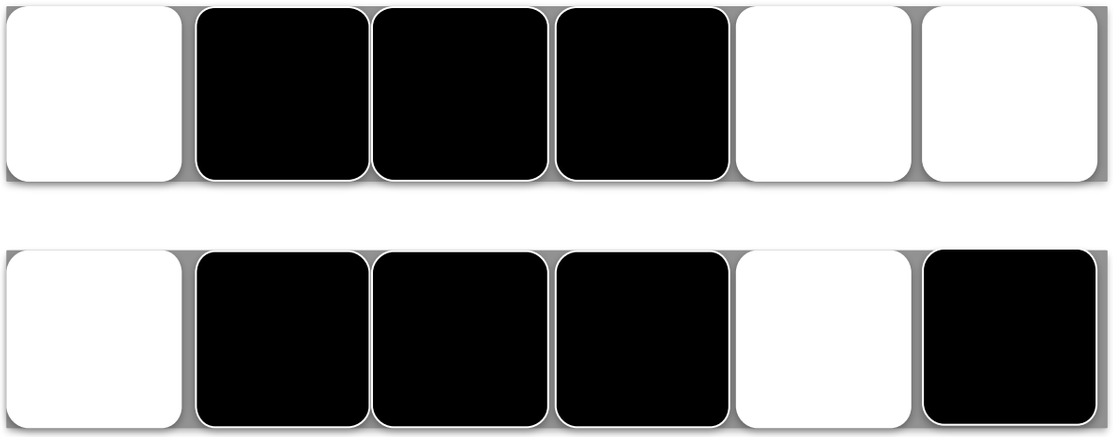


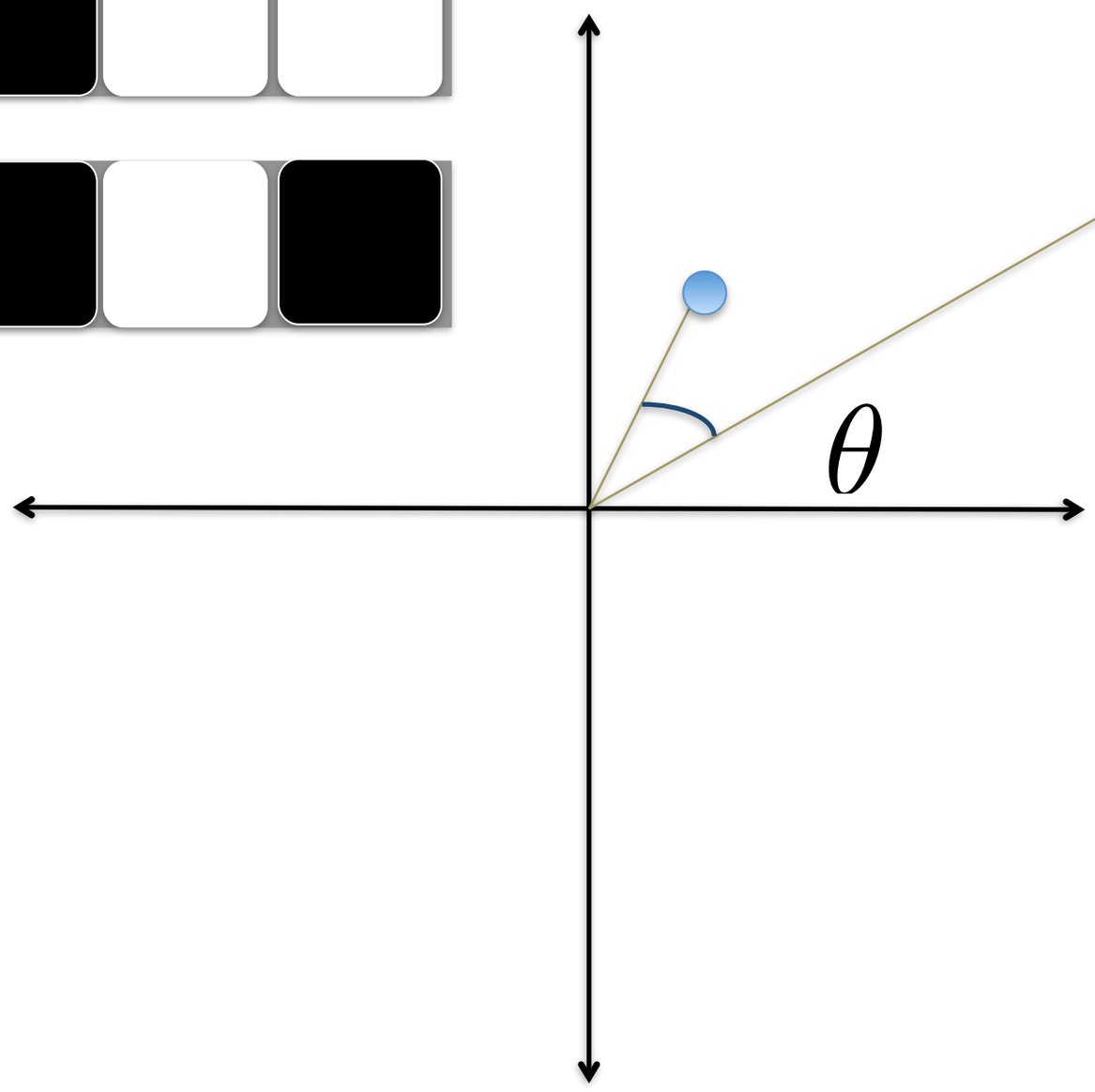


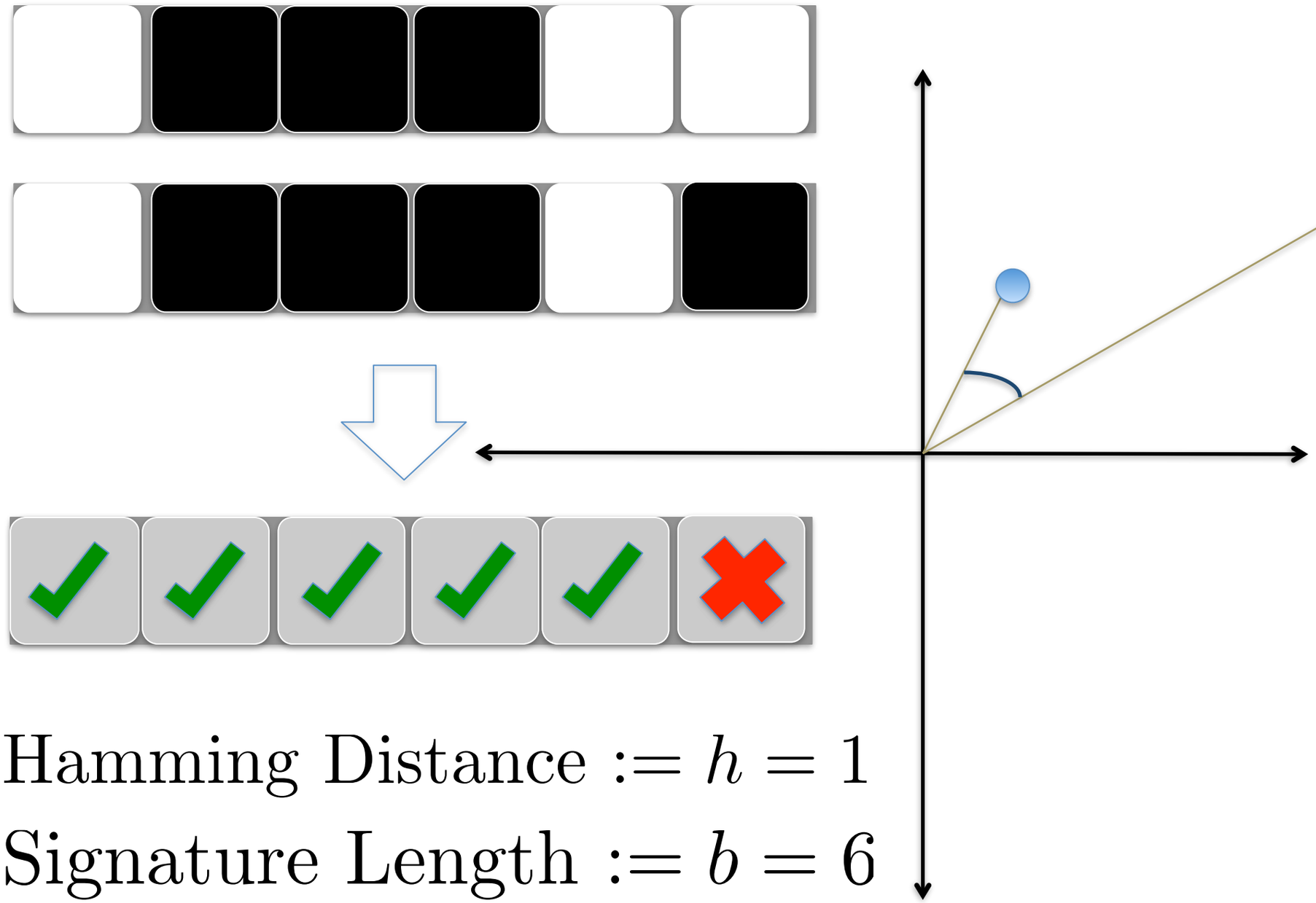






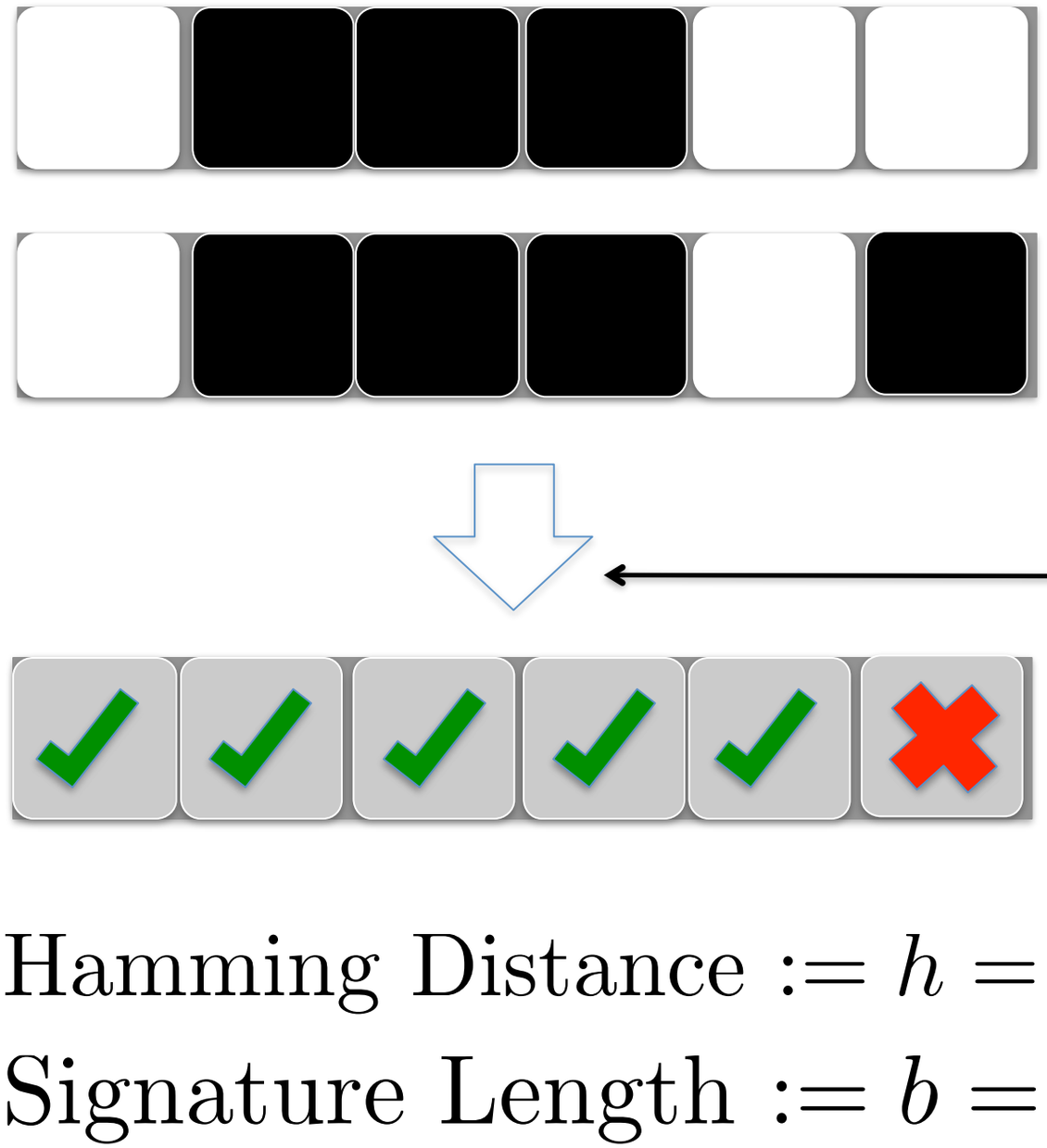



θ



Hamming Distance $:= h = 1$

Signature Length $:= b = 6$



Hamming Distance $:= h = 1$

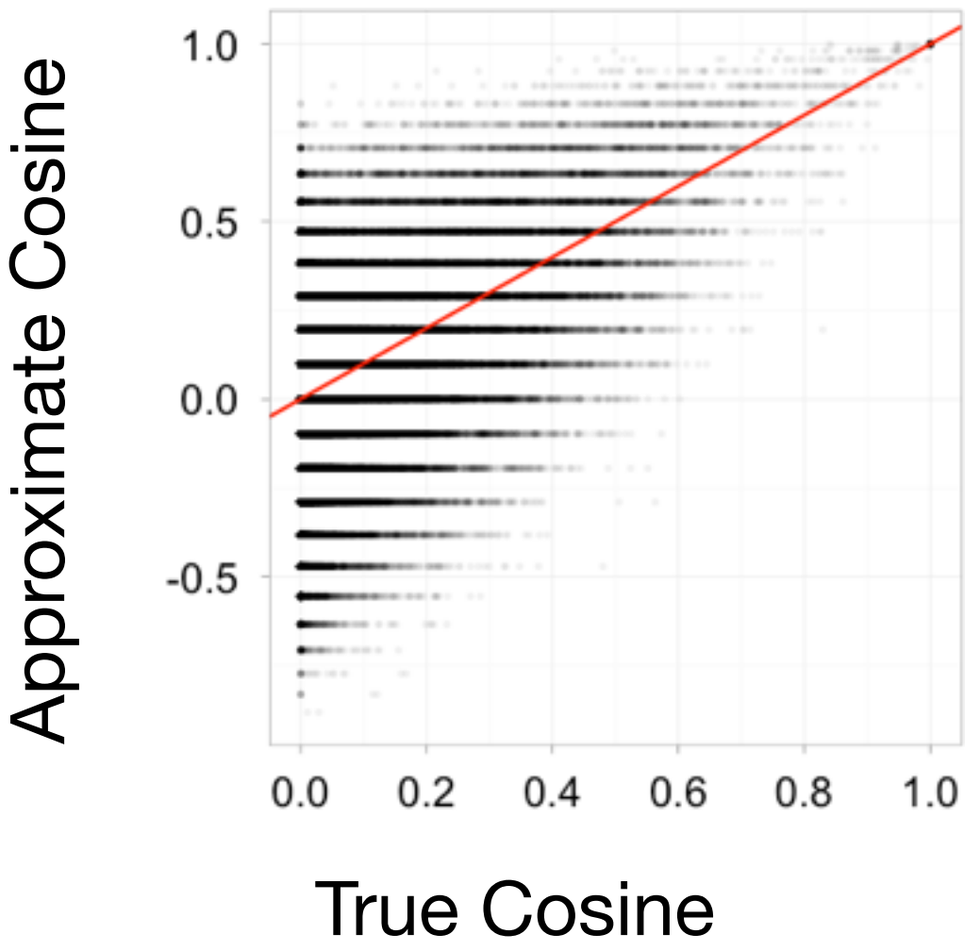
Signature Length $:= b = 6$

$$\cos(\theta) \approx \cos(\tfrac{h}{b}\pi)$$

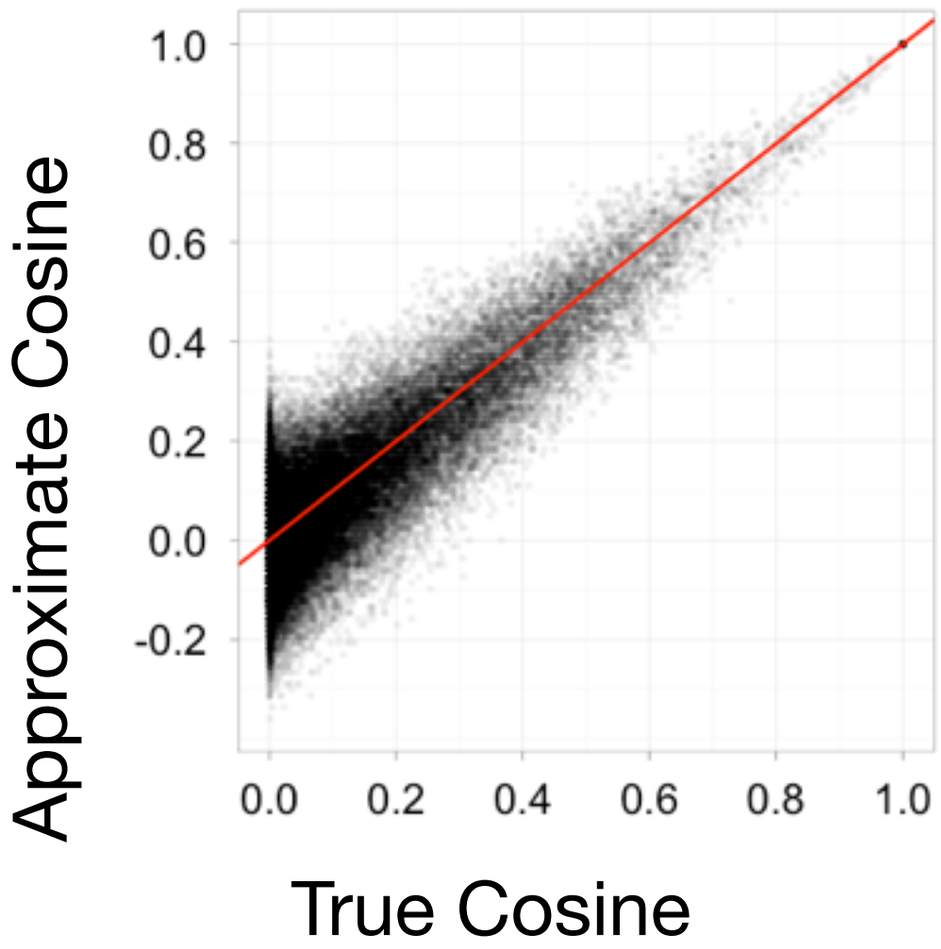
$$= \cos(\tfrac{1}{6}\pi)$$

Accuracy as function of bit length

32 bit signatures

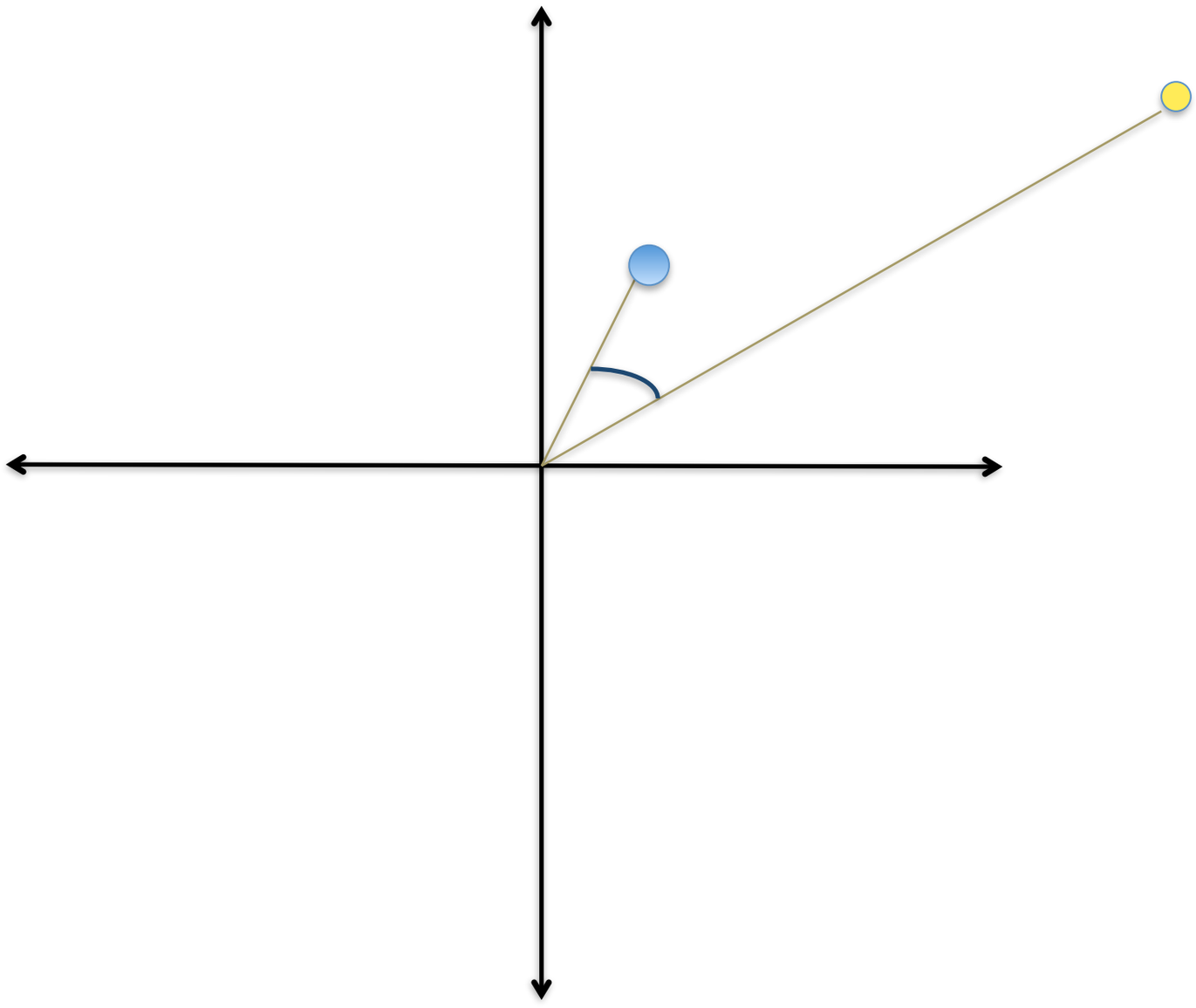


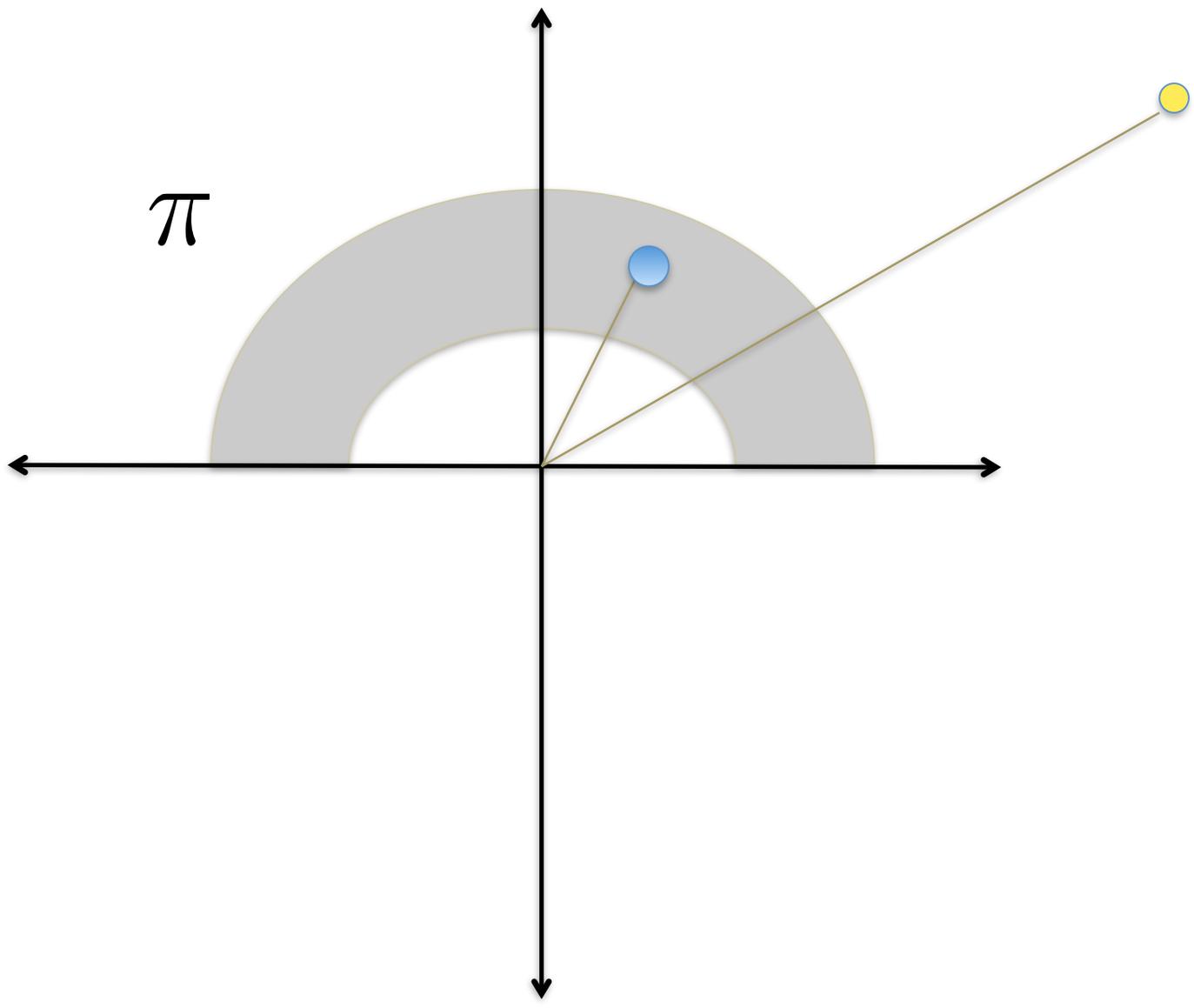
256 bit signatures

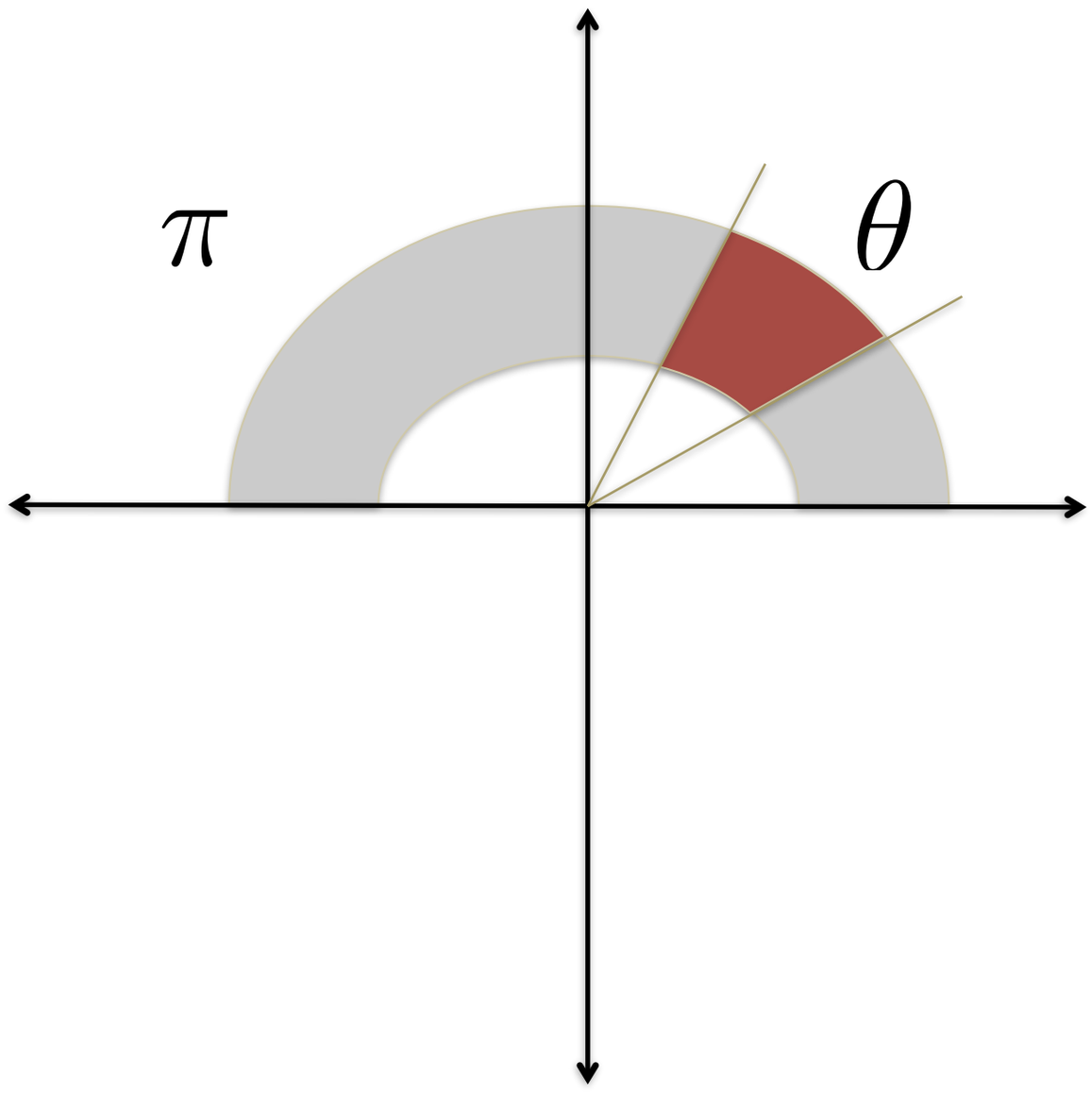


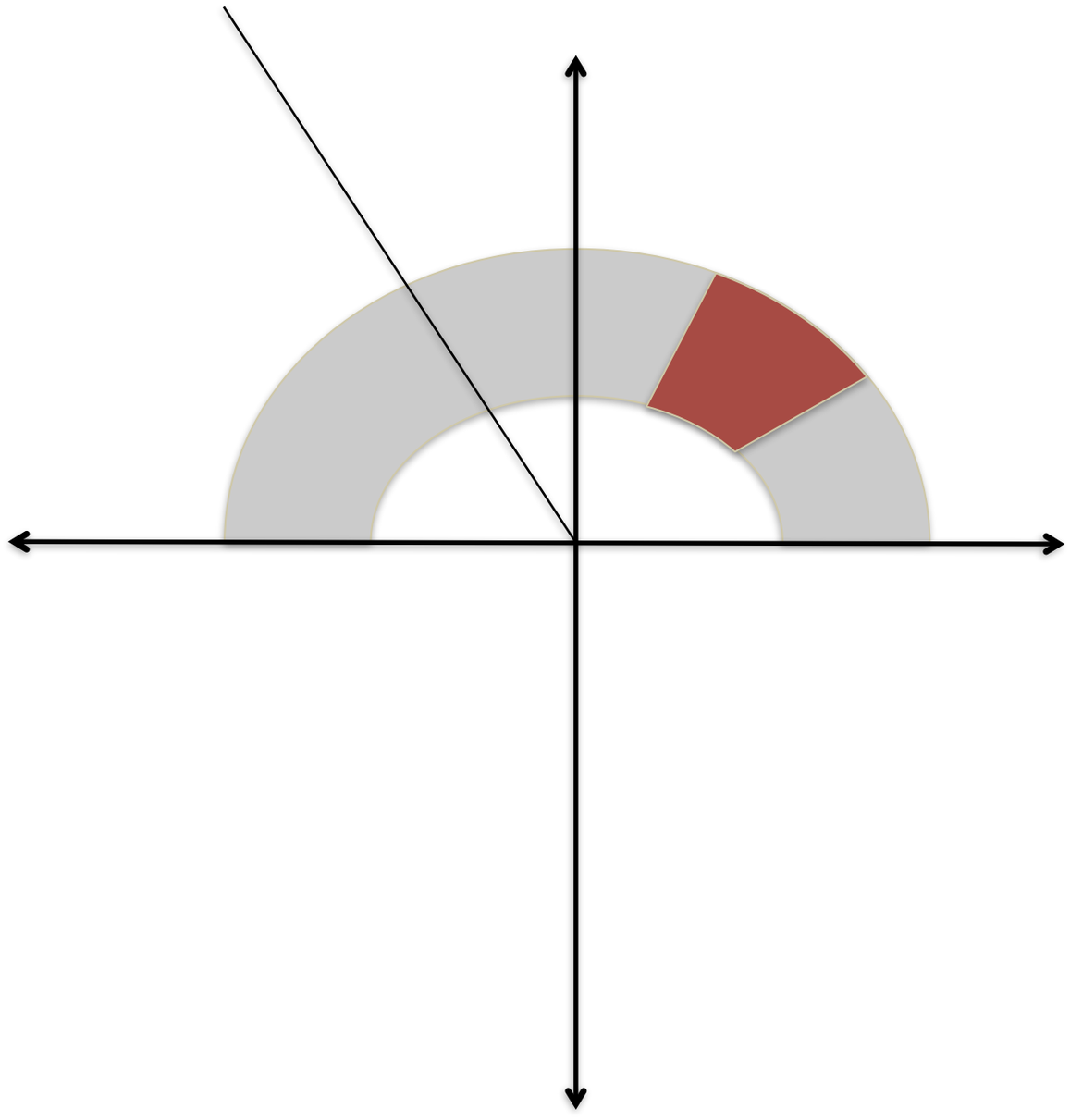
Cheap

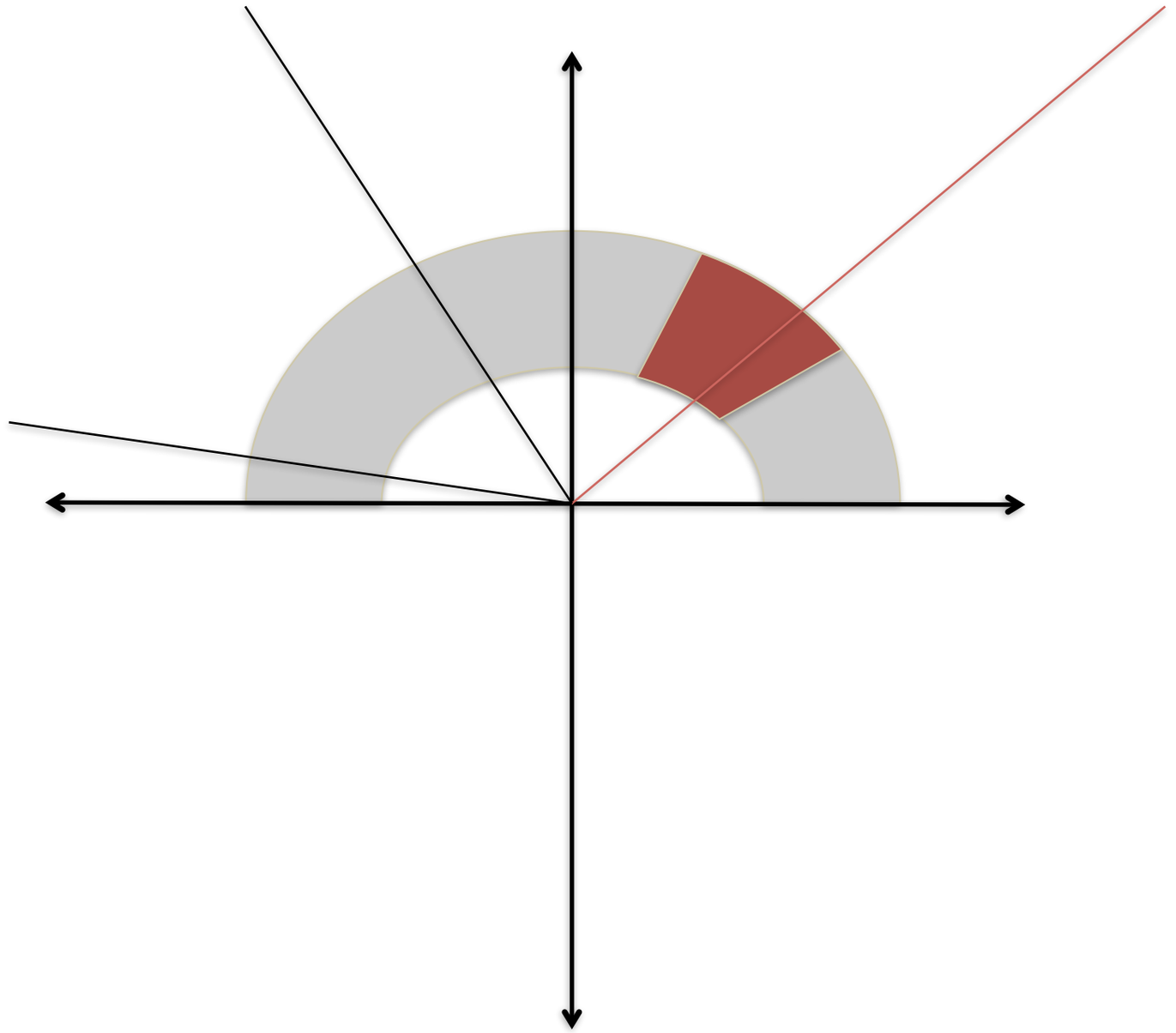
Accurate

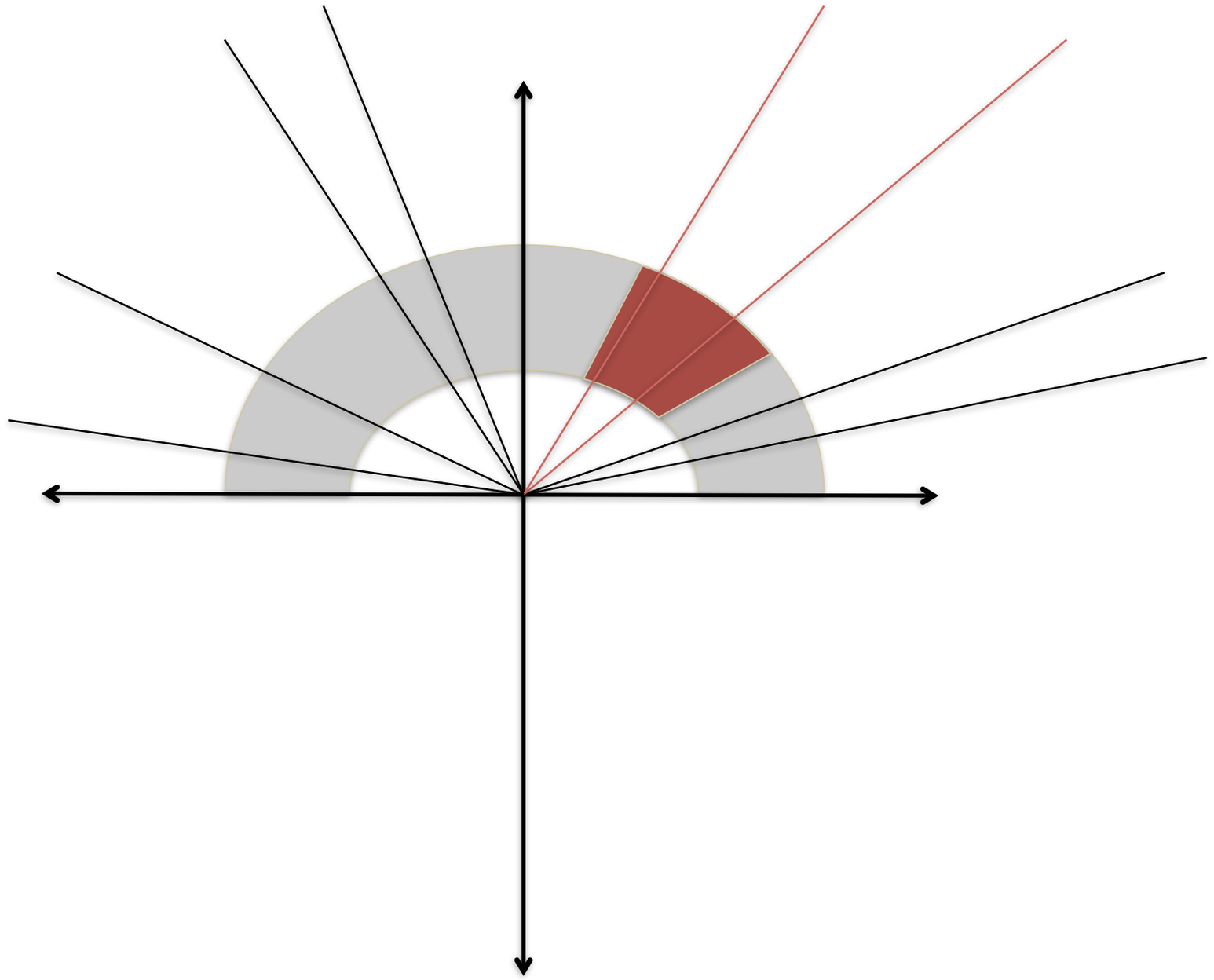


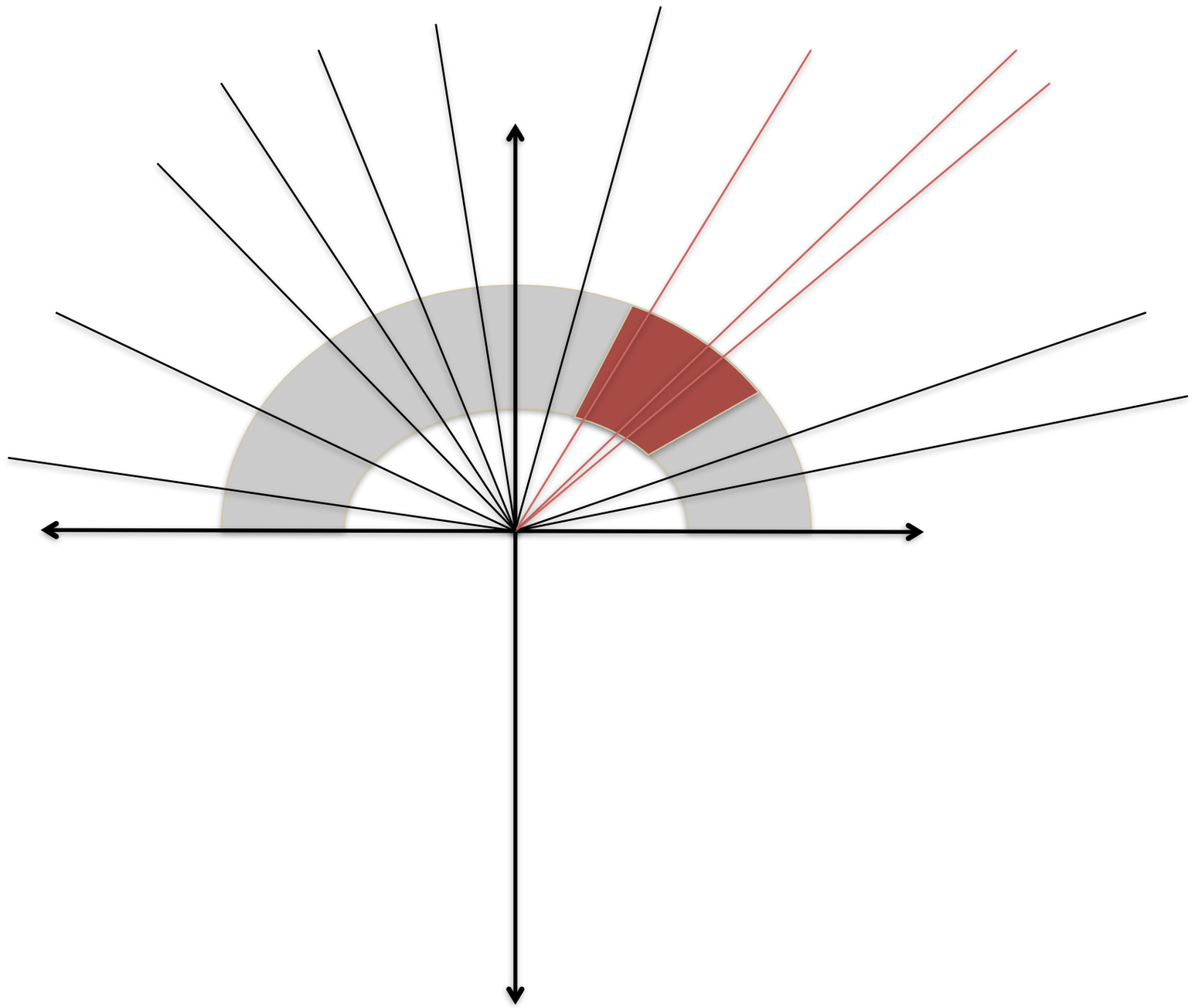

π

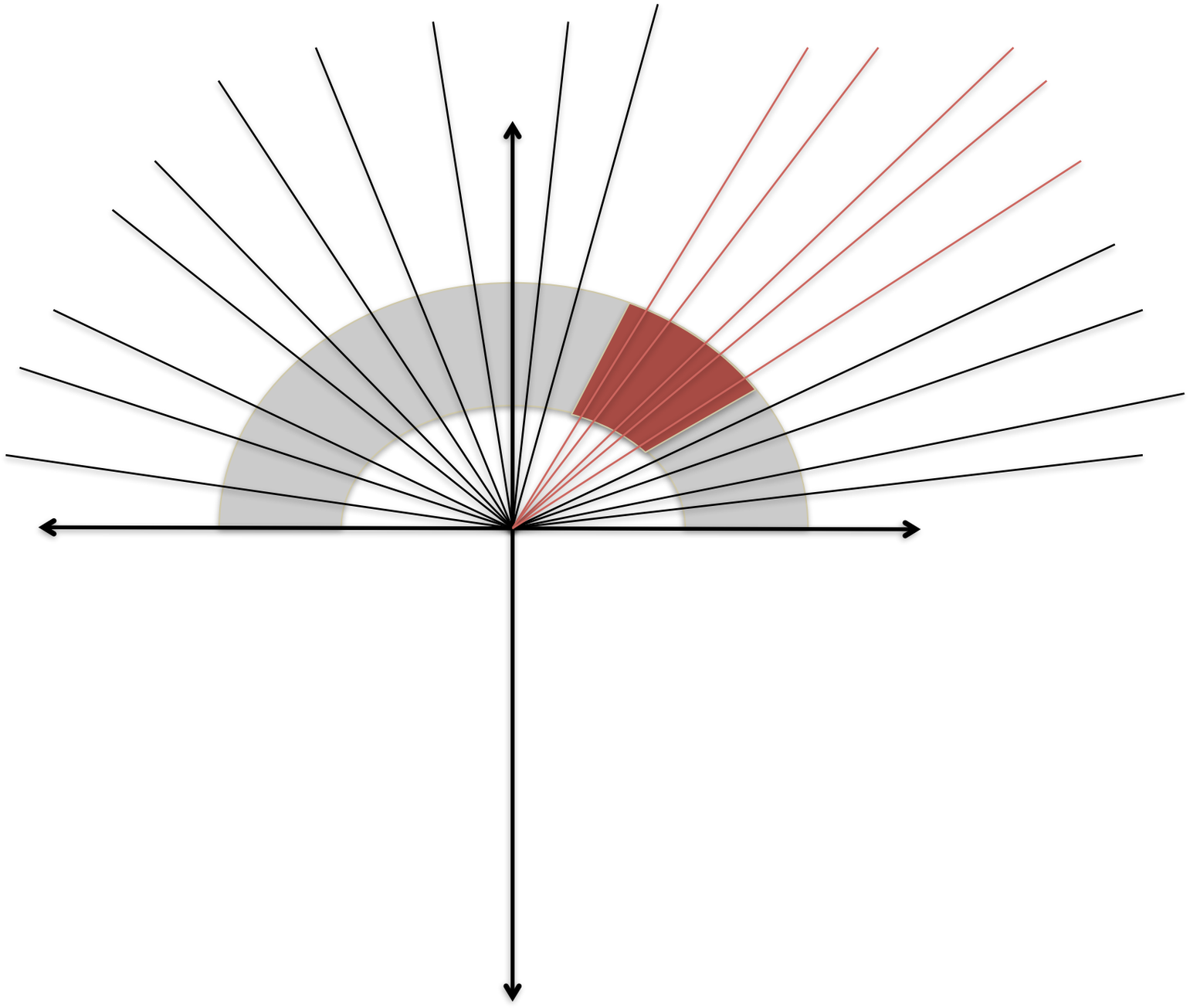

π
θ



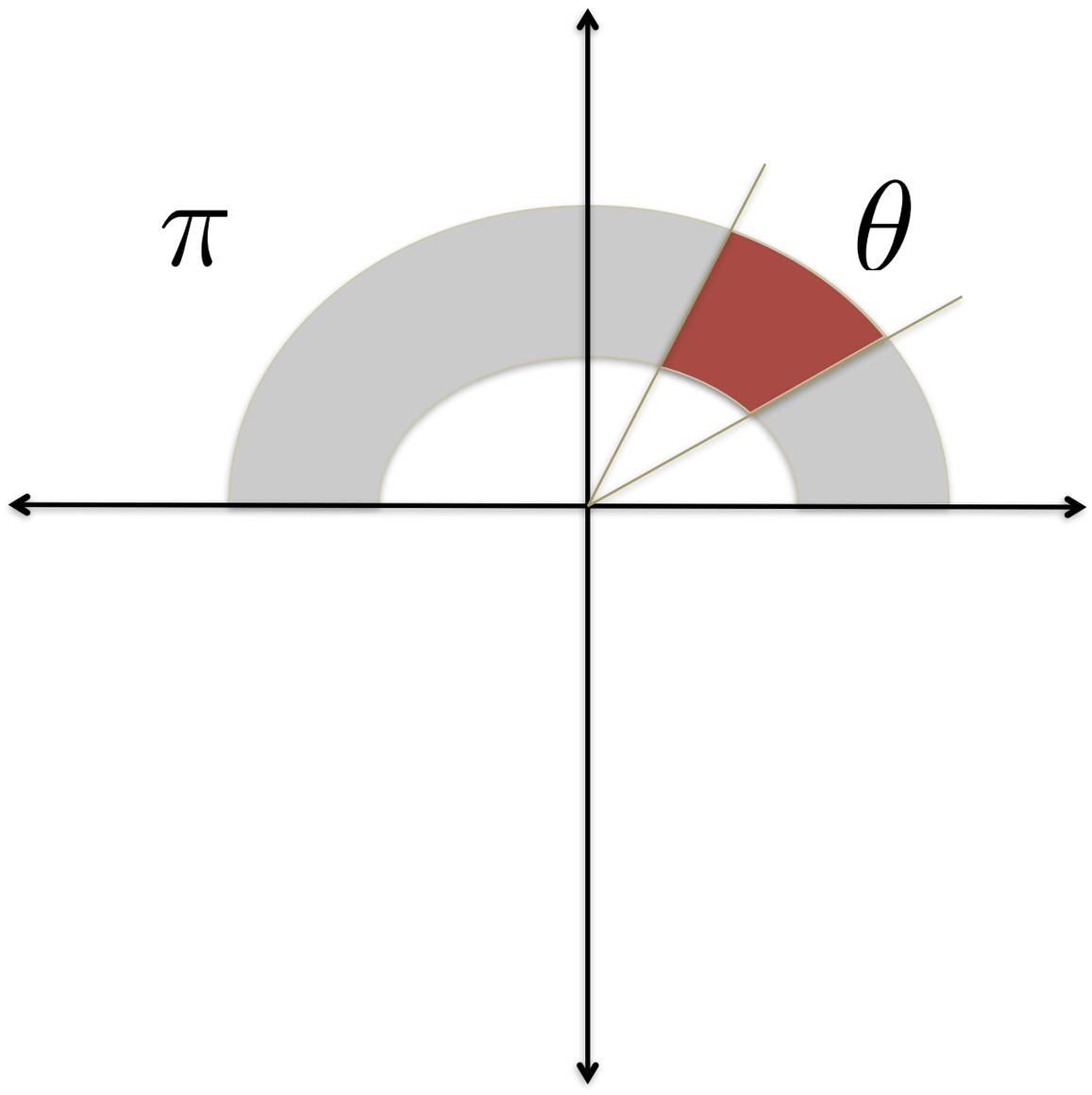








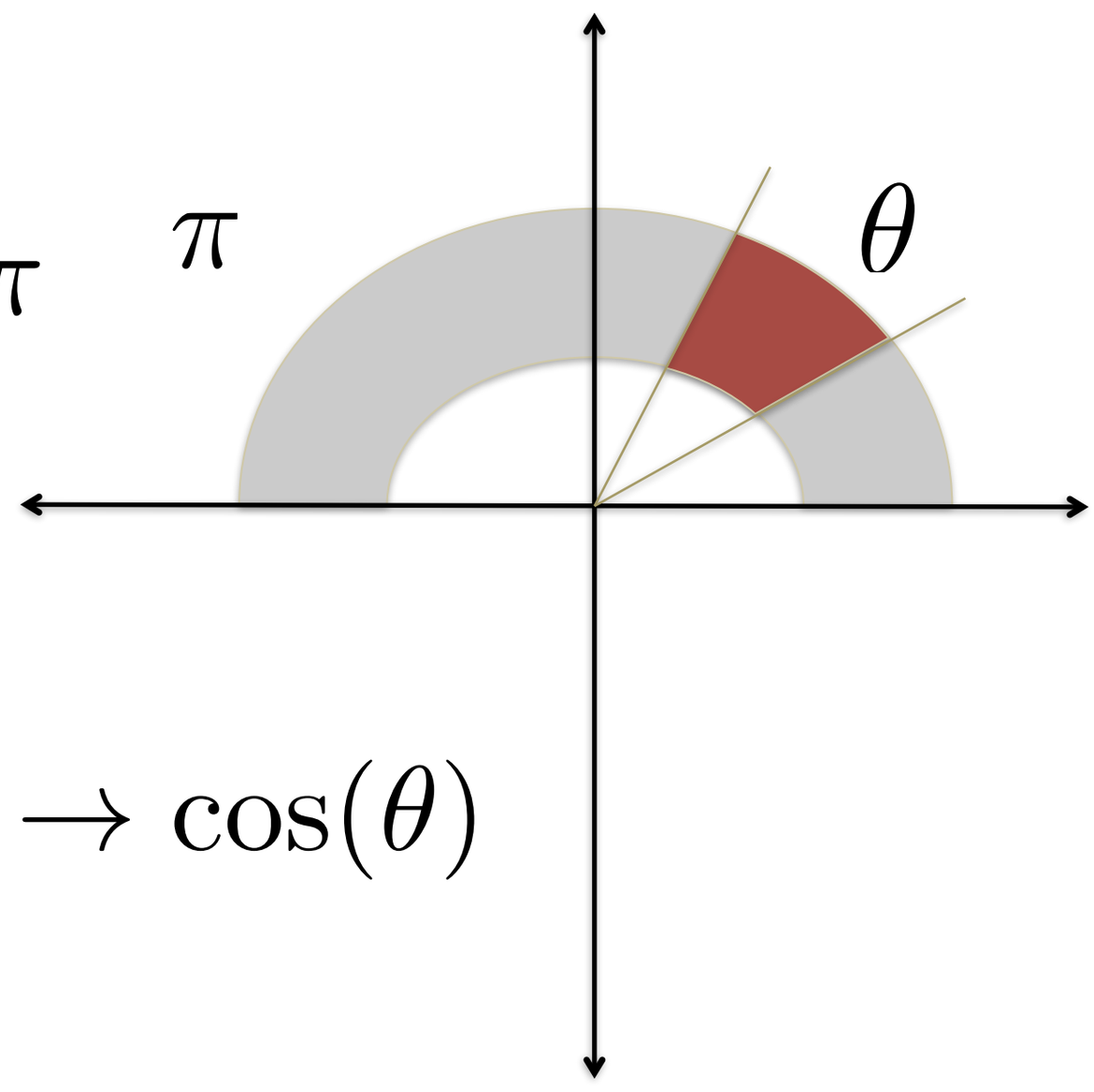
$$\frac{h}{b} \longrightarrow \frac{\theta}{\pi}$$



$$\frac{h}{b} \rightarrow \frac{\theta}{\pi}$$

$$\frac{h}{b}\pi \rightarrow \frac{\theta}{\pi}\pi$$

$$\cos(\frac{h}{b}\pi) \rightarrow \cos(\theta)$$



High dimensional nouns?

- How does this relate to finding “similar” nouns?

	visit x	to x	x Airport	the x	x barked	...
London	100	5,000	250	8	5	...
dog	0	30	0	10,000	7,000	...
...						

A single point in high dimensional “bigram” space

Similarity Clustering

Closest based on true cosine

Accurate

London

Milan$_{.97}$, **Madrid**$_{.96}$, **Stockholm**$_{.96}$, **Manila**$_{.95}$, **Moscow**$_{.95}$

ASHER_0, Champaign_0, MANS_0, NOBLE_0, come_0

Prague_1, Vienna_1, suburban_1, synchronism_1, Copenhagen_2

Frankfurt_4, Prague_4, Taszar_5, Brussels_6, Copenhagen_6

Prague_{12}, Stockholm_{12}, Frankfurt_{14}, Madrid_{14}, Manila_{14}

Stockholm_{20}, Milan_{22}, Madrid_{24}, Taipei_{24}, Frankfurt_{25}

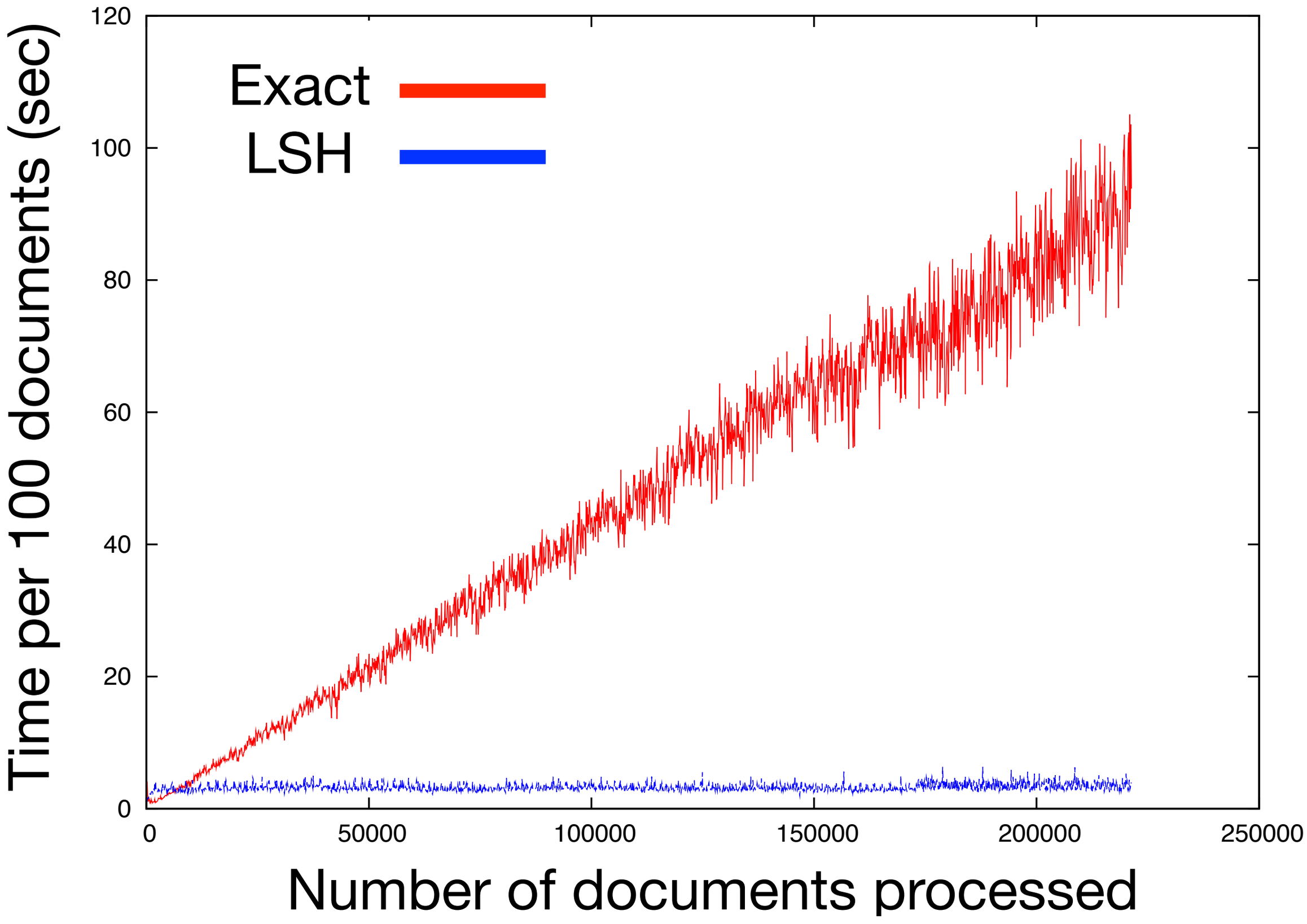
Closest based on 32 bit sig.'s

Cheap

Closest based on 256 bit sig.'s

Cheap-ish

Newswire Experiments



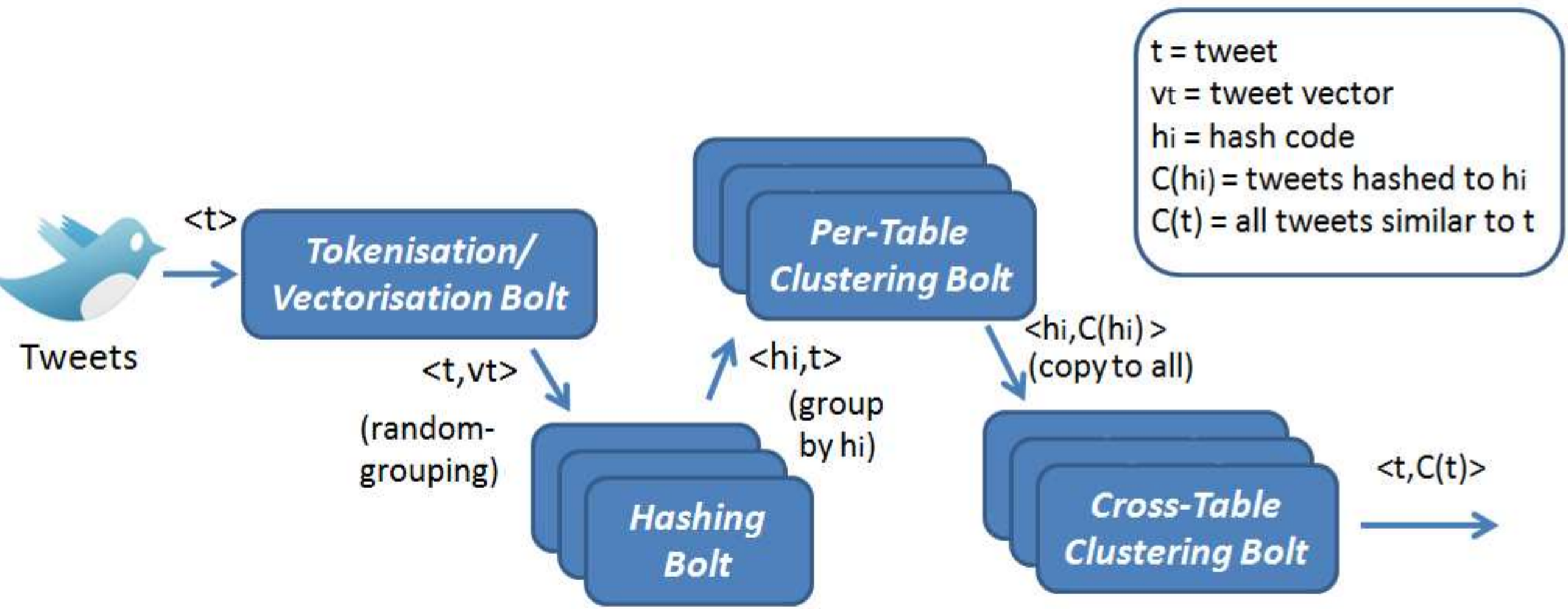
Parallelizing FSD

- LSH enables us to process each incoming post efficiently
- We still need to process thousands of posts per second
- Storm is an in-memory distributed processing streaming infrastructure
  - ‘Real-time Hadoop’
  - Low-latency
  - Suitable for incremental processing

Storm

- Everything runs in-memory, across multiple machines
- Low latency (sub-second response)
- Data is injected into a topology
  - A job is represented as a graph of communicating tasks
- Computation never ends

Storm


t = tweet
vt = tweet vector
hi = hash code
C(hi) = tweets hashed to hi
C(t) = all tweets similar to t
Tweets
<t>
Tokenisation/ Vectorisation Bolt
<t,vt>
(random-grouping)
Hashing Bolt
<hi,t>
(group by hi)
Per-Table Clustering Bolt
<hi,C(hi) >
(copy to all)
Cross-Table Clustering Bolt
<t,C(t)>

Storm Experiment

- Task: process 1 million Tweets, looking for novelty
- Each Tweet is hashed 70 * 13 times
- Results:
  - Linear scaling in terms of the number of machines
  - Approximately 70 cores to deal with the full Firehose (4.5k Tweets per second)

Storm Experiment

- Compared against Hadoop set up with equivalent functionality
- Varied the number of cores
  - Required 24 cores using Hadoop to get the same average throughput as Storm (3 cores).
  - Hadoop has a 24 minute latency; Storm produces results immediately

Event detection in Twitter

- Less than 5% of Tweets carry news-related content
- Running a traditional FSD system on Twitter will produce a tremendous number of false positives
    - Less than 1% of events detected in Twitter are news related

Examples of false positives

- Juicy Couture, Ed Hardy, Coach, Kate Spade and many more! Stay tuned for more brands coming in http://. . .
- i lovee my nephew hair :D
- Going to look at houses tomorrow. One of them is & right behind Sonic Taco Casa. If I live there, I might weigh 400 lbs within a year.
- Hope a bad morning doesnt turn into a bad day...

Quality improvements to TDT on Twitter

- Three strategies:

1. Wait for evidence to accumulate – Event detection trades time for fewer false positives
2. Filter false positives using other streams – If something is interesting it will be seen in multiple places.
3. Classifier – Manually label examples of newsworthy v. not newsworthy, train classifier

Wait for more evidence

- Most spurious events are never noticed by anyone else
- Genuine events tend to attract comments / retweets etc.
- Approach:
  - Wait a short “deferral” period
  - Emit events that are novel and attract follow-ups

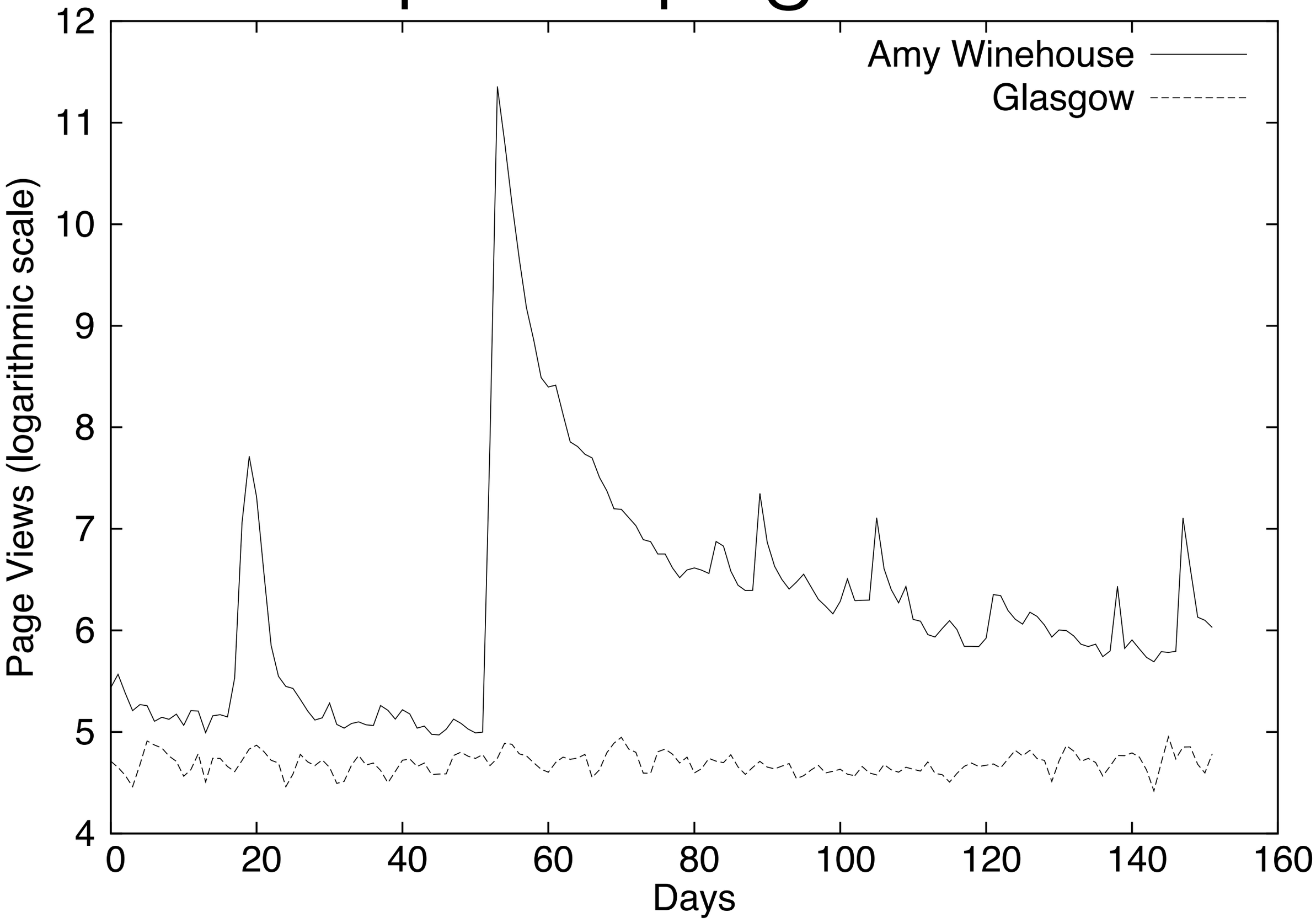
Results for Waiting

- get your free $1000 bestbuy giftcard now! #iloveshopping
- RT @SkyNewsBreak: Sky Sources: 27-year-old singer Amy Winehouse found dead at her flat in North London
- Do you think caylee got justice? #caseyanthony
- Tweeting from my new iPad2!! thank you!! #freestuff
- how dumb are you?-take this quiz and retweet your score

Streaming Data

Stream	Volume Per Day	Total Volume	Units
Twitter	662,000	51 million	Tweets
Wikipedia	240 million	18.5 billion	page requests
Newsire	610	47,000	story posts

Wikipedia page views



Filtering Events using Wikipedia

- Approach:
  - Run FSD system over Twitter.
    - Find all time-synchronous spiking Wiki pages.
    - If a Tweet matches with a spiking page, emit it.

Example Events using Wikipedia Filter

- I love Seth meyers! #ESPYs
- @tanacondasteve amy whinehouse is dead
- RT @katyperry: HAPPY 4TH OF JULY!!!!!!!!!!!!!! . . .
- Yao Ming retired
- Derek jeter 3000 hits.

Wikipedia has a 90 minute latency

Filtering Events using Newswire

- Approach:
  - Run FSD system over Twitter.
  - Find all time-synchronous Newswire
  - If a Tweet is sufficiently similar to an aligned Newswire page, emit it.

Filtering Events using Newswire

- http:www.weshopsongs.com/news.html Amy Winehouse, British Soul Singer With a Destructive Image, Dies at 27
- On Baseball: Jeter Reaches Fabled 3,000, and It's a Blast: At Yankee Stadium, Derek Jeter became the 28th player... http:. . .
- RT @AdamAndEvePR: Japan trade surplus grows in July: Japan's trade surplus widens by more than expected in July, boosting optimism... ht ...
- RT @SkyNewsBreak: Petrol bombs thrown at officers and some cars set alight in Derry, Northern Ireland
- Two arrested over Croydon death: Two men are arrested over the death of Trevor Ellis, who was found with bullet ... http:. . .

Filtering Events using a classifier

- Why not simply filter posts using machine learning?
- Manually labelled 145k events and trained a classifier
- Baseline: 96% accuracy; classifier 98.4%

I will be using http:// to manage and clean my twitter account	not newsworthy
RT @CNN: Gunmen open fire on sleeping college students in Nigeria	newsworthy

Classifier examples

- RT @ajam: BREAKING: President Obama directs Kerry to pursue nuclear weapons deal with Iran: ”diplomatic path must be tested.”
- Baupost hedge fund plans to return some money: sources - Reuters http:
- RT @chrisgeidner: Senate Confirms Todd Hughes, First Openly Gay Federal Appeals Court Judge,
- Exclusive: Secretary of State Kerry to sign Arms Trade Treaty - diplomats http://t. via @reuters
- India’s Political Parties Offer Subsidized Mobile Phones for Votes http: via @mashable
- RT @BreakingNews: The police headquarters in Columbus, Ohio, is evacuated after reports of an explosion; officers are heading down the stai

What strategy is best?

Does Twitter lead Newswire for breaking news?

- Social Media research commonly asserts that Twitter *leads* Newswire
- Usually the same examples are mentioned:
  - Plane crashing on Hudson
  - Earthquakes
- How general is this claim?

Does Twitter lead Newswire for breaking news?

- Identified 27 events.
- Task: find breaking news corresponding with these events.
- Manually search for when they first break.

Event	Newswire	Twitter	Lead
Amy Winehouse dies	**07-23 16:10**	07-23 16:11	-0:01
Atlantis shuttle lands	07-21 09:59	**07-21 09:56**	+0:03
Betty Ford dies	**07-09 00:00**	07-09 00:57	-0:57
Richard Bowes killed in riots in England	**08-11 23:18**	08-11 23:31	-0:14
Flight 4896 crash	**07-13 11:37**	07-13 11:46	-0:09
S&P downgrade US credit rating	**08-06 00:11**	08-06 00:18	-0:07
US increases debt ceiling	08-01 23:06	08-01 23:06	0:00
Terrorist attack in Delhi	09-01 05:12	**09-07 04:53**	+0:19
Earthquake in Virginia	08-23 18:24	**08-23 17:53**	+0:31
First victim of London riots dies	08-09 11:46	**08-09 11:45**	+0:01
War criminal Goran Hadzic arrested	**07-20 07:56**	07-21 05:42	-21:46
India and Bangladesh sign a border pact	**09-06 07:15**	09-06 14:24	-7:09
Plane with Russian hockey team Lokomotiv crashes	**09-07 12:51**	09-07 12:59	-0:08
Explosion in French nuclear plant in Marcoule	09-12 11:42	09-12 11:42	0:00
NASA announces there might be water on Mars	08-04 18:08	08-04 18:08	0:00
Google announces plans to buy Motorola Mobility	08-15 11:43	**08-15 11:38**	+0:05
Car bomb explodes in Oslo, Norway	07-22 13:57	**07-22 13:38**	+0:19
Gunman opens fire in youth camp in Norway	**07-22 16:13**	07-22 16:14	-0:01

Twitter versus Newswire

- No evidence that Twitter generally leads
- On average most important news is reported first on Newswire
- Tweets repost already posted news
- Most important news appears in Twitter

Twitter versus Newswire

- Twitter wins for a certain class of news:
  - Earthquakes, natural disasters
  - Sports news
  - Hyper-local events

Newswire breaks

- Iranian actor Pegah Ahangarani arrested in Teheran
- Court upholds decision to impose control order on terror suspect in London
- Support of Assad government shows signs of weakening
- French socialist project sharing and caring in bid to beat Nicolas Sarkozy
- Dick Cheney autobiography heaps praise on Tony Blair

Twitter breaks

- RIP Rick Rypien. Sad to see another death in the NHL. Too many tragedies in the world lately
- UFC on Versus 5 results: Jacob Volkmann def. Danny Castillo via unanimous decision (29-28, 29-28, 29-28)
- RT NASA: NASA is ready to move forward with Space Launch System, a new capability for human exploration beyond Earth
- Car reg NP05 LPU looting PC World Charlton. Retweet and shame.
- RT DerbysPolice: To reiterate rumours circulating there is disorder or looting in Derby city are untrue. Please RT. #derby #police

Wrap-up

- We can capture collective intelligence by mining Twitter steams to detect important events in the world
- Efficiency is key: without LSH these experiments would be impossible.
- Need for a short deferral period External streams can be useful:
  - Latency is a problem
  - Helps us understand the relationship between streams
  - If we have Newswire do we need Twitter?

Event Detection and Security

- Armed police in stand-off with man threatening to blow himself up at Terminal 5 Heathrow. Terminal evacuated.
- London's Heathrow Terminal 5 partially evacuated http:
- At least three foreign photographers beaten by men believed to be plainclothes police on Wangfujing. One hospitalized.
- RT DerbysPolice: To reiterate rumours circulating there is disorder or looting in Derby city Automatically detected events, 2011

Multilingual Event Detection

- Early Haiti Earthquake-related Tweets in our crawl:
- 22:24:43 Terremoto en haiti 7.3 posible tsunami en el caribe fuente cnn hace 1 min
- 22:23:57 RT @jorr2006: temblor 7 grados en haiti http://earthquake.usgs.gov/earthquakes/recenteqsww/Quakes/ reading the USGS and Nat Weather Service
- 22:17:43 NOT expecting Tsunami on east coast after haiti earthquake. good new

Earthquake struck 21:53 UTC

Bibliography

- Sasa Petrovic, Miles Osborne and Victor Lavrenko. **Streaming First Story Detection with application to Twitter.** NAACL 2010.
- Benjamin Van Durme and Ashwin Lall. **Online Generation of Locality Sensitive Hash Signatures**. ACL 2010
- Piotr Indyk and Rajeev Motwani. **Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality**. Theory of Computing 1998.