Information Retrieval

KST8R

- K-State's new social network!
  - What features should it have?
  - What data do we need?
  - How should it be stored?
  - How can we retrieve it?

Edgar F. Codd

- Worked for IBM in the late 1960s, early 1970s
- Worked on storage of data in computer systems
- “A Relational Model of Data for Large Shared Data Banks”

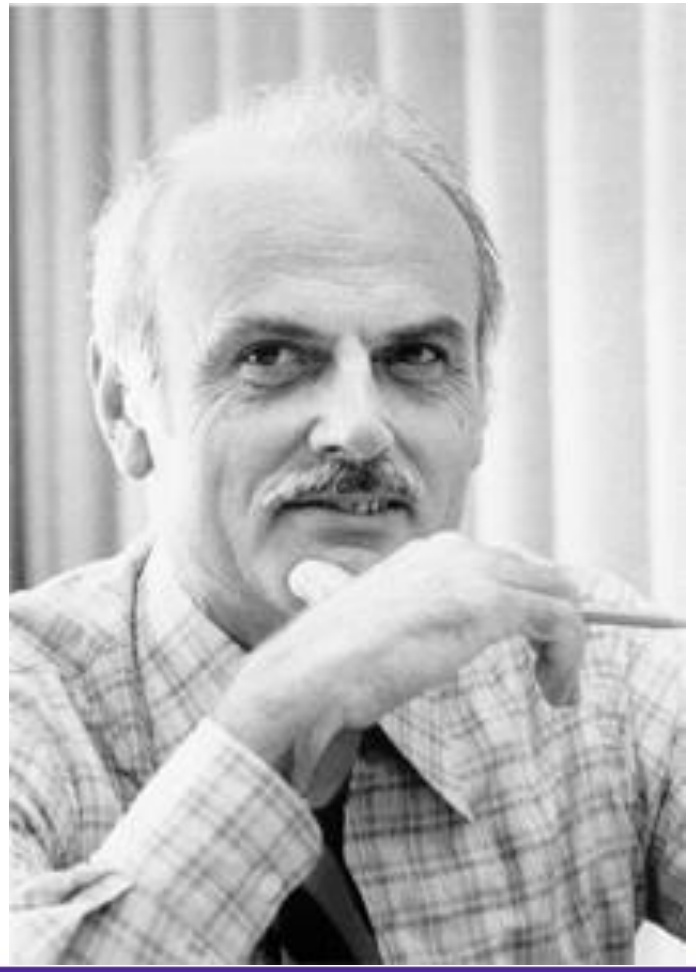


Image Source: Wikipedia

Relational Database

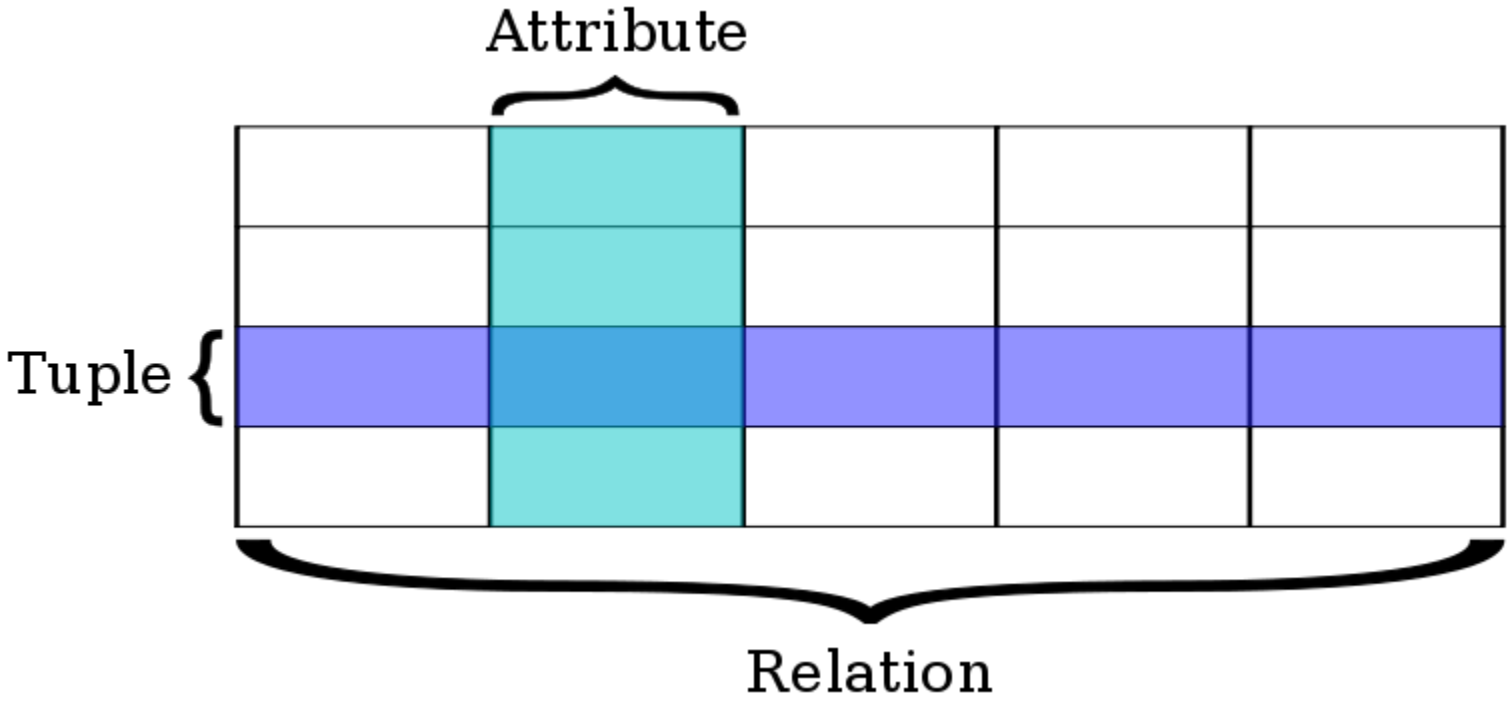


Image Source: Wikipedia

Relational Database

userID	Name	Birthday	Major
weeser	Josh	June 13th	Comp. Sci
johnsmith	John	June 1	Info. Sci
kimc	Kim	February 2nd	Info. Sys
gameguy	Jayson	Dec 26	computersci
sharpie	Reily	Dec. 18th	IS

Why Normalize Data?

- Avoid data anomalies
- Make redesigning easier
- Mirror real-world concepts
- Simplify queries

Related Tables

userID	Name	Birthday	majorID
weeser	Josh	June 13th	1
johnsmith	John	June 1	3
kimc	Kim	February 2nd	2
gameguy	Jayson	Dec 26	1
sharpie	Reily	Dec. 18th	3

majorID	Major	Abbr
1	Computer Science	CS
2	Software Engineering	SE
3	Information Systems	IS

Storing Phone Numbers

userID	Name	Phone1	Phone2
johnsmith	John	555-1234	
smiller	Sheila	555-5134	
gameguy	Jayson	555-1235	555-5134
sharpie	Reily	555-5134	

Many-to-One?

userID	Name
johnsmith	John
smiller	Sheila
gameguy	Jayson
sharpie	Reily

Phone	user1	user2	
555-1234	johnsmith		
555-5134	smiller	gameguy	?sharpie?
555-1235	gameguy		

One-to-Many

userID	Name
johnsmith	John
smiller	Sheila
gameguy	Jayson
sharpie	Reily

userID	phone
johnsmith	555-1234
smiller	555-5134
gameguy	555-1235
gameguy	555-5134
sharpie	555-5134

The World Wide Web

- Where is most of the data in the world stored?
- Internet is like a VERY BIG unstructured database
- Search engines do a decent job
  - ….but how do we go about that?

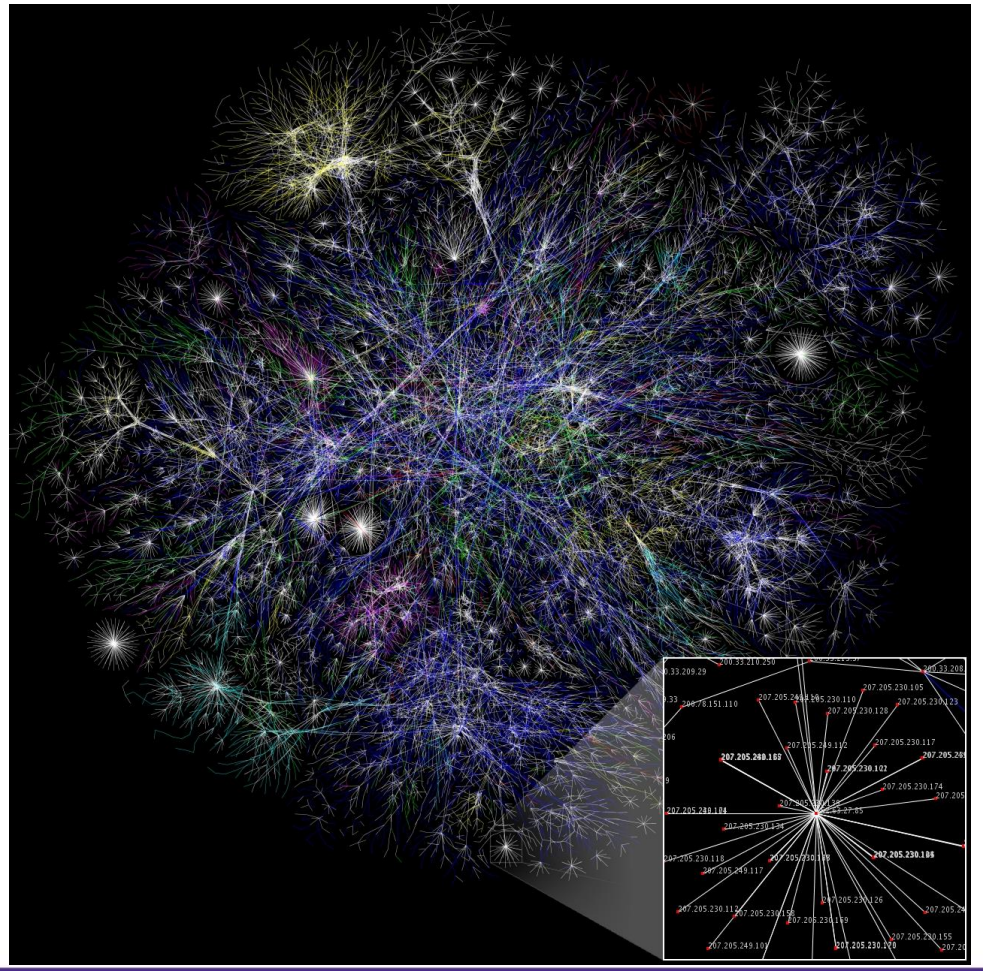
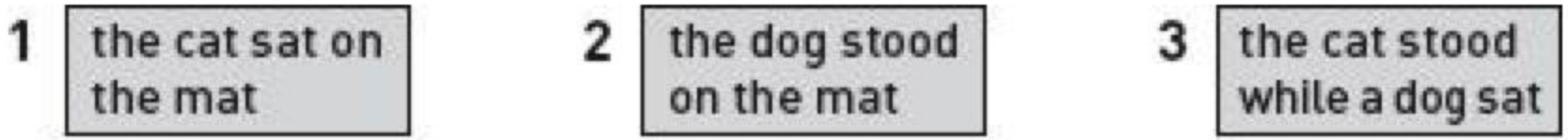


Image Source: Wikipedia

Our World Wide Web

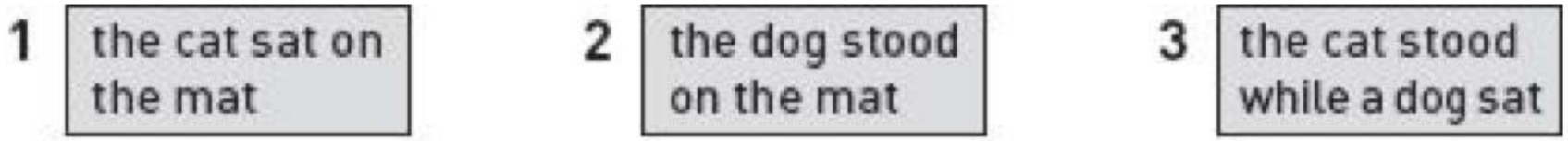
Queries
1. cat
2. dog
3. dog sat
4. "dog stood"
5. cat OR mat
6. cat AND mat



An imaginary World Wide Web that consists of only three pages, numbered 1, 2, and 3.

Image Source: *9 Algorithms that Changed the Future* by John MacCormick

Indexing



An imaginary World Wide Web that consists of only three pages, numbered 1, 2, and 3.

Queries

1. **cat**
2. **dog**
3. **dog sat**
4. ~~"dog stood"~~
5. **cat OR mat**
6. **cat AND mat**

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

World Location

1 | the cat sat on
1 2 3 4
the mat
5 6

2 | the dog stood
1 2 3
on the mat
4 5 6

3 | the cat stood
1 2 3
while a dog sat
4 5 6 7

Queries
1. cat
2. dog
3. dog sat
4. "dog stood"
5. cat OR mat
6. cat AND mat

a	3-5				
cat	1-2	3-2			
dog	2-2	3-6			
mat	1-6	2-6			
on	1-4	2-4			
sat	1-3	3-7			
stood	2-3	3-3			
the	1-1	1-5	2-1	2-5	3-1
while	3-4				

Algorithm

INPUT: a two-word phrase of form, "Word1 Word2"
OUTPUT: an AnswerList, a list of the numbers of the Web pages that contain the phrase

ALGORITHM:
(0. The AnswerList starts with nothing saved in it.)
1. Extract from the table the list of Page#-Position# pairs for Word1. Call it List1.
2. Extract from the table the list of Page#-Position# pairs for Word2. Call it List2.
3. For each page#-pos# pair in List1,
   search List2 to see if there is a pair, page#-(pos#+1).
   (that is, the page# is the same and pos# differs by +1)
   if yes, then include page# in the AnswerList.
   if no, then ignore this pair.
4. Announce all the page numbers in AnswerList

Nearness

- Look up "cat sat" using the list
- Look up:
  - cat stood
  - "cat stood"
  - cat OR stood
- How would that algorithm change?

Ranking

- Which page is more likely to be about dogs? Cats? How could you tell?
- How would you modify the algorithm to account for that?

Metawords

1 <titleStart> my cat <titleEnd> <bodyStart> the cat sat on the mat <bodyEnd>

2 <titleStart> my dog <titleEnd> <bodyStart> the dog stood on the mat <bodyEnd>

3 <titleStart> my pets <titleEnd> <bodyStart> the cat stood while a dog sat <bodyEnd>

a	3-10
cat	1-3 1-7 3-7
dog	2-3 2-7 3-11
mat	1-11 2-11
my	1-2 2-2 3-2
on	1-9 2-9
pets	3-3
sat	1-8 3-12
stood	2-8 3-8
the	1-6 1-10 2-6 2-10 3-6
while	3-9
<bodyEnd>	1-12 2-12 3-13
<bodyStart>	1-5 2-5 3-5
<titleEnd>	1-4 2-4 3-4
<titleStart>	1-1 2-1 3-1

Image Source: *9 Algorithms that Changed the Future* by John MacCormick

AltaVista

- Early web search engine with crawler and indexer
- 1996: 5 servers, 210 GB storage & 4GB RAM on main indexer
- 1998: 20 servers, 500 GB storage & 130 GB RAM, 13 million queries daily
- Bought by Yahoo in in 2003, shut down in 2011

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

Hyperlinks

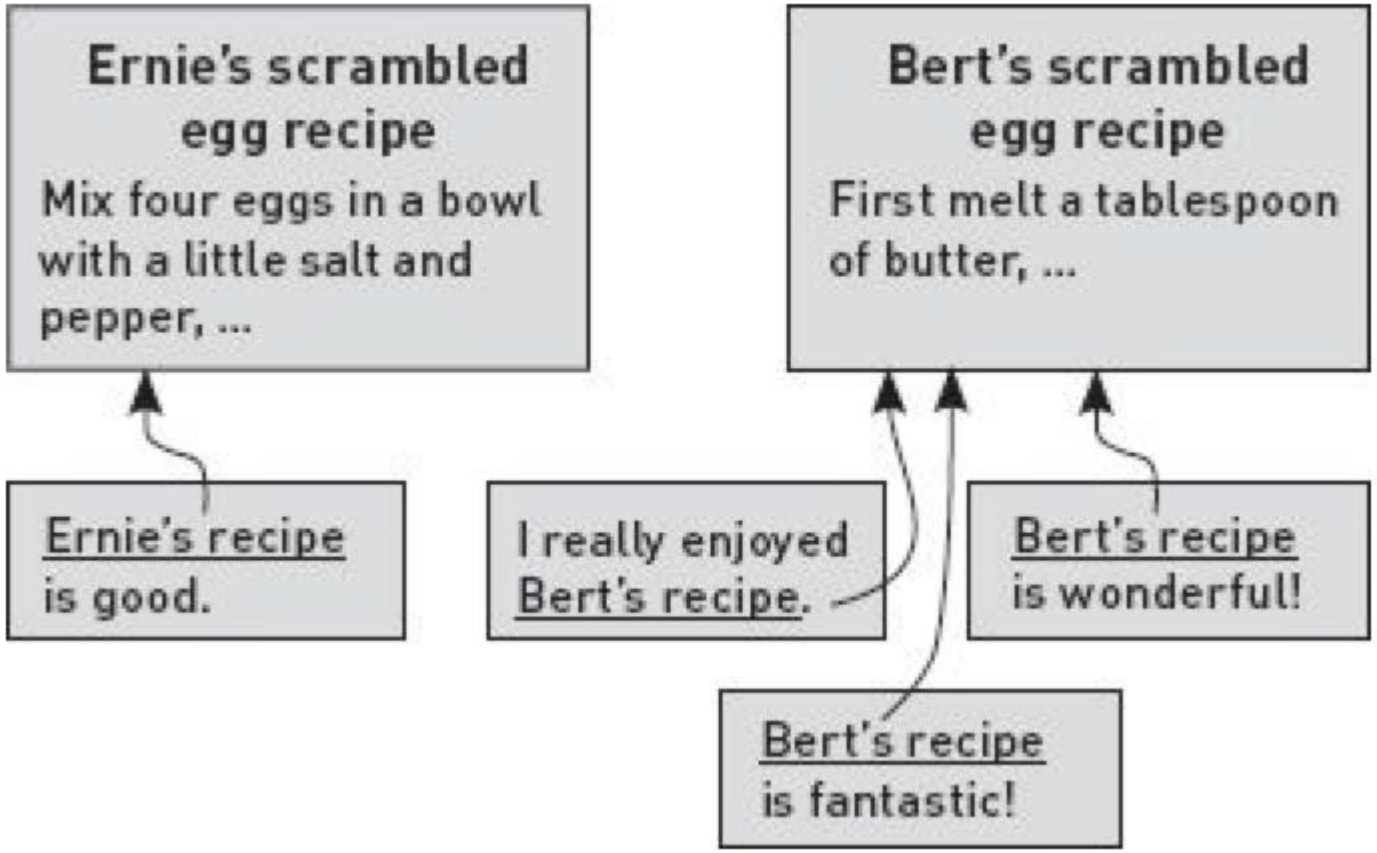


Image Source: *9 Algorithms that Changed the Future* by John MacCormick

Who Wins?

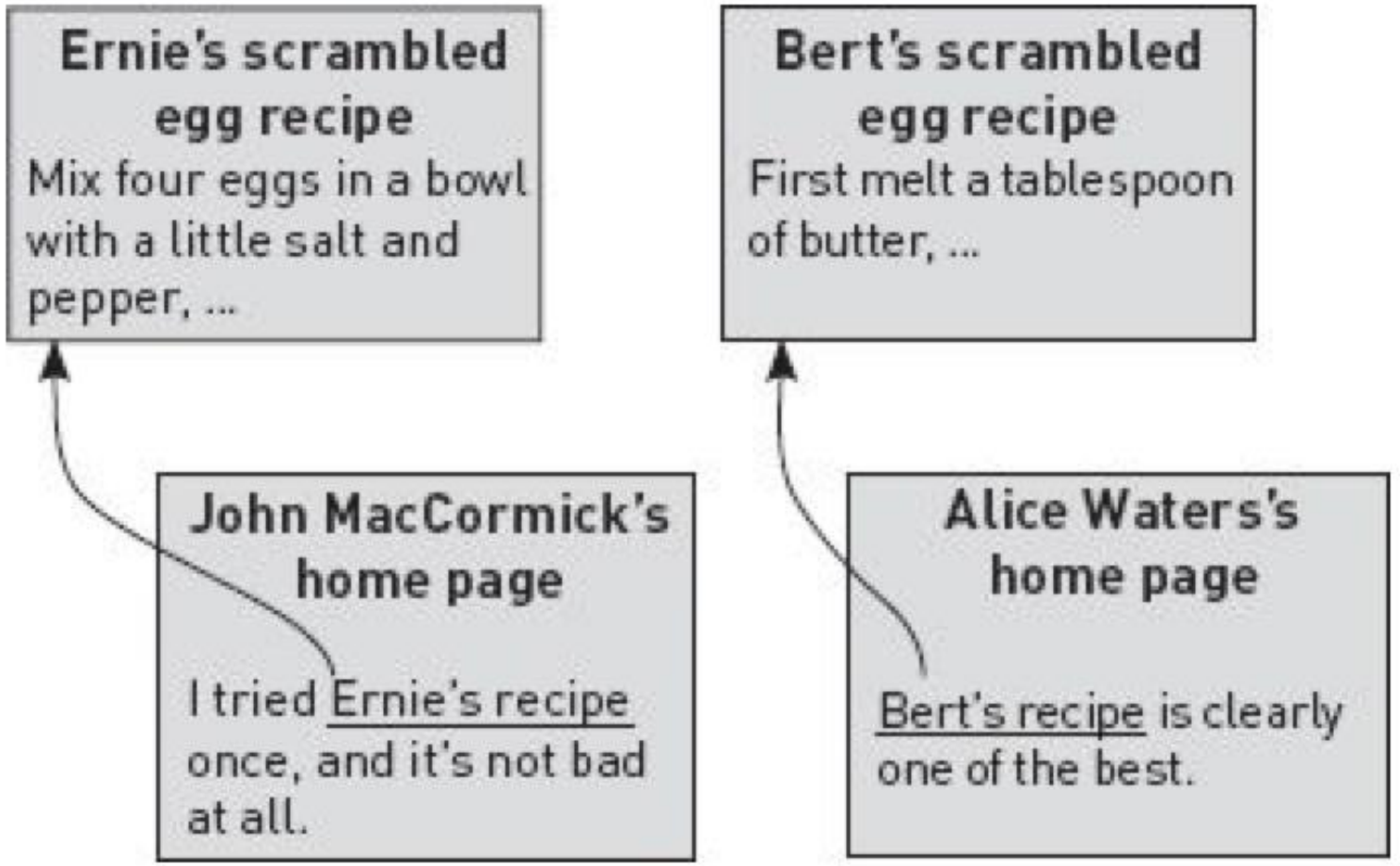


Image Source: *9 Algorithms that Changed the Future* by John MacCormick

Authority

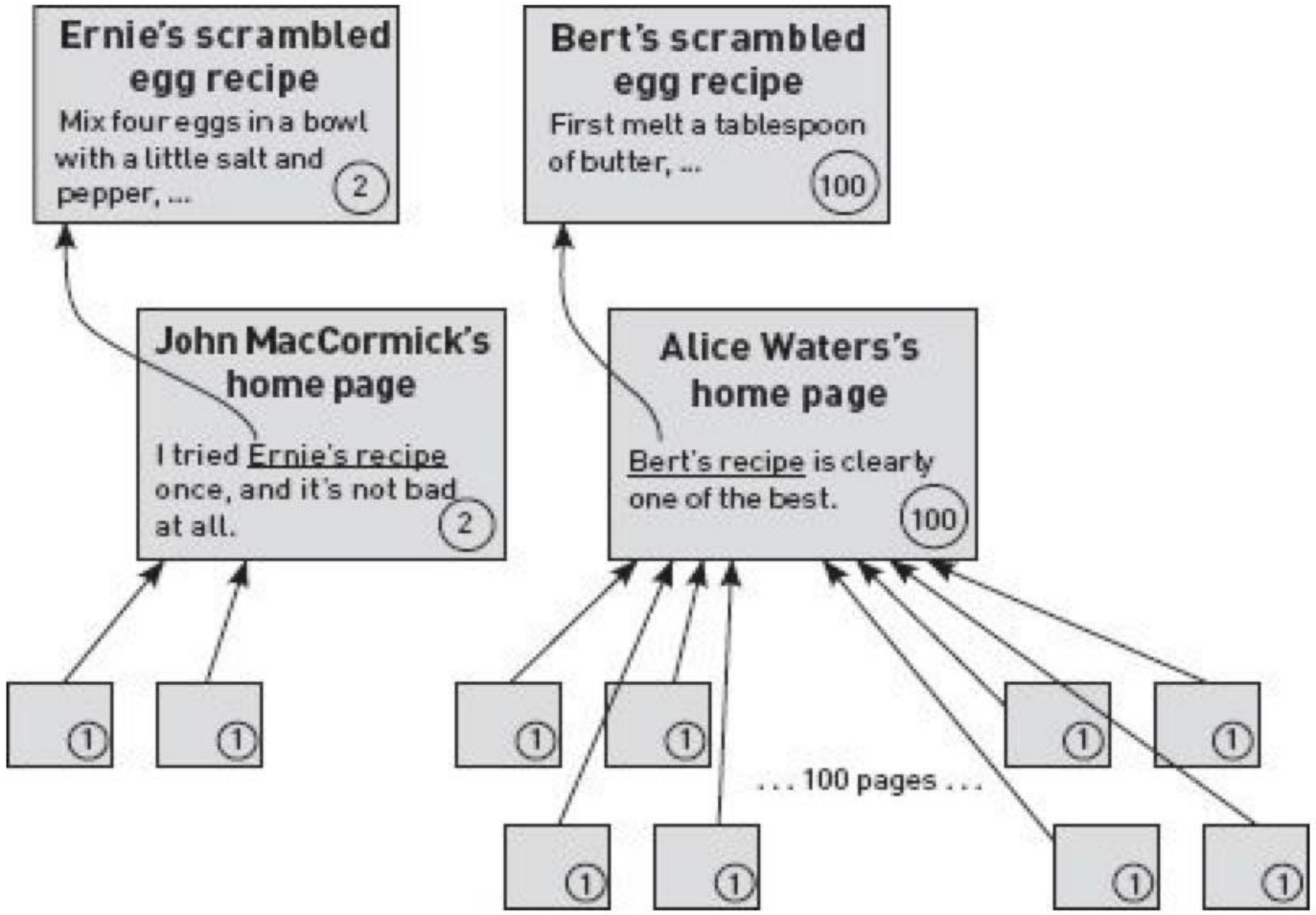


Image Source: *9 Algorithms that Changed the Future* by John MacCormick

Cycles

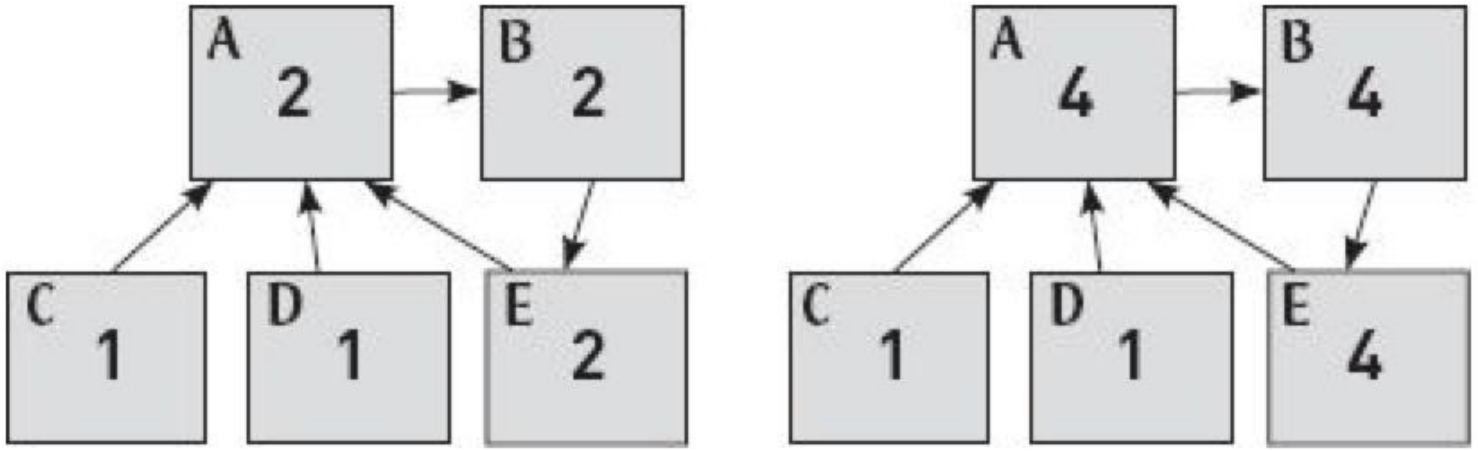


Image Source: *9 Algorithms that Changed the Future* by John MacCormick

Random Surfer

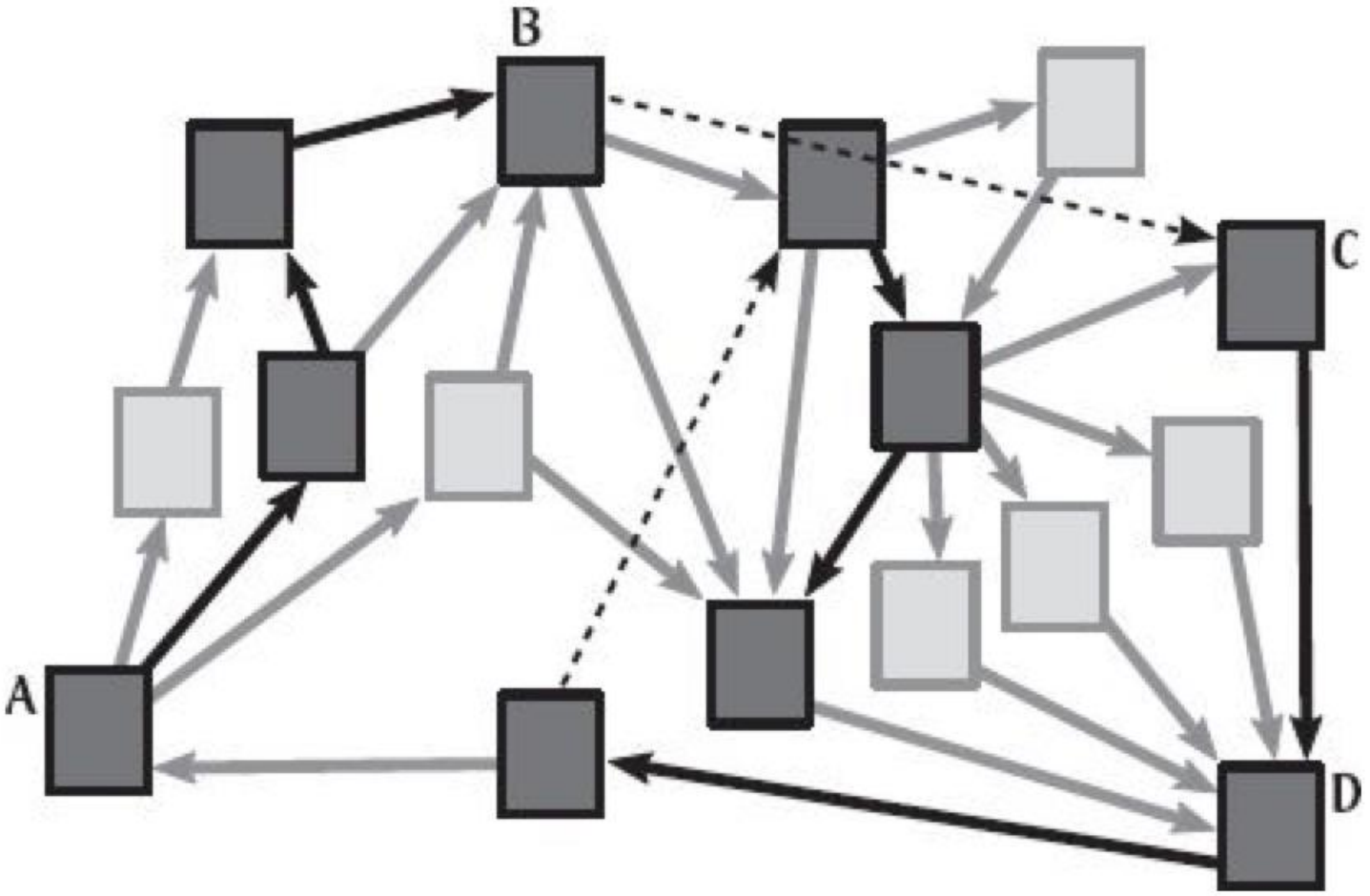


Image Source: *9 Algorithms that Changed the Future* by John MacCormick

Page Rank Activity

https://www.random.org/dice

Page Rank Example

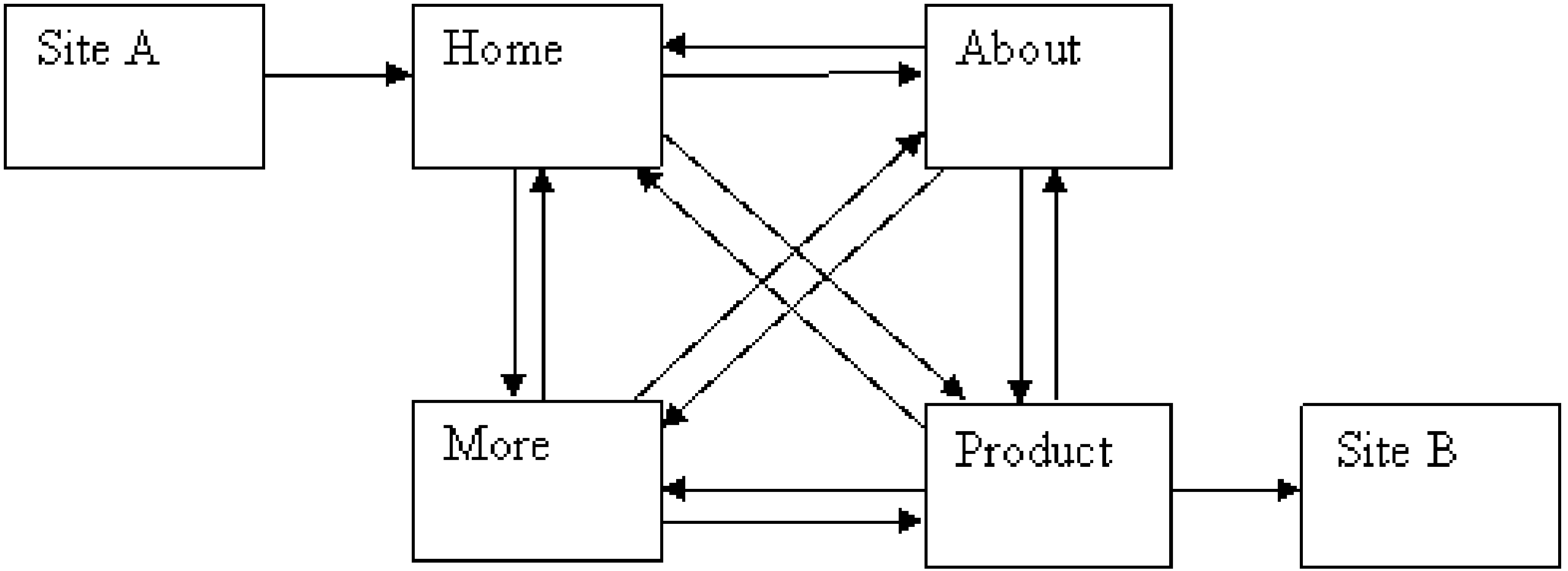


Image Source: Ian Rogers - Princeton

Page Rank Example

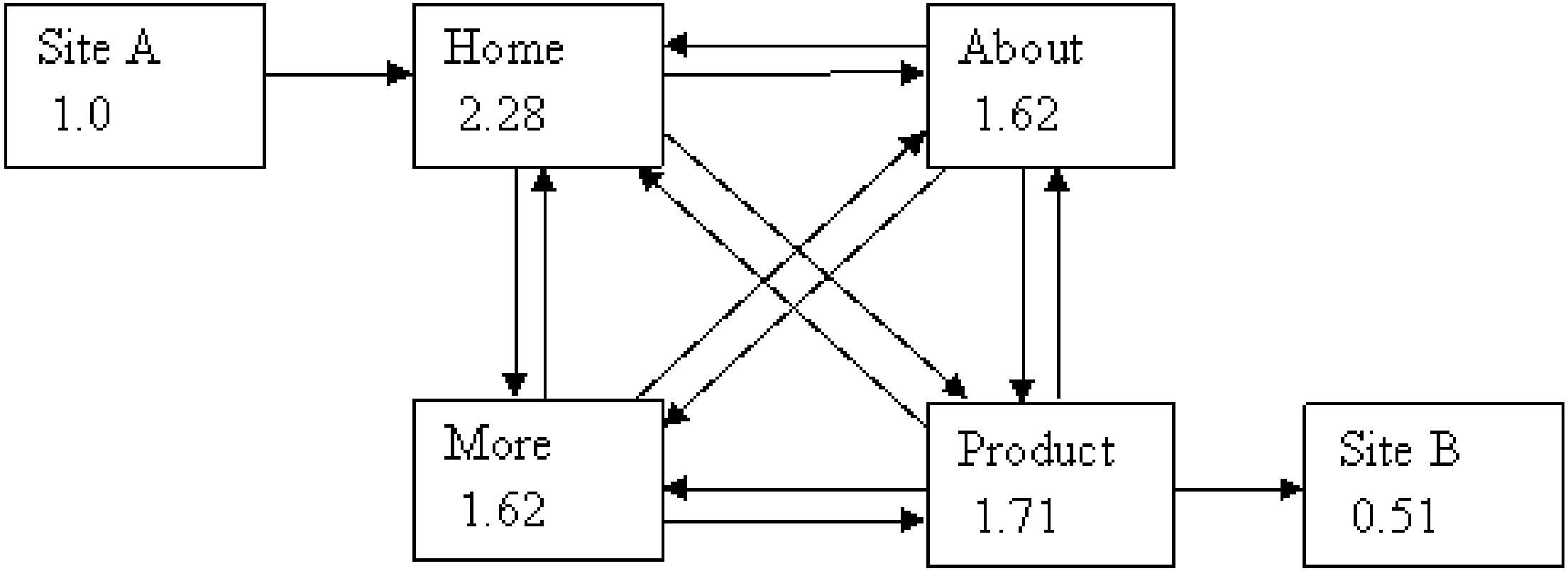


Image Source: Ian Rogers - Princeton

Simulate, Don't Calculate

- Find all paths with length < 5, calculate percentage of times each one appears (count appearances / total paths)

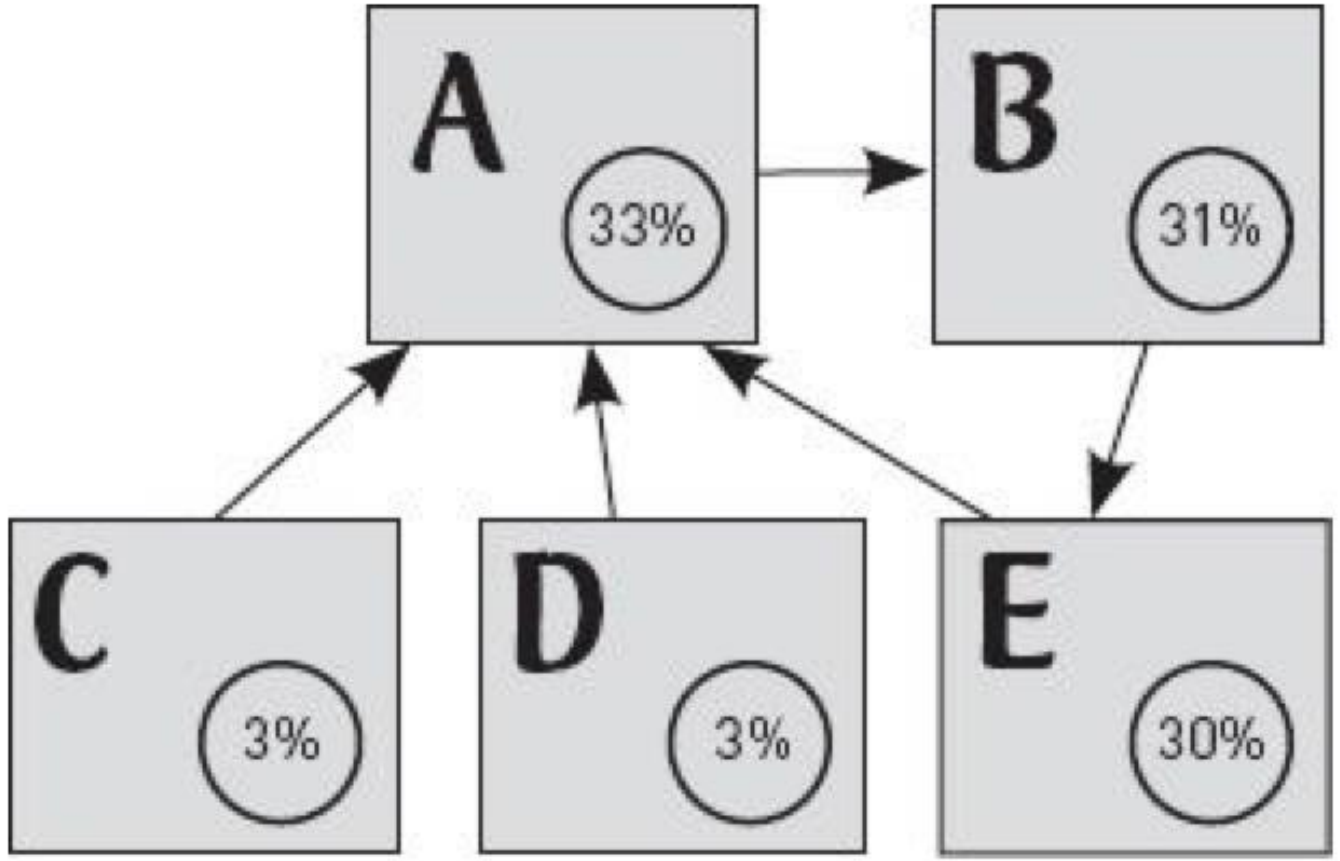


Image Source: *9 Algorithms that Changed the Future* by John MacCormick

Problems?

- Web Spam
- Computation Time
- Non-textual Data
- Structured Data
- Others?

What is Web 2.0?

Web 1.0	Web 2.0
Static web pages	Dynamic pages
Content from few	Content from many
Local software	Web software
Local storage	Web storage
Read only	Write / Collaborate
Text only	Multimedia
Individual thoughts	Collective thoughts
Propietary	Open / Shared

Source: New Tools Workshop