Episode Book

EPISODE #6305

Why Terminal Fuzzy Finders Forgive Different Mistakes

PUBLISHED	27 October 2026 at 18:35 UTC
DURATION	23:30
LISTEN	myweirdprompts.com/episode/fuzzy-matching-algorithms-search-36558/
HOSTS	Corn, Herman, Vex, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
SOURCES	21 cited by the research (listed at the end)
BOOK GENERATED	28 October 2026 at 09:53 UTC

SYNOPSIS

Approximate string matching lives between strict regular expressions and semantic retrieval, but different search tools play by entirely different rules. While web search engines rely on edit distances to forgive typos and transpositions, terminal finders like fzf require strict ordered subsequences. This exploration covers how algorithms from Levenshtein, Damerau, Bitap, and Smith-Waterman compute edit distance, parse text via bitwise word operations, and rank candidate records without burning CPU cycles across massive databases.

TRANSCRIPT



Corn

A search box can tolerate a missing letter and still refuse the word you plainly meant. Daniel wants to open up that gap in our run on search, between exact matching with grep and retrieval that understands meaning. What are terminal fuzzy finders and forgiving app searches actually doing, and where does their forgiveness stop?



Herman

The first surprise is that forgiveness has different rules. A terminal finder might happily accept half the letters in a filename, provided you keep them in order. A shop search might accept all the letters in the wrong order, within limits. Those sound like neighboring settings. They're different algorithms.



Corn

One accepts a shortened confession. The other accepts perjury.



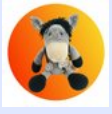
Herman

Take a filename you know well. You probably remember its initials or a distinctive fragment, rather than its spelling with one accidental substitution. That makes skipped characters useful. Product search has a different customer. Someone types the whole name, makes a keyboard error, and expects the system to repair it.



Corn

Before you've pressed Enter, somebody has already decided what sort of mistake you're allowed to make. Herman, I suspect you prefer the system with the most elaborate explanation for why it rejected you.



Herman

Only if the explanation is correct. Several disappointments have excellent documentation.



Corn

Daniel's middle layer needs a slightly untidy definition, then. Approximate matching of strings, with the approximation specified. Not just a search box looking sympathetic.



Herman

Edit distance is the familiar definition. Count the insertions, deletions, or substitutions required to turn one string into another. Some versions also count swapping adjacent characters. Other matchers use shared fragments, sound, or ordered subsequences instead. They don't all produce an edit count.



Corn

Grep gives us a useful boundary. A regular expression can accept many different strings, but each accepted string still satisfies its rules exactly. It doesn't decide that you probably meant a different expression.



Herman

Semantic retrieval crosses another boundary. It compares representations intended to capture meaning, rather than merely counting character differences. The characters still enter the system, of course. Saying it never sees characters would be too strong. Its similarity decision operates on something beyond their literal arrangement.



Corn

A search for password reset can therefore find account access recovery. No sensible allowance of missing letters gets you from one to the other.



Herman

There are also two different jobs hiding under approximate matching. Search can mean finding an approximate occurrence inside a long piece of text. Lookup can mean finding entries in a dictionary that resemble your query. Record linkage extends that second job to comparing customer lists or identifying duplicate people.



Corn

Finding a name in a letter versus deciding whether two names belong to one person. The second one gets dangerous much earlier.



Herman

It also gets expensive differently. Scanning one document and comparing every pair across two databases demand different engineering. A fast distance calculation doesn't rescue you if you insist on calculating it for billions of irrelevant pairs.



Corn

I'd also separate deciding what qualifies from deciding what comes first. My editor can return the command I wanted somewhere underneath twelve plausible alternatives. Technically found. Socially concealed.



Herman

Corn, would you rather have fewer candidates and occasionally miss the command, or have every possible candidate with a strong ranking?



Corn

Strong ranking. Provided exact text remains visible and stable. If the list dances while I'm reaching for it, you've built a typing examination with moving answers.



Herman

That preference will matter when we get inside fzf. Its scoring is full of decisions about what you probably remember.



Corn

The algorithm family should start with what each one permits. Hamming distance is the strict sibling. Same length, substitutions only.



Herman

Levenshtein adds insertion and deletion. It emerged from error correcting codes in nineteen sixty-five. Cot to coat takes an insertion. Coat to cost takes a substitution. The distance is the cheapest sequence of allowed edits, not whichever sequence you happen to notice first.



Corn

Damerau adds the adjacent swap. Cost to cots. A remarkably efficient way to ruin a receipt.



Herman

Damerau's investigation of spelling errors in an information retrieval system found more than eighty percent were a single insertion, deletion, substitution, or transposition. That's a powerful explanation for why modest typo tolerance works. You can capture a large share of ordinary mistakes without admitting everything vaguely nearby.



Corn

It doesn't prove two edits is the right threshold for every language, word length, or database, though. A short surname can acquire an entirely different owner after one edit.



Herman

Nor does it prove that today's typing errors follow exactly that distribution. It gives the approach a strong historical basis, not a universal setting. Short terms usually need tighter treatment. Long terms have more opportunities for mistakes and more surviving evidence about what was intended.



Corn

Two letters wrong in a pharmaceutical name is not a generous customer service policy.



Herman

Name matching has another famous family, Jaro and Jaro-Winkler. They score matching characters and ordering, with Winkler adding a bonus for a common prefix. Usually you'll see similarity from zero to one, higher meaning closer. The distance form reverses that direction.



Corn

Favoring the beginning sounds reasonable until two people share the beginning of a surname and nothing else.



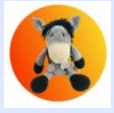
Herman

It's useful evidence, not identity. Also, Jaro-Winkler doesn't satisfy the triangle inequality. That matters when you want an index that relies on metric distances to skip candidates safely. A score being called a distance doesn't automatically give it those mathematical properties.



Corn

Even the distances have misleading names.



Herman

Computing ordinary edit distance means reusing answers to smaller comparisons. You build a table for prefixes of the strings. Each cell considers extending an alignment with an insertion, deletion, or substitution. The cheapest neighboring possibility determines the next answer.



Corn

You're about to describe every cell.



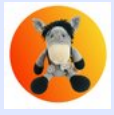
Herman

I am showing restraint.



Corn

I recognize your restrained matrix voice.



Herman

Then appreciate that I'm omitting the recurrence. For approximate substring search, Sellers' approach changes the table's initialization so the match can begin anywhere in the text without paying for the preceding text. A row of zeros allows that free start.



Corn

Otherwise searching a long document would charge me for everything before the word I wanted. A parking meter for paragraphs.



Herman

The work grows with pattern length times text length, rather than repeatedly checking all possible substrings from scratch. Landau and Vishkin later exploited a small permitted edit bound, working along diagonals. Cole and Hariharan improved that line of work. Limiting mistakes can help computation as well as relevance.



Corn

There must be an alternative to carrying a whole table around every time someone mistypes coat.



Herman

Bitap packs matching state into bits. Precompute masks representing where characters occur in the pattern, then update the state using shifts and other bitwise operations. Short patterns can fit their state into a machine word, which lets the processor advance many pattern positions together.



Corn

My misspelling gets a seat on the processor.



Herman

Its lineage begins with Dömölki's exact algorithm in nineteen sixty-four, then Baeza-Yates and Gonnet's reinvention in nineteen eighty-nine. Manber and Wu extended the approach to insertions and deletions; Baeza-Yates and Navarro improved it later. The exact cost depends on the variant, pattern length, and allowed errors. Word packing gives the speed advantage, especially for short patterns and small alphabets.



Corn

That becomes agrep, the member of the grep family willing to negotiate.



Herman

Released in nineteen ninety-two. Approximate patterns, records rather than just lines, and multiple patterns combined with AND or OR. Molecular biology was among its applications. Searching isn't always words on an office computer.



Corn

Longer lists bring a different problem. If I have a million customer records, faster comparisons still leave far too many comparisons.



Herman

Normalize obvious variations, group plausible candidates, then verify them with a more expensive comparison. Tries share prefixes. N-gram indexes share fragments. Metric trees can prune using appropriate distance properties. Hashing and locality-sensitive hashing offer other ways to avoid examining everything.



Corn

Trigrams are three letters at a time, with overlapping groups contributing evidence. Soundex and Metaphone take a different route and compare encodings of pronunciation. Neither is quietly understanding the customer's life story.



Herman

Wildcard expansion is another explicit tool. You describe permissible variation rather than asking the system to infer it. These techniques can cooperate. No requirement says an entity matcher must choose one score and defend it forever.



Corn

Character rules, fragments, sounds. Still no understanding of what you meant. That caller has found the line again. Does screening accept approximate identities now?

V

Vex

Listen, you want to know what fuzzy matching is really about, the escalator handrails, the rubbery black ones, they move at a DIFFERENT SPEED than the steps, always have, that's your first fuzzy match right there, a five percent slip nobody measures.

V

Vex

And every one of those handrails is a rolling tape, and the length of the loop is NOT the length of the ride, so the pattern repeats and drifts, and the drift is the edit distance, and the step between exact and semantic, that gap you keep describing, is EXACTLY the gap between the step and the handrail, and I have the receipts, I have a notebook.

V

Vex

So you tell me why none of your fuzzy finders ever mention the handrail, and I want you to look into it ON AIR, today, before you go back to bitap and trigrams and pretend this is a coincidence.



Herman

Edit distance counts string operations, not differences in mechanical speed.



Corn

The terminal example exposes the naming problem beautifully. With fzf, typing s, d, o can match sudo. Typing an extra w before sudo cannot match that same candidate.



Herman

Every query character must occur in the candidate in order. You may omit candidate characters when constructing your query. You may not introduce a query character that isn't there, or reverse the required order. That's subsequence matching.



Corn

It forgives brevity. It takes invention personally.



Herman

Algolia's typo tolerance uses Damerau-Levenshtein instead. A swapped adjacent pair can qualify there. Identical marketing word, incompatible admission rules. Before comparing how forgiving two tools are, establish what each considers an eligible match.



Corn

Eligibility still doesn't explain why fzf sometimes prefers the longer filename. I type two letters and it appears to develop an opinion about punctuation.



Herman

Its first matcher scans forward until it has found the query, then scans backward to tighten that occurrence. Linear work in candidate length. Fast, but stopping at the first occurrence means it may miss a later alignment with a better score.



Corn

A finder that stops looking once it finds something. I respect the work ethic.



Herman

The second matcher uses modified Smith-Waterman alignment to optimize the score across possible occurrences. Smith-Waterman comes from aligning biological sequences, including DNA and proteins. fzf restricts it so a query character cannot be omitted or mismatched.



Corn

Protein alignment, deployed to locate my shopping list.



Herman

When a match exists, the work can grow with query length times candidate length. If no subsequence exists, a linear scan can reject it first. That's an important distinction. A powerful scoring algorithm doesn't mean filling the expensive table for every impossible candidate.



Herman

Corn, what do you actually remember about a filename? The beginning, the capitals, the last word?



Corn

Usually the word I thought would distinguish it from your other eleven versions. Then the first letter of another word. Almost never the exact number of characters between them.



Herman

That is why word starts and uppercase transitions in camel case receive bonuses. The first query character gets an extra positional bonus because it's especially likely to represent an important anchor. Your remembered initials become useful evidence.



Corn

You're defending the software you installed with my typing habits. A strong negotiating position.



Herman

You keep using it. I accept this as a narrowly defined endorsement.



Corn

I accept the result, not the installation ceremony.



Herman

Gaps receive penalties so initials don't dominate everything. In the documented scoring, a sufficiently long gap cancels the position bonus once it exceeds eight characters. The tension is deliberate. Favor meaningful boundaries, but don't spread three typed letters across an entire sentence without charging for the distance.



Corn

That can make separated letters outrank consecutive ones. I'd have called that a bug before hearing the objective.



Herman

It can be an undesirable side effect. Understanding the objective doesn't make every ranking good. Rofi added an fzf version two sorting method after its older, loosely inspired method could rank the literal query wip last. The actual fzf matcher put that exact candidate first.



Corn

Approximately implementing approximate matching. Two opportunities to miss.



Herman

Editor command palettes have the same user-facing problem, but we shouldn't assign fzf's implementation to every editor. Boundary bonuses, eligibility rules, and ranking weights can differ. Similar behavior on a few examples doesn't establish common internals.



Corn

The forgiving product box has its own expectations. If I type Michael with adjacent letters swapped, I want Michael. I don't want a lecture about subsequences.



Herman

Algolia's worked explanation allows up to two typos, rejecting three or more. Exact matches rank ahead of one typo, then two, within its typo-ranking treatment. Actual search settings and other ranking criteria still matter. A threshold describes admission, not the whole ordering system.



Corn

Michael with the h missing is close. An entirely different word beginning with the same letter isn't redeemed by confidence.



Herman

Synonyms and prefix matching are separate capabilities. Typo tolerance cannot establish that trousers and slacks refer to the same product. A thesaurus applied indiscriminately can make things worse, since words have multiple senses.



Corn

I'd buy discounted trousers. Discounted inactivity is less compelling.



Herman

Surface cleanup helps before you need meaning. Case, punctuation, spacing, and common abbreviations can be normalized. Acme with a hyphen and Acme with a space shouldn't necessarily become a difficult matching problem.



Corn

Saint Mary Hospital versus St Mary Hosp needs abbreviation knowledge, though. Plain edit distance doesn't suddenly know what saint means. I'd put that normalization outside the matcher and keep its decisions visible.



Herman

Yes... well, not every normalization is harmless either. Removing punctuation from a company name is different from stripping characters from a model number. Would you let the semantic system substitute a related product if your exact identifier returned nothing?



Corn

Show related products separately. Don't pretend you've found the specified part. A replacement power supply can be conceptually perfect and electrically disastrous.



Herman

Drug names, account identifiers, and legal clause references have similar precision demands. Meanwhile password reset and account access recovery may share intent without sharing useful character structure. That's where representations of meaning earn their place.



Corn

Neither method should be promoted into universal authority merely because the other disappointed us.



Herman

Hybrid systems preserve both kinds of evidence. Azure's hybrid search runs full-text and vector queries in parallel, then combines rankings with Reciprocal Rank Fusion. Elastic describes the complementary roles as lexical precision and semantic intent. That's broader than typo tolerance, but it answers the same architectural pressure.



Corn

A merged ranking still doesn't authorize merging two customer accounts. Retrieval can offer candidates. Identity needs a decision.



Herman

OpenSanctions Pairs makes that distinction concrete. Its rule-based baseline reached ninety-one point three percent F one, while GPT four o reached ninety-nine percent. That's a substantial improvement on labeled entity pairs, not proof that a language model makes a better interactive filename picker.



Corn

Different task, different penalty for hesitation. I can reject a filename instantly. An incorrect merged identity can survive for years.



Herman

The models and rules also failed differently. Rules overmatched; language models struggled with cross-script transliteration. A locally deployable distilled model approached the strongest result. Accuracy alone doesn't settle the architecture because running a large model on every pair has a cost.



Corn

Especially when most pairs plainly aren't candidates.



Herman

Compact pretrained language model students have competed with much larger models while taking tens to hundreds of times less inference time in entity-matching experiments. That supports selective escalation. Cheap string checks first, embeddings where useful, a larger model for unresolved cases.



Corn

I'd want the unresolved cases retained separately. A system that quietly converts uncertainty into a permanent merged record is going to...



Hilbert

Charge me for the second wedding, Corn. My cousin Pavel runs the archive counter at the scanning shop. I'm in there every Thursday this month. His matching service keeps combining Henrietta Flumingtop and Henriette Flumingtop. They are different people. The combined record now has eleven weddings. Pavel charges three shekels to separate a marriage, and his next import puts it back.



Corn

Recurring revenue from recurring matrimony.



Hilbert

Also, trigrams are three characters, not necessarily three letters. Spaces count if the implementation includes them. You said letters earlier. I'm not disputing the usefulness. I use the search. Pavel sets the distance threshold to two and calls the result verified. Henrietta and Henriette pass. My abbreviated name doesn't pass his other filter because he leaves the full stop in the query. Herman spends considerably longer distinguishing his two filters than Pavel does.



Herman

The full stop would indeed have to occur in a plain subsequence match. Removing it and expanding the initial would be separate processing.



Hilbert

His family server is the laptop beside the garage sink. There is no backup. I am correcting the marriages directly in the archive every Thursday. Pavel doesn't know that. His imports compare new records against my corrections. This month he is asking me to provide an original Henrietta. I am booking her a counter appointment because the form has no box for deceased.



Corn

Keep the appointment separate from the wedding records.



Hilbert

I am trying. The appointment search finds Henriette first. Pavel has put both names in the same Thursday slot at ten fifteen.



Herman

Even candidate suggestions need a visible reason. Was that appointment selected because the spelling was close, because an abbreviation expanded, or because someone confirmed the identity? Those are different claims, and the interface can preserve that distinction.



Corn

The common misconception is that fuzzy means typo tolerant. It can mean that. In fzf it means ordered query characters with gaps allowed in the candidate.



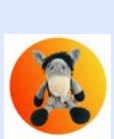
Herman

A substituted letter can therefore fail where a much shorter query succeeds. You need the eligibility rule before the score tells you anything useful.



Corn

Whether hybrid search becomes permanent may depend less on embeddings getting cleverer than on keeping a suggested match distinguishable from a verified identity. That Thursday appointment needs its uncertainty left attached.



Herman

Cheap string checks, then richer comparisons, then expensive judgment where confidence remains low. We haven't established that this cascade wins every interactive task. Entity matching and finding a command are different problems.



Corn

Thanks to Hilbert Flumingtop for producing the show.



Herman

If you want more of this, try episode ten, How ASR Went From Frustration To … Whisper Magic; and episode nineteen, AI Images. This has been My Weird Prompts, the human-AI collaboration podcast.



Corn

If you enjoy the show, leave us a review wherever you're listening.



Herman

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **fzf source, `src/algo/algo.go` (1341 lines) [primary]**
   https://github.com/junegunn/fzf/blob/master/src/algo/algo.go

2. **fzf Reference docs [primary]**
   https://junegunn.github.io/fzf/reference/

3. **Algolia engineering blog [primary]**
   https://www.algolia.com/blog/engineering/fuzzy-search-101

4. **Wu & Manber, agrep [primary]**
   a fast approximate pattern-matching tool, USENIX Winter 1992
   https://www.usenix.org/legacy/publications/library/proceedings/wu.pdf

5. **agrep development chronicle [primary]**
   https://github.com/Wikinaut/agrep/blob/master/agrep.chronicle

6. **Azure AI Search hybrid overview [primary]**
   https://learn.microsoft.com/en-us/azure/search/hybrid-search-overview

7. **Wikipedia, last edited 2026-10-22**
   https://en.wikipedia.org/wiki/Approximate_string_matching

8. **Wikipedia, last edited 2026-09-18**
https://en.wikipedia.org/wiki/Levenshtein_distance

9. **Wikipedia, last edited 2026-08-13**
https://en.wikipedia.org/wiki/Damerau%E2%80%93Levenshtein_distance

10. **Wikipedia, last edited 2025-10-09**
https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

11. **Wikipedia**
https://en.wikipedia.org/wiki/Bitap_algorithm

12. **fzf maintainer discussion**
https://github.com/junegunn/fzf/discussions/3795

13. **rofi PR adding `fzf-v2` sorting**
https://github.com/davatorium/rofi/pull/2306

14. **`norms` crate docs on fzf V2**
https://docs.rs/norms/latest/norm/fzf/struct.FzfV2.html

15. **Jeff Bailey, 2026-04-25**
https://jeffbailey.us/blog/2026/04/25/what-is-fzf/

16. **Manticore Search**
https://manticoresearch.com/blog/fuzzy-search-vs-semantic-search/

17. **fuzzy.direct guide**
https://fuzzy.direct/hybrid-search-vs-fuzzy-search

18. **OpenSanctions Pairs: Large-Scale Entity Matching with LLMs, v2 revised 2026-08-25**
https://arxiv.org/abs/2603.11051v2

19. **Labeling Training Data for Entity Matching Using Large Language Models, v2 revised 2026-10-05**
https://arxiv.org/abs/2606.28823v2

20. **Entity Matching using Decision Models, submitted 2026-10-25**
https://arxiv.org/abs/2610.34626v1

21. **Polecat hybrid search docs**
https://polecat.jasperfx.net/documents/querying/hybrid-search