Leveraging Knowledge Bases for Contextual Entity Exploration

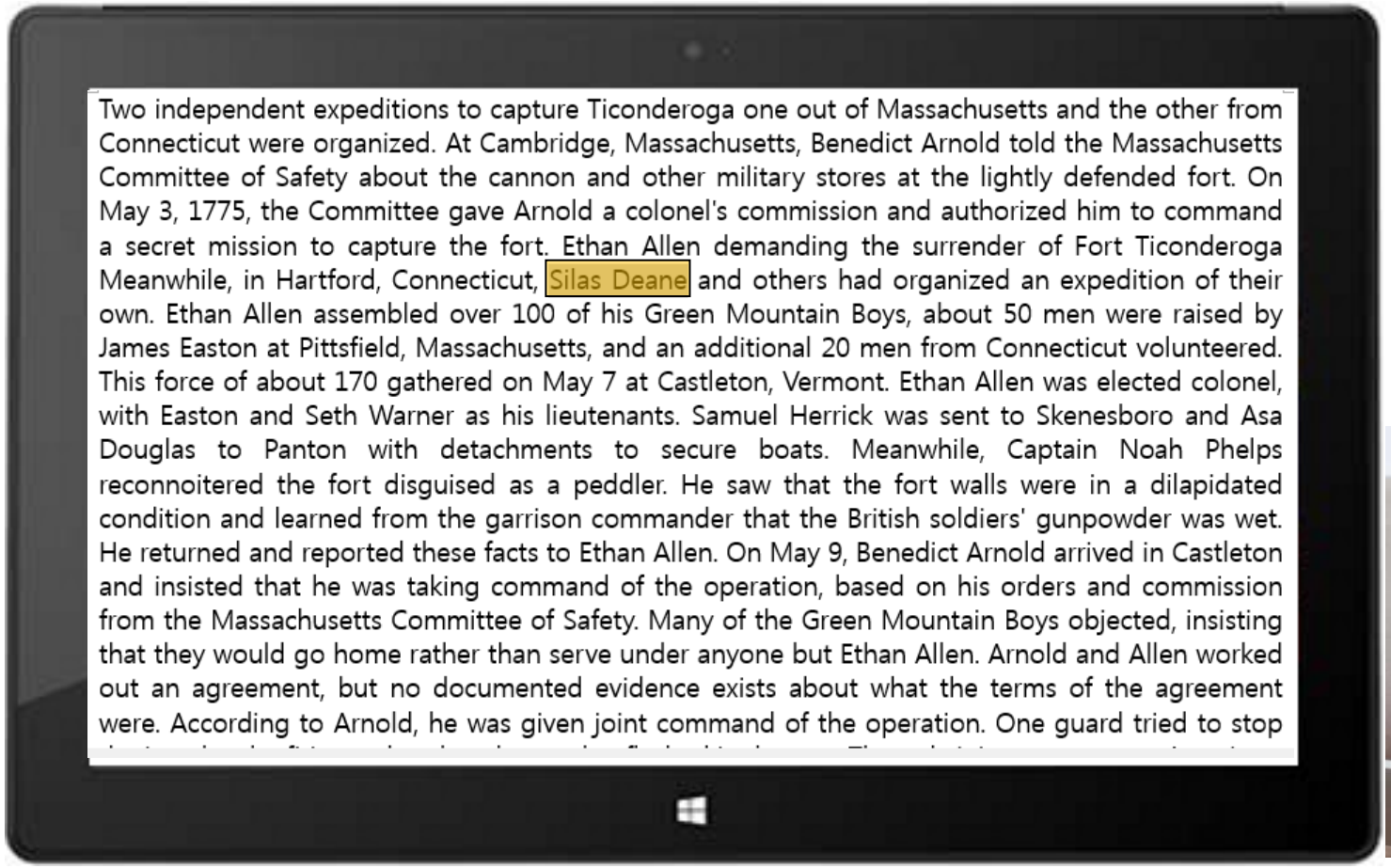
Joonseok Lee, Google

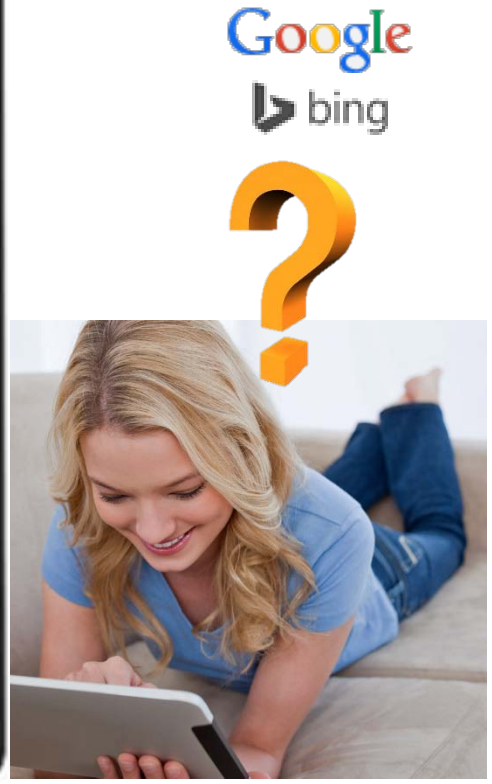
Ariel Fuxman, Google

Bo Zhao, LinkedIn

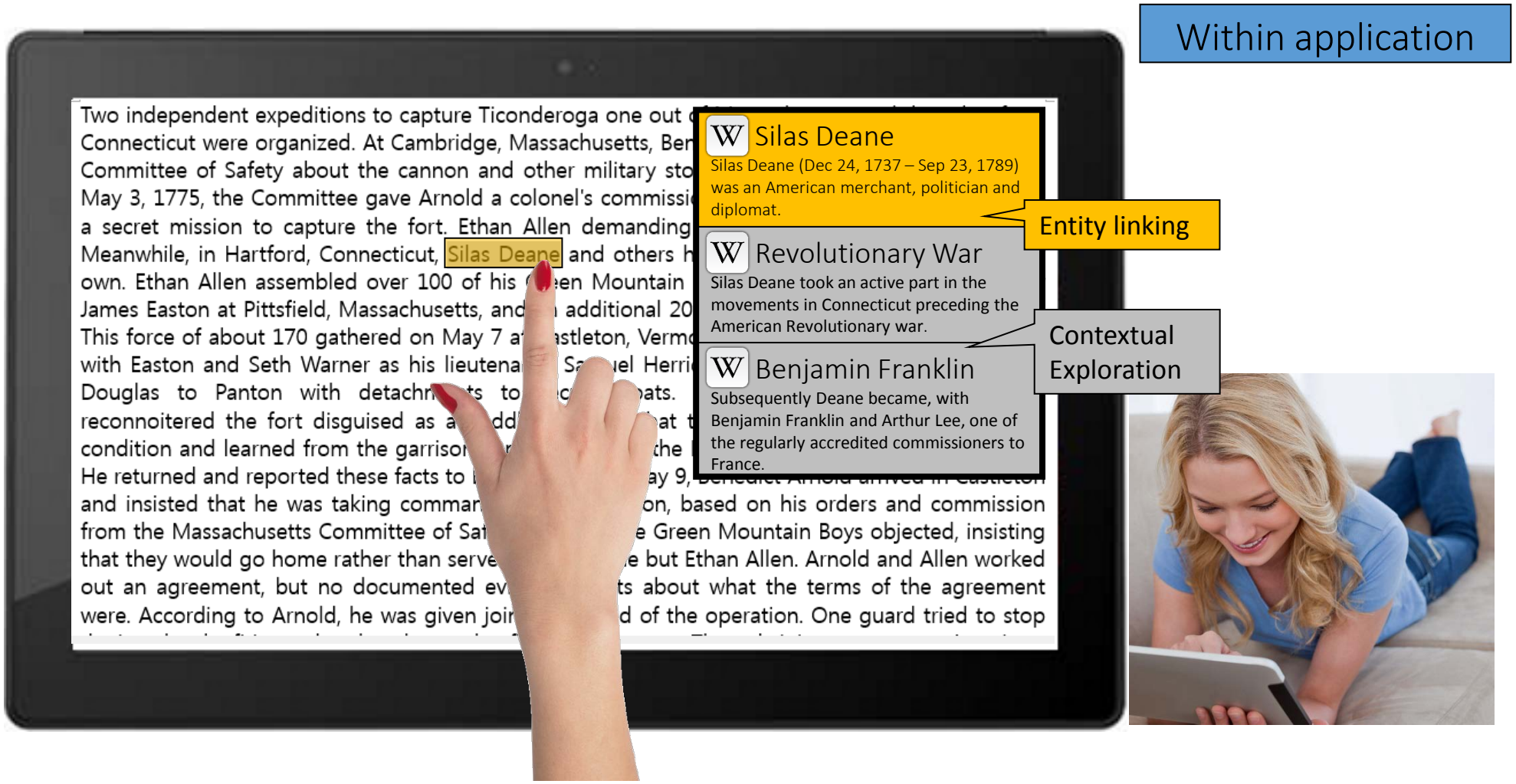
Yuanhua Lv, Microsoft Research

Contextual Insight Problem

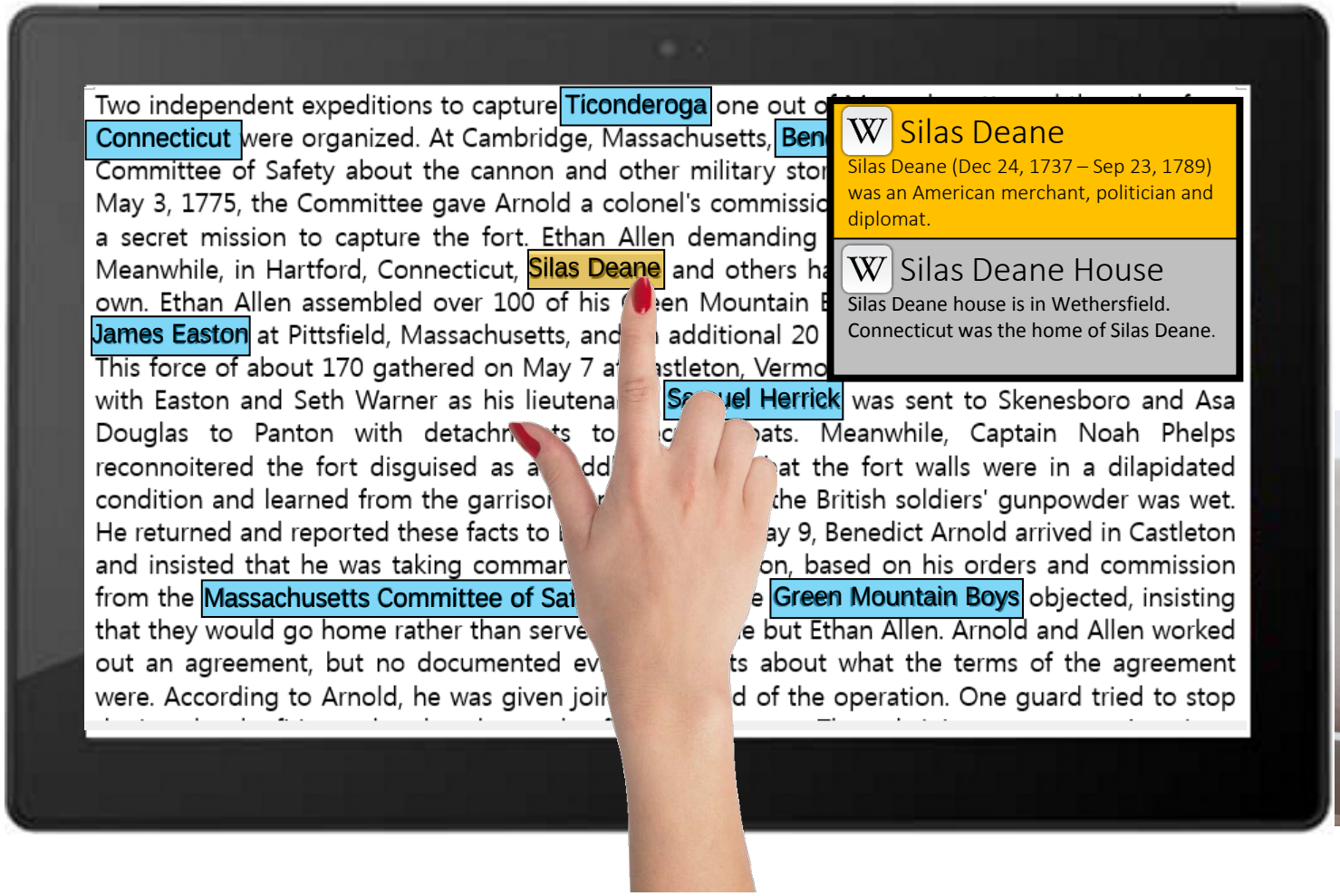


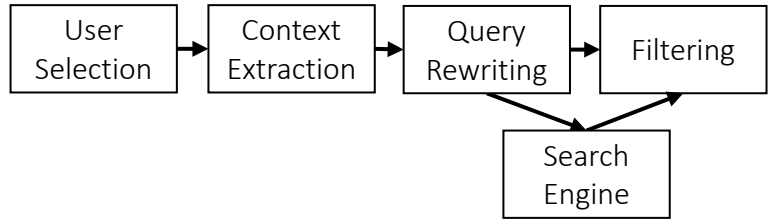


Contextual Insight Problem

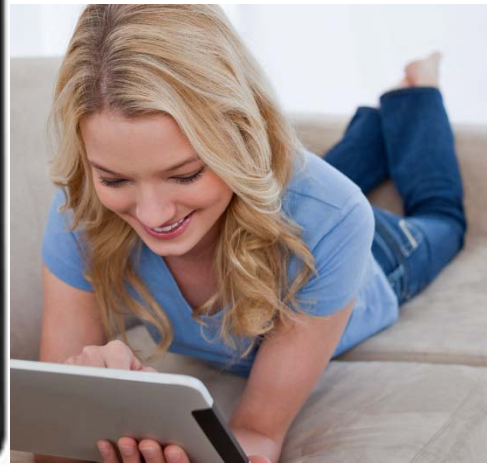


Query Rewriting Approaches

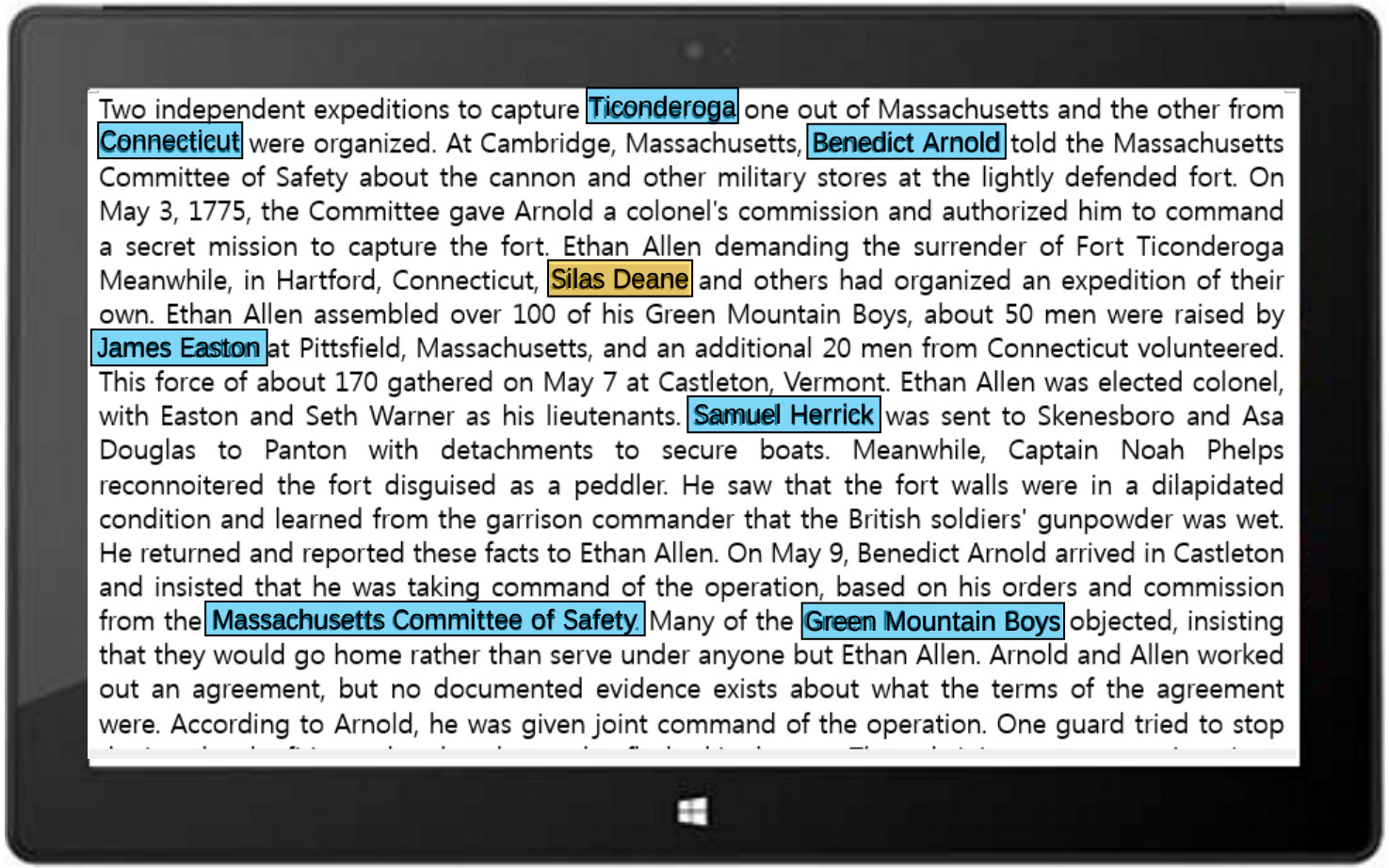


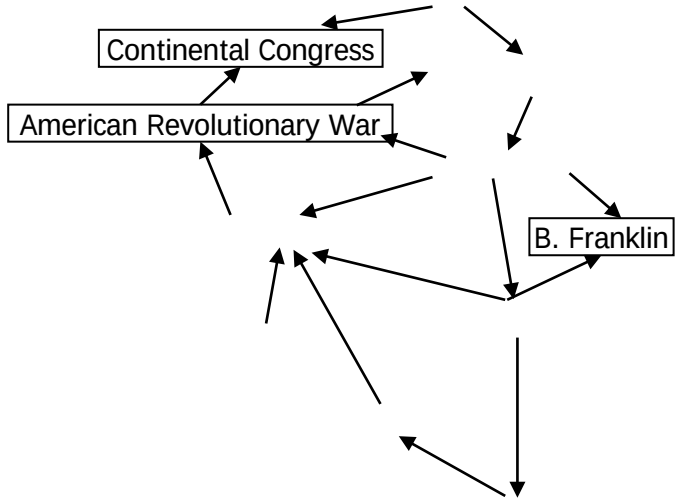


"Silas Deane" rankonly:Connecticut rankonly:Ticonderoga

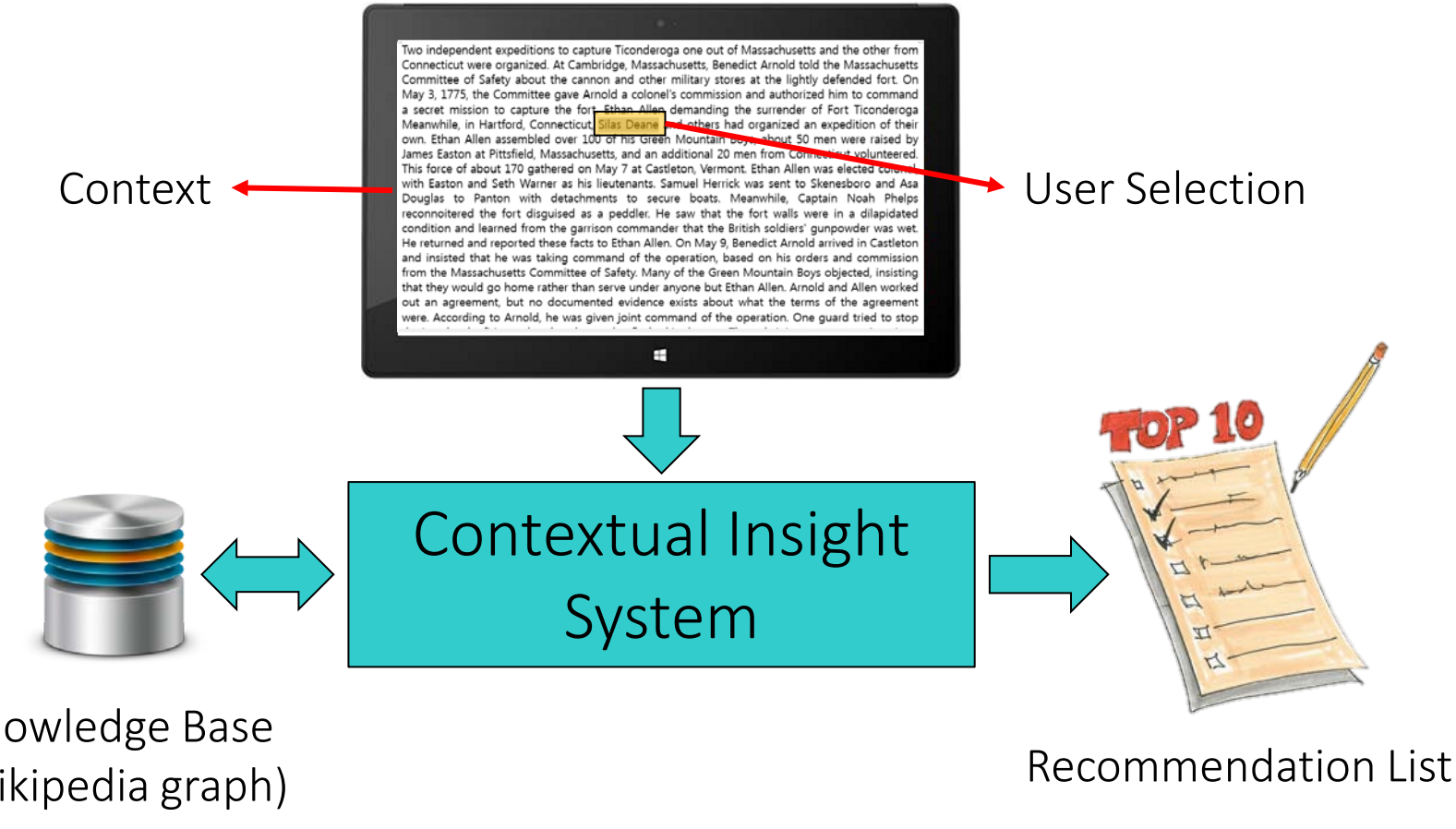


Proposed Approach: Leveraging Semantics in KB





Problem Definition



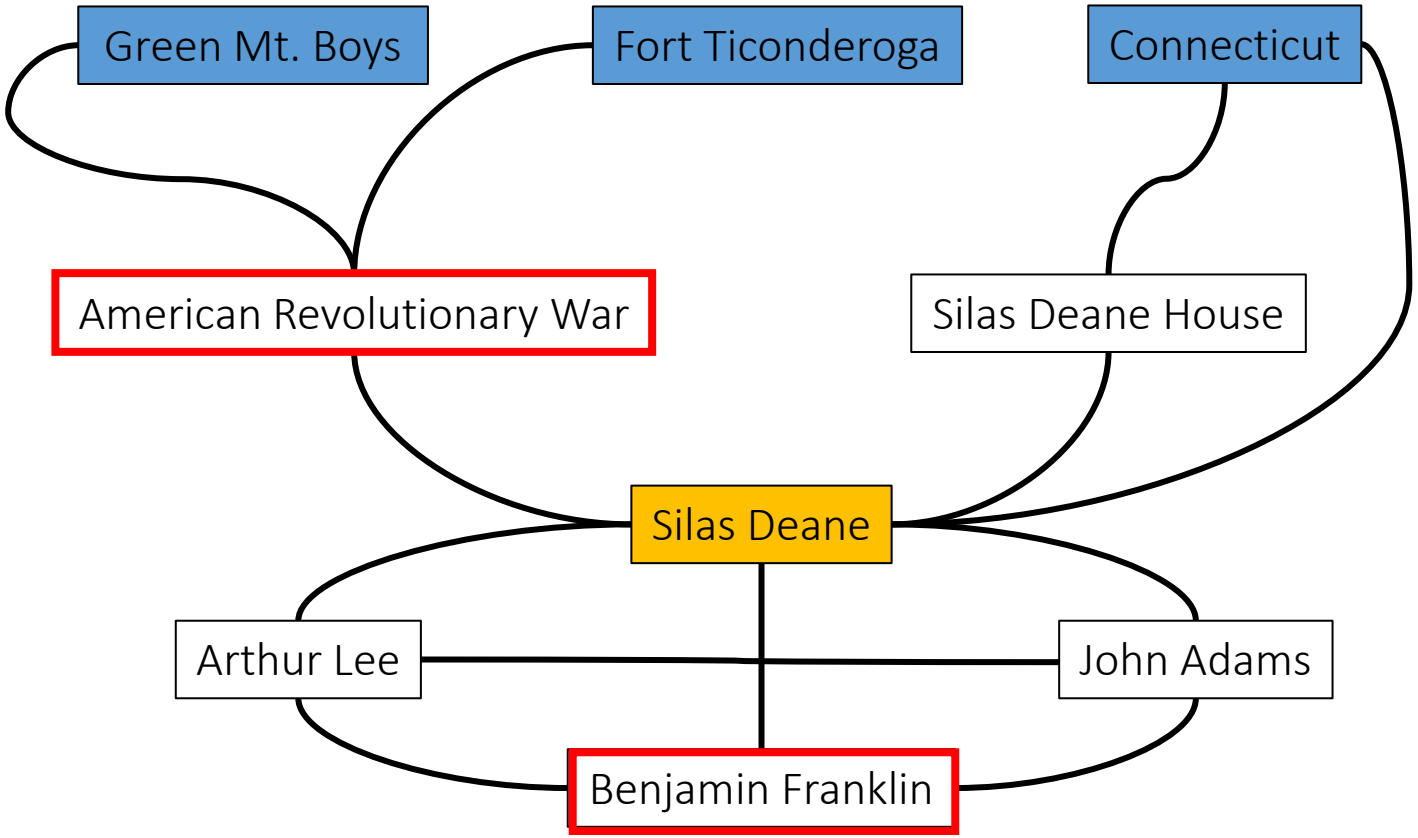
Previous work leveraging Wikipedia graph:

[Yeh2009] WikiWalk: random walks on Wikipedia for semantic relatedness.

[Agrawal2014] Similarity Search using Concept Graphs.

[Milne2008] An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links.

Focused Sub-graph Construction

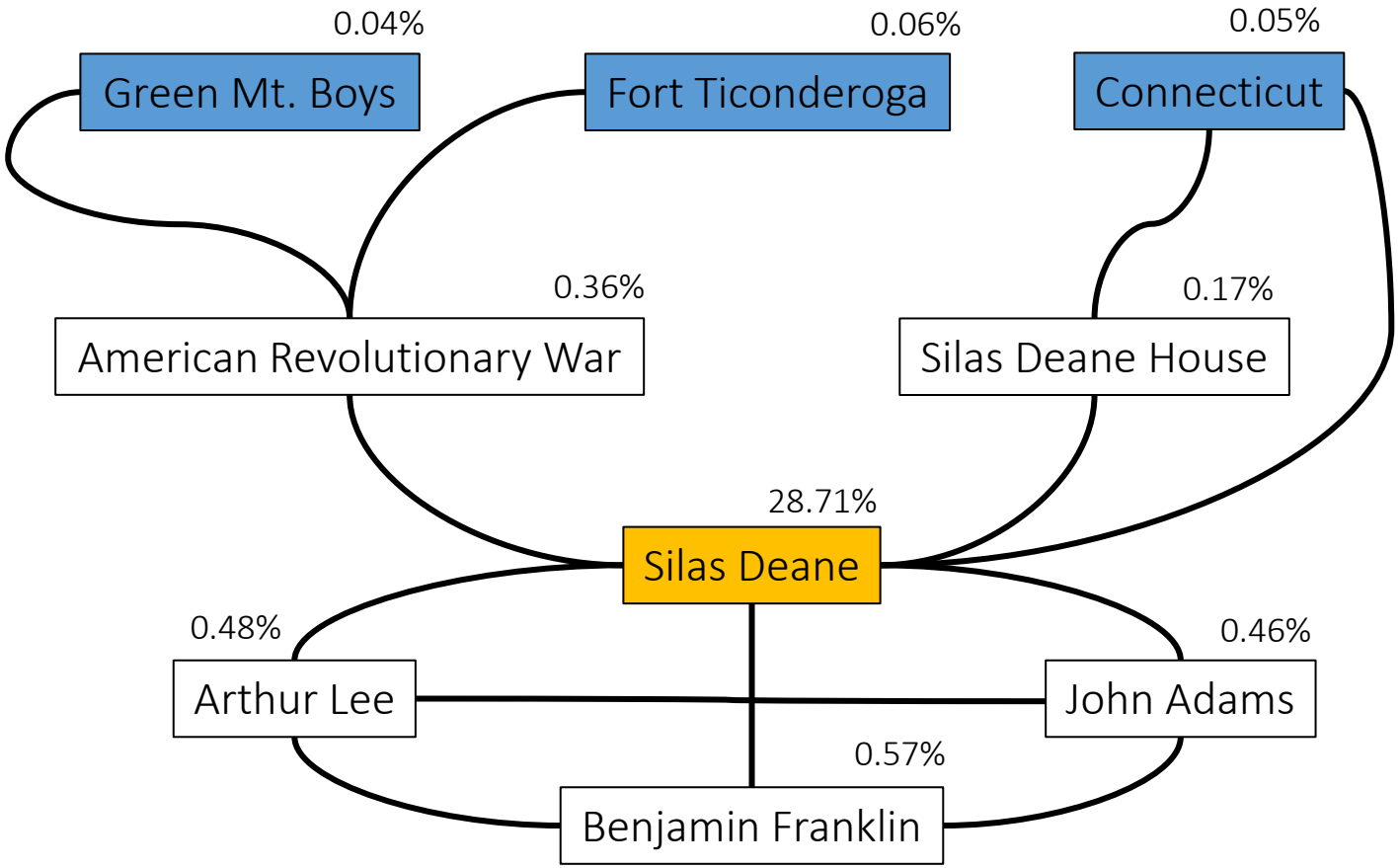


User Selection

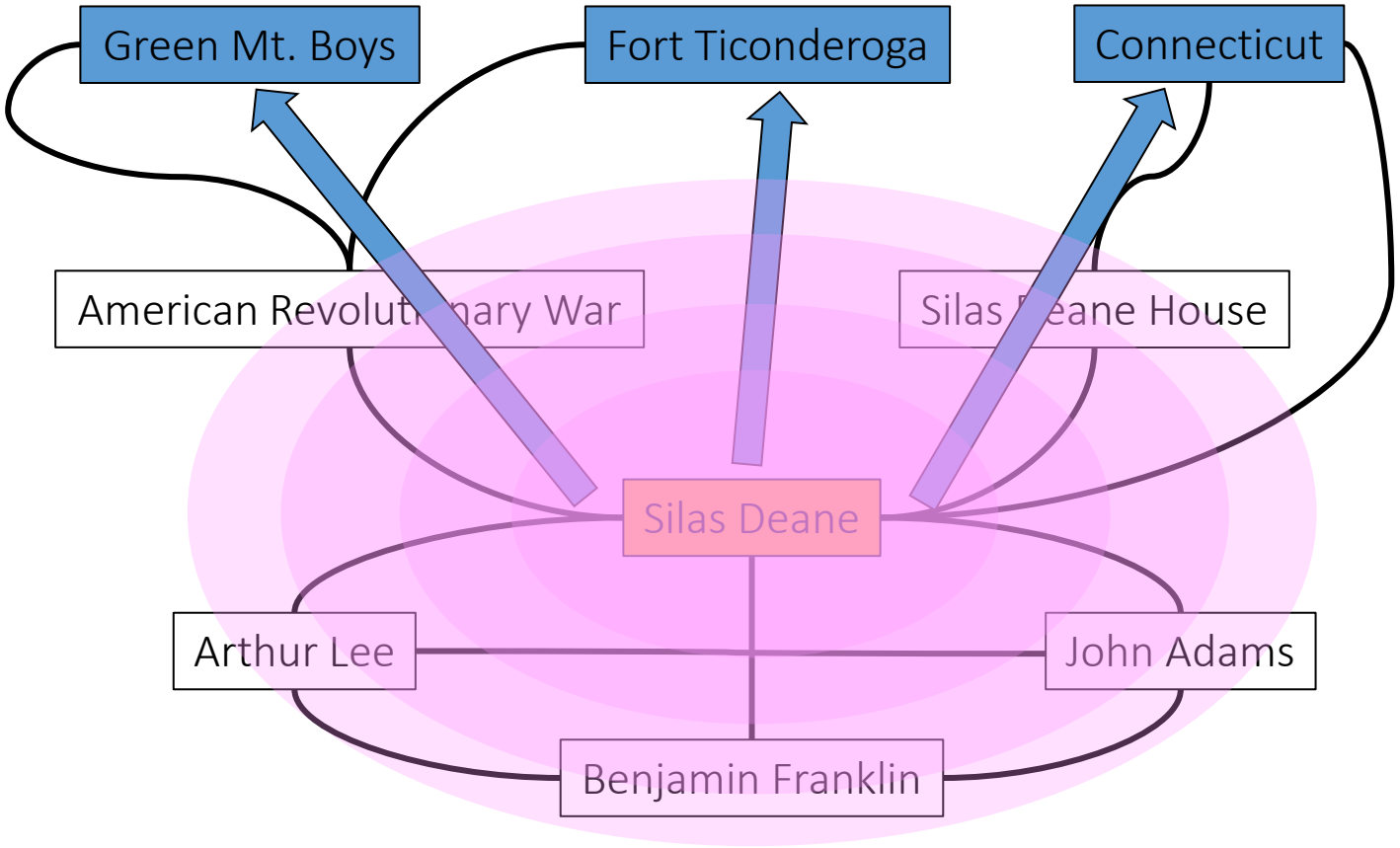
Context Nodes

General Nodes

Personalized Random Walk



Incorporating Contextual Signals



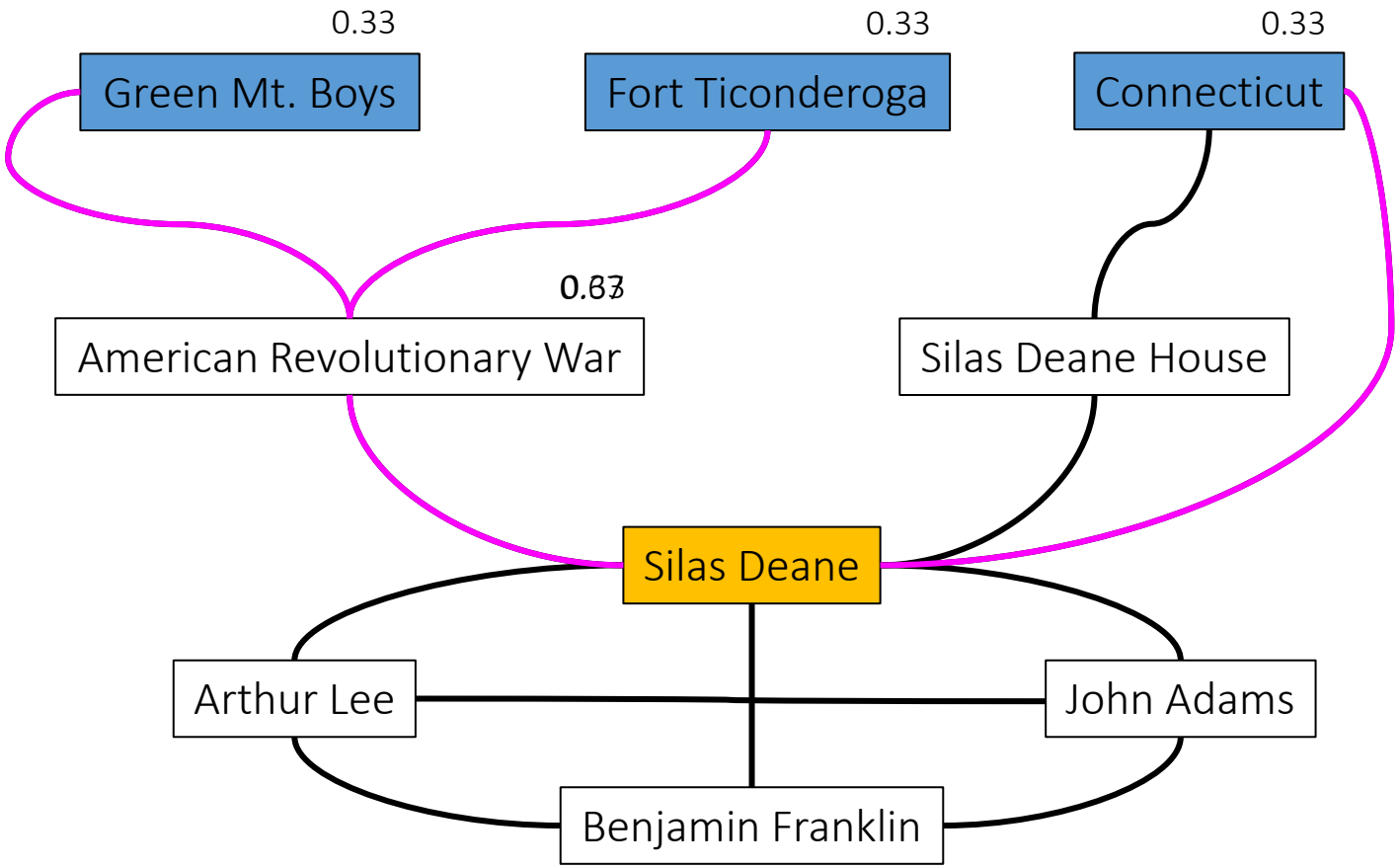
Context-Selection Betweenness

- Retrieve nodes that play an important role connecting the user selection and context nodes.
- Idea from **Betweenness Centrality**: the number of shortest paths from all vertices to all others that pass through that node. [Freeman1977]
- Credit each node when a shortest path from the user selection to each context node passes through it.

$$BC(v) = \frac{1}{Z} \sum_{c:v \in sp(p,c)} \frac{w(p,c)}{k \cdot l(p,c)}$$

$$w(p,c) = \max(\theta - NWD(p,c), 0)$$

Context-Selection Betweenness Example



Score Aggregation

Random Walk score:
How many times the user is likely to visit the page *v*.

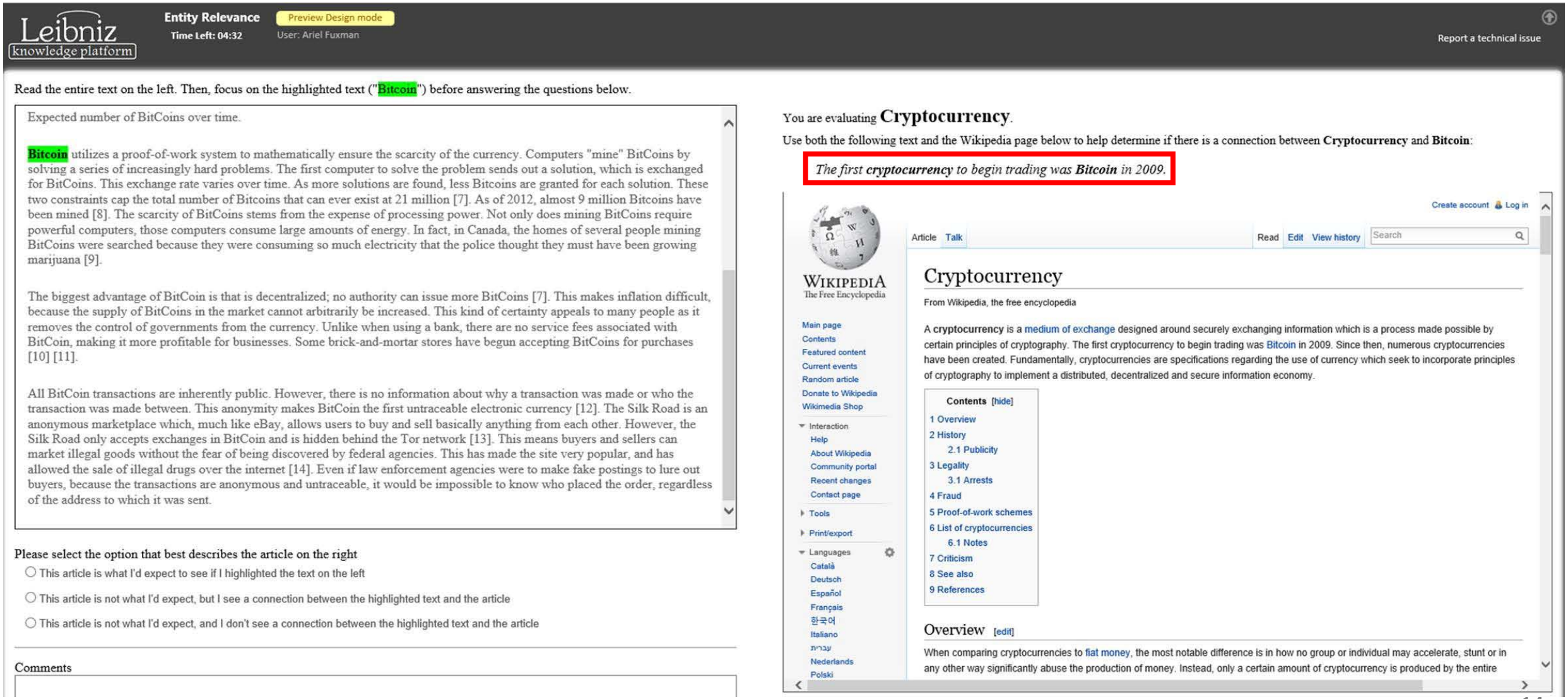
Context-Selection Betweenness score:
Expected number of shortest paths from user selection to any context page through *v*.

$$Relevance(v) = |V|RW(v) + \alpha \frac{|C|}{|V|}|C|CSB(v)$$

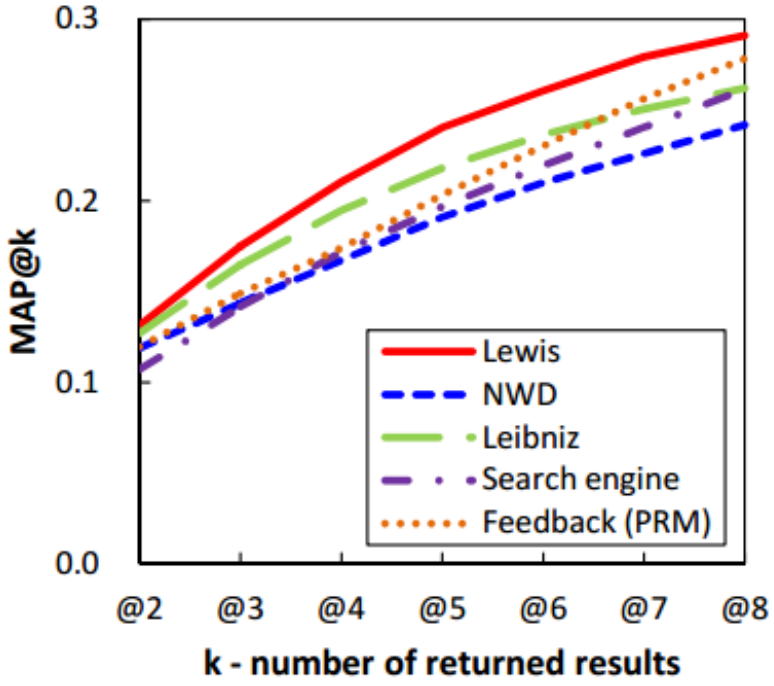
Evaluation: Crowd-sourced Data

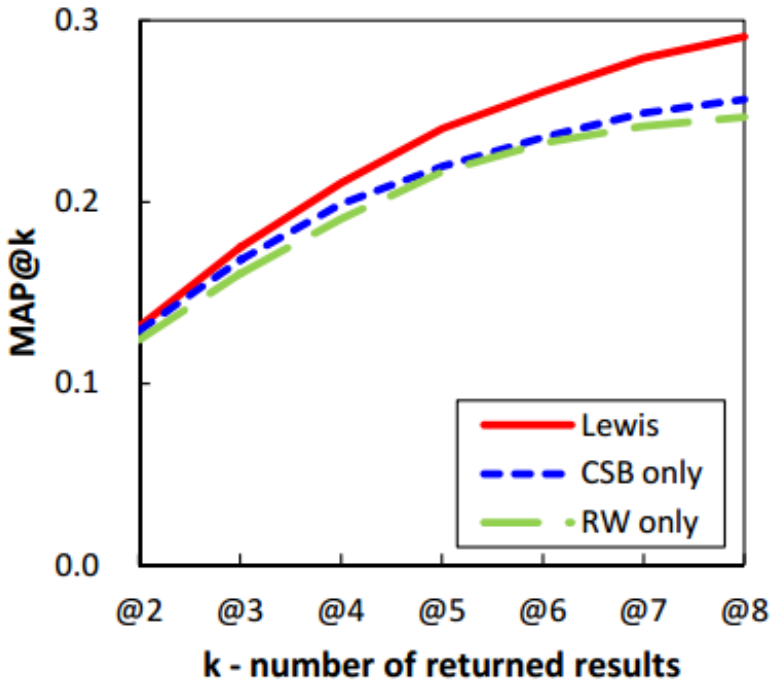
- Corpora
  - 2,600 textbooks from Wikibooks
- Foci of attention
  - Workers selected phrases for which they would like to learn more.
  - 100 workers per paragraph (500 paragraphs per corpus).
- Relevance labeling
  - Each (focus, result) pair evaluated by 10 crowd workers.
  - Large scale: over 145,000 labels.

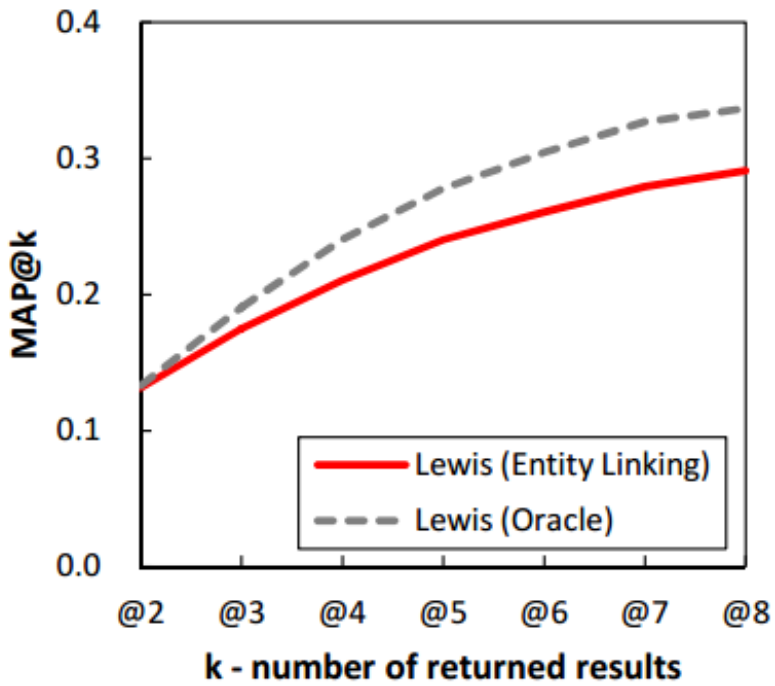
Evaluation


Leibniz knowledge platform
Entity Relevance
Time Left: 04:32
Preview Design mode
User: Ariel Fuxman
Report a technical issue
Read the entire text on the left. Then, focus on the highlighted text ("Bitcoin") before answering the questions below.
Expected number of BitCoins over time.
Bitcoin utilizes a proof-of-work system to mathematically ensure the scarcity of the currency. Computers "mine" BitCoins by solving a series of increasingly hard problems. The first computer to solve the problem sends out a solution, which is exchanged for BitCoins. This exchange rate varies over time. As more solutions are found, less Bitcoins are granted for each solution. These two constraints cap the total number of Bitcoins that can ever exist at 21 million [7]. As of 2012, almost 9 million Bitcoins have been mined [8]. The scarcity of BitCoins stems from the expense of processing power. Not only does mining BitCoins require powerful computers, those computers consume large amounts of energy. In fact, in Canada, the homes of several people mining BitCoins were searched because they were consuming so much electricity that the police thought they must have been growing marijuana [9].
The biggest advantage of BitCoin is that is decentralized; no authority can issue more BitCoins [7]. This makes inflation difficult, because the supply of BitCoins in the market cannot arbitrarily be increased. This kind of certainty appeals to many people as it removes the control of governments from the currency. Unlike when using a bank, there are no service fees associated with BitCoin, making it more profitable for businesses. Some brick-and-mortar stores have begun accepting BitCoins for purchases [10] [11].
All BitCoin transactions are inherently public. However, there is no information about why a transaction was made or who the transaction was made between. This anonymity makes BitCoin the first untraceable electronic currency [12]. The Silk Road is an anonymous marketplace which, much like eBay, allows users to buy and sell basically anything from each other. However, the Silk Road only accepts exchanges in BitCoin and is hidden behind the Tor network [13]. This means buyers and sellers can market illegal goods without the fear of being discovered by federal agencies. This has made the site very popular, and has allowed the sale of illegal drugs over the internet [14]. Even if law enforcement agencies were to make fake postings to lure out buyers, because the transactions are anonymous and untraceable, it would be impossible to know who placed the order, regardless of the address to which it was sent.
Please select the option that best describes the article on the right
This article is what I'd expect to see if I highlighted the text on the left
This article is not what I'd expect, but I see a connection between the highlighted text and the article
This article is not what I'd expect, and I don't see a connection between the highlighted text and the article
Comments
You are evaluating Cryptocurrency.
Use both the following text and the Wikipedia page below to help determine if there is a connection between Cryptocurrency and Bitcoin:
The first cryptocurrency to begin trading was Bitcoin in 2009.
Create account
Log in
Article
Talk
Read
Edit
View history
Search
WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Tools
Print/export
Languages
Català
Deutsch
Español
Français
한국어
Italiano
עברית
Nederlands
Polski
Cryptocurrency
From Wikipedia, the free encyclopedia
A cryptocurrency is a medium of exchange designed around securely exchanging information which is a process made possible by certain principles of cryptography. The first cryptocurrency to begin trading was Bitcoin in 2009. Since then, numerous cryptocurrencies have been created. Fundamentally, cryptocurrencies are specifications regarding the use of currency which seek to incorporate principles of cryptography to implement a distributed, decentralized and secure information economy.
Contents [hide]
1 Overview
2 History
2.1 Publicity
3 Legality
3.1 Arrests
4 Fraud
5 Proof-of-work schemes
6 List of cryptocurrencies
6.1 Notes
7 Criticism
8 See also
9 References
Overview [edit]
When comparing cryptocurrencies to fiat money, the most notable difference is in how no group or individual may accelerate, stunt or in any other way significantly abuse the production of money. Instead, only a certain amount of cryptocurrency is produced by the entire

Comparing Proposed Components





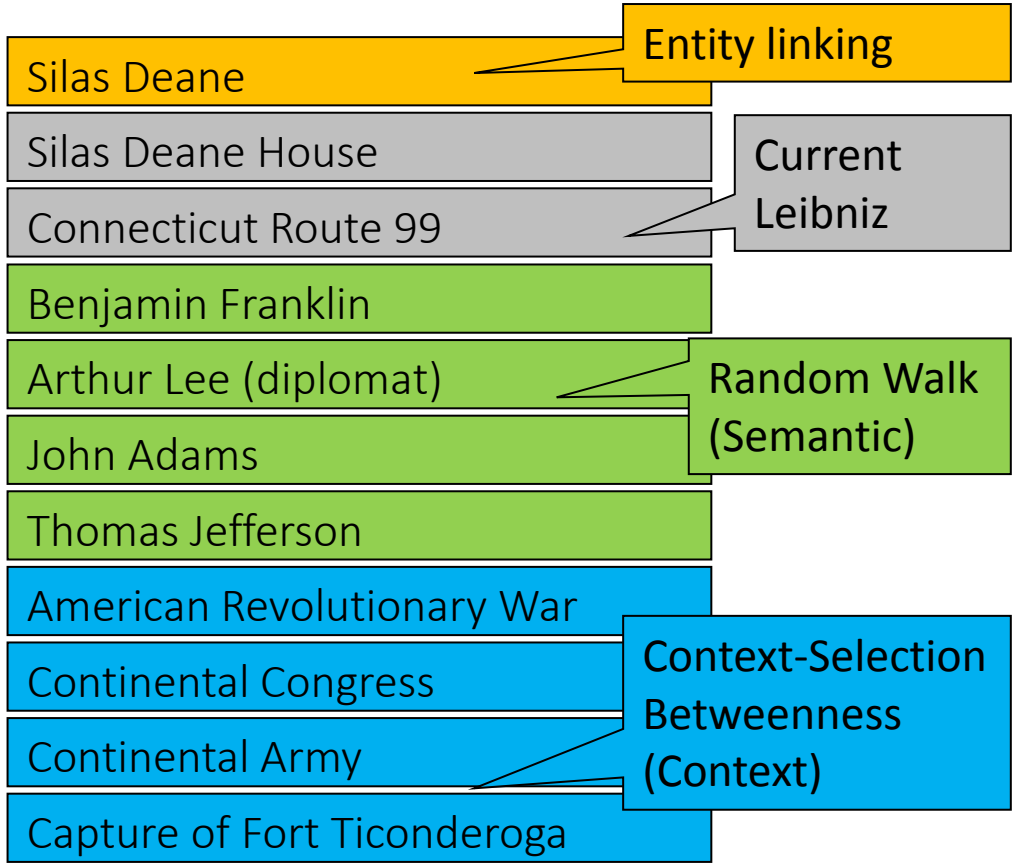


Score decomposition with *Silas Deane* Example

Rank	Page Title	CSB	RW	Total
1	Silas Deane	402.26	6442.53	6844.78
2	American Revolutionary War	49.58	82.42	131.99
3	Benjamin Franklin	2.46	128.17	130.63
4	Arthur Lee (diplomat)	0.00	107.99	107.99
5	Thomas Jefferson	2.46	98.82	101.28
6	Continental Congress	19.58	74.02	93.60
7	Continental Army	30.25	61.45	91.70
8	Capture of Fort Ticonderoga	49.58	39.56	89.14

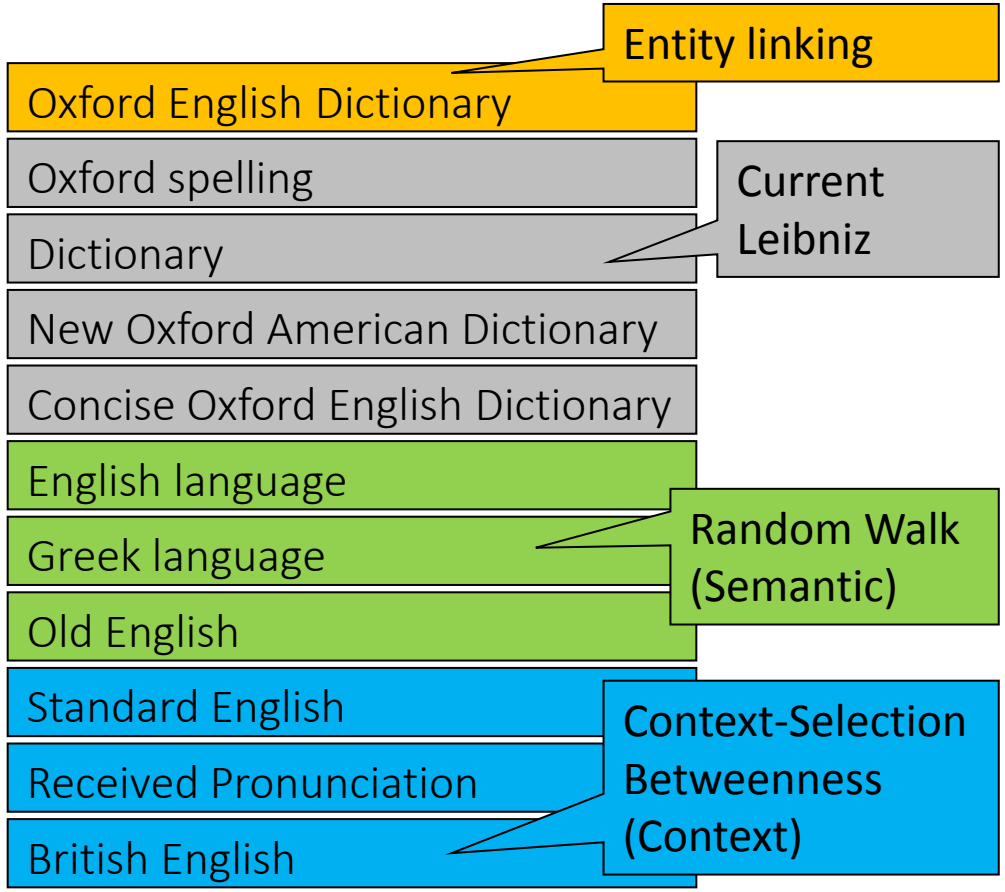
Examples

Two independent expeditions to capture Ticonderoga – one out of Massachusetts and the other from Connecticut – were organized. At Cambridge, Massachusetts, Benedict Arnold told the Massachusetts Committee of Safety about the cannon and other military stores at the lightly defended fort. On May 3, 1775, the Committee gave Arnold a colonel's commission and authorized him to command a secret mission to capture the fort. Ethan Allen demanding the surrender of Fort Ticonderoga Meanwhile, in Hartford, Connecticut, Silas Deane and others had organized an expedition of their own. Ethan Allen assembled over 100 of his Green Mountain Boys, about 50 men were raised by James Easton at Pittsfield, Massachusetts, and an additional 20 men from Connecticut volunteered. This force of about 170 gathered on May 7 at Castleton, Vermont.



Examples

In 1895, when Joyce was in his third year at Belvedere College, he chose Ulysses as his subject for an essay entitled "My Favourite Hero". In English Ulysses is sometimes stressed on the second syllable, and this is the pronunciation required in most verse translations of the Homeric epics. Joyce, however, always referred to his novel as YOOL-i-seez, with the stress on the first syllable. This pronunciation is sanctioned by the Oxford English Dictionary and is used almost universally in Ireland when one is referring to the book. In his design for the cover of the 1949 Random House edition of Ulysses, the American artist Edward McKnight Kauffer emphasized the initial UL, "giving graphic form to the phonetic structure of the title with its accent on the first syllable."



Take-home Messages

- Contextual entity exploration problem
- Knowledge-graph-based solution
  - Semantics incorporated
  - Context-relevant retrieval
  - Nice performance in IR measures
- Crowd-sourced evaluation

Thank you for your attention.