



Andreas Holzinger
185.A83 Machine Learning for Health Informatics
2017S, VU, 2.0 h, 3.0 ECTS
Lecture 12 - Week 24 – 13.06.2017



Evolutionary Computing, Neuroevolution and Genetic Algorithms: Toward Tumor-Growth Simulation

a.holzinger@hci-kdd.org

<http://hci-kdd.org/machine-learning-for-health-informatics-course>



Holzinger Group, hci-kdd.org

1

Machine Learning Health 12

00 Reflection

Holzinger Group, hci-kdd.org

4

Machine Learning Health 12

01 Evolutionary Principles



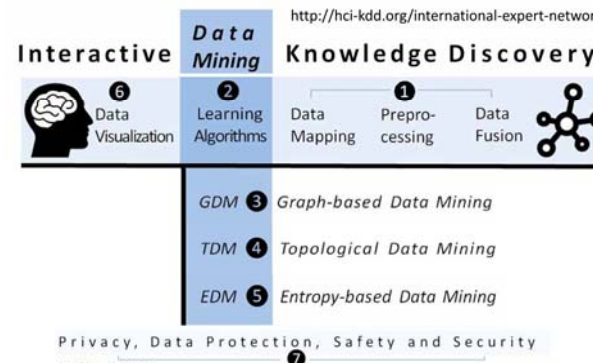
<http://www.interaliomag.org/audiovisual/thomas-ray-aesthetically-evolved-virtual-pets/>

"Evolution is the natural way to program"
Thomas S. Ray, University of Oklahoma,
<http://life.ou.edu/>

Holzinger Group, hci-kdd.org

7

Machine Learning Health 12

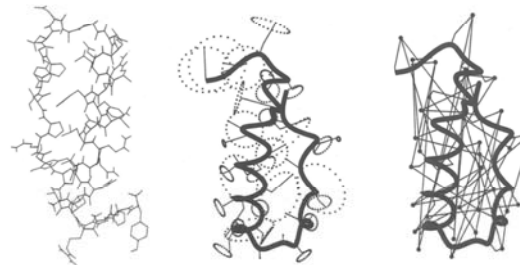


Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: Cognitive Science meets Machine Learning. IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

Holzinger Group, hci-kdd.org

2

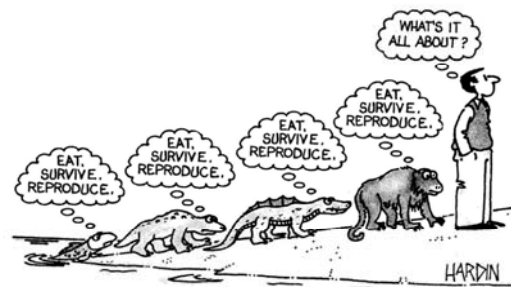
Machine Learning Health 12



Holzinger Group, hci-kdd.org

5

Machine Learning Health 12



Knoll, A. H. & Bambach, R. K. 2000. Directionality in the history of life: diffusion from the left wall or repeated scaling of the right? Paleobiology, 26, 1-14.

Holzinger Group, hci-kdd.org

8

Machine Learning Health 12

- 00 Reflection
- 01 Evolution
- 02 Neuroevolution
- 03 Genetic Algorithms
- 04 Medical Example: Tumor-Growth Simulation

Holzinger Group, hci-kdd.org

3

Machine Learning Health 12



Holzinger Group, hci-kdd.org

6

Machine Learning Health 12



- Jean Baptiste de Lamarck, 1801. Theory of Inheritance of Acquired Characteristics, Paris



- Charles Darwin, 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, London, John Murray.



- James M. Baldwin, 1896. A New Factor in Evolution. The American Naturalist, 30, (354), 441-451, doi:10.2307/2453130.



- Gregor Mendel, 1866. Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brunn 4: 3, 44.

Holzinger Group, hci-kdd.org

9

Machine Learning Health 12

- The goal of aML is to build systems that learn and make decisions *without* the human.
- Early aML efforts, e.g. the perceptron [1], had been truly inspired by human intelligence.
- Today, probabilistic modelling has become the cornerstone of aML [2], with applications in neural processing [3] and human learning [4].




[1] McCulloch, W. S. & Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biology, 5, (4), 115-133, doi:10.1007/BF02459570.

[2] Doya, K., Ishii, S., Pouget, A. & Rao, R. 2007. Bayesian brain: Probabilistic approaches to neural coding, Boston (MA), MIT press.

[3] Deneve, S. 2008. Bayesian spiking neurons I: inference. Neural computation, 20, (1), 91-117.

[4] Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. science, 331, (6022), 1279-1285.

Biological Universe vs. Computational Universe

NOTION	BIOLOGICAL UNIVERSE	COMPUTATIONAL UNIVERSE
Chromosome 	DNA, protein, and RNA sequence in cells	Sequence of information objects
Fitness 	Determines chances of survival and reproduction	Determines chances of survival and reproduction
Gene 	Part of a Chromosome, determines a (partial) characteristic of an individual	Information object, e.g. a bit, a character, number etc.
Generation	Population at a point in time	Population at a point in time
Individual	Living organism	Solution candidate
Population	Set of living organisms	Bag or multi-set of Chromosomes

Holzinger, K., Palade, V., Rabadan, R. & Holzinger, A. 2014. Darwin or Lamarck? Future Challenges in Evolutionary Algorithms for Knowledge Discovery and Data Mining. In: LNCS 8401. Heidelberg, Berlin: Springer, pp. 35-56.

Optimization of a Naive Bayes classifier

- **Naive Bayes** is a very effective classifier
- EAs need parameters that can be modified
- A **Weighted Naive Bayesian (wnb)** [1] classifier offers the possibility of easy optimization:

$$p(a_1, a_2, \dots, a_n | c) = \prod_{i=1}^n p(a_i | c). \quad V_{nb}(E) = \arg \max_c p(c) \prod_{i=1}^n p(a_i | c)$$

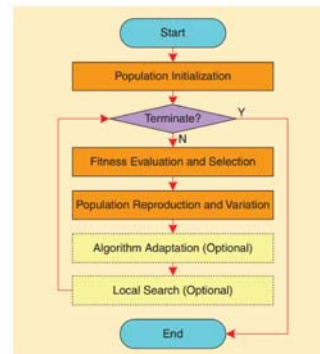
$$V_{wnb}(E) = \arg \max_c p(c) \prod_{i=1}^n p(a_i | c)^{w_i}$$

[1] Zhang, H. & Sheng, S. Learning weighted naive Bayes with accurate ranking. Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on, 2004. IEEE, 567-570.

- Based on the evolutionary theories of **Darwin, Lamarck, Baldwin, Mendel**.
- Since the 1980s, EAs have been used for **optimization** problems
- Exploring the possibility of **optimizing** machine learning algorithms rather recently [1]

[1] Z. Zhang, G. Gao, J. Yue, Y. Duan, and Y. Shi, "Multi-criteria optimization classifier using fuzzification, kernel and penalty factors for predicting protein interaction hot spots," Applied Soft Computing, vol. 18, no. 0, pp. 115-125, 2014.

The General Evolutionary Computation Framework [1]

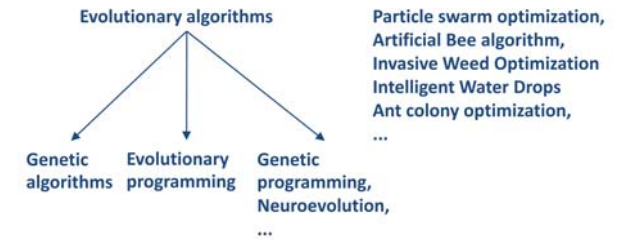


[1] Zhang, J., Zhan, Z.-H., Lin, Y., Chen, N., Gong, Y.-J., Zhong, J.-H., Chung, H. S., Li, Y. & Shi, Y.-H. 2011. Evolutionary computation meets machine learning: A survey. Computational Intelligence Magazine, IEEE, 6, (4), 68-75.

Implementation

- **Dataset:** Pima Indians Diabetes dataset [8]
 - 768 instances (patients)
 - 8 attributes
 - 2 classes
- **Fitness** of an chromosome determined by: number of correctly classified instances in training set
- **Performance** was compared to algorithms in Weka

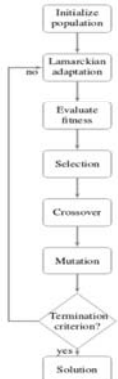
[8] K. Bache and M. Lichman, "UCI machine learning repository," 2013.[Online]. Available: <http://archive.ics.uci.edu/m>



[1] Michalewicz, Z. 1996. Genetic algorithms + data structures = evolution programs, New York, Springer.

Lamarckian/Baldwin Adaptation [1]

- Modify chromosomes to adapt to the environment
 - can be used additionally or instead of mutation process
- A local search optimization is applied (e.g. Hill Climbing)
- **Baldwin** uses only pseudo adaptation



[1] B. J. Ross, "A lamarckian evolution strategy for genetic algorithms," Practical handbook of genetic algorithms: complex coding systems, vol. 3, pp. 1-16, 1999.

Fitness function

Algorithm 1 Fitness function

```

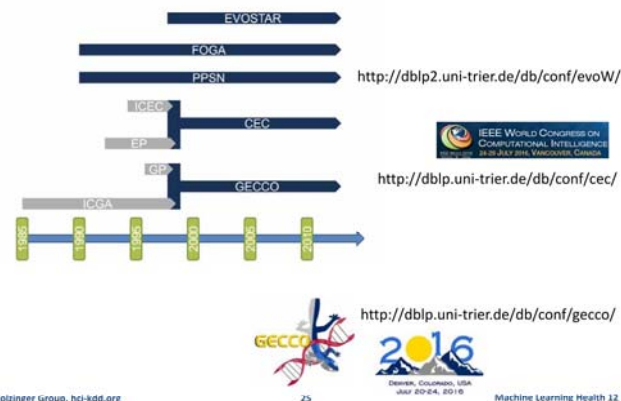
1: procedure FITNESS FUNCTION(weights[], List trainingSet)
2:   for all instances of trainingset do
3:     for i = 1 to NumberOfClasses do
4:       for all attribute to MaxNumberAttributes do
5:         probability[i] *= NORMDISTRIBUTION(attribute + weights[attribute])
6:
7:   index ← INDEX OF MAX(probability[])
8:
9:   if index == CLASS OF(instance) then
10:    INCREMENT(fitness)
11:   else
12:    DECREMENT(fitness)
13:
14: RETURN fitness
  
```


Classifier	Correctly Classified Instances
Evolutionary classifier	522
Naive Bayes classifier	586
Bayes Net classifier	571

- Advantages:
 - Fast to train and fast to classify
 - Not sensitive to irrelevant features
 - Handles real and discrete data
- Disadvantages:
 - Assumes independence of features

- Text mining with EAs on unstructured information:
 - Doctors/Nurse reports
 - Different Medical Records
 - ...
- Sample applications:
 - Categorizing Texts into subject groups [1]
 - Mining “interesting” details [2] like:
 - Gender ▪ Addresses
 - Age ▪ Occupation

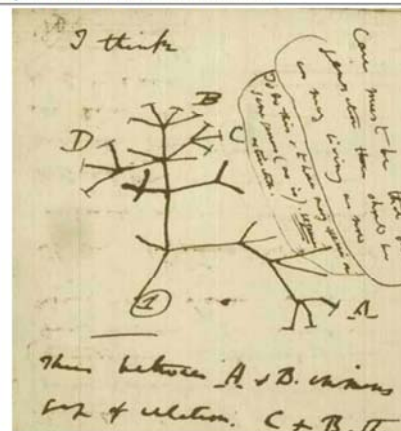
[2] Deepankar B. and Suneet S. Text Mining Technique using Genetic Algorithm. *IJCA Proceedings on International Conference on Advances in Computer Application 2013 ICACA 2013*: 7-10, Feb. 2013. Pub.: Foundation of Computer Science, N.Y., USA.



- Offers many possibilities to improve machine learning algorithms, but finding the right parameters is a difficult task
- Not many machine learning algorithms are suitable for **direct function optimization**
- Implementation of EA:
 - straightforward
 - simple
- EAs are suitable for many tasks in health informatics beyond function optimization

Holzinger, A., Blanchard, D., Bloice, M., Holzinger, K., Palade, V. & Rabadan, R. Darwin, Lamarck, or Baldwin: Applying Evolutionary Algorithms to Machine Learning Techniques. In: Slezak, D., Dunin-Keplicz, B., Lewis, M. & Terano, T., eds. IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 Warsaw, Poland. IEEE, 449-453, doi:10.1109/WI-IAT.2014.132.

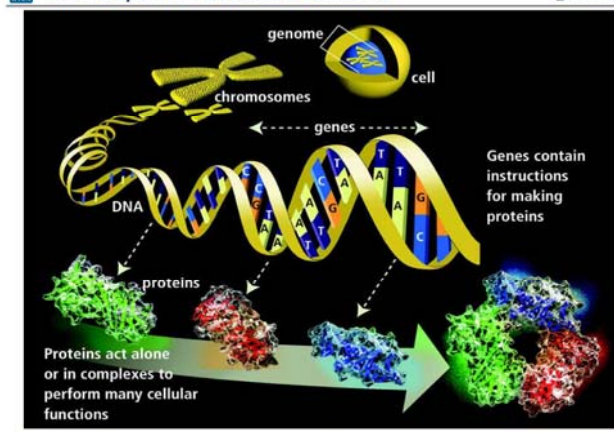
Evolutionary Computing

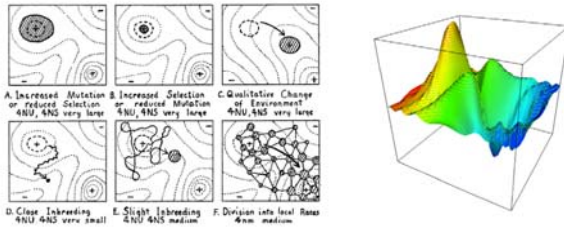


- Improvement of function optimization strategy
- Use EAs in different **fields**
 - Graph Optimization
 - Text Mining [1]
 - Feature selection
- Usage of novel **evolutionary strategies**
 - Intelligent Water Drops
 - Invasive Weed
 - Ant Colony with humans-in-the-loop (Super-Ants)

[1] Mukherjee, Indrajit, et al. Content analysis based on text mining using genetic algorithm. In: Computer Technology and Development (ICCTD), 2010 2nd International Conference on. IEEE, 2010. S. 432-436.2

- 1948 Alan Turing:
"genetical or evolutionary search"
- 1962 Hans-Joachim Bremermann:
optimization through evolution and recombination
- 1964 Ingo Rechenberg:
introduces evolution strategies
- 1965 Lawrence J. Fogel, Owens and Walsh:
introduce evolutionary programming
- 1975 John Holland:
introduces genetic algorithms
- 1992 John Koza:
introduces genetic programming





An evolving population is conceptualized as moving on a surface whose points represent the set of possible solutions = search space

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. 6th International Congress on Genetics. Ithaca (NY). 356-366.

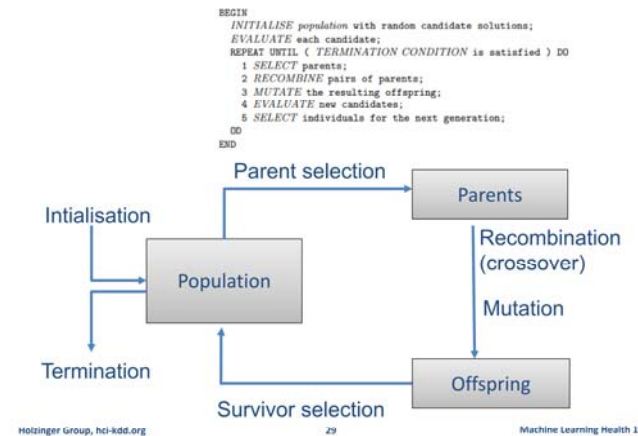
Two competing forces

- 1) **Increasing population diversity** by genetic operators (e.g. mutation, recombination, ...) Push towards creating **novelty**
- 2) **Decreasing population diversity** by selection of parents and survivors Push towards **quality**

Evaluation function = Fitness function

- Role:
 - Represents the task to solve, the requirements to adapt to (can be seen as "the environment")
- Enables selection (provides basis for comparison)
 - e.g., some phenotypic traits are advantageous, desirable, e.g. big ears cool better, these traits are rewarded by more offspring that will expectedly carry the same trait
- A.k.a. *quality function* or *objective function*
- Assigns a single real-valued fitness to each phenotype which forms the basis for selection
 - So the more discrimination (different values) the better
- Typically we talk about fitness being maximised
 - Some problems may be best posed as minimisation problems, but conversion is trivial

General Scheme of an Evolutionary Algorithm



Main EA components: Representation

- Role: provides code for candidate solutions that can be manipulated by variation operators, and leads to two levels of existence:
 - phenotype**: object in original problem context (outside)
 - genotype**: code to denote that object, the inside (chromosome, "digital DNA")
- Implies two mappings:
 - Encoding: phenotype \rightarrow genotype (not necess. 1:1)
 - Decoding: genotype \rightarrow phenotype (must be 1:1)
- Chromosomes contain genes, which are in (usually fixed) positions called loci and have a value (allele)



Basic Model of Evolutionary Process

- Population of individuals
- Each individual has a fitness function
- Variation operators: crossover, mutation, ...
- Selection towards higher fitness by "survival of the fittest" and "mating of the fittest"

Neo Darwinism:

Evolutionary progress towards higher life forms

= Optimization according to some fitness-criterion (optimization on a fitness landscape)

Phenotype \rightarrow Genotype (integers \rightarrow binary code)

- In order to find the global optimum, every feasible solution must be represented in the genotype space

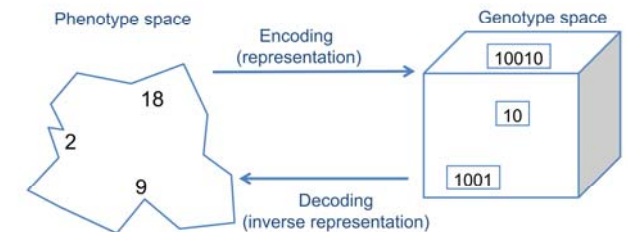


Image credit: Eiben, A. E. & Smith, J. E. 2015. Introduction to evolutionary computing. Second Edition, Berlin, Springer.

Population

- Role: holds the candidate solutions of the problem as individuals (genotypes)
- Formally, a population is a multiset of individuals, i.e. repetitions are possible
- Population is the basic unit of evolution, i.e., the population is evolving, not the individuals
- Selection operators act on population level
- Variation operators act on individual level
- Some sophisticated EAs also assert a spatial structure on the population e.g., a grid
- Selection operators usually take whole population into account i.e., reproductive probabilities are relative to current generation
- Diversity** of a population refers to the number of different fitness / phenotypes / genotypes present (note: not the same thing)

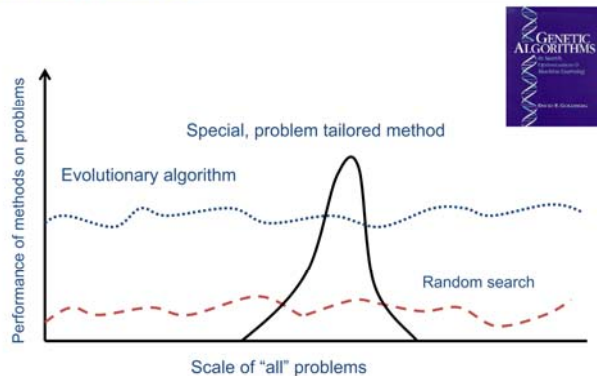
Selection mechanism

Role:

- Identifies individuals
 - to become parents
 - to survive
- Pushes population towards higher fitness
- Usually probabilistic
 - high quality solutions more likely to be selected than low quality
 - but not guaranteed
 - even worst in current population usually has non-zero probability of being selected
- This *stochastic* nature can aid escape from local optima

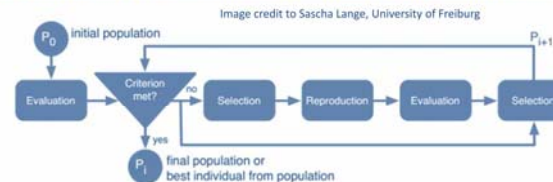
- Most EAs use fixed population size so need a way of going from (parents + offspring) to next generation
- Often deterministic (while parent selection is usually stochastic)
- Fitness based : e.g., rank parents + offspring and take best
- Age based: make as many offspring as parents and delete all parents
- Sometimes a combination of stochastic and deterministic (elitism)

- Role: merges information from parents into offspring
- Choice of what information to merge is stochastic
- Most offspring may be worse, or the same as the parents
- Hope is that some are better by combining elements of genotypes that lead to good traits
- Principle has been used for millennia by breeders of plants and livestock



- Role: to generate new candidate solutions
- Usually divided into two types according to their arity (number of inputs):
 - Arity 1 : mutation operators
 - Arity >1 : recombination operators
- Arity = 2 typically called crossover
- Arity > 2 is formally possible, seldom used in EC
- There has been much debate about relative importance of recombination and mutation
- Nowadays most EAs use both
- Variation operators must match the given representation

- Initialisation usually done at random,
- Need to ensure even spread and mixture of possible allele values
- Can include existing solutions, or use problem-specific heuristics, to "seed" the population
- Termination condition checked every generation
 - Reaching some (known/hoped for) fitness
 - Reaching some maximum allowed number of generations
 - Reaching some minimum level of diversity
 - Reaching some specified number of generations without fitness improvement



- Individuals:** hypothesis x from a hypothesis space X
- Population:** collection P of μ hypotheses $P = \{x_i \mid i = 1, \dots, \mu\}$
- Evaluation:** $f : X \rightarrow R$ (fitness function) to all individuals
- Selection mechanism:** selects individuals $x \in P_i$ for reproduction (mating); selects individuals from off-springs and P_i to form the new population P_{i+1}
- Reproduction:** combination of two or more individuals (Crossover) and random alteration (Mutation).

- Role: causes small, random variance
- Acts on one genotype and delivers another
- Element of randomness is essential and differentiates it from other unary heuristic operators
- Importance ascribed depends on representation and historical dialect:
- Binary GAs – background operator responsible for preserving and introducing diversity
- EP for FSM's / continuous variables – only search operator
- GP – hardly used
- May guarantee connectedness of search space and hence convergence proofs

- Historically different EAs have been associated with different data types to represent solutions
- Binary strings : Genetic Algorithms
- Real-valued vectors : Evolution Strategies
- Finite state Machines: Evolutionary Programming
- LISP trees: Genetic Programming
- These differences are largely irrelevant, best strategy
 - choose representation to suit problem
 - choose variation operators to suit representation
- Selection operators only use fitness and so are independent of representation

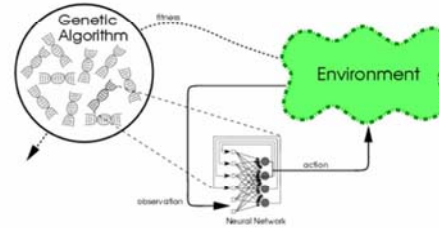
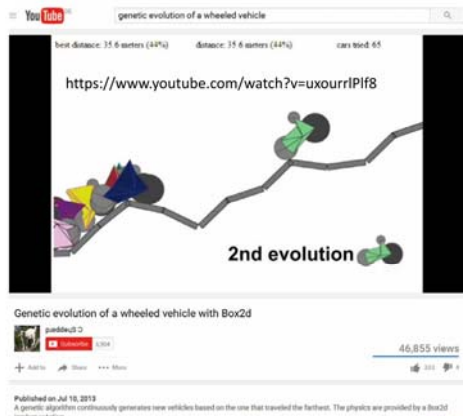
Algorithm 1 Fitness function

```

1: procedure FITNESS FUNCTION(weights[], List trainingSet)
2:   for all instances of trainingSet do
3:     for i = 1 to NumberOfClasses do
4:       for all attribute to MaxNumberAttributes do
5:         probability[i] *= NORMDISTRIBUTION(attribute + weights[attribute])
6:
7:       index ← INDEX OF MAX(probability[])
8:
9:       if index == CLASS OF(instance) then
10:        INCREMENT(fitness)
11:       else
12:        DECREMENT(fitness)
13:
14:   RETURN fitness
    
```

02 Neuro Evolution *)

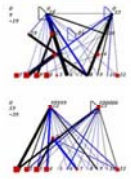
Note: this is ML - not to confuse with neural evolution!



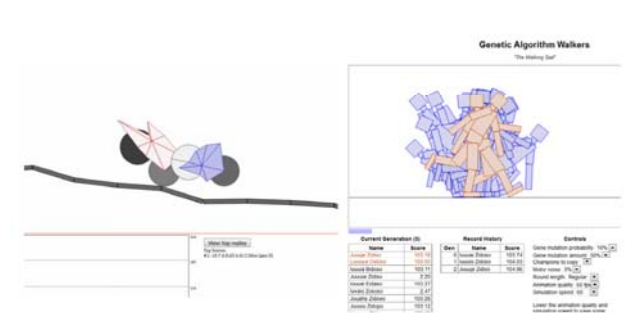
- Future Challenge: to utilize domain knowledge for problem solving
- Sometimes we have knowledge and sometimes random initial behavior is not acceptable
- Grand question: How can domain knowledge be utilized?
- by incorporating rules
- by learning from examples

03 Genetic Algorithms

- is a form of machine learning that uses evolutionary algorithms to train deep learning networks.
- method for optimizing neural network weights and topologies using evolutionary computation. It is particularly useful in sequential decision tasks that are partially observable (i.e. POMDP) and where the state and action spaces are large (or continuous).
- Application: Games, robotics, artificial life
- neuroevolution can be applied more widely than supervised learning algorithms



<http://nn.cs.utexas.edu/keyword?neuroevolution>

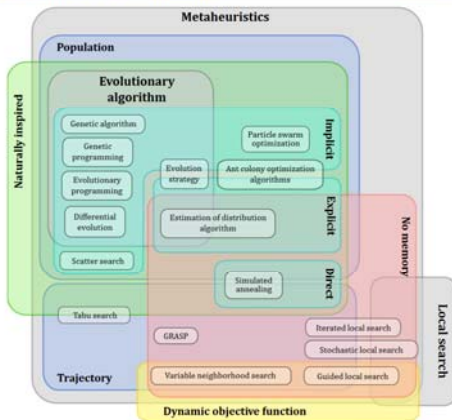


http://rednuht.org/genetic_cars_2/

http://rednuht.org/genetic_walkers/

- Similar to stochastic optimization
- Iteratively trying to improve a possibly large set of candidate solutions
- Few or no assumptions about the problem (need to know what is a good solution)
- Usually finds good rather than optimal solutions
- Adaptable by a number of adjustable parameters

Image Credit to
Johann Dréo,
Caner Candan -
Metaheuristics
classification
CC BY-SA 3.0
<https://commons.wikimedia.org/w/index.php?curid=16252087>



$$N - n^* \approx \sqrt{8\pi \cdot b^4 \cdot \ln(N^2)} \cdot \exp\left(\frac{n^*}{2b^2}\right)$$

- The trials are allocated to the **observed** best arm
- This 2-arm bandit can be generalized to a k-armed bandit, resulting in:
- A) Generalized corollary: The optimal strategy is to allocate an exponentially increasing n of trials to the **observed** best arm
- B) This links-up to Genetic Algorithms because: Minimizing expected losses from k-armed bandits \approx Minimizing expected losses while sampling from order $\log_2(k)$ schemata (=GA's allocate trials opt.)

- 1) What is Cancer: A biological introduction
- 2) The multistep process of cancer
- 3) Key Problems for cancer research
- 4) Overview of Machine Learning for cancer
- 4) Tumor Growth Modeling
- 5) Cellular Potts Model > Tumor Growth Simulation
- 6) Implementation of Tumor Growth Visualization
- 7) Summary and Open Problems

$K = 2 \rightarrow$ **Two-armed bandit problem:**

Arm 1: award μ_1 with variance σ_1^2

Arm2: award μ_2 with variance σ_2^2

$\mu_1 > \mu_2$

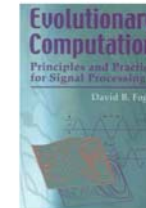
Question: Which arm (left/right) is which index 1, 2?



Can be used for motivation of the Schema Theorem by John Holland (1975): is widely taken to be the foundation for explanations of the power of genetic algorithms: low-order schemata with above-average fitness increase exponentially in successive generations.

Holland, J. H. 1975. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, U Michigan Press (as of 01.06.2016 49,320 citations I)

- Why would this be optimal for global optimization?
- Minimizing expected losses does not always correspond to maximizing potential gains.



- In Silico
- Differentiation
- Benign & Malignant Tumor Cells
- Proliferation
- Migration
- Tissue
- Adhesion
- Tumor Growth Modeling
- Agent Based Modeling
- Cellular Automaton (CA)
- Extra-Cellular Matrix (ECM)
- Cellular Potts Model (CPM)

- N = total number of trials

$$b = \frac{\sigma_1}{\mu_1 - \mu_2}$$

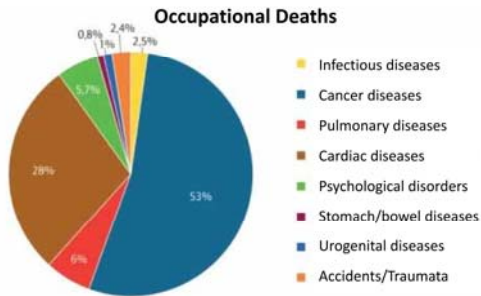
- Conclusion:
Expected loss is minimal if approximately:

$$n^* \approx b^2 \cdot \ln\left(\frac{N^2}{8\pi \cdot b^4 \cdot \ln(N^2)}\right)$$

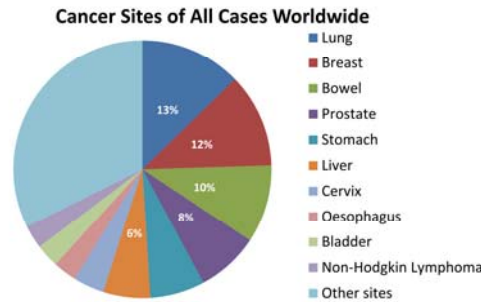
- Consequently, trials are allocated to the observed worst arm

04 Tumor Growth Simulation

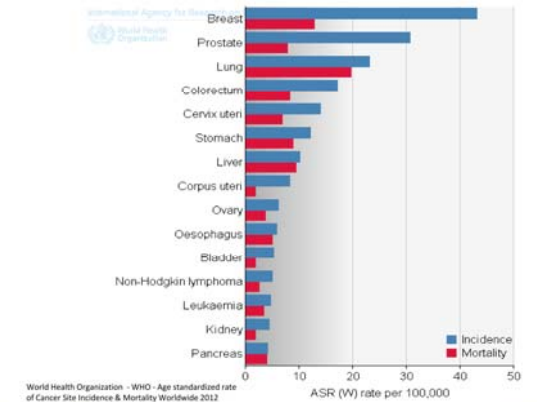
Part 1: Biology



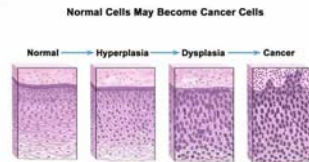
OGB - German Health News, apr 5th 2016 - Annually mortality causes in EU28 and other industrial countries



Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. GLOBOCAN 2013 v1.0: Cancer Incidence and Mortality Worldwide (IARC CancerBase No. 11) [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. Available from: <http://globocan.iarc.fr>, accessed on 16/01/2014.



- **Tumor / Neoplasm**
... Abnormal mass of tissue - cells divide more than they should
- **Cancer**
... group of diseases in which abnormal cells divide without control, can invade nearby tissues
- **Malignant**
... Cancerous: invasive, destroy nearby tissue, spread to other parts of the body
- **Benign / Non-malignant**
... normal (not cancerous): grow larger but do not spread
- **Hyperplasia / Dysplasia**
... increased number / abnormal form



NCI Dictionary of Cancer Terms - US Department of Health & Human Services, National Institutes of Health, National Cancer Institute.

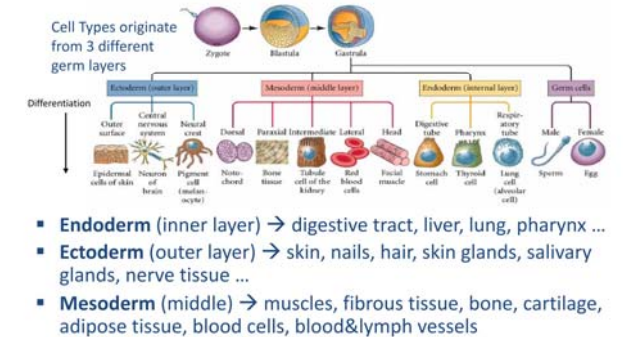
Most common Cancer types, based on their origin (primary manifestation):

- Skin
- Lung
- Breast
- Prostate
- Colon & rectum
- Uterus

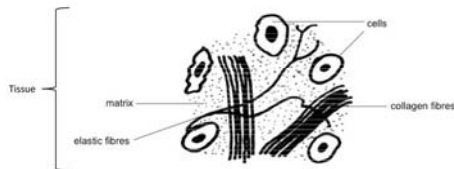


NIH NATIONAL CANCER INSTITUTE Surveillance, Epidemiology, and End Results Program

ICD-O-3 - The International Standard for the classification and nomenclature of histologies is the International Classification of Diseases for Oncology, 3rd Ed. NIH SEER Training Modules - US Department of Health & Human Services, National Institutes of Health, National Cancer Institute.



The three-germ-layers - <http://madmemories.blogspot.co.at> - 06/2015



- Tissue
- Extracellular Matrix (ECM)
- Cell → Organelles
- Microfibril → Fiber → Protein

Loose connective tissue - by Adrigola, Sunshineconelly, Lawison R. 2011

Histological Types: Hundreds of different cancers, summed up to 6 major categories:

- **Carcinoma** (epithelial tissue)
- **Sarcoma** (supportive/connective tissue)
- **Myeloma** (plasma/bone marrow cells)
- **Leukemia** (bone marrow → blood production)
- **Lymphoma** (lymphatic system)*
- **Mixed Types** (eg. Carcinosarcoma)

* Hodgkin/Non-Hodgkin lymphoma depending on presence of Reed-Sternberg cells

ICD-O-3 - The International Standard for the classification and nomenclature of histologies is the International Classification of Diseases for Oncology, 3rd Ed. NIH SEER Training Modules - US Department of Health & Human Services, National Institutes of Health, National Cancer Institute.

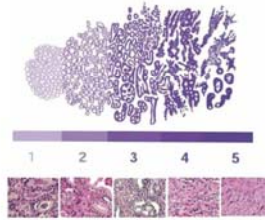
Nomenclature based on tissue type and malignancy/benignancy:

eg. **Adenoma** (benign) & **Adenocarcinoma** (malignant),
Fibroma & **Fibrosarcoma**
Neuroma & **Neuroblastoma**

ICD-O-3 - The International Standard for the classification and nomenclature of histologies is the International Classification of Diseases for Oncology, 3rd Ed. NIH SEER Training Modules - US Department of Health & Human Services, National Institutes of Health, National Cancer Institute.

Grades

- G1 (undetermined)
- G2 (well differentiated)
- G3 (poorly differentiated)
- G4 (undifferentiated)



Cancer type-specific grading

Gleason Scoring - prostate cancer - calculated from pattern 1-5
 → X, 1-6 (well diff.), 7 (moderately), 8-10 (poorly diff.)
 Nottingham system - breast cancer
 (based on tubule formation, nuclear grade, mitotic rate)

NCI - National Cancer Institute, at the National Institutes of Health - About Cancer - Prognosis, May 3rd 2013
 American Joint Committee on Cancer. AJCC Cancer Staging Manual. 7th ed. New York, NY: Springer; 2010.
 The Gleason grading system - Harnden P, et al. The Lancet Oncology, Volume 8, Issue 5, 431 - 439, 2007

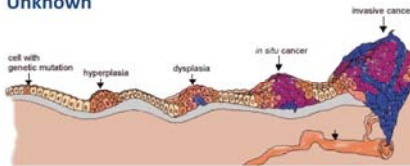
- Tumor location
- Cell type
- Tumor Size
- Spread to lymph nodes
- Spread to different parts of body
- Tumor grade = cell abnormality
 (proliferation rate, nuclear hyperchromasia, mitoses)

NCI - National Cancer Institute, at the National Institutes of Health - About Cancer - Diagnosis and Staging, March 9th 2015

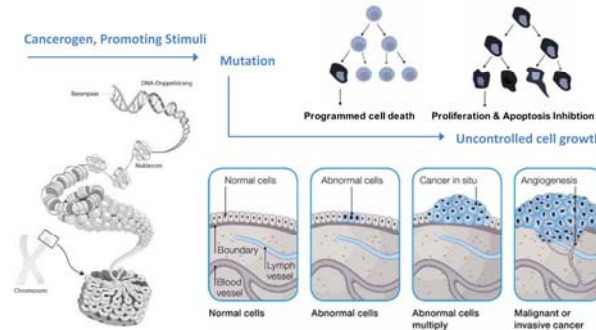
- TNM system (extent/number/metastasis)
 X,0,T1-4, N1-3, M1
 eg. T3N0M0 (large tumor, no cancer in nearby lymph nodes/tissue, not spread to distant body parts)
- Stage
 0 (carcinoma in situ)
 I-III (size and spread to nearby tissue)
 IV (metastasis to distant parts)

NCI - National Cancer Institute, at the National Institutes of Health - About Cancer - Diagnosis and Staging, March 9th 2015

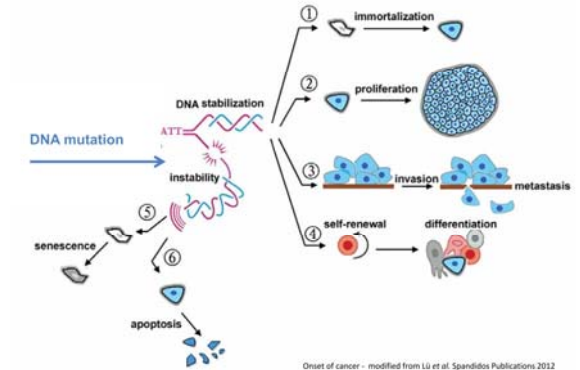
- In situ - abnormal cells present but not spread to nearby tissue
- Localized - cancer limited to origin, not spread
- Regional - cancer spread to nearby lymph nodes, tissues, organs
- Distant - cancer spread to distant parts of the body
- Unknown



NCI - National Cancer Institute, at the National Institutes of Health - About Cancer - Diagnosis and Staging, March 9th 2015
 Cancer progression - <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2910720/>

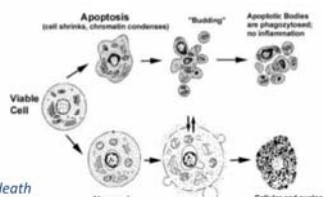


DNA structure - Krebsinformationsdienst, Deutsches Krebsforschungszentrum Jan 2016.
 S. Jorha et al. / Garak76, 2010 cell division - normal vs. cancer.
 Cancer cycle and cancerous cells - © 2015 Oncosera.

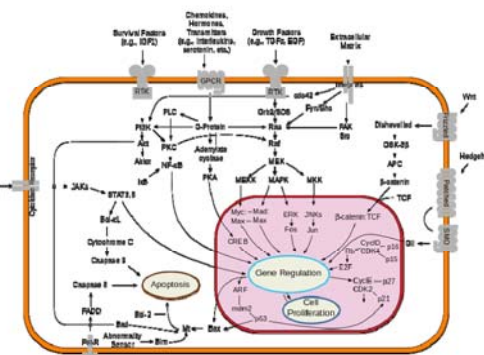


Onset of cancer - modified from Lu et al. Scandinos Publications 2012

- Differentiation
 ... cell changing to a more specialized cell type
- Proliferation
 ... growth: increase in cell number via cell division
- Mitosis
 ... cell division
- Apoptosis
 ... programmed cell death, blocked in cancer cells
- Necrosis
 ... unprogrammed / general cell death

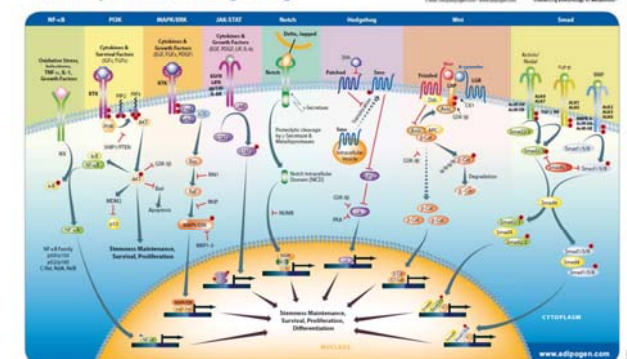


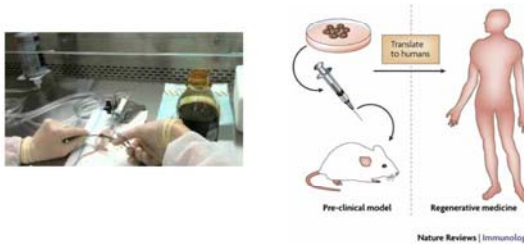
NCI Dictionary of Cancer Terms - US Department of Health & Human Services, National Institutes of Health, National Cancer Institute.
 Intro to Apoptosis - Genes A. Nov. 6th 2014, <http://de.scribd.com/doc/184848488/Intro-to-Apoptosis>



Signal transduction pathways - Commons, by Boho2 sept. 6th 2008

Pathways on Cancer Signaling





Humanized mice in translational biomedical research, Leonard D. Shultz, Fumihiko Ishikawa and Dale L. Greiner, Nature Reviews Immunology 7, 2007
Mouse xenograft surgery – www.youtube.com/watch?v=R2Wka7YhhAo



- Increased frequency in spontaneous formation of tumors
 - Reduced latency time
 - Tumor occurrence in additional tissues
 - Increased number of tumors
- Genotoxicity: direct DNA damage
 - Non-genotoxic: indirect damage on external genetic influence factors



GHS08
Health hazard

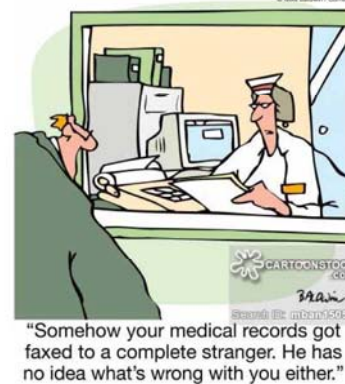
Carcinogenicity (H350, H351),
Germ cell mutagenicity (H340, H341),
Reproduction toxicity (H360, H361)

categories 1A/B, 2 (probability)

BG BAU – Berufsgenossenschaft der Bauwirtschaft, Berlin 2016, GBSAU – Gefahrstoff-Informationssystem

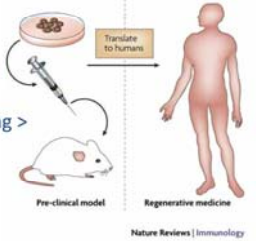


<https://www.youtube.com/watch?v=R2Wka7YhhAo>

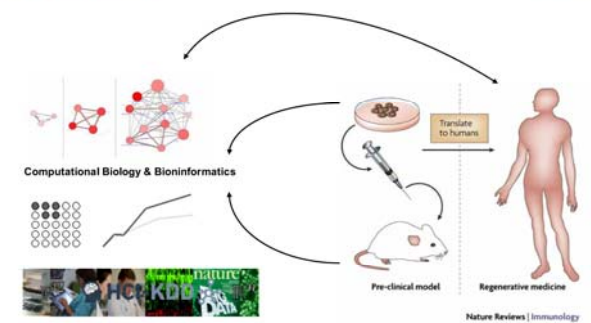


Toxicity / Carcinogenicity studies

- Daily administration of test substance to animal (oral, dermal, inhalative) –18-30 months (live-long for rodents)
- Chron. toxicity by repeated dosing > 12 months
- Histopathological changes (hyperplasia, atypia), rates of cell division



Humanized mice in translational biomedical research, Leonard D. Shultz, Fumihiko Ishikawa and Dale L. Greiner, Nature Reviews Immunology 7, 2007
REACH (EC) No. 1907/2006 – registration, evaluation, approval and limitation of chemical substances, updated (EU) 2015/850.
Regulation 440/2008 – agreement on test methods according regulation 1907/2006, last update 07.12.2016 – (EU) 2016/266.



Humanized mice in translational biomedical research, Leonard D. Shultz, Fumihiko Ishikawa and Dale L. Greiner, Nature Reviews Immunology 7, 2007
Jeanquartier et al. 2016

Part 2 Computational Modelling



Star Trek Voyager - tv series
Are Computers better doctors?
Duerr-Specht, M., Goebel, R. & Holzinger, A. 2015. Medicine and Health Care as a Data Problem: Will Computers become better medical doctors? In: Springer Lecture Notes in Computer Science LNCS 8700. pp. 21-40, doi:10.1007/978-3-319-16226-3_2

- ML in Genomics
 - such as DNA micro array analysis for cancer classification etc.
 - => for *identification & treatment*
- ML in image analysis
 - such as for classifying and/or differentiating benign from malignant samples etc.
 - => for *diagnosis & prognosis*
- ML in cancer research is growing rapidly
 - combination of molecular patterns and clinical data
 - deep text-mining offers new possibilities
 - etc.

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2.

- There are **different kinds of models** in biology, such as spatial ones, space free ones but also cell descriptive models based on density, or cell-based, or sub-cellular or molecular, (relating to their scale of phenomenon)

Szabó, A., & Merks, R. M. (2013). Cellular potts modeling of tumor growth, tumor invasion, and tumor evolution. Frontiers in oncology, 3

- Inter- and intracellular **dynamics**
- avoiding **hard-to-measure** variables
- Inflexible** models
- in silico* **complements in vivo**
- executable (cell) biology*
- reduce** animal experiments (resources)
- boost in silico** for awareness & breakthrough
- patient-personalized** prediction

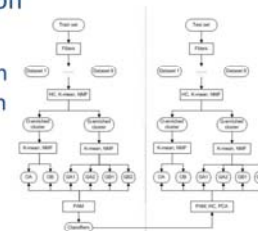


Edelman, L. B., Eddy, J. A. & Price, N. D. 2010. In silico models of cancer. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2, (4), 438-459, doi:10.1002/wsbm.75.

Fisher, J. & Henzinger, T. A. 2007. Executable cell biology. Nature biotechnology, 25, (11), 1239-1249, doi:10.1038/nbt1356.

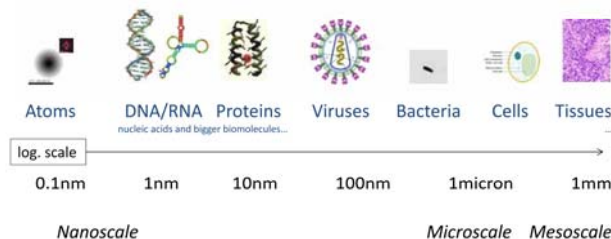
Example: Glioma Classification

- Using gene expression data
- Unsupervised ML approach on genome-wide gene expression profiles of 159 gliomas
- Model predicts
 - two major groups,
 - separated into six subtypes,
 - previously unrecognized prognostic groups within TCGA published data could be found

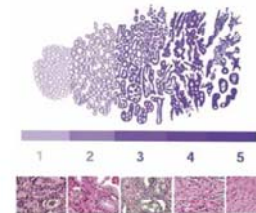


Alguo Li, Jennifer Walling, Susie Ahn, Yuri Kotliarov, Qin Su, M. Quezado, J. C. Oberholtzer, J. Park, J. C. Zenklusen, H. A. Fine: Unsupervised Analysis of Transcriptomic Profiles Reveals Six Glioma Subtypes, DOI: 10.1158/0008-5472.CAN-08-2100 Published 1 March 2009

Visualization Applications At Different Biological Scales



Images of



tissue



wet research

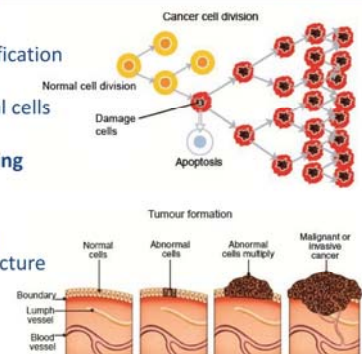
The Gleason grading system - Harnden P. et al. The Lancet Oncology, Volume 8, Issue 5, 411 - 419, 2007

Example: Modeling glioma tumor growth

- Using image data (MRI scans)
 - (g) initial tumor
 - (h) predicted tumor
- Learn** the parameters of a diffusion model
 - Using patient data
 - Preprocessing images
 - noise reduction, linear register and warp to standard coordinate system, reducing inhomogeneity, Intensity standardization, segmentation between grey and white matter...
 - Feature extraction
- => **Prediction** through classification & diffusion

Morris, M., Greiner, R., Sander, J., Murtha, A., & Schmidt, M. (2006). Learning a classification-based glioma growth model using MRI data. Journal of Computers, 1(7), 21-31.

- Tumor growth
- complex disease**: simplification & approximation
- differentiation** of normal cells
- excessive **proliferation**
- either **dormant** or **growing**
- critical mass**
 - growth stops
 - migration (metastasis)
- underlying network structure**
- environmental heterogeneities



Choe, S. C., Zhao, G., Zhao, Z., et al. (2011). Model for in vivo progression of tumors based on co-evolving cell population and vasculature. Scientific reports, 1.

- A tumor can be seen as *spatio-temporal* pattern formation
- Spatial & temporal data exist and can be used for improving existing simulation & analysis tools
- Several attempts have been made to *model* and *predict* malignant tumor

Jeanquartier, F., Jean-Quartier, C., Schreck, T., Cemernek, D. & Holzinger, A.: Integrating Open Data on Cancer in Support to Tumor Growth Analysis. ITBAM. Springer Lecture Notes in Computer Science. LNCS 9832, 2016.

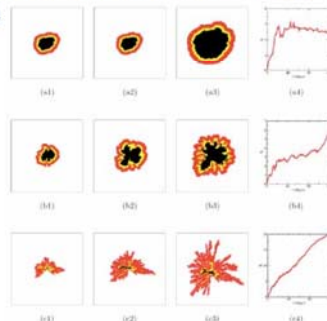
Moreira, J., & Deutsch, A. (2002). Cellular automaton models of tumor development: a critical review. Advances in Complex Systems, 5, 247-267

- Continuum vs. Discrete/Agent-based Modeling

Continuum	Discrete
continuously distributed variables	discrete entities in discrete time intervals
interactions between factors representing several effects of physiological/biochemical events	interactions in a single space representation
f.i.: simulating population dynamics, combinatoric effects of several nutrient availability and other parameters etc.	f.i.: simulating agent dynamics, probabilistics of each time step, a small number of individuals

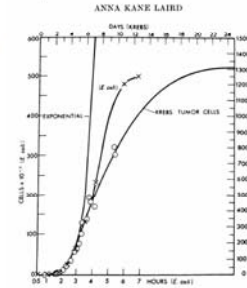
Edelman, L. B., Eddy, J. A., & Price, N. D. (2010). In silico models of cancer.

- Simulated Growth of Solid Tumors in Confined Heterogeneous Environment



Jiao, Y., & Torquato, S. (2011). Emergent behaviors from a cellular automaton model for invasive tumor growth in heterogeneous microenvironments. PLoS Comput Biol, 7(12)

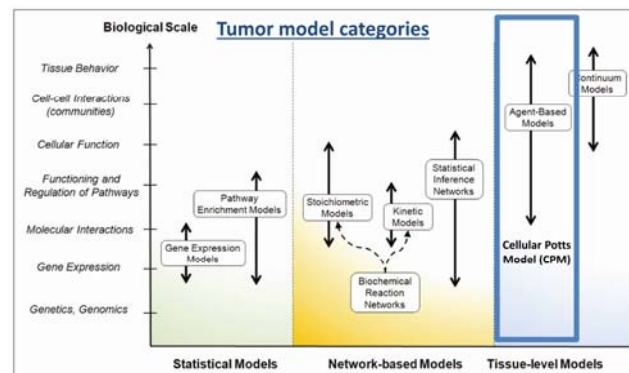
- Tumor growth kinetics follow simple laws
- Mathematical models exist f.i. Gompertz or power law
- No universal law
- Prediction rate low and/or distinct



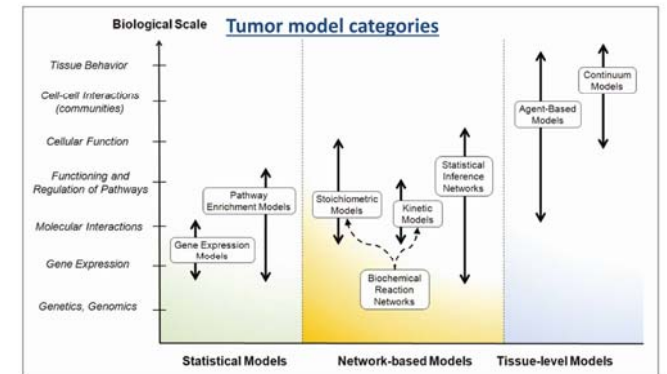
Benzekry, S., Lamont, C., Beheshti, et al. (2014). Classical mathematical models for description and prediction of experimental tumor growth.

- Cellular Automaton (CA) approach to modeling biological cells
- CAs are:
 - Discrete
 - Abstraction of a system
 - Computational
 - At each time, each cell instantiates one state of a finite set of states
- CA for tumor growth with rules:
 - Cell division, movement, change or not change state...
- "On-lattice" modelling (see next slide)

Moreira, J., & Deutsch, A. (2002). Cellular automaton models of tumor development: a critical review. Advances in Complex Systems, 5(02n03), 247-267.



Edelman, L. B., Eddy, J. A., & Price, N. D. (2010). In silico models of cancer.



Edelman, L. B., Eddy, J. A., & Price, N. D. (2010). In silico models of cancer.

- In abstract algebra it is a fundamental algebraic structure, consisting of a partially ordered set in which every two elements have a unique supremum (join) and a unique infimum (meet). An example is given by the natural numbers, partially ordered by divisibility, for which the unique supremum is the least common multiple and the unique infimum is the greatest common divisor.
- In geometry a lattice in \mathbb{R}^n is a subgroup of \mathbb{R}^n , which is isomorphic to \mathbb{Z}^n , and which spans the real vector space \mathbb{R}^n , i.e. for any basis of \mathbb{R}^n the subgroup of all linear combinations with integer coefficients of the basis vectors forms a lattice. A lattice may be viewed as a regular tiling of a space by a primitive cell.



CPM

- Model for cell sorting
- Describes cell-cell interaction, motion, rearrangement, pressure inside tissue
- Suitable for pathol. developmental mechanisms in cancer
- Cell-based method on the lattice
- 2D lattice represents tissue
- Collection of particles to represent the cell
- Each cell is represented as an object with a possible adhesive state, spatially extended
- cells are composed of adjacent lattice sites with similar id nr.
- system tends to minimize overall surface energy (energy per unit of area)

Szabó, A., & Merks, R. M. (2013). Cellular potts modeling of tumor growth, tumor invasion, and tumor evolution. Frontiers in oncology, 3

CPM originally developed by Graner & Glazier 1992

$$p(\sigma_{i,j} \rightarrow \sigma_{i',j'}) = \begin{cases} e^{-\frac{\Delta H}{T}} & \text{if } \Delta H > 0; \\ 1 & \text{if } \Delta H \leq 0; \end{cases}$$

Probability of accepting/rejecting a spin copy

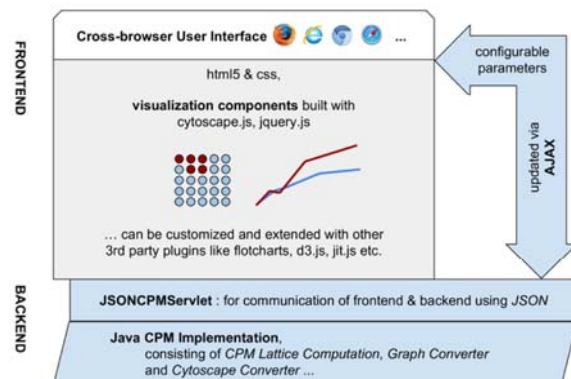
Szabó, A., & Merks, R. M. (2013). Cellular potts modeling of tumor growth, tumor invasion, and tumor evolution. *Frontiers in oncology*, 3

Note that this is an approach with Reals – not discrete – but we are working on discrete multi-agent approaches in the future!

Our idea:

- Reducing animal experiments
- Visualizing tumor dynamics towards better understanding
- Easy-to-use
- Easy-to-extend
- Implementation of Cellular Potts Model
 - visualized with cytoscape.js (web application)
 - other client rendering frameworks
 - ... based on network visualization in biology
- Support biologists and clinical scientists
 - ultimate goal

Jeanquartier, F., Jean-Quartier, C., Cemernek, D. & Holzinger, A. In Silico Modeling For Tumor Growth Visualization. BMC In revision.
 Jeanquartier, F., Jean-Quartier, C., Schreck, T., Cemernek, D. & Holzinger, A. Integrating Open Data on Cancer in Support to Tumor Growth Analysis. *Information Technology in Bio- and Medical Informatics*, LNCS 9832, 2016.
 Jeanquartier, F., Jean-Quartier, C., Cemernek, D. & Holzinger, A. Tumor Growth Simulation Profiling. *LNCS* 9832, 2016.



CPM originally developed by Graner & Glazier 1992

$$H = J \sum_{i,j} (1 - \delta_{\sigma_i, \sigma_j}) + \lambda \sum_{\sigma} (v(\sigma) - V_t(\sigma))^2$$

$$\text{Kronecker delta } \delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

σ spin of a cell
 J surface energies between spins (adhesion)
 λ cellular constraint, function of elasticity
 $v(\sigma)$ area/volume of a cell
 $V_t(\sigma)$ target area for cells of type

Szabó, A., & Merks, R. M. (2013). Cellular potts modeling of tumor growth, tumor invasion, and tumor evolution. *Frontiers in oncology*, 3

```
for Number of MCS do
  for Appropriate number of samples (substeps) do
    Calculate the Hamiltonian in current state, H0;
    Select a lattice site, i, from the domain at random;
    Select a neighbour, j, of this site at random;
    Change config so that site i refers to same cell as site j
    (if not ecm)
    Calculate the Hamiltonian in new configuration, H1;
    if DeltaH = H1 - H0 <= 0, then
      Accept change;
    else
      Evaluate p = exp((-DeltaH)/(T));
      Sample a number u from U(0, 1);
      if p < u, then
        Accept change;
      end
    end
    If change is rejected, then restore original configuration.
  end
end
```

J. M. Osborne. (2015). Multiscale Model of Colorectal Cancer Using the Cellular Potts Framework. *Cancer Informatics*, 14(Suppl. 4), p83-93

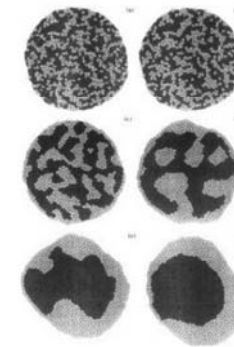
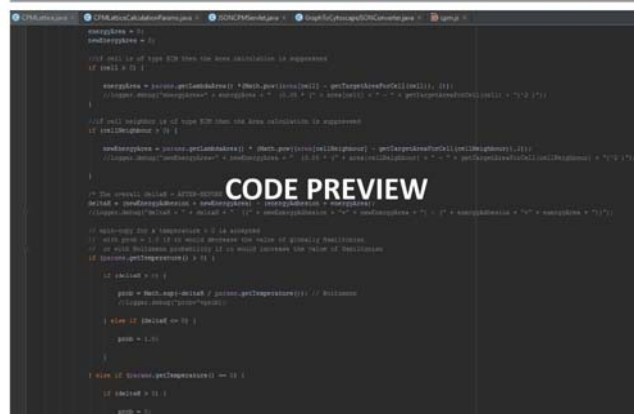


Image of a cell sorting time series

- initial: random assigned cell types
- each step represents a growing number of Monte Carlo Step (MCS)
- figure shows pattern

Graner, F., & Glazier, J. A. (1992). Simulation of biological cell sorting using a two-dimensional extended Potts model. *Physical review letters*, 69(13), 2013.

CPM implementations already exist:

- CompuCell3d
 - Tissue Simulation Toolkit
- However
- though „community-driven“, not maintained
 - context-specific
 - static
 - lack of re-usability
 - hard to be combined with visualization libraries
 - no web implementation
 - not useful for interactive visualization

Jeanquartier, F., Jean-Quartier, C., Cemernek, D. & Holzinger, A. In Silico Modeling For Tumor Growth Visualization. BMC In revision.
 Szabó, A., & Merks, R. M. (2013). Cellular potts modeling of tumor growth, tumor invasion, and tumor evolution. *Frontiers in oncology*, 3.

Implementation of Tumor growth visualization

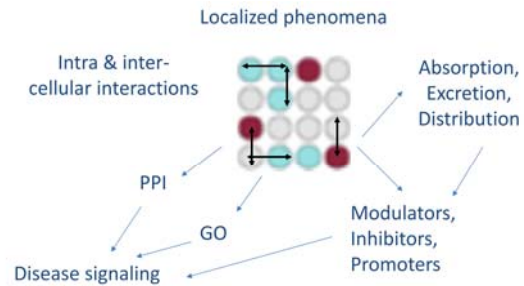
- "cpm-cytoscape" already on GitHub:
<https://github.com/davcem/cpm-cytoscape>
- and available as online DEMO:
<http://styx.cgv.tugraz.at:8080/cpm-cytoscape/>



Jeanquartier, F., Jean-Quartier, C., Cemernek, D. & Holzinger, A. In Silico Modeling For Tumor Growth Visualization. BMC, Manuscript in rev

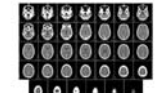
Nodes as Cellular bricks

→ Compartmental states:



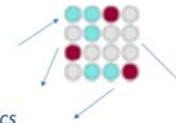
Hot topics:

- *Learning from image data for Initialization*
 - image preprocessing & feature extraction
 - comparison, refinement & optimization
- *ML for tumor growth profiles and model validation*
- *On using open tumor growth data for ML*
 - Histologic data
 - Drug targeting data etc
- *On multi-scale trends in cancer modelling*
 - Compare results of different models
 - Link between different scales
 - Combining microscopic characteristics
 - with macroscopic parameters
- *Sensitivity plots for tumor modelling*
- *Trajectory visualization of tumor dynamics*



initialize random or upload image

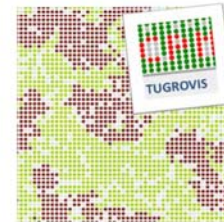
Choose a profile: brain cancer type II



Thank you!

Sample Questions (1)

- What is the difference between tumor and cancer?
- What does the term differentiation in biological context stand for? Give an example.
- What means in vivo, in vitro and in silico?
- What types of computational tumor growth models exist?
- What is a cellular automaton?



In Silico Modeling for Tumor Growth Visualization

<http://styx.cgvtugraz.at:8080/cpm-cytoscape/>

Appendix

Remember: Many problems in health informatics are hard

- **P:** algorithm can solve the problem in polynomial time (worst-case running-time for problem size n is less than $F(n)$)
- **NP:** problem can be solved and any solution can be verified within polynomial time ($P \subseteq NP$)
- **NP-complete:** problem belongs to class NP and any other problem in NP can be reduced to this problem
- **NP-hard:** problem is at least as hard as any other problem in NP-complete but solution cannot necessarily be verified within polynomial time

