Consistency without consensus: CRDTs in production at SoundCloud

Consistency without consensus: CRDTs in production at SoundCloud

Me

Some guy

Embedded, sensor networks

Distributed systems

SoundCloud infrastructure

Theory

Distributed programming

> “The art of solving the same problem that you can solve on a single computer using multiple computers.”
>
> —*book.mixu.net*

Distributed programming

“Generally a bad idea,
best avoided.”

—*me*

```
>>> x = 1
>>> print x
1
```

```
$ curl -XPOST -d'{"val": 1}' http://db/vars/x

HTTP 502 Bad Gateway

$ curl -XGET http://db/vars/x

HTTP 503 Service Unavailable
```

Idioms

1980s — RPC

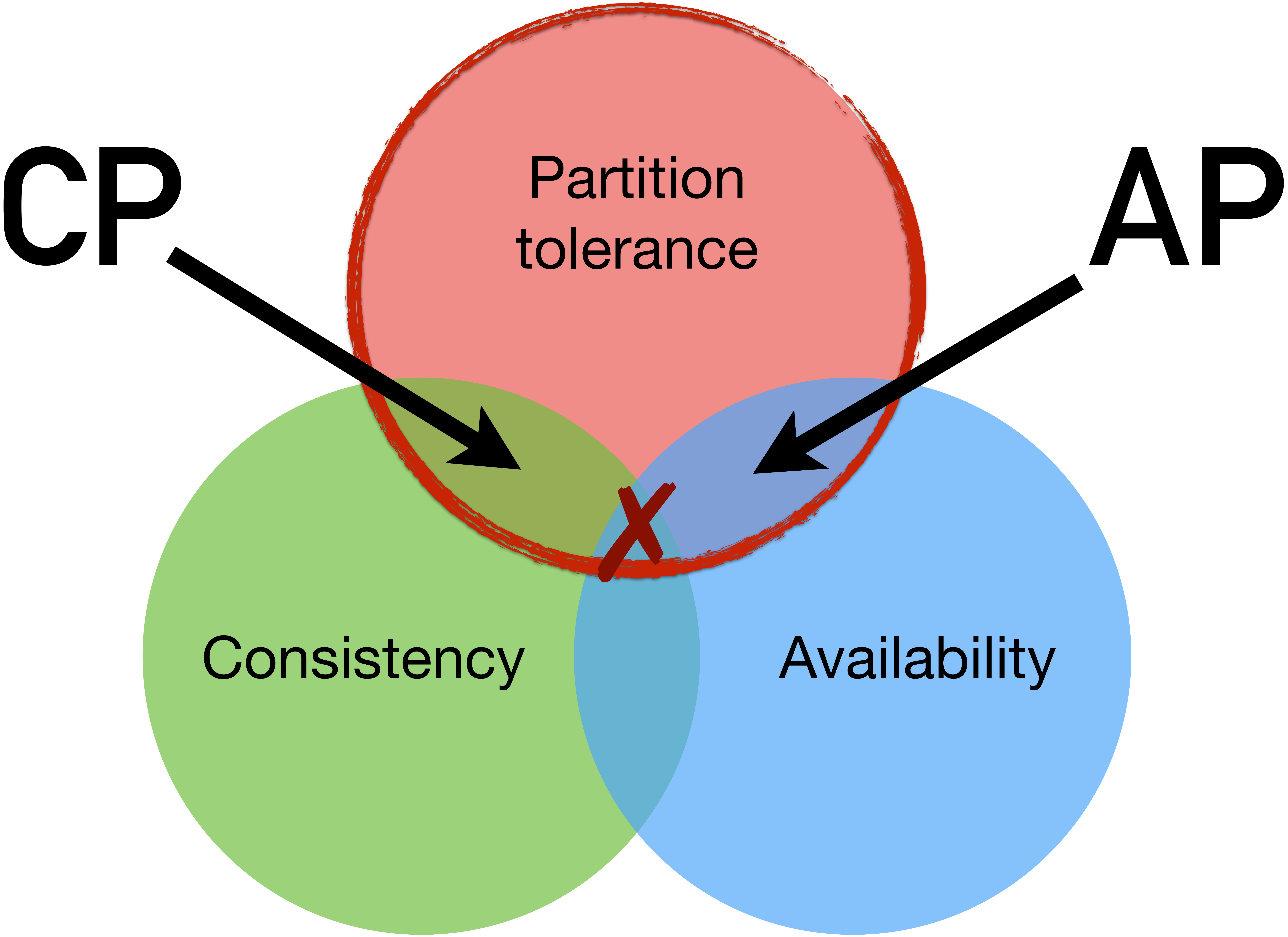
1990s — CORBA

2000 — CAP

Partition tolerance

“The system continues to operate despite message loss due to network and/or node failure.”

—book.mixu.net


CP
Partition tolerance
AP
Consistency
Availability

CP

Chubby, Doozer — Paxos

ZooKeeper — Zab

Consul, etcd — Raft

? — Viewstamped Replication

AP

Cassandra

Riak

Mongo

Couch

Message failure

Delayed

Dropped

Delivered out-of-order

Duplicated

CALM principle

Consistency

As

Logical

Monotonicity

ACID 2.0

Associative

Commutative

Idempotent

Distributed, sure, whatever

CRDT

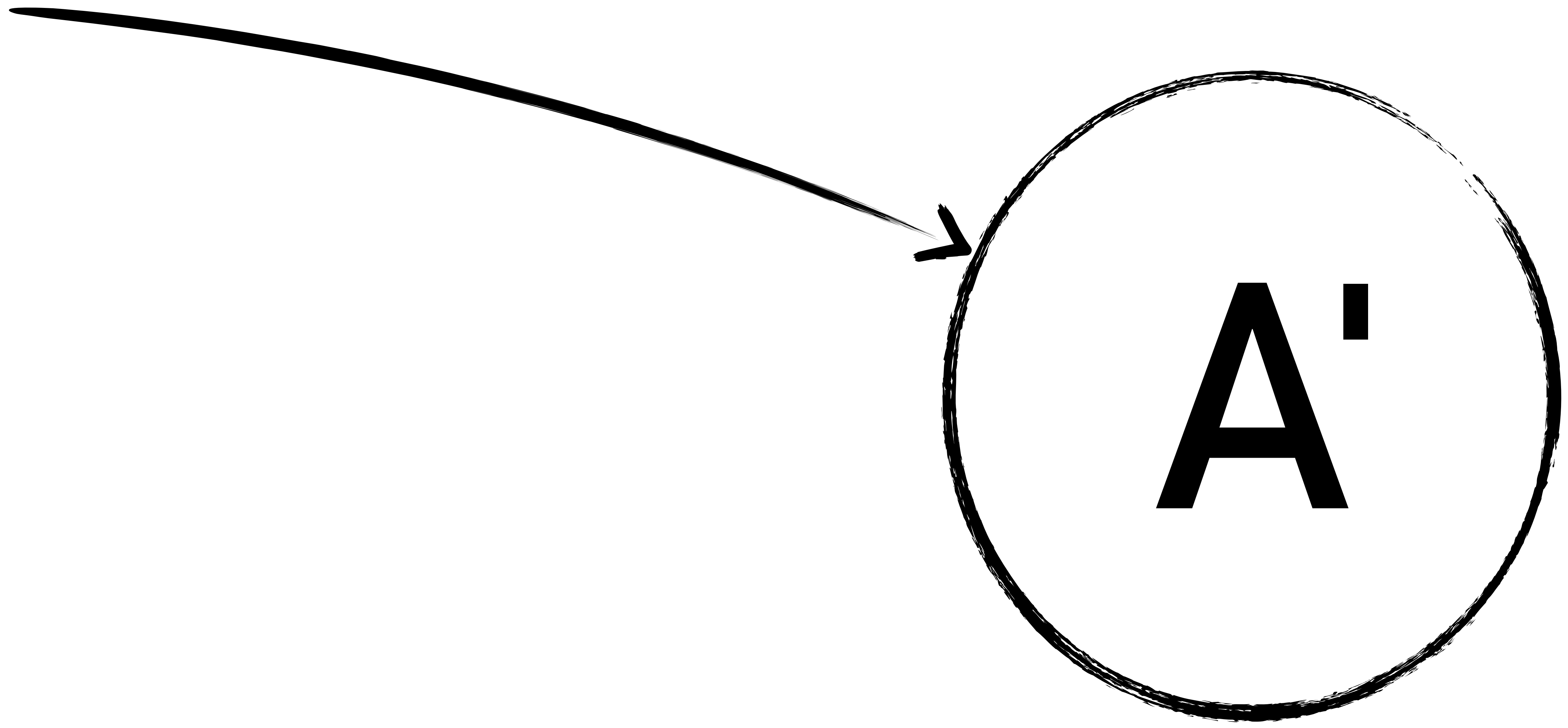
Conflict-free

Replicated

Data

Type

Increment-only counter



A → A'

+

$$(1 + 2) + 3 = 1 + (2 + 3)$$

$$1 + 2 = 2 + 1$$

$$1 + 1 \neq 1$$

∪

$$(1 \cup 2) \cup 3 = 1 \cup (2 \cup 3)$$

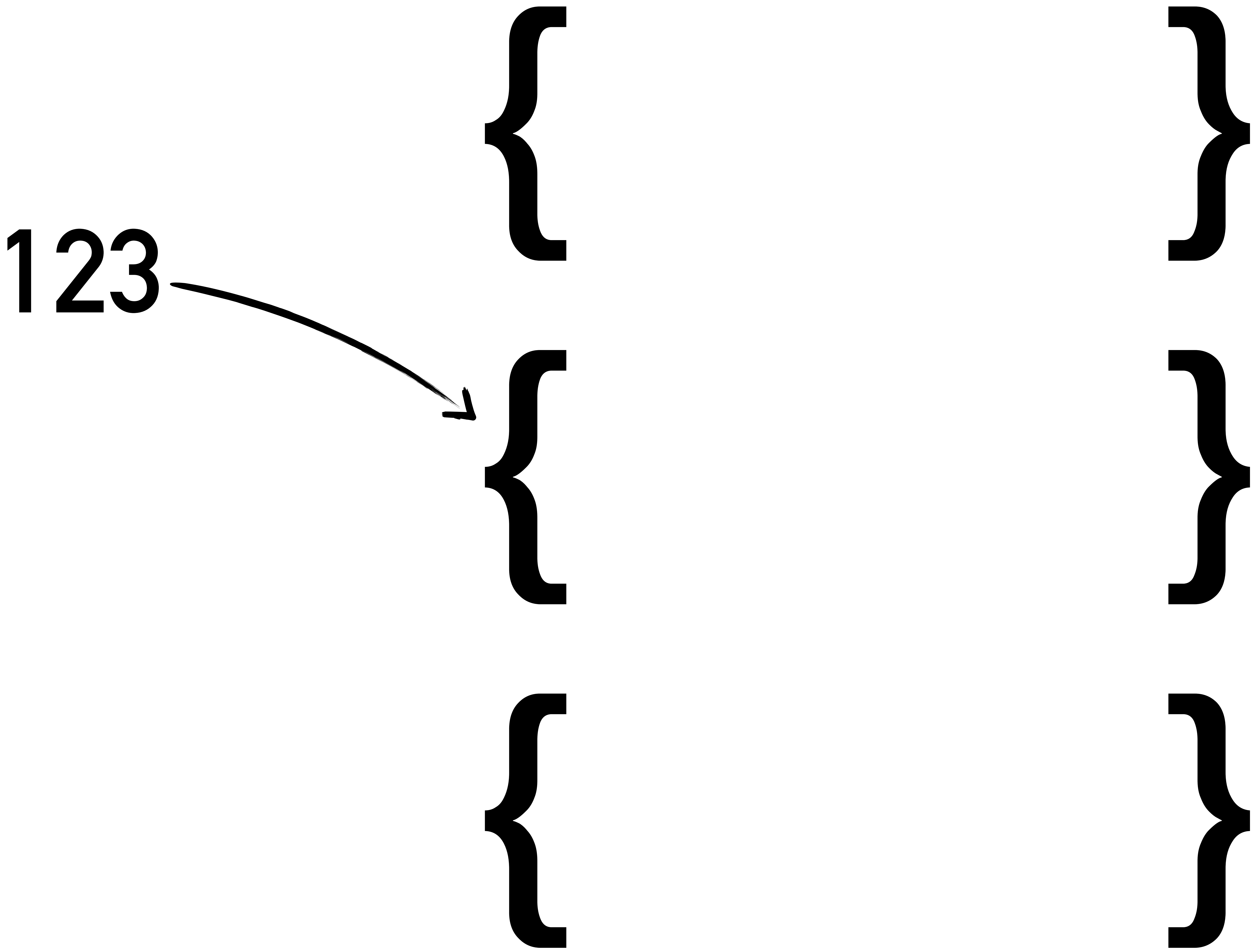
$$1 \cup 2 = 2 \cup 1$$

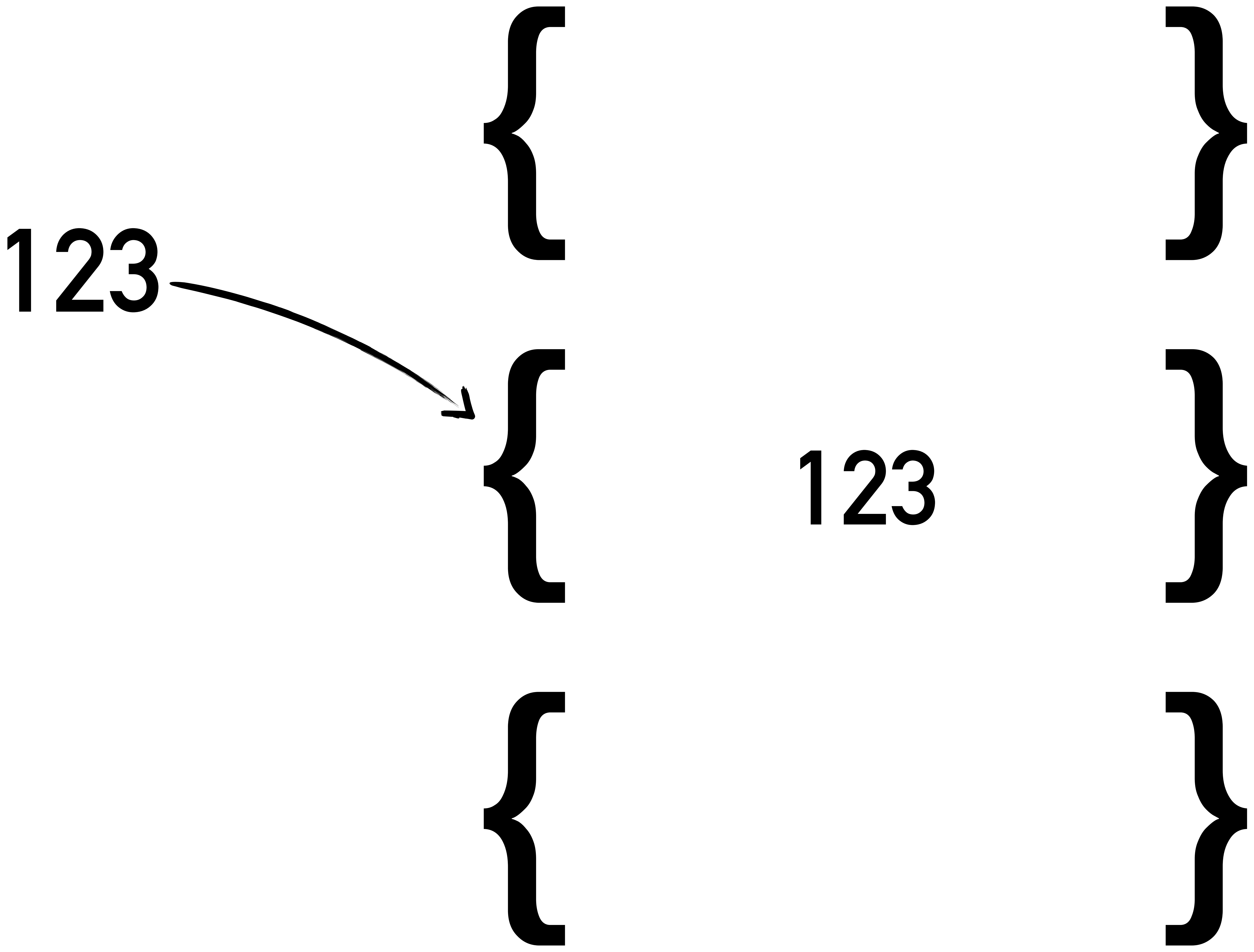
$$1 \cup 1 = 1$$

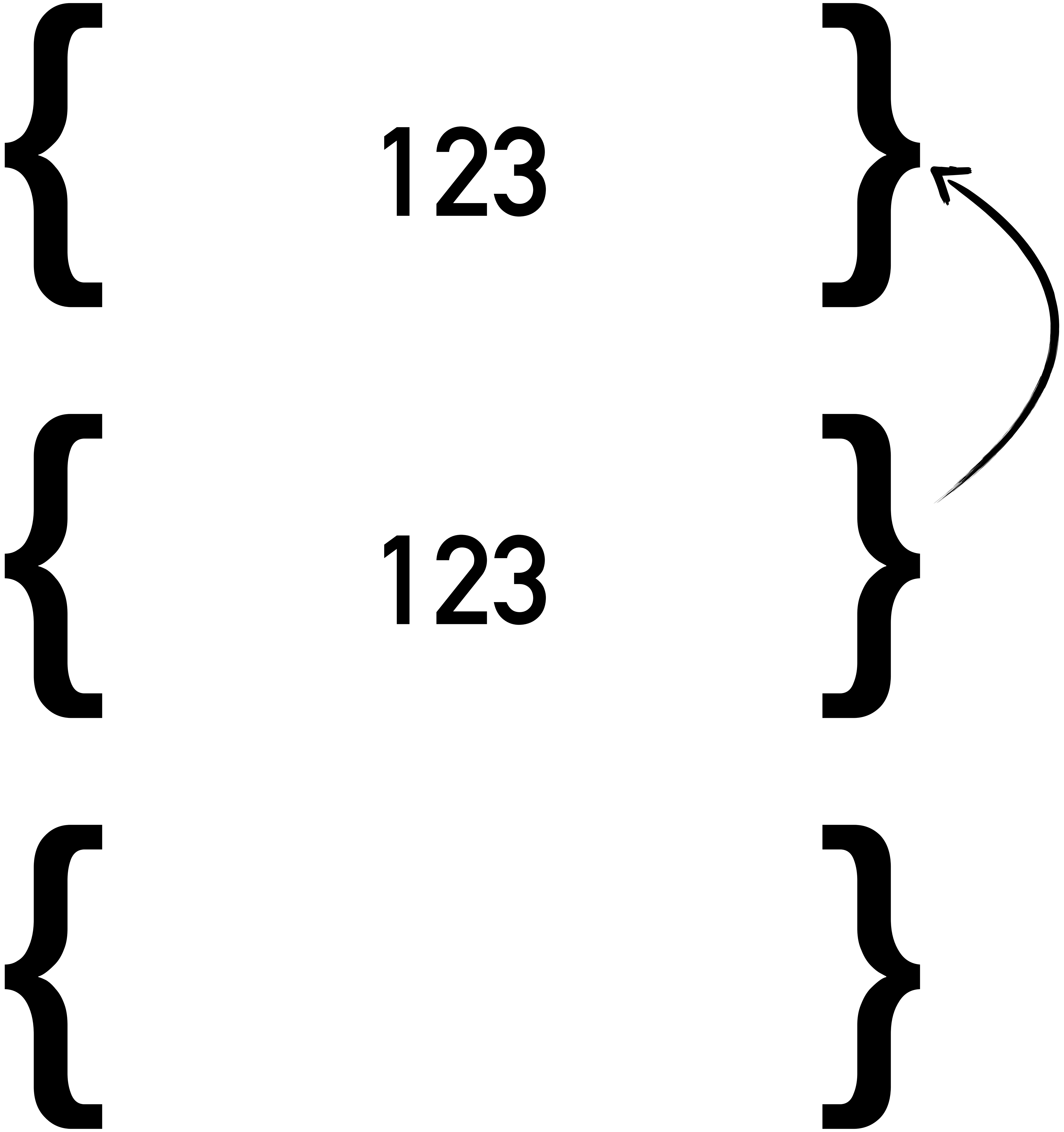
{ }

{ }

{ }

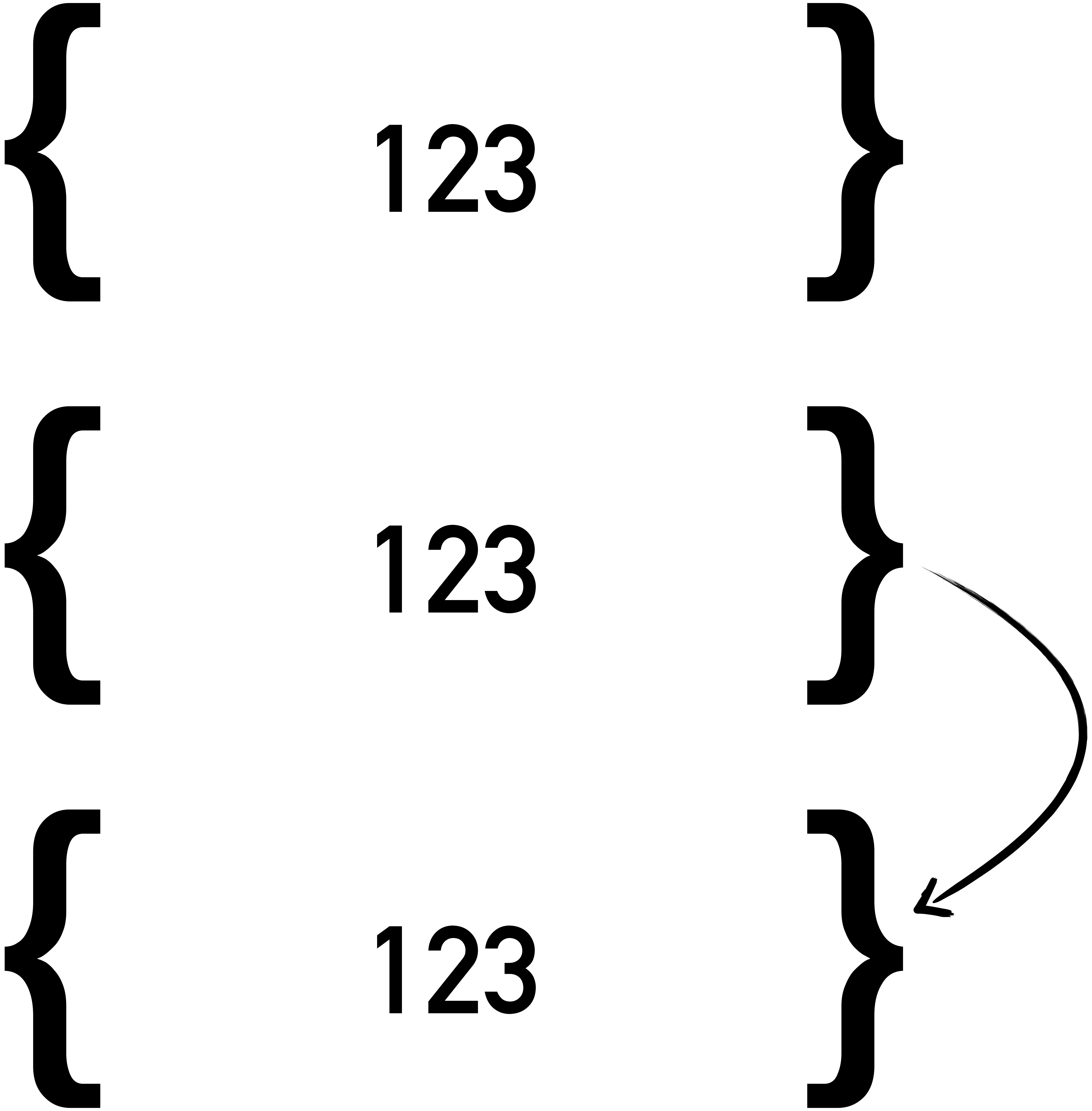

123


123
123



123

123



{ 123 }

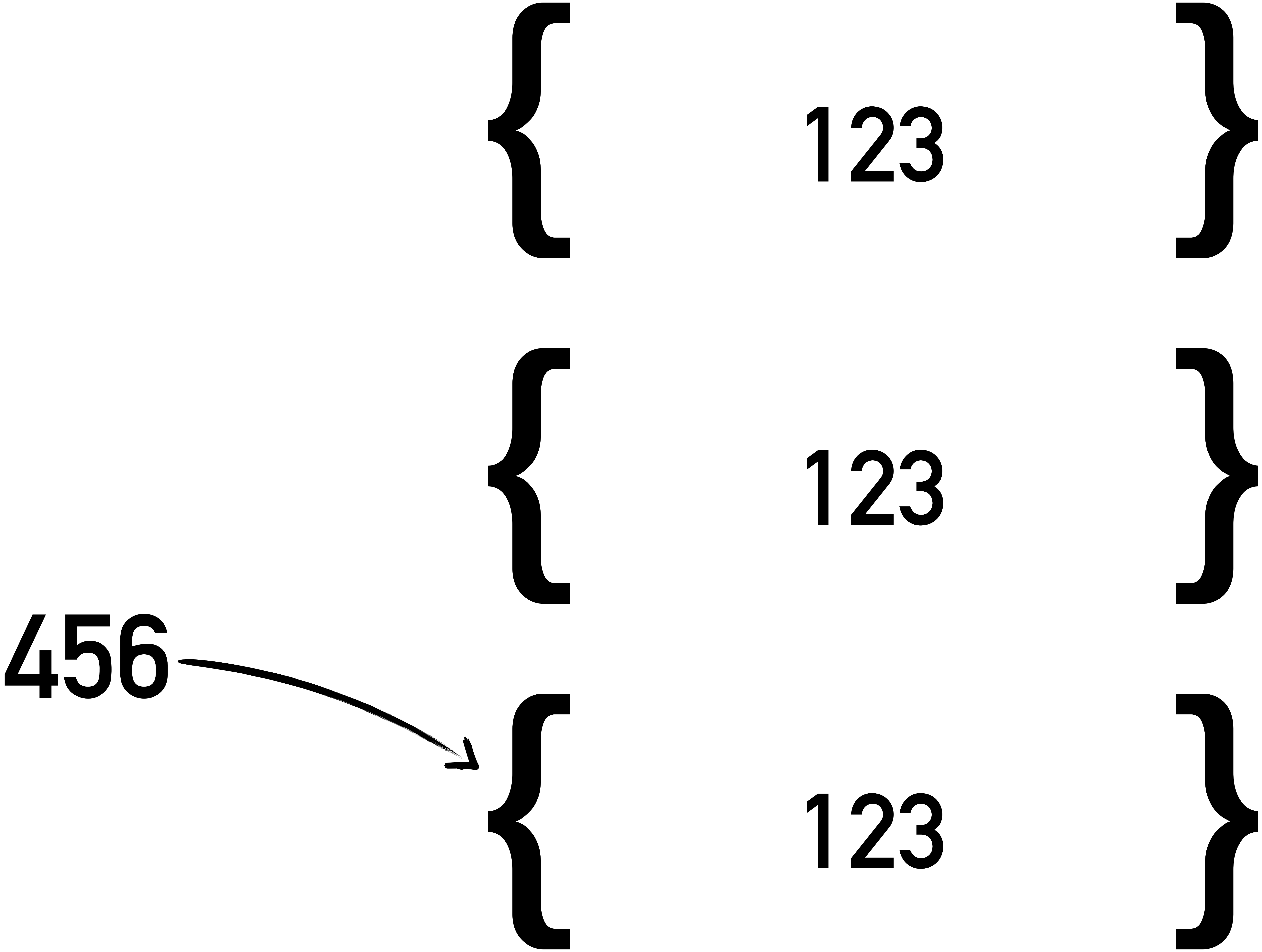
{ 123 }

{ 123 }

{ 123 }

{ 123 }

{ 123 }

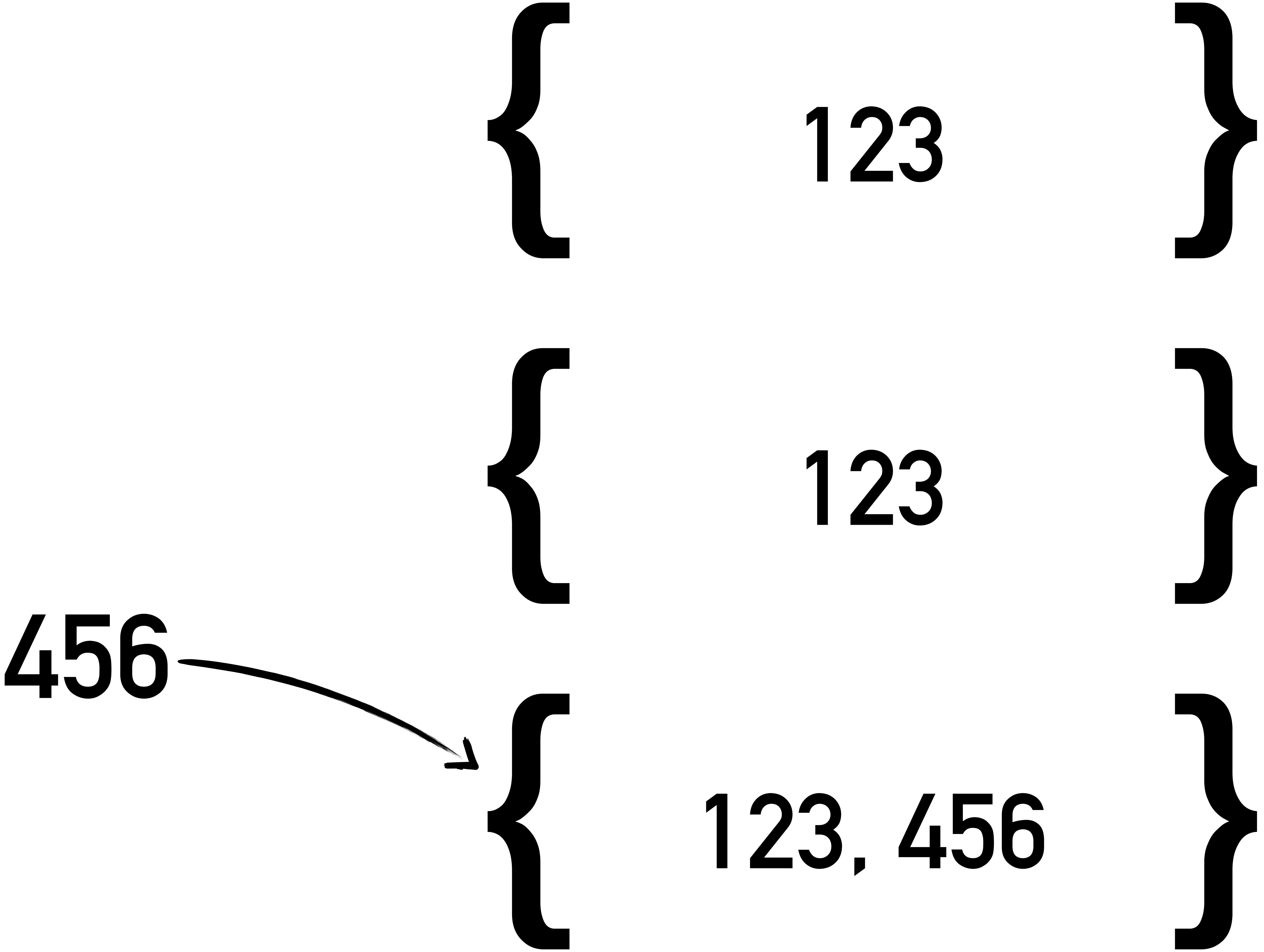


123

123

456

123

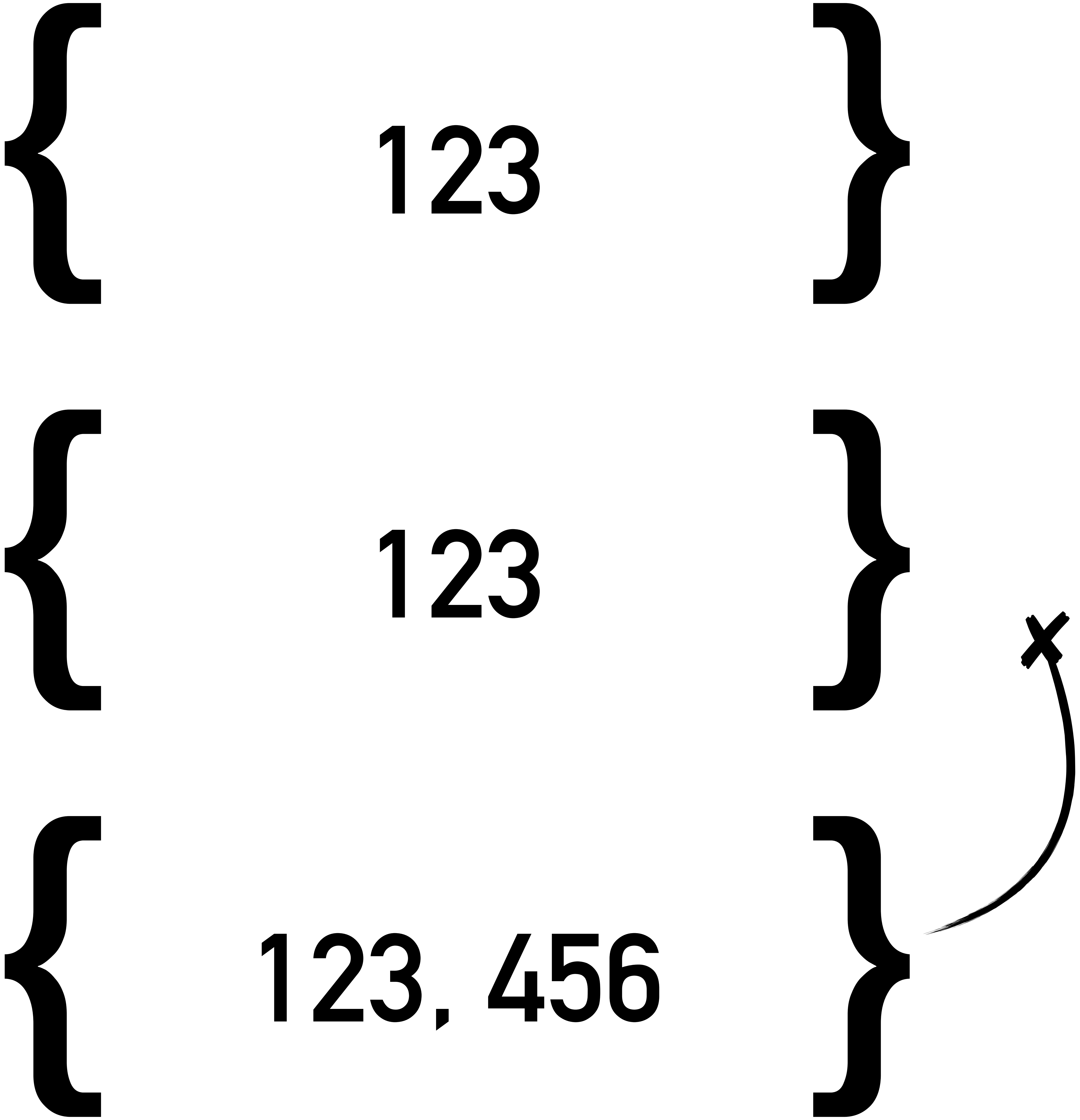


{ 123 }

{ 123 }

456

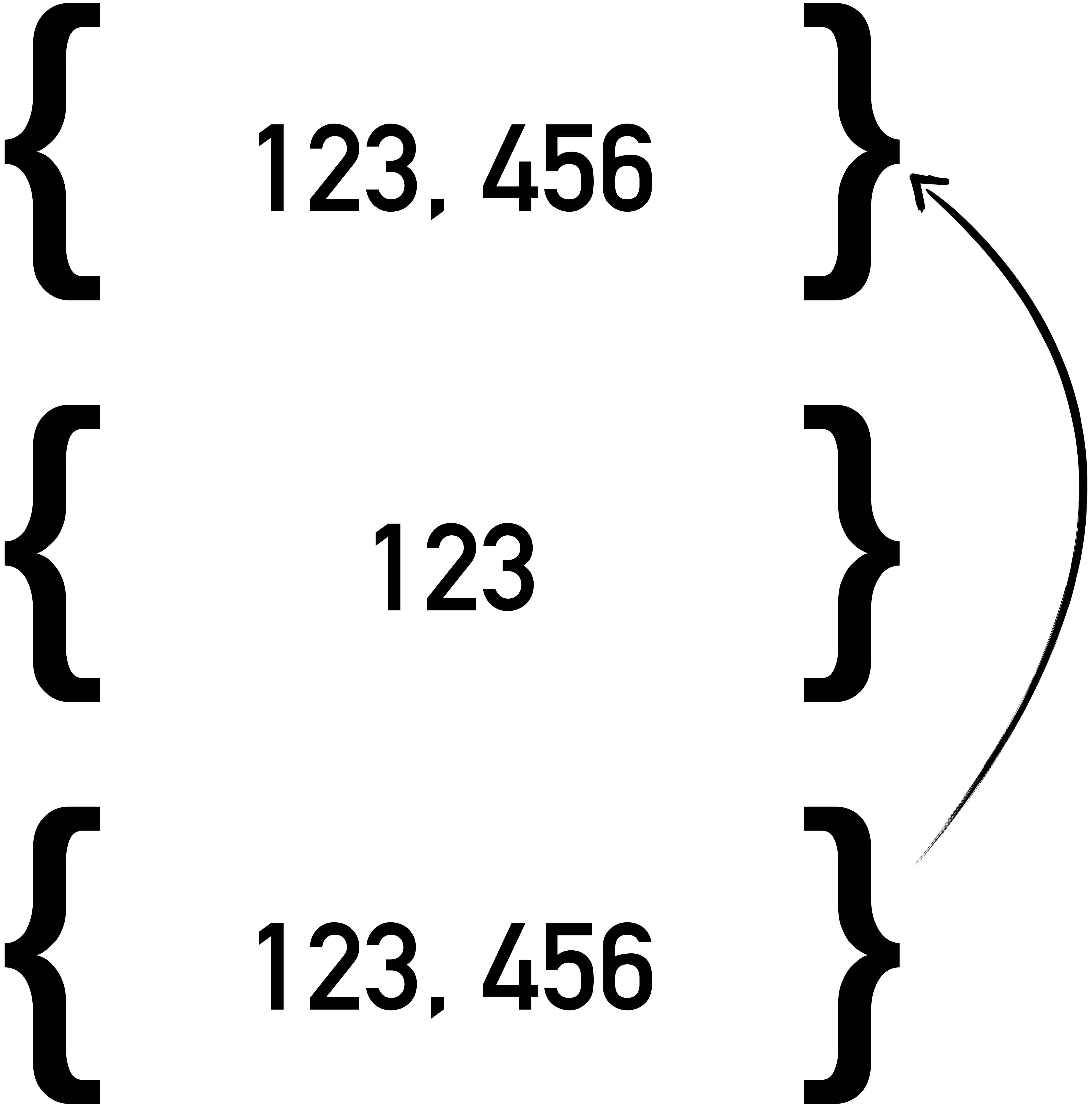
{ 123, 456 }



{ 123 }

{ 123 }

{ 123, 456 }



{ 123, 456 }

{ 123 }

{ 123, 456 }

{ 123, 456 }

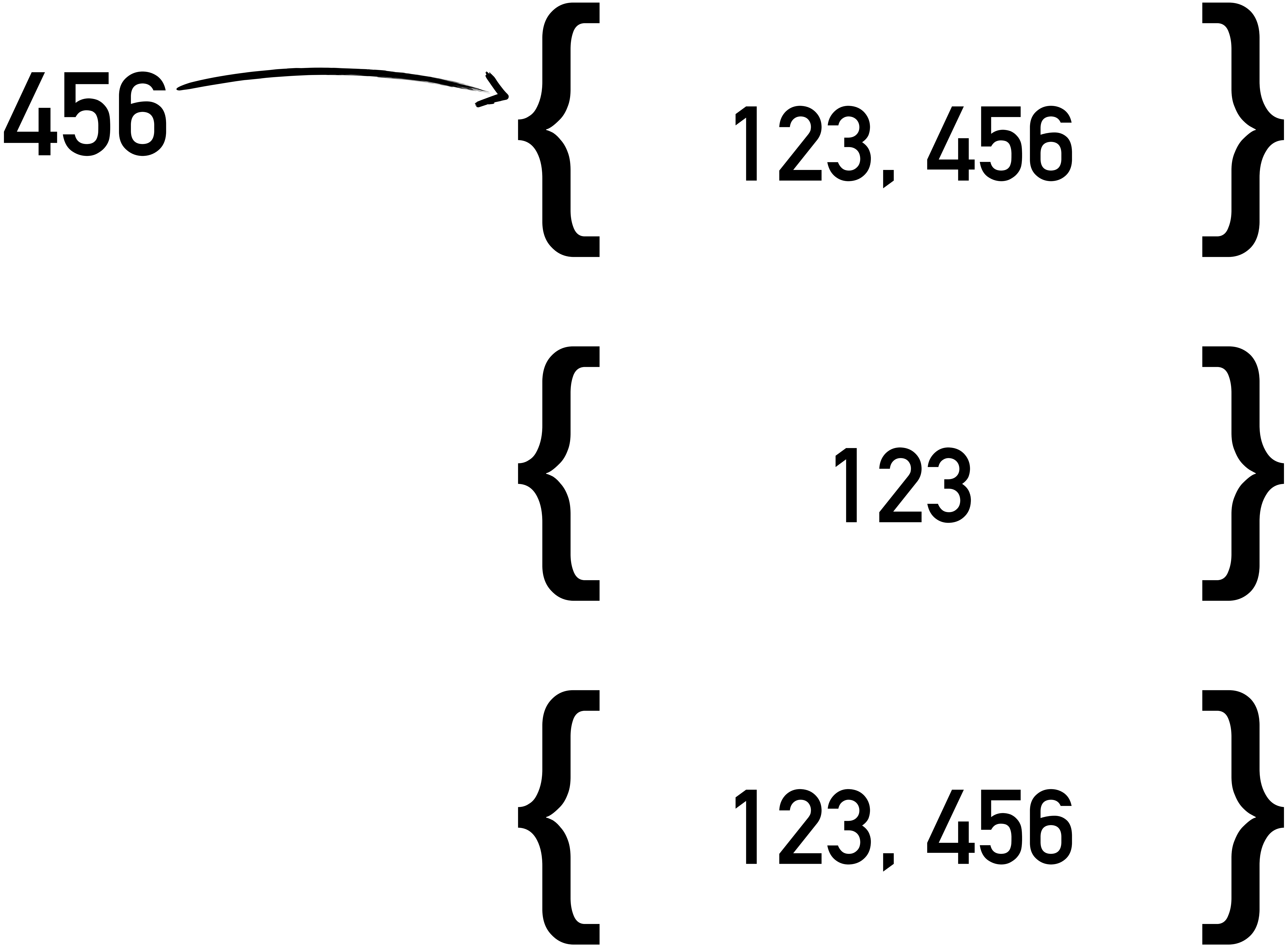
{ 123 }

{ 123, 456 }

Read

$$\{123, 456\} \cup \{123\} \cup \{123, 456\} = \{123, 456\}$$

$$\{123, 456\} \Delta \{123\} \Delta \{123, 456\} = \mathbf{\{456\}}$$

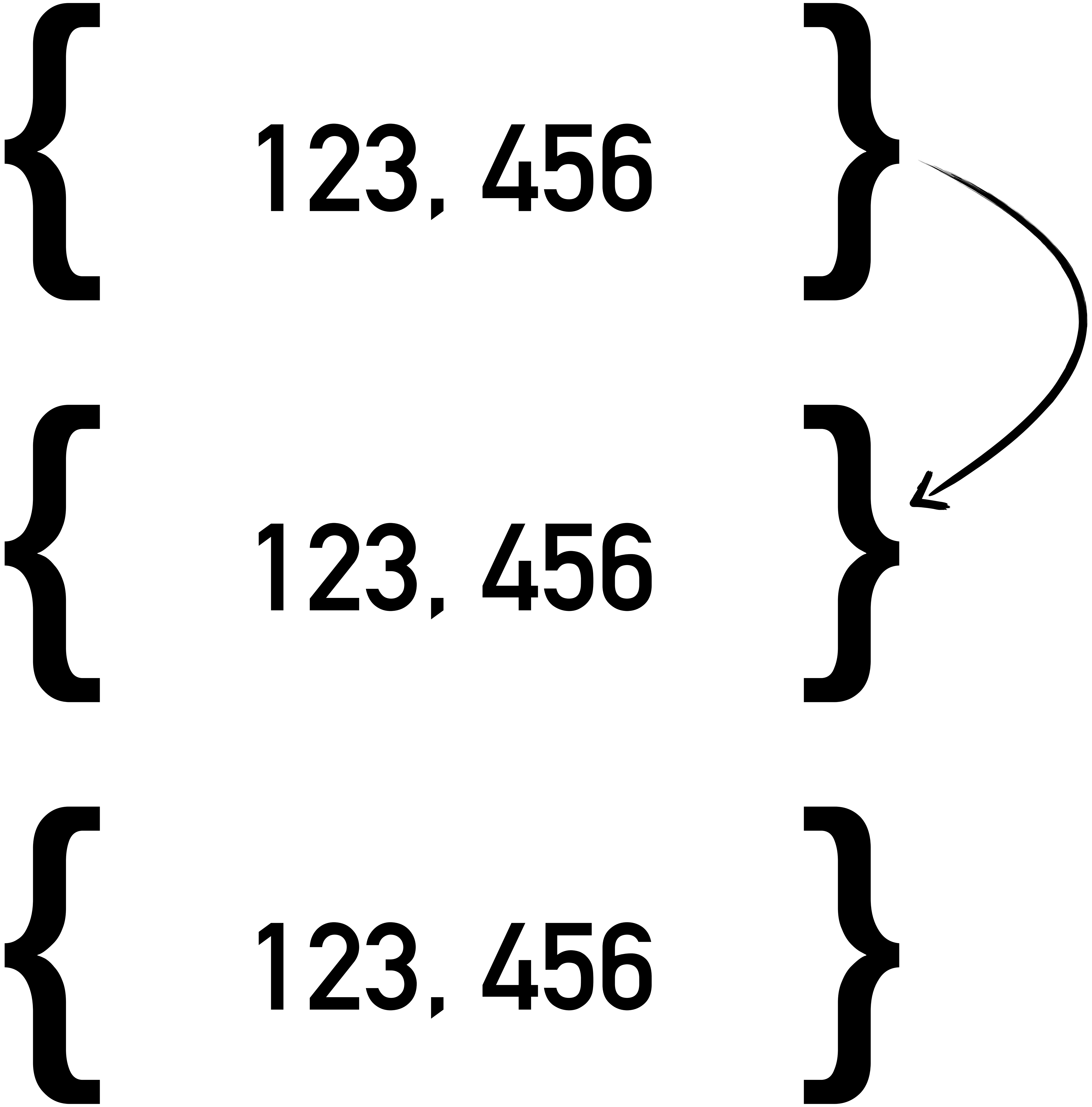


456

{ 123, 456 }

{ 123 }

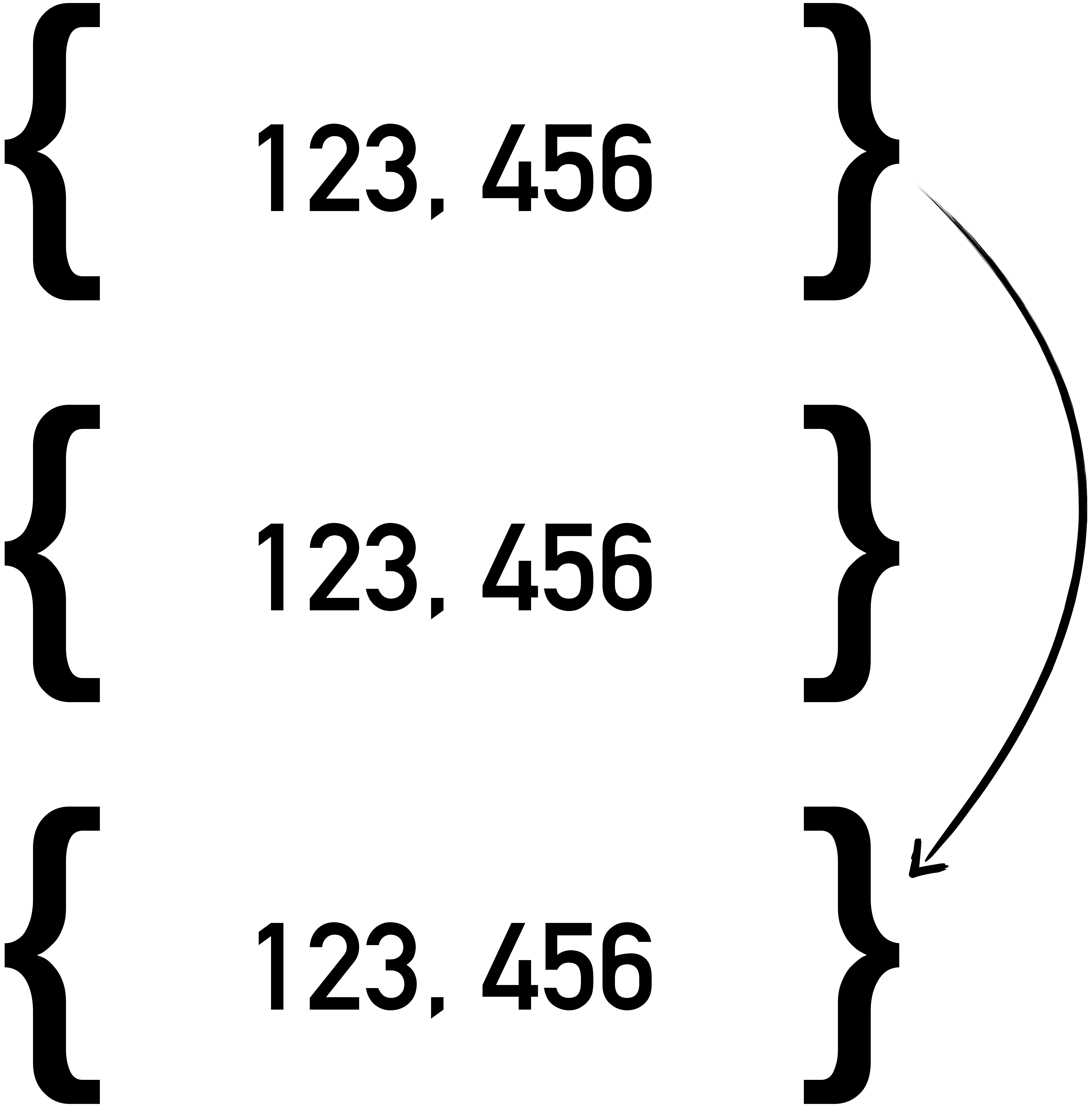
{ 123, 456 }



{ 123, 456 }

{ 123, 456 }

{ 123, 456 }



{ 123, 456 }

{ 123, 456 }

{ 123, 456 }

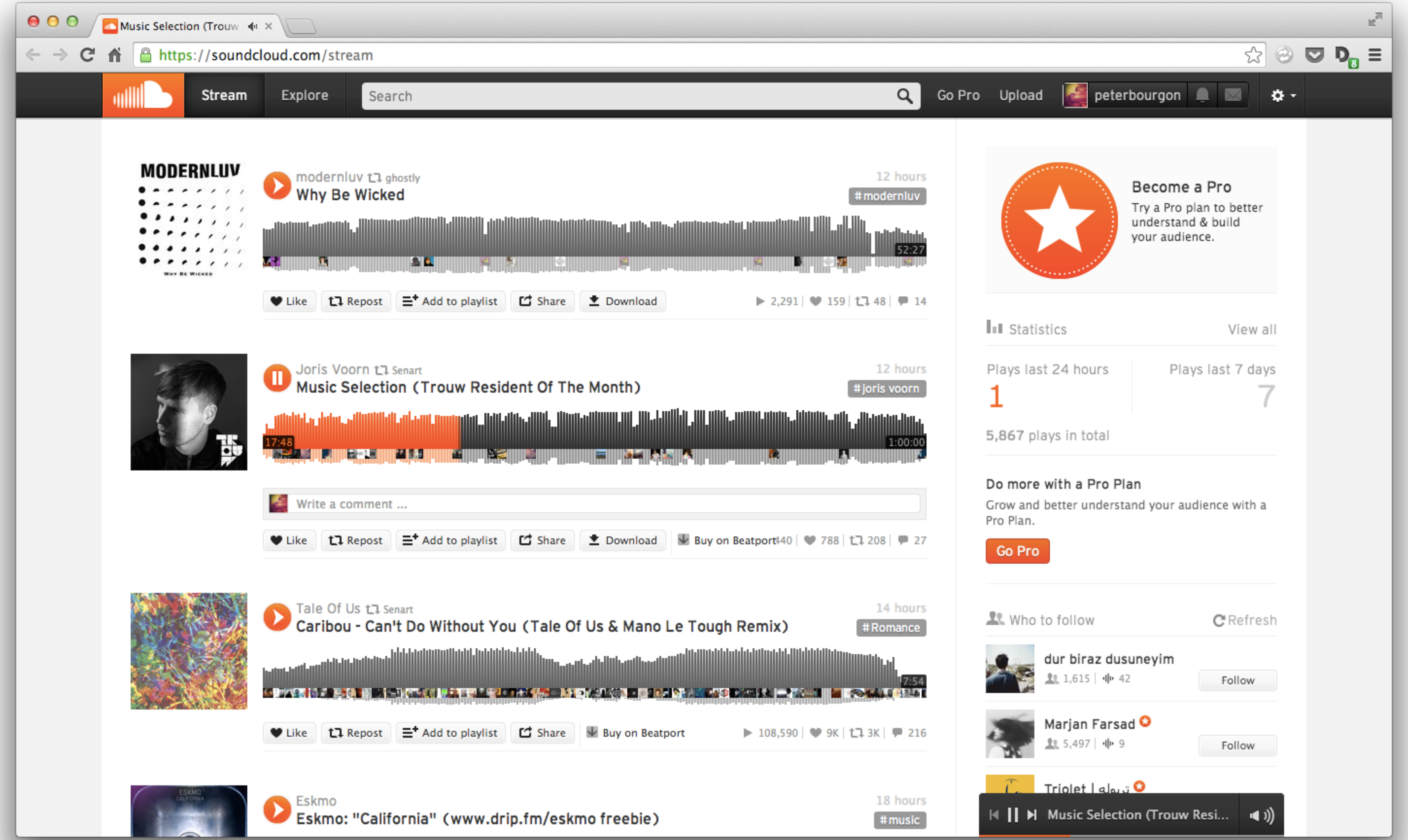
{ 123, 456 }

{ 123, 456 }

{ 123, 456 }

Interlude — Bending the problem

CRDTs in production



Music Selection (Trouw

https://soundcloud.com/stream

Stream · Explore · Search · Go Pro · Upload · peterbourgon

moderniuv ghostly
Why Be Wicked
12 hours
#modernluv
52:27
Like · Repost · Add to playlist · Share · Download
2,291 · 159 · 48 · 14

Joris Voorn Senart
Music Selection (Trouw Resident Of The Month)
12 hours
#joris voorn
17:48 · 1:00:00
Write a comment ...
Like · Repost · Add to playlist · Share · Download · Buy on Beatport
140 · 788 · 208 · 27

Tale Of Us Senart
Caribou - Can't Do Without You (Tale Of Us & Mano Le Tough Remix)
14 hours
#Romance
7:54
Like · Repost · Add to playlist · Share · Buy on Beatport
108,590 · 9K · 3K · 216

Eskmo
Eskmo: "California" (www.drip.fm/eskmo freebie)
18 hours
#music

Become a Pro
Try a Pro plan to better understand & build your audience.

Statistics · View all
Plays last 24 hours: 1
Plays last 7 days: 7
5,867 plays in total

Do more with a Pro Plan
Grow and better understand your audience with a Pro Plan.
Go Pro

Who to follow · Refresh
dur biraz dusuneyim · 1,615 · 42 · Follow
Marjan Farsad · 5,497 · 9 · Follow
Triolet | تریوله

Music Selection (Trouw Resi...

Event

Timestamp

Actor

Verb

Thing

Event

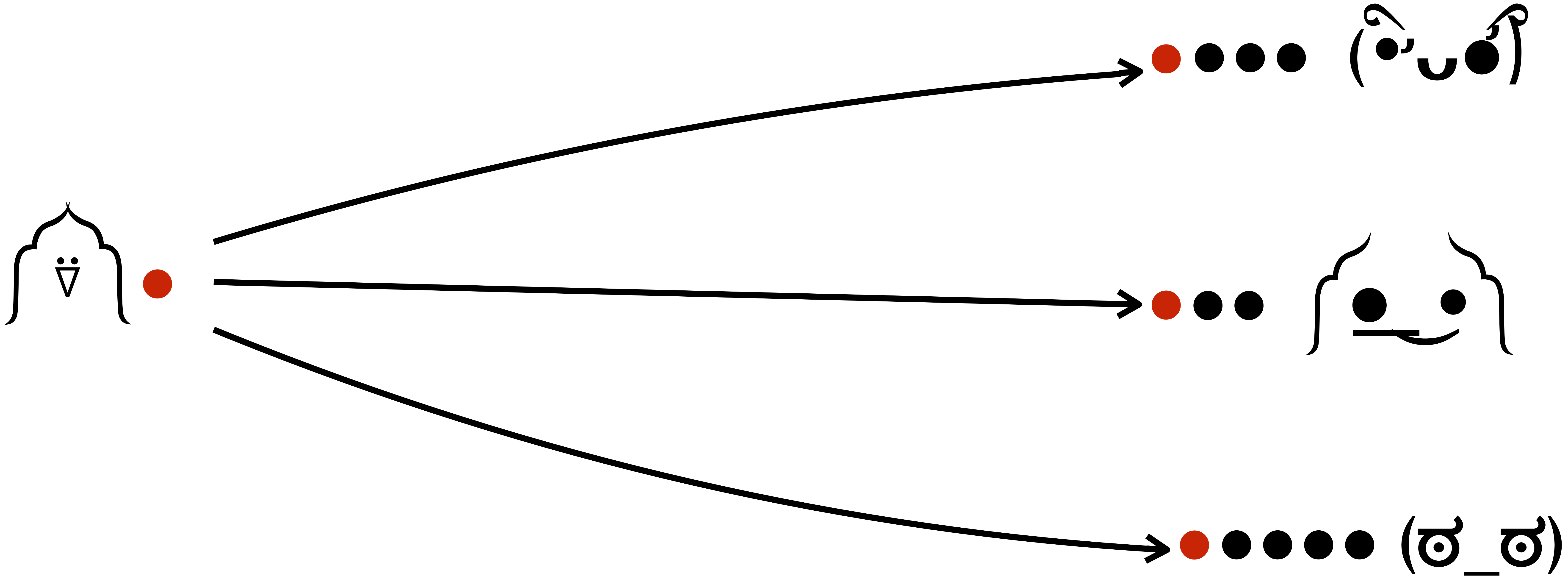
2014-04-01T15:16:17.187Z

snoopdogg

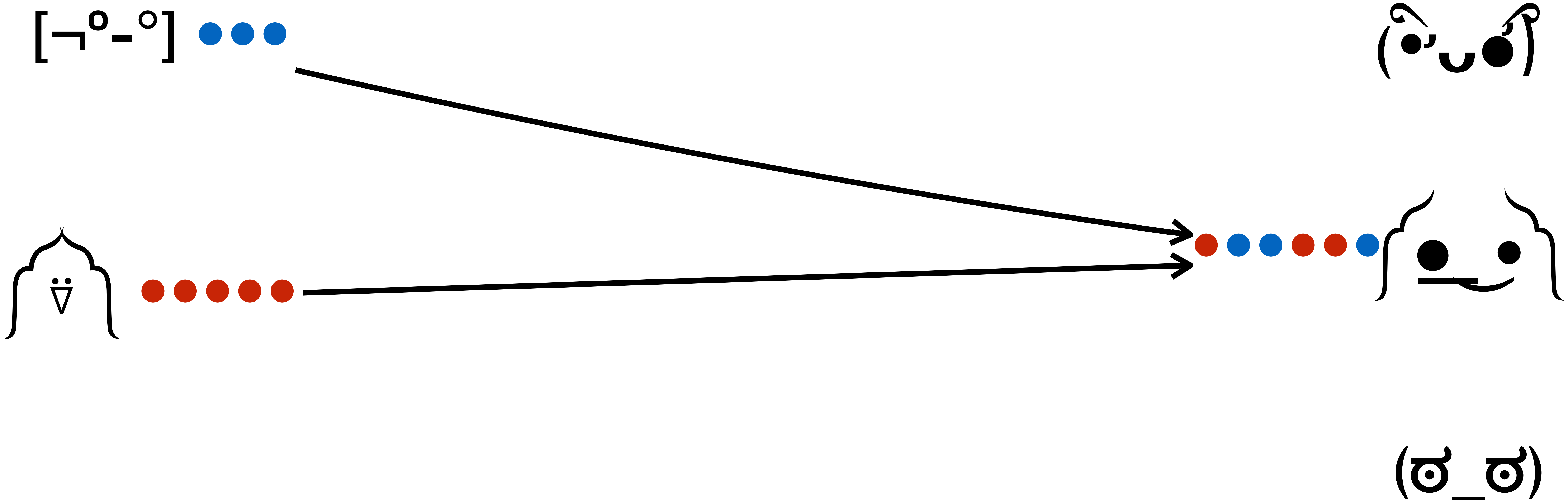
reposted

theeconomist/election-day

Fan out on write



Fan in on read



Unique events — use a set

G-set — can’t delete

2P-set — add, remove once

OR-set — storage overhead

A wild set appears

Roshi set

S+ { A/1 B/2 C/3 }

S– { D/4 }

S { A B C }

Roshi set

S = actor’s outbox key

snoopdogg·outbox

A/B/C/D = actor+verb+thing

snoopdogg·repost·theeconomist/election-day

1/2/3 = timestamp

2014-04-01T15:16:17.187Z

Reading is easy

Writing is interesting

Insert

- If either *key*+ or *key*– already contains *element*, and the existing score >= *score*, **no-op and exit**.

- Insert (*element*, *score*) into add set *key*+.

- Delete (*element*) from remove set *key*–.

Delete

- If either *key*+ or *key*– already contains *element*, and the existing score >= *score*,
  no-op and exit.

- Insert (*element*, *score*) into add set *key*–.

- Delete (*element*) from remove set *key+*.

Example

S+	{ A/1 B/2 }
S–	{ C/3 }

Insert D/4

S+	{ A/1 B/2 }
S–	{ C/3 }

Insert D/4

S+	{ A/1 B/2 D/4 }
S–	{ C/3 }

S+ { A/1 B/2 D/4 }

S– { C/3 }

Insert D/4

S+	{ A/1 B/2 D/4 }
S–	{ C/3 }

Insert D/4

S+ { A/1 B/2 D/4 }

S– { C/3 }

S+ { A/1 B/2 D/4 }

S– { C/3 }

Delete D/3

S+	{ A/1 B/2 D/4 }
S–	{ C/3 }

Delete D/3

S+ { A/1 B/2 D/4 }

S– { C/3 }

S+ { A/1 B/2 D/4 }

S– { C/3 }

Delete D/5

S+ { A/1 B/2 D/4 }

S– { C/3 }

Delete D/5

S+ { A/1 B/2 ~~D/4~~ }

S– { C/3 D/5 }

S+	{ A/1 B/2 }
S–	{ C/3 D/5 }

Delete D/6

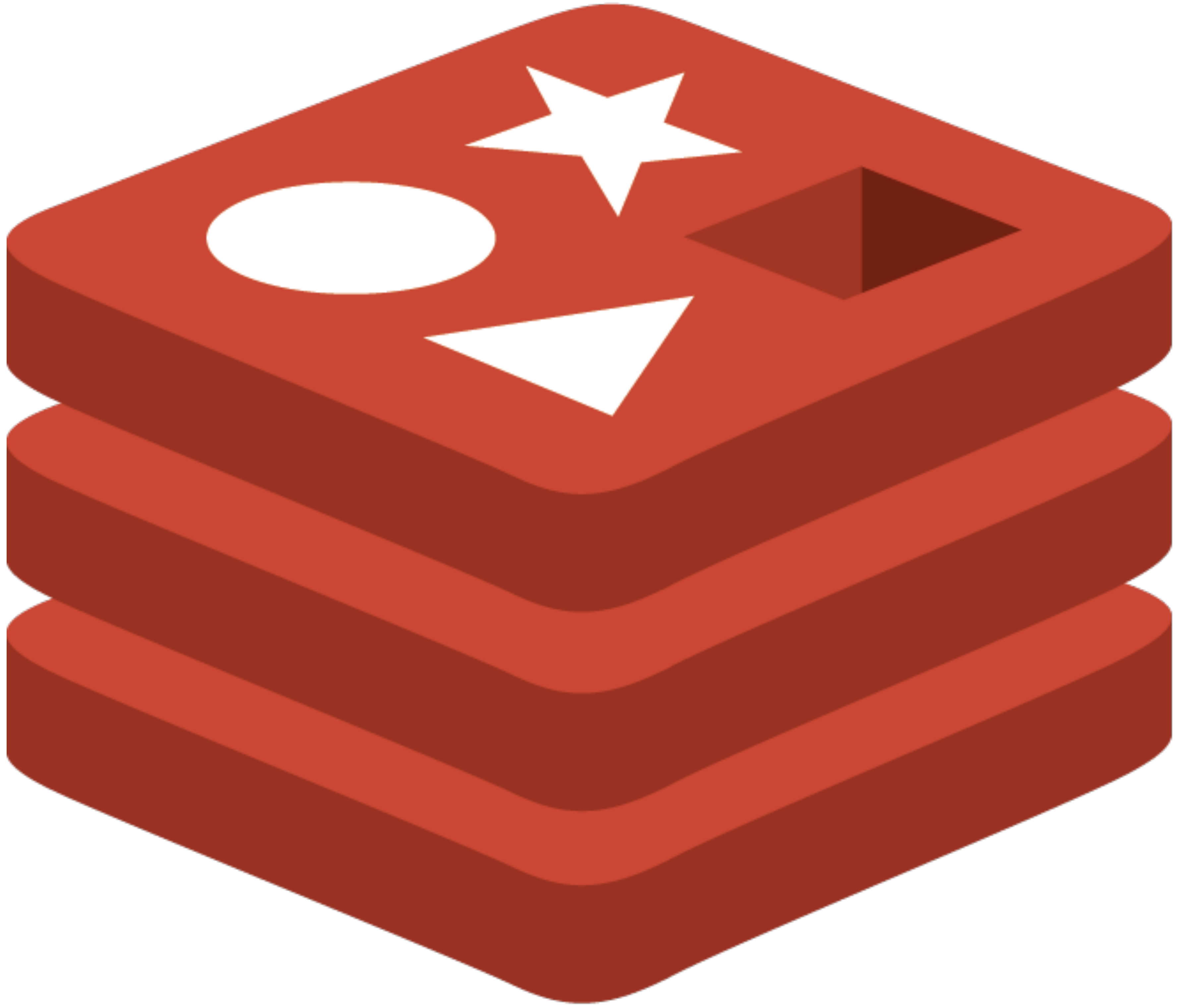
S+	{ A/1 B/2 }
S–	{ C/3 D/5 }

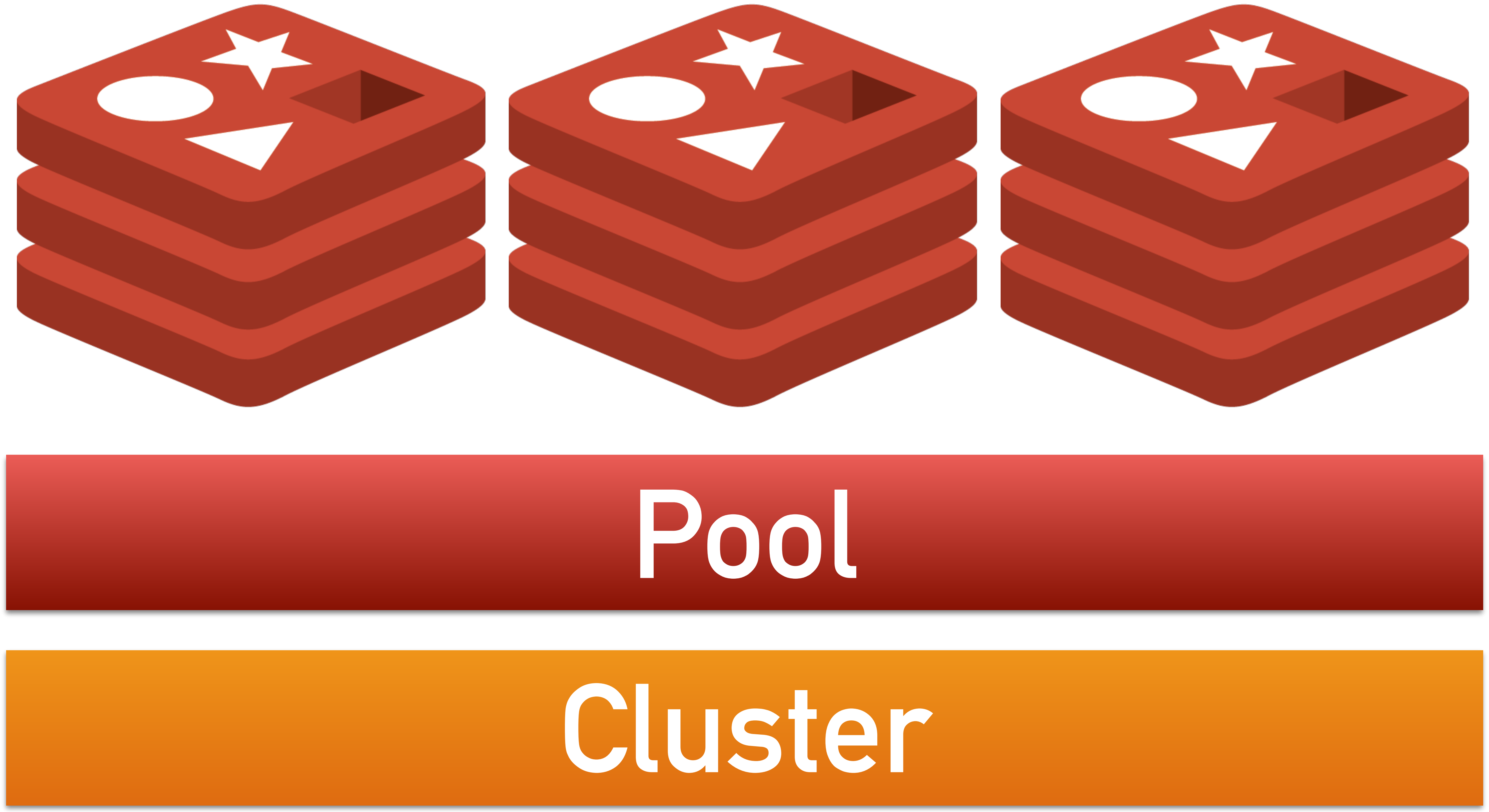
Delete D/6

S+	{ A/1 B/2 }
S–	{ C/3 D/6 }

S+	{ A/1 B/2 }
S–	{ C/3 D/6 }

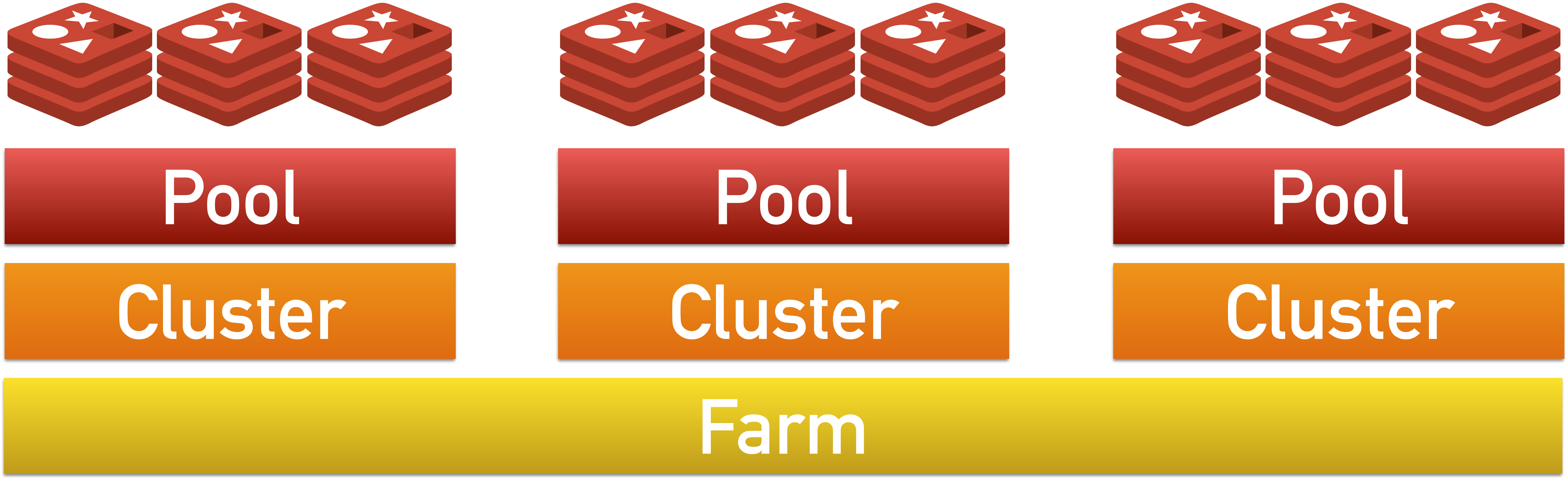
Making it real





Pool

Cluster



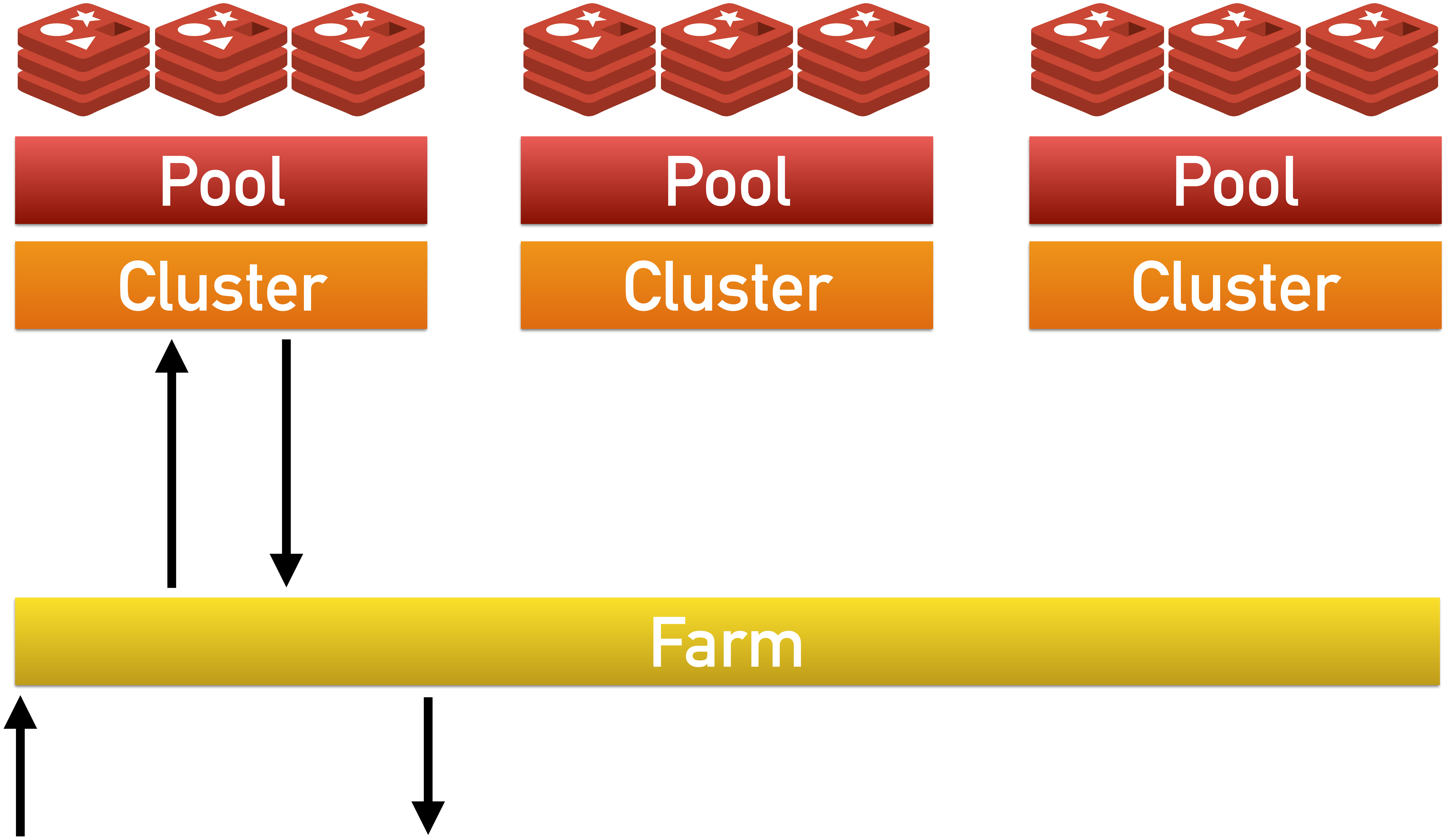
Pool Pool Pool

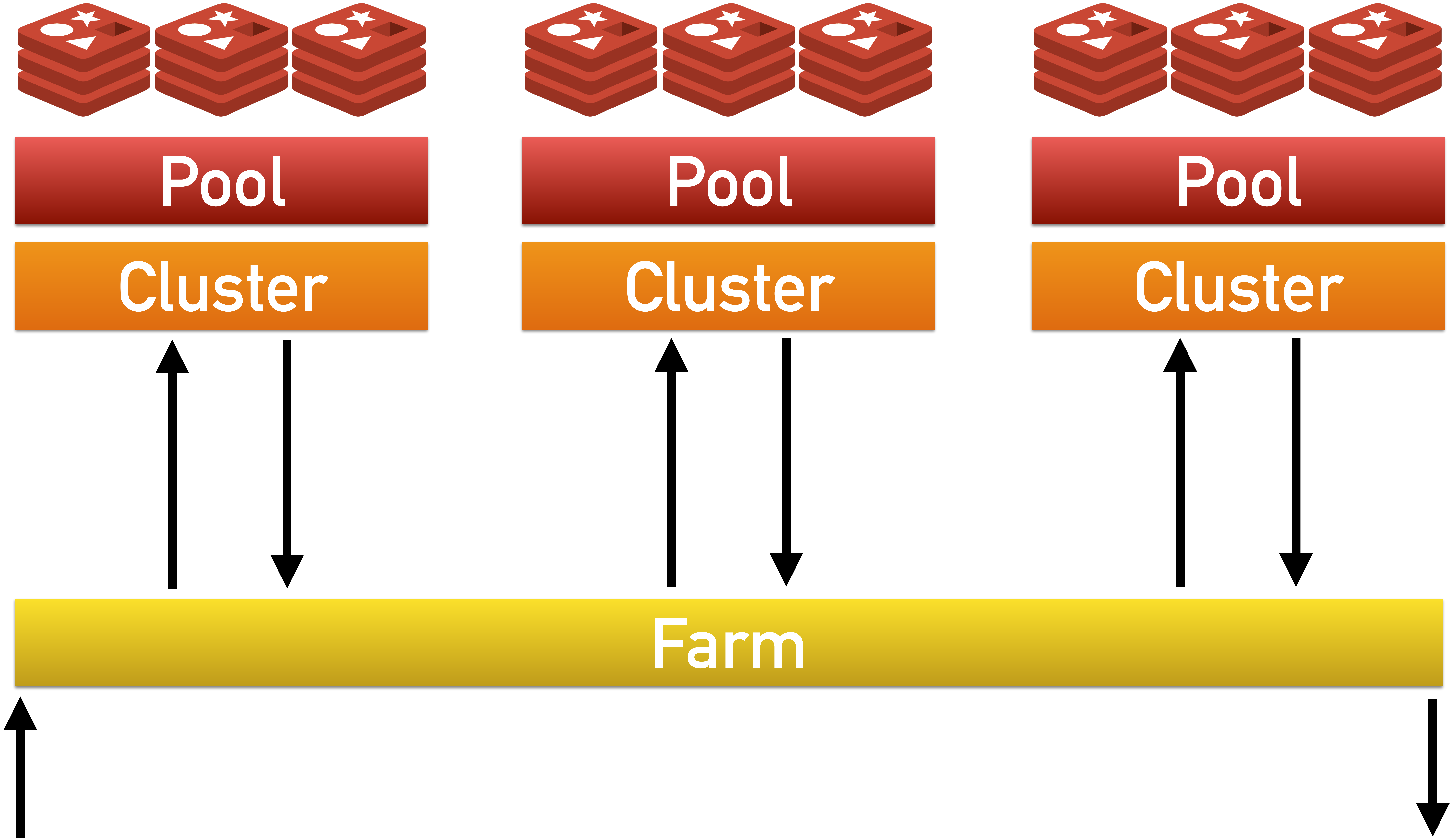
Cluster Cluster Cluster

Farm

Writing is easy

Reading is interesting


Pool
Cluster
Pool
Cluster
Pool
Cluster
Farm



Pool

Cluster

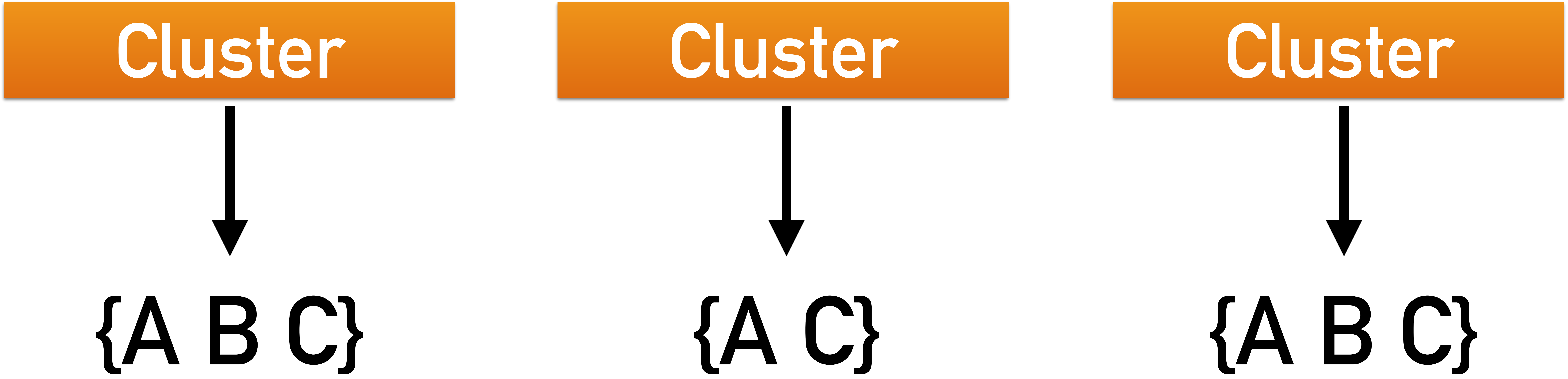
Pool

Cluster

Pool

Cluster

Farm



Cluster

Cluster

Cluster

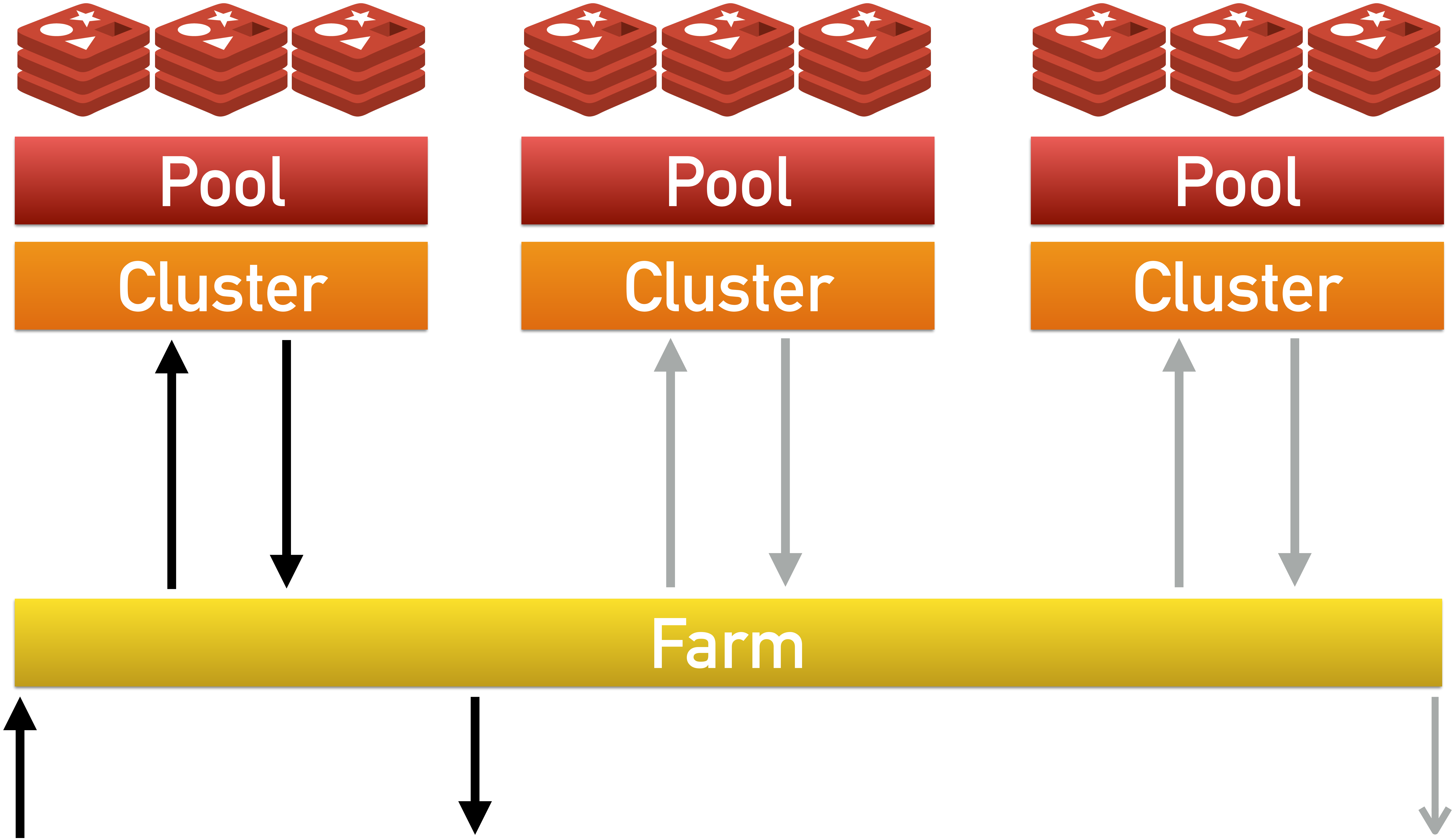
{A B C}

{A C}

{A B C}

$$\cup = \{A\ B\ C\}$$

$$\Delta = \{B\}$$


Pool
Pool
Pool
Cluster
Cluster
Cluster
Farm

github.com/soundcloud/roshi

In conclusion,

Consistency without consensus = CRDT.

Embrace your invariants.

Maybe bend your problem, not your solution.

Thanks!

☞ soundcloud.com/jobs ☜

@peterbourgon