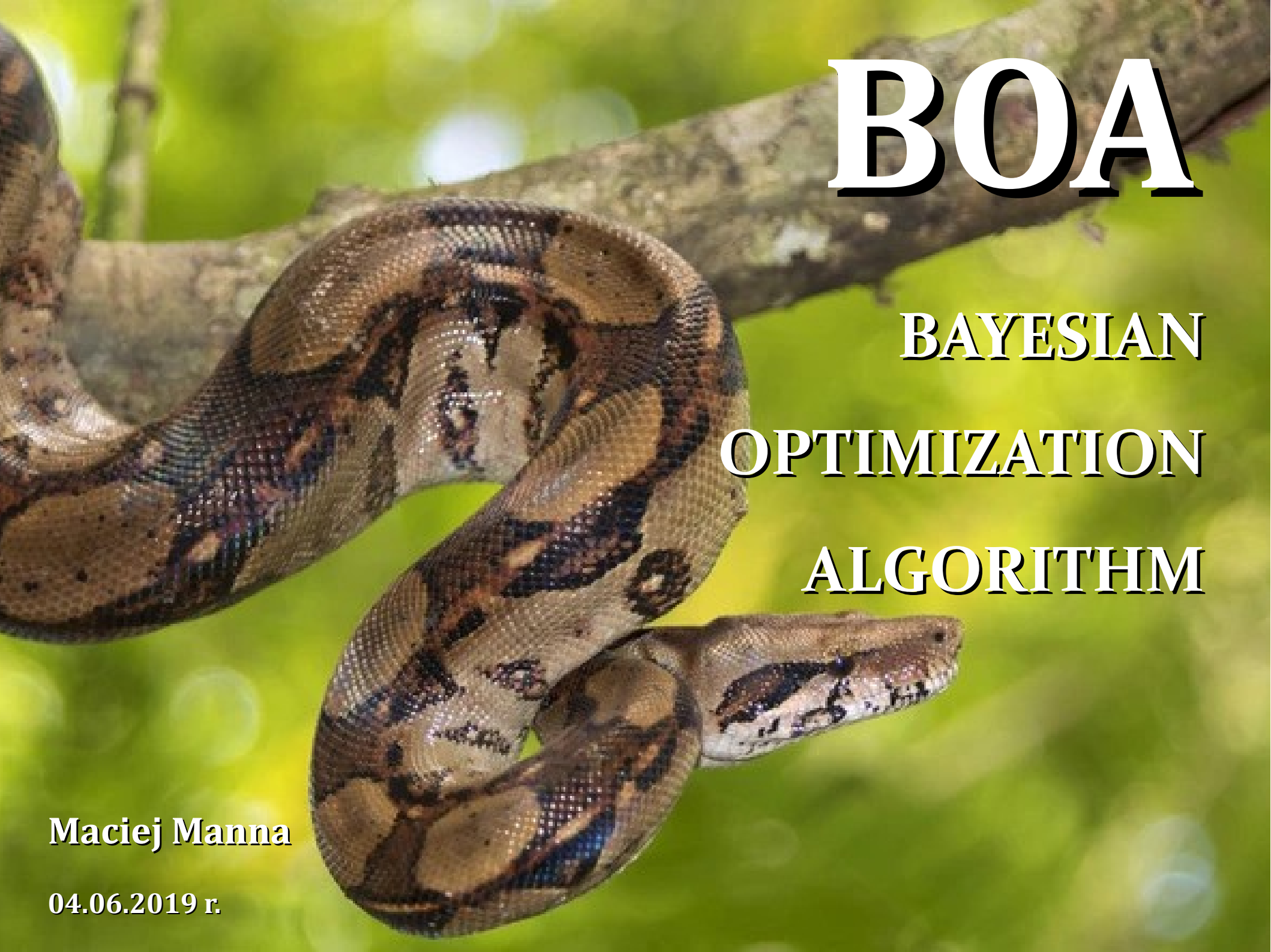


BOA

BAYESIAN OPTIMIZATION ALGORITHM

Maciej Manna

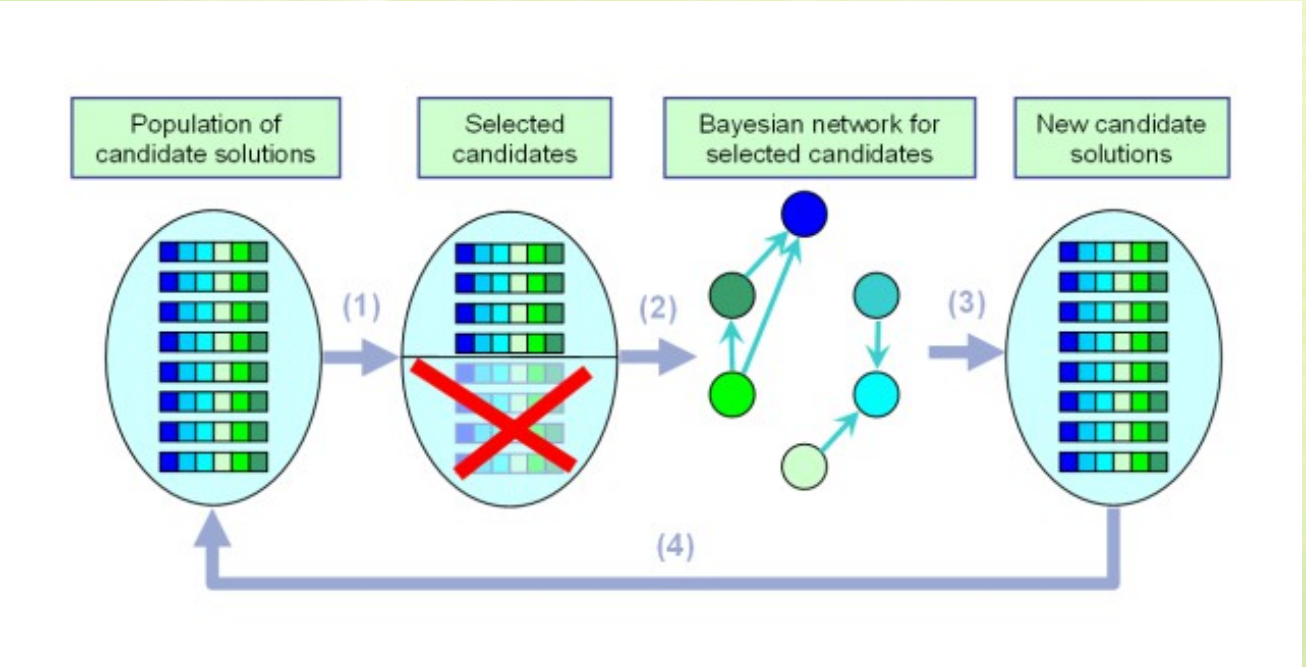
04.06.2019 r.

Uwagi wstępne – BOA vs BO

- **Bayesian Optimization**
  vs
  Bayesian Optimization Algorithm

BOA – Ogólny zarys

- Algorytm autorstwa Martina Pelikana (1998)
- "[...] oparty na koncepcji algorytmów genetycznych (GA)" oraz "wykorzystujący technikę sieci bayesowskich (BN) do modelowania zależności wyższych rzędów w obrębie danych".
- W uproszczeniu:



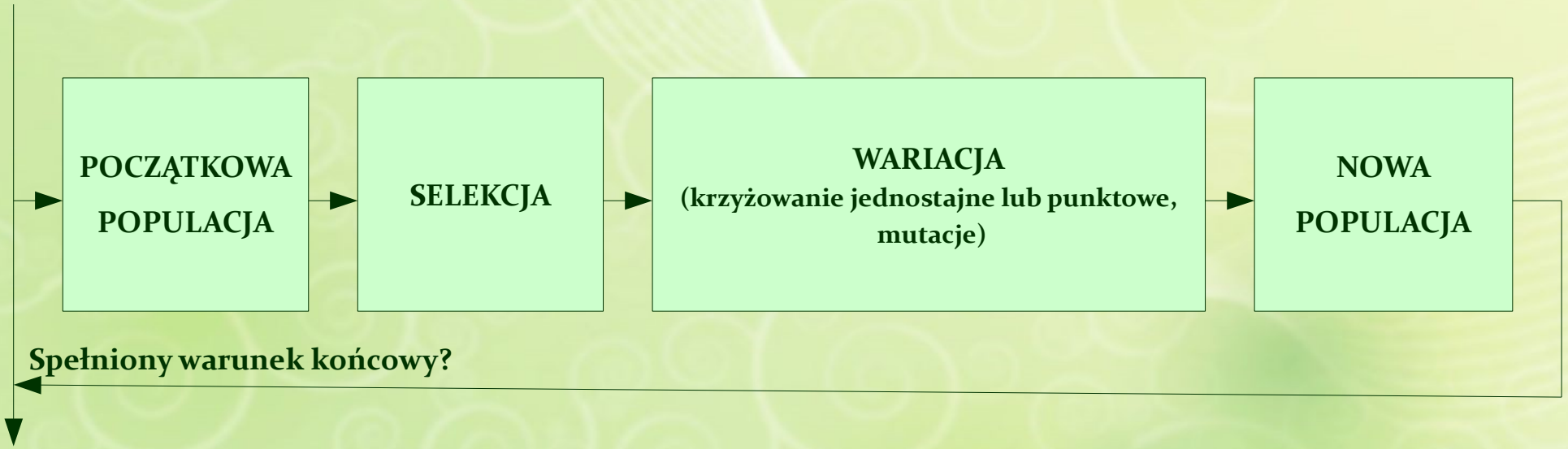
Plan prezentacji

- Wprowadzenie
- Algorytmy genetyczne (GA)
- Algorytmy genetyczne budujące model probabilistyczny (PMBGA)
- Algorytm optymizacji bayesowskiej (BOA)
- Warianty i alternatywy
- Podsumowanie

ALGORYTMY GENETYCZNE

GA - przypomnienie

- **Black-Box optimization…**
- **Ogólny schemat:**



GA – podstawowa terminologia

- **Populacja składa się z chromosomów (rozwiązania), które są ciągami pojedynczych genów (zmiennych)**

$$\boldsymbol{x} = (x_1,\ldots,x_n),$$

które należą do wybranego alfabetu (zazwyczaj binarne, {0,1})

- **W kontekście probabilistycznym, mozemy je rozważać jako wektory i zmienne losowe:**

$$X = (X_1,\ldots,X_n)$$

Twierdzenie o schematach

- Schematy – np.: (* 1 1 * * 0)
  – rząd schematu – liczba ustalonych pozycji w schemacie (dla ww. przykładu rząd wynosi 3)
  – długość definiująca – odległość między skrajnymi ustalonymi pozycjami (dla ww. przykładu – 4)
- Twierdzenie o schematach [Holland] – niskiego rzędu schematy z lepszym dopasowaniem średnim przyrastają wykładniczo w kolejnych pokoleniach
- Łączenie i niszczenie rozwiązań częściowych (zniszczenie jest bardziej prawdopodobne dla schematów o większej długości definiującej; sprzyja mu krzyżowanie jednostajne)

Hipoteza bloków budujących

- **Próby wyjaśnenia skuteczności i sposobu działania GA**
- **Hipoteza bloków budujących (BBH):**
  GA realizuje adaptację przez pośrednie i efektywne wykorzystanie abstrakcyjnych mechanizmów adaptacji realizowanych poprzez rekombinację "bloków budujących", czyli schematów niskiego rzędu, o niewielkiej długości definiującej oraz wysokim dopasowaniu. Takie schematy są próbkowane, zachowywane i łączone z rozwiązaniami o potencjalnie wyższym dopasowaniu w procesie selekcji. W ten sposób, dzięki przetwarzaniu poszczególnych bloków budujących dochodzi do redukcji złożoności zadania. Innymi słowy, łącząc ciągi o wysokiej częstości pojawiania się, tworzymy coraz lepsze rozwiązania spośród najlepszych rozwiązań cząstkowych w poprzednich próbkach.
- **Krytyka BBH (niespójność teorii; większa skuteczność GA, które wykorzystują krzyzowanie jednostajne zamiast punktowego)**

Jak poprawić skuteczność GA?

- Zakładamy, że BBH dobrze opisuje mechanizm odpowiedzialny za skuteczność GA
- Jednym z głównych problemów z wydajnością i stabilnością GA jest niszczenie już powstałych rozwiązań częściowych
- Jak temu zapobiec?
- Rozwiązania doraźnie – znając strukturę rozwiązywanego problemu, tworzenie specjalnie dopasowanych operacji rekombinacji, które mają mniejsze szanse na rozrywanie i niszczenie powstałych rozwiązań częściowych
- Czy możemy stworzyć taką operację rekombinacji, która będzie to robić automatycznie?

ALGORYTMY GENETYCZNE BUDUJĄCE MODEL PROBABILISTYCZNY

PMBGA - podstawy

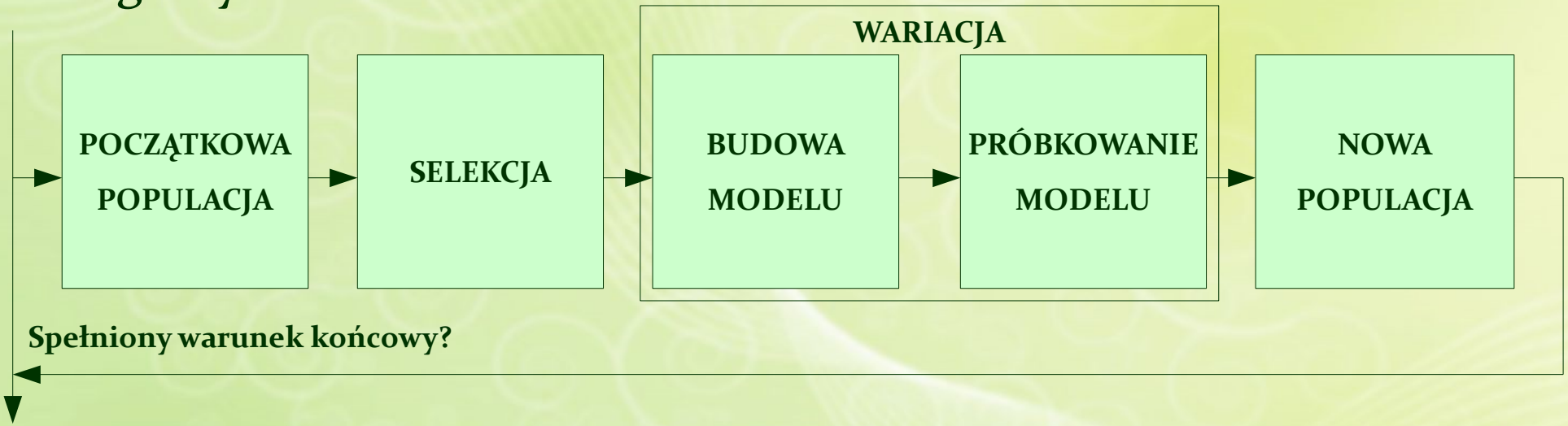
- **Model probabilistyczny "ukryty" w GA**
- **Możemy taki model *explicite* zdefiniować lub tworzyć na bieżąco – – przed utworzeniem nowej populacji budujemy model, który opisuje cały zbiór obiecujących rozwiązań (po selekcji)**
- **Model probabilistyczny – łączny rozkład prawdopodobieństwa dla wektora losowego reprezującego rozwiązanie problemu:**

$$p(X) = \ldots$$

- **Ta idea stoi u podstaw algorytmów typu PMBGA (*Probabilistic Model-Building GAs*), nazywane również EDA (*Estimation of Distribution Algorithms*), zaliczane do grupy algorytmów ewolucyjnych**

PMBGA - schemat

- **Ogólny schemat:**



- **Dwa istotne elementy definiowane przez konkretne algorytmy:**
  – *budowa modelu* (może być ustalony odgórnie lub być dynamicznie tworzony w trakcie działania algorytmu, wówczas musimy ustalić miarę, wg której oceniamy jakość modelu oraz strategię poszukiwania nowych modeli)
  – *próbkowanie modelu* (tworzenie rozwiązań nowej populacji)

UMDA

- ***Univariate Marginal Distribution Algorithm*** **(UMDA)**
- **Model estymujący jedynie rozkłady brzegowe poszczególnych zmiennych losowych składających sie na rozwiązania:**

$$p(X) = \prod_{i=0}^{n-1} p(X_i),$$

- **Zalety i wady:**
  (+) bardzo prosty model, który zachowuje rozwiązania częściowe pierwszego rzędu
  (-) nie ma możliwości pozyskania informacji o zależnościach między zmiennymi
- **W praktyce, zachowuje sie tak samo jak GA z krzyżowaniem jednostajnym**

BMDA

- ***Bivariate Marginal Distribution Algorithm*** **(BMDA)**
- **Model estymujący oprócz rozkładów brzegowych pojedynczych zmiennych losowych, również rozkłady warunkowe dla niektórych par zmiennych (wybranych na podstawie testu statystycznego wykrywającego niezależnośc zmiennych, np. *chi-kwadrat*):**

$$p(X) = \prod_{j=0}^{r} p(X_{i_j}) \prod_{j=r+1}^{n-1} p(X_{i_j}|X_{e_j}),$$

- **Zalety i wady:**
  (+) zachowuje również niektóre rozwiązania rzędu drugiego
  (-) nie ma możliwości uchwycenia bardziej złożonych schematów
- **W praktyce, zachowuje sie podobnie do GA z krzyżowaniem jednopunktowym (zależnie od długości definiującej schematów)**

FDA

- ***Factorized Distribution Algorithm*** **(FDA)**
- **Model wykorzystujący rozkłady brzegowe i warunkowe otrzymane przez (dokładną lub przybliżoną) faktoryzacje znanego rozkładu**

$$p(X) = \prod_{j=0}^{l-1} p(X_{i_j} | \Pi_{X_{i_j}}),$$

- **Zalety i wady:**
  (+) Jest w stanie w pełni zachowywać częściowe rozwiązania dla odpowiedniej klasy problemów (*addytywnie rozkładalnych*)
  (-) Faktoryzacja rozkładu wymaga albo pełnej wiedzy na temat problemu lub próby jej przybliżonego wykonania (co jest samo w sobie bardzo złożonym i trudnym zadaniem)
- **Ma znaczenie teoretyczne (idealne rozwiązanie, rzadko mozliwe)**

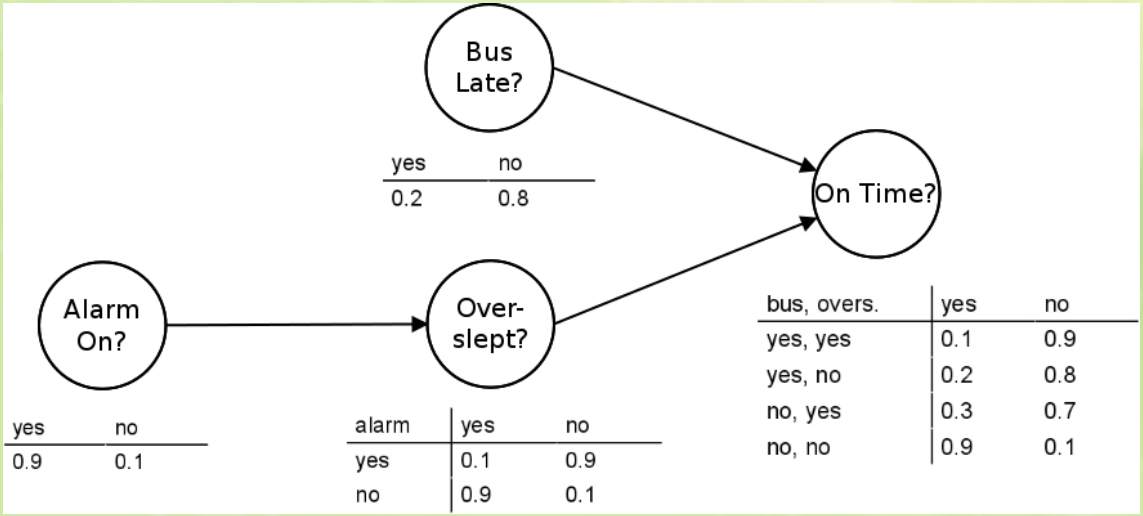
Dynamiczna struktura modelu

- Poprzednie rozwiązania stosują z góry ustaloną strukturę modelu
- Czy mozliwe jest stworzenie modelu, którego struktura naturalnie rozwija sie wraz z kolejnymi pokoleniami, coraz bardziej dopasowując się do uaktualnianego zbioru obiecujących rozwiązań?
- Czy możliwe jest połączenie zalet powyższych rozwiązań, w szczególności FDA, tj. możliwość uchwycenia relacji wyższych rzędów między zmiennymi oraz wykorzystanie dostępnej wiedzy odnośnie rozważanego problemu, ale gdzie posiadanie takiej uprzedniej wiedzy nie jest wymagane?

ALGORYTM OPTYMIZACJI BAYESOWSKIEJ

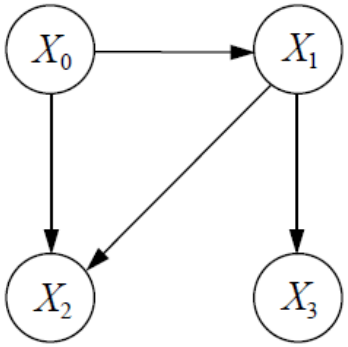
Sieci bayesowskie

- **Sieci bayesowskie (BN) – grafowa reprezentacja zależności pomiędzy zmiennymi losowymi**
- **Konkretnie: skierowany graf acykliczny (DAG), którego wierzchołki reprezentują poszczególne zmienne losowe, a krawędzie zależności między nimi (wyrażane przy pomocy odpowiednich prawdopodobieństw warunkowych)**



Sieci bayesowskie – c.d.

- **Każda sieć bayesowska definiuje pewnien model probabilistyczny:**

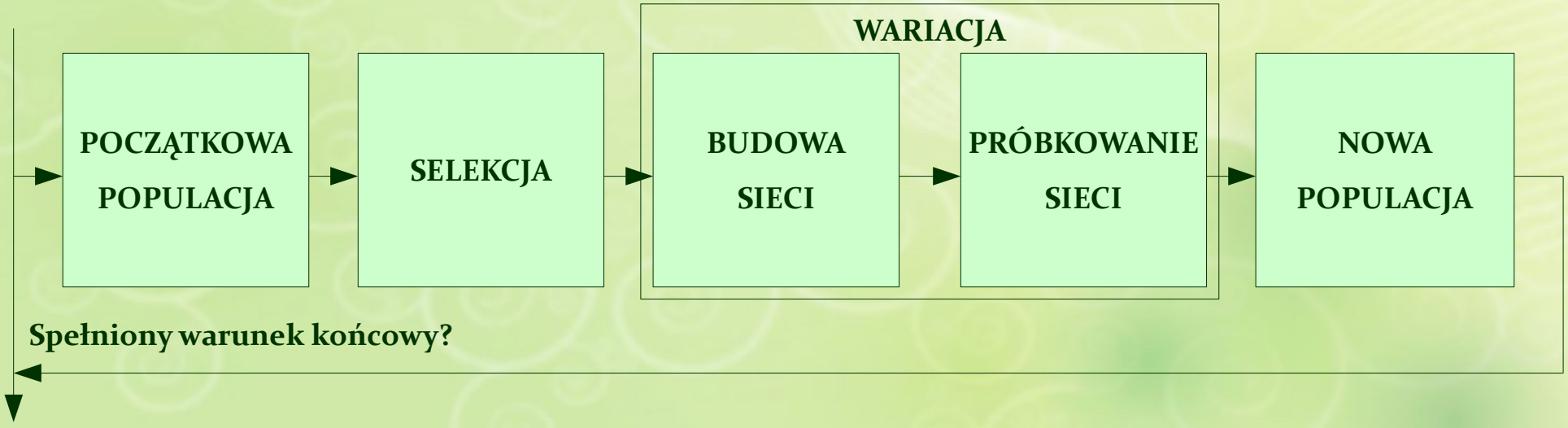


$$p(X) = p(X_0).p(X_1|X_0).p(X_2|X_0, X_1).p(X_3|X_1)$$

- **Prawdopodobieństwo danej zmiennej losowej jest zależne jedynie od zmiennych, które są jej rodzicami w grafie, a niezależna od pozostałych (tzw. *lokalna własność Markowa*)**
- **Powyższa własność istotnie ułatwia budowanie i używanie z sieci bayesowskich, ale uniemożliwia modelowanie cyklicznych relacji między zmiennymi**

BOA - schemat

- Sieci bayesowskie są bardzo dobrym kandydatem do reprezentowania przyrastającej wiedzy na temat wzajemnych relacji pomiędzy zmiennymi należącymi do obiecujących rozwiązań
- Algorytm wykorzystujący to rozwiązanie – algorytm optymizacji bayesowskiej (BOA, *Bayesian Optimization Algorithm*):



Budowa sieci bayesowskiej

- **Chcąc używać sieci bayesowskich do reprezentowania zmieniającego się w trakcie działania algorytmu modelu, musimy rozwiązać dwie kwestie:**

  (1) Jaką metryką będziemy rozstrzygać, która z proponowanych sieci lepiej modeluje bieżące dane?

  (2) W jaki sposób (tj. przy pomocy jakiego algorytmu) będziemy przeszukiwać przestrzeń wszystkich możliwych sieci w poszukiwaniu lepszej?

Metryki oceniające jakość BN

- **W kontekście zastosowania BN w BOA, stosowane są dwa typy metryk do oceny jak dobrze sieć modeluje bieżące dane:**
  (1) *bayesowskie* – pozwalają łączyć wiedzę a priori dotyczącą problemu (pozyskaną z zewnątrz lub podczas wcześniejszych iteracji algorytmu) z danymi statystycznymi pochodzącymi z obecnego zbioru obiecujących rozwiązań
  (2) oparte na *minimalnej długości opisu* (MDL) – preferująca sieci prostsze, których struktura (połączenia) i dane (tabele prawdopodobieństw) zajmują mniej miejsca

- **Metryki bayesowskie sa podatne na szum i wymagają pewnych ograniczeń (np maksymalnej ilości krawędzi wchodzących do wierzchołka), a MDL preferują modele często zbyt proste, które wymagają dużej populacji rozwiązań, aby działać skutecznie**

Bayesowska metryka Dirichleta

- **Przykład metryki bayesowskiej – bayesowska metryka Dirichleta (BD, *Bayesian Dirichlet Metric*), która pozwala łączyć uprzednie informacje o problemie z bieżącymi danymi:**

$$p(D,B|\xi) = p(B|\xi) \prod_{i=0}^{n-1} \prod_{\pi_{X_i}} \frac{m'(\pi_{X_i})!}{(m'(\pi_{X_i}) + m(\pi_{X_i}))!} \prod_{x_i} \frac{(m'(x_i,\pi_{X_i}) + m(x_i,\pi_{X_i}))!}{m'(x_i,\pi_{X_i})!},$$

prawdopodobieństwo a priori sieci *B*
(pozwala uwzględnić w ocenie fakt na ile nowa sieć przypomina poprzednią)

Zlicza przypadki, w których dla $X_i = x_i$ również zachodzi podobna równość dla wszystkich rodziców X_i w sieci *B*
(m'() wyraża podobną wartość ale odnoszącą się do wiedzy a priori o takiej współzależności)

Przeszukiwanie przestrzeni BN

- W jaki sposób, mając sieć z poprzedniej iteracji, poszukiwać nowej, lepszej?
- W ogólnym przypadku znalezienie optymalnej sieci względem pewnej metryki, to problem NP-zupełny
- W praktyce stosuje się algorytmy zachłanne, które stosują proste operacje na sieci (dodanie krawędzi; opcjonalnie również usunięcie oraz odwrócenie krawędzi) i wybierają rozwiązanie o najlepszym wyniku metryki w każdym kroku
- Często stosuje się górne ograniczenie dla ilości krawędzi wchodzących do wierzchołka (k), aby otrzymać dopuszczalną w praktyce złożoność takiego algorytmu przeszukiwania

Przeszukiwanie przestrzeni BN – c.d.

- Na początku algorytmu sieć może być pusta (tj. bez krawędzi), jeśli nie mamy żadnej uprzedniej wiedzy o zależnościach między zmiennymi w danym problemie – wtedy algorytm buduje ją od podstaw; można też rozpocząć algorytm z przygotowaną siecią przy użyciu zewnętrznej wiedzy dotyczącej problemu – wówczas algorytm to uwzględnia, ale jeśli dane tego wymagają, dokonuje zmian.

- Uwagi:
  – dla $k = 0$, sieć pozostaje pusta i BOA redukuje się do UMDA
  – dla $k = 1$, tzn. każdy wieżchołek może mieć co najwyżej jedną krawędź wchodzącą, przy użyciu odpowiedniej metryki BOA działa w podobny sposób jak BMDA (dla $k = 1$ problem poszukiwania sieci optymizującej metrykę jest wielomianowy)

Próbkowanie sieci bayesowskiej

- Struktura sieci bayesowskiej pozwala na bardzo proste tworzenie nowych rozwiązań, które realizują reprezentowany model
- Ponieważ sieć jest reprezentowana przez DAG, można jej wierzchołki (tj. zmienne losowe) posortować topologicznie
- Przechodząc zmienne w takiej kolejności – dzięki lokalnej własności Markowa BN – mamy gwarancję, że każda zmienna jest niezależna od tych, które po niej następują
- Każdą zmienną możemy więc kolejno spróbkować i wykorzystać te wartości w kolejnych, aż otrzymamy całe nowe rozwiązanie

Próba oceny

- Pomysł oparty na argumentacji związanej z kontrowersyjną hipotezą (BBH), choć to nie przesądza o jego wartości
- W głównej mierze wykorzystywany przez autora (i nie tylko) jako rozwiązanie pozwalające sprostać pewnym teoretycznym problemom, z którymi trudności mają GA (tzw. *fully deceptive subfunctions*)
- Nie wydaje się żeby to rozwiązanie zyskało popularność w rozwiązywaniu proktycznych problemów
- Rozwinęta wersja algorytmu (hBOA) może być wg autora wykorzystana do skutecznego rozwiązania pewnych problemów (*Ising spin glasses*, *maximum satisfiability*)

Próba oceny – c.d.

- **Bardziej współczesne analizy pokazują słabość rozwiązań bazujących na BOA dla niektórych problemów (np. wielo-wymiarowy problem plecakowy), gdzie wymagają znacznie większej populacji, aby osiągnąć podobny rezultat co metody alternatywne należące do tej samej kategorii algorytmów (np. LTGA, eCGA).**

WARIANTY I ALTERNATYWY

BOA Hierarchiczne

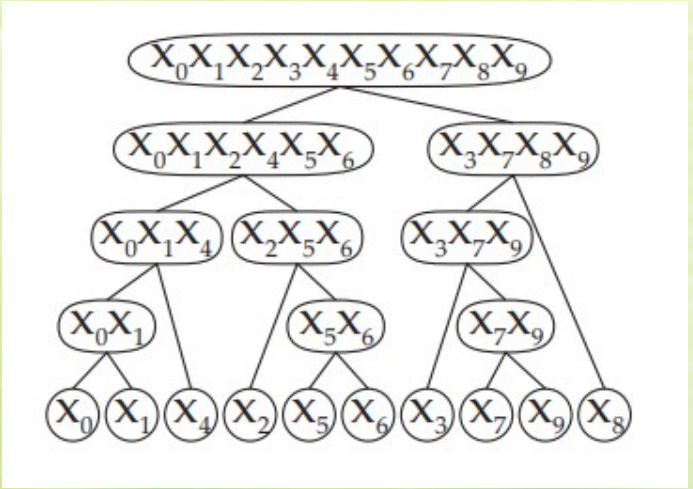
- BOA hierarchiczne (hBOA) – wykorzystuje dodatkowe struktury danych do reprezentowania hierchii zależności dla poszczególnych zmiennych
- Przykładem jest BOA z drzewami lub grafami decyzyjnymi, gdzie każdej zmiennej przypisana jest taka struktura, w której znajdują się informacje o jej zależnościach od innych zmiennych
- Taki algorytm nie tworzy sieci bayesowskiej wprost, ale może być ona w każdej chwili odtworzona na podstawie drzew/grafów wszystkich zmiennych

BOA Hierarchiczne – c.d.

- **Korzyści wynikające z takiego podejścia:**
  – kompresja danych (potrzebne jest mniej prawdopodobieństw warunkowych do określenia relacji między zmiennymi)
  – większa klasa problemów skutecznie rozwiązywalnych (pozwala rozwiązywać pewne problemy, z którymi w teorii zwykłe BOA sobie radzi, ale wymaga populacji wykładniczo rosnącej wraz z jego rozmiarem)
  – skuteczniejsze uczenie się modelu

Metody alternatywne – LTGA

- **Linkage Tree GA (LTGA) – również PMBGA pozwalający tworzyć modele zdolne uchwycić relacje wyższych rzędów między zmiennymi**
- **Zamiast sieci bayesowskiej wykorzystuje tzw. drzewo połączeń (linkage tree), również budowane w trakcie działania algorytmu**
- **Punktem wyjścia są same węzły z jedną zmienną (rozkłady brzegowe), a następnie dodawane są kolejne (od dołu) przy wykorzystaniu algorytmu klasteryzacji hierarchicznej**



- **Nie buduje pełnego modelu (z pr. war.), ale tylko gromadzi informacje o występujących zależnościach.**

Źródła i podsumowanie

- **Wybrane źródła:**

– M. Pelikan, D. E. Goldberg, E. Cantu-Paz, *Linkage Problem, Distribution Estimation, and Bayesian Networks*, IlliGAL Report 1998.

– M. Pelikan, *Bayesian Optimization Algorithm: From Single Level to Hierarchy*, Ph.D. Thesis, University of Illinois–Ubrana 1998.

– M. Pelikan, D. E. Goldberg, S. Tsutsui, *Hierarchical Bayesian Optimization Algorithm: Toward a New Generation of Genetic Algorithms*, SICE 2003.

– D. Thierens, *The Linkage Tree Genetic Algorithm*, PPSN 2010.

– J. P. Martins et al., *A comparison of Linkage-learning-based Genetic Algorithms in Multi-dimensional Knapsack Problems*, IEEE Congerss on Evolutionary Computatation 2013.

- **Dziękuję za uwagę. Pytania?**