How to determine which algorithm, index, data Layouts is better?

What we considered:

amount of data read and written

random I/O vs sequential

What we ignored:

DB buffer scheduling strategy

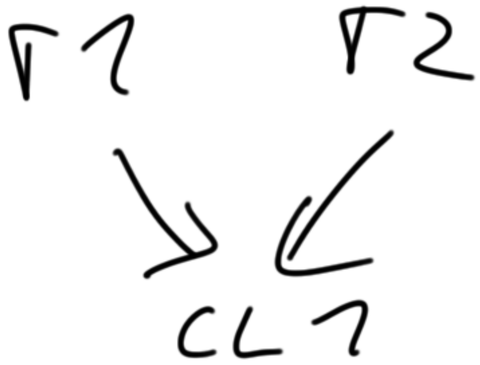
L1, L2, L<x> cache misses

different bandwidth among CPU, caches, and main memory

CPU costs

multiple cores

multithreading



What we ignored:

DB buffer scheduling strategy

L1, L2, L<x> cache misses

different bandwidth among CPU, caches, and main memory

CPU costs

multiple cores

multithreading

hyperthreading

8 = 4 + 4

4 physical cores

4 hyperthread

What we ignored:

DB buffer scheduling strategy

L1, L2, L<x> cache misses

different bandwidth among CPU, caches, and main memory

CPU costs

multiple cores

multithreading

hyperthreading

scheduling effects

multiple disks/RAID level

data distributions/data skew

uniform → 42 sec

skewed → 720 sec

What we ignored:

DB buffer scheduling strategy

L1, L2, L<x> cache misses

different bandwidth among CPU, caches, and main memory

CPU costs

multiple cores

multithreading

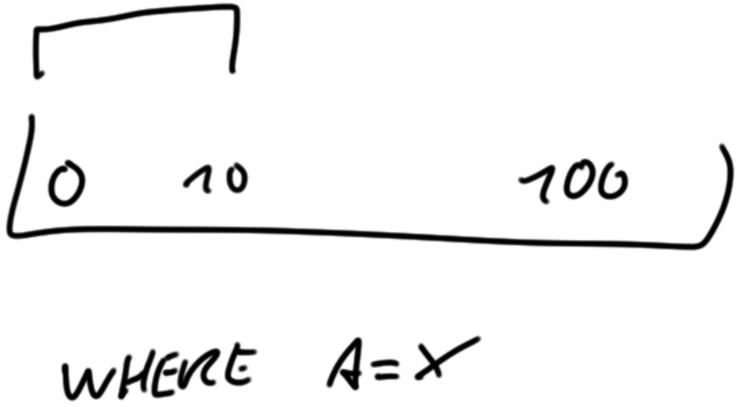
hyperthreading

scheduling effects

multiple disks/RAID level

data distributions/data skew

workload distribution/workload skew



What we ignored:

DB buffer scheduling strategy

L1, L2, L<x> cache misses

different bandwidth among CPU, caches, and main memory

CPU costs

multiple cores

multithreading

hyperthreading

scheduling effects

multiple disks/RAID level

data distributions/data skew

workload distribution/workload skew

SIMD

...

How to Measure Performance?

Three ways to measure the performance of a computer program:

(1.) Analytical Modelling, O-Notation

→ mathematical model

How to Measure Performance?

Three ways to measure the performance of a computer program:

(1.) Analytical Modelling,

(2.) Simulation,

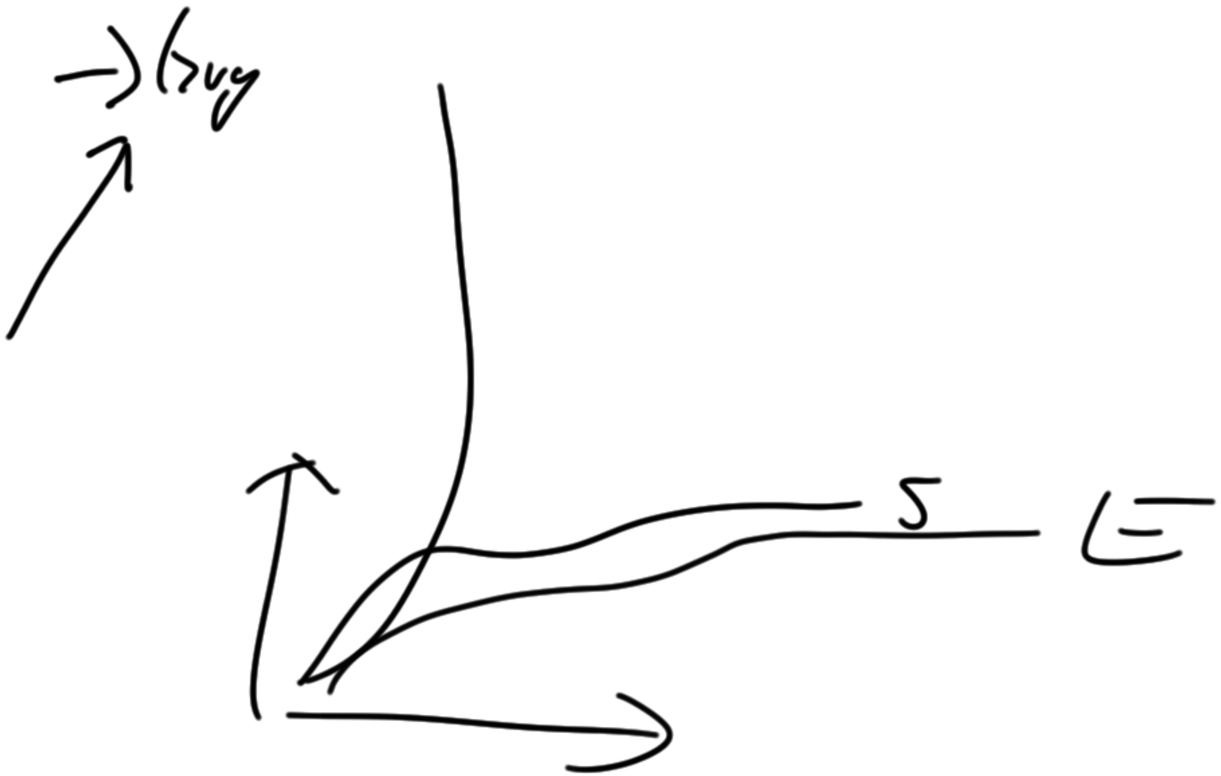
How to Measure Performance?

Three ways to measure the performance of a computer program:

(1.) Analytical Modelling,

(2.) Simulation,

(3.) Experiment.

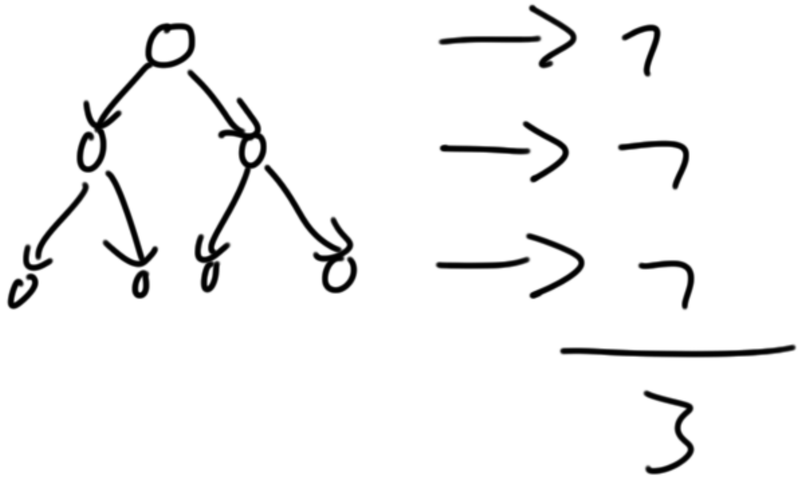


1. AND 2.

2. AND 3.

(1a) Asymptotic Complexity, O-Notation

count each primitive operation as “1“
(arithmetic operations, node hop, memory accesses, page accesses, etc.)



(1a) Asymptotic Complexity, O-Notation

count each primitive operation as “1“
(arithmetic operations, node hop, memory accesses, page accesses, etc.)

determine asymptotic complexity of the index

$$F(N) = N^2 + \underline{N \cdot \log N}$$

$$O(N^2)$$

(1a) Asymptotic Complexity, O-Notation

count each primitive operation as “1“
(arithmetic operations, node hop, memory accesses, page accesses, etc.)

determine asymptotic complexity of the index

worst, best and/or average case

example:

AVL-tree search has O(log n) average and worst case *time*

model time ≠ reality time

(1a) Asymptotic Complexity, O-Notation

count each primitive operation as “1“
(arithmetic operations, node hop, memory accesses, page accesses, etc.)

determine asymptotic complexity of the index

worst, best and/or average case

example:

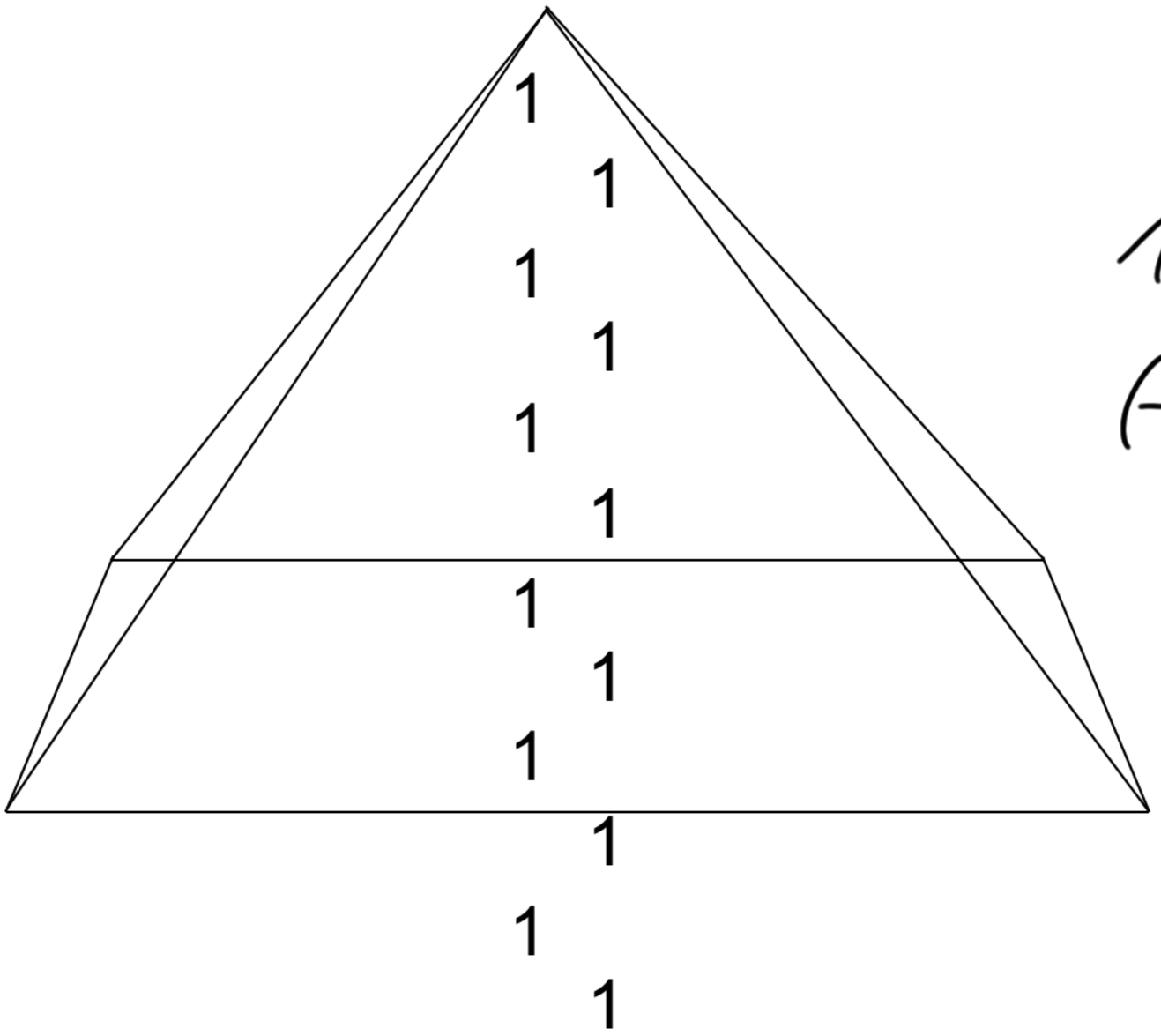
AVL-tree search has O(log n) average and worst case *time*

problem:

completely ignores memory hierarchy

Scenario

(1a) Asymptotic Complexity, O-Notation

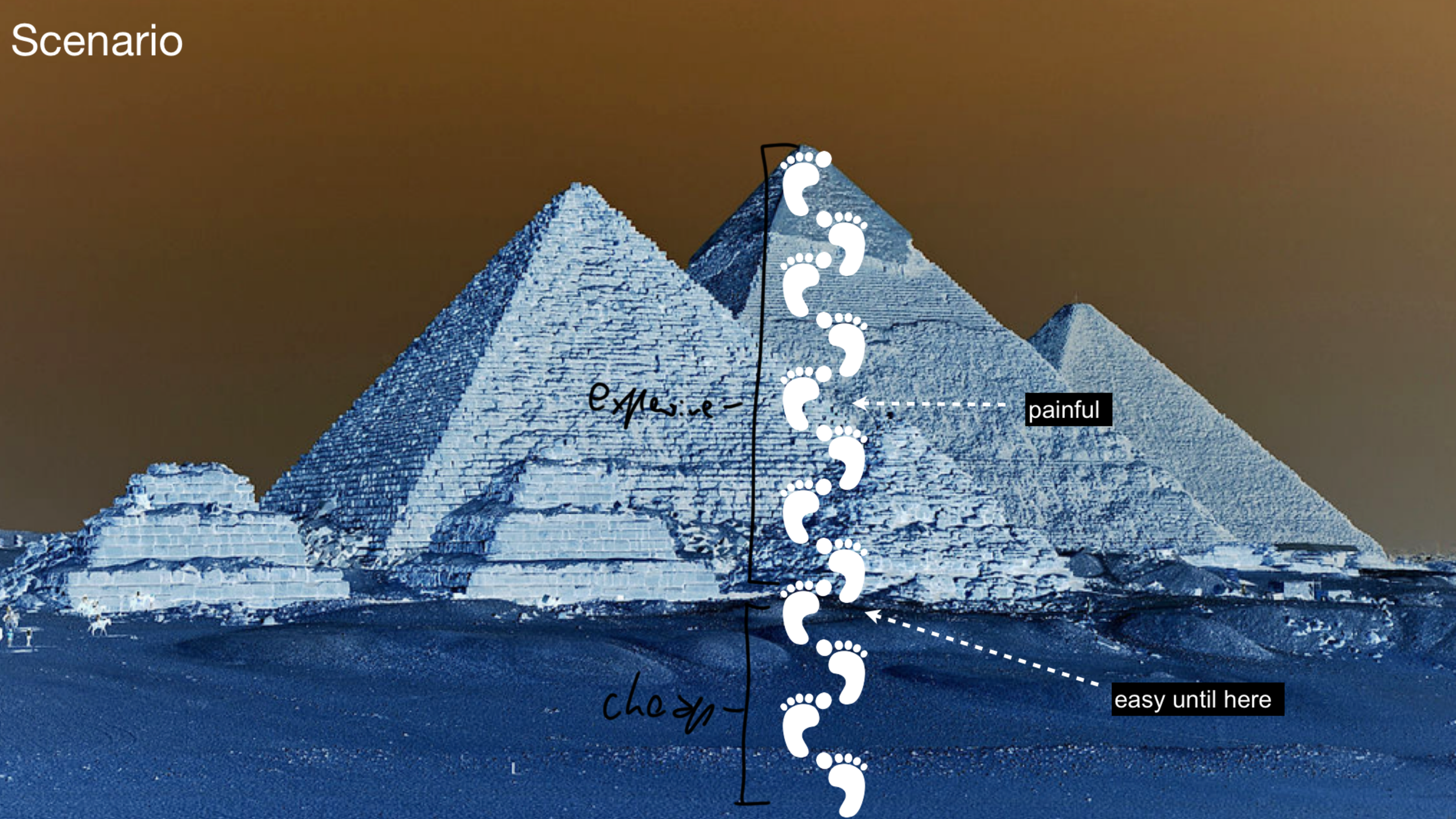


12 steps

$f(N) = N$

$O(N)$

Scenario



(1b.1) Cost Models

also consider number of random I/Os

example:

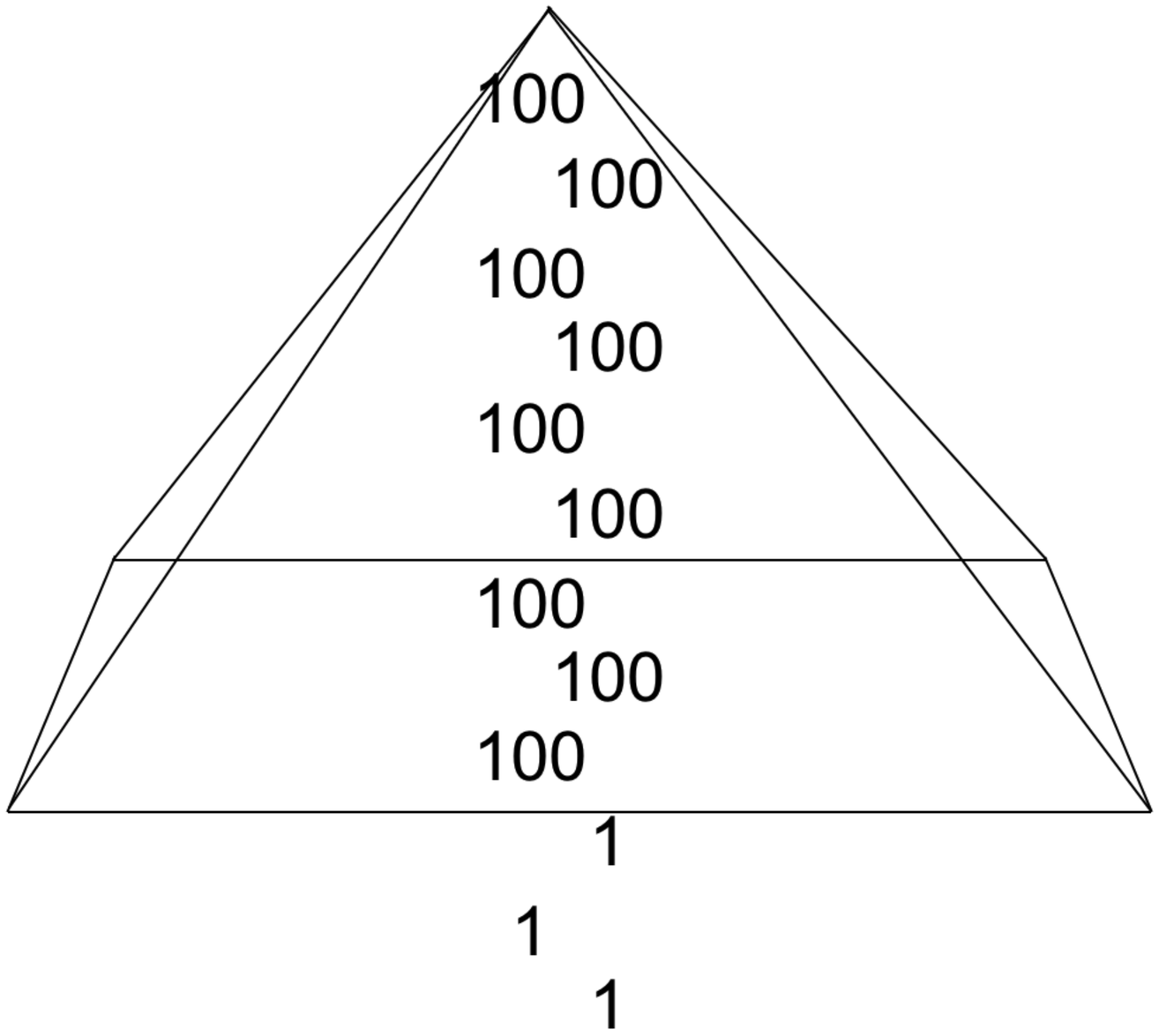
count “100“ for each random read or write of a page

count “1“ for each sequential read or write of a page

problem:

still ignores CPU time and cache misses...

(1b.1) Cost Models



903

(1b.2) Cost Models

count number of cache misses as well

example:

CSS-tree has $O(\log_m n)$ cache misses (m=fan-out)

→ B-Tree in main memory

(1b.2) Cost Models

count number of cache misses as well

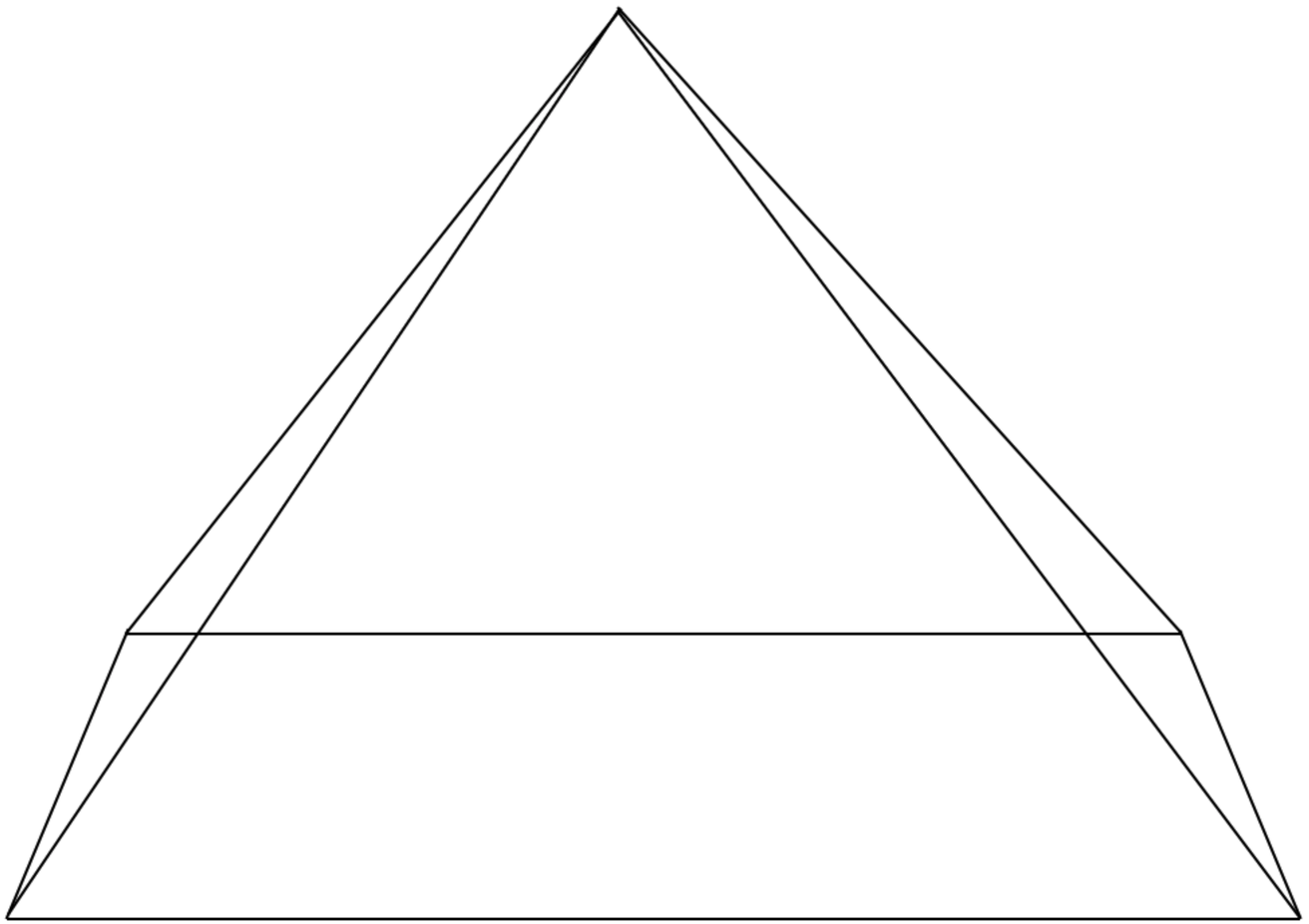
example:

CSS-tree has $O(\log_m n)$ cache misses (m=fan-out)

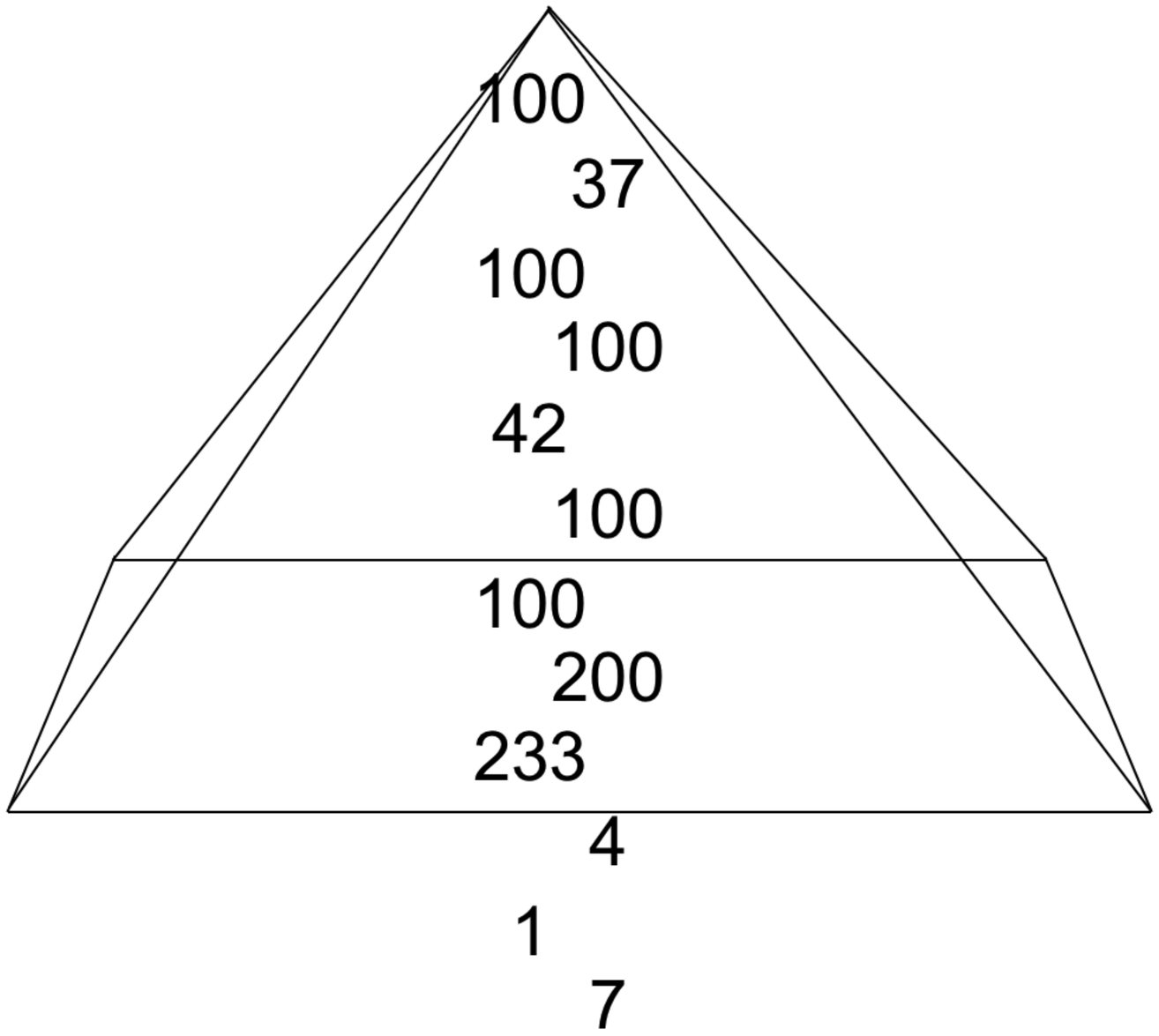
problem:

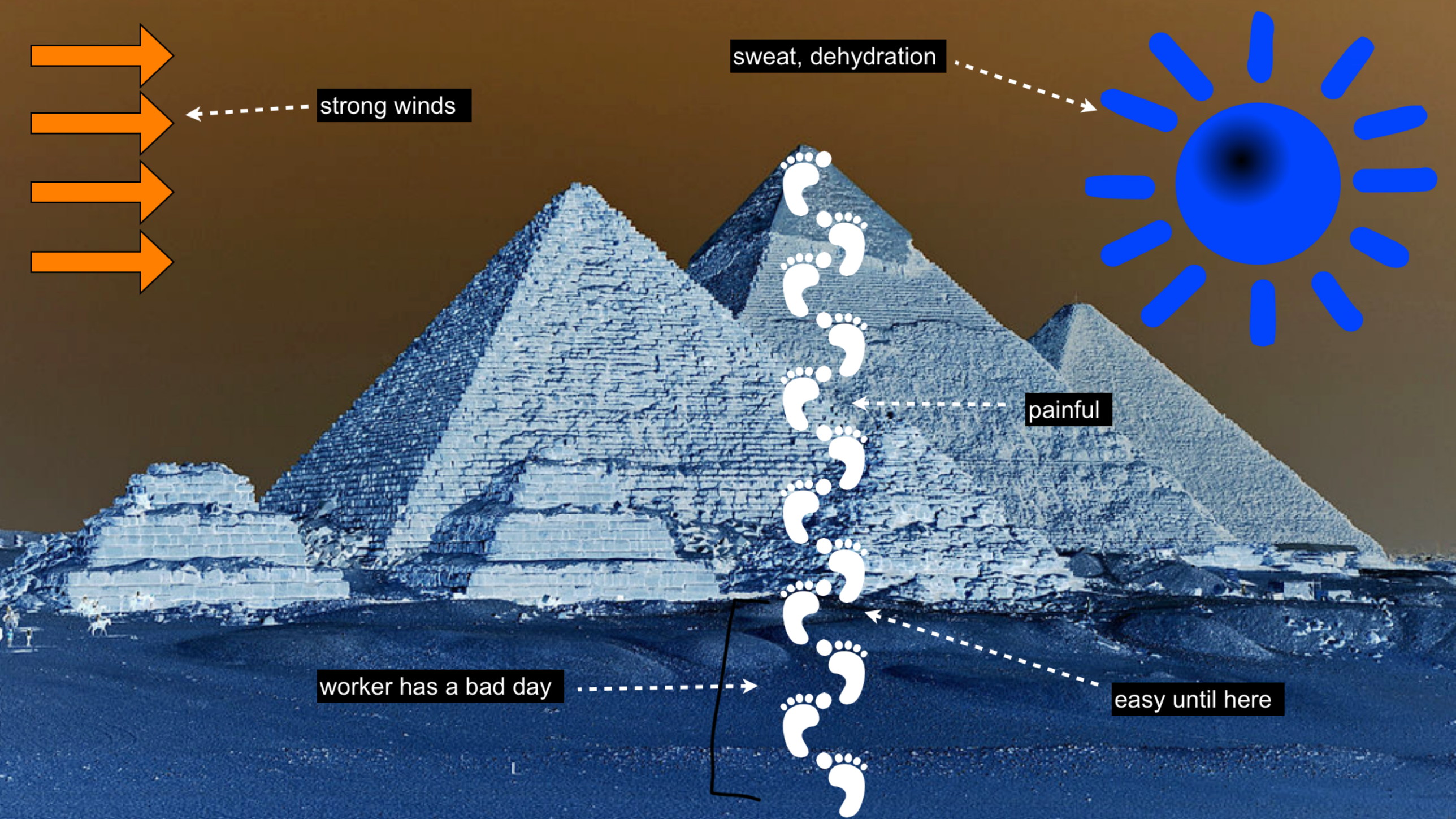
ignores CPU time...

(1b.2) Cost Models



(1b.2) Cost Models




strong winds
sweat, dehydration
painful
worker has a bad day
easy until here

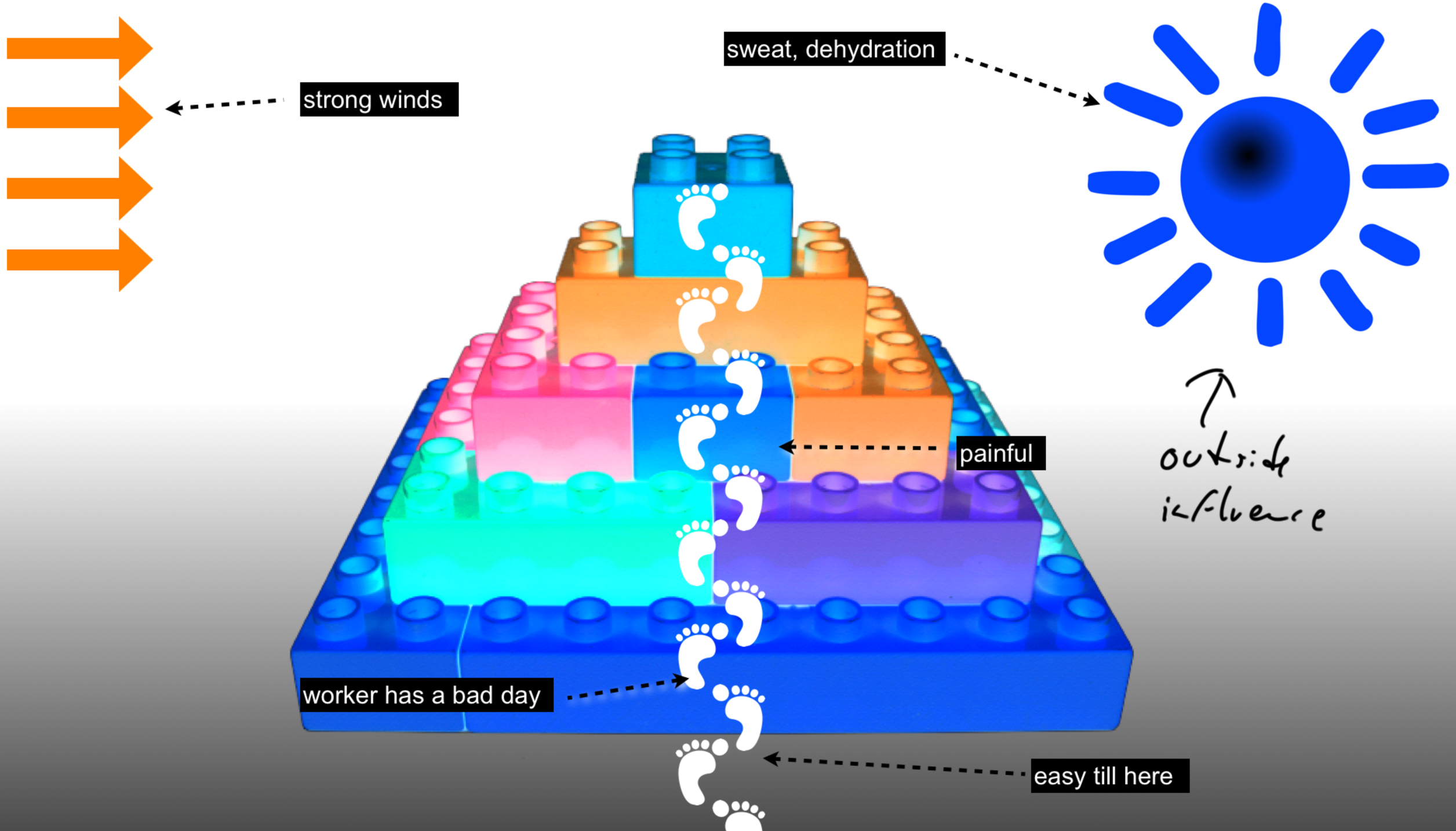
(2) Simulation

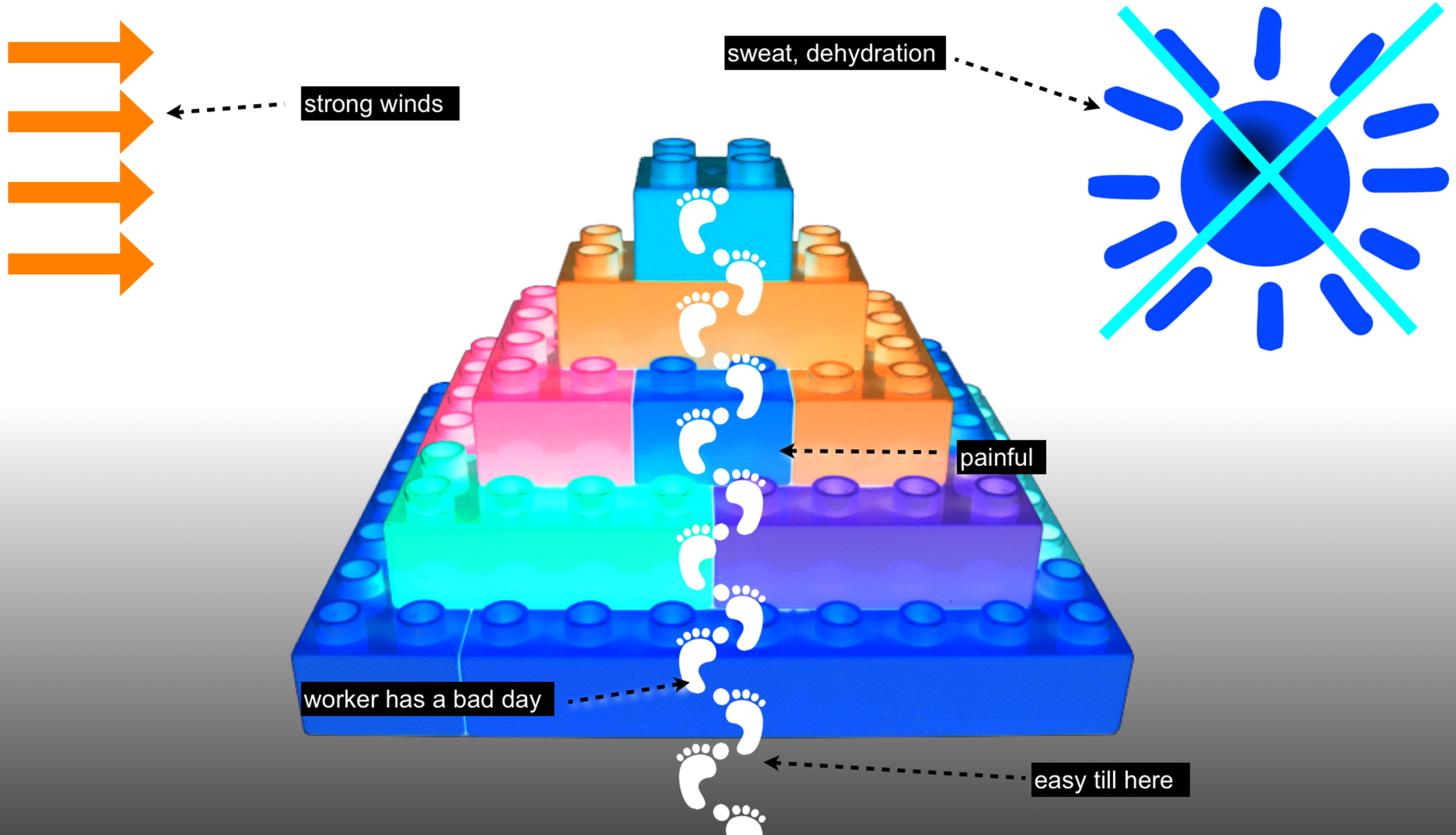
run part of the actual algorithm or system

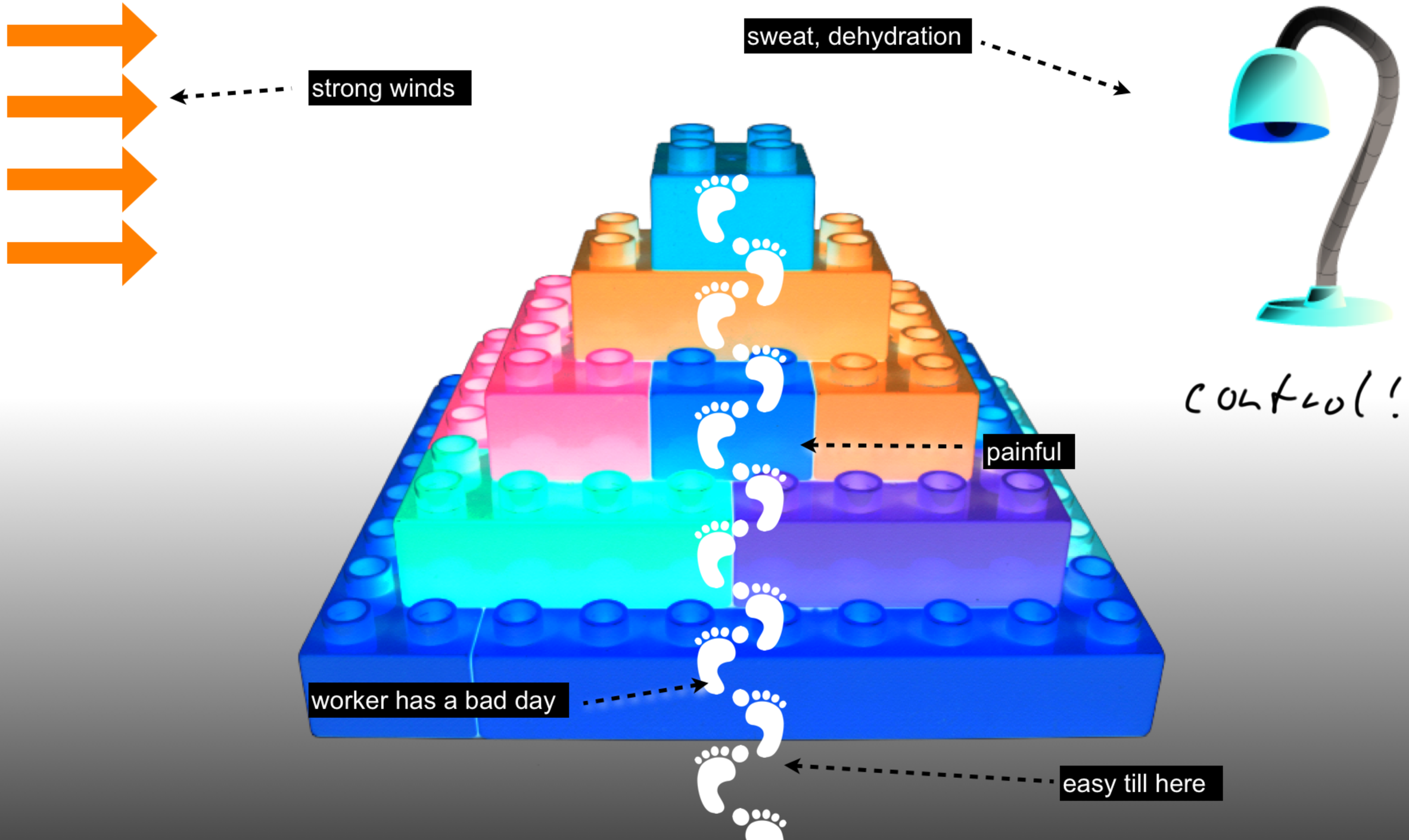
simulate other parts

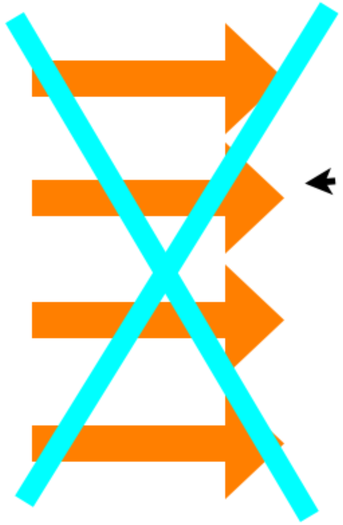
problem:

- might still oversimplify reality
- might miss some important effect from reality...


strong winds
sweat, dehydration
outside influence
painful
worker has a bad day
easy till here

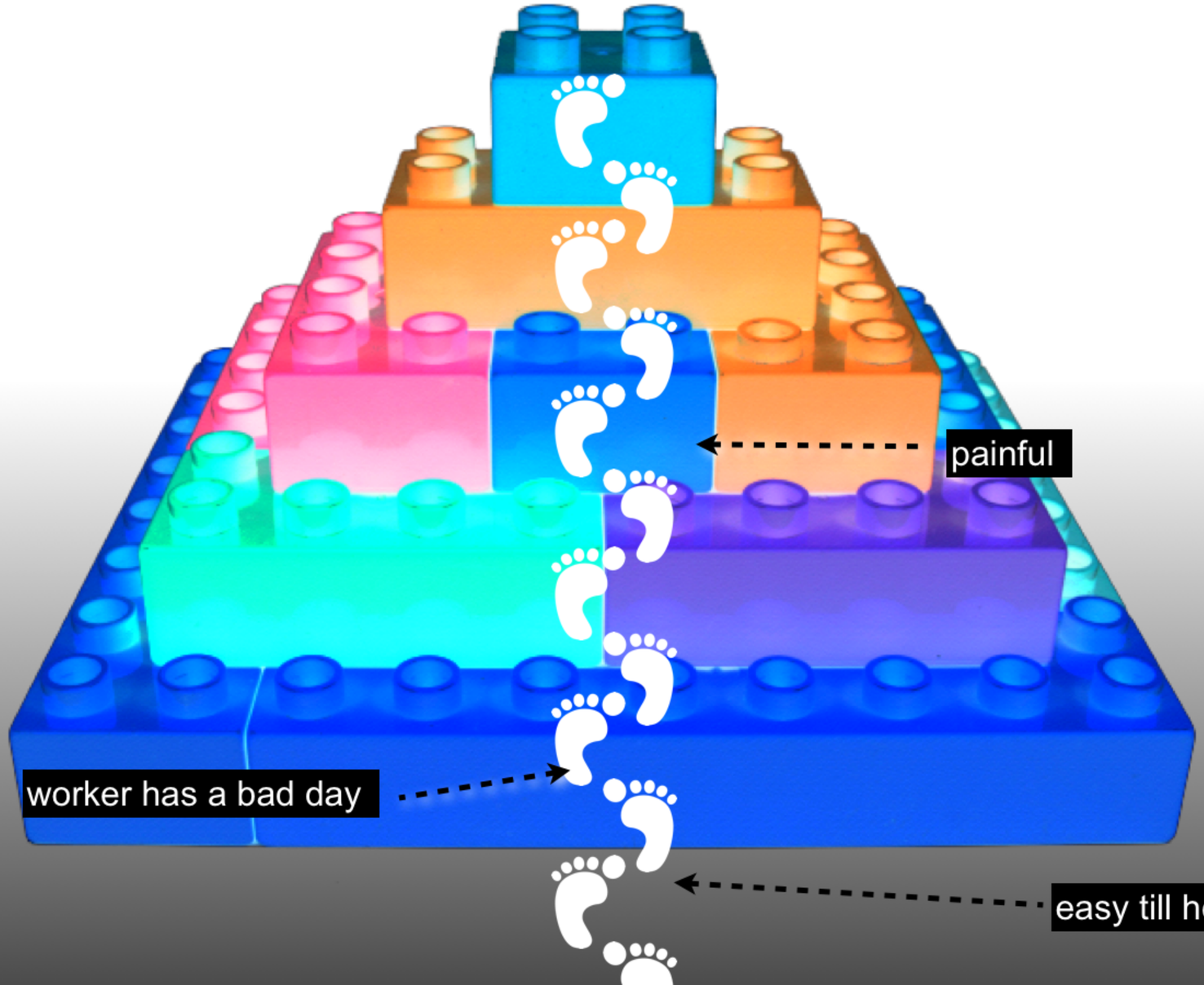

strong winds
sweat, dehydration
painful
worker has a bad day
easy till here


strong winds
sweat, dehydration
control!
painful
worker has a bad day
easy till here



strong winds

sweat, dehydration



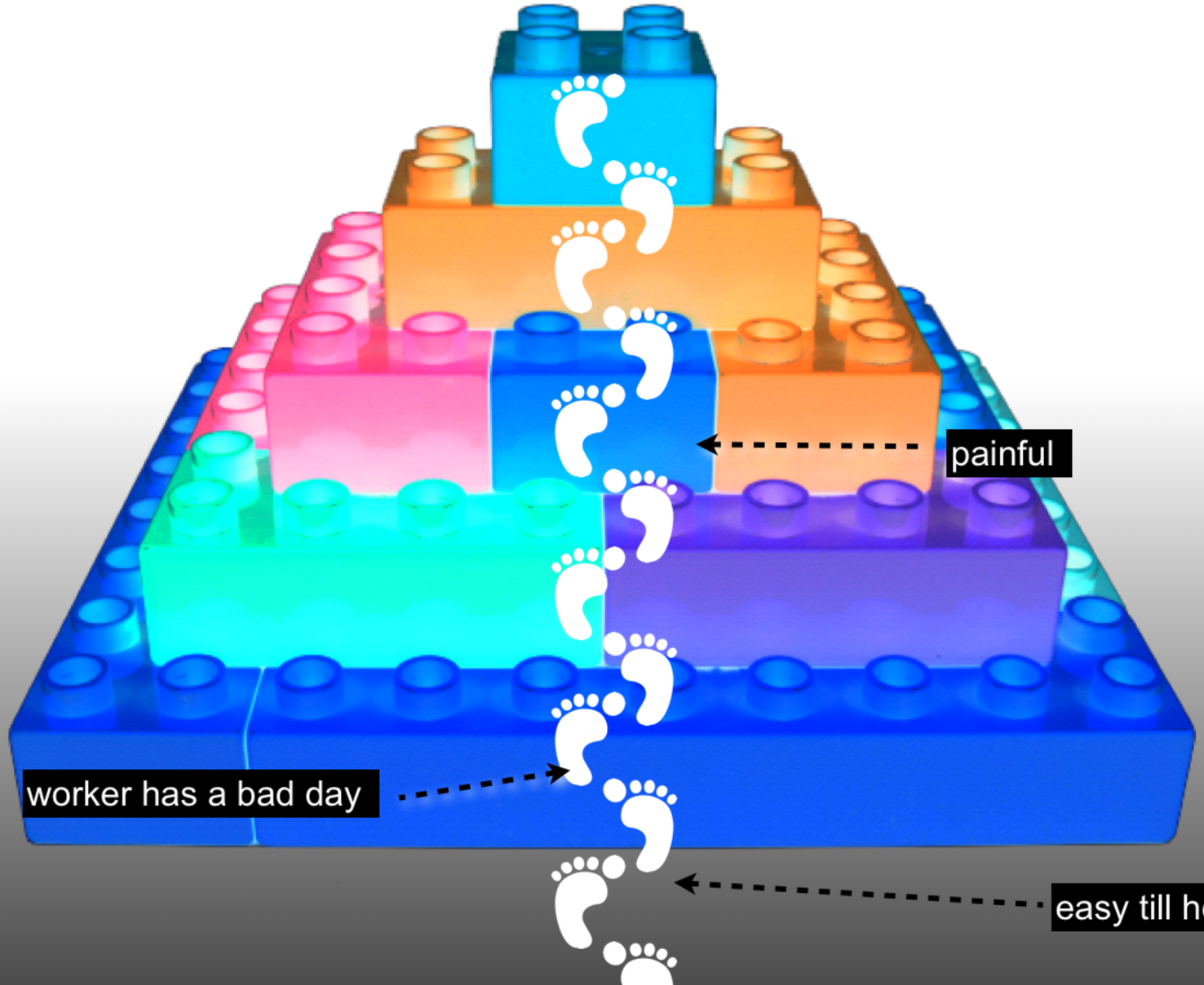
painful

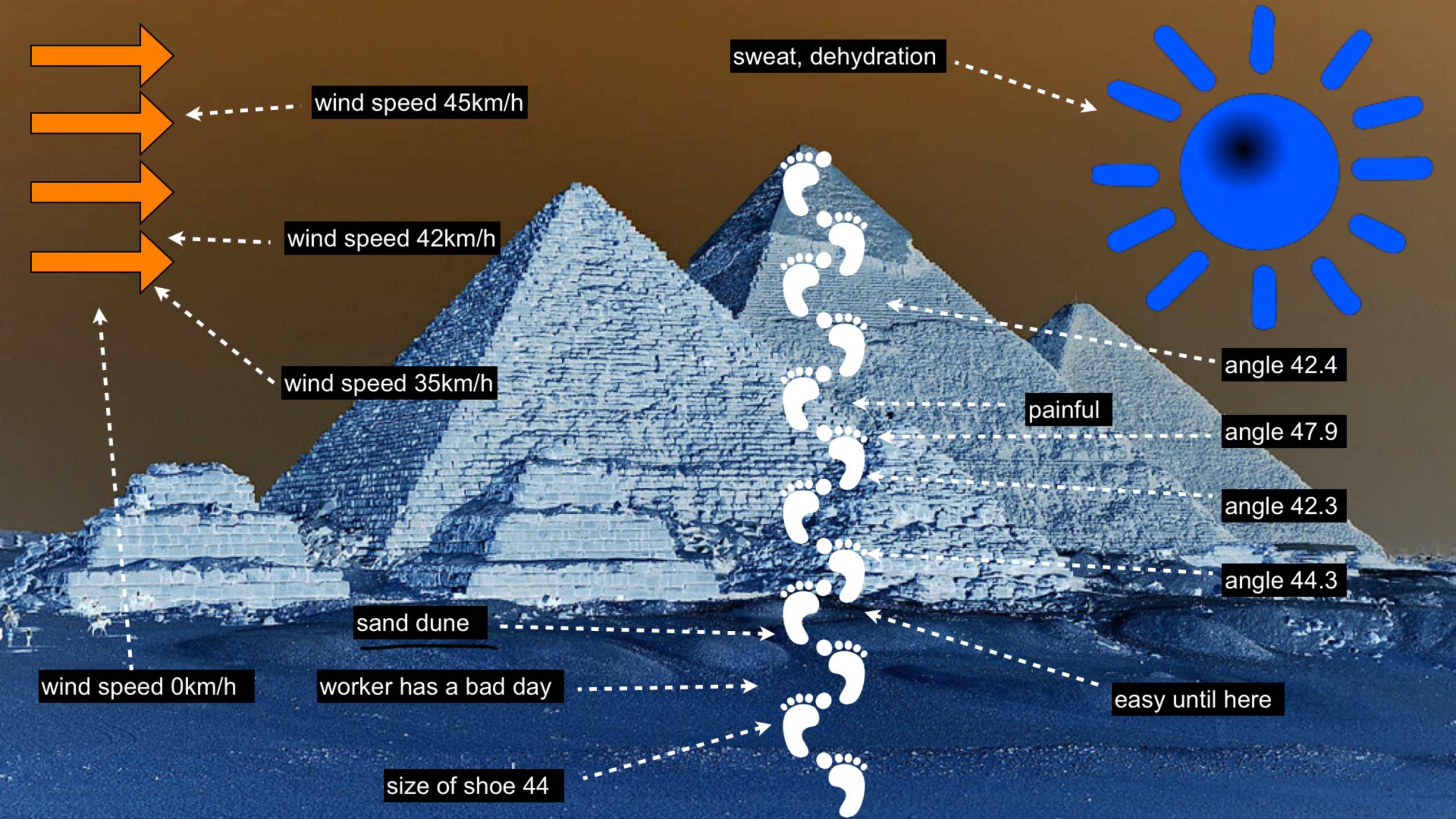
worker has a bad day

easy till here

strong winds

sweat, dehydration




wind speed 45km/h
wind speed 42km/h
wind speed 35km/h
wind speed 0km/h
sweat, dehydration
angle 42.4
painful
angle 47.9
angle 42.3
angle 44.3
sand dune
worker has a bad day
easy until here
size of shoe 44

(3) Experiment

implement it

run it

measure it

problems:

lacking abstraction → overfitting

(3) Experiment

implement it

run it

measure it

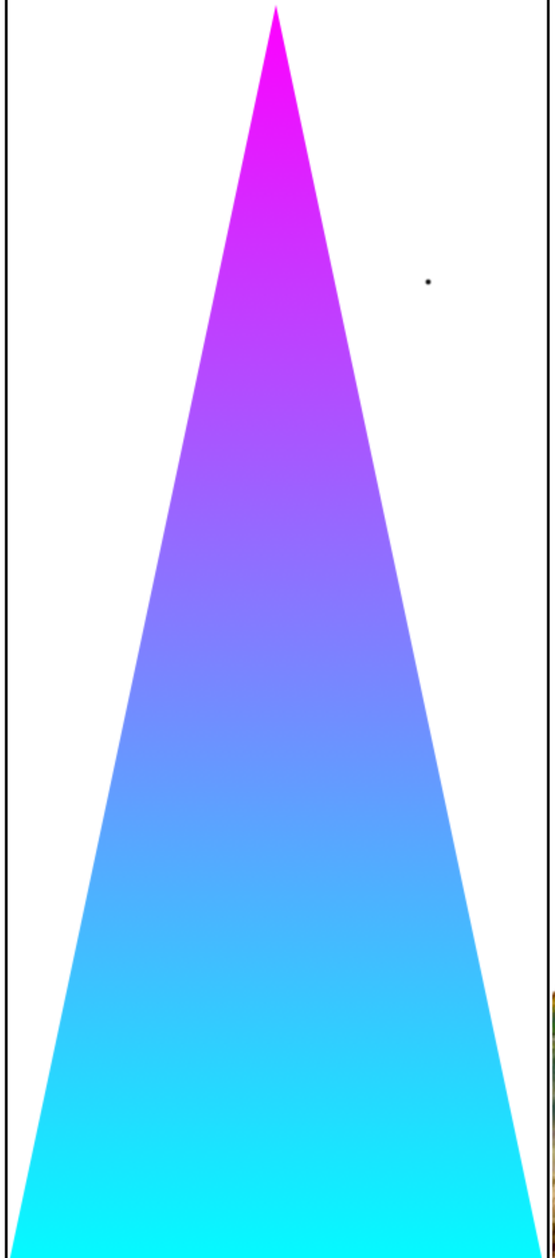
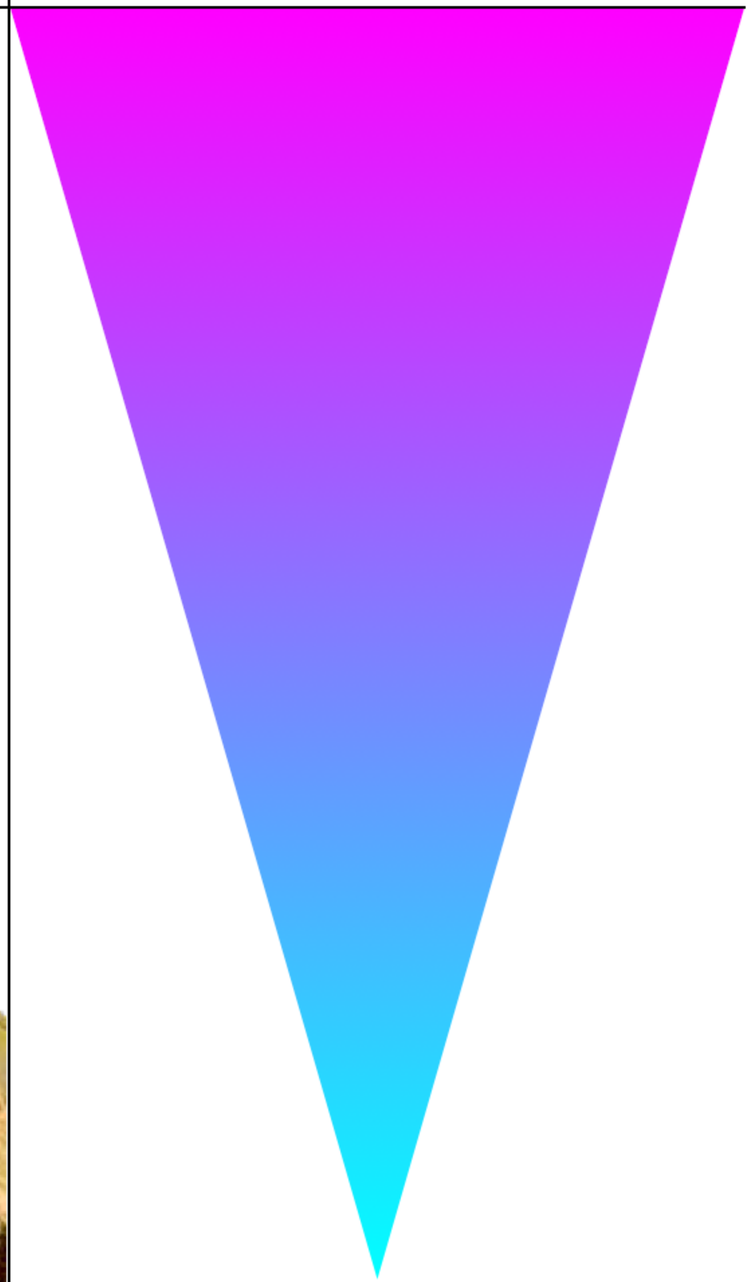
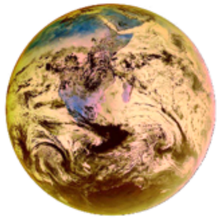
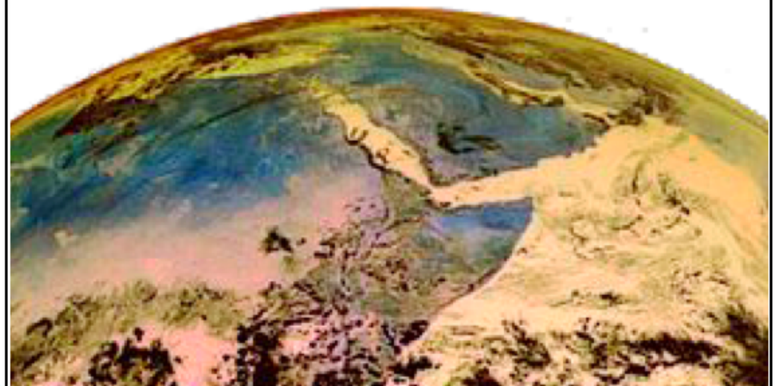
problems:

- lacking abstraction
- lacking bounds
- lacking theory
- (lacking insight)

End2End timer

42 sec

5 | 30 | 7

method	effort/cost	reality	generalizability
Analytical Modelling			
Simulation			
Experiment			

Copyrights and Credits