CS 365
Introduction to Scientific Modeling
Review for Midterm

Stephanie Forrest
Dept. of Computer Science
Univ. of New Mexico
Fall Semester, 2014

THE UNIVERSITY *of* NEW MEXICO

Exam Format

- Closed book, closed note
- 1 hour
- Usually 5-8 questions, with sub-parts
  - Some short answer
  - I give you a CA rule, you show a sequence of states
  - I describe desired behavior, you give me some CA rules to implement it
  - 1 longer essay question
- Intended to be easy if you have done all the readings and did the homework yourself
  - A sanity check

Topics

- Models
  - What is a model?
  - Styles of modeling
    - Aggregate vs. individual
    - State-based vs. process-based
    - Deterministic vs. probablistic
  - How do we use models?
- Cellular automata
- Data analysis
  - Statistical distributions
  - Testing for power laws and other distributions
    - Curve fitting
    - Maximum likelihood estimates
- Power laws and scaling
  - Power laws in nature and technology
  - Mechanisms for generating power laws

Modeling

- How do we use models?
- How do we evaluate models?
- Different approaches to modeling
- Examples of different kinds of models and what they are used for
- Pros and cons of different modeling methods
- Limitations of modeling?

What is a model?

- A hypothetical description, often based on an analogy, used to study a behavior of interest.
- A representation of a system of interest.
- The representation is reduced---what to throw away?
- Examples?

Models for Complex Systems

- Concept-driven
  - Mathematical equations
  - Computational and simulation models
    - Agent-based modeling
- Data-driven
  - Bayesian networks

Model Representations

- Verbal descriptions
  - The “invisible hand” in economics
  - Clonal selection theory in immunology
  - Bohr’s model of the atom

Model Representations

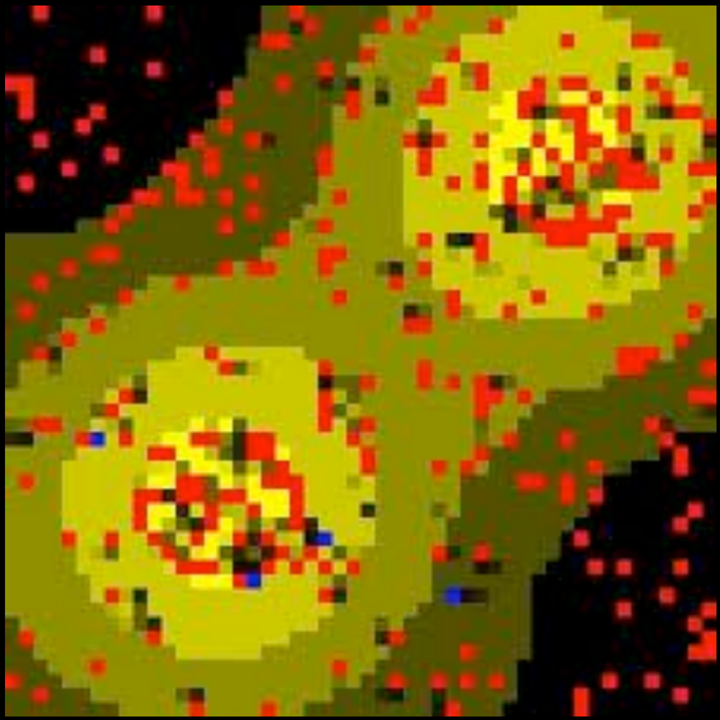
- Verbal descriptions
- Mathematical equations

$$\frac{dx}{dt} = Ax - Bxy$$

$$\frac{dy}{dt} = -Cy + Dxy$$

Model Representations

- Verbal descriptions, e.g.,
- Computer programs (model vs. simulation)
  - The real system
  - The model
  - The computer

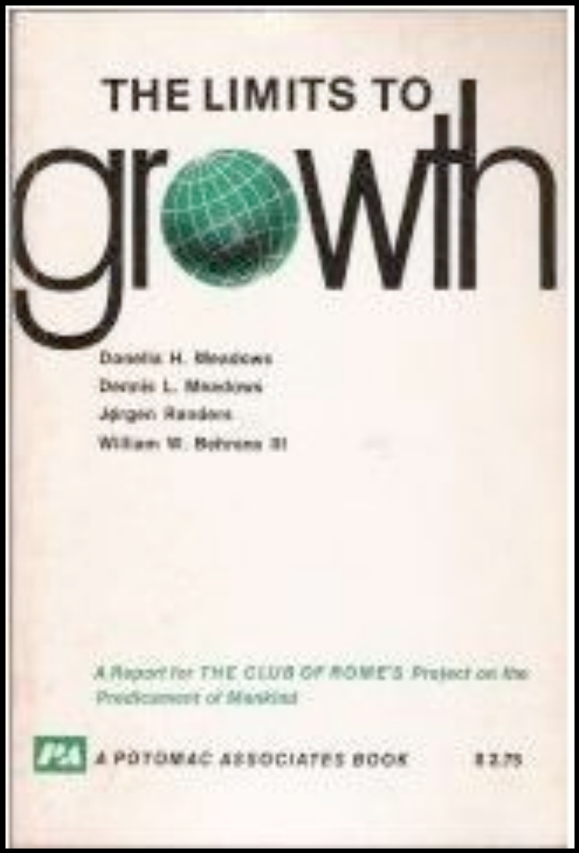


Examples of Models and Representations???

Example: Limits to growth model

Meadows et al. 1972

- Computer model:
  - 5 variables: populartion, industrialization, pollution, food production, resource depletion
- Goal: Not to make specific predictions, but to explore how exponential growth interacts with finite resources. Because the size of resources is not known, only general behavior can be explored.
- This process of determining behavior modes is "prediction" only in the most limited sense of the word. … These graphs are *not* exact predictions of the values of the variables at any particular year in the future. They are indications of the system's behavioral tendencies only.



Common Modeling Assumptions

- Homogeneity (all agents are identical / stateless)
- Equilibrium (no or very simple dynamics)
- Random mixing
- No feedback (learning)
- Deterministic
- No connection between micro and macro phenomena

Models with these assumptions can produce some interesting features, e.g., tipping points (R_0).

Hallmarks of Complex Systems

- Heterogeneous agents
- Nonlinear dynamics
  - Nonequilibrium systems
- Contact structure: networks, nonrandom mixing
- Stochastic behavior (interesting behavior in the tails)
- Learning/feedback
  - Agents can change their behavior
- Emergence
  - Multi-scale phenomena

What are the Challenges in Modeling Complex Systems?

- Closed form solutions rarely exist:
  - Features from previous slide
- Detailed simulations are problematic:
  - Can never hope to get all the details correct.
  - Because systems are nonlinear, small errors can have large consequences.
- Evolution is key:
  - Basic components change over time.
  - Individual variants matter (hard to do theory).
- Discreteness (e.g., time, state spaces, and internal variable values)
  - Techniques developed to study nonlinear systems are not always directly applicable
- Spatial heterogeneity

Classes of Scientific Models

- Continuous vs. discrete
- Deterministic vs. probabilistic
- Spatial vs. nonspatial
- Data-driven vs. theory-driven

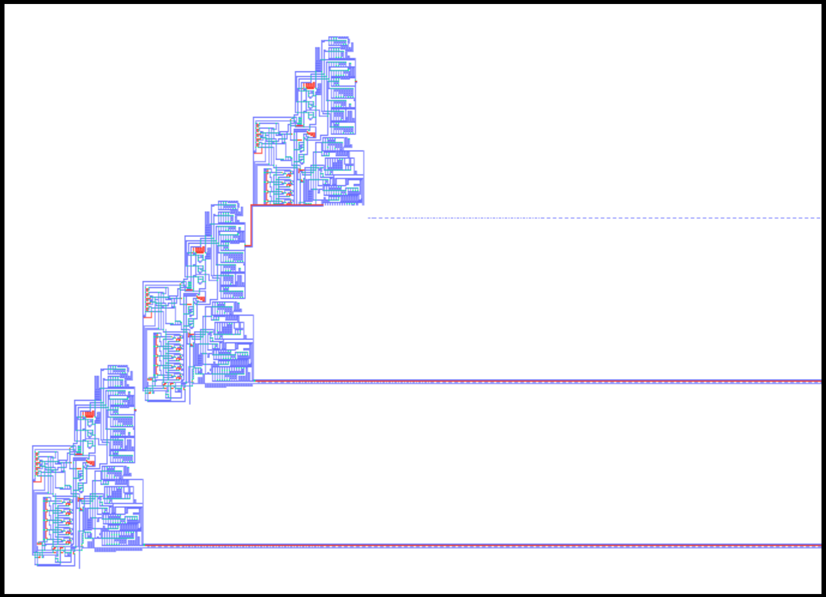
Classes of Scientific Models

- Continuous vs. discrete
  - e.g., differential equation vs. cellular automata
- Deterministic vs. probabilistic
  - Dynamical system vs. Markov chain
  - Cellular automata vs. genetic algorithm
- Spatial vs. nonspatial
- Data-driven vs. theory-driven
  - Bayseian networks vs. expert system

How do we use models?

- Making predictions (conventional use)
  - Quantitative (most analytical models)
  - Qualitative, e.g., critical parameters, regions of stability or instability
  - Validation: Accurate predictions

How do we use models?



ftp://ftp.ira.uka.de/pub/cellular-automata/jvn

- Existence proofs
  - Demonstrate that something is possible (e.g., von Neumann's self-reproducing automaton, patents)
  - Validation: it works

How do we use models?

- Building intuitions about how complex systems work (exploratory):
  - Examples: Flight simulators, Sim City.
  - Study how patterns of behavior, how resources flow, how co-operation arises, arms races.
  - Sensitivity analysis.
  - Discovering lever points for intervention, e.g., vaccines
  - Validation? (Trust, allows a pilot to land safely).

"*The contemplation in natural science of a wider domain than the actual leads to a far better understanding of the actual.* " A.S. Eddington. (first test of Einstein's Theory of Relativity experimentally)

How do we evaluate models?

- Parsimony and simplicity
  - Occam's Razor: select the hypothesis that makes the fewest assumptions when they are equal in other respects
- Accuracy of predictions
  - R^2 and other statistical tests
- Model works as claimed
  - Run it. e.g. patented devices
- Falsifiability
- Consistency---formalize notion of model as a *homomorphic map.*

How do we build models?

- Three elements in a computational model:
  - The system of interest
  - The model
  - The computer
- Modeling: The relationship between the real system and the model
- Simulation: The relationship between the computer and the model

Basic Modeling Approaches

- Top-down
  - Analyze problem / system of interest
    - Collect data: Live with your data!
  - Formulate a model
  - Make simplifying assumptions
  - Determine variables and model relationships between them, and submodels
  - Solve/run model
  - Compare to data
- Bottom-up
  - Collect data (training data)
  - Formulate a model to account for the data
  - Test model on additional data for goodness of fit

Case Studies

- Cancer
- Forest fire model
- Metabolic scaling in chips and biology
- Spam/malware
- Student presentations

CELLULAR AUTOMATA

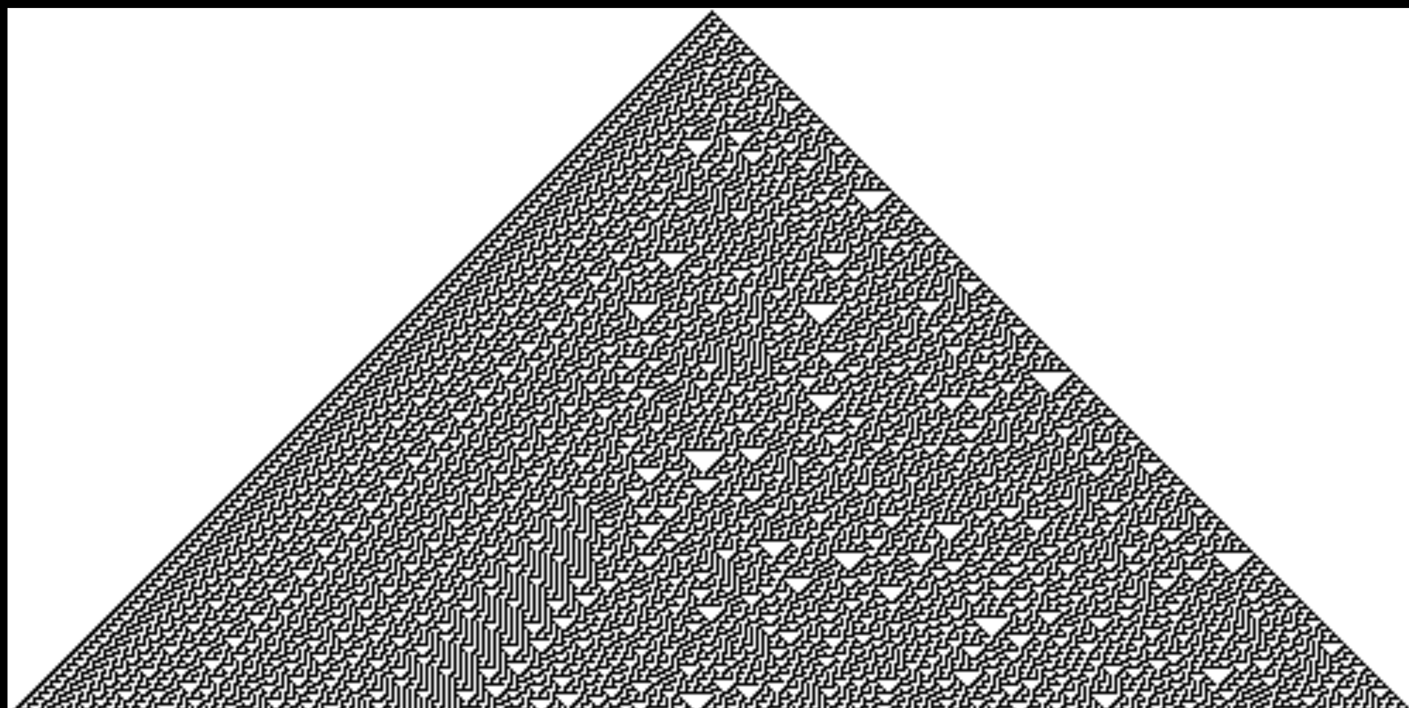
Cellular Automata Topics

- 1-dimensional
  - Space time plots
- Wolfram classes
- 2-dimensional
- Case studies
  - Forest fire model
  - Game of life

CAs as Discrete Simulation Models

- Cellular Automata are discrete
  - Time changes in incremental steps
    - Differential equations for continuous time, typically
  - Space is represented explicitly, in regular arrangements of cells
  - Each cell is in one of a finite number of states at any given time
- Deterministic
  - Initial states of cells determine the rest of the computation
- Each of these assumptions can be relaxed

Rule 30



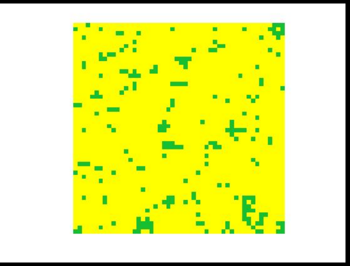
current pattern	111	110	101	100	011	010	001	000
new state for center cell	0	0	0	1	1	1	1	0

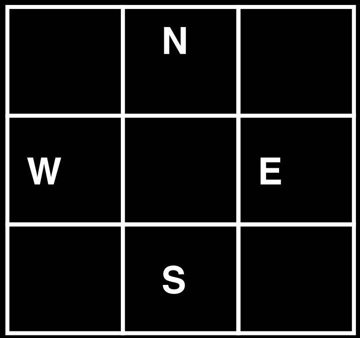
CA can exhibit complex behavior Wolfram's Classification

- Class I: Eventually every cell in the array settles into one state, never to change again
  - Analogous to computer programs that halt after a few steps and to dynamical systems that have fixed-point attractors
- Class II: Eventually the array settles into a periodic cycle of states, called a limit cycle
  - Analogous to computer programs that execute infinite loops and to dynamical systems that fall into limit cycles
- Class III: The array forms "aperiodic" random-like patterns
  - Analogous to computer programs that are pseudo-random number generators (pass most tests for randomness, highly sensitive to seed, or initial condition).
  - Analogous to chaotic dynamical systems. Almost never repeat themselves, sensitive to initial conditions, embedded unstable limit cycles
- Class IV: The array forms *complex* patterns with localized structure that move through space and time
  - Difficult to describe. Not regular, not periodic, not random
  - Speculate: this is *interesting* computation, the edge of chaos
  - Example: Rule 110

Forest Fire Model

- 4 transition rules
  - A burning cell turns into an empty cell
  - A tree will burn if at least one neighbor is burning
  - A tree ignites with probability *f* even if no neighbor is burning
  - An empty space fills with a tree with probability *p*
- Neighborhood (von Neumann)





How are forest fire models related to power laws?

The Game of Life

John Conway (1970)

- The number of states, k = 2
- The alphabet, Σ = {0,1}   // {‘dead’,’alive’}
- The Moore neighborhood

NW	N	NE
W		E
SW	S	SE

- Transition rules:
  – Loneliness: A live cell with less than 2 live neighbors, dies
  – Overcrowding: A live cell with more than 3 live neighbors, dies
  – Birth: A dead cell with exactly 3 live neighbors becomes a live cell
  – Survival: A live cell with 2 or 3 live neighbors stays alive

Periodic Objects

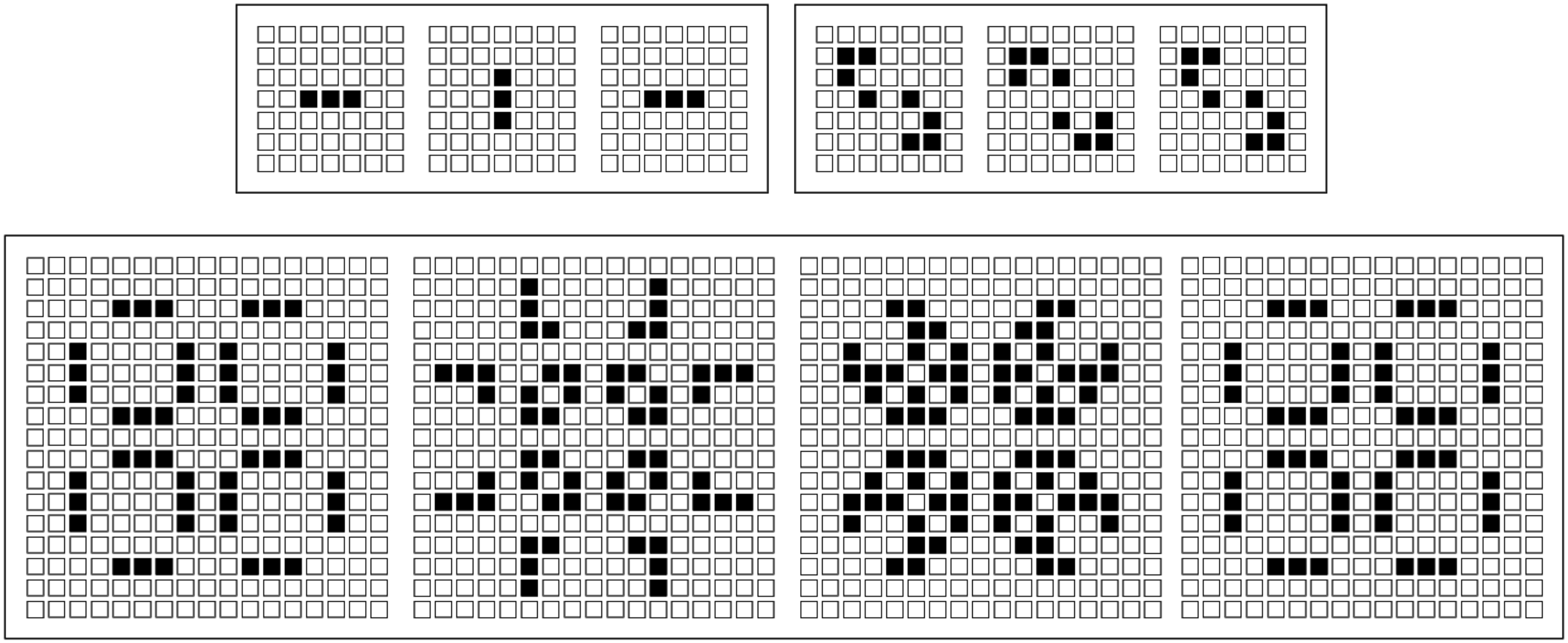
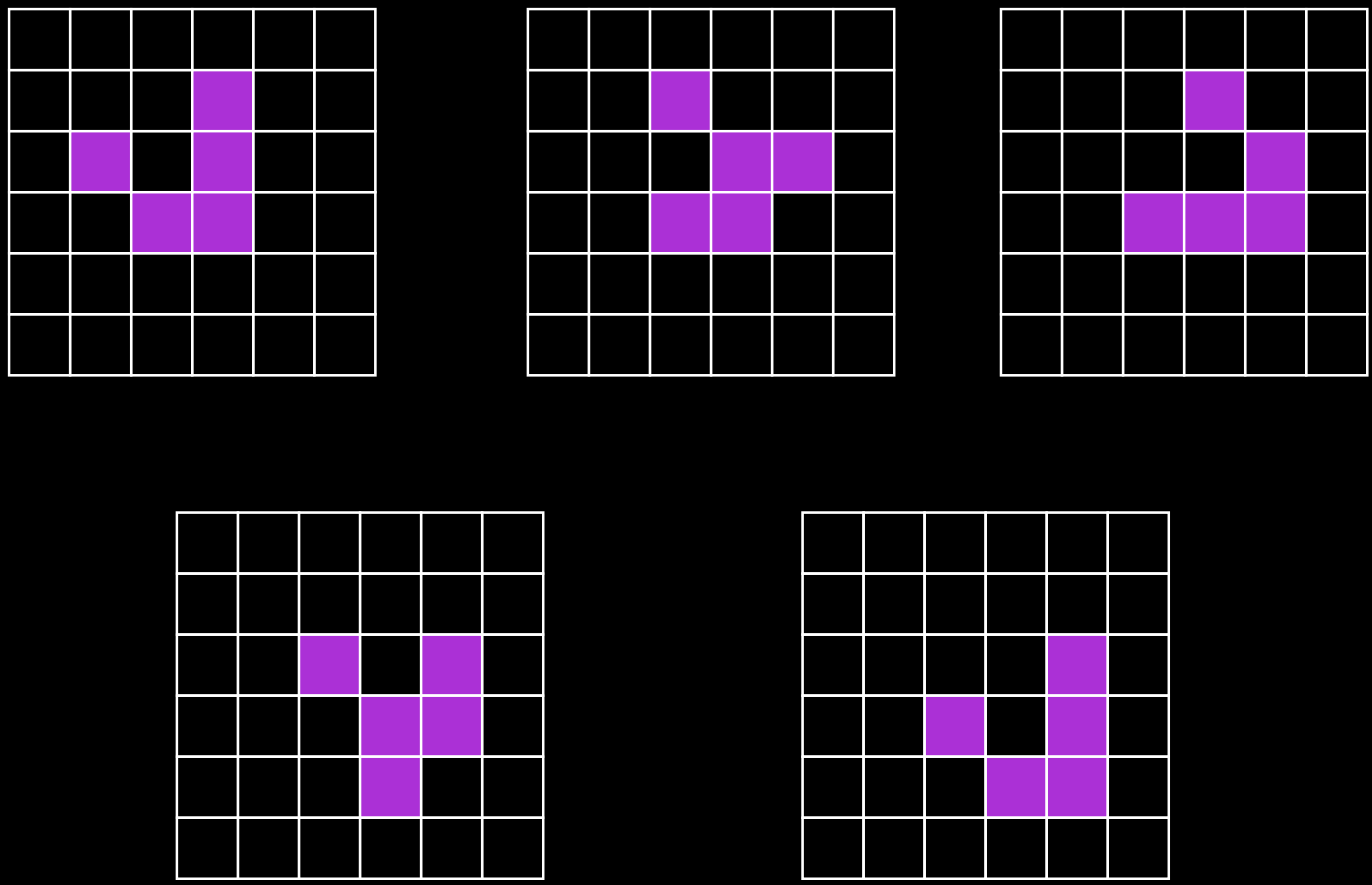


Figure 15.12 Examples of simple periodic objects in Conway's Game of Life

Gliders (moving objects)

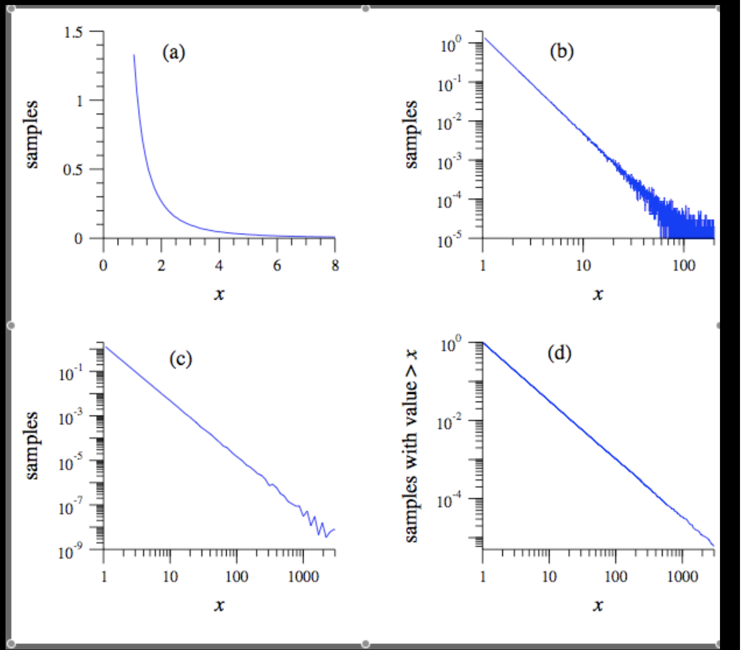


Topics

- Statistical distributions
  - Normal
  - Log normal
  - Exponential
  - Power law
- Power laws in nature and computing
  - Complex networks
- Detecting power laws in data
- How power laws are generated
- Special topics
  - Metabolic scaling theory

Nice properties of power laws

- Scale invariant
  - Distribution can range over many orders of magnitude
- Log log plot is a straight line
  - Noise in tail
- CCDF is also a power law
  - Exponent is –b + 1



Testing for distributions: Review

- Two basic strategies
  - Plot data, fit curves, measure goodness of fit
  - Maximum Likelihood calculations
- Normal
  - Plot on semilogy (look for a quadratic curve)*
  - Mean gives MLE
- Log normal
  - Log data, then treat like a normal distribution
- Exponential
  - Plot on semilogy and look for linear fit
  - MLE is given in the assignment
- Power law
  - Plot on log-log scale and look for lines (++)

* Don't try this at home

How to decide which distribution best fits your data?

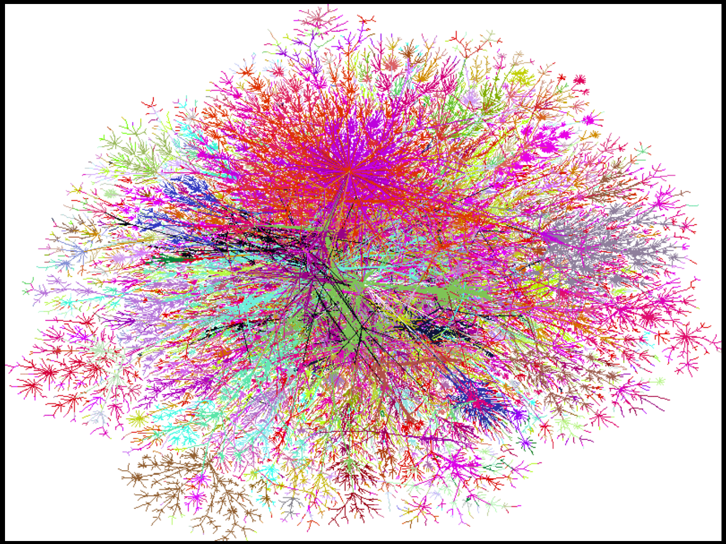
1. Estimate the Maximum Likelihood Estimate (MLE) parameters for each distribution
   1. Power law, log normal, exponential
2. Compute the loglikelihood for each distribution using the MLE parameter estimates
3. Compare the loglikelihoods

Basic Mechanisms that Produce Power Laws

- Preferential attachment (previous slides)
- Combinations of exponentials
- Random walks
- Phase transitions (next slide)
  - Critical points
- Optimizations
  - Evolution
  - Engineering

Complex Networks
Important properties

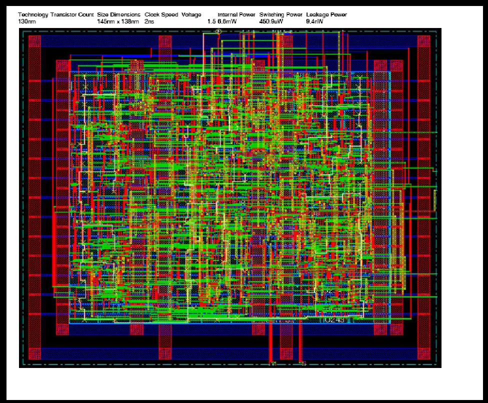
- Degree distribution is often power law (ish)
  - Centrality
- Low diameter
  - 6 degrees of separation
  - Small world effect
- Community structure
  - Cliques, triangles, etc.



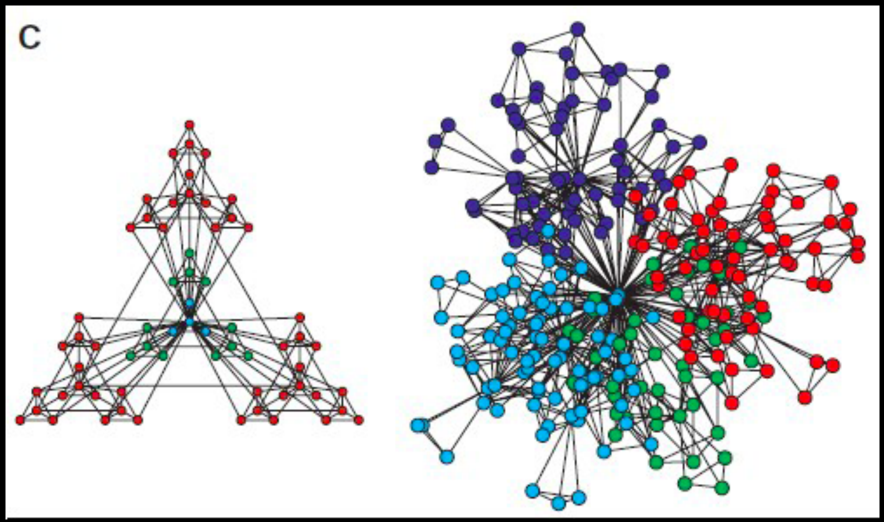
Examples?

Physical and Geometric constraints determine network architecture and growth

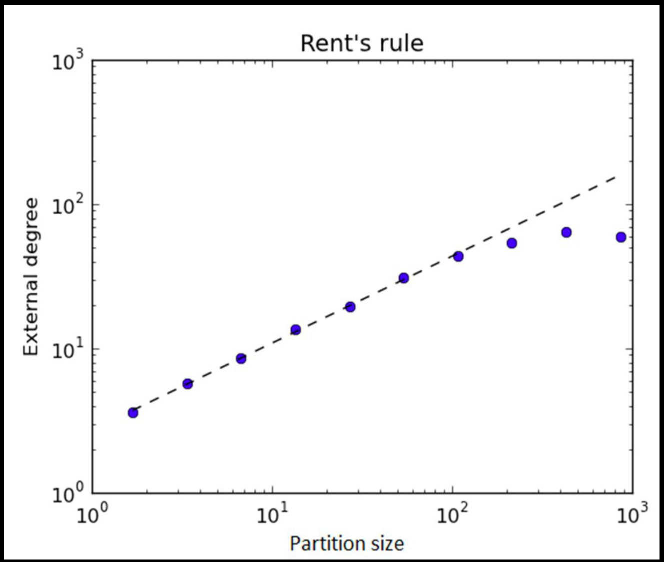
- Network capacity limits performance as systems scale
- Metabolism, response times, power consumption
- Are universal patterns in system behavior predictable from the scaling properties of distribution networks?



Hierarchical Modularity



Log-log plot of C vs. N

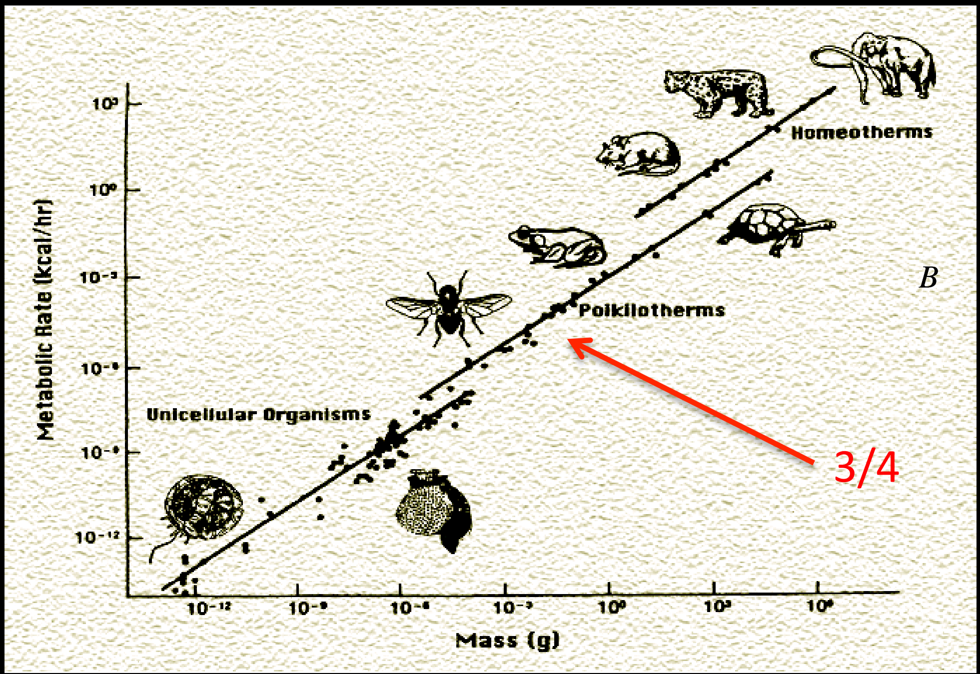


C = communication
N = circuit size
P = Rent's exponent, in 0,1

$$C \propto N^p$$

Rent's rule for benchmark circuit c3540

Kleiber's Law

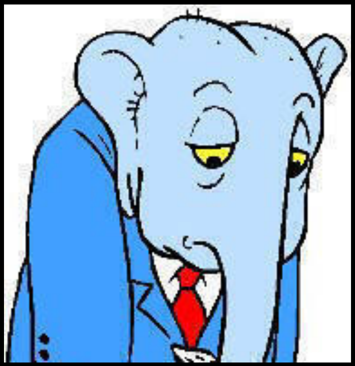


Hemmingson, 1960

- Observed metabolic scaling

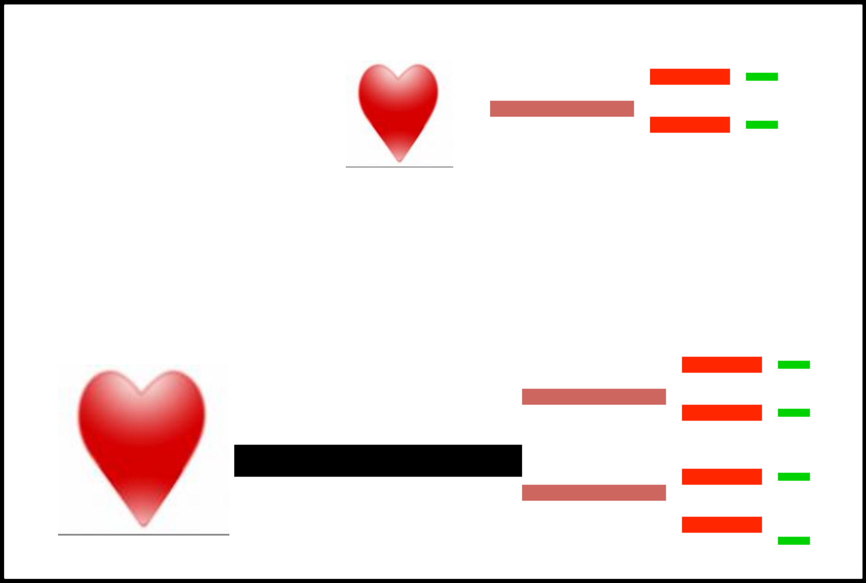
$$B \propto M^{3/4}$$

- B is the rate of energy (oxygen) use
  - Mass specific scaling
- B is the master biological rate that governs
  - Ecological interactions
  - Food webs & ecosystem dynamics $\propto M^{-1/4}$
- Other biological rates
  - Biological times $\propto M^{1/4}$



Metabolic Scaling Theory

- Larger organisms require larger networks
  - Pipe lengths (L) are longer
  - Cross-section areas (A) are larger
  - # capillaries increases more slowly than pipe volume: $N = cV^{3/4}$
  - Metabolism: $B = cM^{3/4}$



Increasing volume (mass) 100 times increases delivery rate 30 times

Diminishing returns: Network size grows faster than network delivery rate