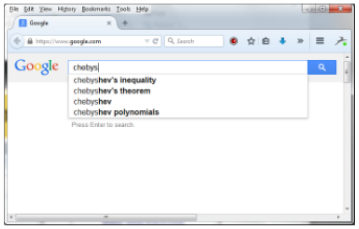
Auto correction/completion

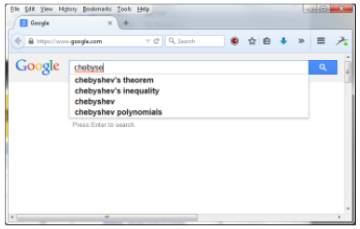
••

University of Southern California

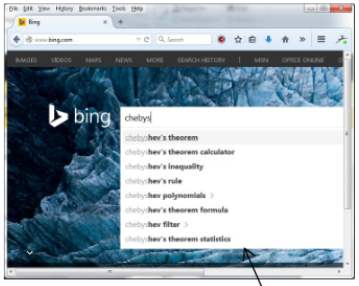
USC Viterbi School of Engineering

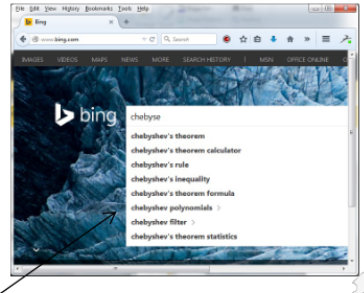
Some Google/Bing Examples



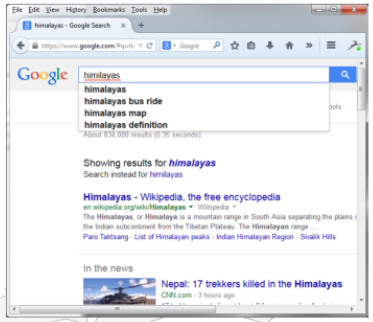


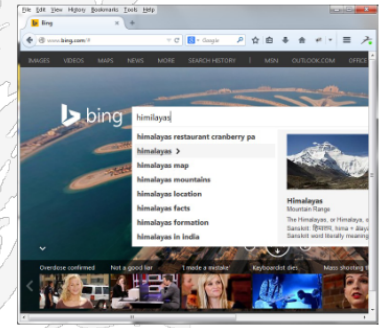
Russian mathematician, notice red underline appears as soon as the first incorrect character is typed





Bing also combines autocomplete with spelling correction but there is no red underline





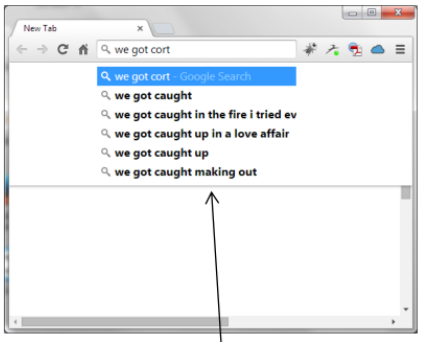
Himilayas

••

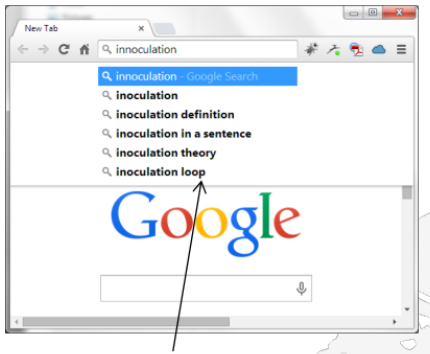
University of Southern California

USC Viterbi School of Engineering

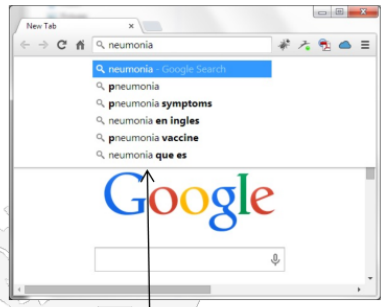
More Examples



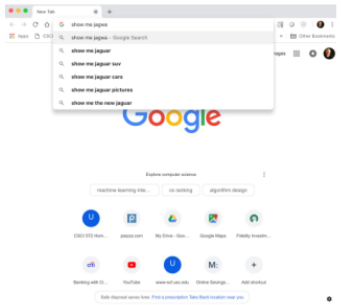
easy for people, but harder for computers to correct, **likely use of n-grams**

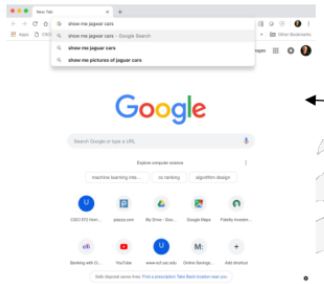


easy for a computer to correct **likely use of a database** of words



computer needs to both identify the error and correct the misspelling





Google combines spelling correction with the **most likely terms** as it comes up with "cars" in autocomplete for the query "jagwa" (misspelled) but leaves the user's misspelling for a while

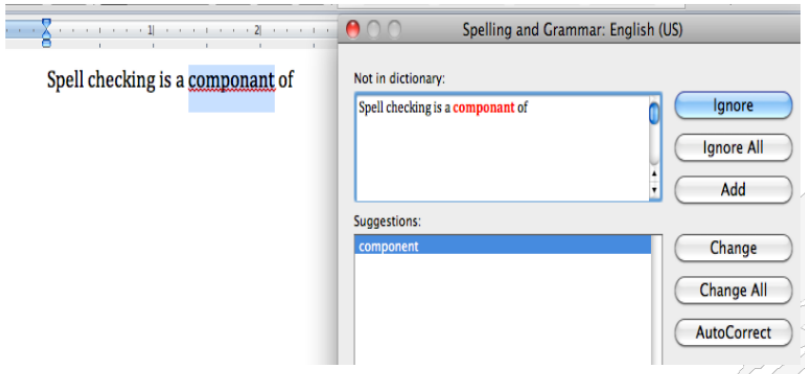
••

University of Southern California

USC Viterbi School of Engineering

Spelling Correction is Done in Many Places

1. Word processing



2. Smartphone input

Word processing is the classic application for spelling correction

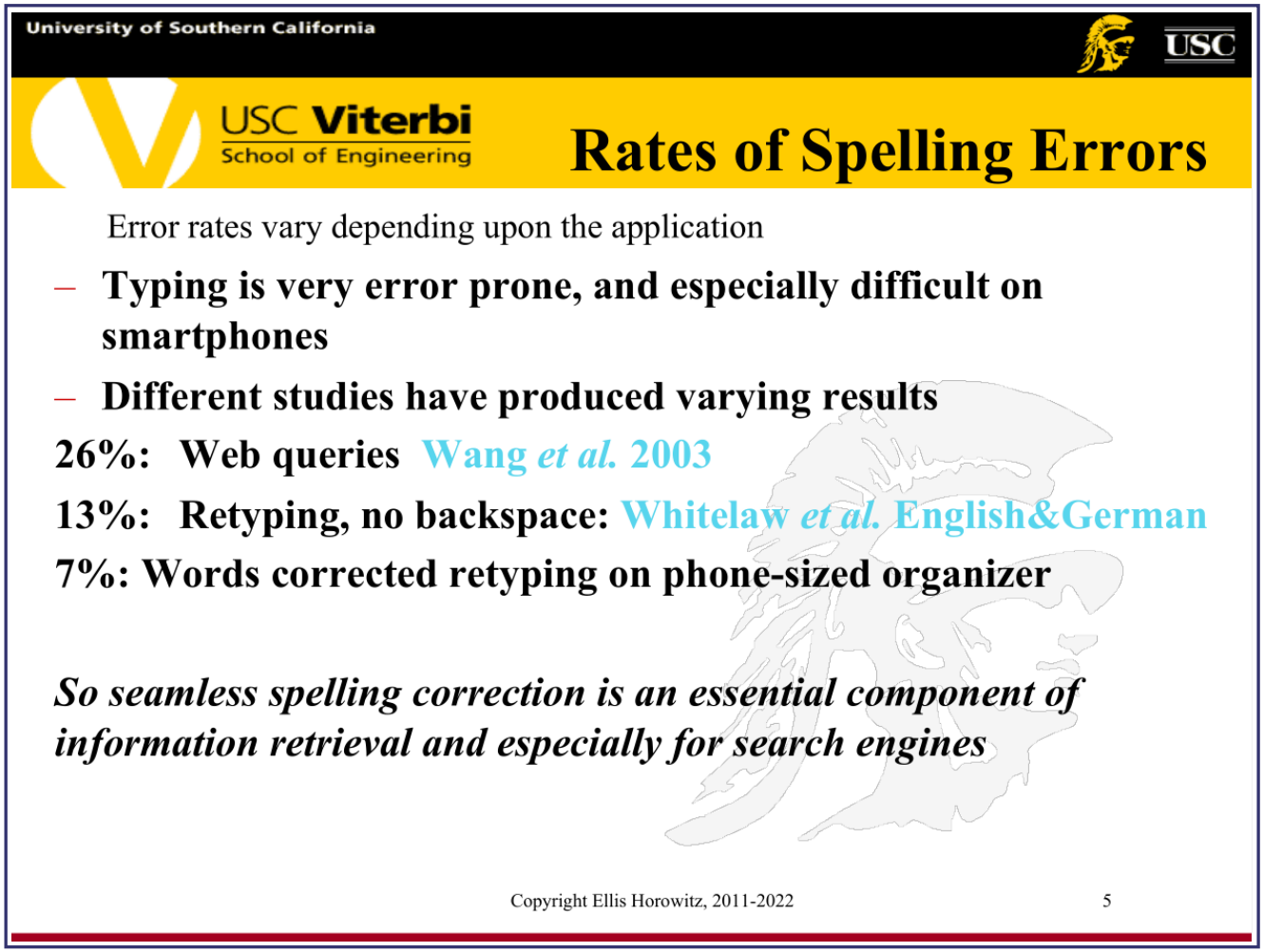
Word and PowerPoint have mode to auto-correct

- set as the default
- the spell dictionary can be modified

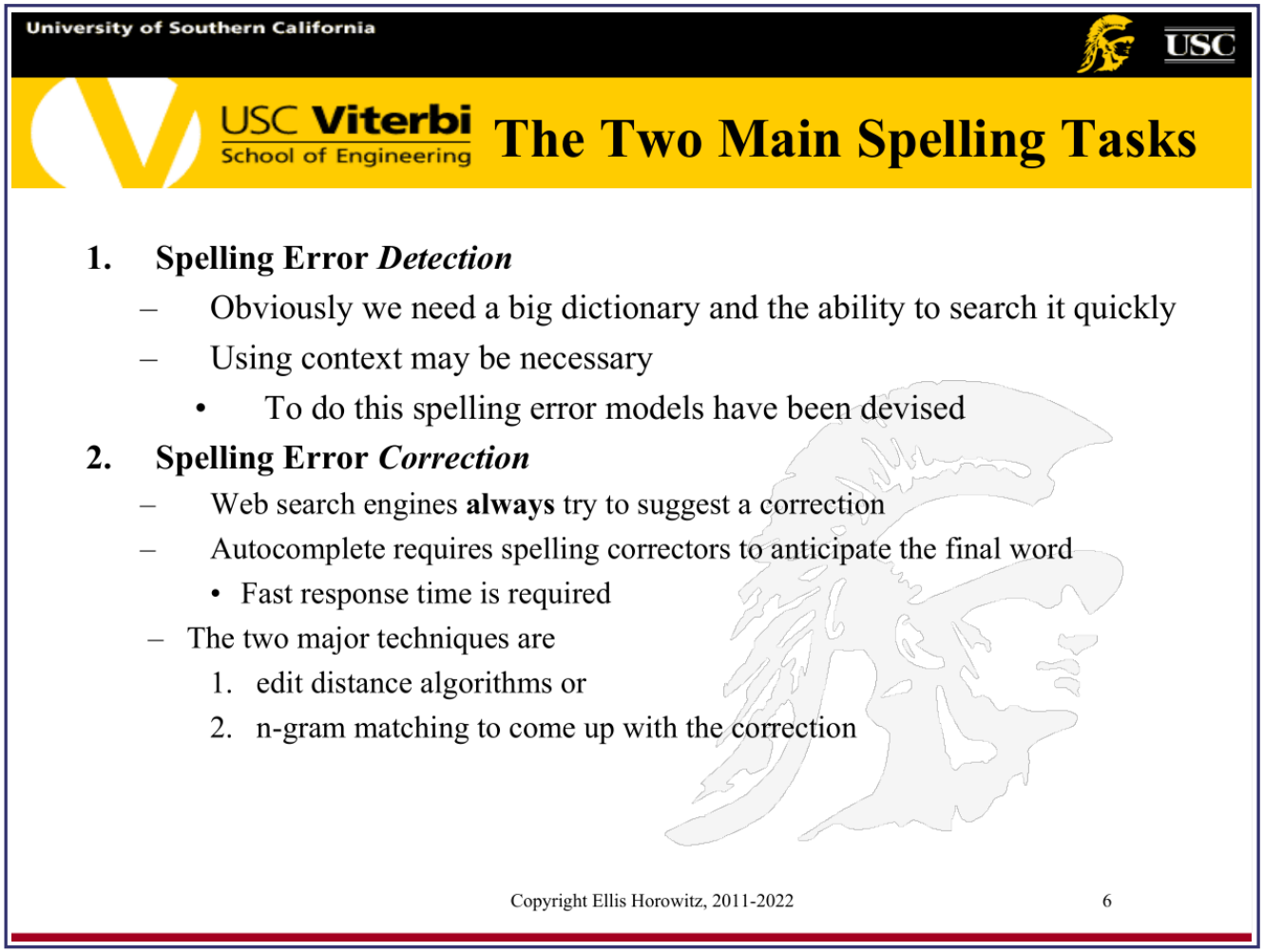
Typing on a virtual keyboard can be doubly difficult (for seniors)

 4

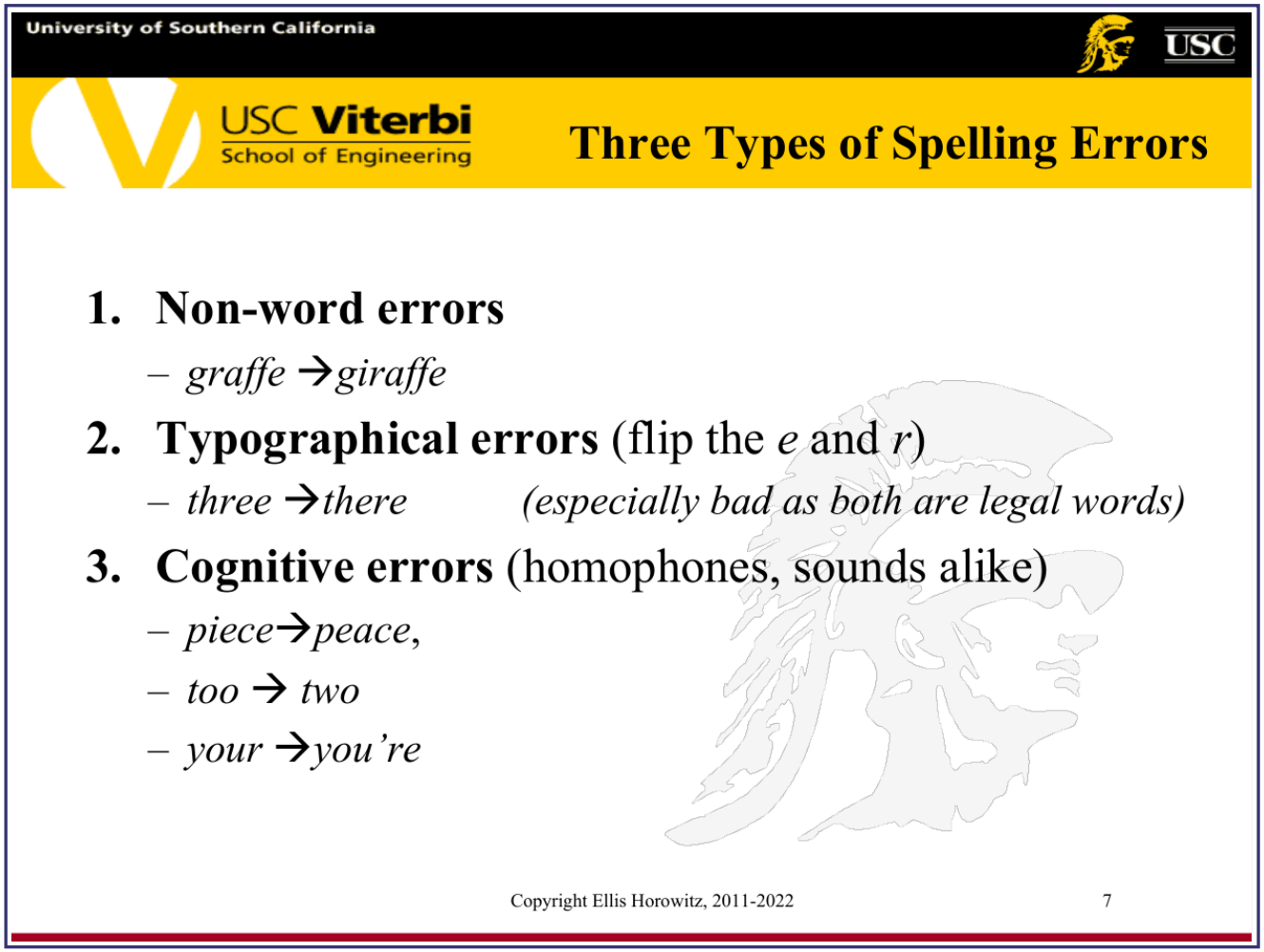
••


University of Southern California
USC
USC Viterbi
School of Engineering
Rates of Spelling Errors
Error rates vary depending upon the application
– Typing is very error prone, and especially difficult on smartphones
– Different studies have produced varying results
26%: Web queries Wang et al. 2003
13%: Retyping, no backspace: Whitelaw et al. English&German
7%: Words corrected retyping on phone-sized organizer
So seamless spelling correction is an essential component of information retrieval and especially for search engines
Copyright Ellis Horowitz, 2011-2022
5

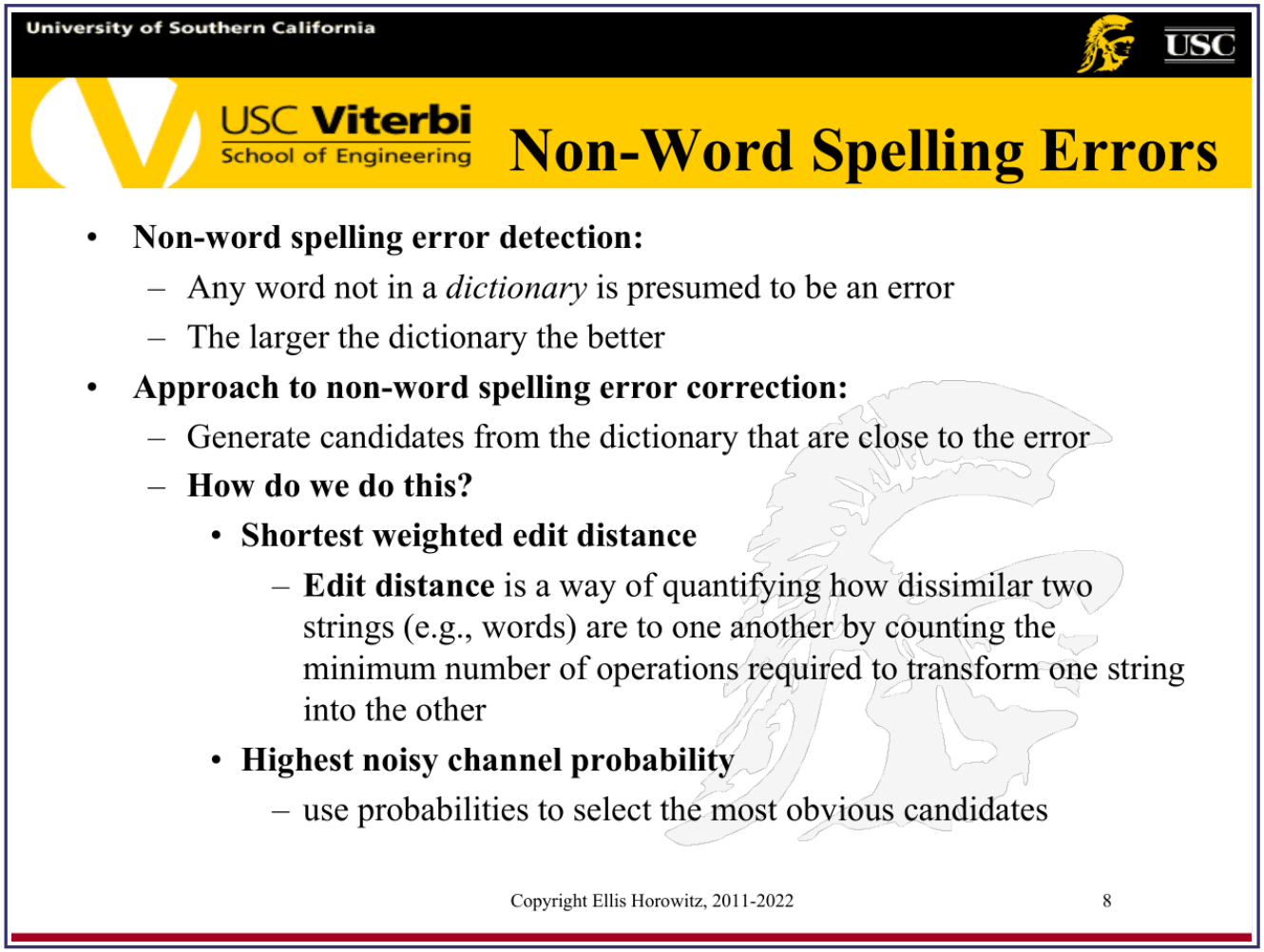
••


University of Southern California
USC
USC Viterbi
School of Engineering
The Two Main Spelling Tasks
1. Spelling Error Detection
– Obviously we need a big dictionary and the ability to search it quickly
– Using context may be necessary
• To do this spelling error models have been devised
2. Spelling Error Correction
– Web search engines always try to suggest a correction
– Autocomplete requires spelling correctors to anticipate the final word
• Fast response time is required
– The two major techniques are
1. edit distance algorithms or
2. n-gram matching to come up with the correction
Copyright Ellis Horowitz, 2011-2022
6

••


University of Southern California
USC
USC Viterbi
School of Engineering
Three Types of Spelling Errors
1. Non-word errors
– graffe →giraffe
2. Typographical errors (flip the e and r)
– three →there (especially bad as both are legal words)
3. Cognitive errors (homophones, sounds alike)
– piece→peace,
– too → two
– your →you're
Copyright Ellis Horowitz, 2011-2022
7

••


University of Southern California
USC
USC Viterbi
School of Engineering
Non-Word Spelling Errors
• Non-word spelling error detection:
– Any word not in a dictionary is presumed to be an error
– The larger the dictionary the better
• Approach to non-word spelling error correction:
– Generate candidates from the dictionary that are close to the error
– How do we do this?
• Shortest weighted edit distance
– Edit distance is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other
• Highest noisy channel probability
– use probabilities to select the most obvious candidates
Copyright Ellis Horowitz, 2011-2022
8

••

University of Southern California

USC Viterbi School of Engineering

Causes of Misspellings

Cause	Misspelling	Correction
typing quickly	exxit mispell	exit misspell
keyboard adjacency	importamt	important
inconsistent rules	conceive concierge	conceive concierge
ambiguous word breaking	silver light	silverlight
new words	kinnect	kinect

e before i
i before e

According to Cucerzan and Brill, **more than 10% of search engine queries are misspelled**
"*Spelling Correction as an iterative process that exploits the collective knowledge of web users*"
http://csci572.com/papers/Cucerzan.pdf (advocates using query logs to guess the correct spelling)

 9

••

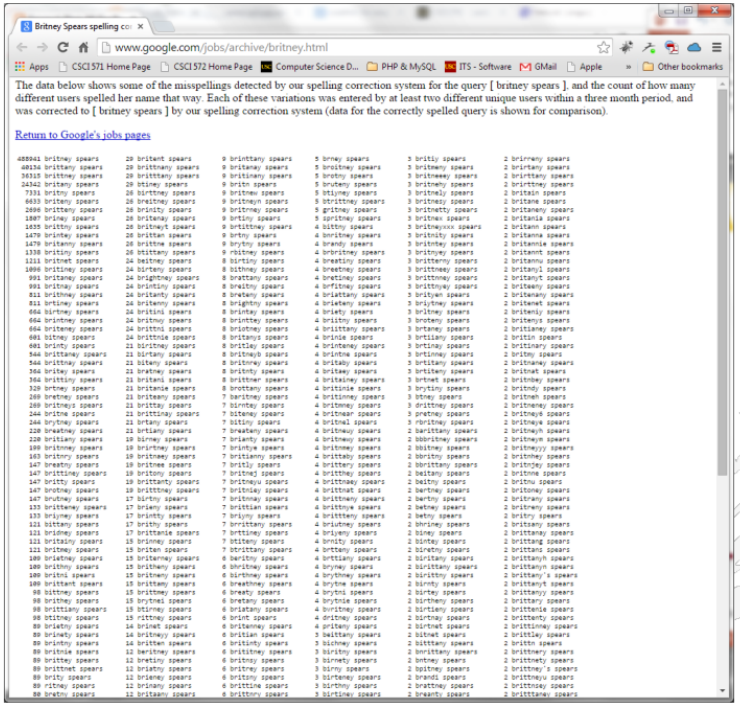
University of Southern California

USC

USC Viterbi School of Engineering

How Many Error Variations of a Word May Actually Appear

http://www.netpaths.net/blog/britney-spears-spelling-variations/



Google's list of spelling errors for "Britney Spears" includes a count of how many different users spelled her name in various way, e.g.
488,941 people used the correct spelling "britney spears" while 40,134 used "britnay spears"

there are 593 different variations that actually occurred

••

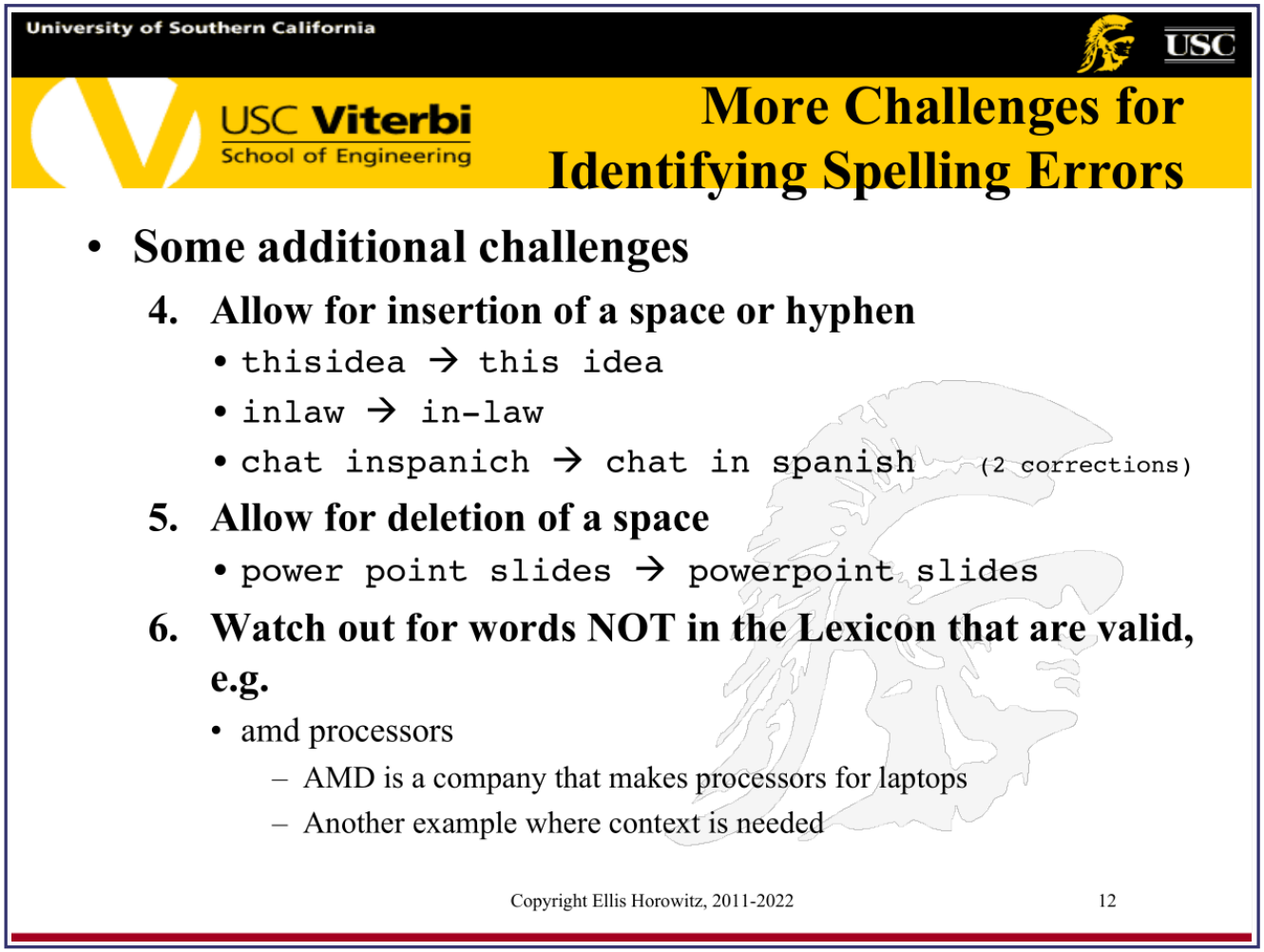
University of Southern California

USC Viterbi School of Engineering

Spelling Errors Needing Context

- **Some misspellings require context to disambiguate**
  1. **consider whether the surrounding words "make sense" for your candidate set, e.g.**
     - *Flying form Heathrow to LAX → Flying from Heathrow to LAX*
     - *Power crd → power cord versus Video crd → video card*
  2. **For candidate words with similar sounds but different spellings and different meanings, context is needed**
     - *e.g. there, their*
     - *N-grams are most useful here*
  3. **To resolve the above, candidate words must be found with similar *pronunciations***
     - *use the Soundex algorithm*

••


University of Southern California
USC
USC Viterbi
School of Engineering
More Challenges for Identifying Spelling Errors
• Some additional challenges
4. Allow for insertion of a space or hyphen
• thisidea → this idea
• inlaw → in-law
• chat inspanich → chat in spanish (2 corrections)
5. Allow for deletion of a space
• power point slides → powerpoint slides
6. Watch out for words NOT in the Lexicon that are valid, e.g.
• amd processors
– AMD is a company that makes processors for laptops
– Another example where context is needed
Copyright Ellis Horowitz, 2011-2022
12

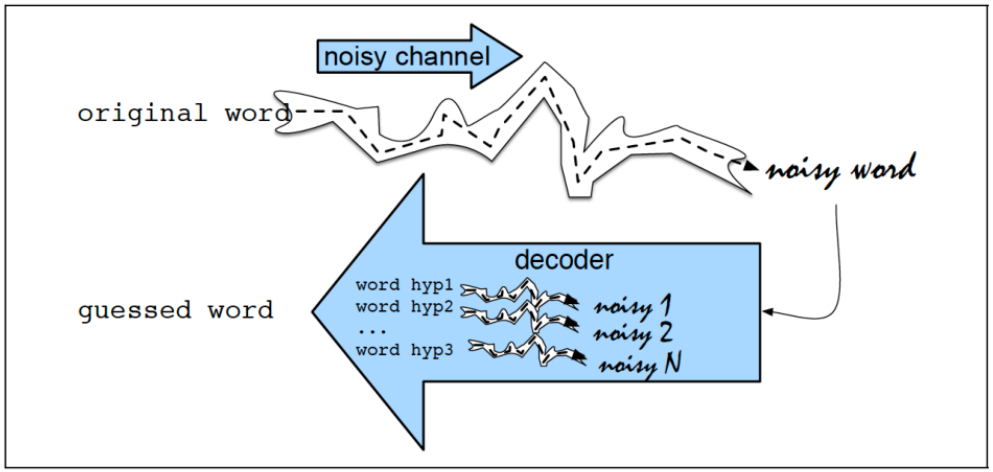
••

University of Southern California

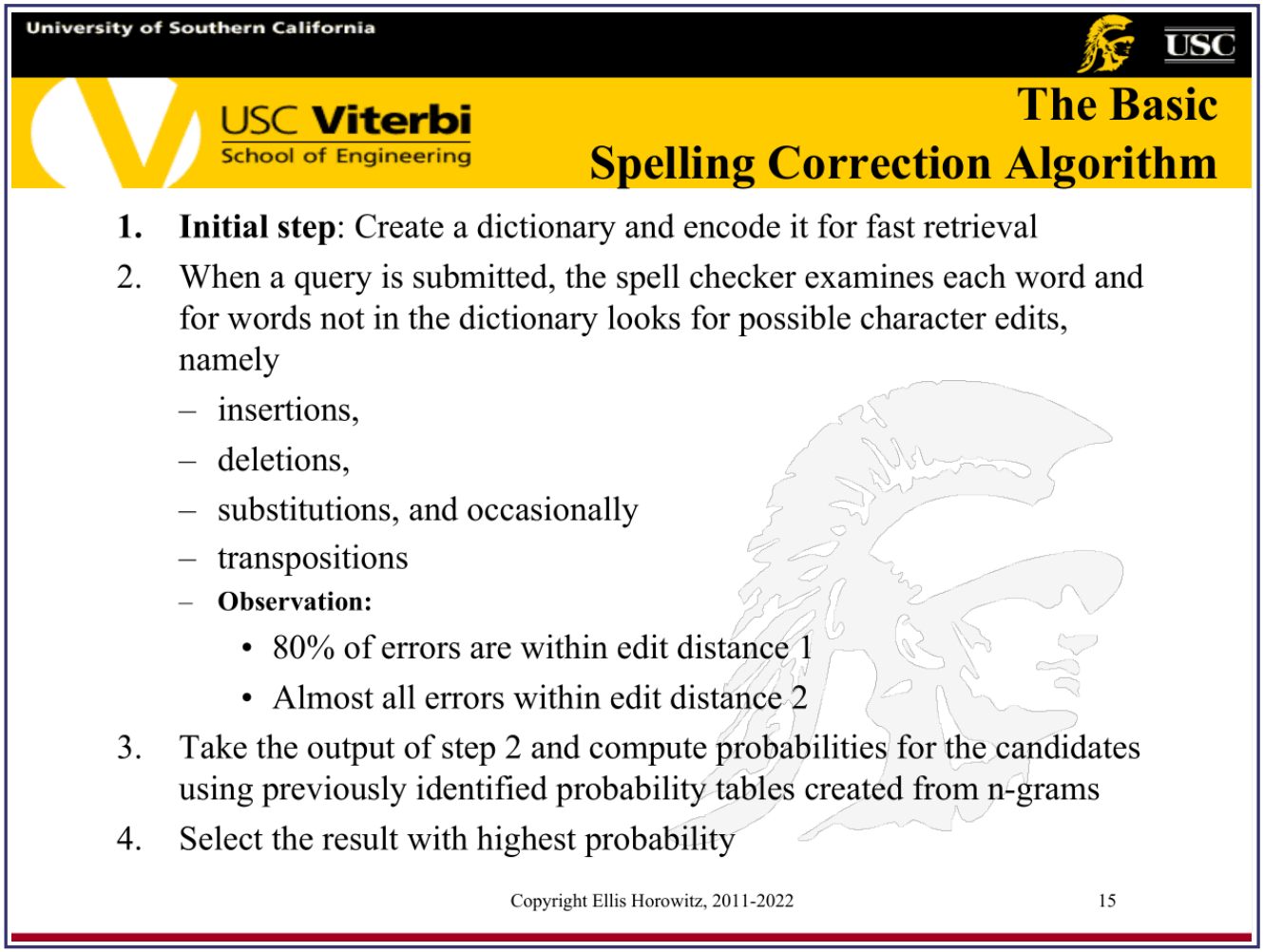
USC Viterbi School of Engineering

The Noisy Channel Model

- This model suggests treating the misspelled word as if a correctly spelled word has been distorted by being passed through a noisy communication channel
- Noise in this case refers to substitutions, insertions or deletions of letters



••


University of Southern California
USC
USC Viterbi
School of Engineering
The Basic
Spelling Correction Algorithm
1. Initial step: Create a dictionary and encode it for fast retrieval
2. When a query is submitted, the spell checker examines each word and for words not in the dictionary looks for possible character edits, namely
– insertions,
– deletions,
– substitutions, and occasionally
– transpositions
– Observation:
• 80% of errors are within edit distance 1
• Almost all errors within edit distance 2
3. Take the output of step 2 and compute probabilities for the candidates using previously identified probability tables created from n-grams
4. Select the result with highest probability
Copyright Ellis Horowitz, 2011-2022
15

••

University of Southern California

USC Viterbi School of Engineering

The Basic Spelling Correction Algorithm Refined

- **Edit distance** is a way of quantifying how dissimilar two strings (e.g. words) are to one another by counting the minimum number of operations required to transform one string into the other
  - different algorithms assume slightly different operations
  - e.g., Levenshtein uses: removal, insertion, substitution of a character

1. **Check each query term against the dictionary**
2. **For each term NOT found in the dictionary, generate all words within edit distance ≤ *k* (e.g., *k* = 1 or 2) and then intersect them with dictionary**
   - Compute them fast with a Levenshtein algorithm
3. **Use a character *n*-gram index and find dictionary words that share "most" *n*-grams with word**
4. **Select the word with the highest probability** (largest n-gram count)
5. **For speed, have a pre-computed map of words to possible corrections**

 16

••

University of Southern California

USC Viterbi School of Engineering

Use Edit Distance To Produce Candidate Corrections

Input	Candidate Correction	Correct Letter	Error Letter	correction Type
acress	actress	t	-	insertion
acress	cress	-	a	deletion
acress	caress	ca	ac	transposition
acress	access	c	r	substitution
acress	across	o	e	substitution
acress	acres	-	s	deletion

Six words within 1 of acress

Context check is necessary to choose the appropriate word

(try this yourself in Google/Bing)

For the word "acress" there are six dictionary words all within edit distance 1

 17

••

University of Southern California

USC Viterbi School of Engineering

Now Apply Probabilities

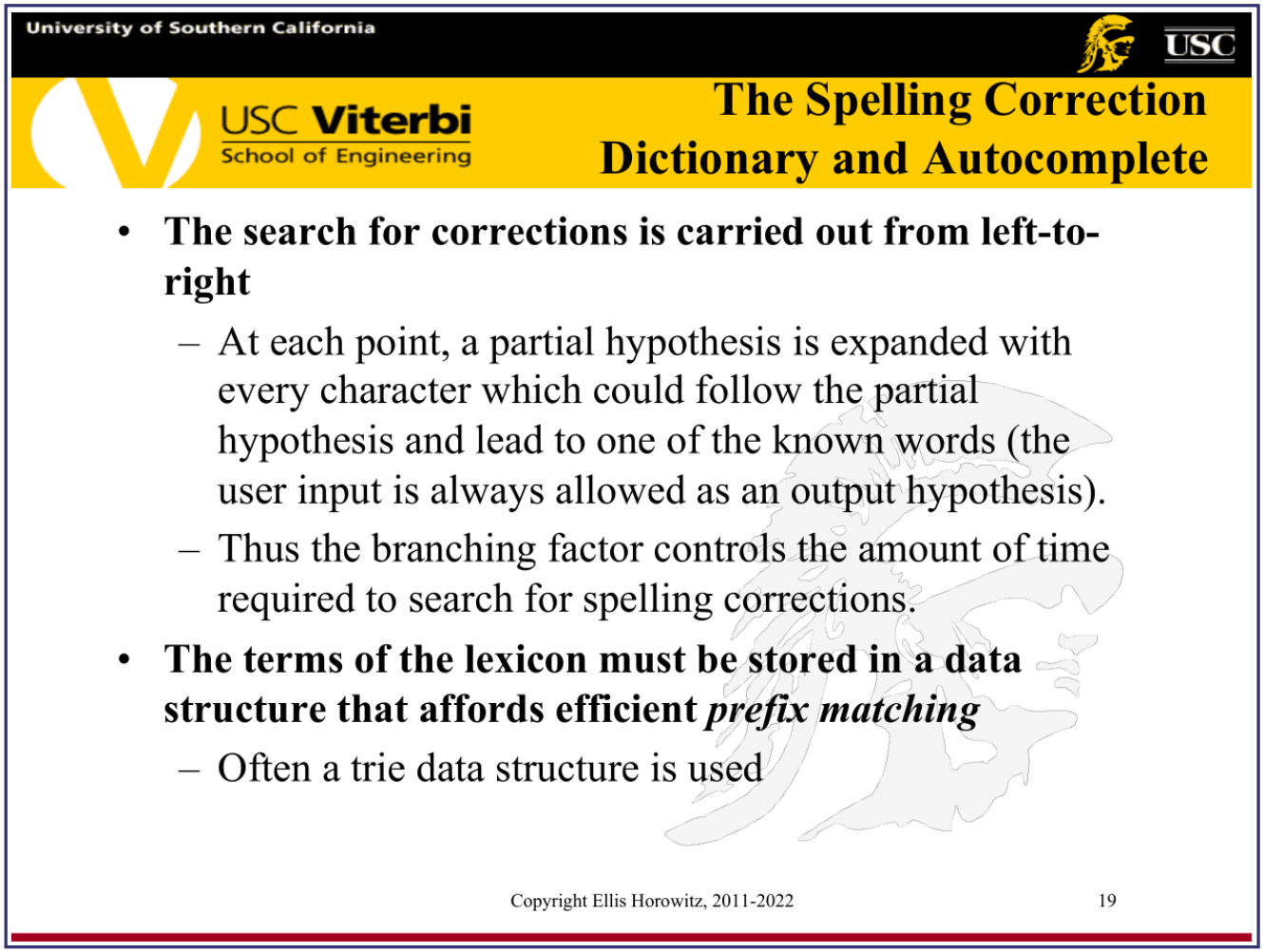
- **We now need to compute the prior probability of each occurrence**
- **We can do this using unigrams, bigrams, trigrams, etc**
- **Using the Corpus of Contemporary English, 404,253,213 words we get the following**
- ***Across* is the most likely choice, followed by**

For fun try:
I need acress to . . .
I love the acress . . .
a kiss and a acress . . .

word	Frequency of word	*P(w)*
actress	9,321	.0000230573
cress	220	.0000005442
caress	686	.0000016969
access	37,038	.0000916207
across	120,844	.0002989314
acres	12,874	.0000318463

"***across***" is the most likely correction →

••


University of Southern California
USC
USC Viterbi
School of Engineering
The Spelling Correction
Dictionary and Autocomplete
• The search for corrections is carried out from left-to-right
– At each point, a partial hypothesis is expanded with every character which could follow the partial hypothesis and lead to one of the known words (the user input is always allowed as an output hypothesis).
– Thus the branching factor controls the amount of time required to search for spelling corrections.
• The terms of the lexicon must be stored in a data structure that affords efficient *prefix matching*
– Often a trie data structure is used
Copyright Ellis Horowitz, 2011-2022
19

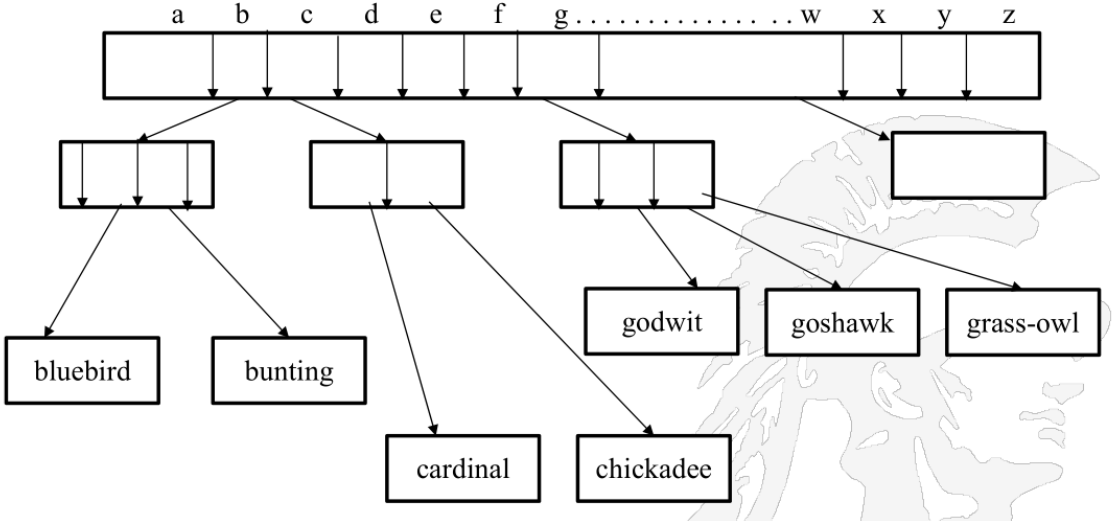
••

University of Southern California

USC Viterbi School of Engineering

The Spelling Correction Dictionary Example of a Trie

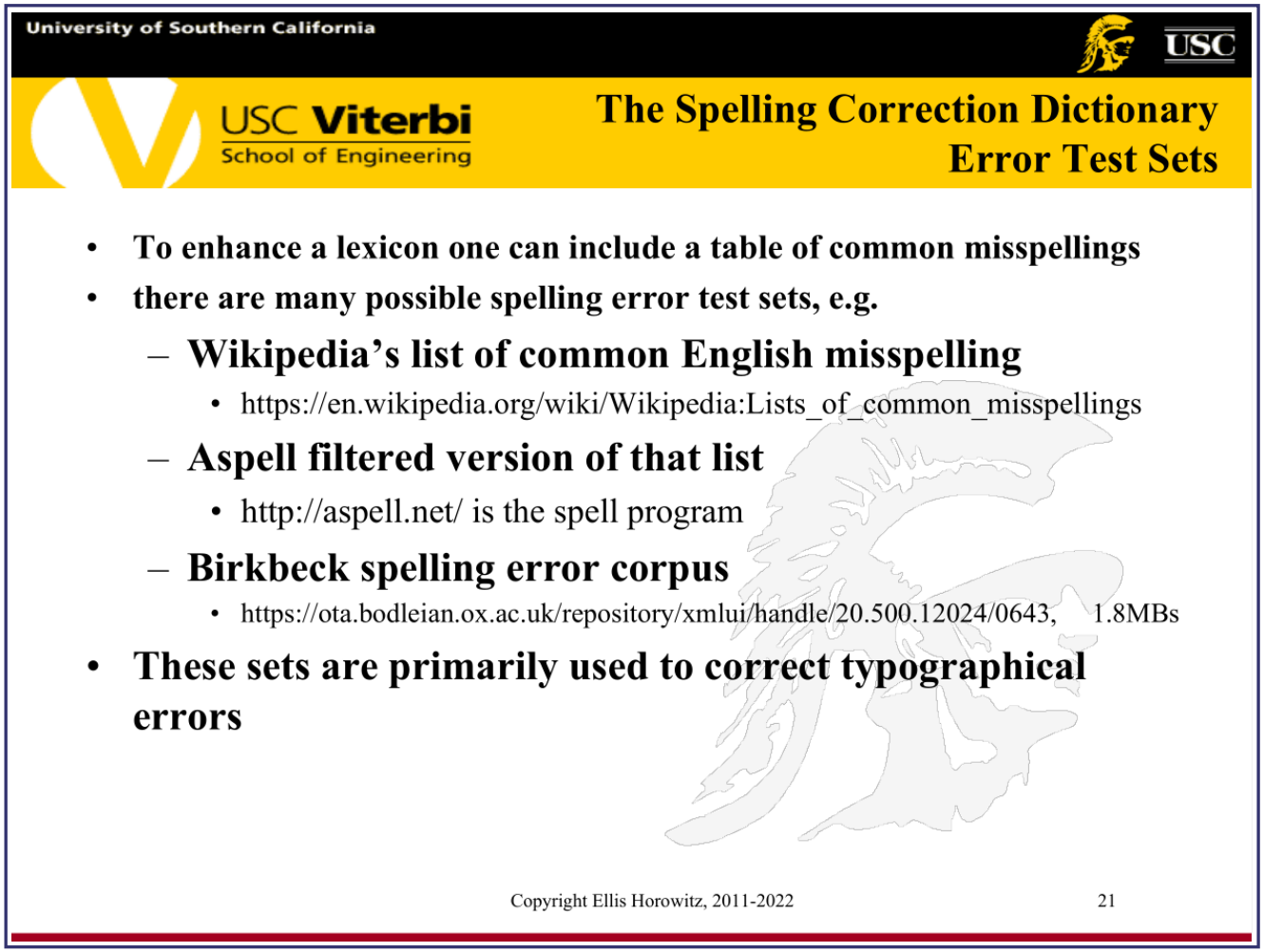
- **a prefix tree (sometimes called a trie from the word retrieval) is a tree of degree >=2 in which the branching at any level is determined by a portion of the key**



branch nodes take you down the tree to element nodes; At any stage one is pointing at all keyword matches that contain the same prefix; Computing time for retrieval is *O(m)* where *m* is the length of the string, at the expense of increased storage

 20

••


University of Southern California
USC
USC Viterbi
School of Engineering
The Spelling Correction Dictionary
Error Test Sets
• To enhance a lexicon one can include a table of common misspellings
• there are many possible spelling error test sets, e.g.
– Wikipedia's list of common English misspelling
• https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings
– Aspell filtered version of that list
• http://aspell.net/ is the spell program
– Birkbeck spelling error corpus
• https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0643, 1.8MBs
• These sets are primarily used to correct typographical errors
Copyright Ellis Horowitz, 2011-2022
21

••

University of Southern California

USC Viterbi School of Engineering

Using N-Grams For Spelling Correction

- An ***n*-gram model** is a type of probabilistic language model for predicting the next item in a sequence
- Two benefits of *n*-gram models (and algorithms that use them) are simplicity and scalability – with larger *n*, a model can store more context with a well-understood space–time tradeoff, enabling small experiments to scale up efficiently

- **Sample of 3-gram sequences**
- ceramics collectables fine (130)
- ceramics collected by (52)
- ceramics collectible pottery (50)
- ceramics collectibles cooking (45)

Sample of 4-gram sequences and # of times appeared

serve as the incoming (92)
serve as the incubator (99)
serve as the independent (794)
serve as the index (223)
serve as the indication (72)
serve as the indicator (120)

a query such as "serve as the indapendant" would match the above

 22

••

University of Southern California

USC Viterbi School of Engineering

Google's N-Gram Data

- Google has collected and uses a great deal of N-gram data
- Google is using the Linguistics Data Consortium to distribute more than one trillion words they have extracted from public web pages
- Below is a statistical summary of the data they are distributing

File sizes: approx. 24 GB compressed (gzip'ed) text files

Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

••

University of Southern California

USC Viterbi School of Engineering

Some References

- ***How Difficult is it to Develop a Perfect Spell-checker? A Cross-linguistic Analysis through Complex Network Approach*, Monojit Choudhury1, Markose Thomas2, Animesh Mukherjee1, Anupam Basu1, and Niloy Ganguly1**

http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=52A3B869596656C9DA285DCE83A0339F?doi=10.1.1.146.4390&rep=rep1&type=pdf

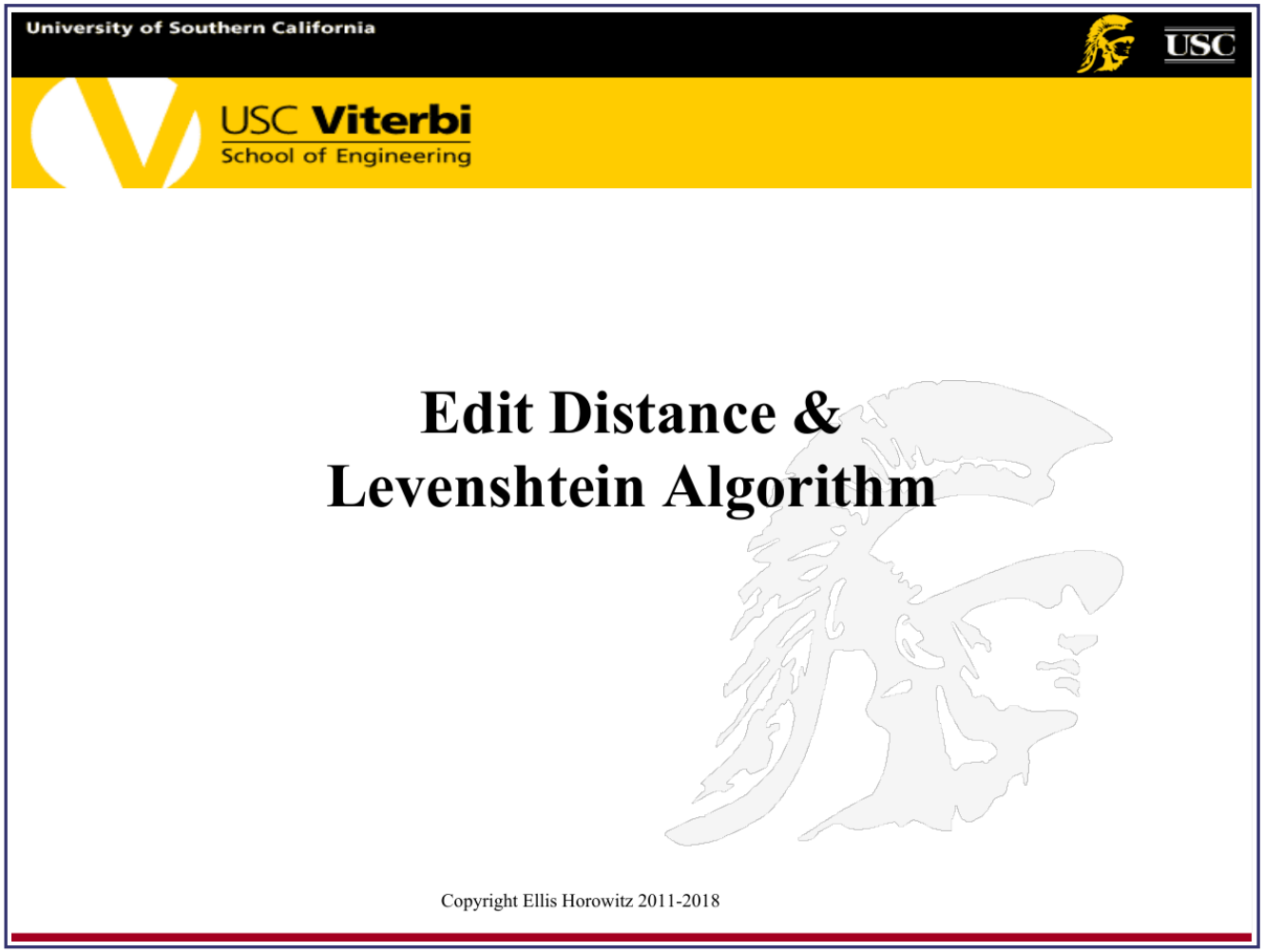
- ***Using the web for language independent spellchecking and autocorrection* by C. Whitelaw et al Proc. 2009 Conf. on Empirical Methods in Natural Language Processing, pp890-899**

http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/pubs/archive/36180.pdf

Spell Checking by Computer, by Roger Mitton,

http://www.dcs.bbk.ac.uk/~roger/spellchecking.html

••


University of Southern California
USC
USC Viterbi
School of Engineering
Edit Distance &
Levenshtein Algorithm
Copyright Ellis Horowitz 2011-2018

••

University of Southern California

USC Viterbi School of Engineering

Edit Distance

- the minimum edit distance between two strings is the minimum number of editing operations
  - insertion
  - deletion
  - substitution

  needed to transform one into the other

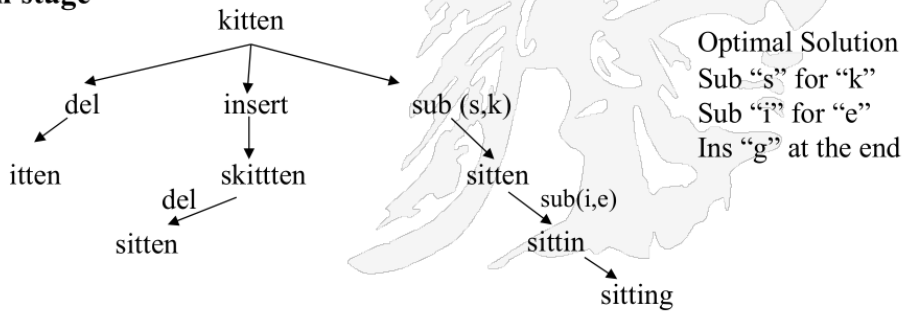
••

University of Southern California

USC Viterbi School of Engineering

How to Find the Minimum Edit Distance

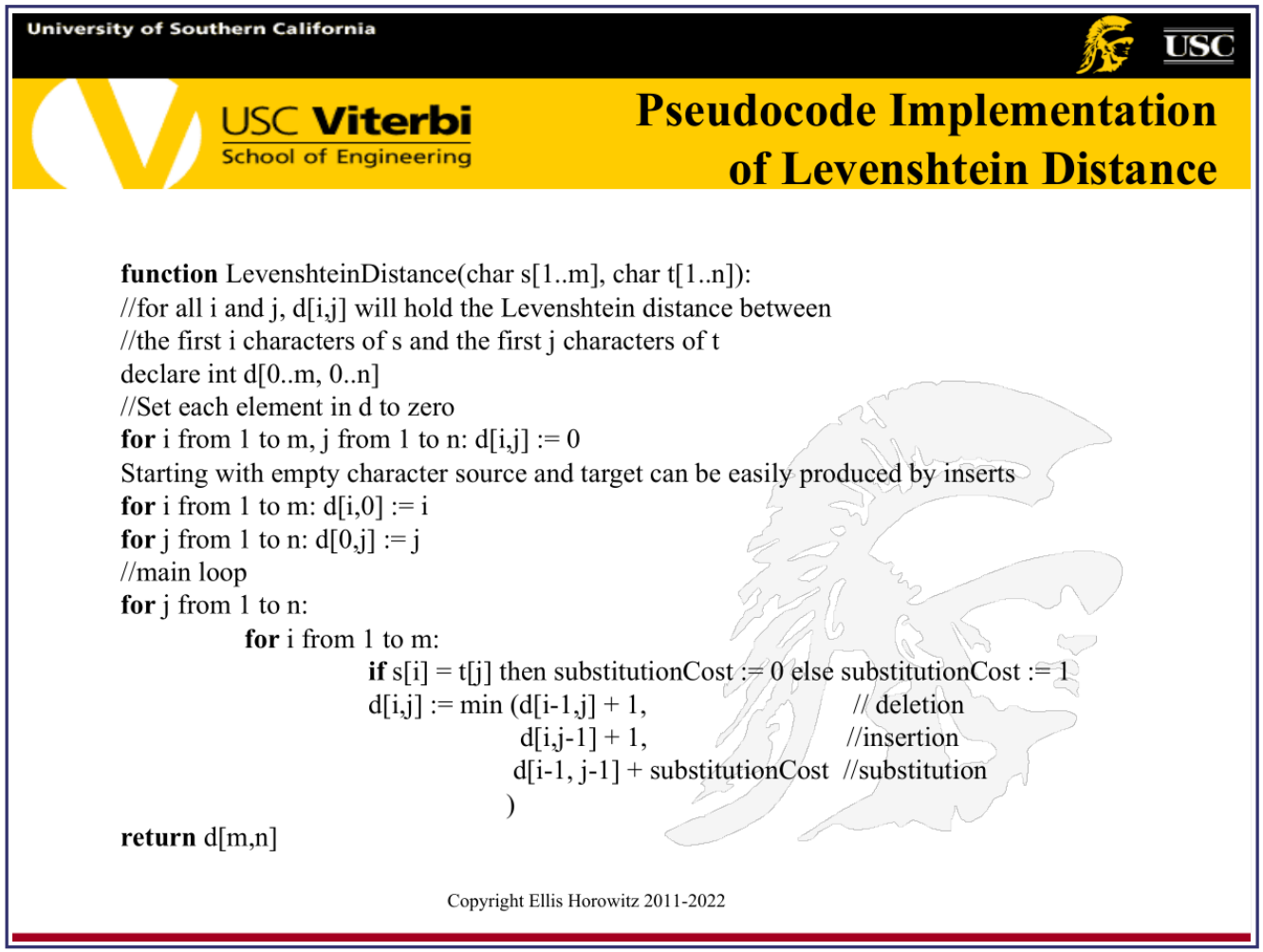
- Searching for a path (sequence of edits) from the start string to the final string
  - initial state: word we're transforming (e.g. kitten)
  - operators: insert, delete, substitute
  - goal state: the word we're trying to get to (e.g. sitting)
  - path cost: what we want to minimize, the number of edits
- If we blindly generate all possible paths in an effort to produce the goal state, our algorithm will take exponentially long
- But we realize that we needn't do that, we may only follow the path that is optimal at each stage



Optimal Solution
Sub "s" for "k"
Sub "i" for "e"
Ins "g" at the end

 34

••



You can learn more about the Levenshtein algorithm here.

And, you can run the algorithm here, to understand how the edit distance is incrementally computed; this is a minimal version.

This is an example of how the Levenshtein distance is used in practice.

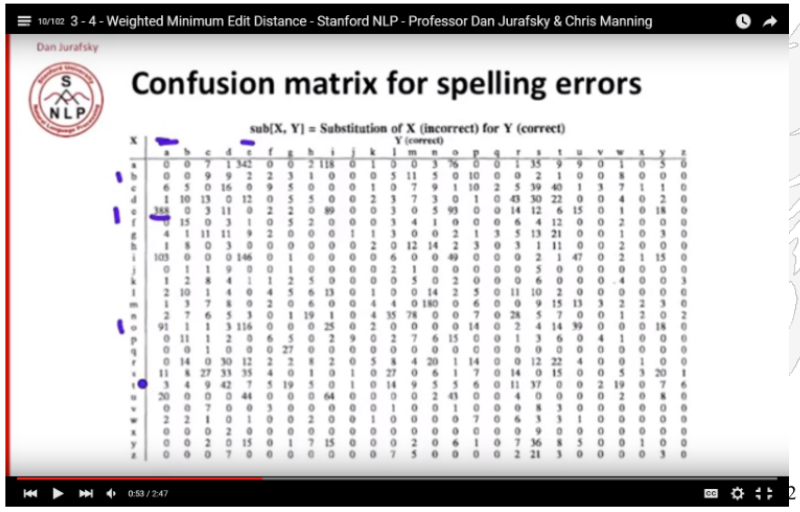
••

University of Southern California

USC Viterbi School of Engineering

Weighted Edit Distance

- **why would we add weights to the computation?**
  - **spell correction: some letters are more likely to be mistyped than others**
- a **confusion matrix** is a specific table layout that allows visualization of the performance of an algorithm; each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice-versa)



the confusion matrix for spelling errors shows us, e.g. that "e" is most often confused with "a",
and that "i" is often confused with both "e" and "a"

42