IA of AI

Convey UX 2017

Daedalus Information Systems

Agenda

Introduction

What is Artificial Intelligence (AI)

Where do information architecture (IA), user experience (UX) and content strategy (CS) fit in

How do we get involved

Why should we care

Daedalus Information Systems

2

INTRODUCTION

Daedalus
Information Systems
3

Me

Serial Career Changer

Transitioned to Technology 1995

Became Information Architect 1998

Microsoft 1999 – 2006

Discovered Information Retrieval 2003

Ascentium Digital Agency 2006 – 2010

Daedalus Information Systems 2010 – 2012

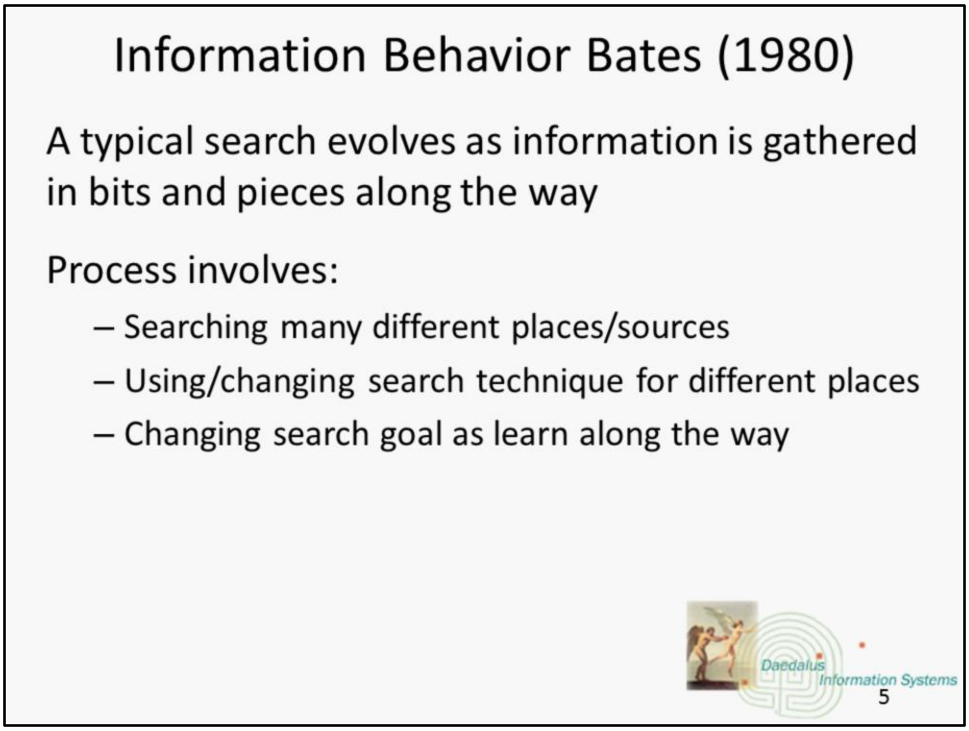
Portent Interactive January 2012 – 2 015

Brightedge 2016 -

Daedalus Information Systems

4

Me
Them
Shared Vision for workshop

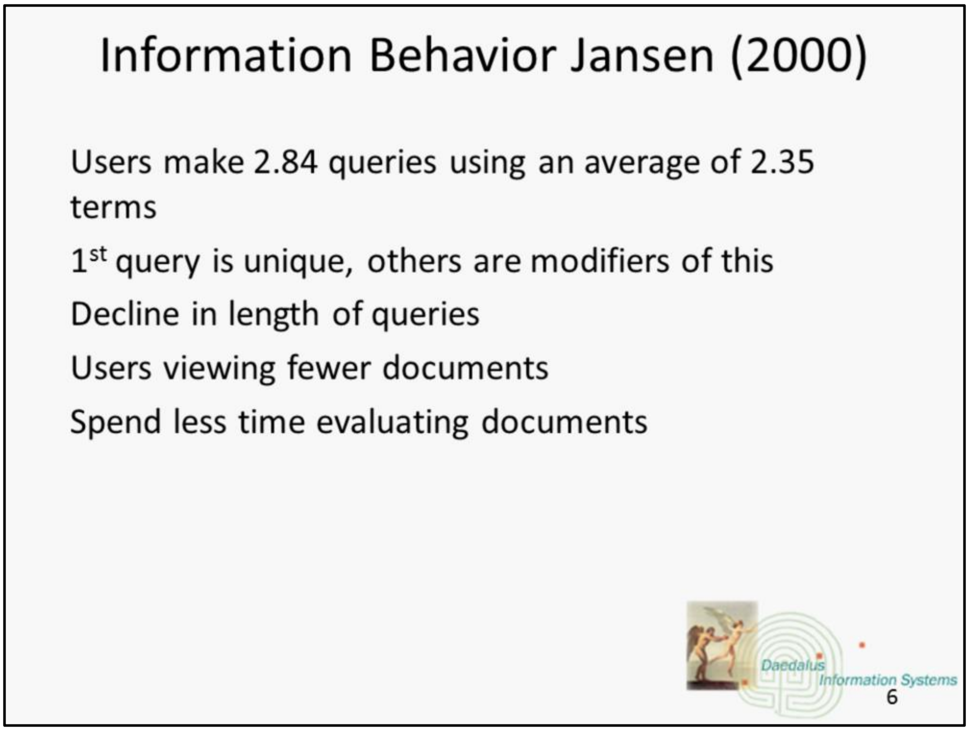


(1989)
Marcia Bates is an information scientist at UCLA. Her model tells us that the user starts out with a query and traverse to a point where they find something of interest. Their query is reformulated with this new information and continues in a changed state until the next piece of information affects a change which… well, you get the idea. We used to call this "browsing."

"An important thing we learned early on is that successful searching requires what I called "berrypicking." It is usually a fallacy to think that everything you want is going to be found in one place. Good searchers "berry pick," that is, they search for a topic in numerous places, just as you would go around in the woods picking blueberries or huckleberries. (I drew this example from actually searching for berries in the woods in the Northwest, when I lived in Seattle.)

Berrypicking involves 1) searching many different places/sources, 2) using different search techniques in different places, and 3) changing your search goal as you go along and learn things along the way. (See Bates, Marcia J. "The Design of Browsing and Berrypicking Techniques for the Online Search Interface." Online Review 13 (October 1989): 407-424.)

This may seem fairly obvious when stated this way, but, in fact, many searchers erroneously think they will find everything they want in just one place, and second, many information systems have been designed to permit only one kind of searching, and inhibit the searcher from using the more effective berrypicking technique. For example, in a technical report database, you might be able to search by word but not browse by the name of the laboratory that produced the report."
https://www.quora.com/profile/Marcia-J-Bates/Online-Search-and-Berrypicking/An-important-thing-we-learned-early-on-is-that-successful-searching-requires-what-I-called-berrypicking-It-is-usu-1



Possible reasons for shorter queries:

- Higher precision in search results
- Users more sophisticated about information needs
- Users forming better queries

We're in the web world albeit new:
Free text rules
Advanced search scares people

Begin the decay of discernment: ease of search, plentitude of results without effort, PageRank novelty

Information Behavior Mai (2016)

Quality of Information

Part of a spectrum

Data >> Information >> Knowledge

Quality depends on individual characteristics

- Contextual
- Situational
- Environmental
- Emotional

Daedalus Information Systems

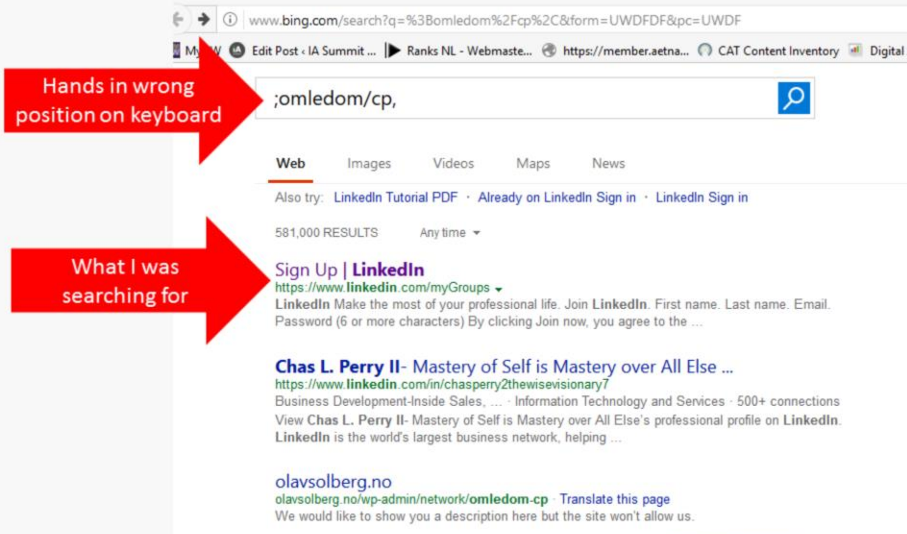
7

Jens Erik Mai, 2011

WHAT IS ARTIFICIAL INTELLIGENCE

Daedalus Information Systems

AI Manifestation 1



Daedalus Information Systems

9

An Algorithm is…

Set of instructions (procedure or function) that is used to accomplish a certain task

"A computable set of steps to achieve a desired result. " National Institute of Standards and Technology

"Math that computers use to decide stuff." Kevin Slavin

Daedalus Information Systems 10

algorithms = math that computers use to decide stuff
Algorithm is a recipe for performing a certain task. A data structure is a way of arranging values (array, tree, graph, list, etc.)

hash table: hash marks on twitter but computer numbers, quick way to store and retrieve values/items

arrays: chunk of contiguous memory a program can access b using indices >> one index per dimension >> arrangement of "boxes" where program stores values

Algorithms lock into behavior with no human supervision

Designed for a machine dialect: happens all the time on Google, adjust one thing - something else happens = unintended consequences

Algorithms

Algorithms process instructions according to formulas and rules that produce desired results if followed.

Daedalus Information Systems

11

Roger Martin, Rotman School of Management, University of Toronto
http://rogerlmartin.com/

3 Features of Any Algorithm

Correctness

Maintainability

Efficiency

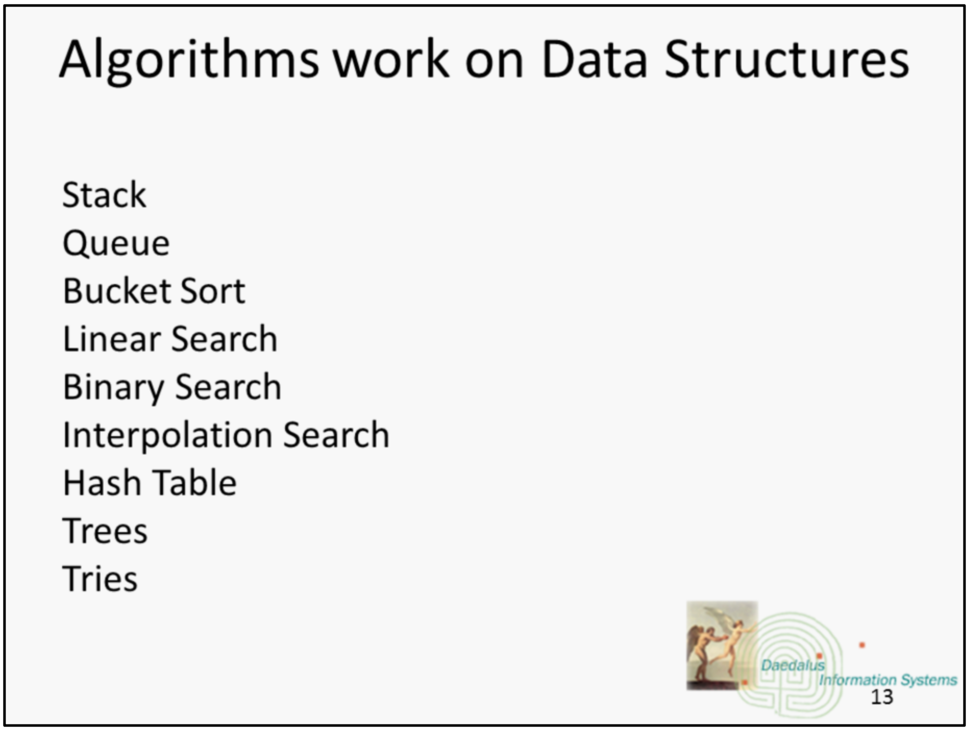
Daedalus Information Systems

12

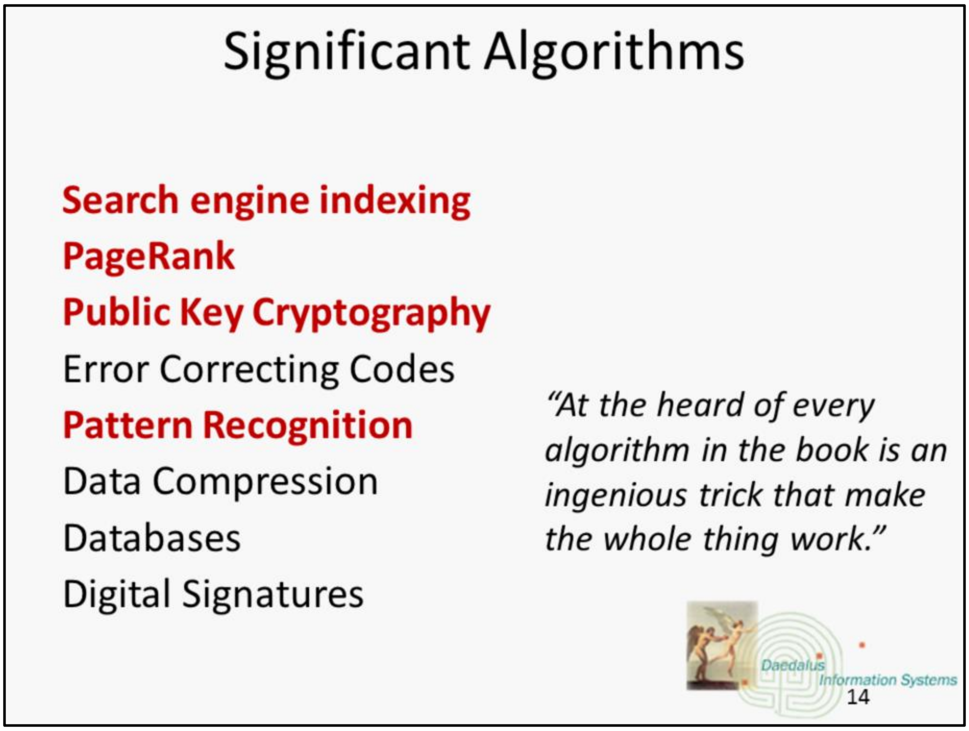
Correctness: Does it produce the right answer

Maintainability: can be kept operational

Efficiency: uses resources wisely, e.g. binary tree there are only 2 options



Stack: LIFO (last in first out) data structure with prescribed order for additions and deletions
Queue: FIFO (first in first out) data structure used mostly for additions
Bucket Sort: Algorithm divides items into bins/buckets then sorts buckets and concatenates the bucket content for sorted results
Linear Search: examines every item in array until it finds a match
Binary Search: algorithm calculates index hallway between minimum and maximum and divides, searches the array for the target, keeps dividing and searching until it finds the right value
Interpolation Search: uses the value of the target item to guess which array has the value
Hash Table: maps data to location in data table and assigns unique value – hashing function to map keys to the location, collision resolution policy for when keys collide (2 keys that hash the same value)
Trees: ordered (matters) and unordered (doesn't matter) = goal is traverse all nodes (leaves on the tree) in some order and perform the operation. Traversal types: pre order (node then children), in order (left child, node, right child), post order (left child, right child, node) depth first (all nodes left then all nodes right). Sorted trees: nodes arranged so that an in order traversal process them in sorted order.
Tries: tree that holds strings. Each internal node represents a single letter – leaf nodes can represent more than one letter , B-Trees are used to store large records, organized by key value – position in the tree defines the key value (not the reverse) – children of node share string prefix of parent leaf



From: 9 Algorithms that Changed the Future – John MacCormick

What makes a great algorithm?

- Used by computers every day
- Address concrete, real-world problems
- Relate primarily to the THEORY of computer science [efficiency with which problems can be solved with a model of computation supported through algorithms]

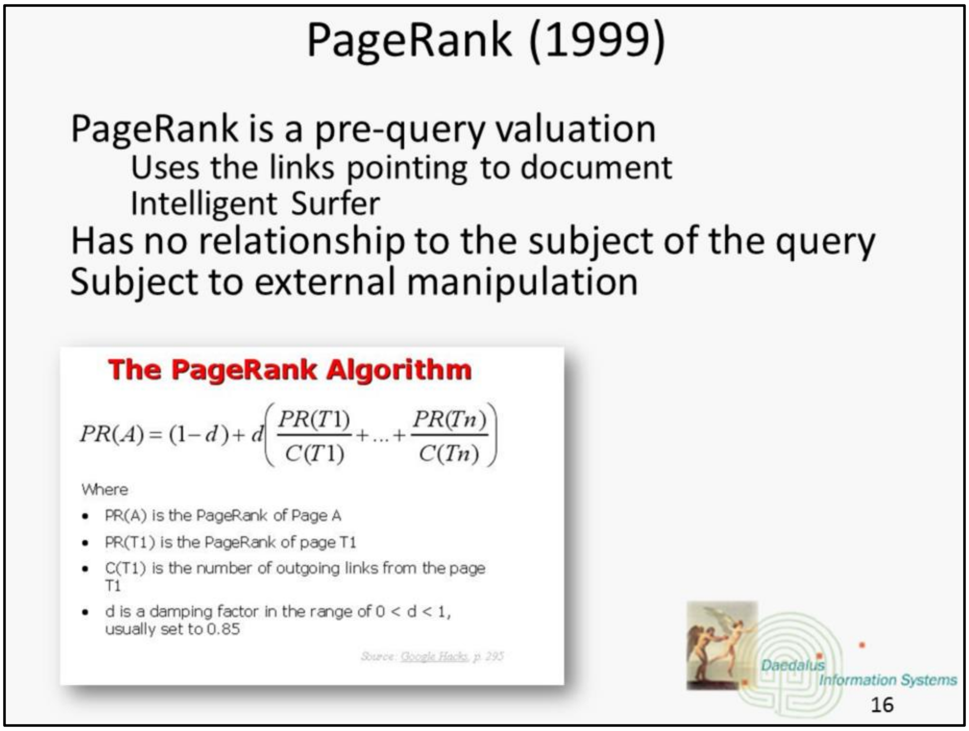
Search Engine Indexing

Internal & External Algorithms

1. Crawl
2. Convert
3. Label
4. Store
5. Retrieve

Daedalus Information Systems

15

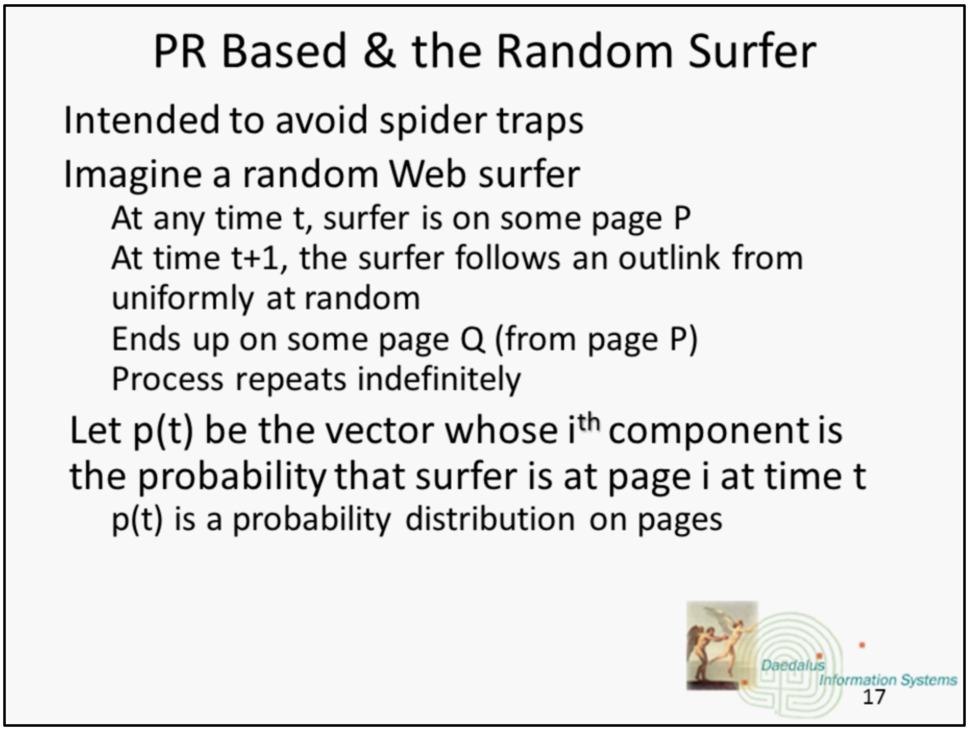


Based on academic citation model
1998 named one of the top 100 Websites by PC Magazine “uncanny knack for returning extremely relevant results”
Ranking based on number of links to the page
Used link structure of Web to build a Markov Chain with a primitive transition probability

Intelligent Surfer: algorithm picks a link to exit page – one that is topically related to the query
First introduction of “loose authority” determined by adding up the “authority” scores of the pages linking in
Discounted pages linking to each other (black hat link ring)

Complications:

- Assumes link vote of authority, does not consider commercial value of links
- Ability to link limited to subset of users
- Orphan pages
- Users no longer “surf” randomly
- Does not scale

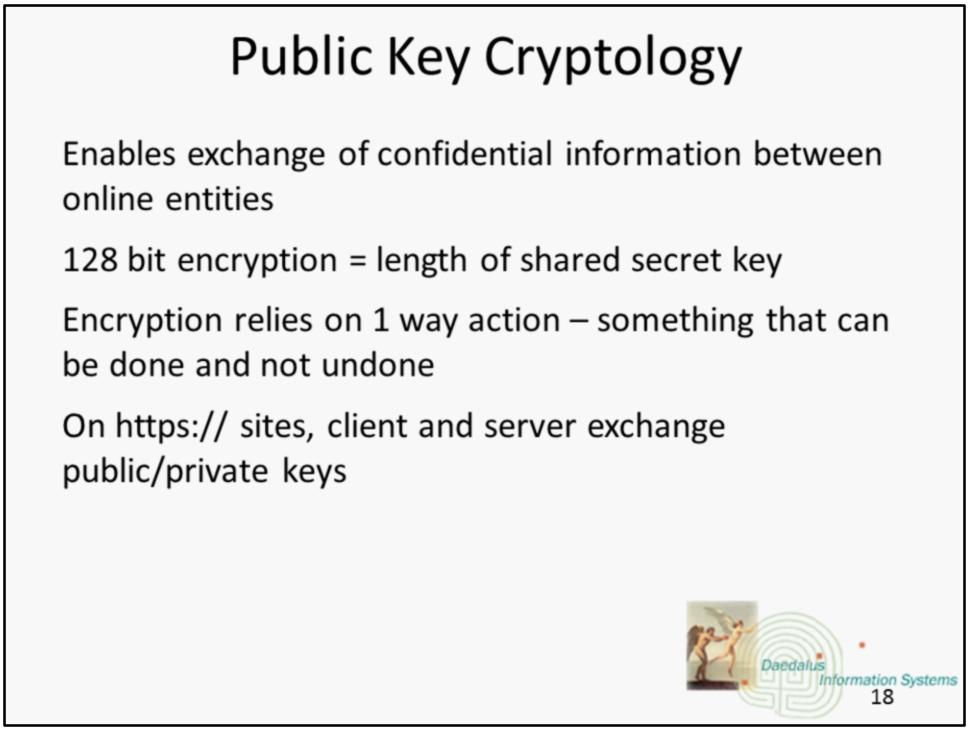


Slide curtsey of Dr Jun Wang, University College London, BCS Search Solutions IR Tutorial November 2011

Random Surfer (spider follows "randomly selected links) examines all of the links and follows one to destination, does that at destination

- Random Surfer authority score: % of time random surfer would spend visiting the page (added to the hyperlink score)
- Restart probability = 15%, surfer does not select a link and instead "jumps" to another page

Since updated to the "reasonable surfer" to reflect choice that is semantically related to original request



Uses sequence of algorithms to generate number sequences that are hard to predict

2 large prime numbers for security – prime # is one that is greater than 0, only factors are 1 and itself, composite number is counting number greater than 0 that is not a prime number

Public # (known to both)
Private # (known to each)
Encrypted # is the combination of all of both pubic and private = shared secret

Diffie Hellman Key algorithm 1976
RSA famous public key encryption system

Discrete exponentiation – combining of encryption keys
Discrete Logarithm – deconstruction of encrypted message

Pattern Recognition

Ability for computer to act intelligently based on input data with a lot of variability

- Decision Trees
- Nearest neighbor classification
- Neural Networks

Classification

Ideal replaced by practical

Constant decision what problem to work on

- Value based

Pandemonium

Daedalus Information Systems

19

Decision trees: run through series of questions where answer determines outcome
Nearest neighbor: find in training data and use mot similar to predict the unsorted data
Neural networks: based on biochemistry, electric and chemical signals

- some connections dedicated to send, others to receive
- neurons are either idle or firing
- stretch of incoming signals determines the neuron firing
- 2 types of inputs: excitatory (adds up to total) and inhibitory (subtracted from total)
- each neuron assigned a threshold
- signal here is data related to a pre-assigned condition

Explicit teaching based on user data
Learning from example based extracted characteristics from training set of documents

3 Information Retrieval Models

Boolean Model: framework of document sets and standard set of operations on sets

Vector Model: framework composed of t-dimensional vector space representing queries and documents as vectors

Probabilistic Model: framework based on probability distributions of terms on documents and queries (Bayes theorem)

Daedalus Information Systems

20

These are for unstructured text

Boolean Model

Framework: considers query terms as either present in or absent from collection

Term frequencies in document seen as binary

Query is a set of terms linked by operators: AND, OR, NOT

Predicts every document is either relevant or not relevant

Weights assigned to terms to compute rank

Term-term correlation to establish relationship between terms

Document diversity addressed by:

- TF/IDF weight
- Document length normalization

Daedalus Information Systems

21

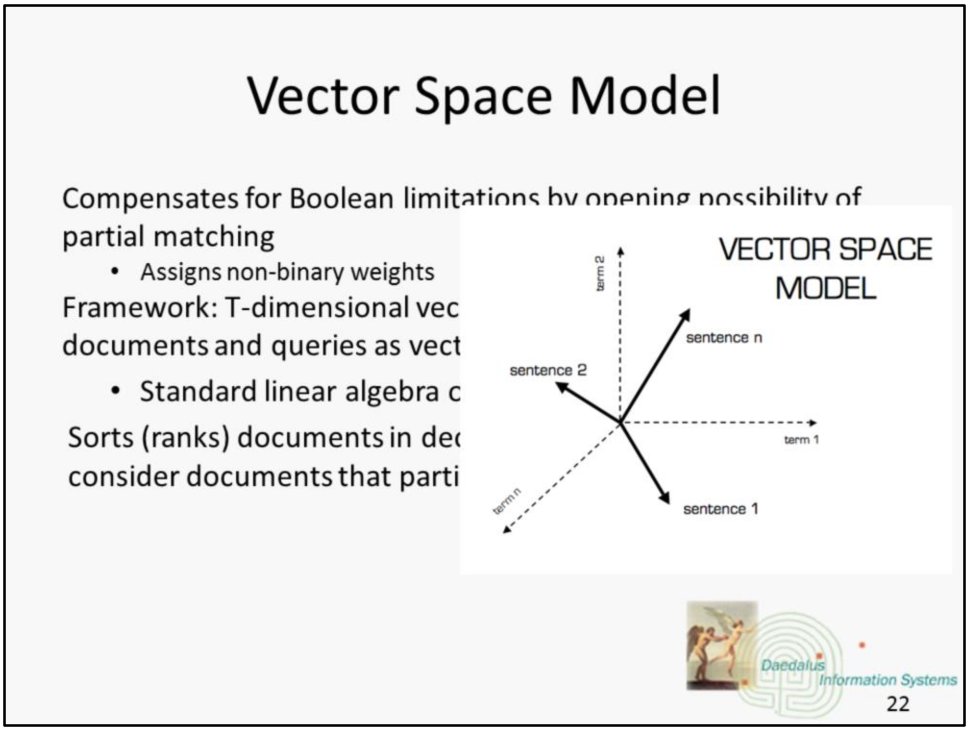
Assigning weights to terms allows computing a numeric rank to each document with regard to query
Term term correlation establishes a relationship between any two terms based on their join co-occurrence inside documents of the collection (phrase searching)

Term frequency weights: value or weight of term simply proportional to frequency in document
Exhaustivity: property of document descriptions – exhaustivity of document description quantified as number of index terms contained
Specificity: property of index terms – specificity of a term can be quantified as inverse function of the number of documents in which it occurs

Zipf’s Law: power law equation

TF/IDF: document normalization divides document rank by length so that longer docs are not at an advantage strictly by size (makes frequency of term more meaningful)

Normalized by size in bytes, number of words or vector norms
very important ranking principle

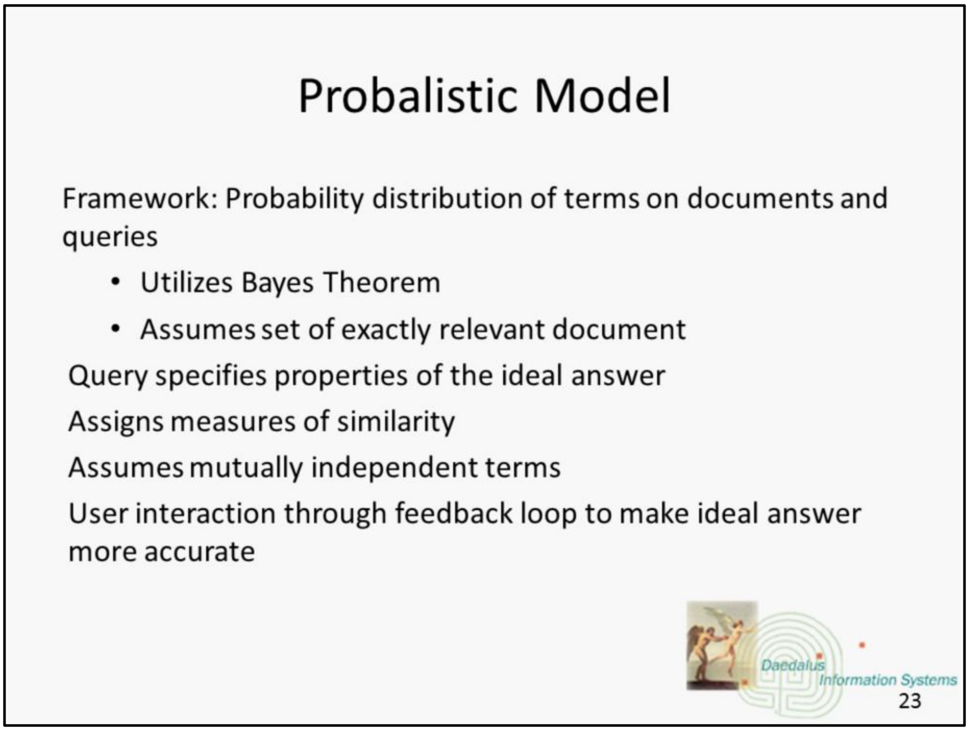


Advantages

- Term weighting scheme
- Partial match strategy
- Rank by degree of similarity
- Natural document length normalization

Disadvantages

Index terms presumed to be semantically independent



Query is a subset of index terms, document is represented by a vector of binary weight indicating the presence or absence of index terms

Advantages
Optimality as documents are ranked in decreasing order of probability

Disadvantages
- Ideal often not realized because user relevance often relies on non-document factors
- Need to guess the initial similarity of documents
- Term frequency within document not considered
- No document length normalization

ORIGINS OF AI

Daedalus Information Systems

Two Schools of AI

Symbol Processing

Neural nets

Symbolic AI

Intelligence = symbol manipulation

Fixed and formal rules

Assume: all intelligent processes are forms of information processing

Computer processes symbolic representations (1s/0s) according to formal rules (program)

Plato's rationalism

GOFAI

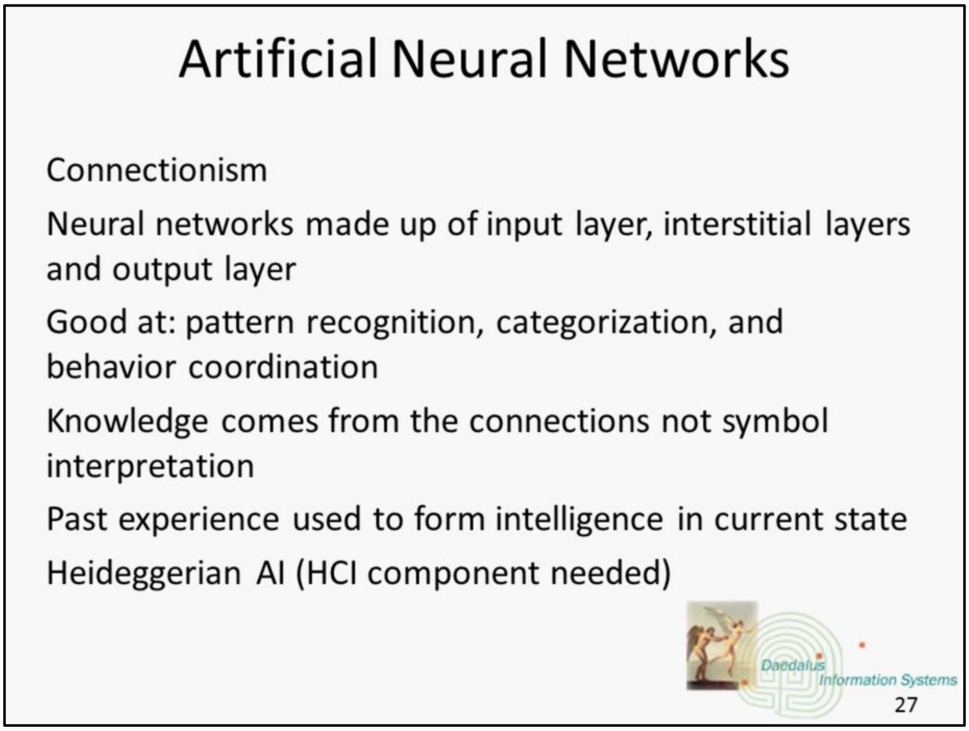
Daedalus Information Systems

26

Intelligent processes = perceiving, reasoning, calculating, language use

Language is symbolic: eg a dog does not look like the word that represents it
3 characteristics of Plato's rationalism: Psychological assumption that human intelligence is symbol-manipulation according to formal rules, Epistemological assumption that knowledge is formalized and can be expressed in a context-independent, formal rules or definitions, Ontological that reality has a formalized structure built on objective, determinant elements each of which exists independent of the other .
Dreyfus added the Biological assumption, rules and symbols implemented by the human brain in the same way as by a machine

GOFAI = good old fashioned AI – meat and potatoes AI – train the computer without the need for understanding



Re-emerged in 1980's
Layers of data – decisions inform up the line (backpropagation)
Autonomy: without human supervision
Automate: replace human effort

Intelligent processing modeled on structure and operation of human brain instead of digital computer – neurons and synapses, receptors and reactors
Neurons as processors with input/output functions
Intelligence is a product of the neuron connections

The ANNs of the 1980s could never conceive of the vast amount of personal and behavioral data used in today's neural networks (deep mind, Watson). Examples: IoT (intelligent machines), Watson (expert systems)

Cannot generalize as humans do, cannot perform functions that require "common sense" (must be programmed)
Heideggerian AI: intelligence is situated in the world and does not require rules.
Terry Winograd (Stanford): design of computers must include consideration that computers must function in a human world and communicate with human users and not impose their own rationalistic logic on surroundings.

Turing (1948)

Imitation Game

Objections to machine intelligence

- Theological
- Head in Sand
- Mathematical/philosophical
- Consciousness
- Lady Lovelace

Reward Signal increases probability of repetition events leading to it

Suitable imperative: one that regulates the order in which the rules of the logical system are applied

Program the learning

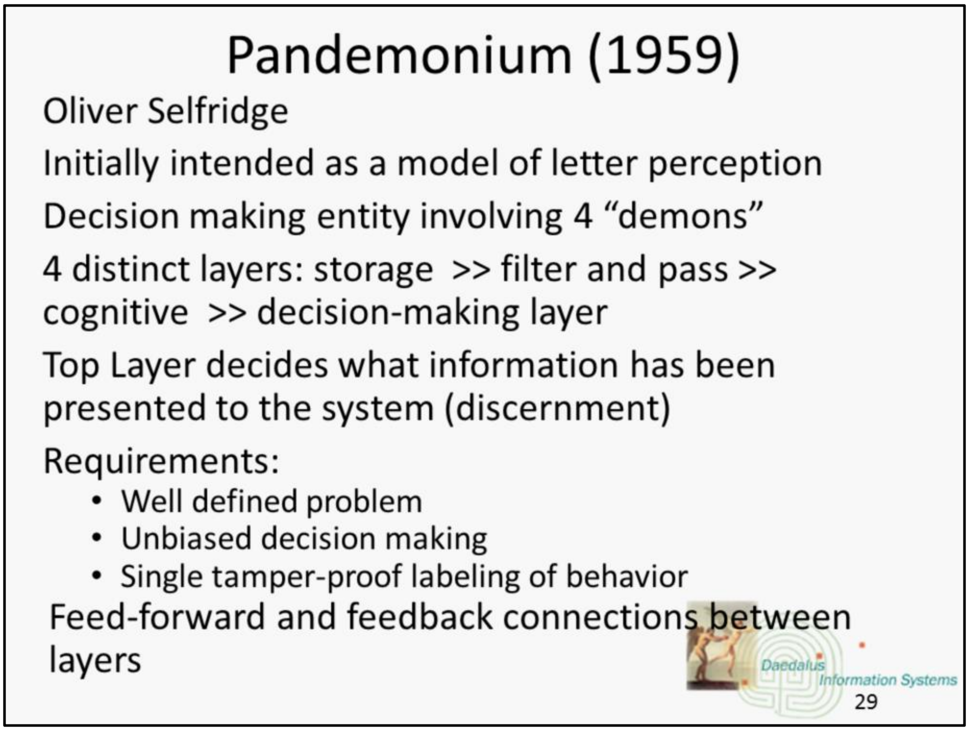
Daedalus Information Systems

28

Computing Machinery and Intelligence: A.M. Turing; Mind: Quarterly Review of Psychology and Philosophy ; October 1950

Imitation Game: 3 individuals, man, woman and interrogator (apart from the other two). Object of game is for interrogator to decide which of the other two is a man or a woman based on answers. Substitute machine for one of the individuals, can interrogator determine machine from person based on answers
AI for Turning is a digital machine able to play the Imitation Game

Some objections addressed by Turning;
Theological: machines have no soul and neither do high-level animals
Head in Sand Objection: too scary to accept
Mathematical Objection: Godel’s theorem on limitations of discrete state machines (there are questions that cannot be answered, certain things a machine will never be able to do) - all humans are also incapable of answering some questions
Consciousness Argument: dismissed as solipsistic (solipsism holds that knowledge of anything outside one's own mind is unsure)
Lady Lovelace Objection: machines cannot originate anything – humans take risks, machines cannot



Wikipedia calls him the “Father of Machine Perception” – related to Selfridges stores in England
Supervisor of Minsky at MIT
Associated with neural networks and pattern recognition
Intellectual processes can be carried out by hierarchy of simultaneously functioning submachines (demons)

Cognitive demons and data demons use abstracted data for evidence of specific propositions. Demons record events and recognize patterns
Demons = entities that perform intellection processes that can be carried out by a hierarchy of simultaneously functioning submachines.
Demons receive input through continuous interaction that leads to conclusion on what the input “is”
Layers group demons by specific task

- Bottom: collect and stores information
- 3rd Layer: selects and passes along good candidates
- 2nd layer: filters and weighs >> “shrieks” information up to top layer where decision is made
- Top layer: biased towards loudest shrieks and draws own conclusion regarding the output

Used human behavioral psychology for models
Hierarchical letter identification: visual features are mapped onto abstract letter identities through a series of increasing invariant representations
Layers of feature and letter detectors – building blocks of inputs and outputs
Template matching model

Unbiased decision making: encourage random behavior for a wide-range of decision-making behavior

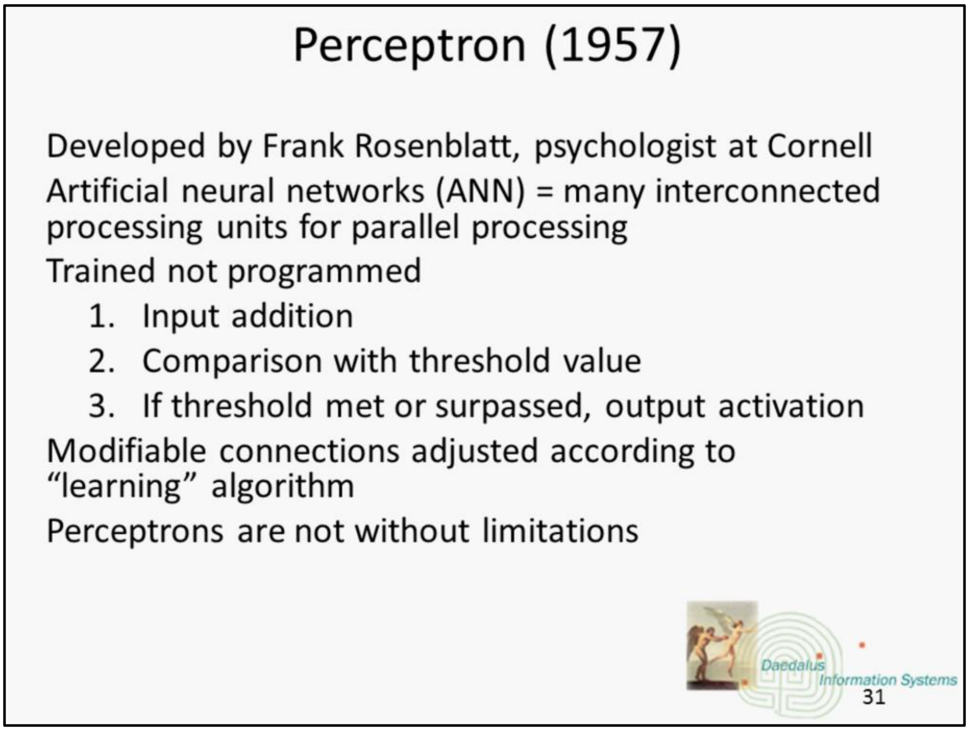
“The visual world can be described in terms of variations in special frequency, that changes in luminance across space…The key to both approaches [measures of identification thresholds and measures of variations in identification thresholds a function of the characteristics] involves comparison of human performance with that of an ideal observer.” Perception: From Pixels to Pandemonium: Jonathan Granger, Arnaud Rey, Stephane Dafau August 2008

Pandemonium (1959)

Developed by Oliver Selfridge and Frank Rosenblatt

Selfridge Rosenblatt character (letter) recognition based on component features

Daedalus Information Systems



Signal transmission network consisting of sensory units, association units and response units
Types: single layer, multilayer, cross coupled, multilayer back couples
Backpropagation algorithm (Rumelhard 1986) adaptively corrects weights based on adjusted pattern classification from training set.
Birth of Symbolic AI
Single layer: one way, no hidden components limited pattern recognition capacity
Multilayer: feed forward with multiple hidden elements
Cross coupled; connections made to join units of the same type
Multilayered back coupled: feedback paths from elements further back in the process (near the output) – possesses a universal approximation property

Error correction learning algorithm

- Binary outputs
- Feed forward

Limitations

- Theoretical learning curves for error correction procedure
- Determination probability that solution exists
- Representation of complex environments
- Efficient reinforcement
- Recognition of abstract concepts

Debunked by Marvin Minsky and Samuel Papert in 1960's – they favored the Symbolic AI model developed in 1956)
Symbolic AI: capabilities of computer to manipulate symbolic representations in a ways sensitive to logico-syntactical (discrete) structure
Manipulated and transformed according to rules and strategies (logical, programmed and rationalistic
Neural Networks seen as self-organizing based on training
Cannot process anomalous problems (those resistant to an acceptable solution)

Pandemonium Layers

Bottom layer: store data

3^{rd} Layer: select, weigh, filter and pass along data

2^{nd} Layer: “cognitive demons” decide which information from 3^{rd} layer process

Decision layer: single decision demon on what information is presented to the system for processing

Daedalus Information Systems

32

Identified behaviors for Pandemonium

- Ability to define problem
- Unbiased decision making
- Single tamper=proof way of labeling “demon” behavior

Marvin Minsky (1960)

Cognitive computer scientist

Co-founder MIT AI Laboratory

Symbolic AI

With Seymour Papert brought forth a 20 year "AI winter" with criticism of early AI Artificial Neural Network (ANN) approach

Daedalus Information Systems

33

1927 – 2016
A humanist computer scientist

Symbolic AI = rule based system of symbol processing and manipulation based on selected training. Uses a representational structure, applies rules - trained
Neural Net AI = interconnected processing units producing cycles of input feeding and weight adjustment. 3 part processing operation: input addition, comparison with threshold value, output firing

ELIZA (1963)

Human mediated

Early computer/human conversation (NLP)

Heuristic programming

- Keyword identified (input)
- Sentence transformed according to rule associated with identified keyword
- Choose appropriate transformation – if none available, choose most likely/earlier transformation
- Generate responses.

Keyword dictionary contains composition, assembly and decomposition rules

Daedalus Information Systems

34

Recognition of semantic patterns in text
NLP = natural language processing

Name chosen because of incremental learning by “teacher” example – e.g. Eliza Doolittle in My Fair Lady

Input sentences analyzed on basis of composition rules triggered by keywords that appear in input text. Responses generated by assembly rules associated with decomposition rules (a data structure that searches text for specific patterns and then decomposes text into disjointed constituents) and reassembly rules (specification for construction of new text by means of recombination of old and possible (new) constituents).
No understanding

There are elements in human conversation that do not take place in text, available only to humans in F2F meeting

Eliza Global Context

Global Context is key to understanding

Sub-contexts emerge as conversation continues for consequential richness

Individual participants bring Belief Structure

ELIZA scripts (previous learnings) establish a global context for future “understandings”

Broad context framework only

Daedalus Information Systems

35

Global context assigns meaning to what is being said in most general way.
Belief Structure: emerges from individual’s intellectual life (highly logically organized) and life experiences

Broad contextual understanding due to general nature of scripts development

Dreyfus (1964)

Rationalist assumption of “ordered reality” is flawed

Knowledgeable reality itself lacks rational structure

Inter-relatedness between humans and the world

Human world filled with experience structures- neither subjective or objective

Intelligence is discovering meaningful structures and applying meaningful behaviors independent of fixed rules

Daedalus Information Systems

36

Hubert Dreyfus (AI critic) computer scientist at MIT teaching philosophical theories of knowledge and perception
Recruited by RAND Corp in 1960’s as a philosophical consultation to their AI program
Published highly critical article on AI “Alchemy and Artificial Intelligence” and “What Computers Can’t Do” 1972, republished in 1992 and “Mind over Machine” 1986

Principle focus is unique and intuitive way humans experience the world and develop manners of getting around in it

Intelligence is situated – co-determined by situation/environment
Experience structures = smells, feelings, threats, obstacles, goals

Humans experience the world as a single whole before its individual components
Sensorimotor intelligence: human skill used in perceiving, recognizing, moving and manipulating objects as well as coordinating and integrating perception and movement (localized complex feedback system of nervous system, senses, glands and muscles


Perceptrons & AI Winter
WINTER IS COMING
Information Systems

Norvig & Russell (2004)

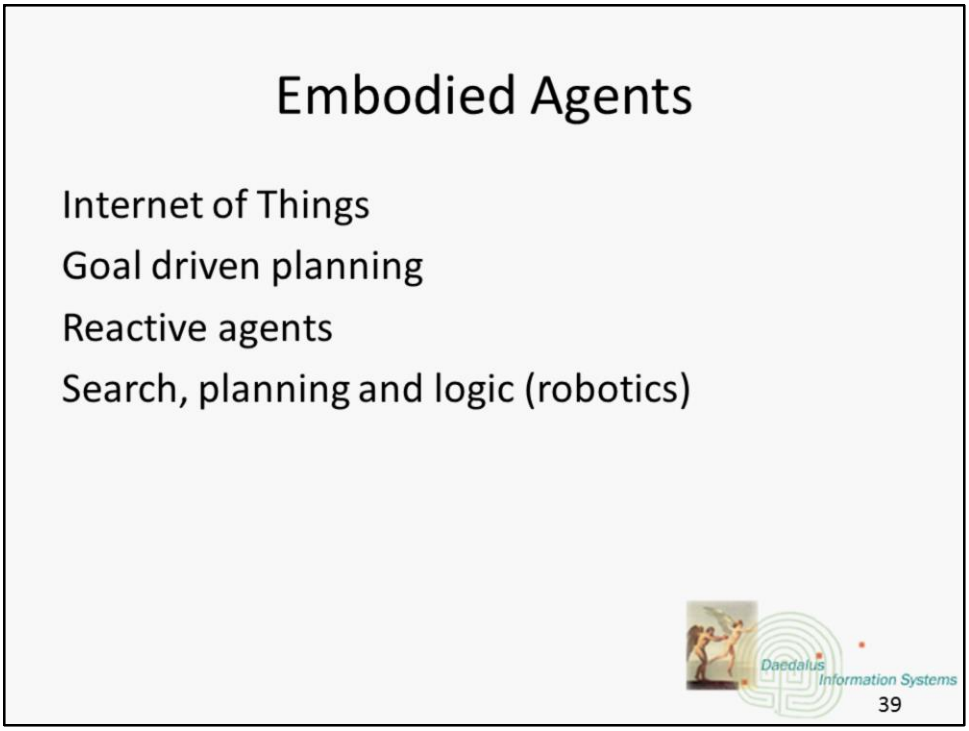
Artificial Intelligence: A Modern approach

Types

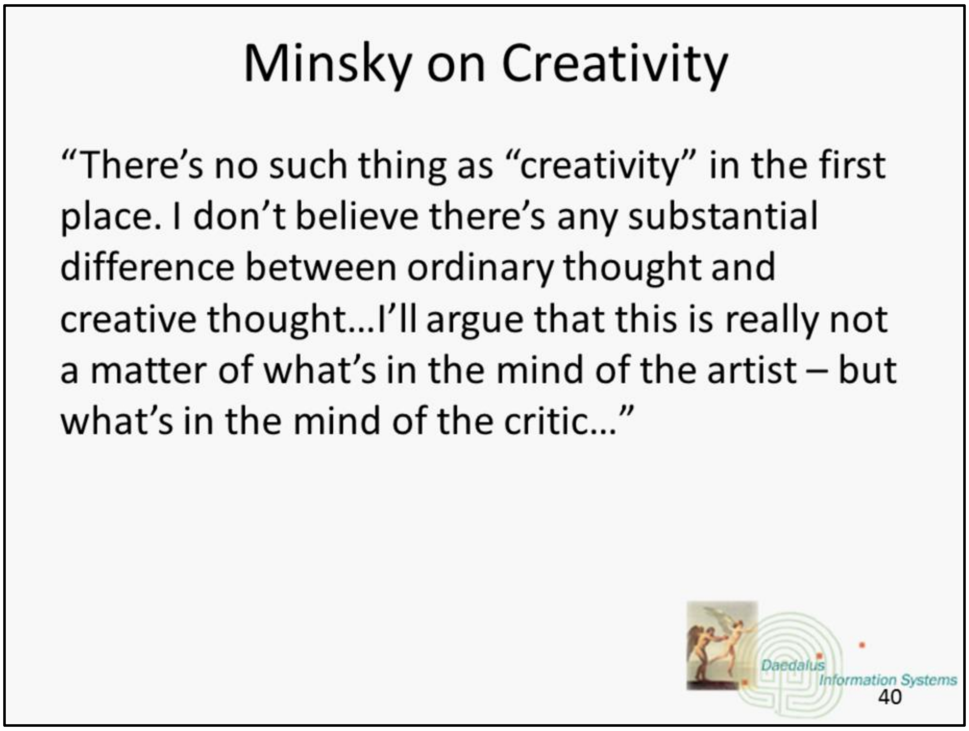
- Systems that think like humans (neural)
- Systems that act like humans (Turing)
- System that think rationally (logic solvers)
- Systems that act rationally (perception, NLP, Planning, Navigation)

Daedalus Information Systems

38


Embodied Agents
Internet of Things
Goal driven planning
Reactive agents
Search, planning and logic (robotics)
Daedalus
Information Systems
39

Roomba
GPS



Minsky: Why People Think Computers Cannot

He goes on to elaborate that some humans are better at learning than others and this accounts for creativity
Creativity could be just the consequence of childhood accidents in which a person’s learning is more self-applied than others.
If machines are made to learn better they can be creative
Would require computers to be programmed for abstract thinking in addition to logical reasoning

Minsky on What AI Best Suited To

Search

Learning Systems

Pattern Recognition

Planning

Induction

Daedalus Information Systems

41

Minsky
Search: search engines
Learning Systems:
Pattern Recognition: fraud detection
Planning: GPS
Induction: IBM Watson

Search

Requires additional structure

Near to/close to

Solve for one, solve for many

Personalization

Daedalus Information Systems

42

AI and Search: Mrs. Dewey

Librarian with an attitude
Product of Microsoft Live Search
Processing results slow
No user influence opportunities
Novelty act?

Daedalus Information Systems

43

AI and Search: Wolfram

Developed Mathematica software

Powers iPhone 4 Siri

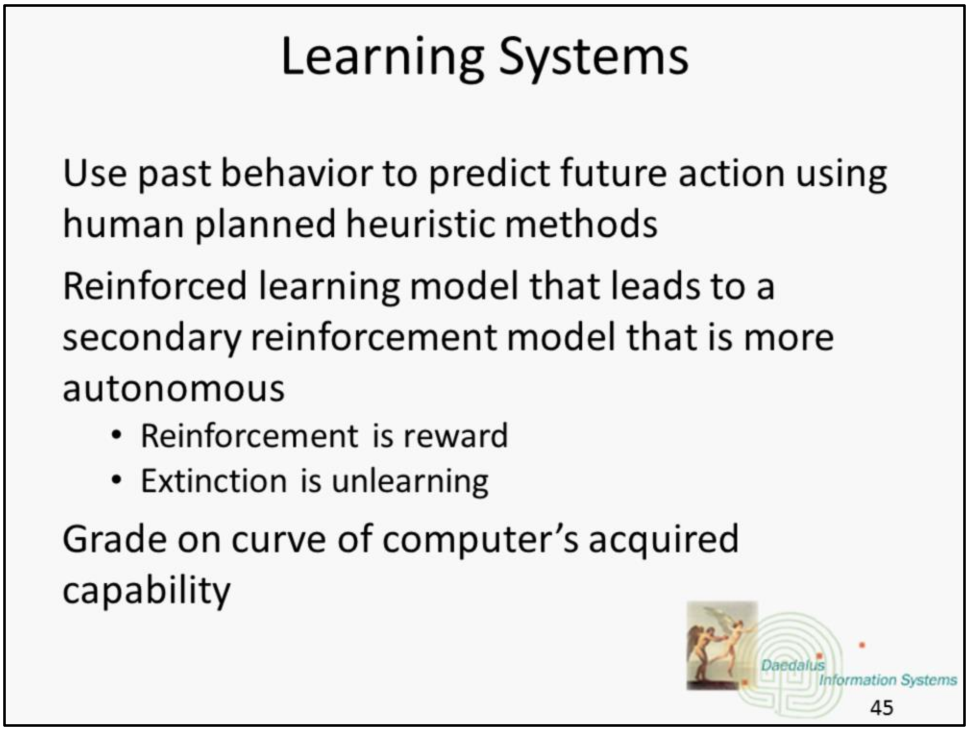
Searching own "knowledge base" (limited corpus)

Not for every type of search

WolframAlpha™
computational™
knowledge engine

44 Information Systems

Knowledge base is not an index – not crawled and refreshed with new data from outside resources



Generalized past experiences
Success is reinforced decision models
- Can have secondary reinforcement models (more autonomous)

Reward for partial goals (local reinforcements)
Grade on curve of computers acquired capacity
Reinforcement = reward
Unlearning = extinction

Pattern Recognition

Ability for computer to act intelligently based on input data with a lot of variability

- Decision Trees
- Nearest neighbor classification
- Neural Networks

Classification

Ideal replaced by practical

Constant decision what problem to work on

- Value based

Pandemonium

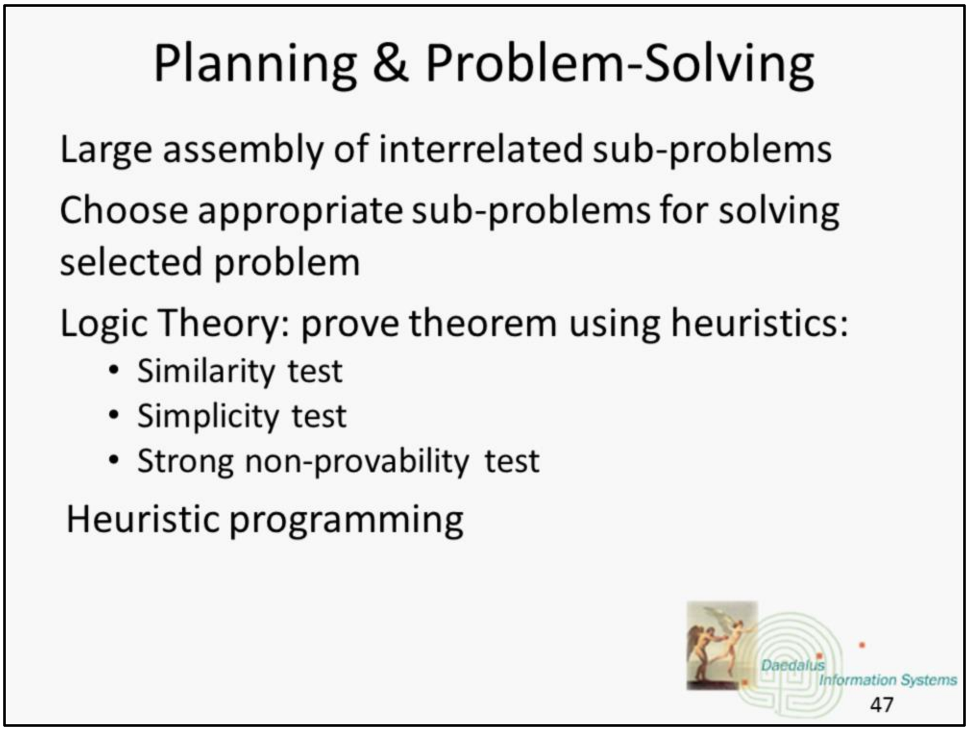
Daedalus Information Systems

46

Decision trees: run through series of questions where answer determines outcome
Nearest neighbor: find in training data and use mot similar to predict the unsorted data
Neural networks: based on biochemistry, electric and chemical signals

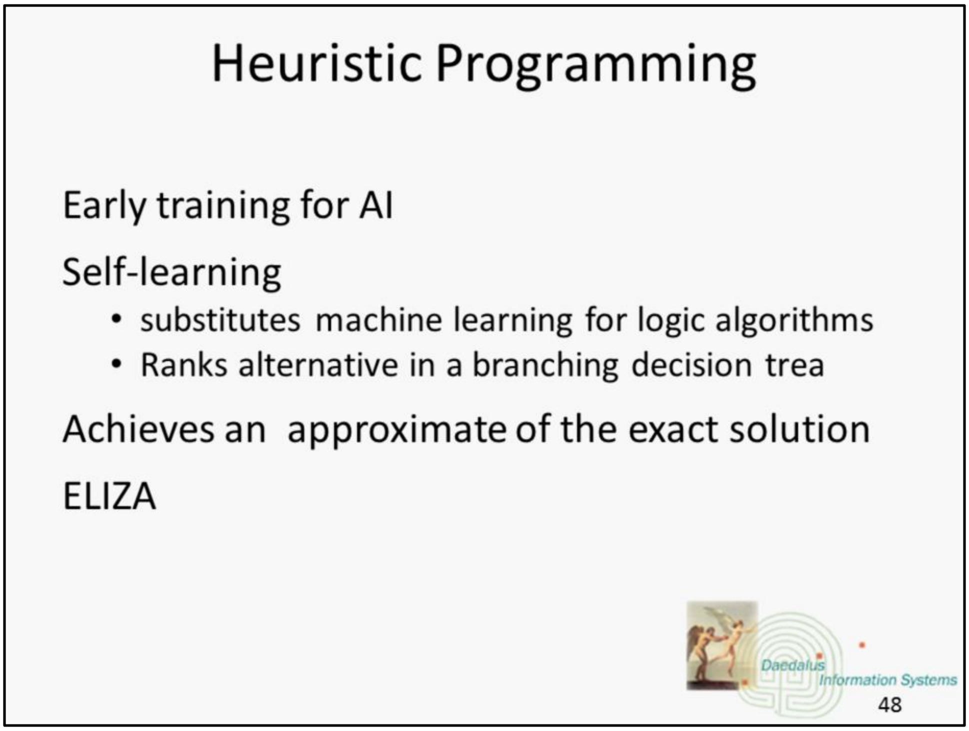
- some connections dedicated to send, others to receive
- neurons are either idle or firing
- stretch of incoming signals determines the neuron firing
- 2 types of inputs: excitatory (adds up to total) and inhibitory (subtracted from total)
- each neuron assigned a threshold
- signal here is data related to a pre-assigned condition

Explicit teaching based on user data
Learning from example based extracted characteristics from training set of documents



AKA Goal Seeking or Problem Solving
Intelligent systems that decide for themselves
Action and resource management

Given description of start state, a goal state and a sequence of actions. Outcome is to find the most efficient set of actions to achieve the goal
Transportation, scheduling
Interactive decision making: military planning,



Heuristics here mean common sense rules from experience (in this case, the programmer) , likely using vast personal data trove to build these use cases
Counter part is Algorithmic programming that uses mathematically proven features, quantifying, logic driven

AI and Induction

IBM Watson

Question answering machine

Performance: speed, confidence, estimation, clue selection, betting strategy

- Massive parallelism
- Pervasive confidence estimation
- Integrated shallow and deep knowledge

Lexical Answer type

Daedalus Information Systems

49

Deep QA Project at IBM
Precursor Deep Blue, chess database that beat Gary Kasparov
3 years in development
Stores millions and millions of documents on room full of servers
Lexical answer type: word or noun phrase in the question that specifies the type of answer without deconstructing the question semanticsa
Processing

- Parsing
- Classification
- Decomposition
- Automate answer source
- Evaluation
- Confidence estimation

Critical thinking vs. computational math
Computes and chooses between possible answers (module 5 will look at text classification and training more closely)

Wolfram’s comment from NYT article: can only answer “factual” questions (db questions), cannot for an answer that has judgment

Discussion Question

WHAT IS INTELLIGENCE?

Daedalus Information Systems

50

Machine Learning

Derives rules from a data set

A programming approach to problem-solving – composite of not a single algorithm

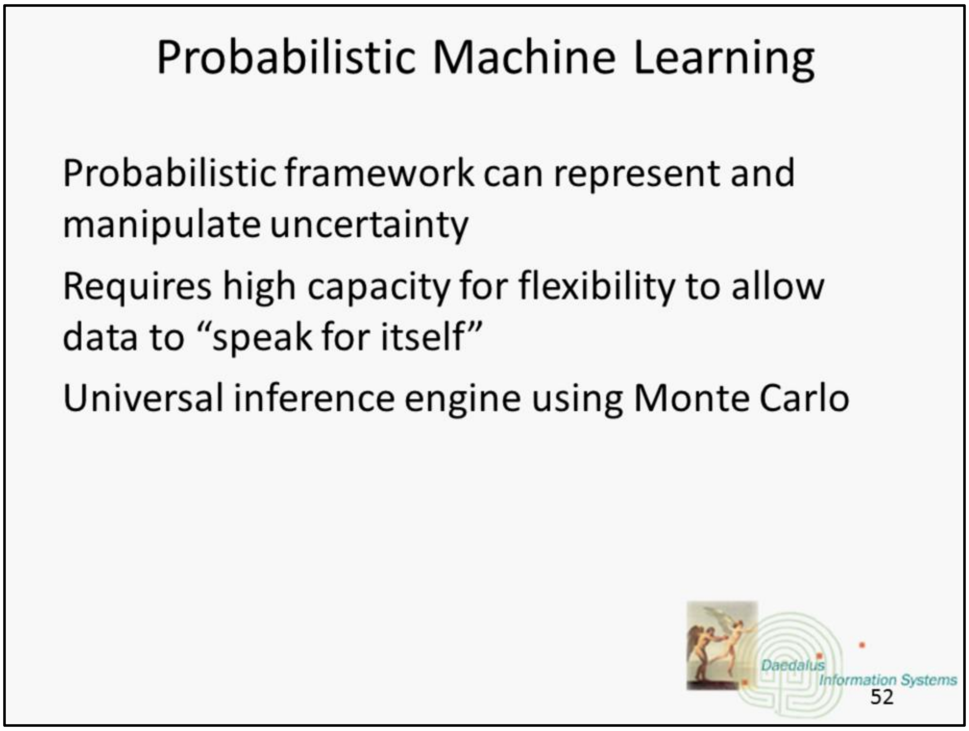
Model of real world using mathematic structure with decision-making rules

Objective function = desired outcome

Training set with adjusted parameters until goal achieved

Test set used to validate accuracy and effectiveness

Daedalus Information Systems



Infers plausibility models to explore observed data
Inference prediction for forecasting using Cox Axiom (spectrum: impossible to absolutely certain)
Dutch Book theorem used for degrees of uncertainty
Universal inference engine infers imputes that match a certain output

Supervised Learning

Uses document-class pairs to indicate proper classes for given documents

Used human specialists for classification of “training set” used to “teach” system

- Assigns classes to documents
- Reviews machine classification performance

6 Algorithm types

- Decision Trees
- Nearest neighbor
- Relevance Feedback
- Naïve Bayes
- Support Vector Machine
- Ensemble

Daedalus Information Systems

53

Labeled data

Regression (estimating relationships between variables for prediction)

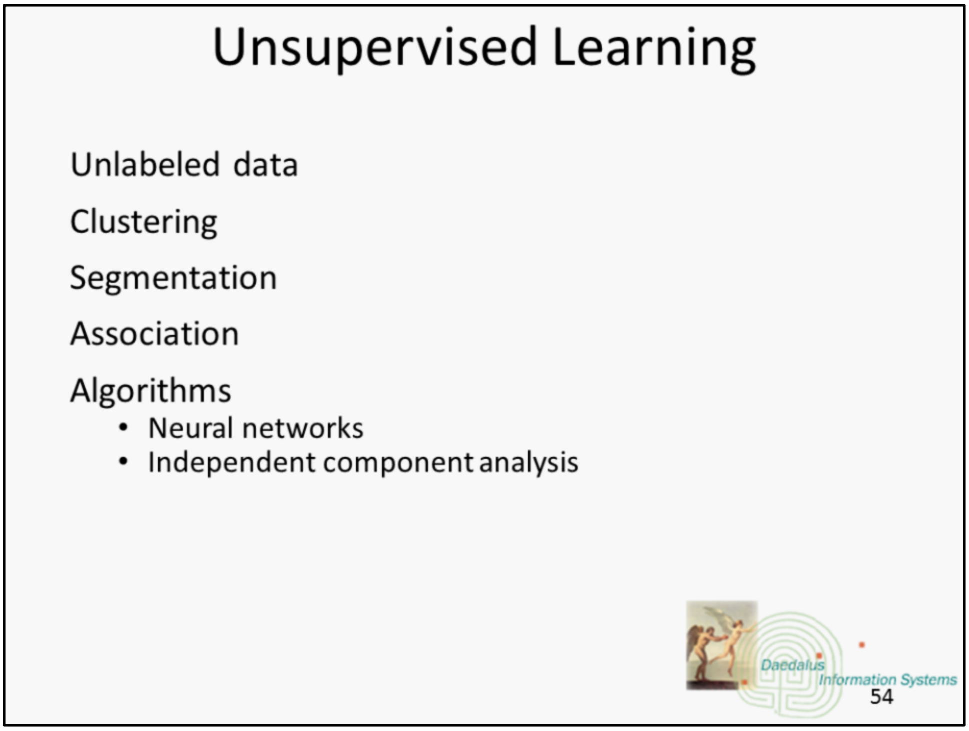
Classification

Ranking

Nearest Neighbor (aka k-NN): no established classification model, done on the fly, classification decision based on nearest neighbor in predefined metric space, more focused on document features and less on global values application (bottom up, document based, classification)

Relevance Feedback (Rocchio): vector space model that allows modification based on user feedback (training set is the feedback mechanism)

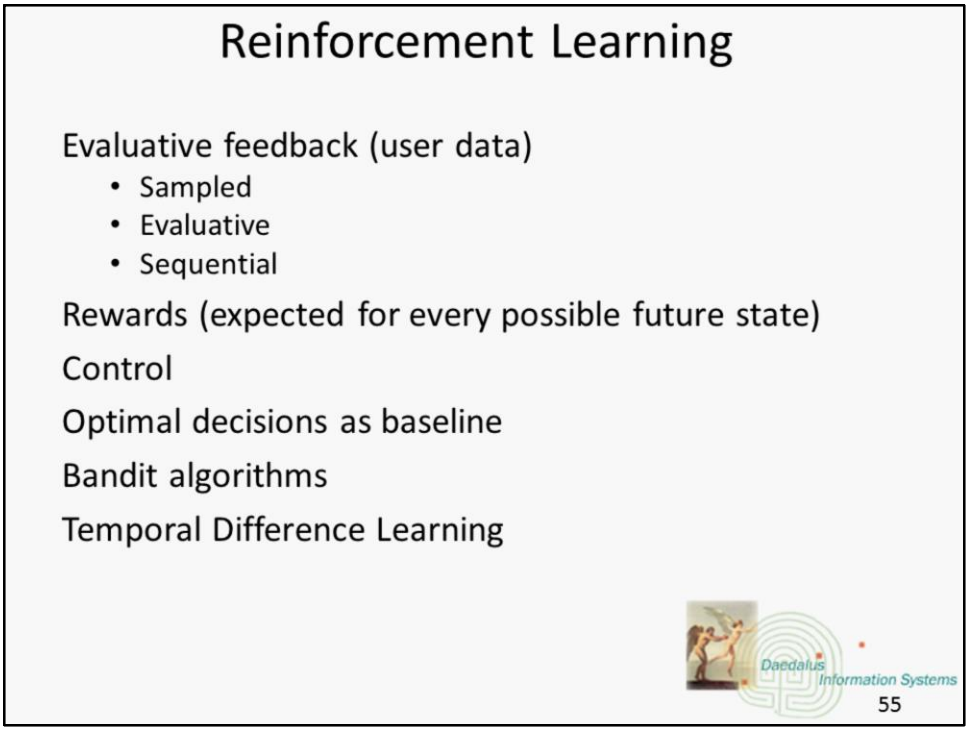
Ensemble: Combines the output of independent classifiers, accuracy = better than random guessing. Meta classifiere takes various classifiers prediction output for document and combines into a single prediction



Neural Network: approximates human brain neural network of nodes and electrons

- Composed of 3 layers: query terms, document terms, actual documents
- Query terms nodes initiates inference process with sent signals to document term nodes
- Uses BM25 Probalistic models that use term weighting (inverse document frequency, term frequency and document length normalization)

Independent component analysis:
Wikipedia: (ICA) is a computational method for separating a multivariate signal into additive subcomponents supposing the mutual statistical independence of the non-Gaussian source signals



Inspired by human decision making
Evaluative feedback is based on decision effectiveness and appropriateness of available alternatives

Control: moving through all possible combination of environments/states with a sequence of actions that maximizes potential for reward

Bandit algorithms: based on slot machine "one armed bandit." Algorithm must decide which decision to make in what sequence to maximize rewards
Collect a large, static, set of bandit decisions using real problems with a random selection algorithm
Supervised learning is a contextual bandit set of algorithms

Uses Stochastic gradient Descent (SGD) This consists of showing the input vector for a few examples, computing the outputs and the errors, computing the average gradient for those examples, and adjusting the weights accordingly.

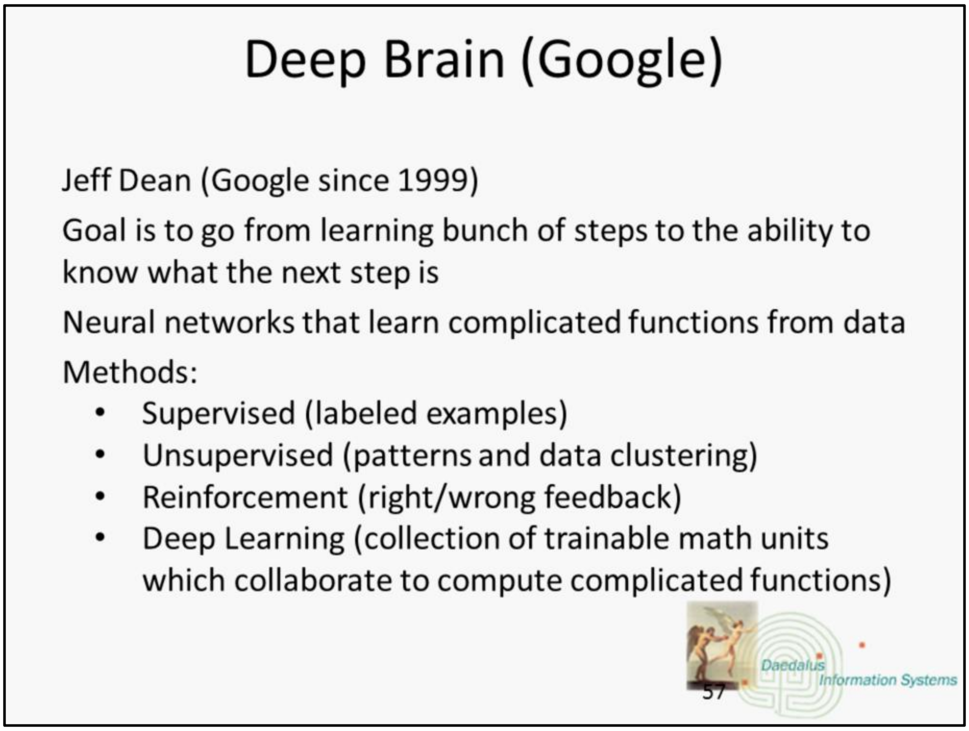
Temporal difference learning: makes predictions about long-term implications. Errors are used as learning signals and later incorporates into a generalized learning objective [Google likely uses temporal difference learning for geo-based local profiles to inform decisions.] – modeled on human brain neuron firing of dopamine signals

Deep Learning

"Deep learning, a form of machine learning based on layered representations of variables referred to neural networks, has made speech-understanding practical on our phones and in our kitchens, and its algorithms can be applied widely to an array of applications that rely on pattern recognition"

Stanford 100 AI Study

Daedalus Information Systems



Google acquired machine learning companies over the last years: Deep Mind 2013
Cannot write algorithms for all of the tasks, must write algorithms that learn from each other
Data Set

- Text (trillions of words in all languages)
- Visual (billions of images)
- Audio (thousands of hours of speech)
- User activity (searches, clicks, time on site)
- Knowledge Graph (billions of labeled relationships)

Deep Learning

HUGE raw data training set

Results get better with more data, new/better algorithms based on observation and insight

Powerful processing hardware is key

Machine Learning requirements

- Ease of expression for many types of ideas/algos
- Scalable
- Portable
- Reproducible
- Extensible

RankBrain search engine relevance model

Daedalus Information Systems

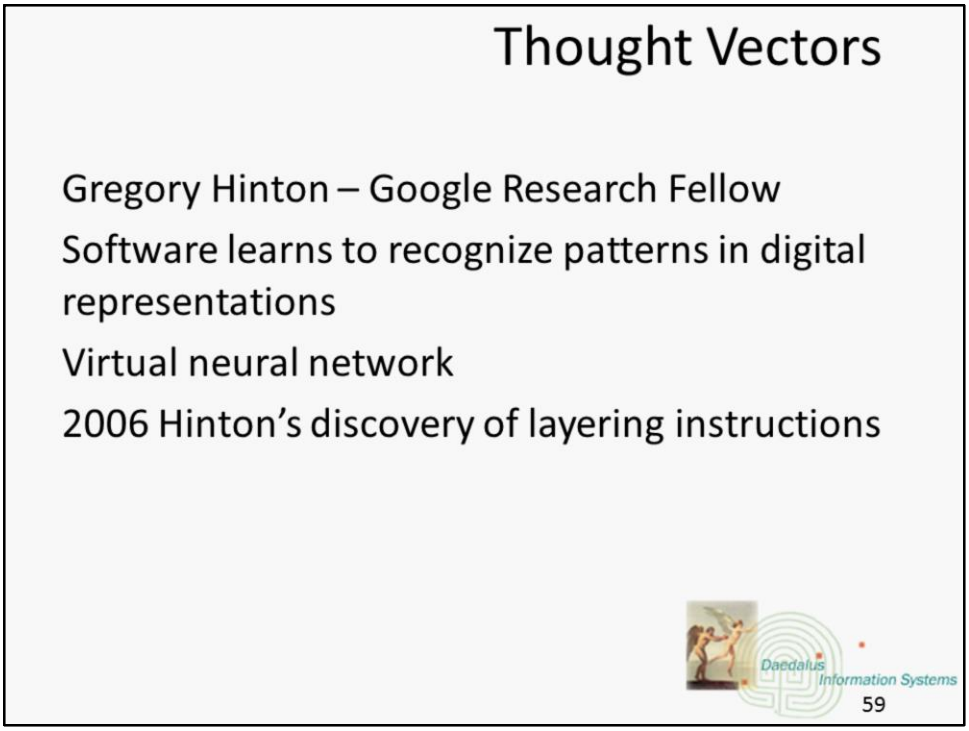
58

Text (documents, queries)
Visual data (images)
Audio (speech, music, sounds)
User activity (mark spam, engagement metrics, etc)
Knowledge Graph (acquisition of structured repository Freebase)

Hardware: Tensor CPU

Portable: use across many platforms
Reproducible: by others
Extensible beyond research phase

Uses: search, robotics, self-driving cars, healthcare, video retrieval, voice search, personal assistance (chat bots)

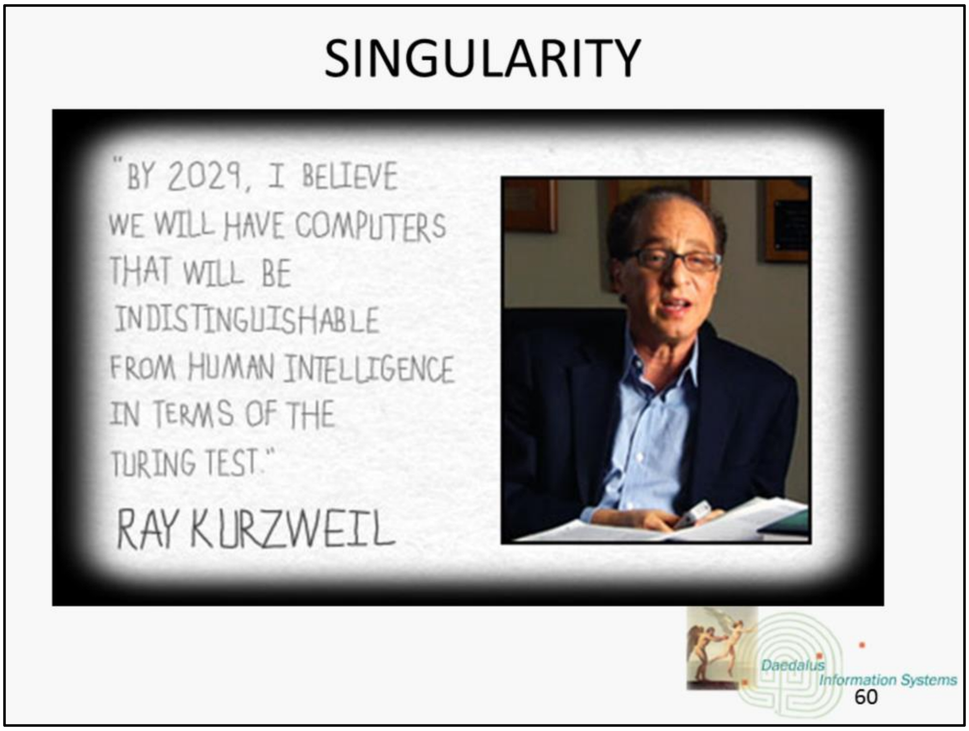


System maps out set of virtual neurons and assigns random values or weights to connections – weight determines how each stimulated neuron responds to a digitized feature and adjusted until system responds correctly
Layering: first layer learns primitive features by finding pixel combinations that occur more often then should by chance then feeds to next layer on recognized features learned – repeated

Google produces an image of a cat: 10 million randomly selected videos with cats along with other subjects – used 16000 processors for parallel processing – 16% success rate

Deployed on android voice search resulted in 25% reduction in errors

Deep mind = software that can learn using a deep learning method (reinforcement learning)



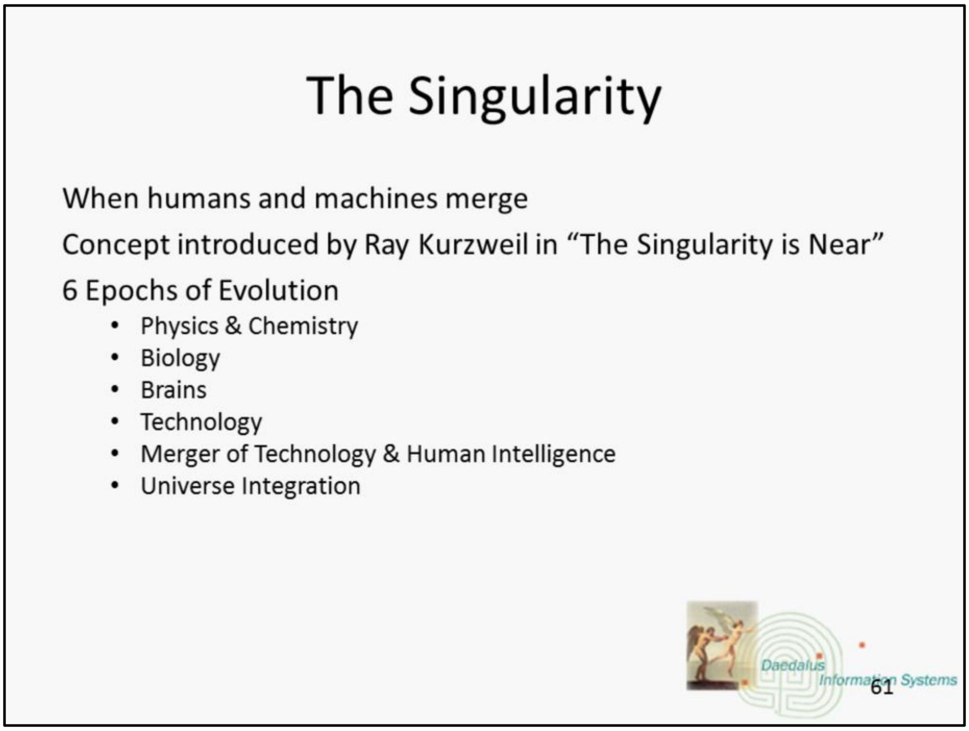
http://singularity.com/
Concept introduced by Ray Kurzweil in "The Singularity is Near"

When humans and machines merge – "one in which humans gained near immortality by becoming one with robotic technology" [Why the Future Doesn't Need Us: Bill Joy: Wired Magazine August 2000]

Jaron Lanier refers to the Singularity as being like the Rapture in that it cannot be verified by the living. He makes the point in Apocalypse of Self-Abdication that information is an artifact of human thought (much like Jens-Erik Mai).

Bill Joy, creator of Java and Jini, was so disturbed by its line of thinking that he wrote "The Future Does Not Need Us" for Wired magazine (2000)

A recent article (Dec 2015) from Technology Review [https://www.technologyreview.com/s/544606/can-this-man-make-aimore-human/] that claims AI will need to "learn" like children, by example and also by storing, manipulating and applying rules and making conclusions achieved through experience.



http://singularity.com/
E1: Physics and chemistry: information in atomic structures
E2: Biology: information in DNA
E3: Brains: Information in neural paths
E4: Technology: Information in hardware and software design
E5: Merger of Technology & Human Intelligence: methods of biology (including human intelligence) are integrated into the (exponentially expanding) human technology base
E6: Universe Wakes Up: Patterns of matter and energy in the universe become saturated with intelligent processes and knowledge

Gelernter (2016)

Tides of Mind: Uncovering the Spectrum of Consciousness

Key question going unanswered: What is the human mind without the human being?

The human mind is not just creation of thought and collection of data; also a product of feelings, composite of sensations, memories, ideas that are worked and reworked over a lifetime.

Computer science Yale University
Artist and writer

http://time.com/4236974/encounters-with-the-archgenius/

Discussion Question

IS EMBODIMENT A CRITICAL COMPONENT OF EXPERIENCE

Daedalus Information Systems

63

PERSONALIZATION: USER DATA SET

Daedalus Information Systems

In 2002, Google acquired personalization technology Kaltix and founder Sep Kamver who has been head of Google personalization since. Defines personalization: “product that can use information given by the user to provide tailored, more individualized experience”

Query Refinement

- System adds terms based on past information searches
- Computes similarity between query and user model
- Synonym replacement

Dynamic query suggestions - displayed as searcher enters query

Results Re-ranking

- Sorted by user model
- Sorted by Seen/Not Seen

Personalization of results set

Calculation of information from 3 sources

- User: previous search patterns
- Domain: countries, cultures, personalities
- GeoPersonalization: location-based results

Metrics used for probability modeling on future searches

- Active: user actions in time
- Passive: user toolbar information (bookmarks), desktop information (files), IP location, cookies

Privacy Paradox

Privacy risk is weighed against value of object, interaction, end result
- Research assumes user calculates an internalized value
- Basis for choice to reveal personal identification information (PII)

Value is determined by the smoothness of the interaction (Groupon, Amazon Local)
- Value proposition overrides security/privacy concerns

Higher level of user control over PII reduces the perception of risk

Daedalus Information Systems

65

Personalization Privacy Paradox: An exploratory study of decision making process for Location-aware marketing: Xu, Luo, Carroll, et.al.
Study focused on location-aware marketing (LAM) – targeting ads, groupons based on awareness of user location, preferences, etc.

Users share private information in exchange for some THING of perceived value and based on assumptions
Agency will deliver (paper, goods, etc.)
They will not share the information indiscriminately
Will protect the data

Users assume a social contract on the part of the agency that they will be responsible

The ease of usability influences the willingness to proceed – Obama campaign online voter registration 2008 – long form split into small, digestible chunks
Previous privacy risk is more influential in covert model (e.g. tracking without user awareness)

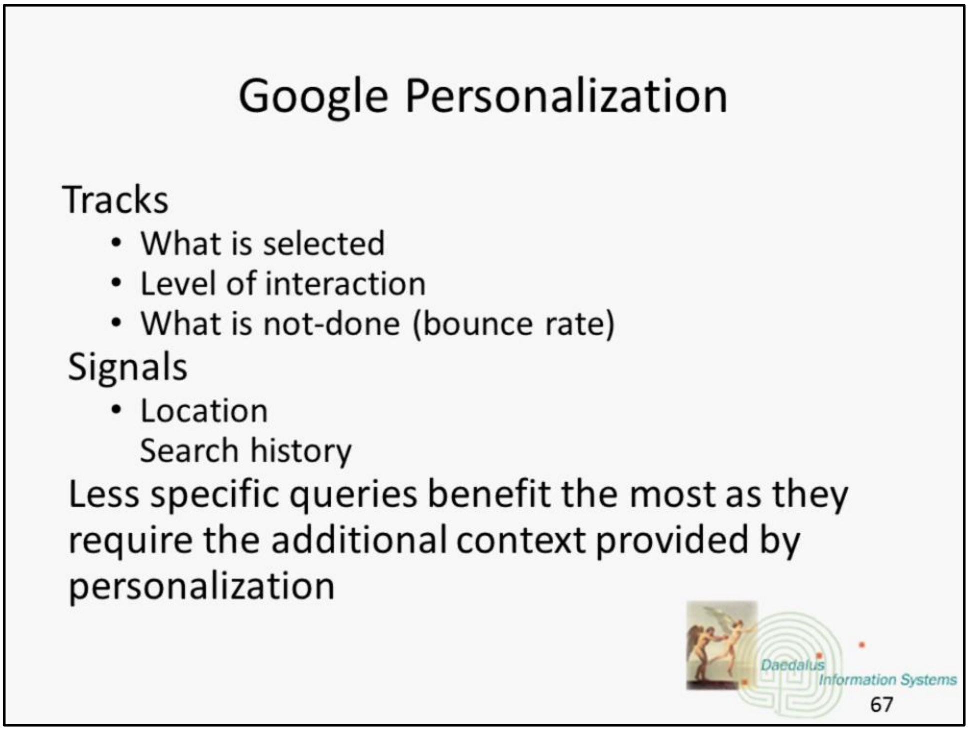
Google on Privacy

"There was a small trade off on privacy but they're going to get dramatically better search results. That was something that made sense to us over time."

Melissa Mayer
VP User Experience
Google

Daedalus Information Systems
66

www.google.com/history
Web history tied to the Google toolbar (first launched in 2000) and the ability to track what user looked at across the Web



Metrics used for probability modeling on future searches

- Active: user actions in time
- Passive: user toolbar information (bookmarks), desktop information (files), IP location, cookies

In 2002, Google acquired personalization technology Kaltix and founder Sep Kamver who has been head of Google personalization since
Defines personalization: “product that can use information given by the user to provide tailored, more individualized experience”
Personalization enables shorter, less specific queries set to change user behavior (easier, more natural queries) = search shorthand

Tied direct user interaction with results (ability to promote/demote in results set, add comment) discontinued because too noisy & interest did not always equal searching for topic and used by SEO community for other purposes

- Only enable if signed in
- Only impacted future searches (if signed in)

T

Google Collected Data Types

MICRO: collected over milliseconds

MACRO: data collection over time

MESO: Collected over minutes

Daedalus Information Systems

68

MICRO:
Eye tracking
Visual complexity study showed that users make a decision on attractiveness of a website within 17 milliseconds
Used for feature optimization

MACRO:
Search logs, user profiles
Aggregated over time
Used to learn about search strategies over time

MESO:
Ethnographic studies
Field studies
Hands on user studies
Used to learn about search behavior and habits

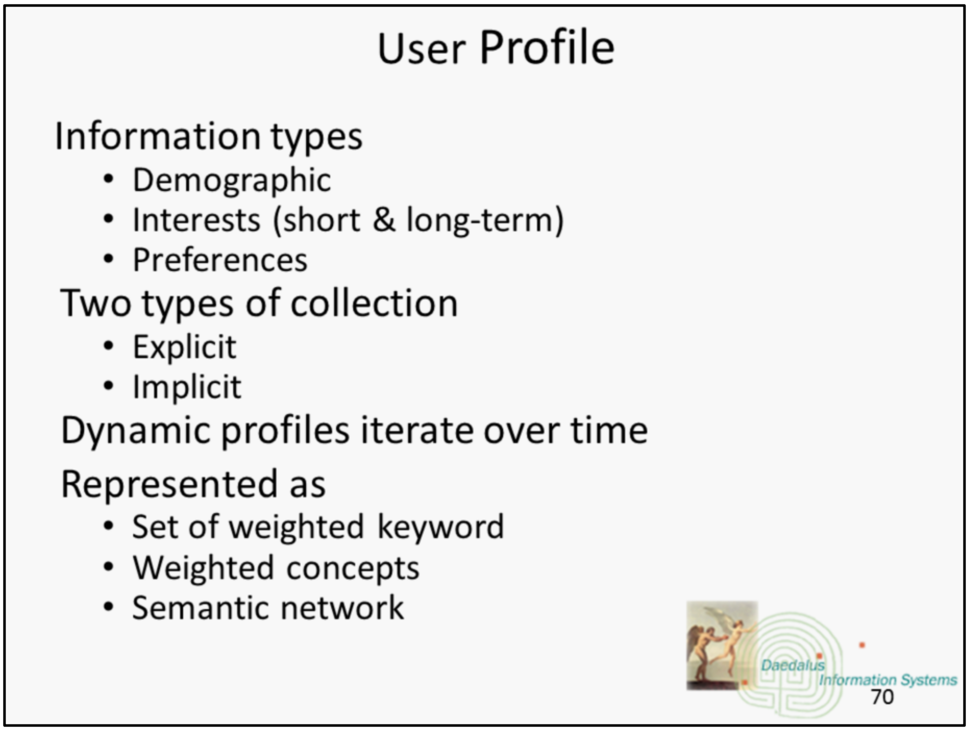
What Google Collects

Profile information
Use information
Device information
Log information
Location information
Unique application information
Local storage
Cookie data

Daedalus Information Systems
69

Google Privacy Policy http://www.google.com/policies/privacy/
shared across services

- Profile information: Information you give us. For example, many of our services require you to sign up for a Google Account. When you do, we'll ask for personal information, like your name, email address, telephone number or credit card. If you want to take full advantage of the sharing features we offer, we might also ask you to create a publicly visible Google Profile, which may include your name and photo.
- Use information: Information we get from your use of our services. We may collect information about the services that you use and how you use them, like when you visit a website that uses our advertising services or you view and interact with our ads and content. This information includes:
- Device information: We may collect device-specific information (such as your hardware model, operating system version, unique device identifiers, and mobile network information including phone number). Google may associate your device identifiers or phone number with your Google Account.
- Log information "When you use our services or view content provided by Google, we may automatically collect and store certain information in server logs. This may include:
    - details of how you used our service, such as your search queries.
    - telephony log information like your phone number, calling-party number, forwarding numbers, time and date of calls, duration of calls, SMS routing information and types of calls.
    - Internet protocol address.
    - device event information such as crashes, system activity, hardware settings, browser type, browser language, the date and time of your request and referral URL.
    - cookies that may uniquely identify your browser or your Google Account.
- Location information: When you use a location-enabled Google service, we may collect and process information about your actual location, like GPS signals sent by a mobile device. We may also use various technologies to determine location, such as sensor data from your device that may, for example, provide information on nearby Wi-Fi access points and cell towers.
- Unique application numbers" Certain services include a unique application number. This number and information about your installation (for example, the operating system type and application version number) may be sent to Google when you install or uninstall that service or when that service periodically contacts our servers, such as for automatic updates.
- Local storage: We may collect and store information (including personal information) locally on your device using mechanisms such as browser web storage (including HTML 5) and application data caches.
- Cookies and anonymous identifiers: We use various technologies to collect and store information when you visit a Google service, and this may include sending one or more cookies or anonymous identifiers to your device. We also use cookies and anonymous identifiers when you interact with services we offer to our partners, such as advertising services or Google features that may appear on other sites.



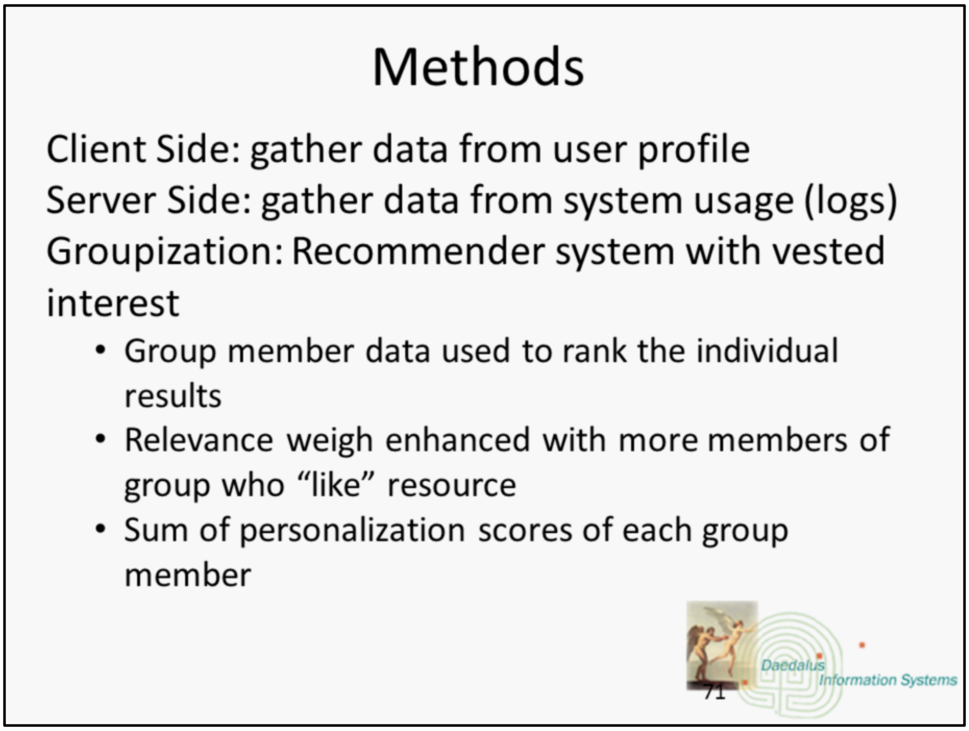
User profile phases

1. Gather raw information
2. Construct profile from user data
3. Allow application to exploit profile to construct personal results

Keywords profiles represent areas of interest

- Extracted from documents or directly provided by user, weights are numerical representation of user interest
- Polysemy is a big problem for KW profiles

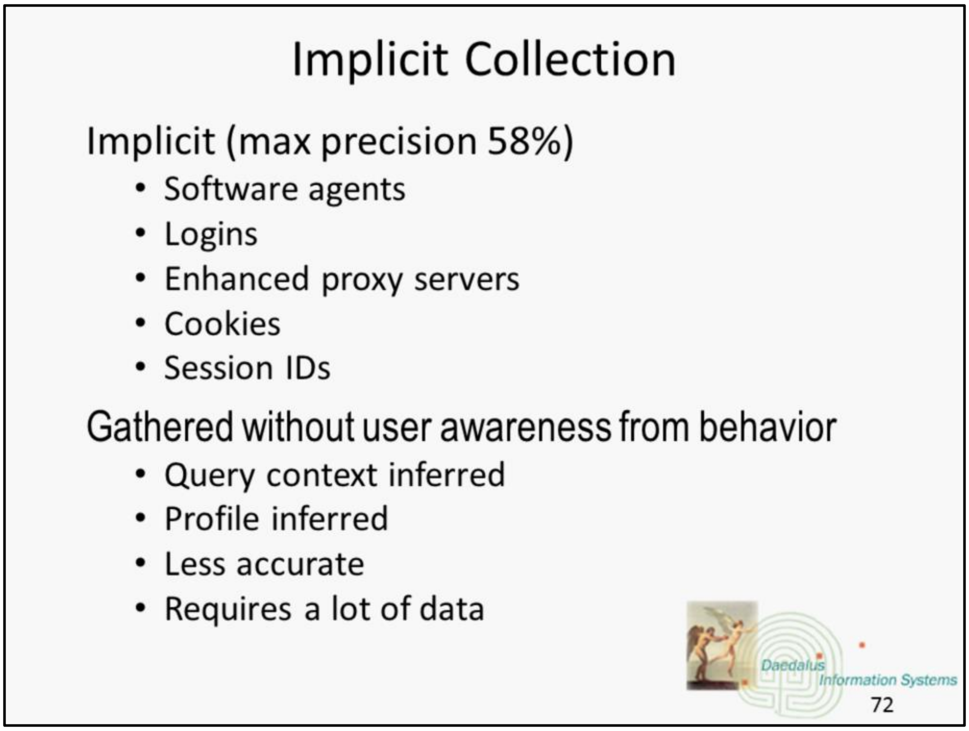
Semantic networks
Filtering system
Network of concepts – unlinked nodes with each node representing a discrete concept
Used by alta vista (used header that represented user personal data, set of stereotypes (prototypical user comprised of a set of interests represented by a frame of slots
Each “slot” (made up of domain, topic & weight (domain =area of interest, topic = specific term used to identify area of interest, weight = degree of interest) that makes up frame weighted for relevance



Source: Information Retrieval: Personalization 2 Fernando Diaz Yahoo! Labs April 25, 2011

Use group level information rather than individual
Groups intersect with work or social interests (or both)

Advantages: can protect group privacy, potentially richer user signals
Disadvantages: no external data used, client side resource challenges (processing/space)



Jaime Teevan MS Research
(http://courses.ischool.berkeley.edu/i141/f07/lectures/teevan_personalization.pdf)

Tools used
Software agents: most reliable as more control over install and application
Cookies: least invasive
Login: more pervasive across machines and time
Proxy Servers: limited to user register of machine with server
Session IDs: limited to a single session

Advantages: more data, better data (easier for system to consume and rationalize)
Disadvantage: user has no control over what is collected

Explicit Collection

Explicit (max precision 63%)

- HTML forms
- Explicit user feedback interaction (early Google personalization with More Like This)

Provided by user with knowledge

More accurate as user shares more about query intent and interests

Daedalus Information Systems

73

Advantage: User has more control over personal and private information

Disadvantage: compliance, users have a hard time expressing interests, burdensome on user to fill out forms, false info from user

Context

Context becomes what the system can measure

- Environmental features
- Interactions
- Ubiquitous computing
- Internet of things (IoT)

Non-methodical approach that brings in containment (social through local) interactions

- Adaptive/reactive interaction in situ
- Context as perceived and used by actor

Daedalus Information Systems

74

AI ETHICS & SAFETY

Daedalus Information Systems

75

Adversarial AI

Used in adversarial setting

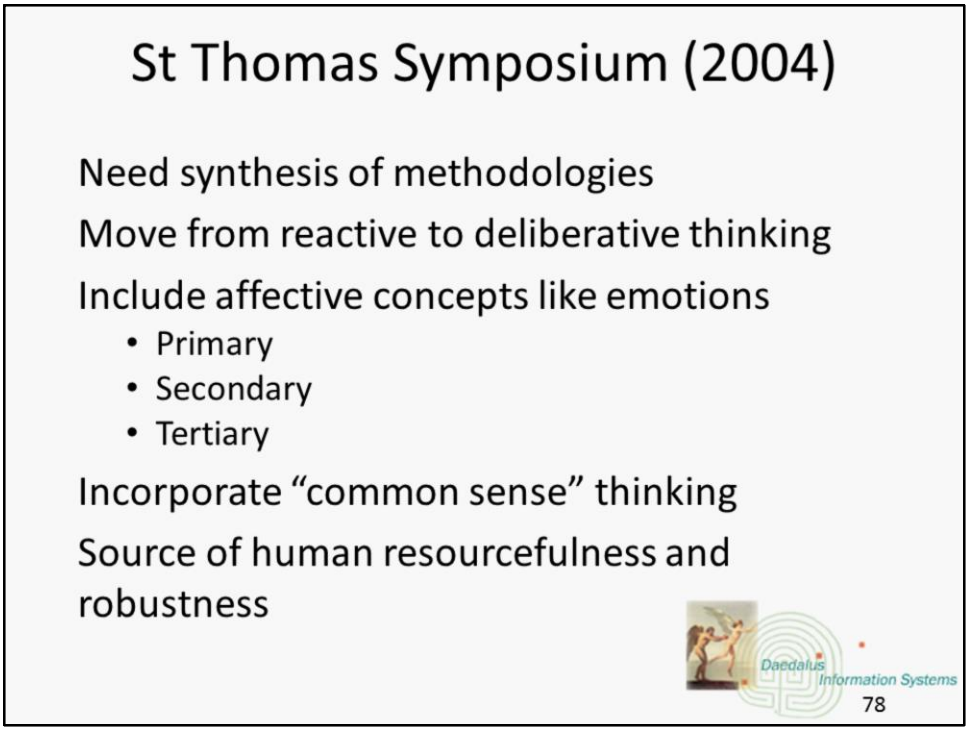
Used for malicious means

Adversarial setting example: Russian hacking of US election
Malicious means: Fall 2016 IoT hack that took down part of the internet

3 Laws of Robotics

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm
2. A robot must obey orders give in to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Isaac Asimov in I, Robot



Designing Architectures for Human Level Intelligence - St Thomas VI 2004

Marvin Minsky MIT
Push Singh MIT
Aaron Sloman University of Birmingham UK – Cognition and Effects project

Levels of control
Emphasizes the role of motives (generally distinguished from emotions) – motives have both emotional and cognitive correlates

Emotional Machine Architecture
http://www.aaai.org/ojs/index.php/aimagazine/article/view/1764/1662

Human level intelligence requires a machine architecture that can support multiple ways to represent | acquire | apply common sense knowledge
Requirements

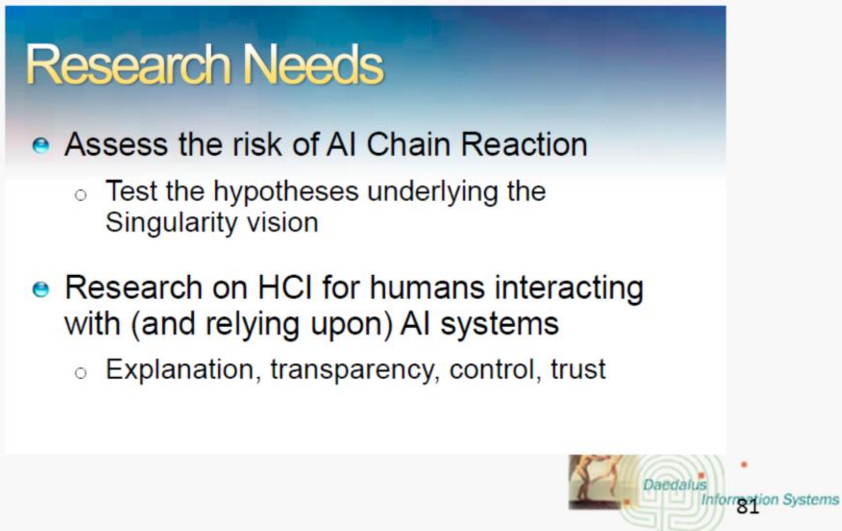
- Development of an "intention-based" programming languages
- Vigilance for programmer bias
- Virtual model world
- Develop and organize mini scenarios
- Organize and evaluate progress
- Create interoperable protocols
- Catalogue all "ways to think" (reasoning about prediction, explanation, generalization, exemplification, abstraction)
- Debug learning

http://www.aaai.org/Organization/presidential-panel.php

Called together computer scientists to consider AI successes, address changes and opportunities in light of the success.
Reflect on potential socioeconomic, legal and ethical issues surrounding machine intelligence
Review concerns about control of computer-based intelligences
Consider proactive actions that could enhance long-term societal outcomes

Participants: Margaret Boden, Craig Boutilier, Greg Cooper, Tom Dean, Tom Dietterich, **Oren Etzioni (UW, Allen Institute of Brain Science)**, Barbara Grosz, **Eric Horvitz (Microsoft technical Fellow)** , Toru Ishida, Sarit Kraus, Alan Mackworth, David McAllester, Sheila McIlraith, Tom Mitchell, **Andrew Ng (Baidu)**, David Parkes, Edwina Rissland, Bart Selman, Diana Spears, Peter Stone, Milind Tambe, Sebastian Thrun, Manuela Veloso, David Waltz, Michael Wellman

Asilomar AI Research Needs: HCI



Source: Asilomar Study on Long-Term AI Futures: Highlights of 2008-2009 Study: Presidential Panel on Long-Term AI Futures IJCAI 2009 Invited Panel, July 2009

AI Risks

Mis-specified Objectives

Negative Side Effects that extend to wider application

Hacking: rewards, devices

Bad extrapolation of the real world

Poor training data

Privacy

Fairness

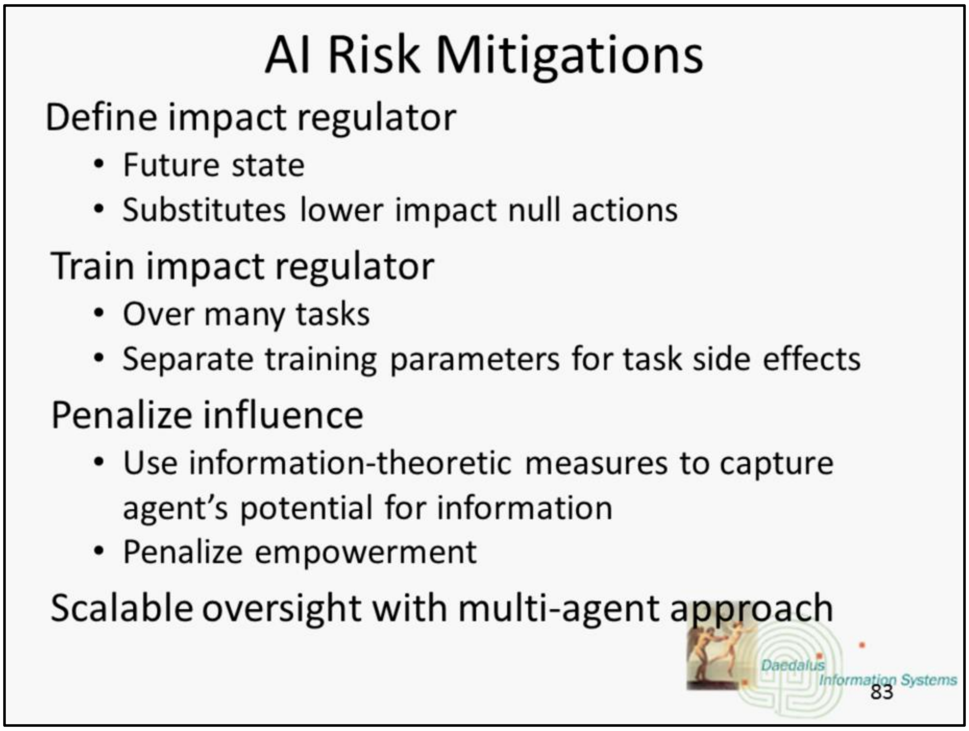
Abuse

Transparency

Daedalus Information Systems

82

Google Report on AI Safety
Privacy: right to be forgotten
Fairness: digital divide
Security: IoT takedown of internet, GM self-driving car
Transparency: common understanding of complex engineering



Multi-agent approach = human and agent working together

Reward Hacking: adversarial reward function, careful programming to avoid adversarial blindness

Scalable oversight: distant supervision, hierarchical reinforcement learning

AI Risk Mitigations 2

Use Objective functions to capture designer informal intent

- No partially observed goals
- Concrete, not abstract rewards
- Deep correlation between tasks and functions

Feedback loops

- Model look ahead
- Reward capping
- Counter example resistance – combination of rewards

Daedalus Information Systems

84

Correlation between tasks and rewards: do not base cleaning robot reward on amount of cleaning supplies used

AI Risk Mitigations 3

Safe exploration

- Risk sensitive performance criteria
- Use demonstration
- Simulated exploration

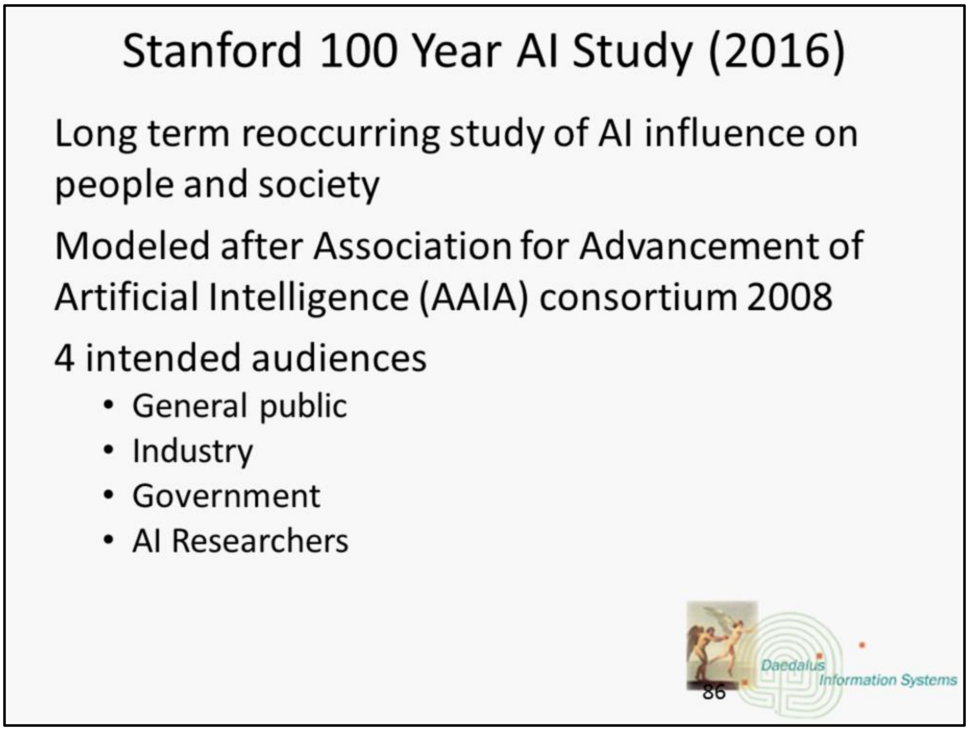
Well defined models

- Train on multiple distributions
- Program for out-of-distribution situations

Daedalus Information Systems

85

Simulated exploration: bounded exploration, trusted policy oversight, human oversight



AAAI organized by Eric Horvitz – AI experts along with scholars from cognitive science, philosophy and law

Salient Domains: transportation, service robots, healthcare, education, low-resource communities, public safety and security, employment and workplace and entertainment

Discussion Question

ARE THE TRADE-OFFS WORTH IT?

Daedalus Information Systems

87

THE IA OF AI

Daedalus Information Systems

88

AI Design According Computer Science

Components

- Variables
- Domains (environment)
- Constraints (limits)

Goal of AI Design = satisfy constraints

Admissible heuristic: if it costs too much to reach solution state then revise or reject

Daedalus Information Systems

89

AI and **Design**
Krzysztof Gajos
Harvard University CS 182, Fall 2011

http://isites.harvard.edu/fs/docs/icb.topic958294.files/CS%20182%20-%20AI%20and%20Design%20-%202011.pdf

Information Architecture and AI

Site Structure

Connections

Proto-typicality (mental models)

Visual complexity (text over images)

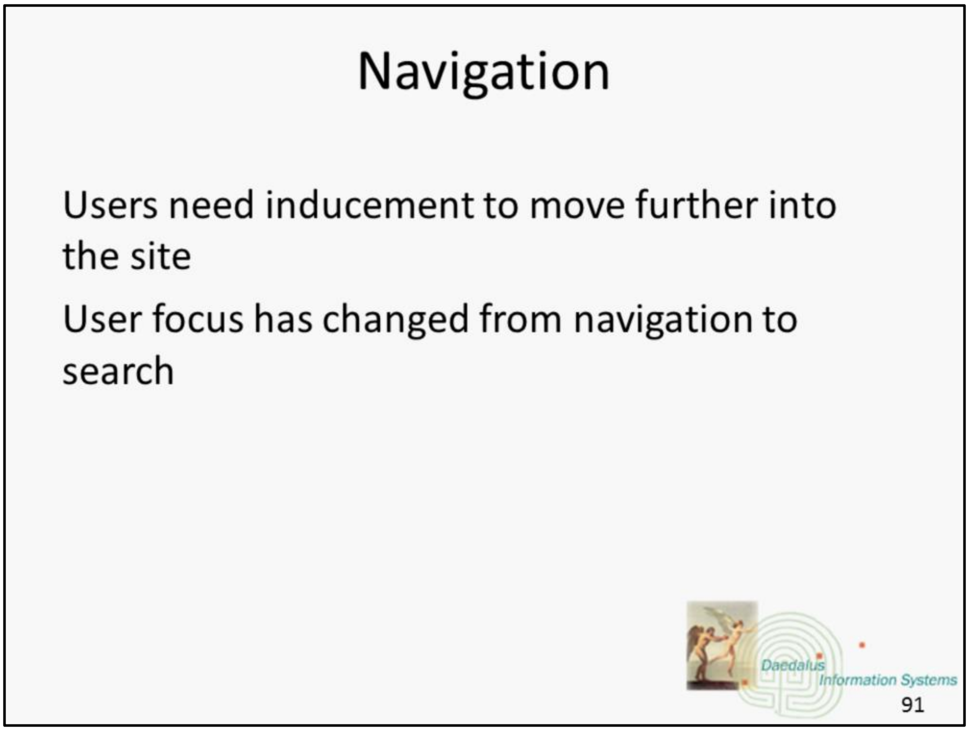
Daedalus Information Systems

90

Legacy newspaper structure of “the fold.”

Proto-typicality: user mental models

Visual complexity: ratio of images to text favors text



Information Architecture: Structure is the Search Aphrodisiac

Distance reflects relevance
- URL Depth: the further from the homepage, the less important it must be
- Click Distance: the further from an authority page, the less important it must be

Page Structure Now a Factor
- Google Page Segmentation Patent: Determining Semantically Distinct Regions of a Document
- Based on eye-tracking studies and user behavior
- Similar Yahoo patent

Enhanced Navigation

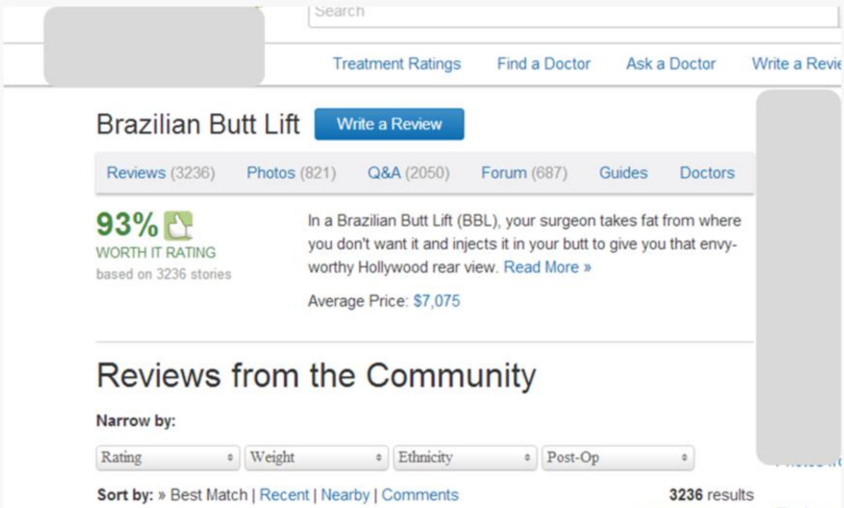
Machine readable text

Related content model

Schema markup

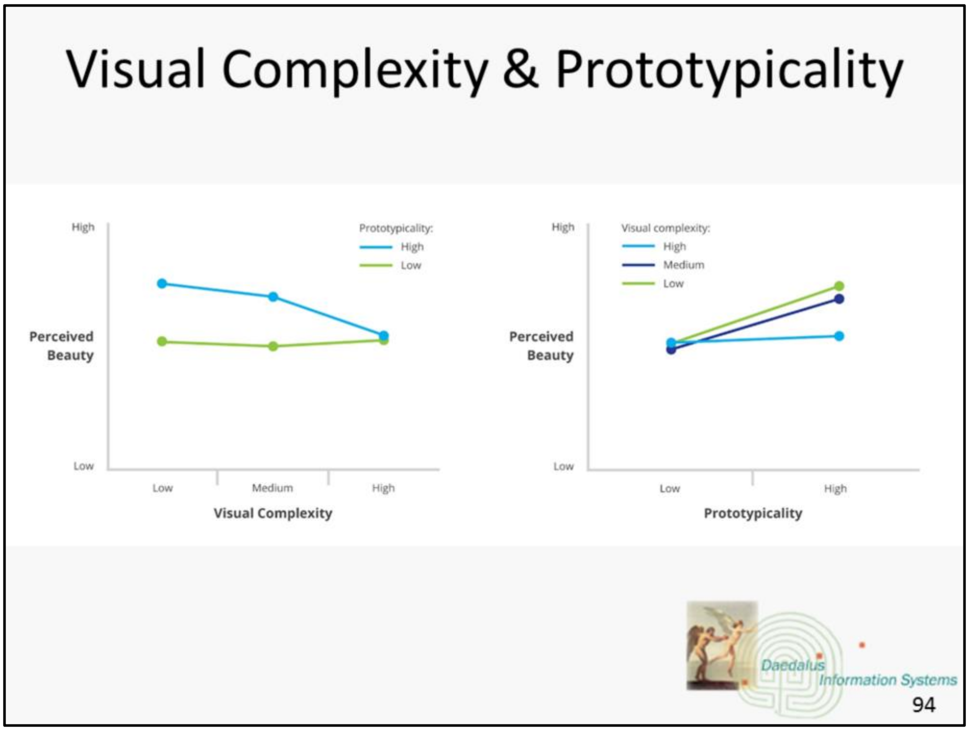
Put the sidewalks where the footprints are
Resource: Stuart Brand: How Buildings Learn

Too Many Choices



Daedalus Information Systems

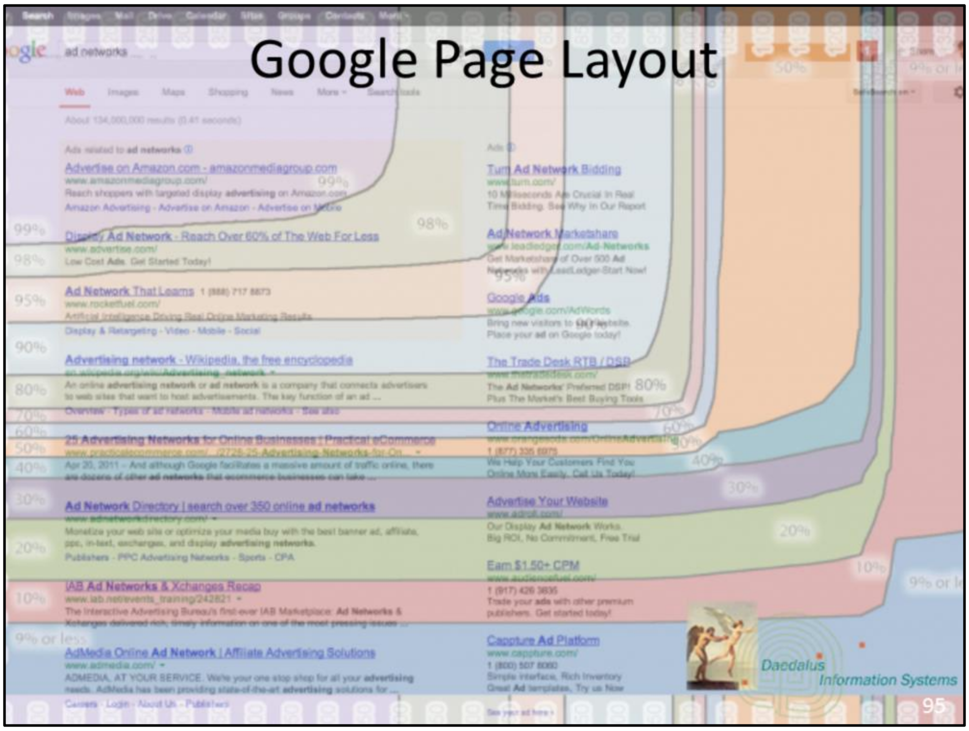
93



VISUAL COMPLEXITY & PROTOTYPICALITY
The results show that both visual complexity and proto-typicality play crucial roles in the process of forming an aesthetic judgment. It happens within incredibly short timeframes between 17 and 50 milliseconds. By comparison, the average blink of an eye takes 100 to 400 milliseconds.

In other words, users strongly prefer website designs that look both *simple* (low complexity) and *familiar* (high prototypicality). That means if you're designing a website, you'll want to consider both factors. Designs that contradict what users typically expect of a website may hurt users' first impression and damage their expectations.
August 2012

Resource: http://googleresearch.blogspot.com/2012/08/users-love-simple-and-familiar-designs.html



From Patent: Techniques for approximating the visual layout of a web page and determining the porting of the page containing significant content.

“As we’ve mentioned previously, we’ve heard complaints from users that if they click on a result and it’s difficult to find the actual content, they aren’t happy with the experience. Rather than scrolling down the page past a slew of ads, users want to see content right away. So sites that don’t have much content “above-the-fold” can be affected by this change.”
http://googlewebmastercentral.blogspot.com/2012/01/page-layout-algorithm-improvement.html

Resources
http://www.seobythesea.com/2011/12/10-most-important-seo-patents-part-3-classifying-web-blocks-with-linguistic-features/
http://www.seobythesea.com/2008/03/the-importance-of-page-layout-in-seo/

THE UX OF AI

Daedalus Information Systems

96

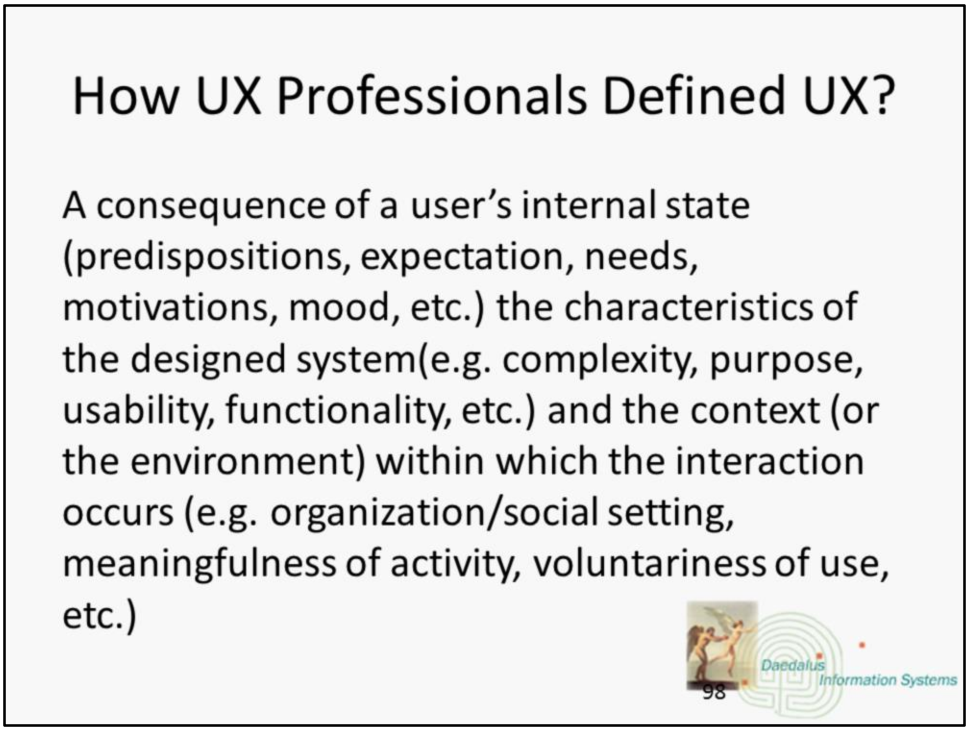
What is the ISO Definition of UX?

A person's perceptions and responses that result from the use or anticipated use of a product, system or service.

Daedalus Information Systems

97

Draft definition



Understanding, Scoping and Defining User Experience: A Survey Approach: Law, Roto, Hassenzahl, Vermeeren, Kort (CHI 2009)

No shared, clear definition of User Experience

- UX is associated with broad/fuzzy concepts that encompass emotion, hedonic, aesthetic , experiential values
- Units of measure are too squishy (quantitative more than qualitative)
- UX discipline is influenced by too many and too diverse theoretical models (again too squishy)

Survey of UX professionals picked this definition

Panda Algorithm Negative Signals

High % of deep content
Low amount of original content
High amount of ads or gratuitous images
Large quantity of boiler-plate text
Over-optimized (too many links)
High bounce rate
Low visit duration
Low CTR from Google search results
No/Low quality in-links
No/Low social mentions

Daedalus Information Systems
99

February 2011
Multiple updates over the ensuing years
Focused on getting rid of “low quality” or “thin sites” so that high quality sites are at the top of the results

User Metrics Training Data

Frequency of access

Click-through (selection from results set)

Time on site

Pages per session

Bounce Rate

Conversion (fulfilled information need)

Profile data

Daedalus Information Systems

100

Why Google Focused on UX for Ranking

Silo-i-zation of Web design and development

Fragmentation of UX community

UX lack of interest in SEO

Daedalus Information Systems

101

Google does not care about UX (just look at android)
Like it or not, part of Google’s evil strategy in selecting the UX community is because they think that we have our heads in the clouds.

Key UX Data Points

Conversions

Unique Visitors

Bounce rate

Social Actions

Number of Pages/visited

Average time on page (exclude bounces)

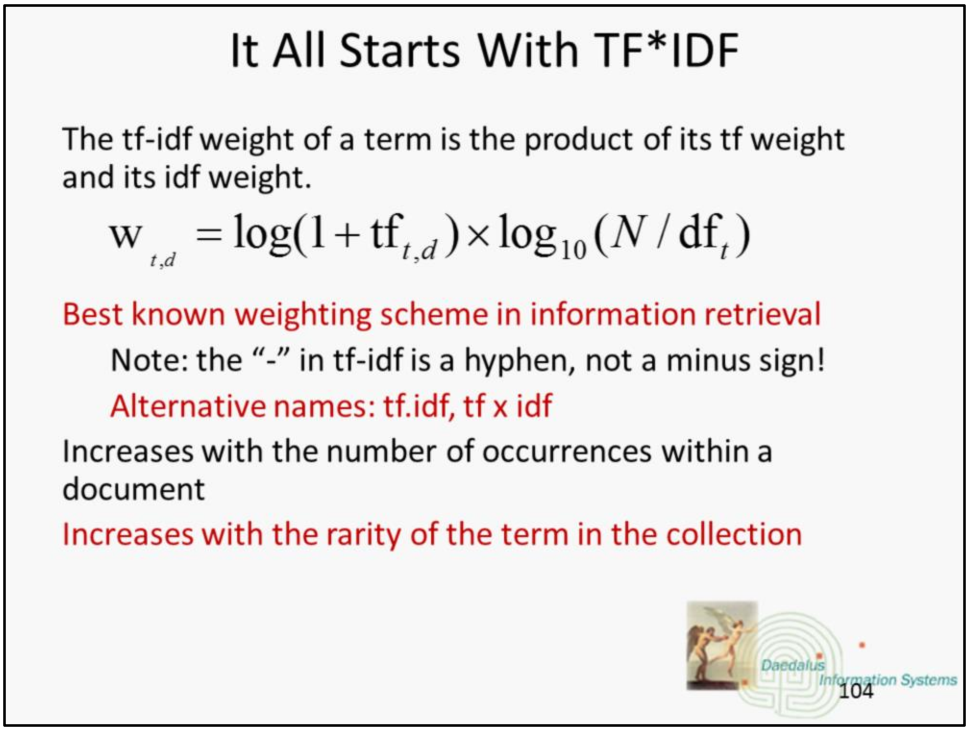
Exit rate

Daedalus Information Systems

102

THE CONTENT STRATEGY OF AI

Daedalus
Information Systems



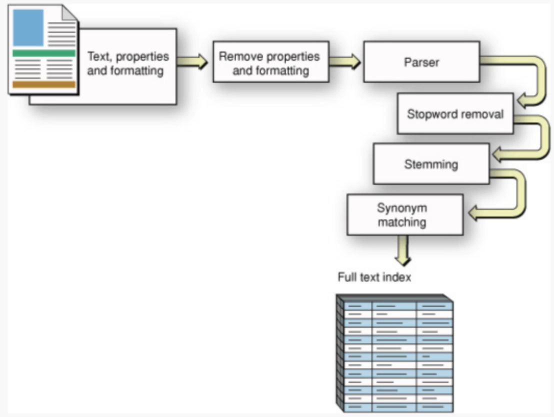
t = how many times the query appears in a document
T = total number of terms in a document
D = set of all documents
d = number of documents with that term

Highest score wins!

The document with the highest proportion of terms which are part of the query is most relevant

- Documents containing more of the term(s) scored higher
- Longer documents discounted
- Rare terms weighted higher

Text Processing by Algorithm



Notes formatting

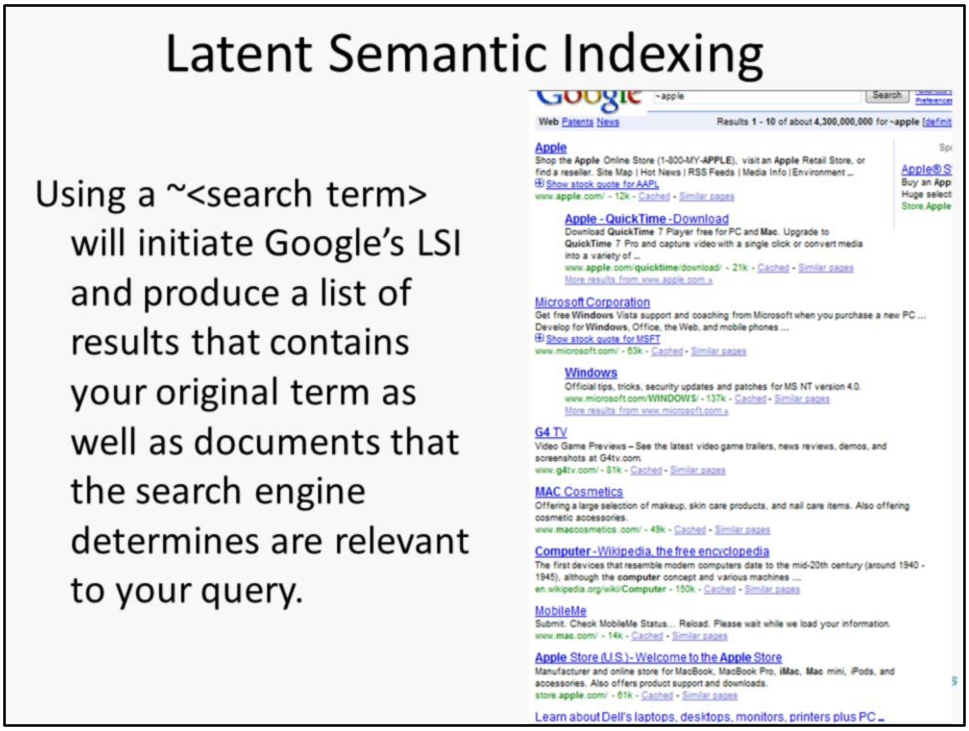
Removes stop words

Stems (reduces to base grammatical concept)

Matches synonyms

105

Minimum term length – eliminate stop words (generally articles, and, if, the, but an, or…) and those too short (unlikely candidates for search)
Synonyms: manually done for smaller indexes, LSI for Web search engines
Stemming: reduces to most basic root, also known as lemmatization, some use of n-grams (***n*-gram** is a contiguous sequence of *n* items from a given sequence of text or speech, can be syllables, letters, words or base pairs collected from corpus)

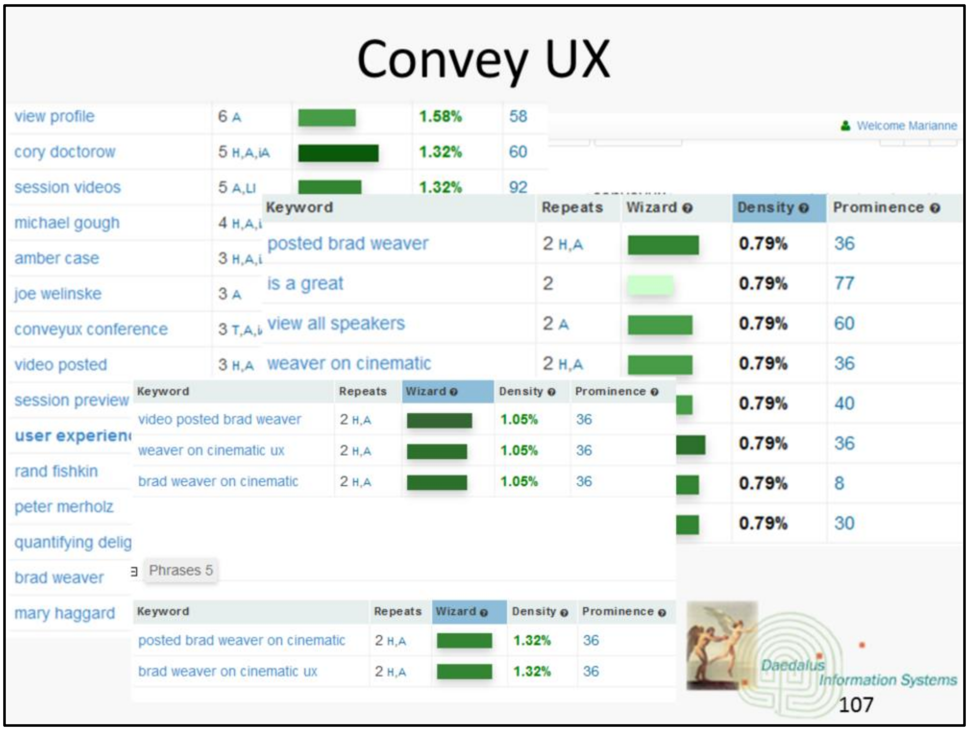


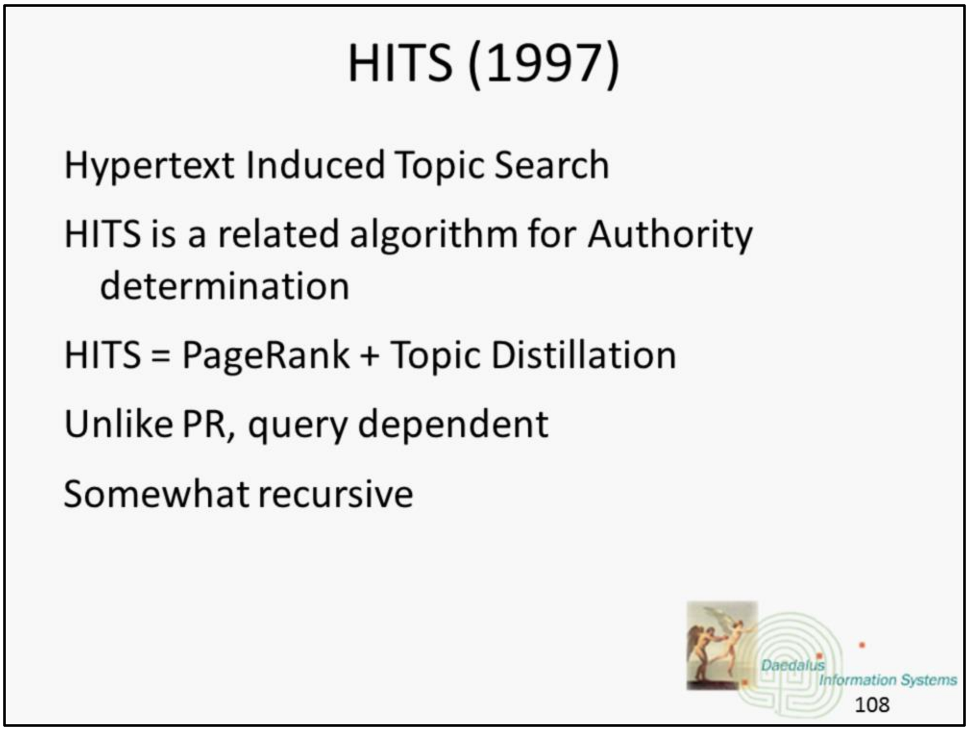
Uses single value decomposition to determine relationships between terms and concepts

- In linear algebra, the singular value decomposition (SVD) is a factorization of a real or complex matrix

Appropriate to Boolean model
First application in 1980's at Bell Laboratories
Sandboxed on Google in mid2000's

In example:
There are no listings for "apple" the fruit – the search engine has been trained to associate "apple" singular with the computer company
The #2 result is Microsoft – the search engine has been trained to associate the term Apple with Microsoft
The #4 result is for Mac cosmetics because the search engine does not know the difference between Mac computers and Mac mascara.


Convey UX
view profile 6 A 1.58% 58
cory doctorow 5 H,A,IA 1.32% 60
session videos 5 A,LI 1.32% 92
michael gough 4
amber case 3
joe welinske 3 A
conveyux conference 3
video posted 3 H,A
session preview
user experien
rand fishkin
peter merholz
quantifying delig
brad weaver
mary haggard
Welcome Marianne
Keyword Repeats Wizard Density Prominence
posted brad weaver 2 H,A 0.79% 36
is a great 2 0.79% 77
view all speakers 2 A 0.79% 60
weaver on cinematic 2 H,A 0.79% 36
0.79% 40
0.79% 36
0.79% 8
0.79% 30
Keyword Repeats Wizard Density Prominence
video posted brad weaver 2 H,A 1.05% 36
weaver on cinematic ux 2 H,A 1.05% 36
brad weaver on cinematic 2 H,A 1.05% 36
Phrases 5
Keyword Repeats Wizard Density Prominence
posted brad weaver on cinematic 2 H,A 1.32% 36
brad weaver on cinematic ux 2 H,A 1.32% 36
Daedalus
Information Systems
107



Jon Kleinberg – CS Professor at Cornell

Simultaneous development to PR (HITS is query dependent). PageRank its sibling using another facet of academia: citation chaining

Calculates: root set, outlinks and inlinks, results in focused neighborhoods (sub graphs) , calculates authority weight (Distillation of broad search topics with authority and Hubs (pages that links to many authority pages on topic)

- Define topics
- Detect authorities
- Detect hubs

Web emerging with pages that were access points to other pages on particular topic

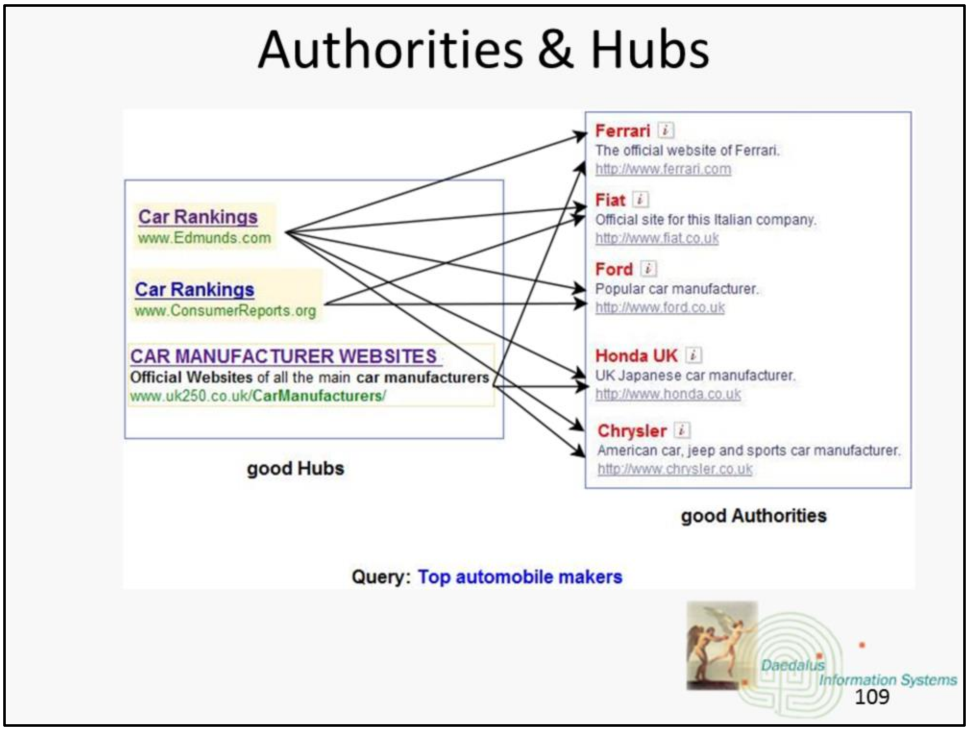
- Uses hubs and authorities to define a recursive relationship between web pages
- An authority is a page that many hubs link to
- A hub is a page that links to many authorities

Subset analysis of link graph to split out topics, then authority pages, then hub pages that consolidate links to authority pages on the topic

- Collect top ## of results (based on occurrences) = root set
- Construct a small link graph with pages pointing to pages in root set
- Collect set of pages that either link to authority pages in set or are included in hubs that link to pages in set

"HITS algorithm is in the same spirit as PageRank. They both make use of the link structure of the Web graph in order to decide the relevance of the pages. The difference is that unlike the PageRank algorithm, HITS only operates on a small subgraph (the seed SQ) from the web graph. This subgraph is query dependent; whenever we search with a different query phrase, the seed changes as well. HITS ranks the seed nodes according to their authority and hub weights. The highest ranking pages are displayed to the user by the query engine."
http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

Recursive in that good authorities receive links from good hubs link out to good authorities



http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

Authority Score is based on Web link structure (using representational sample set of WWW pages to distill structure)
Hubs: WW form of biblio-metrics – collection of thematically related authority pages

Hilltop Algorithm (2001)

Topic segmentation algorithm = query dependent

Introduces concept of non-affiliated “expert documents” to HITS

Quality of links more important than quantity of links

Segmentation of corpus into broad topics

Selection of authority sources within these topic areas

Daedalus Information Systems

“Our approach is based on the same assumptions as the other connectivity algorithms, namely that the number and quality of the sources referring to a page are a good measure of the page's quality. The key difference consists in the fact that we are only considering "expert" sources - pages that have been created with the specific purpose of directing people towards resources.”
http://ftp.cs.toronto.edu/pub/reports/csrg/405/hilltop.html

Components

Quality of links more important than quantity of links

Only links from Segmentation of corpus into broad topics

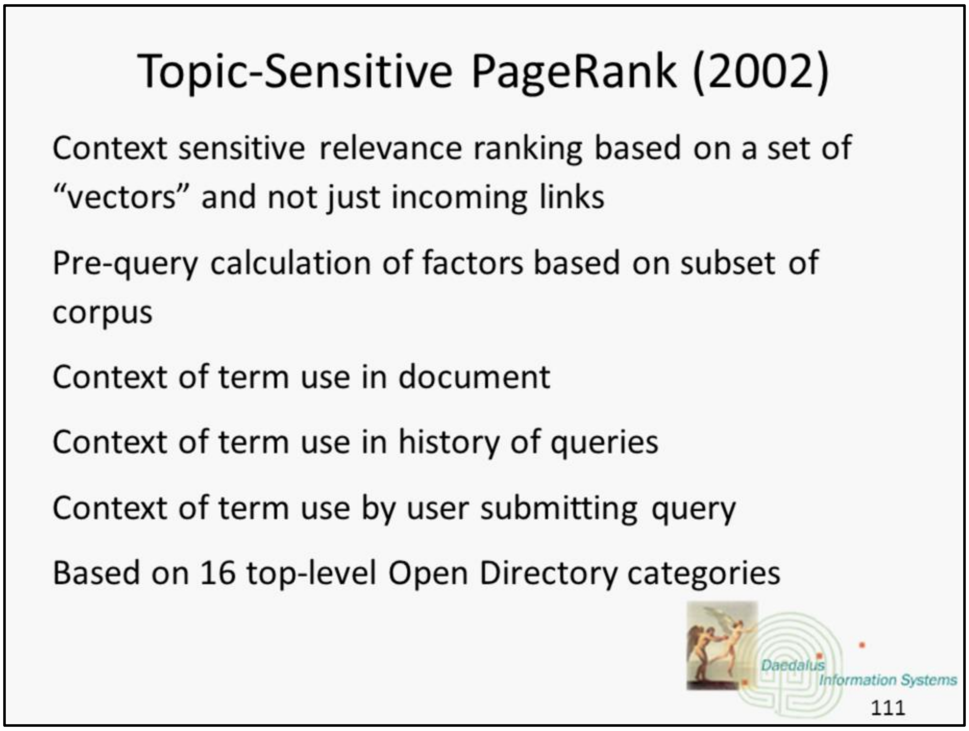
Subset that is then extrapolated to Web as a whole

Selection of authority sources within these topic areas with authorities have lots of unaffiliated expert document on the same subject pointing to them

Hubs are navigation pages that point to several authority resources on a certain topic

HITS is a related algorithm for Authority determination

The beauty of Hilltop is that unlike PageRank, it is query-specific and reinforces the relationship between the authority and the user’s query. You don’t have to be big or have a thousand links from auto parts sites to be an “authority.”



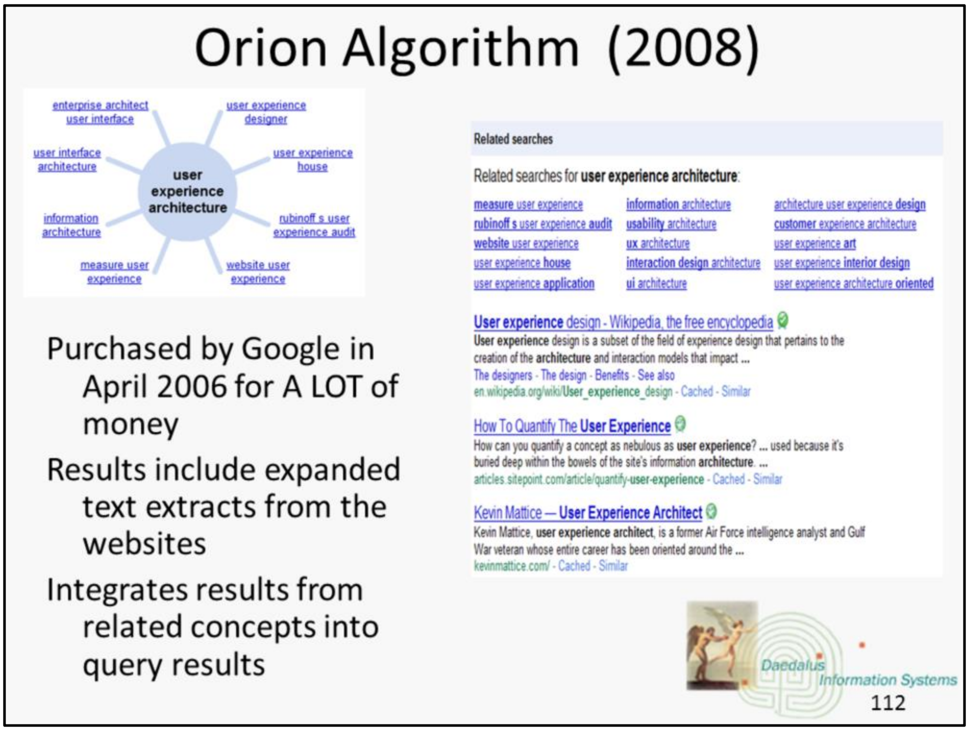
PR: single vector computed to capture relative importance + set of vectors based on document topics = TSPR

Instead of the single PR vector TSPR uses a set of ranking vectors: Pre-query selection of topics + at-query comparison of the similarity of query to topics
Topic-Sensitive PageRank computes PR based on a set of representational topic vectors
Augments PR with content analysis

At query time, calculates similarity of query to the topics
Computes PR based on a set of representational topics [augments PR with content analysis], rank vectors determine importance

Topic specific vectors derived from the 16 top-level categories from the Open Source directory (pre-query processing done offline)

Top-level categories (from Wikipedia): *Adult*, *Arts*, *Business*, *Computers*, *Games*, *Health*, *Home*, *News*, *Recreation*, *Reference*, *Regional*, *Science*, *Shopping*, *Society*, *Sports* and "World".



http://googleblog.blogspot.com/2009/03/two-new-improvements-to-google-results.html

Developed by a computer science student, this algorithm was the subject of an intense bidding war between Google and Microsoft that Google one. The student, Ori Allon, went to work for Google in April 2006 and has not been heard from since. There is no contemporary information on the algorithm or it's developer.
Allon left Google in 2010 to found Julpan, social network analysis tool acquired by Twitter (analyzes social activity for what is being shared)

Relational content modeling done by machines-usually contextualized next steps.

http://searchengineland.com/google-implements-orion-technology-improving-search-refinements-adds-longer-snippets-17038
Orion finds pages where the content is about a topic strongly related to the key word. It then returns a section of the page, and lists other topics related to the key word so the user can pick the most relevant.

"The results to the query are displayed immediately in the form of expanded text extracts, giving you the relevant information without having to go the website–although you still have that option if you wish," said Israeli-born Allon, who completed a Bachelor and Masters degree at Monash University in Melbourne before moving to UNSW for his PhD. By displaying results to other associated key words directly related to your search topic, you gain additional pertinent information that you might not have originally conceived, thus offering an expert search without having an expert's knowledge.

Hummingbird: Entity detection

Comparison of search query to general population search behavior around query terms

Revises query and submits both to search index

- Confidence score
- Relationship threshold
- Adjacent context
- Floating context
- Results a consolidation of both queries

Daedalus Information Systems

113

Entity=anything that can be tagged as being associated with certain documents, e.g. Store, news source, product models, authors, artists, people, places thing

The entity processing unit looks at “candidate strings and compares to query log to extract: most clicked entity, most time spent by user)

Referring queries data taken away

User Behavior information: user profile, access to documents seen as related to original document, amount of time on domain associated with one or more entities, whole or partial conversions that took place

AI Content Components

Traditional IR (tf*idf)
Link analysis
Page Layout
Query type
Uniqueness
Authoritative
Freshness
Well Written

Daedalus Information Systems
114

Link analysis (matches context of query)
Page layout (content above fold, not to many ads/images)
Authority (site and author)
Query Type: Informational queries account for 63% of studied with transactional at 22% and navigational at 15%)
Well written: Fleishman Kincaid scale, grammar and spelling

Content Creation | Curation

Dense, subject-specific content is what is indexed

Opening paragraphs most important for subject determination

Newspaper model

Trim the garden: get rid of the old or unpopular items

Daedalus Information Systems
115

Home page
The more content, the stronger the representation in the search engine index

- More content = Authority = aboutness
- People will scroll - If they don't scroll, they will print it out
- Visible text on a page is what counts
- Spiders cannot “see” = cannot read text images

Consistency in terminology and emphasis in topicality on page is good however search engines are sensitive to over optimization

Headings are a user’s and the spider’s friend. Extra credit for having them and for having topic terms in there

Search engines are:
Semantic (LSI)
Judgmental
Evaluate content based on non-content criteria (bounce rate, click through, conversion)

DESIGN FOR AI

Daedalus Information Systems

116

https://www.oreilly.com/learning/machine-learning-for-designers

Human-centered design has expanded from the design of objects (industrial design) to the design of experiences (encompassing interaction design, visual design and the design of spaces). The next step will be the design of system behavior; the design of algorithms that determine the behavior of automated intelligent systems

Harry West
CEO, Frog Design

Daedalus Information Systems

117

Machines As Users Are Different

Logic: exacting, context independent, conditional logic

Development: uses explicit rules to define possible behaviors

- Heuristics
- Intuition derived from huge data sets

Daedalus Information Systems

118

Reasoning

Deductive	**Inductive**
Theory	Theory
Hypothesis	Hypothesis
Observation	Pattern
Confirmation	Observation

Daedalus Information Systems

119

Design Thinking for Data Science 1

Reach out to the development staff

Embrace design thinking

Transform "my idea" into "our idea" with early stage collaboration

Daedalus Information Systems

120

https://www.linkedin.com/pulse/design-thinking-data-science-george-roumeliotis
http://www.intuitlabs.com/page/2/?s=design+for+delight

Design Thinking for Data Science 2

Customer Empathy Stage

- Understand the problem solving
- Define the solution
- Map the environment (customer journey)
- Define the characteristics of a good solution (heuristics)

Outputs

- Personas (use cases)
- Problem statements
- Environment description (include systems and processes)
- Benchmark success (quantitative, qualitative)

Daedalus Information Systems

121

https://www.linkedin.com/pulse/design-thinking-data-science-george-roumeliotis
http://www.intuitlabs.com/page/2/?s=design+for+delight

Design Thinking for Data Science 3

Go Broad, Go Deep Stage

Brainstorm solution ideas across silos

Diversify contributors

Post all artifacts and review as a group

Organize ideas into themes

Include “leap of faith” assumptions

Take the best and formulate a solution hypothesis

Daedalus Information Systems

122

https://www.linkedin.com/pulse/design-thinking-data-science-george-roumeliotis
http://www.intuitlabs.com/page/2/?s=design+for+delight

Design Thinking for Data Science 4

Rapid experimentation with Customers

Paper prototyping, sketches, storyboard

Build stable testing methodology into plan

Start small (project | testing) to achieve collective wins

Daedalus Information Systems 123

Algorithm-Based Design 1

Designer as art director, algorithm as apprentice

Determine “well designed” site for learning model

Create mood board for algorithm to deconstruct

Use algorithm for simple tasks

- Color match up
- Image assembly (movie poster app)
- Styling videos
- Extract usage patters from data sets

Daedalus Information Systems

124

https://www.smashingmagazine.com/2017/01/algorithm-driven-design-how-artificial-intelligence-changing-design/

Algorithm-Based Design 2

Designer and Developer define the logic to consider content, context and user data

AEM (behavior targeted UI)

Brightedge DataMind

Vox Media Homepage Generator

Daedalus Information Systems

125

https://www.smashingmagazine.com/2017/01/algorithm-driven-design-how-artificial-intelligence-changing-design/

Vox generator: algorithm pulls the page layout from pattern library based on user profile, display dictated by # of words, paragraphs, images, inserts

Generative Design

AKA Mutative Design, Parametric Design

Designer defines rules for algorithm

Algorithm generates variations using the predefined rules

Algorithm filters the results based on design quality and requirements

Designer chooses the best variants and polishes as needed

System runs A|B tests for variant(s)

Test results used to choose most effective design

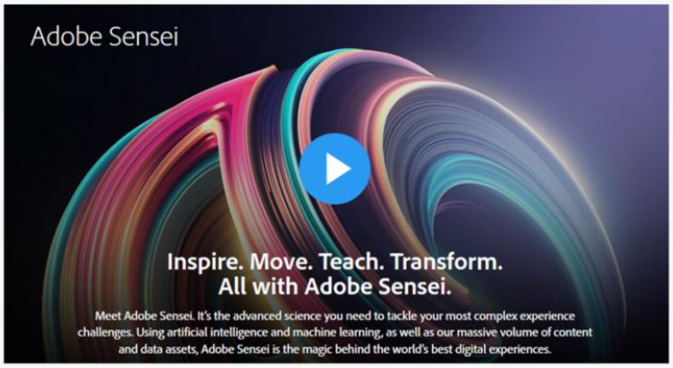
Daedalus Information Systems

126

Idealized collaboration
Iterative
Zora Hadid Architects

Project Phoebe: mutative design
https://medium.com/project-phoebe/meet-project-phoebe-a-moonshot-concept-for-mutative-design-88d997f7ff14#.umwh0ksh3

Smart Tools & Platforms



Sensei uses machine learning and artificial intelligence

Semantic image segmentation

Font recognition

Intelligent audience segmentation

Daedalus Information Systems

127

http://www.adobe.com/sensei.html

Machine Learning Design Process

Define learning problem

- Inputs
- Outputs
- Training Data

Generate good data

- Completeness
- Accurate
- Consistent
- Timely

Sketch out user and data flow (decision trees)

Test assumptions against prototype

Start with simple mechanism and move to complex

Daedalus Information Systems

128

Completeness: indicative of a range of possible behaviors
Accurate: true to real world behavior
Consistent: various data points within the set are not contradictory
Timely: relevant to current system state

IF NOT US, WHO?

Daedalus Information Systems

Why Not Google?

Without a person at (or near) the helm who thoroughly understands the principles and elements of Design, a company eventually runs out of reasons for design decisions... When a company is filled with engineers, it turns to engineering to solve problems. Reduce each decision to a simple logic problem. Remove all subjectivity and just look at the data. Data in your favor?... And that data eventually becomes a crutch for every decision, paralyzing the company and preventing it from making any daring design decisions.

Daedalus Information Systems

130

http://stopdesign.com/archive/2009/03/20/goodbye-google.html

Douglas Bowman
on leaving as head of Google Visual Design (2009)

Tested 41 shades of blue

Prediction Is Not Infallible

AI algorithms rely on past behavior to predict future behavior

Programming and test set must define "normal" for the system to detect "abnormal"

Daedalus Information Systems

131

Cannot predict what has not already occurred
Flash Crash 2009
Taleb's Black Swans

Prediction Drawbacks

Cannot predict what has not already occurred

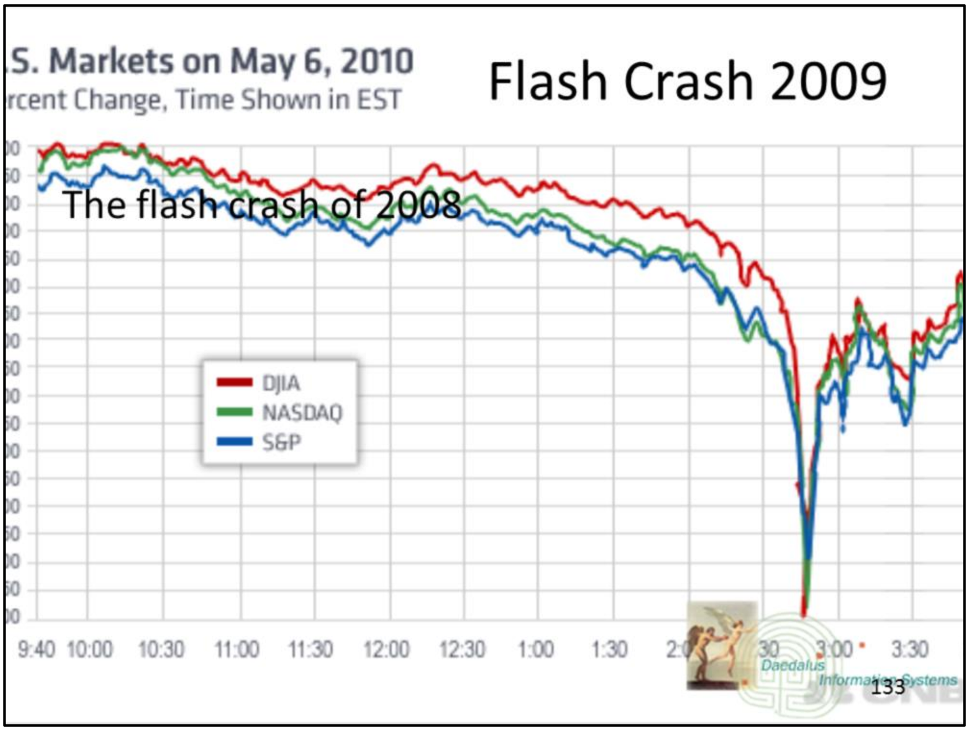
- Taleb's black swans
- Flash Crash of 2009

Past behavior prediction ignores present environment and emotional influences

Must define normal to program for abnormal detection

Daedalus Information Systems

132

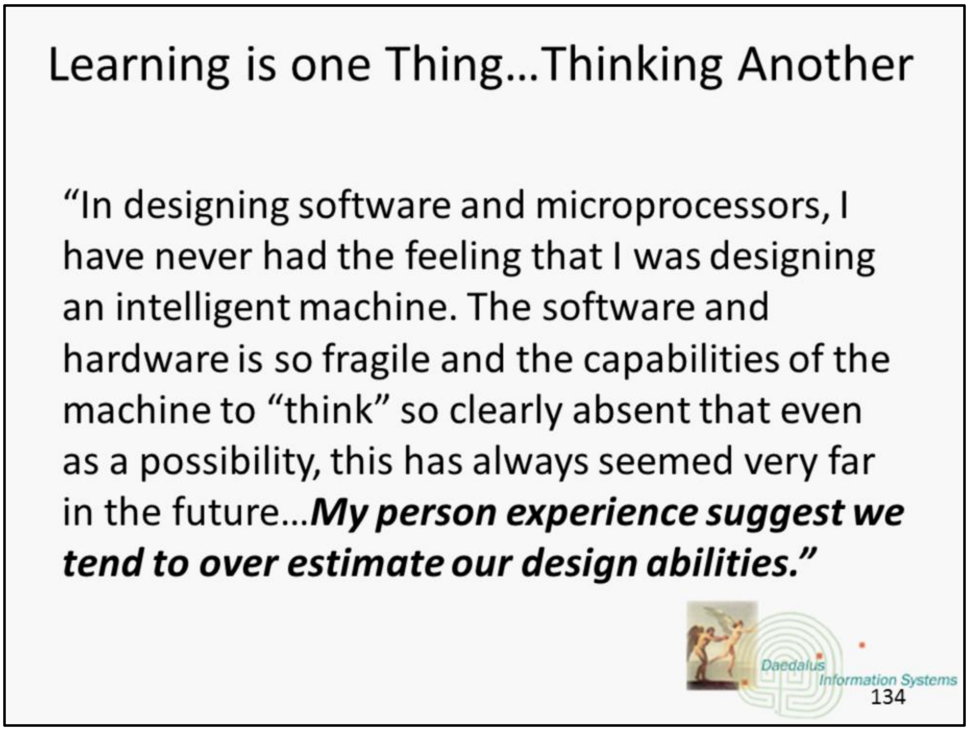


ttp://www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world.html

Kevin Slavin (Assistant Professor, MIT Media Lab): How algorithms share our world and the Flash Crash of 2010 during which the stock market dropped 1000 points in moments and took with it billions of dollars due to conflicting algorithmic trades.

Algorithms were used to break up big things (huge stock trades) into many little things (individual trades) and then recombine - Flash crash of 2:45 where 9% of entire stock market disappeared - went away - no one can explain because they did not do anything...algorithms did it

Epagogix: story algorithms to predict movie success before writing (Netflix Pragmatic Chaos algorithm influences 60% of rentals)
Elevators where you push floor before getting in and that determines what elevator - inside there are no buttons (so cannot change your mind)



Bill Joy, cofounder Sun Microsystems, creator Java and Jini

Sometimes They Learn the Wrong Things

Bill Slawski @bill_slawski · 12s

Microsoft unveils a better-behaved chatbot after its last one became a NAZI

Microsoft unveils a better-behaved chatbot after its last one became ...

Tech giant takes another pop at the artificial intelligence game with the release of a politer(and slightly stupider) machine mind

135

You Are Not A Gadget

"Whenever a computer is imagined to be intelligent, what is really happening is ***that humans have abandoned aspects of the subject at hand in order to remove from consideration whatever the computer is blind to****."*

Daedalus Information Systems

136

Cyber-totalitarianism/Digital Maoism: the focus is on abstraction of the network more than the real people that use it. Lanier believes that emphasizing the crowd deemphasizes the individual. He sustained an online debate over this in 2006 (The Edge) where he referenced the "hive mind" or noosphere (collective brain).

Emphasizing the crowd means de-emphasizing the individual

Lanier goes on to observe that instances of intelligence in a machine are ambiguous and that we degrade our sense of personhood to make machine seem more intelligence or capable of learning.

Lanier’s concept of computationalism:

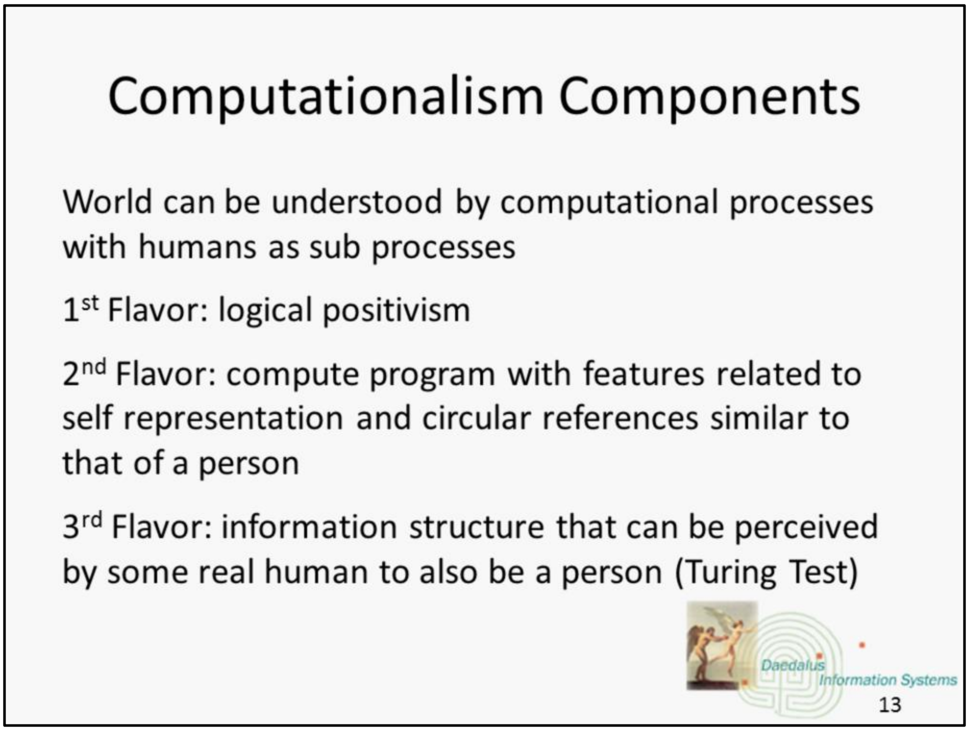
Logical Positivism: a sentence or other fragment that can be put into a computer file will mean something in a freestanding way that does not require invoking the subjectivity of a human reader. EG: a computer can figure out the “meaning” of a sentence if the instructions are correct and comprehensive

World can be understood by computational processes with humans as sub processes

1st level: logical positivism

2nd level: a compute program with features related to self representation and circular references similar to a person

3rd level: an information structure that can be perceived to be human by a human (Turing Test)



Meta knowledge becomes authority
Collective is good at parameters, bad at user experience
World can be understood by computational processes with humans as sub processes

Relies on Logical Positivism: a sentence or other fragment that can be put into a computer file will mean something in a freestanding way that does not require invoking the subjectivity of a human reader.
EG: a computer can figure out the “meaning” of a sentence if the instructions are correct and comprehensive

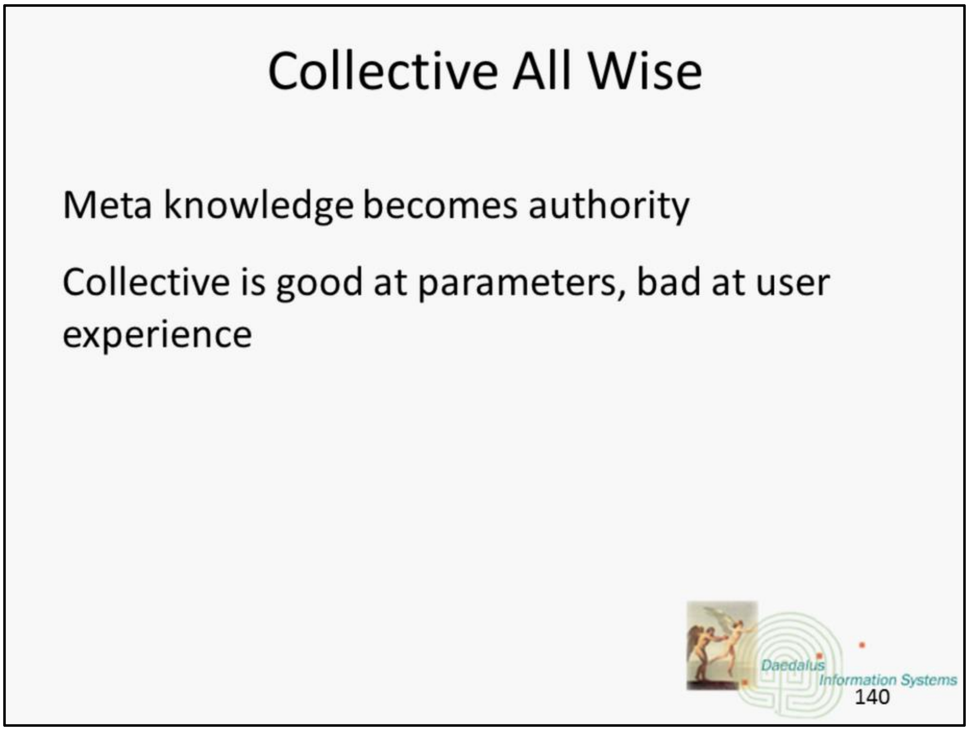
Logical Positivism

To have meaning, a given statement had to be connected to either empirical data or analytic truth.

Daedalus Information Systems

139

Only something empirically verifiable logically or empirically is meaningful. Facts trump all else.
Later thinkers distinguished between two classifications of verifiability: "strong" and "weak" verification, the former being something that is conclusively established by experience, the latter only being rendered probable by experience.



Lanier often refers to the “hive mind” with regard to online information
Those consolidation of bits becomes information (e.g. Wikipedia over true authority sites because edited by the collective)

Solutionism

Every problem has a definite an computable solution (borrowed from architecture and urban planning)

False notion that Internet is a coherent and stable influence in our lives

Sometimes right algorithms can lead to wrong answers

Grasping easy digital solutions often ignores complex causes behind

Questions Silicon Valley assumption that quantifiable self is the truer self

Daedalus Information Systems

141

Morozov often referenced as an intellectual “hit man.” He wants us to:

- Resist oversimplification of techno-optimism and techno-pessimism to assess each case of technology intervention on its own merits
- Not to fixate on what technology can do without inquiring if it is worth doing (what problem are they solving)
- Do not lose sight of the benefits of subjectivity
- Promote new way of thinking that is technologically literate” – attentive to details, mindful of legal and economic circumstances, historically informed to question appropriateness in each and every situation

Quantification of the self is a crime – it forecloses possibilities and narrows vision (numbers replace other possible interpretations: “quantitative and linear casual explanations that have little respect for the complexities of the actual human world”)
Quantifiable information is the low hanging fruit
Quantifiable Self Movement: “fundamental assumption is that numbers can reveal a core and stable self if only we get the technology right.”

“Recasting all complex social situations either as neat problems with definite, computable solutions or as transparent and self-evident processes that can be easily optimized—if only the right algorithms are in place!—this quest is likely to have unexpected consequences that could eventually cause more damage than the problems they seek to address.”

Solutionism

Sometimes right algorithms can lead to wrong answers

Grasping easy digital solutions often ignores complex causes behind

Silicon Valley assumption is of a quantifiable self that is the truer self

Daedalus Information Systems

142

Solutionism
Presented in To Save Everything Click: Evgeny Morozov

Solutionism: there is an app for everything
Morozov sees the dot com elite as rewriting the code of social contract largely without public awareness let alone consent

“recasts all of the complex social situations either as neat problems with definite, computable solutions or as transparent and self-evident processes that can be easily optimized if only the right algorithms are in place.”

Key Takeaways

Broaden scope of awareness

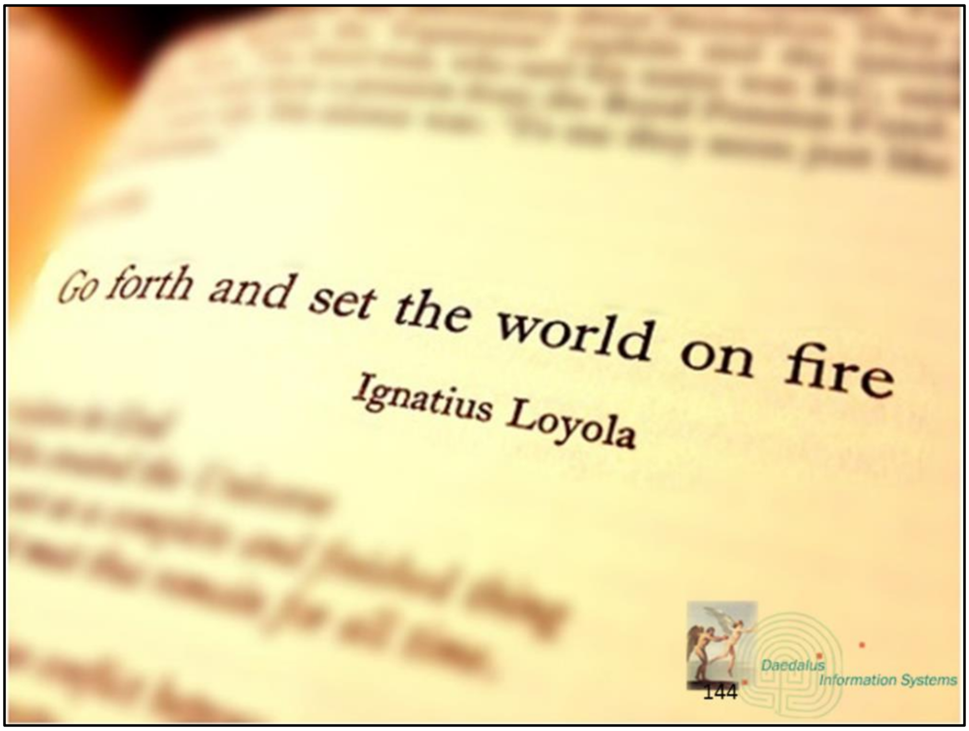
Understand the landscape and influences

Embrace new tools and methodologies

We are the representatives of the qualitative self that is truer than the quantitative self represented by AI

Daedalus Information Systems

143


Go forth and set the world on fire
Ignatius Loyola
Daedalus
Information Systems
144

Thank You

Marianne Sweeny
Principal
Daedalus Information Systems
sweeny48@uw.edu
@msweeny