# Extending KenLM Pruning

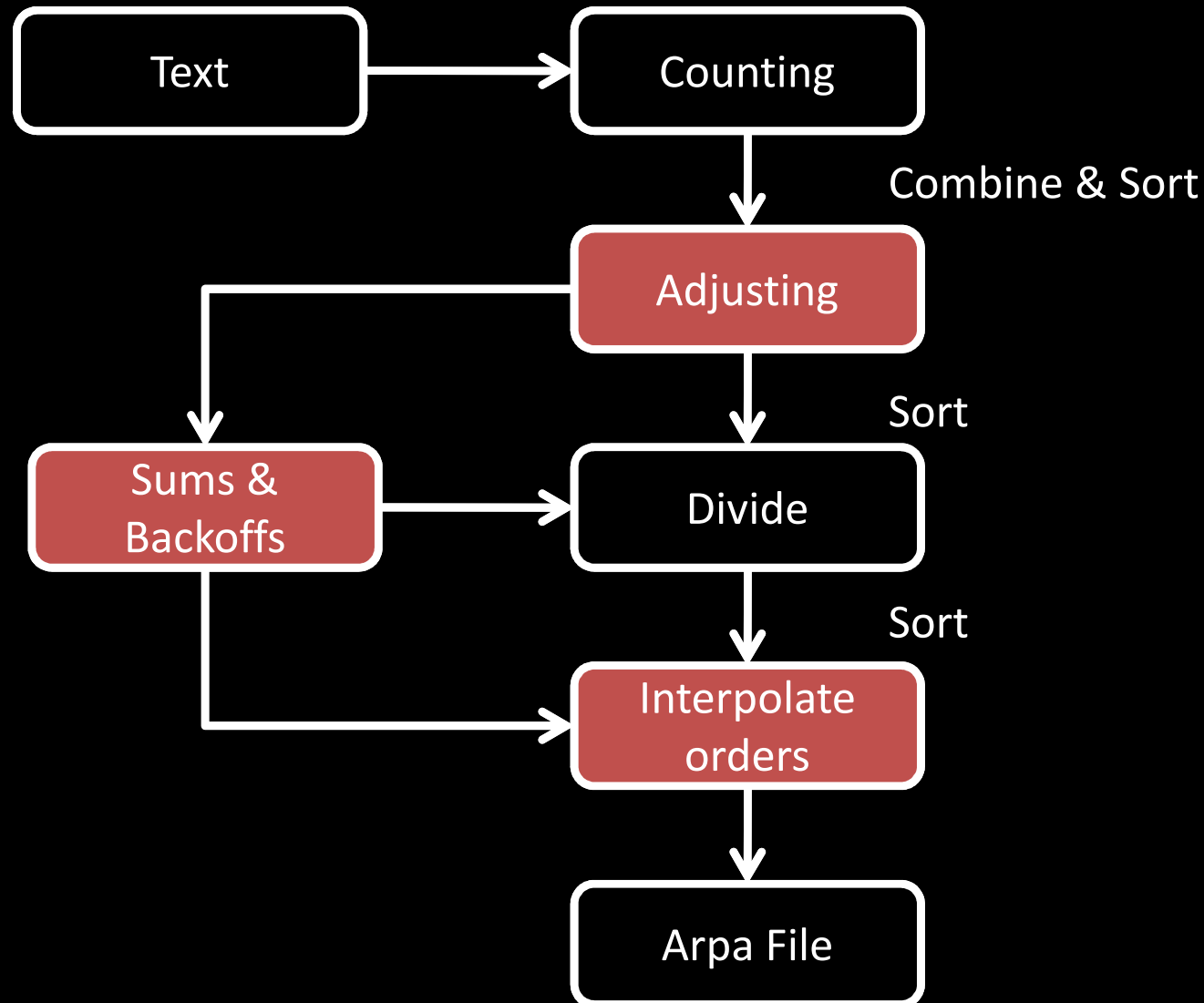Kenneth Heafield, Marcin Junczys-Dowmunt, Valters Sics

# Language Model Pruning

- Count n-grams
- Delete n-gram w with count lower or equal to threshold.

Problems w Modified Kneser-Ney Smoothing:
- To calculate adjusted counts we need original counts
- To calculate backoffs we need to synchronize streams

# Kenneth's Fancy Streams: Now even more fancy

```
┌─────────────┐          ┌─────────────┐
│    Text     │ ───────> │  Counting   │
└─────────────┘          └─────────────┘
                                │
                                │  Combine & Sort
                                ▼
                         ┌─────────────┐
        ┌────────────────│  Adjusting  │
        │                └─────────────┘
        │                       │  Sort
        ▼                       ▼
┌─────────────┐          ┌─────────────┐
│   Sums &    │ ───────> │   Divide    │
│  Backoffs   │          └─────────────┘
└─────────────┘                 │  Sort
        │                       ▼
        │                ┌─────────────┐
        └──────────────> │ Interpolate │
                         │   orders    │
                         └─────────────┘
                                │
                                ▼
                         ┌─────────────┐
                         │  Arpa File  │
                         └─────────────┘
```

# Pruning Results

| Method | Size (Gb) | Time (s) |
|---|---|---|
| No pruning | 3,3 | 153 |
| 5-grams at C(w) = 1 | 1,8 | 110 |
| 4-grams at C(w) = 1 | 0,8 | 76 |
| 3-grams at C(w) = 1 | 0,5 | 60 |
| 2-grams at C(w) = 1 | 0,4 | 59 |

# SUCCESS!

# SUCCESS?

**But wait, what about perplexity?**

# Perplexing Perplexity

| Method | SRILM | KenLM |
|---|---|---|
| No pruning | 48.8 | 48.8 |
| 5-grams at C(w) = 1 | 48.7 | 50.0 |
| 4-grams at C(w) = 1 | 49.0 | 53.4 |
| 3-grams at C(w) = 1 | 50.2 | 57.0 |
| 2-grams at C(w) = 1 | 51.3 | 58.3 |

# It's an ugly bug!
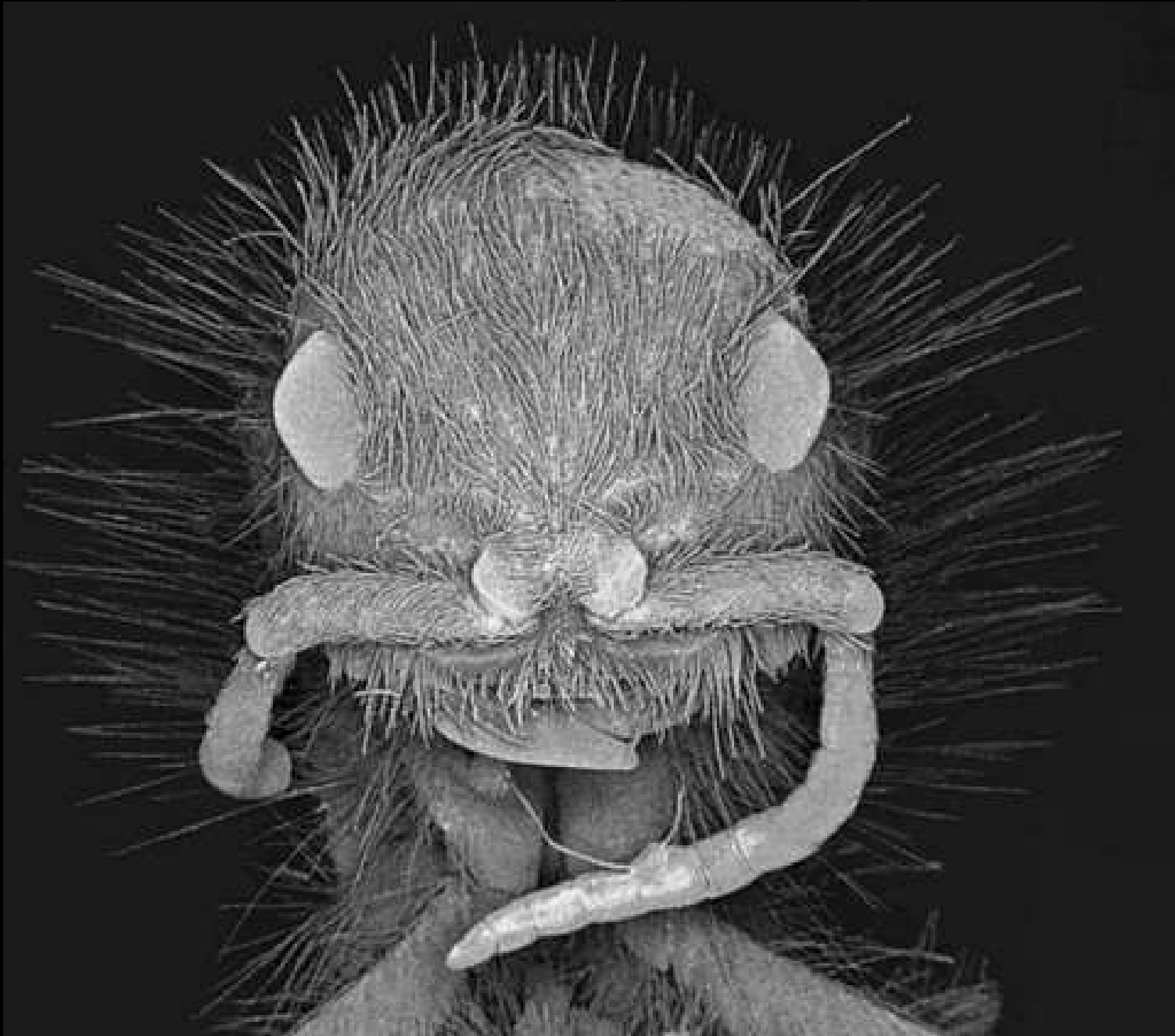
# It's a math problem! (also ugly)
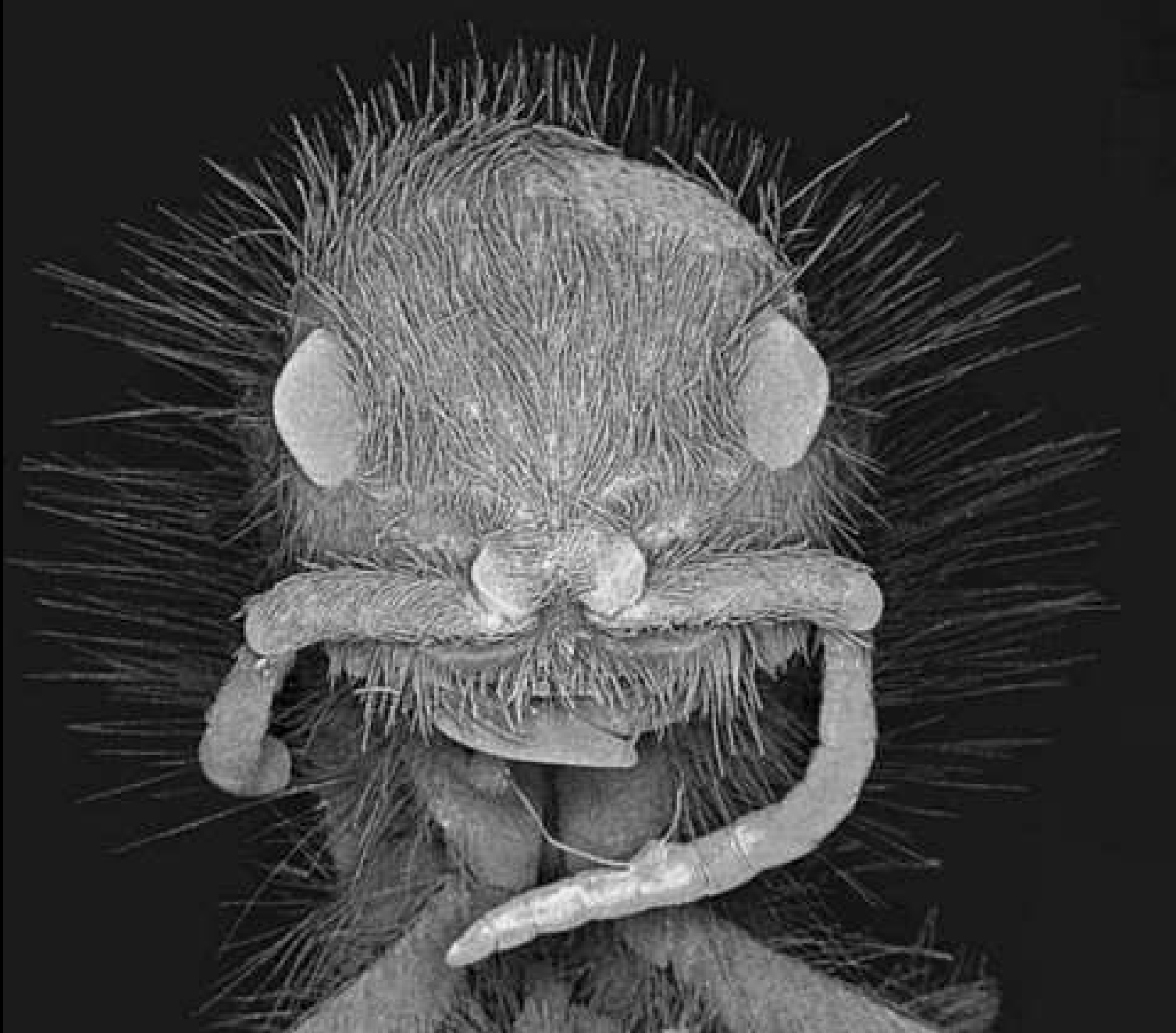
# Still some work left

# Worst Case:
# Copy SRI's Method