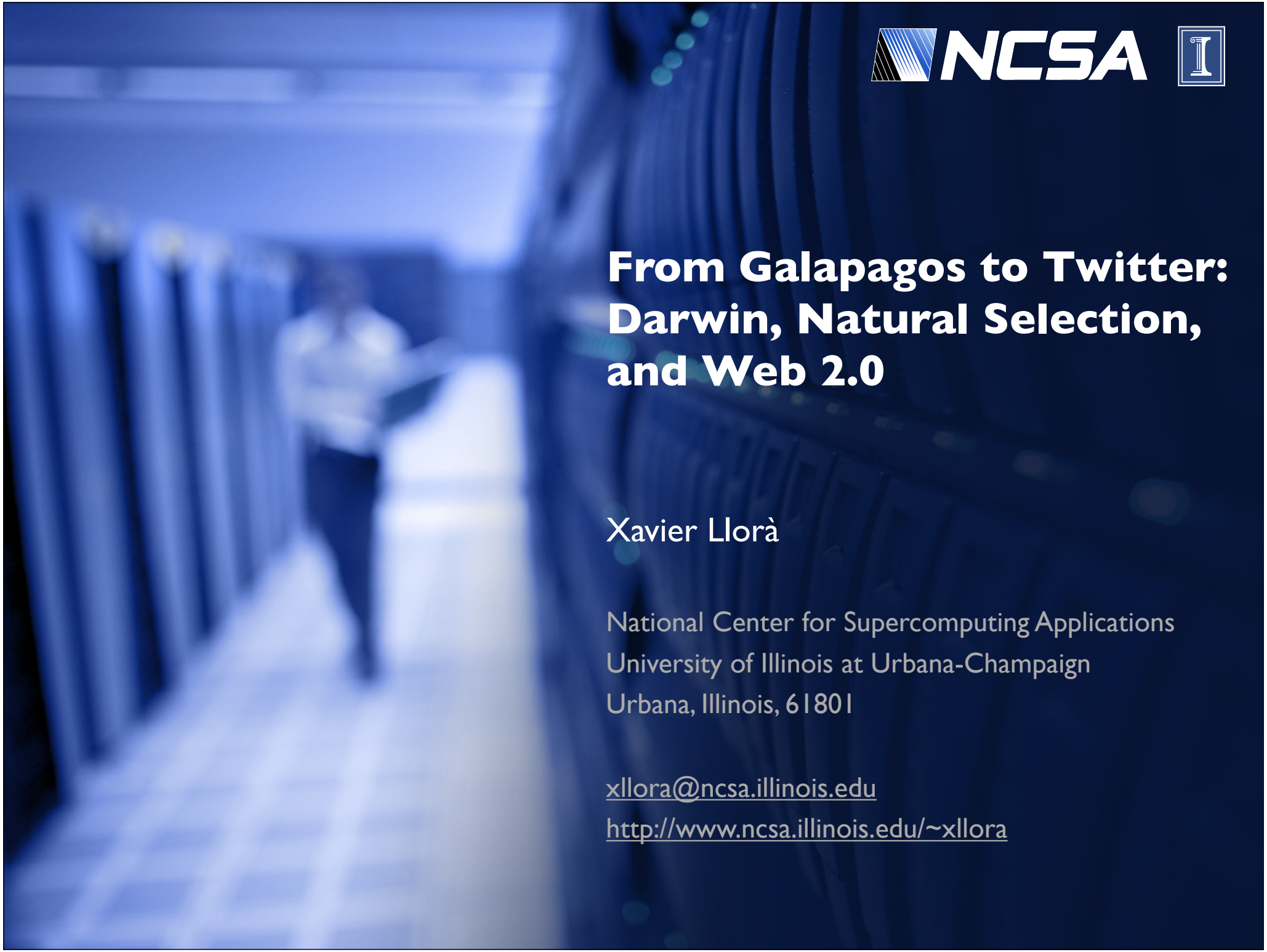


From Galapagos to Twitter: Darwin, Natural Selection, and Web 2.0

Xavier Llorà

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Urbana, Illinois, 61801

xllora@ncsa.illinois.edu
http://www.ncsa.illinois.edu/~xllora

What is this talk about?

- Simple answer:
  Darwin's natural selection concept and information technologies

- Natural selection?

  Natural selection is the process by which heritable traits that make it more likely for an organism to survive and successfully reproduce become more common in a population over successive generations.

- Remember, back on Darwin's time genetics were unknown

Darwin on HMS Beagle

A quick Darwin background check

- In Edinburgh Charles Darwin
  - Learnt standard geology (Robert Jameson)
  - Exposed to British catastrophism of 1820s
  - Run into Lamarckian ideas with Robert Grant
- Cambridge provide him a chance to
  - Under John Henslow, explored boundaries between varieties and species
  - Read and accepted traditional theological explanations of the world
- Evolution existed *before* and *after* Darwin
- HMS Beagle
  - Left Plymouth 2pm on December 27, 1831
  - Returned October 2, 1836

HMS Beagle

http://www.solarnavigator.net/history/hms_beagle.htm

And Galapagos happened

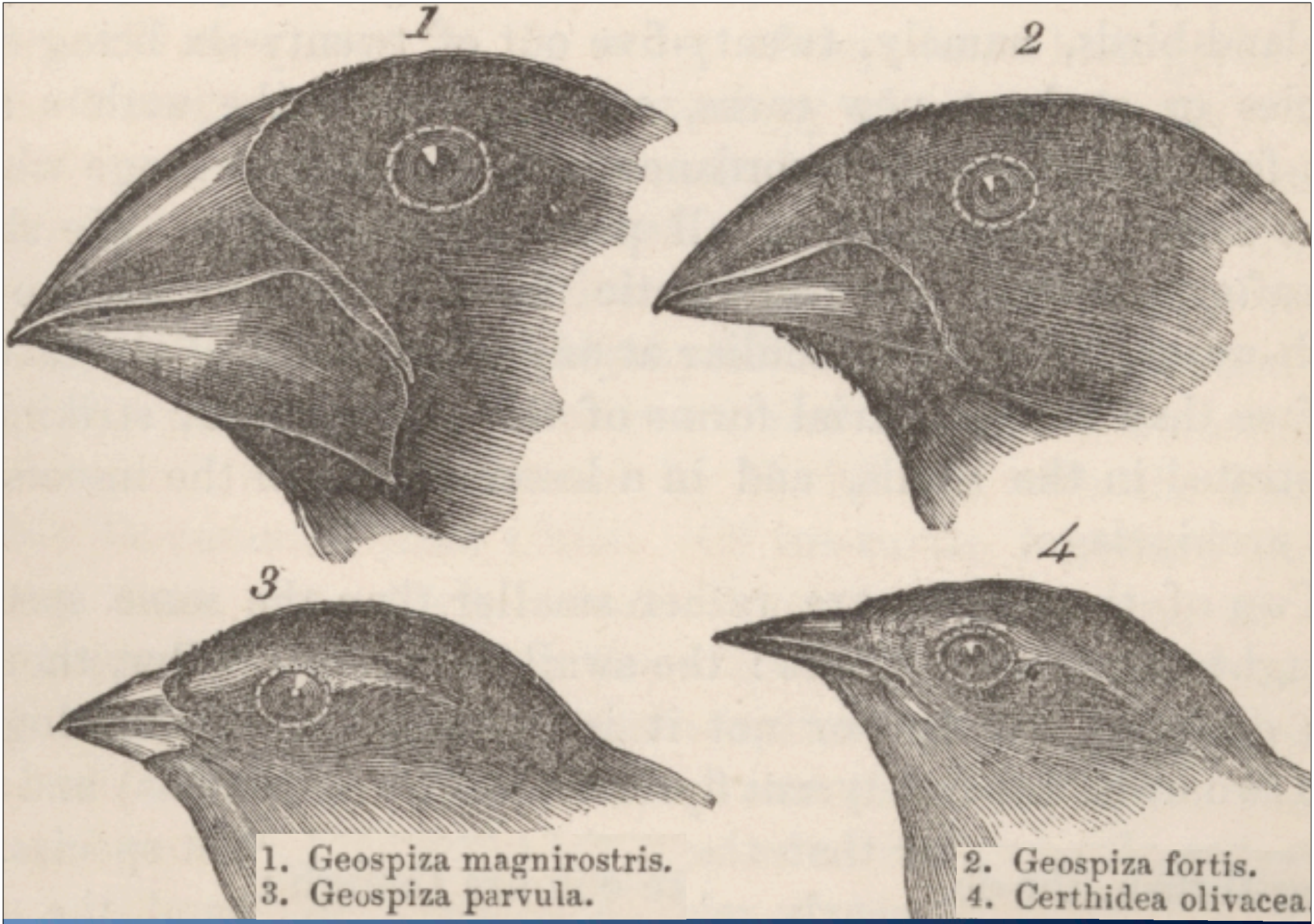
Massive diversity

http://www.flickr.com/photos/ole/1040941928/

NCSA

A key realization

- Knowledge on the boundaries of variation/species
- The facts
  - Large number of islands
  - Islands are spatially separated
  - Wide diversities of species (different sources of food)
  - Very close species (beyond variation)
- Start accumulating descriptions of specimens


1
2
3
4
1. Geospiza magnirostris.
3. Geospiza parvula.
2. Geospiza fortis.
4. Certhidea olivacea.

The basic building blocks

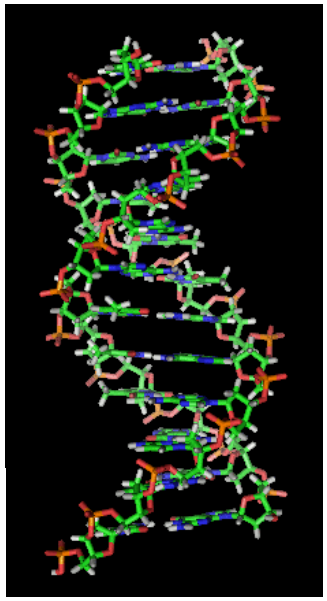
- Large variation of features
- Highly correlated to food sources
- The basic sketch of a crazy thought:
  - Assume you start with a diverse population of As
  - Split As and place them in different environments
  - Individual a_i features may or may not help live in the environment
  - If individual a_i features help in the environment a_i as a result has increased chances to reproduce
- The underlying distilled idea of natural selection
  - Variation will arise in a population
  - Selection by environment adaptation
- The uncanny idea: Variation can be random

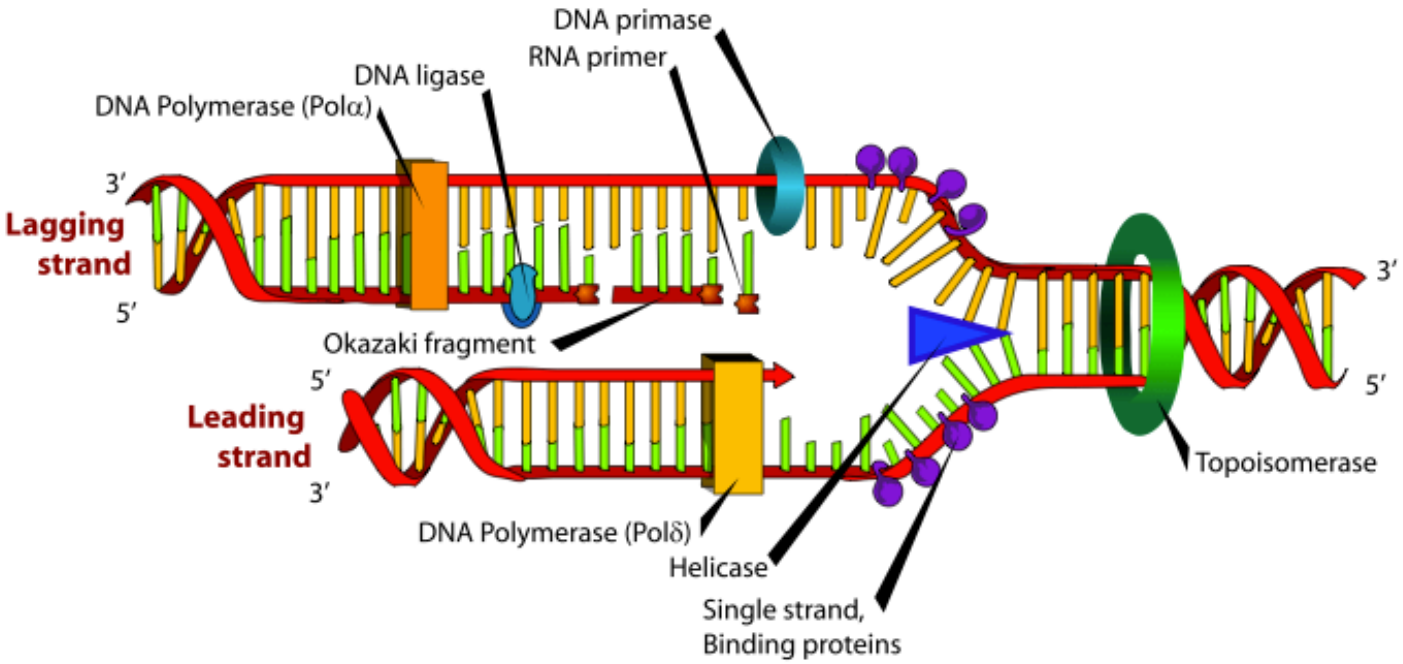
A simple abstraction

[Fast forward] ... and genomics arise

DNA and basics of reproduction

- Watson & Crick (1953)
- Four bases:
  Adenine (A), cytosine (C), guanine (G), thymine (T)
- Groups in to base pairs: A-T, C-G
- Reproduction of DNA (*crossover* and *mutation*)





[Fast forward] ... and computers happen

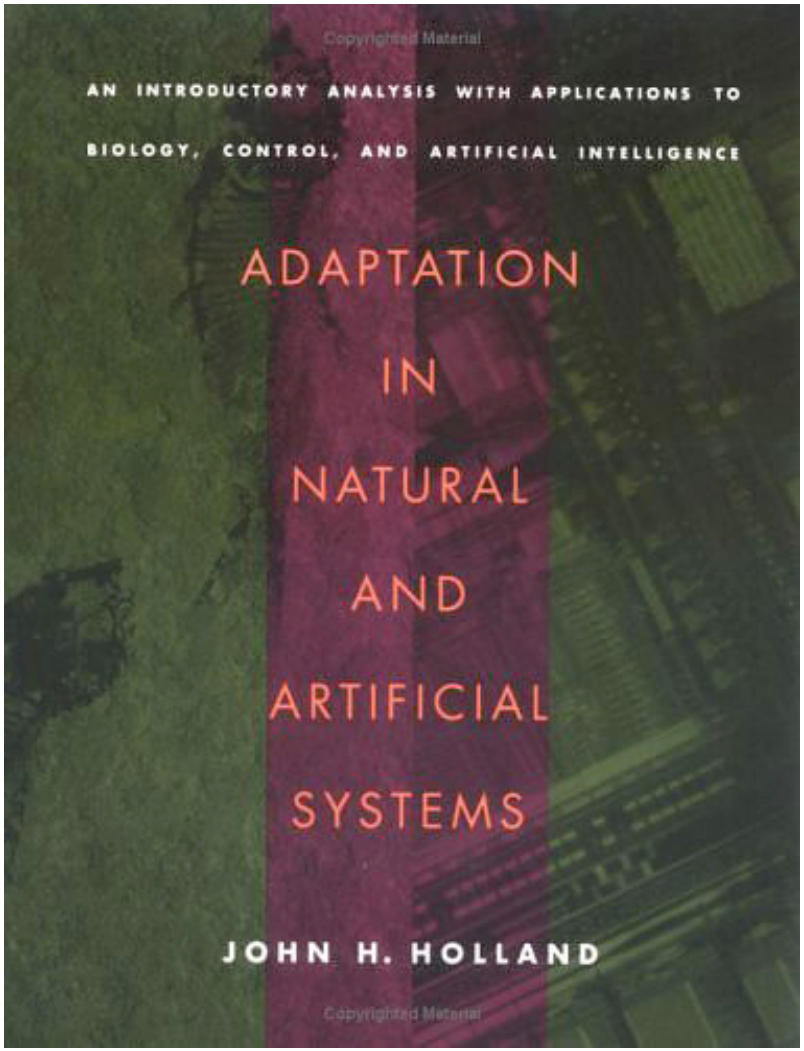
A quick stop in the late 60s

- Computers were the next cool thing
- Genetics were on the rise
- A professor at University of Michigan
- Natural evolution
  - Diverse population of individual
  - Individual a_i features may or may not help live in the environment
  - If individual a_i features help in the environment a_i as a result has increased chances to reproduce
- Artificial evolution?
  - Model of such evolutionary process
  - Use it to solve some real problem

Prof. John H. Holland

What is a genetic algorithm?

- ANAS published in 1975
- Chromosomes encode a solution to a problem
- Evaluation of fitness based on the quality solving the problem
- The process
  - Survival of the fitness
  - Mutation
  - Crossover
  - Repeat until good enough

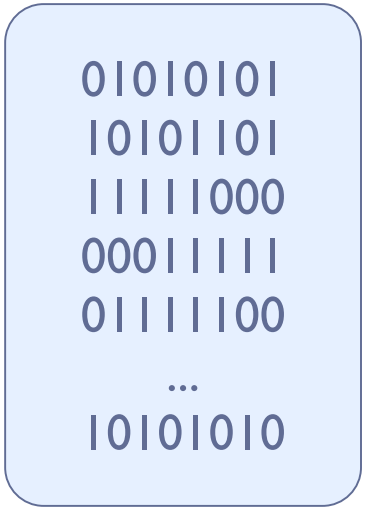


A simple example

- Imagine you have a box with an array of switches
- 0 off, 1 on
- Depending on the switch arrangement you get some payoff
- Payoff is equivalent to the number of switches that are on
- The goal: Maximize the payoff

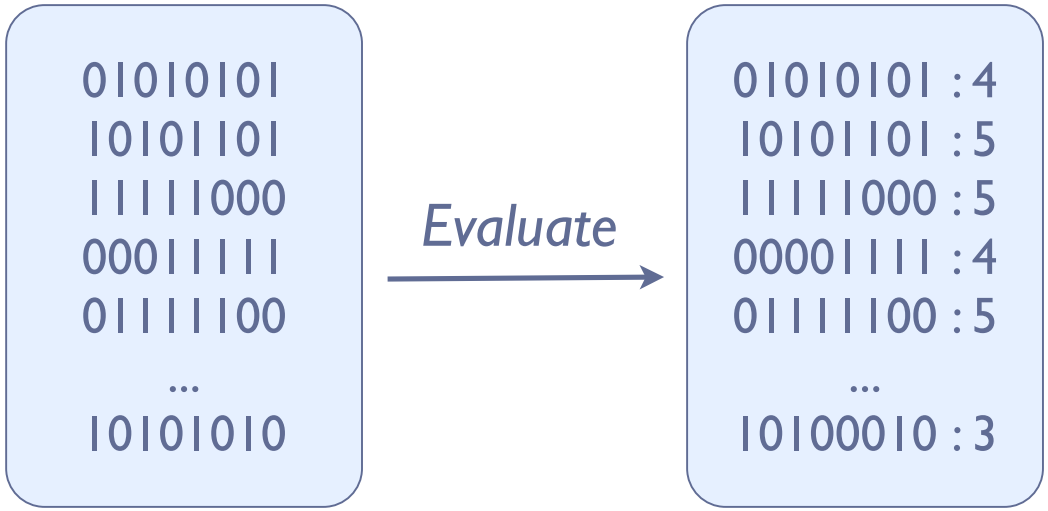
Encoding the problem

- Solutions are a string of 0s and 1s.
- We need a population



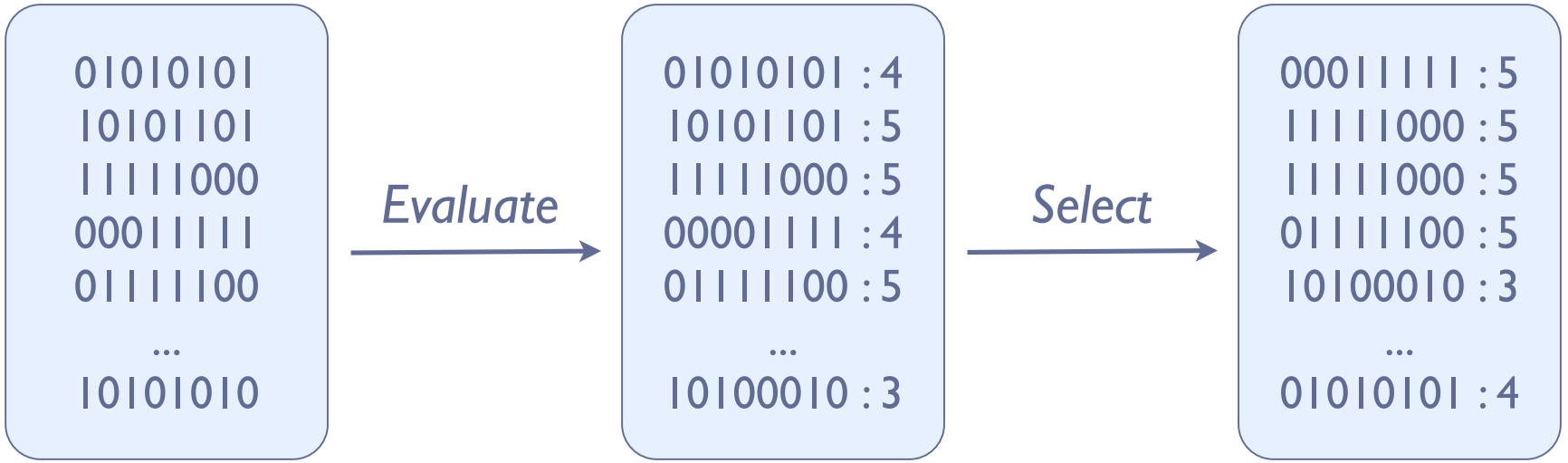
Evaluate solutions performance

- Solutions represent a solution to a particular environment
- Solutions get measured based on their performance



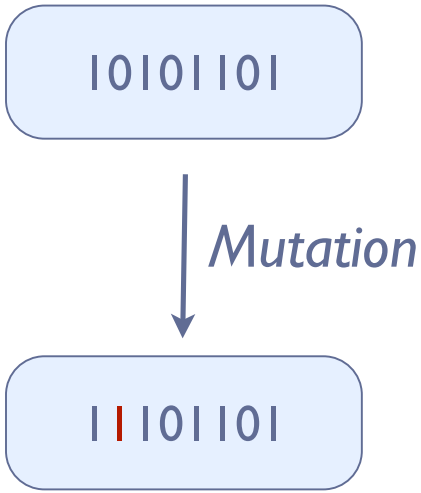
Selection based on performance

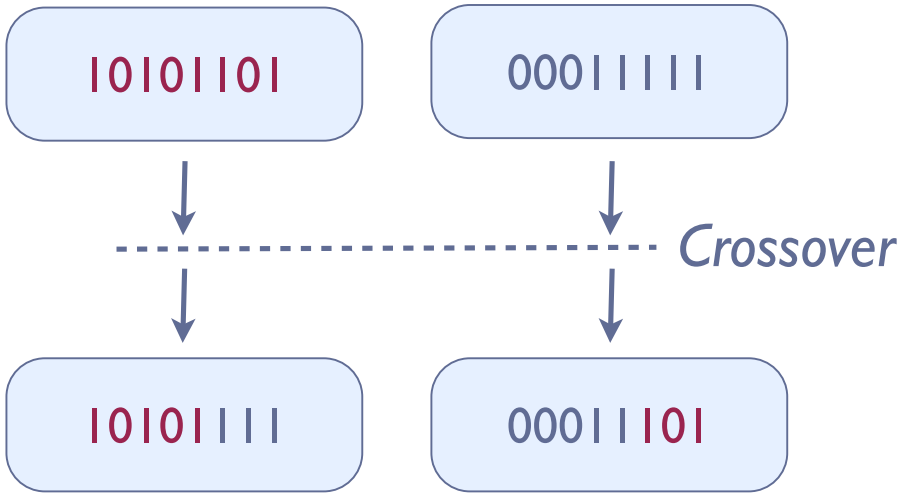
- Proportional to their performance
- Proportionate roulette wheel selection



Reproduction

- Mutation introduce random changes in the solution
- Crossover recombines parent solutions





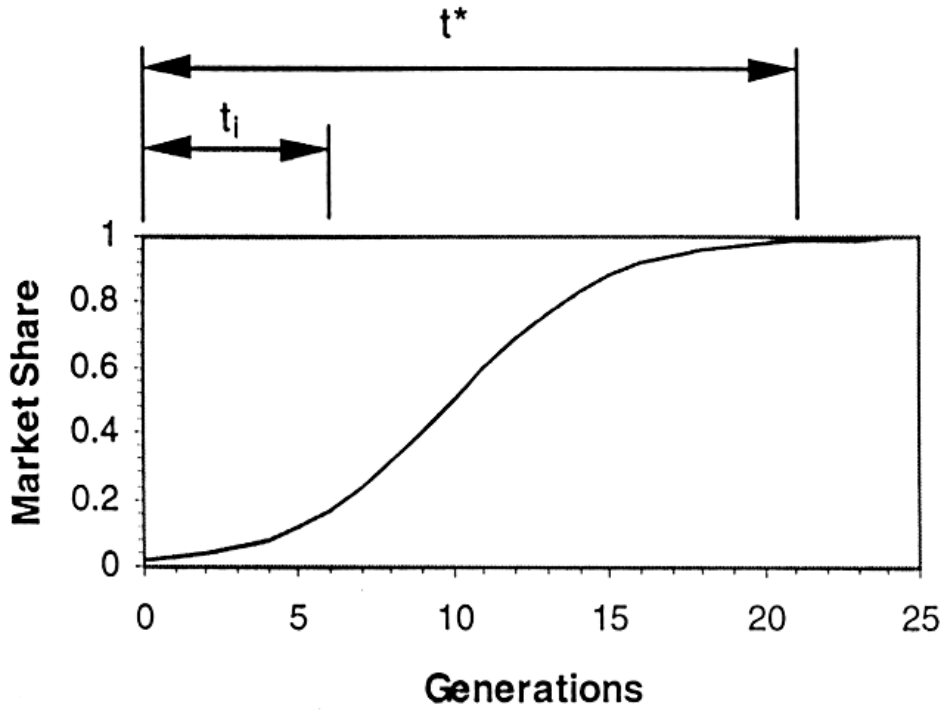
A nagging question

- In early 80s by David E. Goldberg
- How do uninteresting operators yield interesting behavior?
- 1983 proposed the innovation intuition:
  "Genetic algorithm power like that of human innovation"
- The underlying elements:
  - Selection + mutation as hillclimbing or *kaizen*
  - Selection + crossover = *innovation*
- In another words
  - Introducing small changes and keeping them if the improve performance
  - [...] Takes two to invent anything. One makes up combinations; the other one chooses [...] (*Paul Valéry*)

Remember the simple abstraction

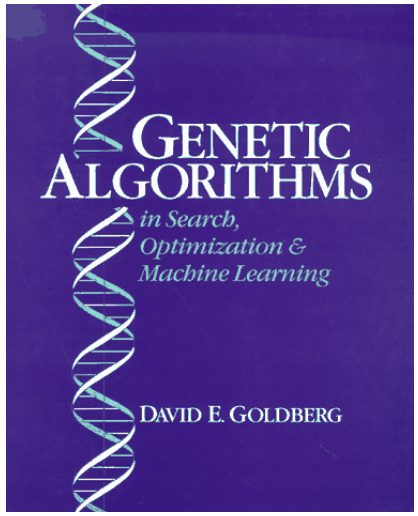
Goldberg's intuition

- Interesting behavior (emergence) is a balance
- Two components
  - Takeover time (t*)
  - Time to innovate (t_i)



Goldberg's intuition

- A design theory for scalable and efficient GAs
  - Understanding building blocks (concepts or ideas)
  - Ensure building block supply
  - Ensure building block growth
  - Control building block speed
  - Ensure good building block decisions
  - Ensure good building block mixing
  - Know building block challengers

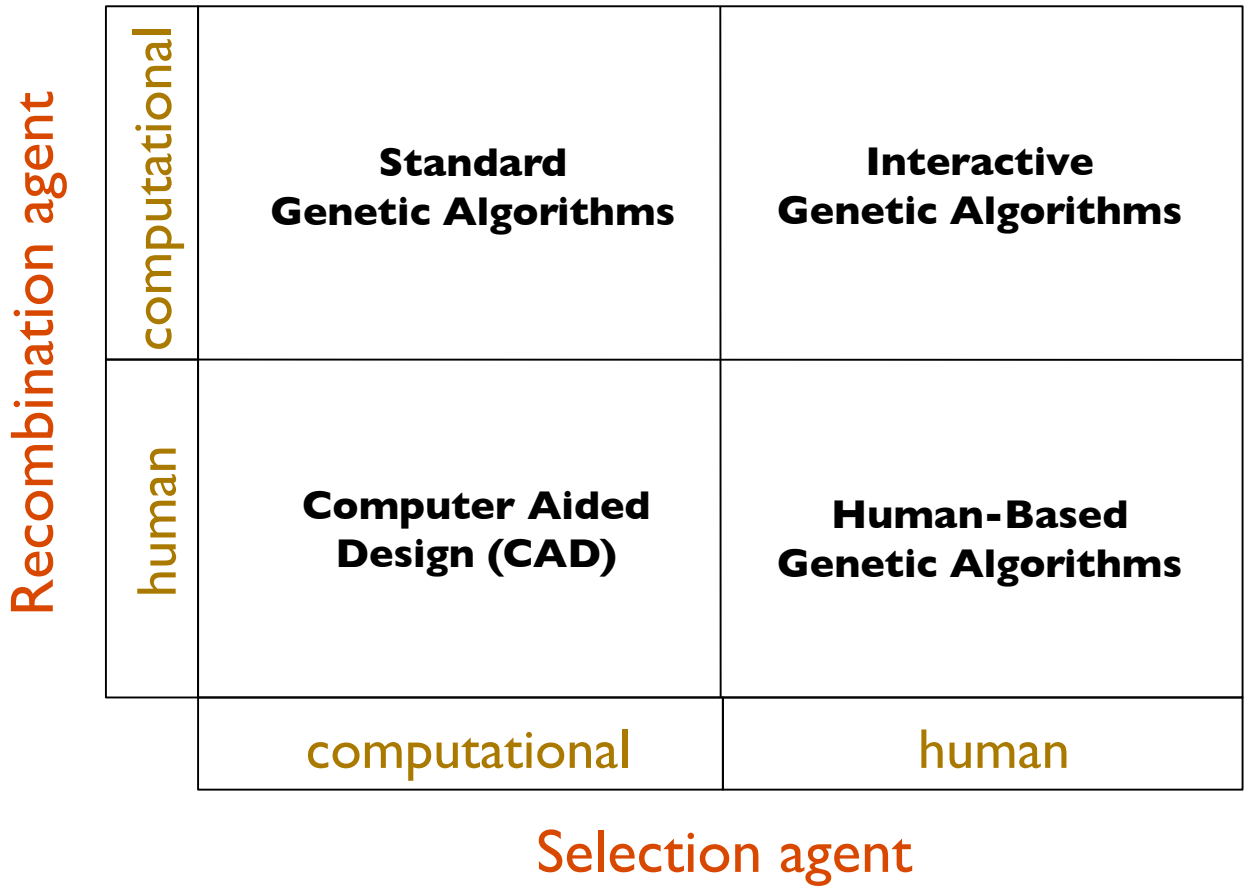


An interesting social insight

- GAs used for optimization, search, machine learning, ...
- But Valéry's quote also implied a social aspect to innovation
  - Solutions may be created in the vacuum, randomly, or carefully crafted
  - There is a social element of acceptance/rejection
- Paul Valéry (1871-1945) lived after Darwin published his ideas

[Fast forward] … The 2.0 information age

Adding social to genetic algorithms

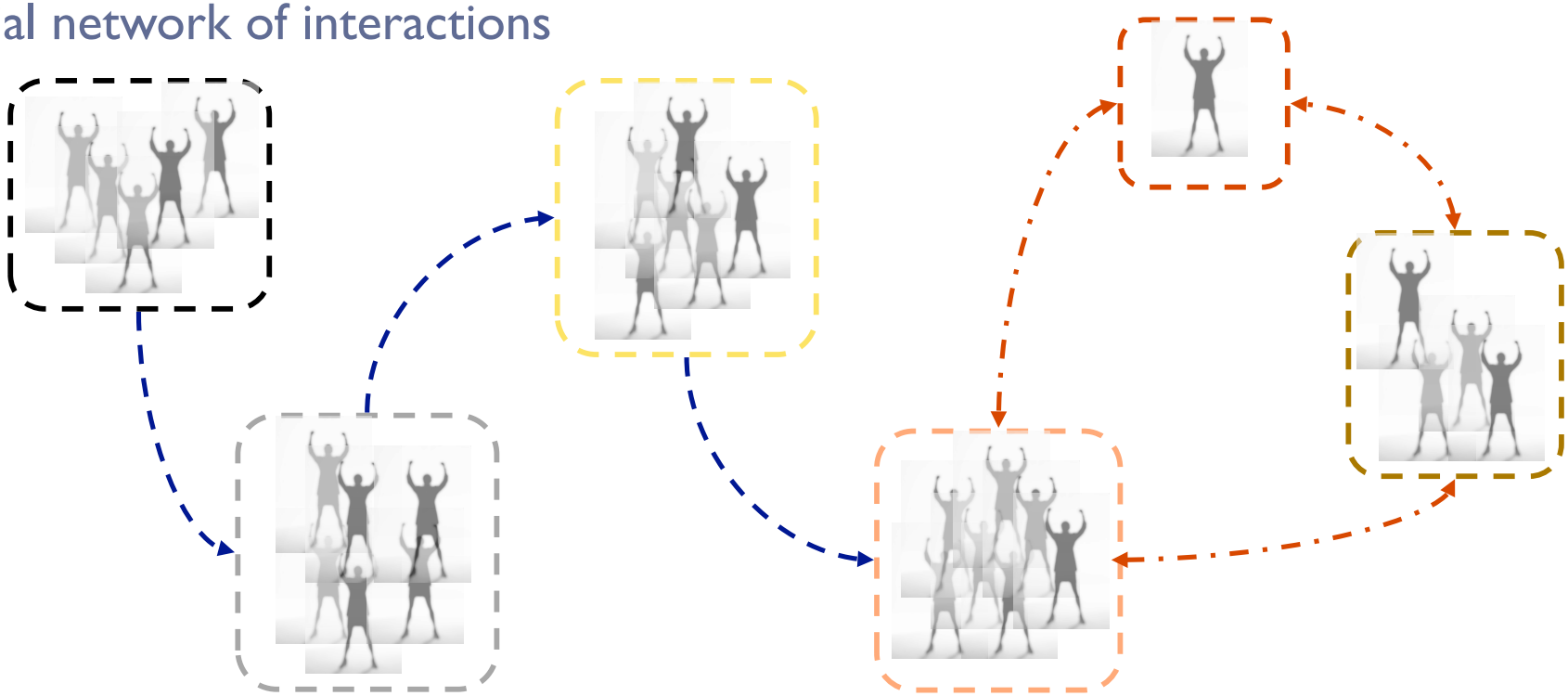


(Kosorukoff & Goldberg, 2002)

NCSA

Social aspects of genetic algorithms

- Population-based method
- Parallel genetic algorithms
  - Groups of working units
  - Communication between groups for a common goal
  - Models for collaborative work
- Social network of interactions



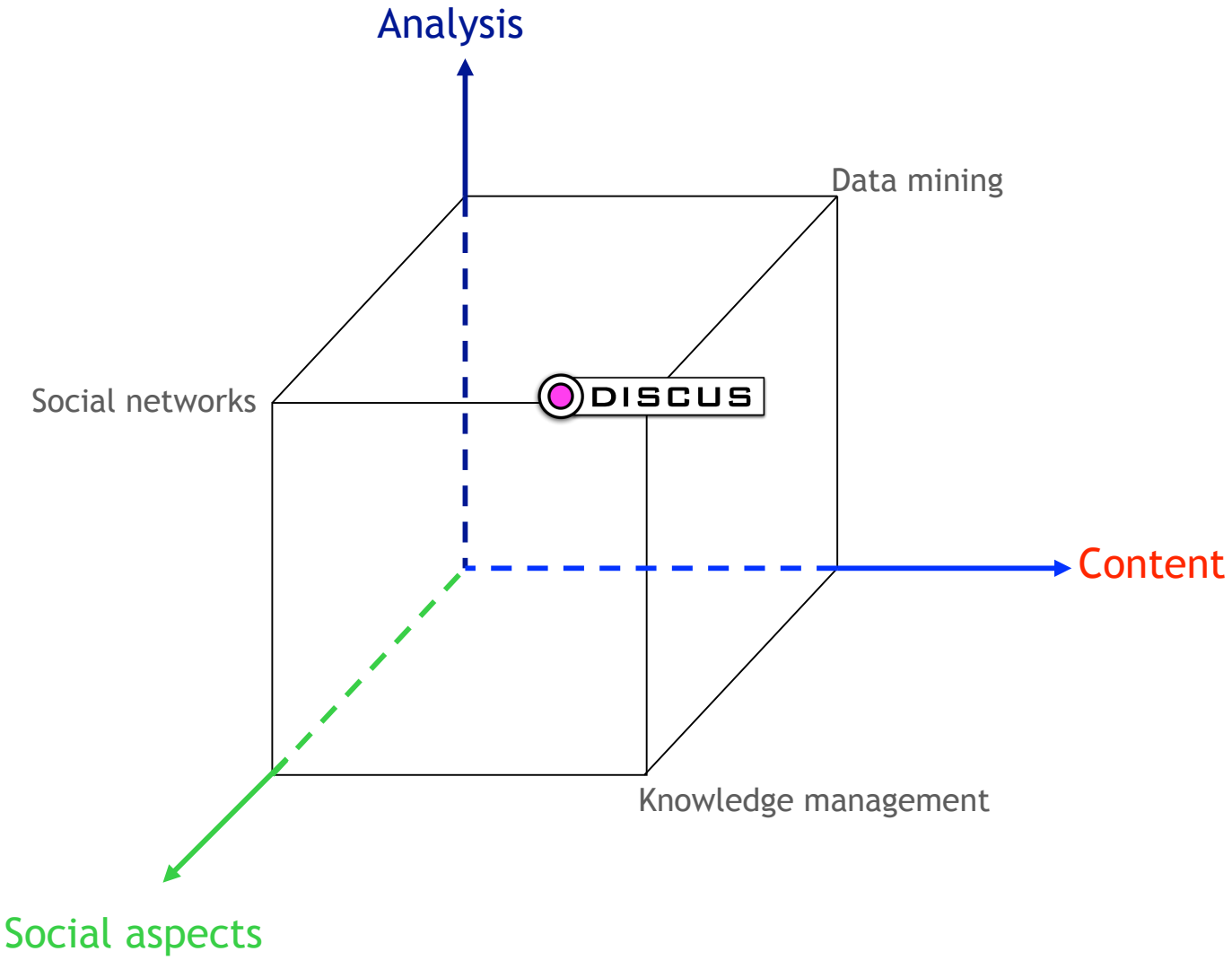
The DISCUS project

- IlliGAL & NCSA collaboration started in 2003
- Computers have become mediators of collaborations
  - Email, chat rooms, blogs, wikis…
  - A flood of available information
  - Different modes of communication
- Let's take advantage of such information
  - Logs of conversations
  - Archive of documents (email attachments, blogs, personal web pages…)
  - Human-computer interactions
  - Social aspect of the communication and collaboration

The question

- Could we help build a system using genetic algorithms/natural selection ideas that could foster innovation and creativity?
- The era of available Web 2.0 tools
  - Web portals
  - Collaboration and communication tools (message boards, chats, blogs, Twitter…)
  - Genetic algorithm (Interactive, Human-based, and Parallel GAs)
  - Chance discovery (KeyGraphs and Influence Diffusion Models)
  - Data and text mining
  - Information retrieval
  - Web services
  - Data-intensive computing

In a picture

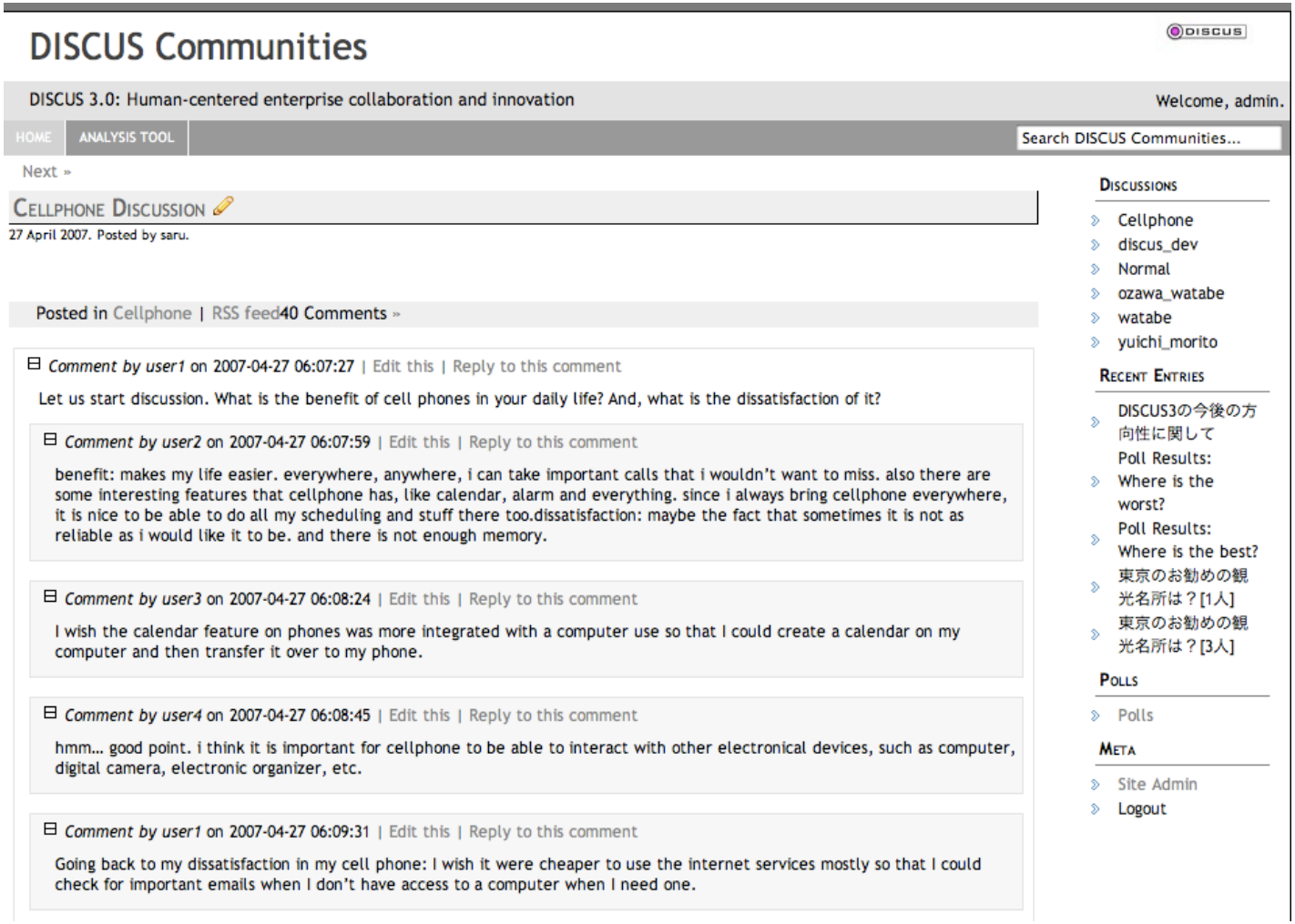


(Llorà et al., 2004)

A simple exercise

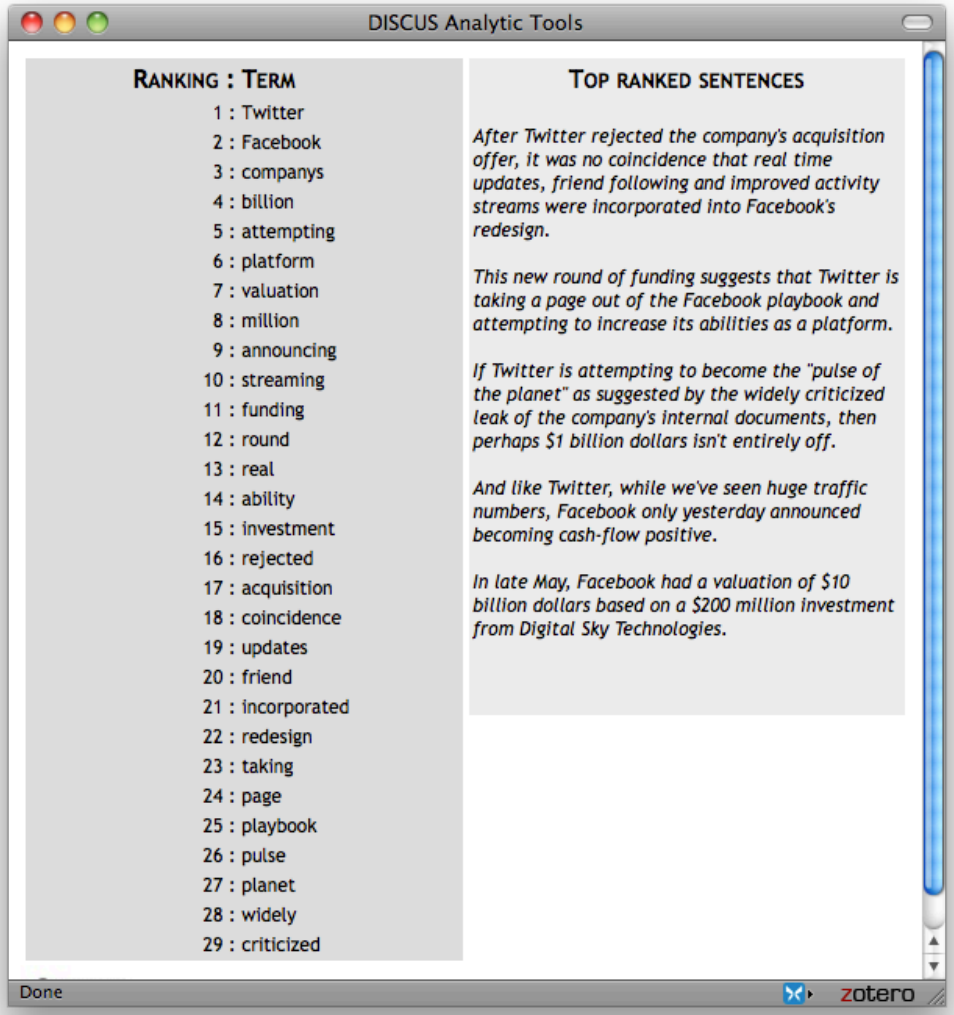
- Imagine you are chatting with your friends
- Your ideas are bundled in messages
  - You pass them your your friends
  - They read them, think about them
  - Your friends reply to you
- Could we model it using natural selection?
  - Valéry suggest so
  - Goldberg proposed a simple model using the human-based GA analogy
- If so, Goldberg's methodology should apply

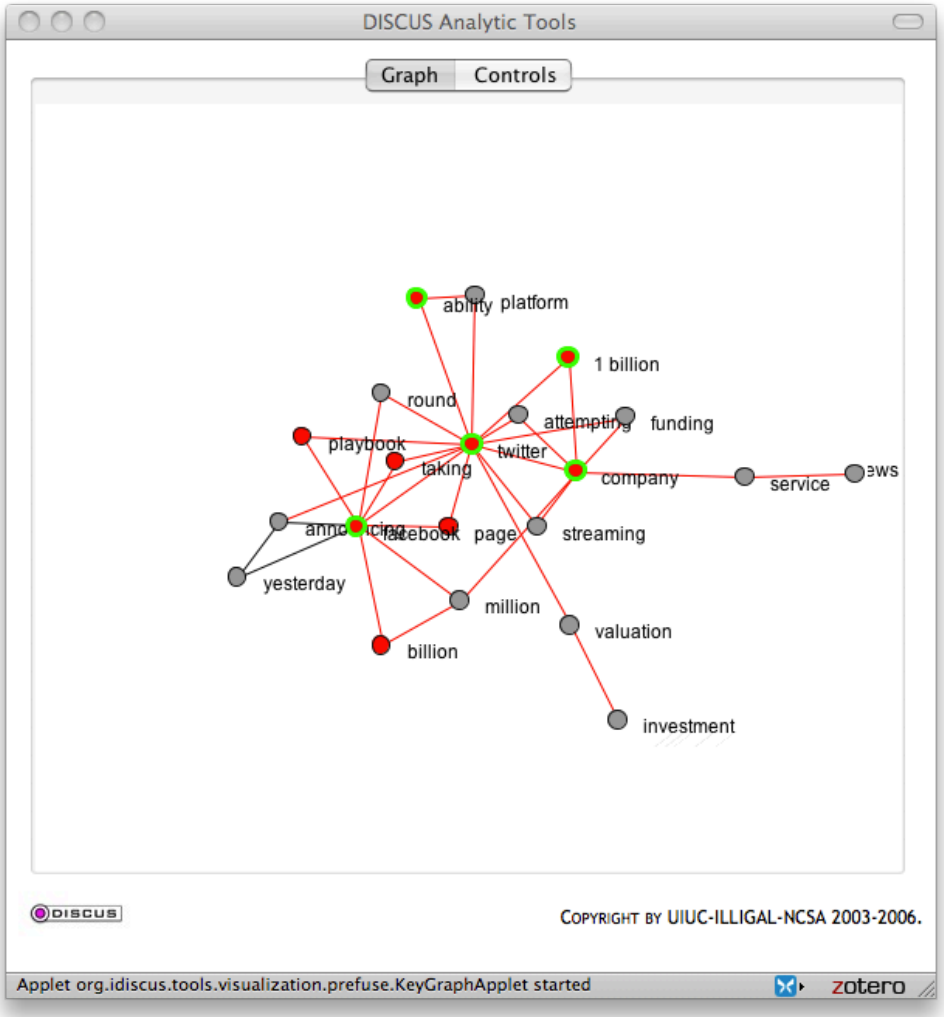
People interacting online



NCSA

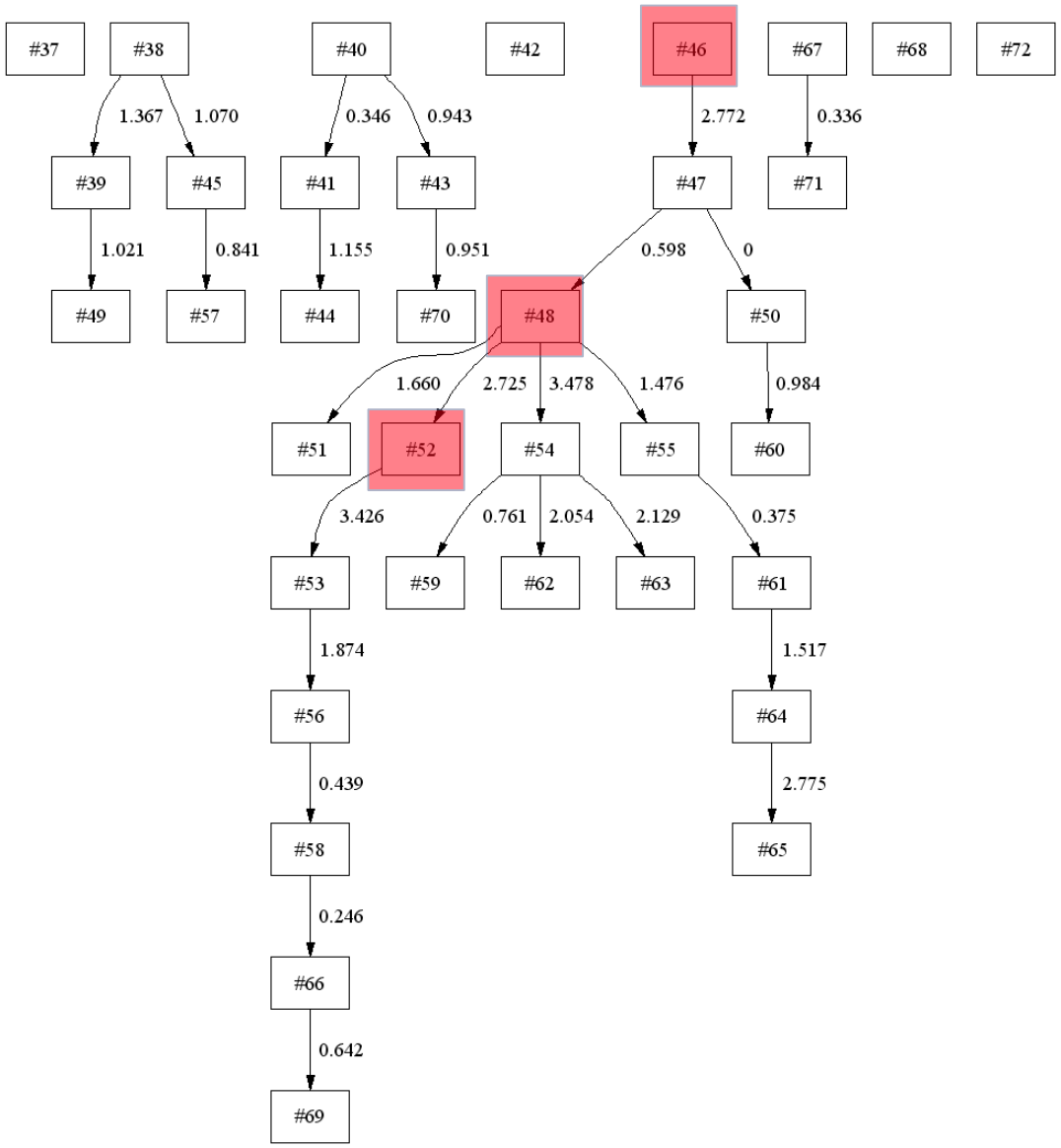
Concepts as building blocks





(Llorà et al., 2004)

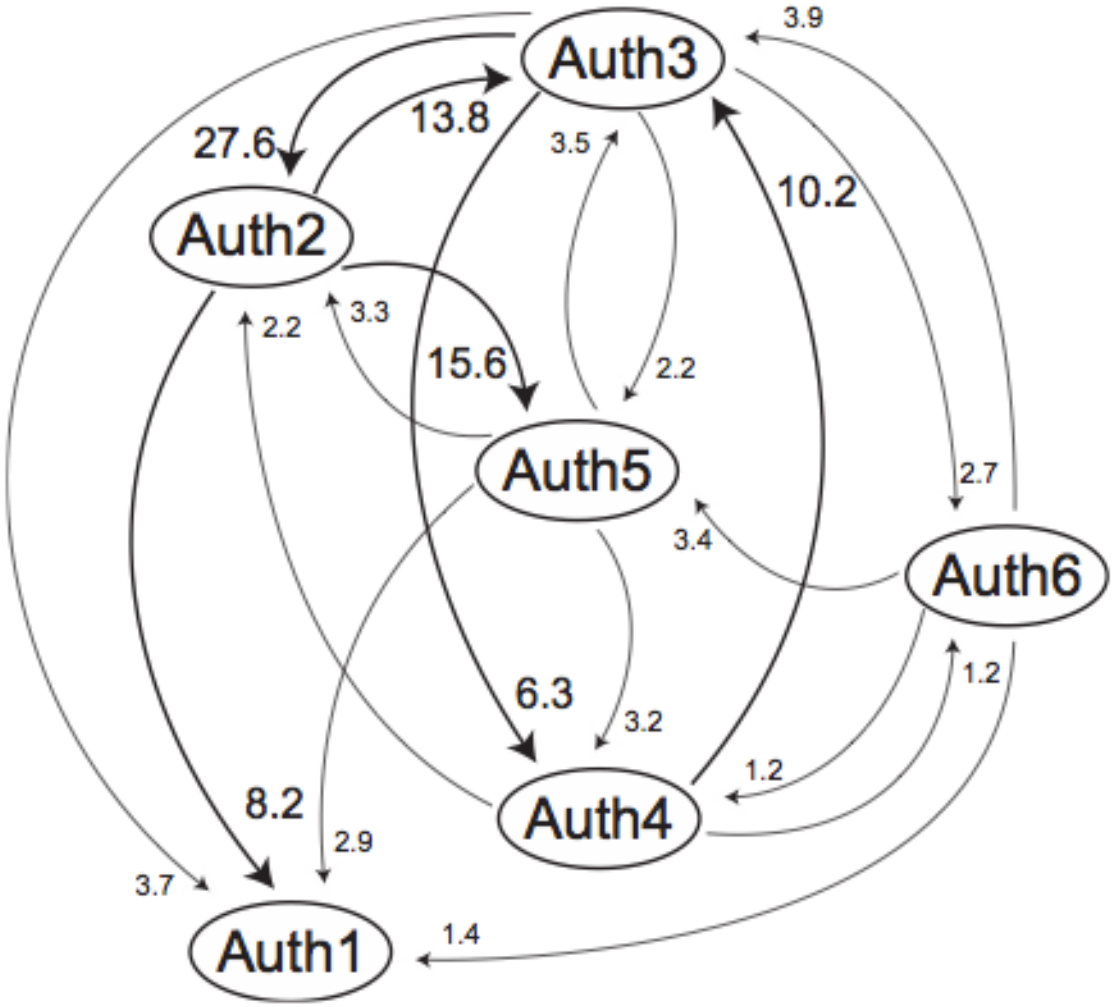
Things you can measure: Concept diffusion



(Llorà et al., 2005)

NCSA

Social acceptance and roles

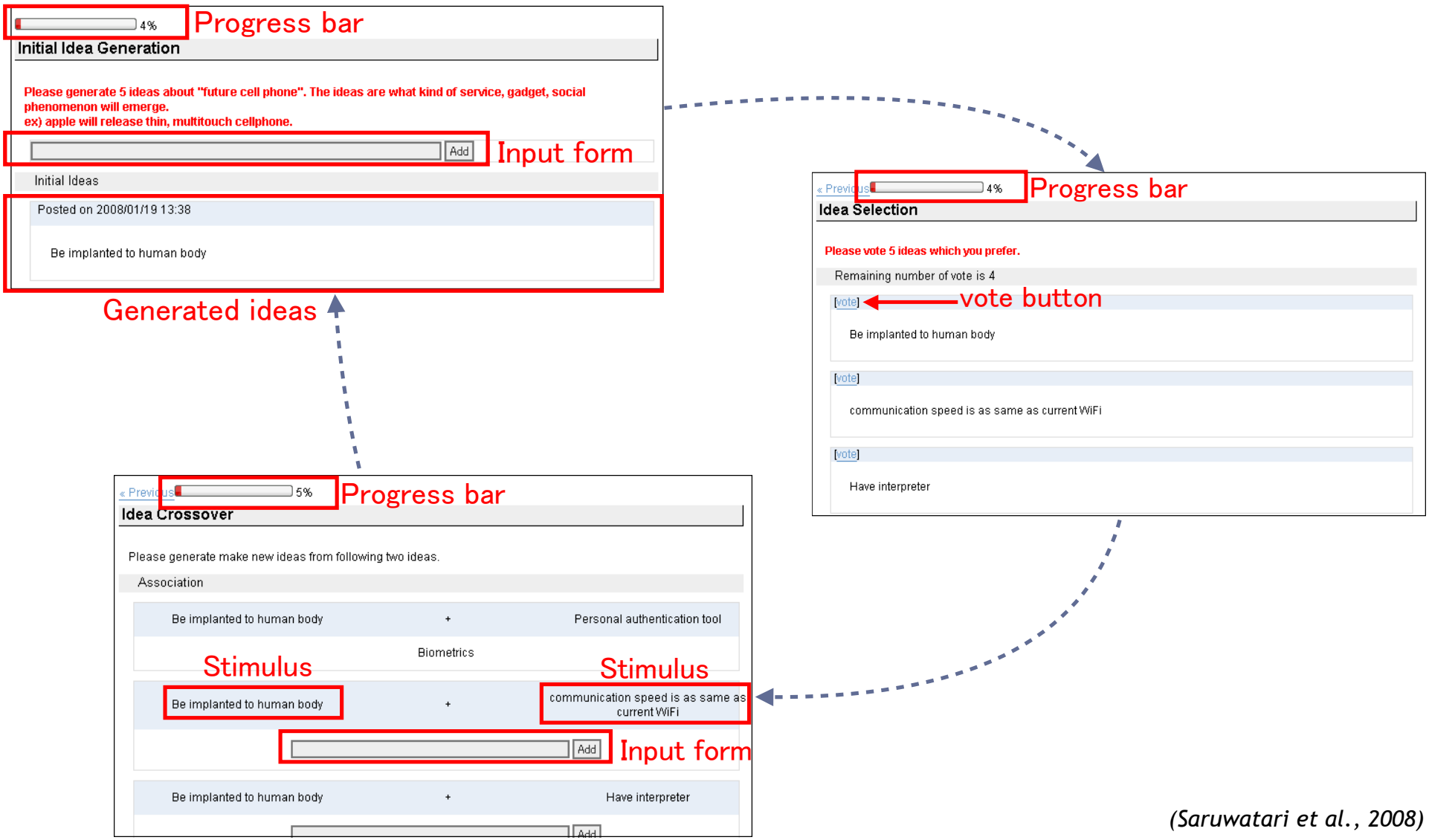


(Yasui et al., 2006)

Everything together

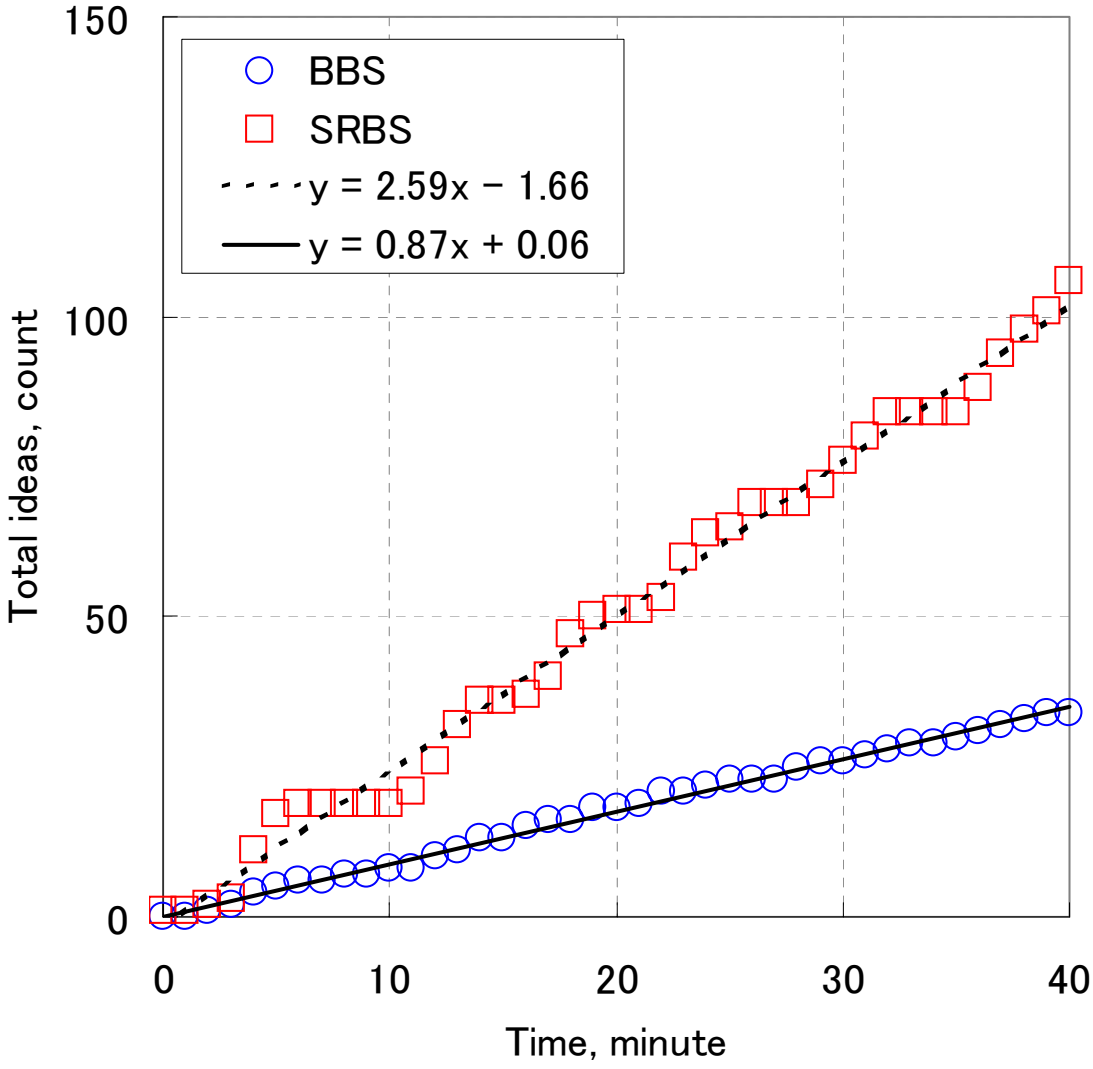
(Saruwatari et al., 2008)

Everything together



(Saruwatari et al., 2008)

In action



(Saruwatari et al., 2008)

Other possible usages

- Tracking influence on social media (e.g. Twitter)
- User preference modeling
- Topic overlap
- Topic/trend dynamics (how conversation is changing on the fly)
- Blogosphere and company message misalignment
- ...

... and beyond

SW development and Darwin's ideas?

- Data deluge, processing large volumes of data starting to become the norm, not the exception
- Another nagging question back on 2007
- Could we use the basic design principles by Goldberg to build software?
- A subset of those could have the answer
  - Understanding building blocks (concepts or ideas)
  - Ensure building block supply
  - Ensure building block growth
  - Ensure good building block mixing
- Also
  - Exploiting parallelism efficiently is becoming a more pressing issue
  - SW is a social activity

A little thinking

- Understanding building blocks (concepts or ideas)
  - Basic units of data processing
  - Building blocks should be easy to group and swap
- Ensure building block supply
  - Create a basic set of building block
- Ensure building block growth
  - Provide a set of tools for developers to add more building blocks
- Ensure good building block mixing
  - Facilitate the reproduction of solutions (programs)
  - Provide means to make those recombination possible

Modeling Meandre

- The basic outline
  - Data-flow execution paradigm
  - Semantic-web driven
  - Web oriented
  - Facilitate distributed computing
  - Support publishing services
  - Promote reuse, sharing, and collaboration
- Fall 2007 Meandre was adopted as the data-intensive platform for a collection of projects
- More information at http://seasr.org/meandre

Data flow execution in Meandre

- A simple example $c \leftarrow a+b$
- A traditional control-driven language

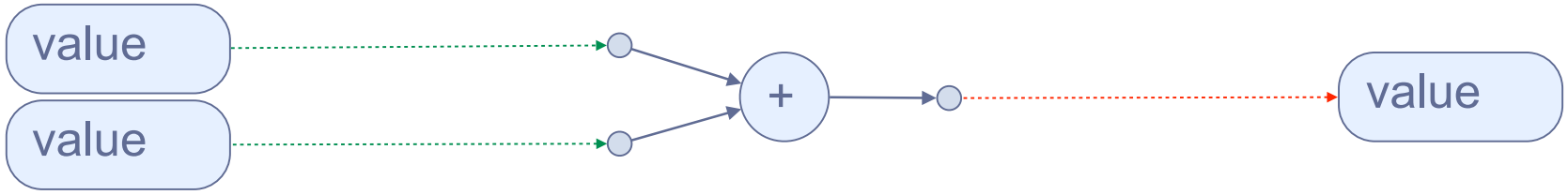
```
a = 1
b = 2
c = a+b
```

- Execution following the sequence of instructions
- One step at a time
  - a+b+c+d requires 3 steps
  - Could be easily parallelized

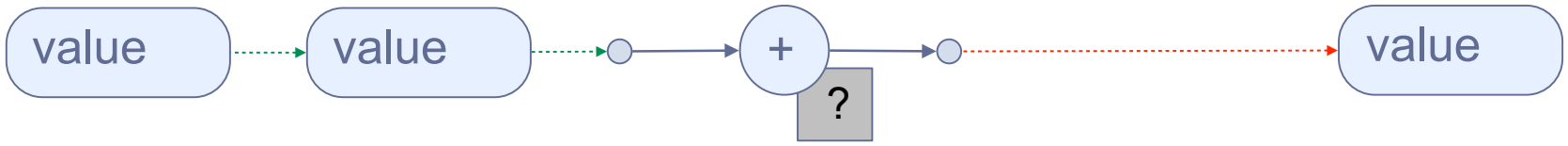
Data flow execution in Meandre

- Data flow execution is driven by data
- The previous example may have 2 possible data flow versions

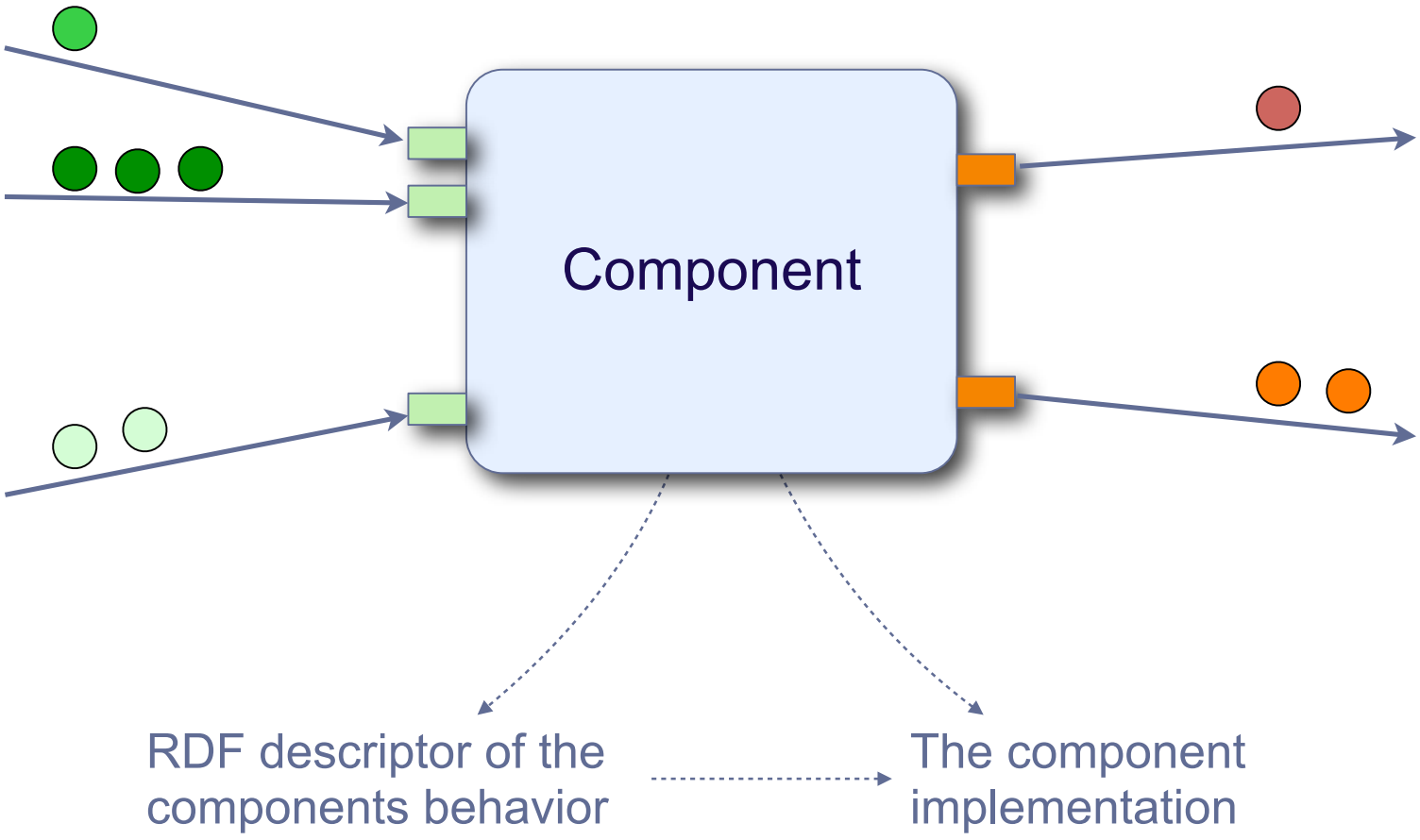
Stateless data flow



State-based data flow

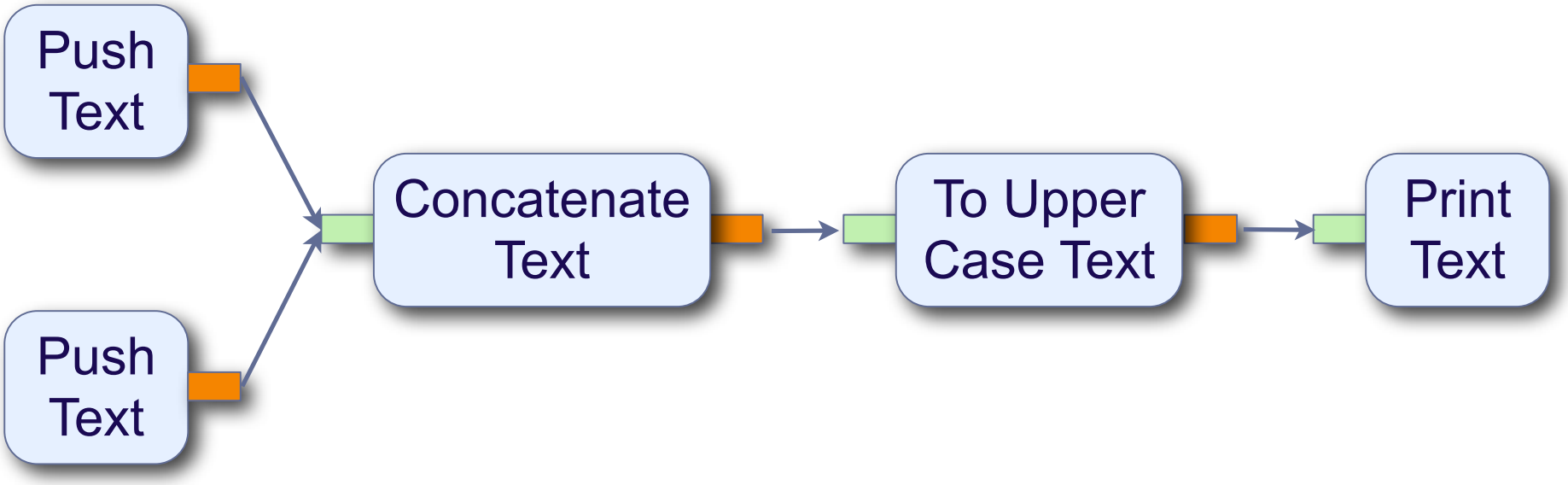


The basic building blocks: Components



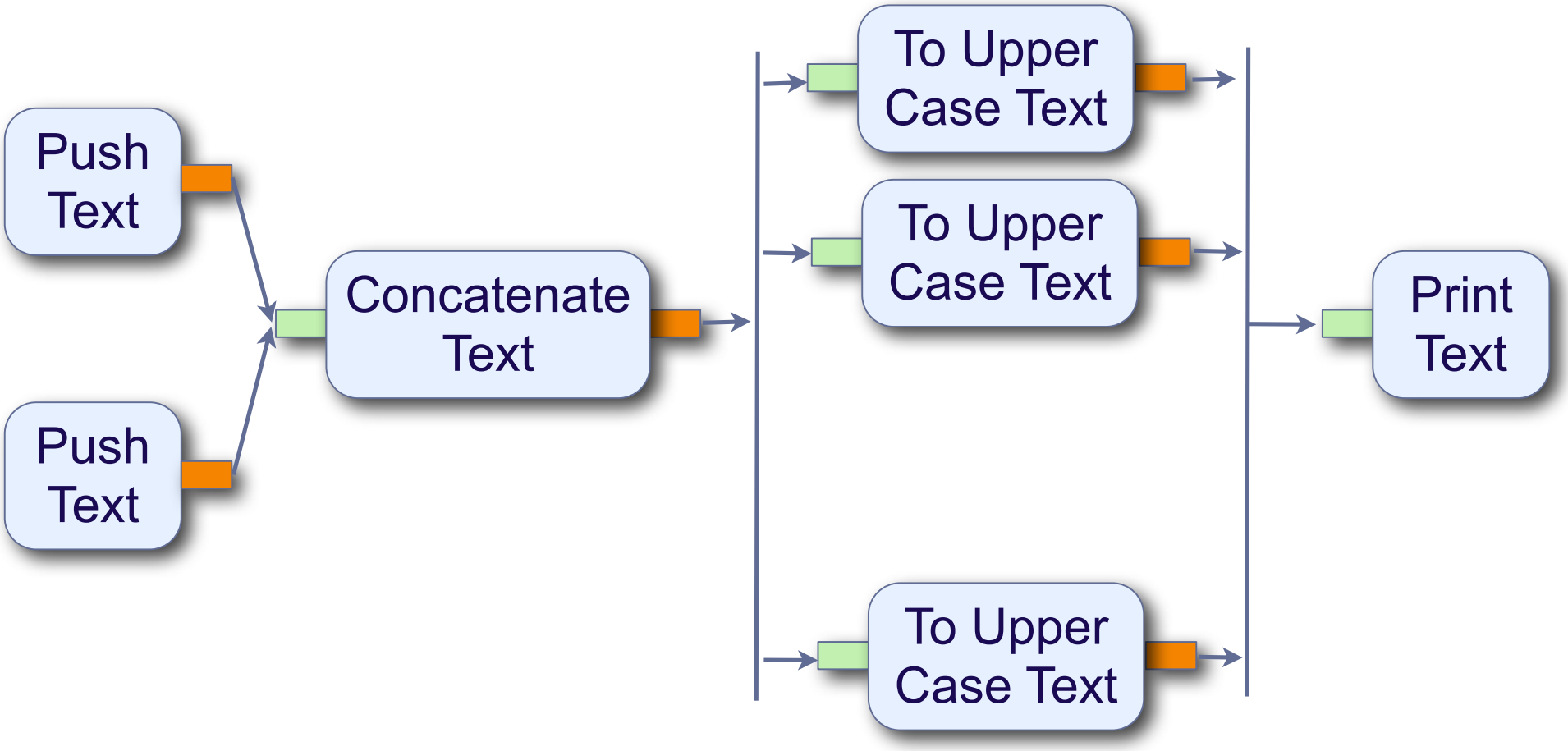
Go with the flow: Creating complex tasks

- Directed multigraph of components creates a flow

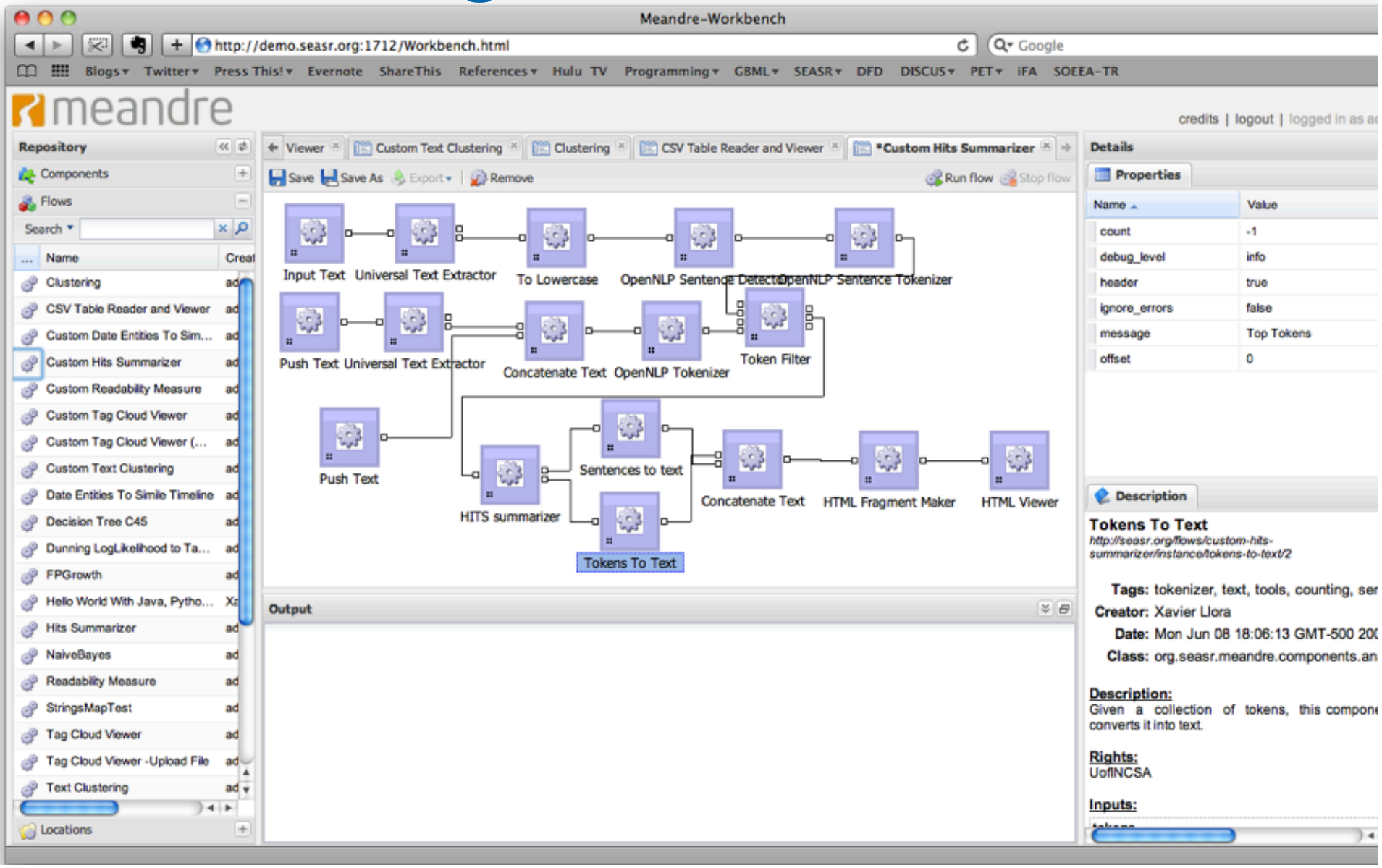


Automatic parallelization: Speed and robustness

- Meandre ZigZag language allow automatic parallelization



Visual building block recombination



Where is natural selection?

- More DISCUS-like evolutionary modeling
- Variations (components) in a population of solutions (flows)
- Selection by usage/adoption
- Recombination of flow available (encapsulation)

The take away message

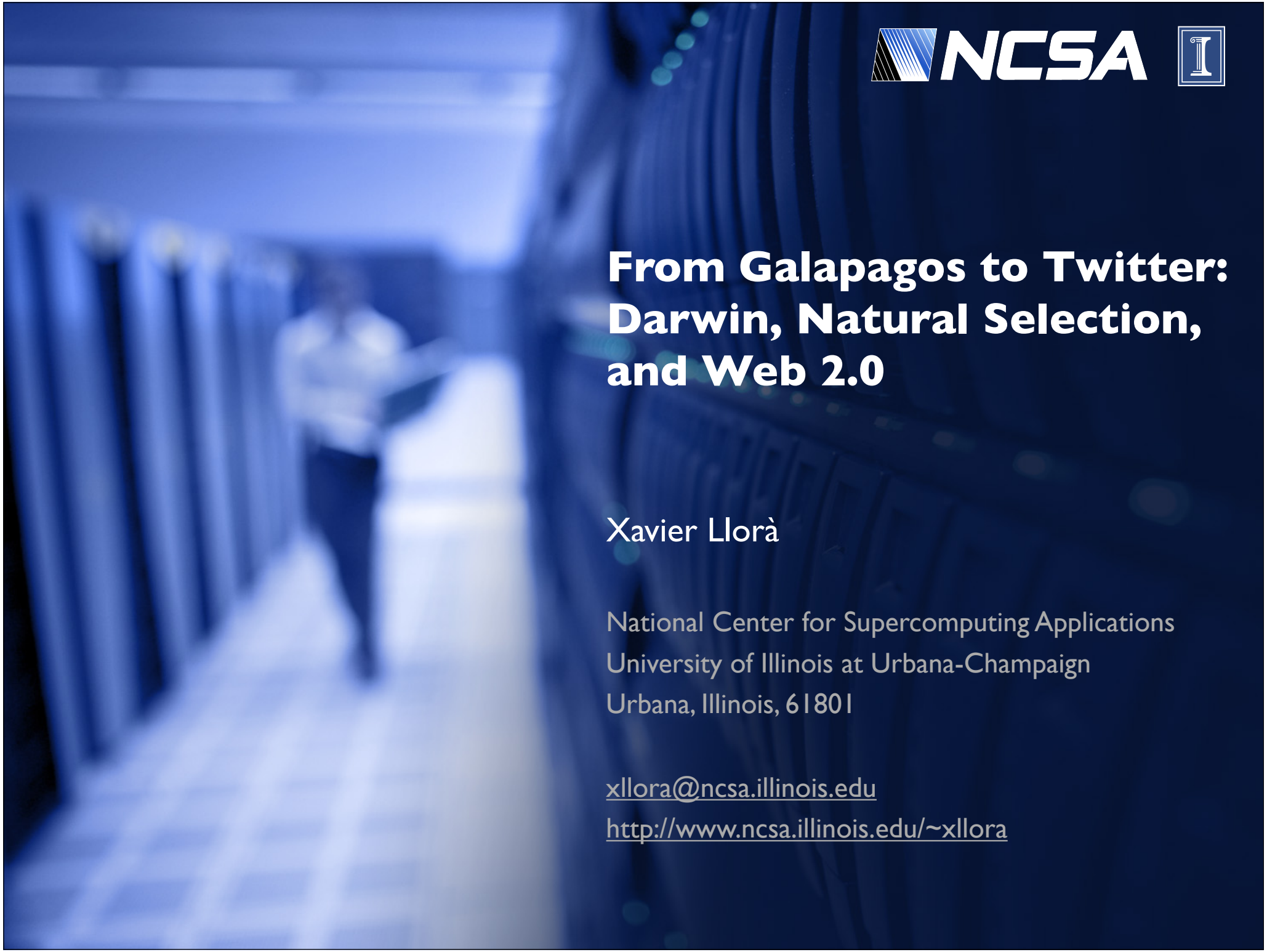
Basic food for thought

- The underlying distilled idea of natural selection
  - Variation will arise in a population
  - Selection by environment adaptation
- Powerful modeling tool
  - Problem solving (optimization and search)
  - Machine learning
  - Human social interaction modeling
  - Defining new ways to explore computer programming
  - ...
- Don't be afraid to try and explore

More information

- Illinois Genetic Algorithms Laboratory
  - http://www.illigal.uiuc.edu/web
- The DISCUS project
  - http://www.i-discus.org
- The Meandre project
  - http://seasr.org/meandre
  - http://seasr.org
- National Center for Supercomputing Applications
  - http://www.ncsa.illinois.edu
- Homepage
  - http://www.xavierllora.net
- Some related Twitter users
  - People: @xllora, @deg511
  - Labs and projects: @illigal, @projectmeandre, @seasrproject

NCSA



From Galapagos to Twitter: Darwin, Natural Selection, and Web 2.0

Xavier Llorà

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Urbana, Illinois, 61801

xllora@ncsa.illinois.edu
http://www.ncsa.illinois.edu/~xllora