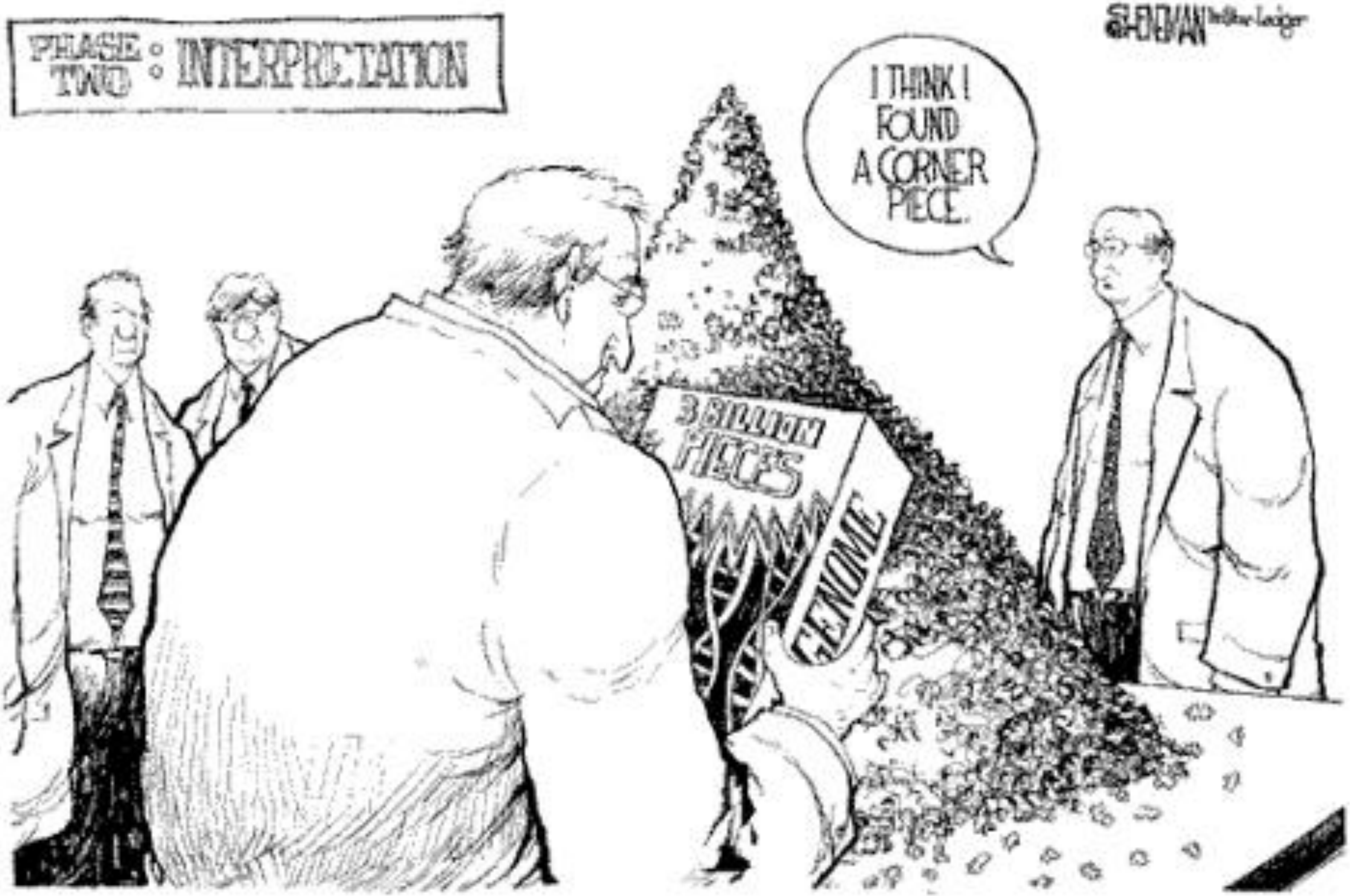
CS342: Bioinformatics
Assembling a Genome

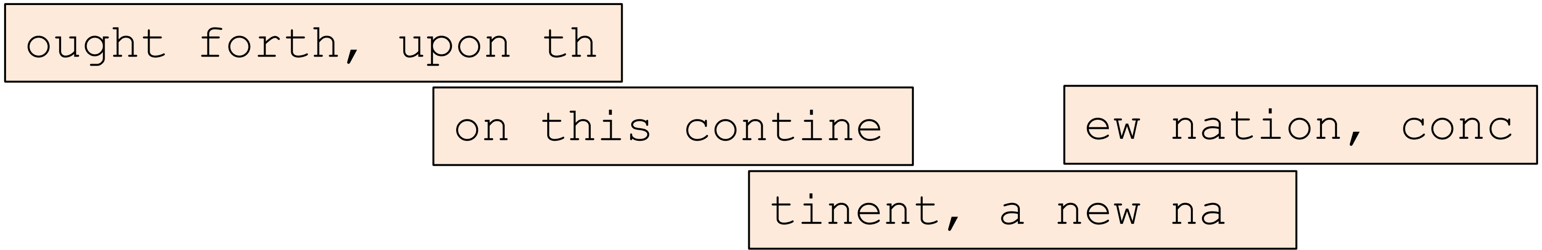

PHASE TWO : INTERPRETATION
SHENEMAN The Star-Ledger
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

Shortest Common Superstring Problem: Example

Set of strings: `{ “on this contine”, “hers brought for”, “r score and sev”, “ears ago our father”, “ew nation, conc”, “ought forth, upon th”, “four sco”, “tinent, a new na”, “and seven years” }`

Concatenation superstring: **“**`on this continehers brought forr score and sevears ago our fatherew nation, concought forth, upon thfour scotinent, a new naand seven years”`

Shortest superstring: **“**`four score and seven years ago our fathers brought forth, upon this continent, a new nation, conc”`



ought forth, upon th

on this contine

ew nation, conc

tinent, a new na

Shortest Common Superstring: Greedy

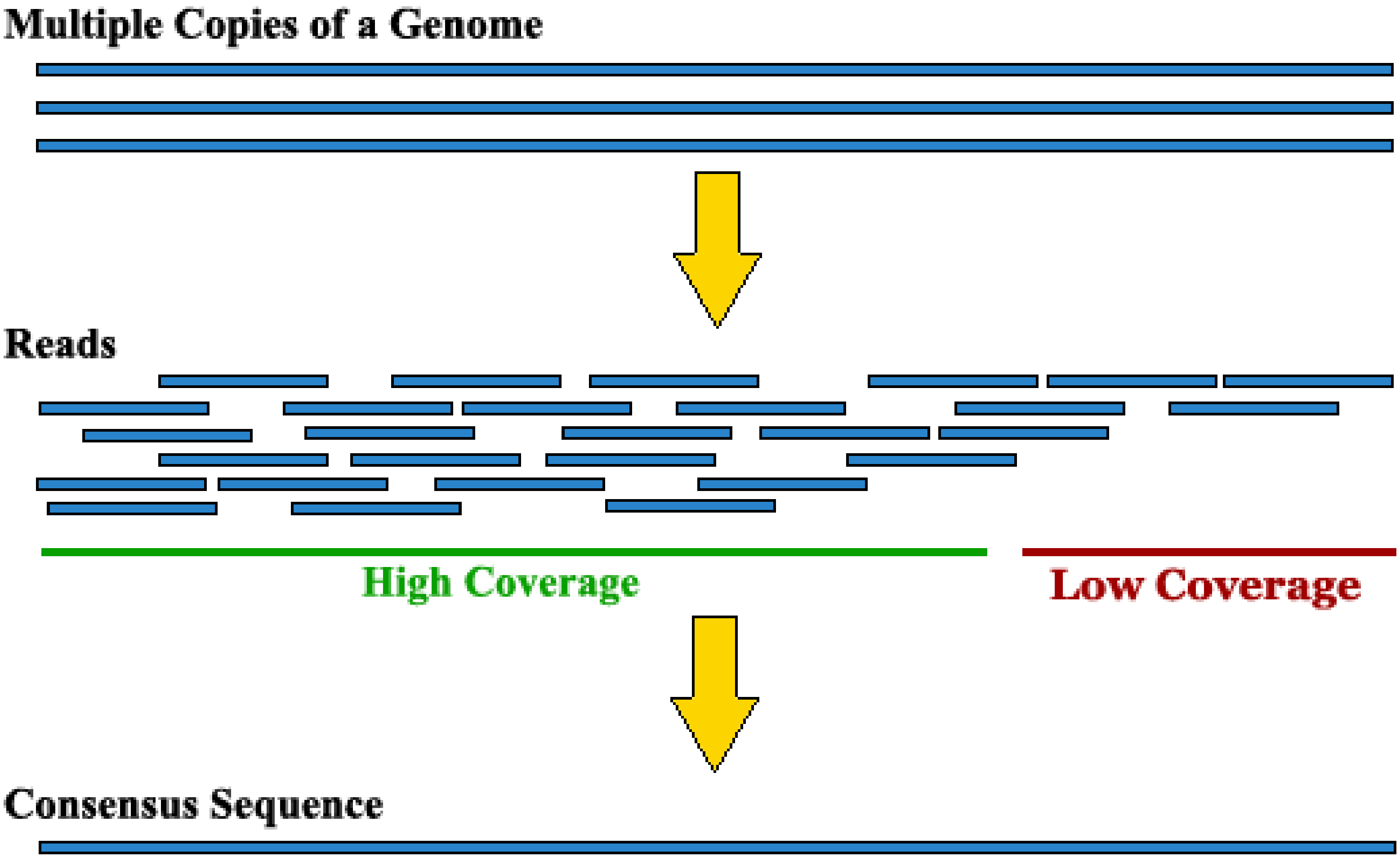
Greedy-SCS assembling all substrings of length 6 (so, all 6-mers) from:

```
a_long_long_long_time
```

```
ng_lon _long_ a_long long_l ong_ti ong_lo long_t g_long g_time ng_tim
ng_time ng_lon _long_ a_long long_l ong_ti ong_lo long_t g_long
ng_time g_long_ ng_lon a_long long_l ong_ti ong_lo long_t
ng_time long_ti g_long_ ng_lon a_long long_l ong_lo
ng_time ong_lon long_ti g_long_ a_long long_l
ong_lon long_time g_long_ a_long long_l
long_lon long_time g_long_ a_long
long_lon g_long_time a_long
long_long_time a_long
a_long_long_time
```

↑

Foiled by a repeat!!!



Multiple Copies of a Genome

Reads

High Coverage

Low Coverage

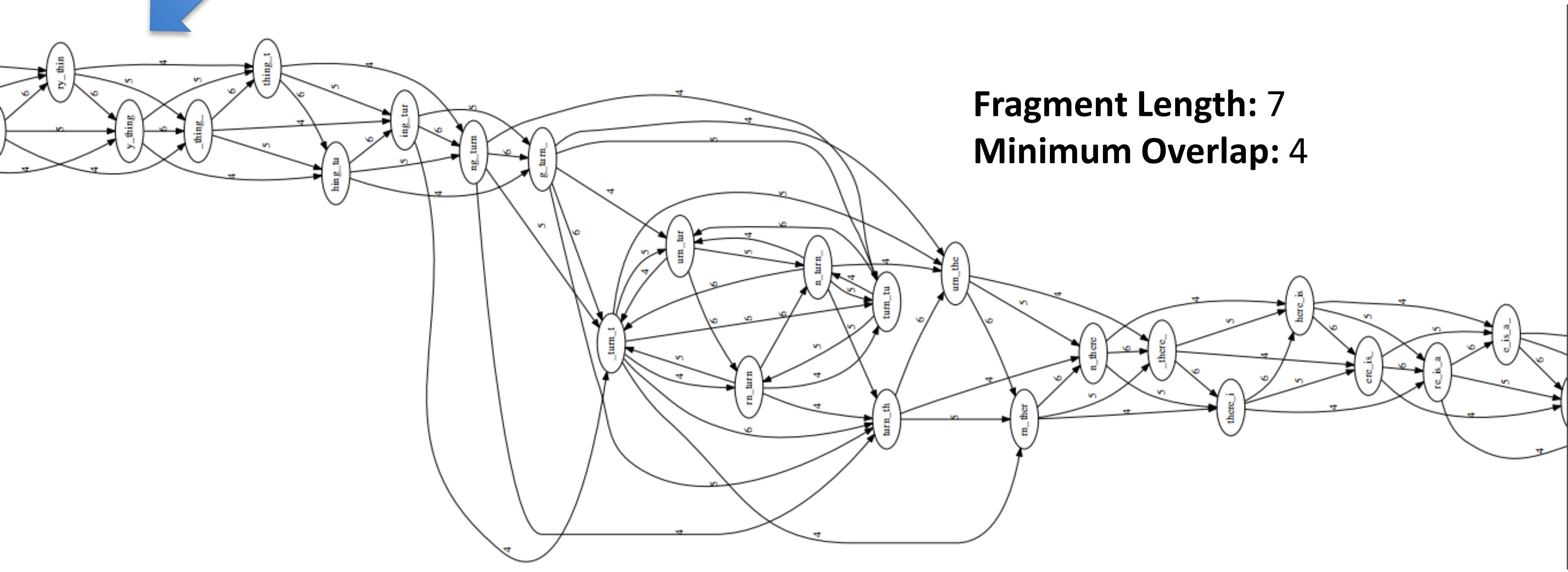
Consensus Sequence

Overlap Layout Consensus

Overlap graph is big and messy!!! Contigs don't just pop out.

Part of the overlap graph for:

`To_every_thing_turn_turn_turn_there_is_a_season`



Fragment Length: 7
Minimum Overlap: 4