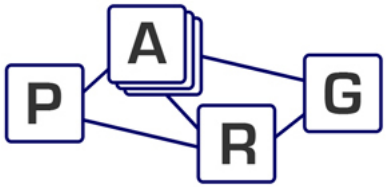
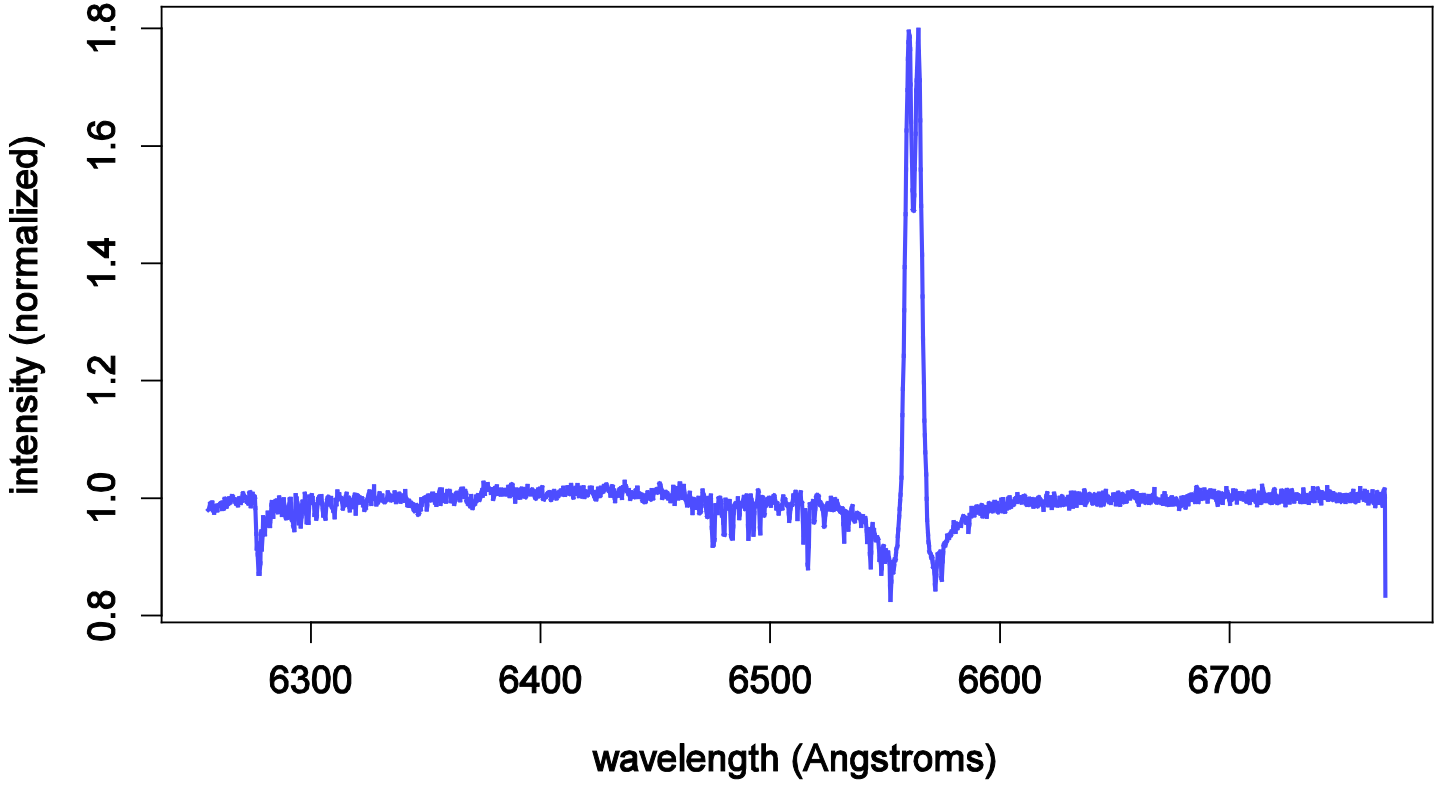


Employing Evolutionary Algorithms for Classification of Astrophysical Spectra

Cíle

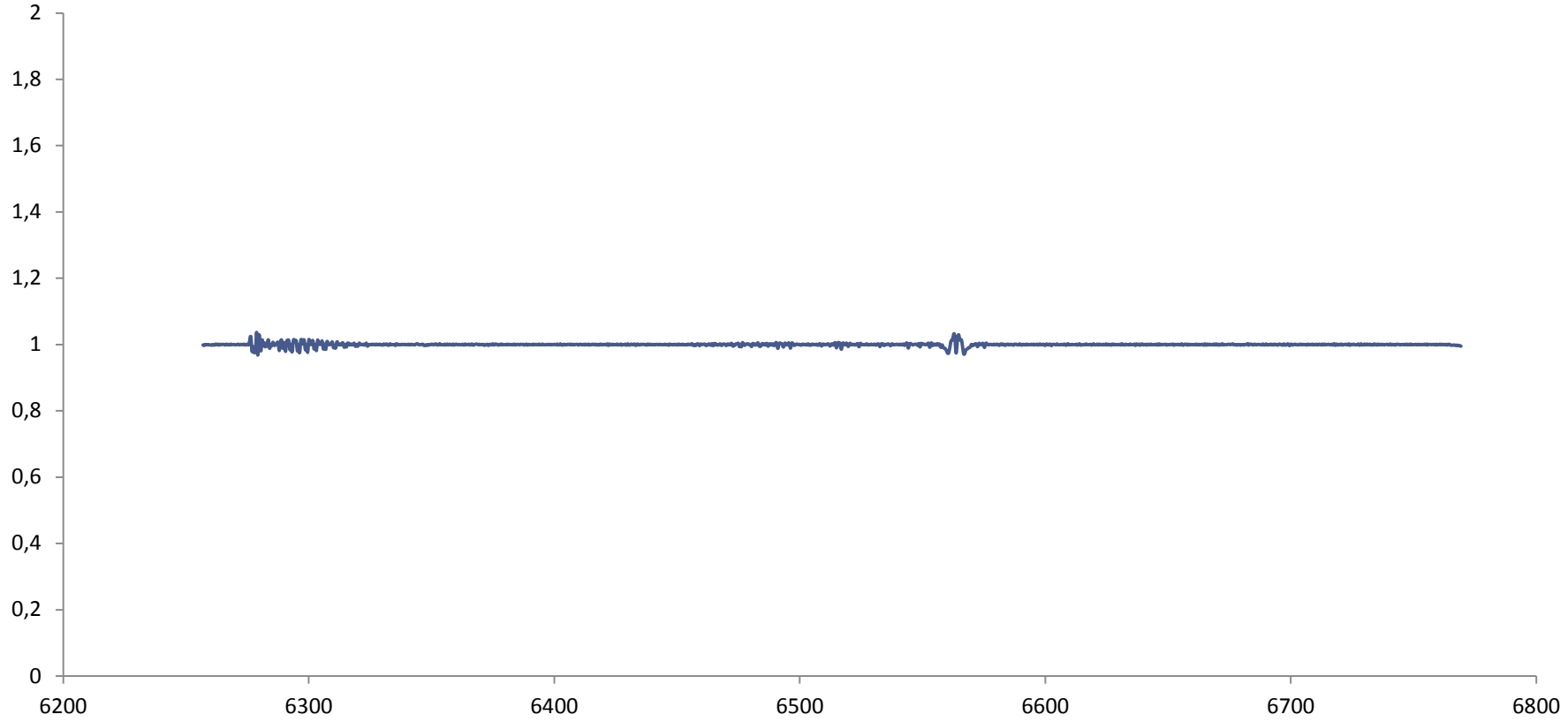
- Astronomické cíle
  - Klasifikace Be hvězd do podtříd
    - Na základě spektroskopie
    - Odlišné tvary spekter odpovídají odlišné geometrii objektu
- Informatické cíle
  - Automatizovat metody klasifikace
  - Aplikovat na velké kolekce (SDSS)
  - Důraz na výkon klasifikačního mechanismu
    - Lze v rozumném čase spočítat klasifikaci na statisícových kolekcích?
  - Generalizace na jiné spektroskopické problémy

Ideální tvar spektra Be hvězdy



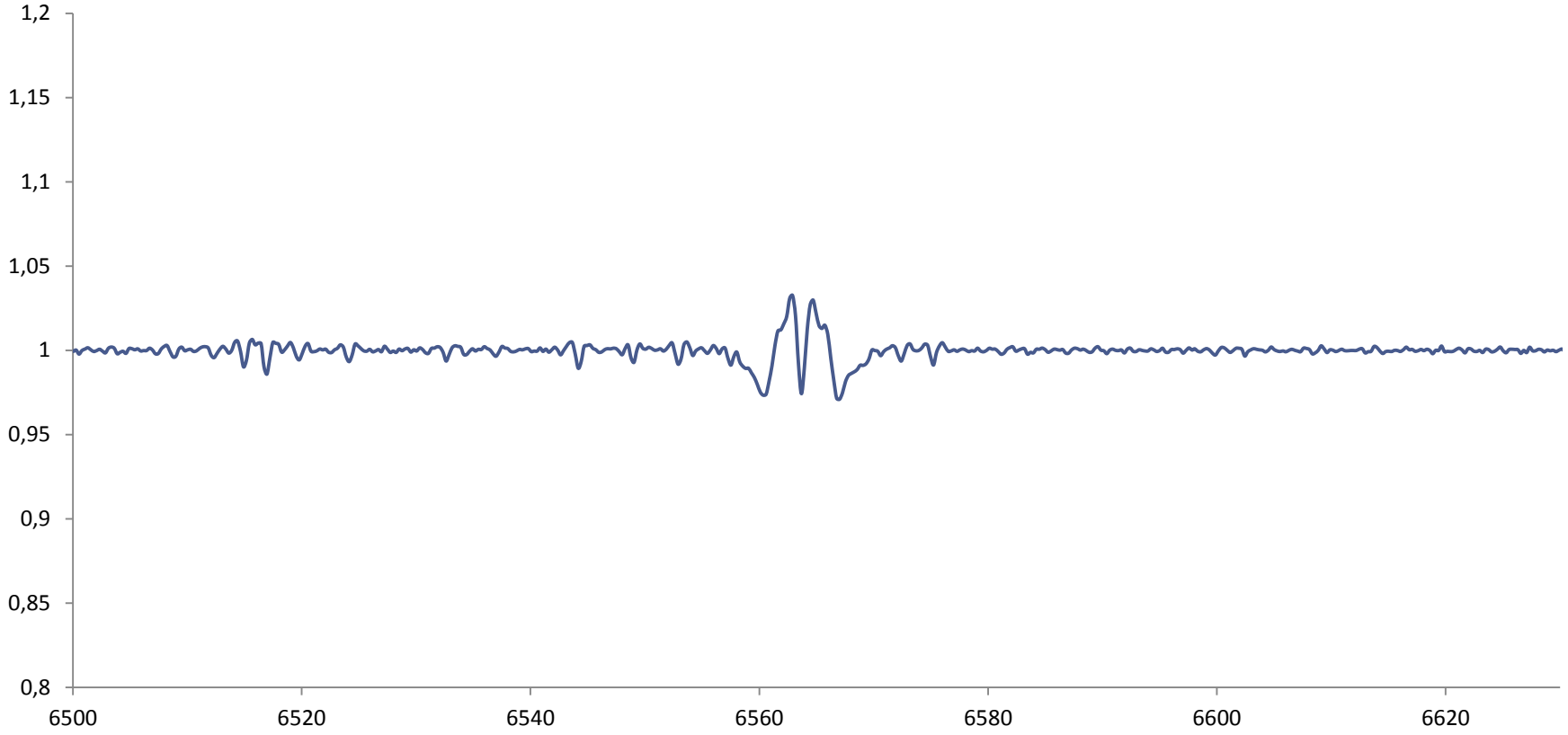
- Jasně viditelná emisní čára (H-alpha – atomární vodík)
  - Dvojitý vrchol daný rotací zářícího plynového disku
- Les absorpčních čar vlevo od emisní je relativně nevýrazný
  - Většinou voda v zemské atmosféře

Méně ideální tvar spektra Be hvězdy



- Vzdálenější hvězda – horší odstup od pozadí
  - Efekt atmosféry výrazně ovlivňuje tvar spektra
  - Rozpoznání emisní čáry je stále možné
  - Přesné měření parametrů emisné čáry je ztíženo

Méně ideální tvar spektra Be hvězdy



- Vzdálenější hvězda – horší odstup od pozadí
  - Přesné měření parametrů emisní čáry je ztíženo
  - FWHM = šířka čáry v polovině výšky
    - Kde to je?

Zdroje spektroskopických dat

Ondřejovská spektra

- 2m dalekohled
  - Dedikovaný pro spektroskopii
  - Velmi přesná měření
- Manuální výběr objektů
  - Stovky jasných objektů
- Automatické zpracování
  - Normalizace

Sloane Digital Sky Survey

- 2.6 m dalekohled
  - Vícevláknová spektroskopie
  - Měří cca 1000 objektů najednou
    - Společné expoziční parametry
    - Nelze měřit jasné objekty
  - Menší rozlišení
- Automatický výběr objektů
  - Statisíce slabých objektů
- Automatické zpracování
  - Odečtení pozadí oblohy
    - Volná vlákna
    - Zdvojnásobuje šum

Původní snaha

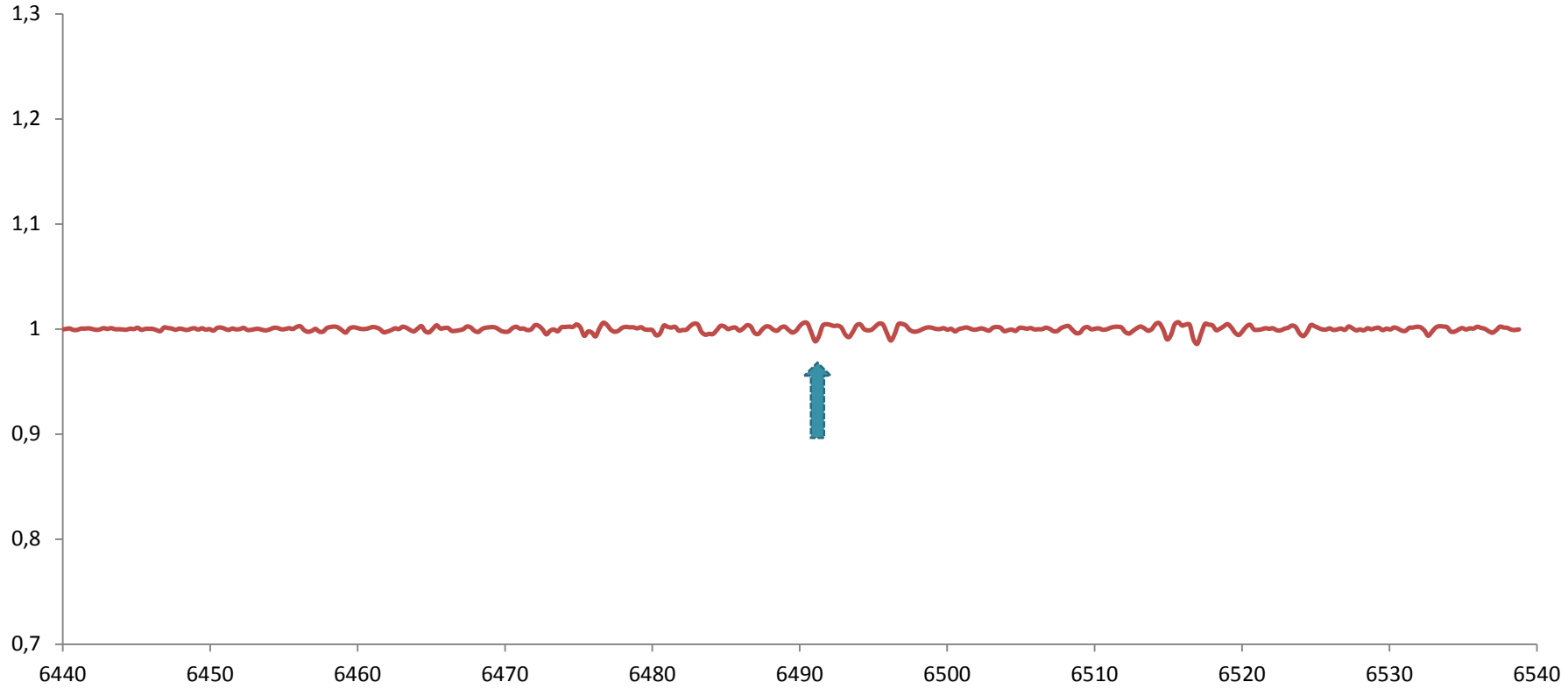
- Přenesení Ondřejovských metod na SDSS
  - Nutná plná automatizace
  - Objekty jsou slabší
    - Původní techniky nefungují
    - Neexistuje objekt měřený oběma kolekcemi
  - Jiné techniky předzpracování, jiné rozlišení
  - Důsledek: Nemáme učitele pro případné učení

  - Závěr: Hledejme zcela jiné přístupy
    - Učení bez učitele

Nové cíle

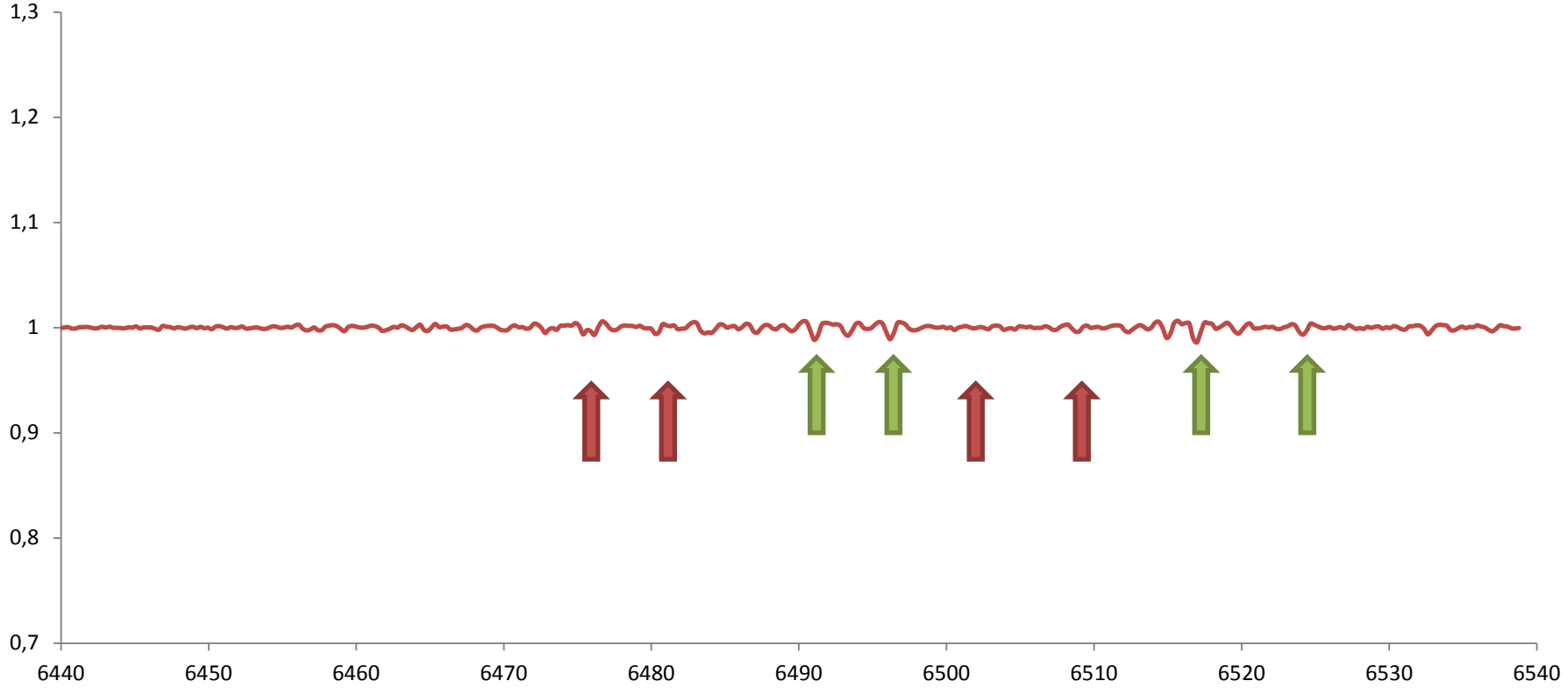
- Hlavní problém velkých kolekcí: Slabé objekty
  - Photon-counting noise – měříme Poissonovo rozdělení
    - Rozptyl úměrný odmocnině intenzity signálu
    - Stejně šumí i odečítaný jas oblohy – rozptyl nezávislý na intenzitě užitečného signálu
    - Rozptyl ale známe - můžeme odhadnout spolehlivost extrahovaných informací

Rozpoznávání slabých signálů



- Je to čára nebo náhodná fluktuace
  - Slabší čáry jsou řádově stejně vysoké jako směrodatná odchylka
  - Pro jednotlivé čáry neumíme odpovědět

Rozpoznávání slabých signálů



- Pro jednotlivé čáry neumíme odpovědět
- Pro soubory čar je odpověď spolehlivější
- Dokážeme rozhodnout, zda měření odpovídá předpokládanému spektru
  - Existují syntetická spektra generovaná na základě fyzikálních zákonů
  - Prostor parametrů je ovšem dost velký

Nové cíle

- Jiná myšlenka:
  - Kdybychom měli několik podobných objektů, můžeme měření sečíst
    - Užitečný signál roste rychleji než šum
    - Ve velkém množství objektů se nějaké podobné objekty najdou
  - Problémy:
    - Jak najdeme podobné objekty?
      - To je náš hlavní cíl – jsme v kruhu
    - Úplně stejné objekty nenajdeme
      - Odlišná intenzita daná různými vzdálenostmi
      - Odlišný Dopplerovský posuv daný různými rychlostmi

Návrh řešení

- Syntetická spektra
  - Parametry spekter generovány evolučními algoritmy
  - Fitness = odchylka od naměřeného spektra
- Evoluce trefuje všechna naměřená spektra najednou
  - Syntetické spektrum přežívá, pokud je dobrou aproximací alespoň jednoho naměřeného spektra
- Hledání podobných spekter
  - Příbuzenský vztah mezi syntetickými spektry
  - Jsou-li měřená spektra dobře aproximována příbuznými syntetickými spektry, budou podobná
    - Obrácená implikace neplatí, to nám ale nemusí vadit
    - Skutečnou klasifikaci uděláme později, na základě syntetických spekter

Komplikace

- Nedokážeme porovávat všechna syntetická spektra se všemi měřeními
  - Náhodný výběr + evoluce
- Měřené objekty nejsou identické
  - Při porovnávání se syntetickým spektrem je nutná další transformace
    - Odpovídá vzdálenosti, rychlosti a teplotě objektu
      - Silně zjednodušená fyzika dostatečně pokrývá malé odchylky
  - Parametry transformace podléhají evoluci

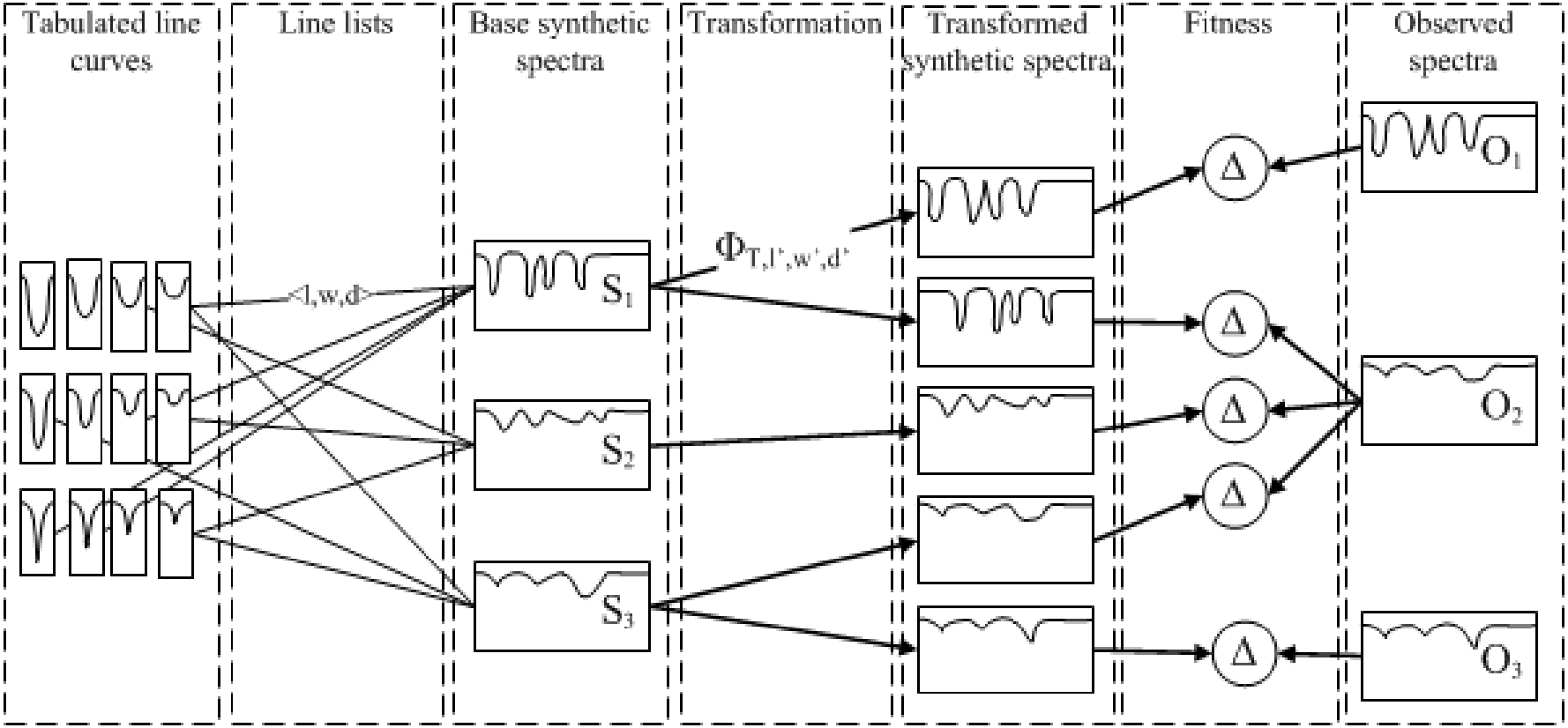
Evoluční mechanismus

- Koevoluce dvou tříd organismů
  - Syntetická spektra
    - Chromozom = množina čar (pozice, šířka, intenzita)
  - Párování
    - Dvojice syntetické + naměřené spektrum
    - Tři číselné parametry transformace mezi syntetickým a naměřeným spektrem

Evoluční mechanismus

- Výpočet fitness
  - Porovnání transformovaného syntetického spektra s naměřeným
    - Fitness = normalizovaná kvadratická odchylka
  - Každé naměřené spektrum leží v několika párech
    - Páry s nejlepší fitness jsou vybrány k přežití
  - Každé syntetické spektrum se účastní v několika párech
    - Fitness spektra = počet přežívajících párů

Schéma výpočtu fitness



- Základní spektra se ve skutečnosti nepočítají
  - Transformace probíhá přímo na seznamech čar
- Syntetická spektra se škálují v rámci výpočtu fitness
  - Multiplikační koeficient hledán metodou nejmenších čtverců
  - Minimální residuum = fitness

Evoluční mechanismus

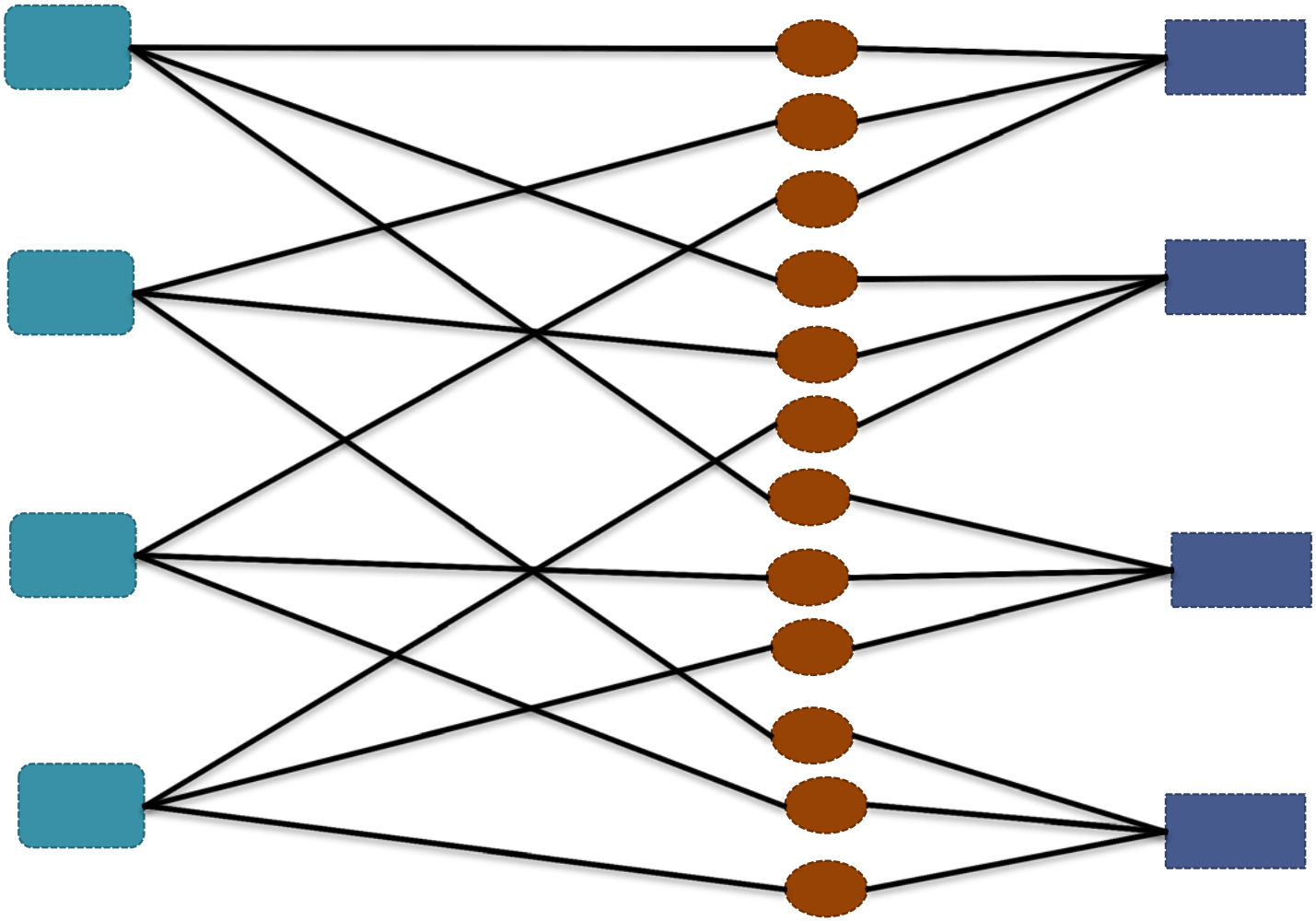
- Generování populace
  - Nepřeživší páry zanikají
  - Syntetická spektra se špatnou fitness nezanikají, ale nemají potomky
  - Syntetická spektra s dostatečnou fitness generují potomky
    - Náhodná mutace, případně křížení
    - Vztah předek-potomek je zaznamenán
  - Přeživší páry generují potomky
    - Náhodná mutace parametrů transformace
    - Přestěhování k jinému syntetickému spektru
      - Páry s dobrou fitness k náhodně vybranému potomku syntetického spektra
      - Ostatní páry k náhodně vybranému sourozenci syntetického spektra

Evoluce

Syntetická spektra

Párování

Naměřená spektra

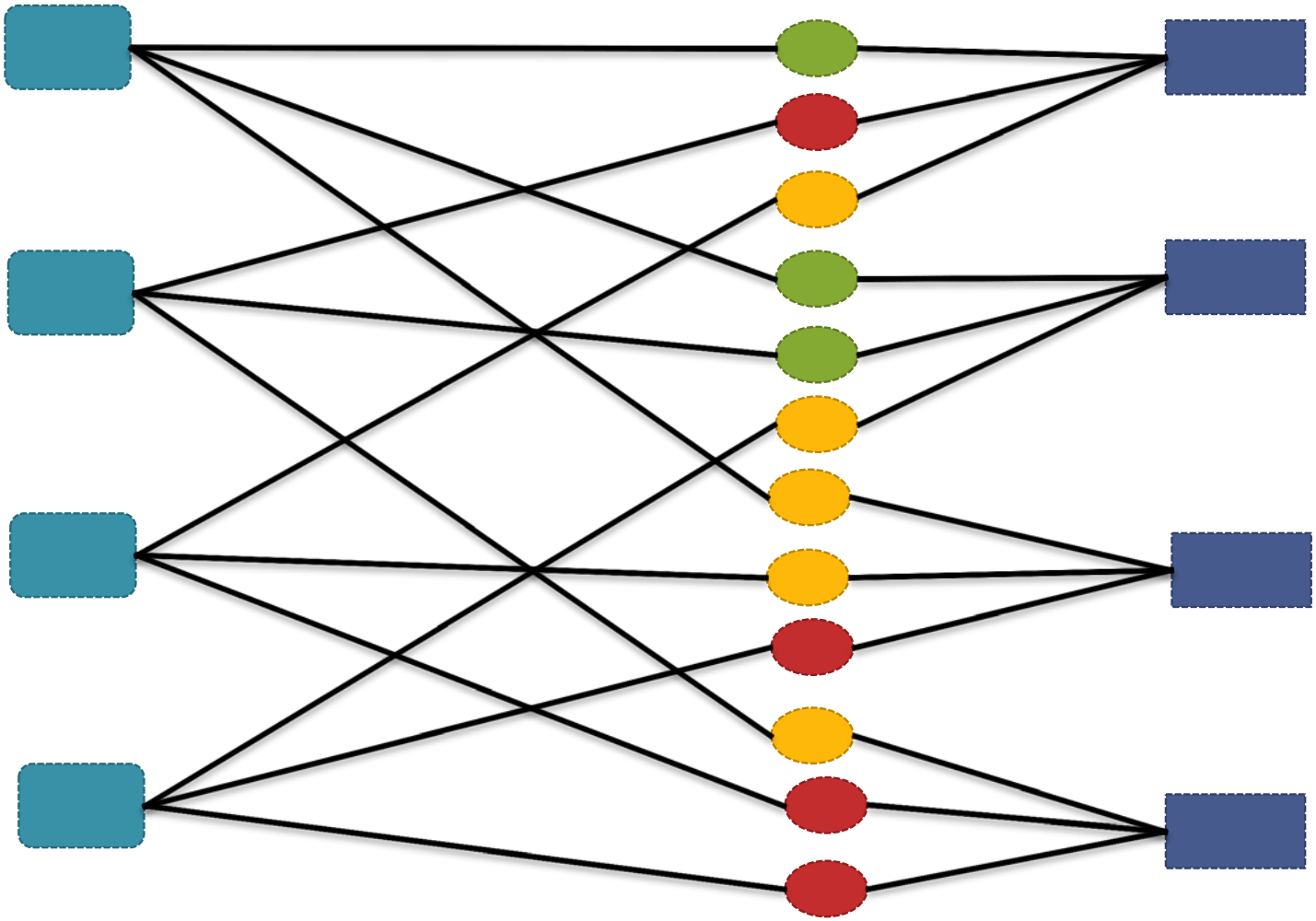


Evoluce

Syntetická spektra

Fitness párování

Naměřená spektra

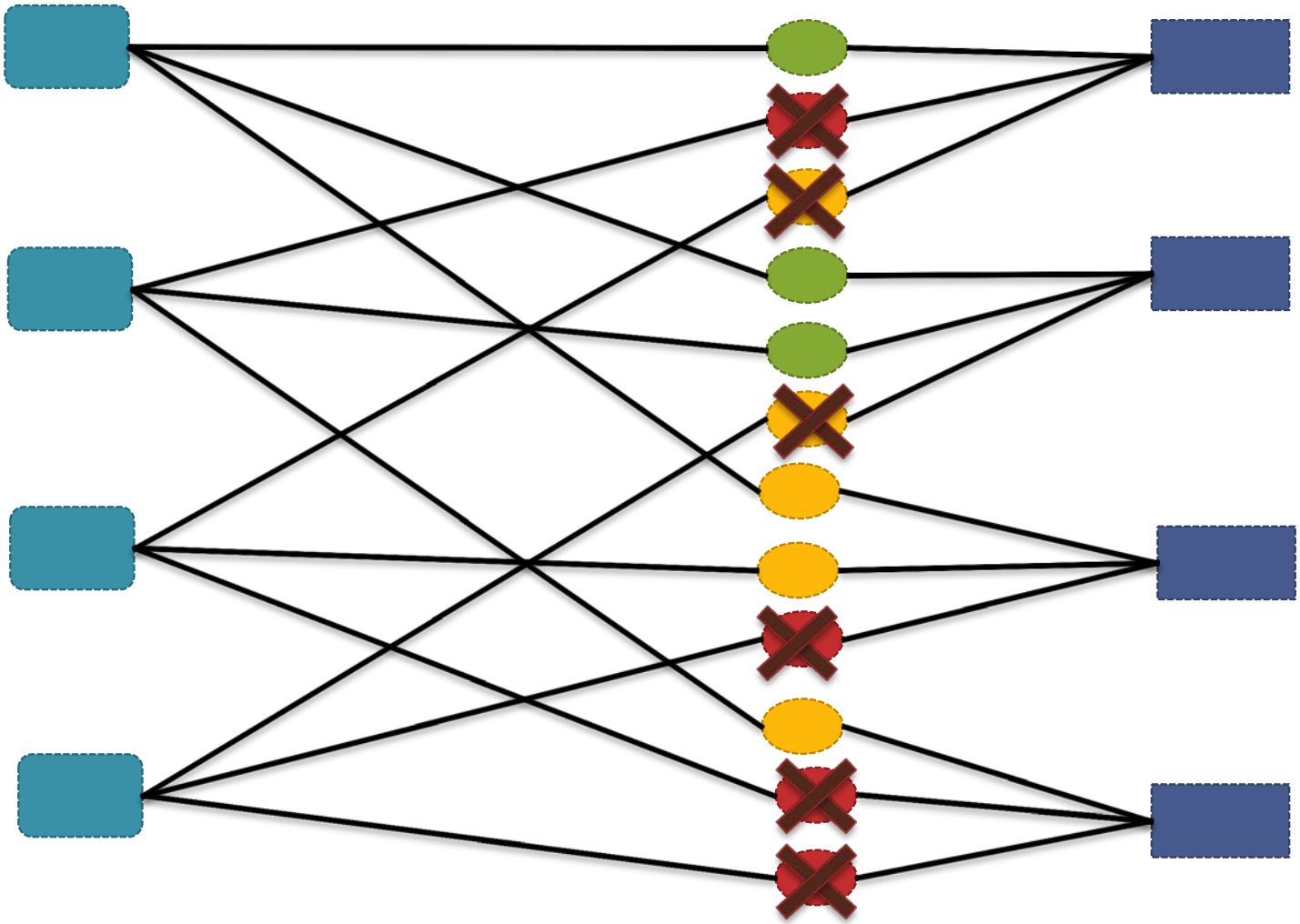


Evoluce

Syntetická spektra

Eliminace párování

Naměřená spektra

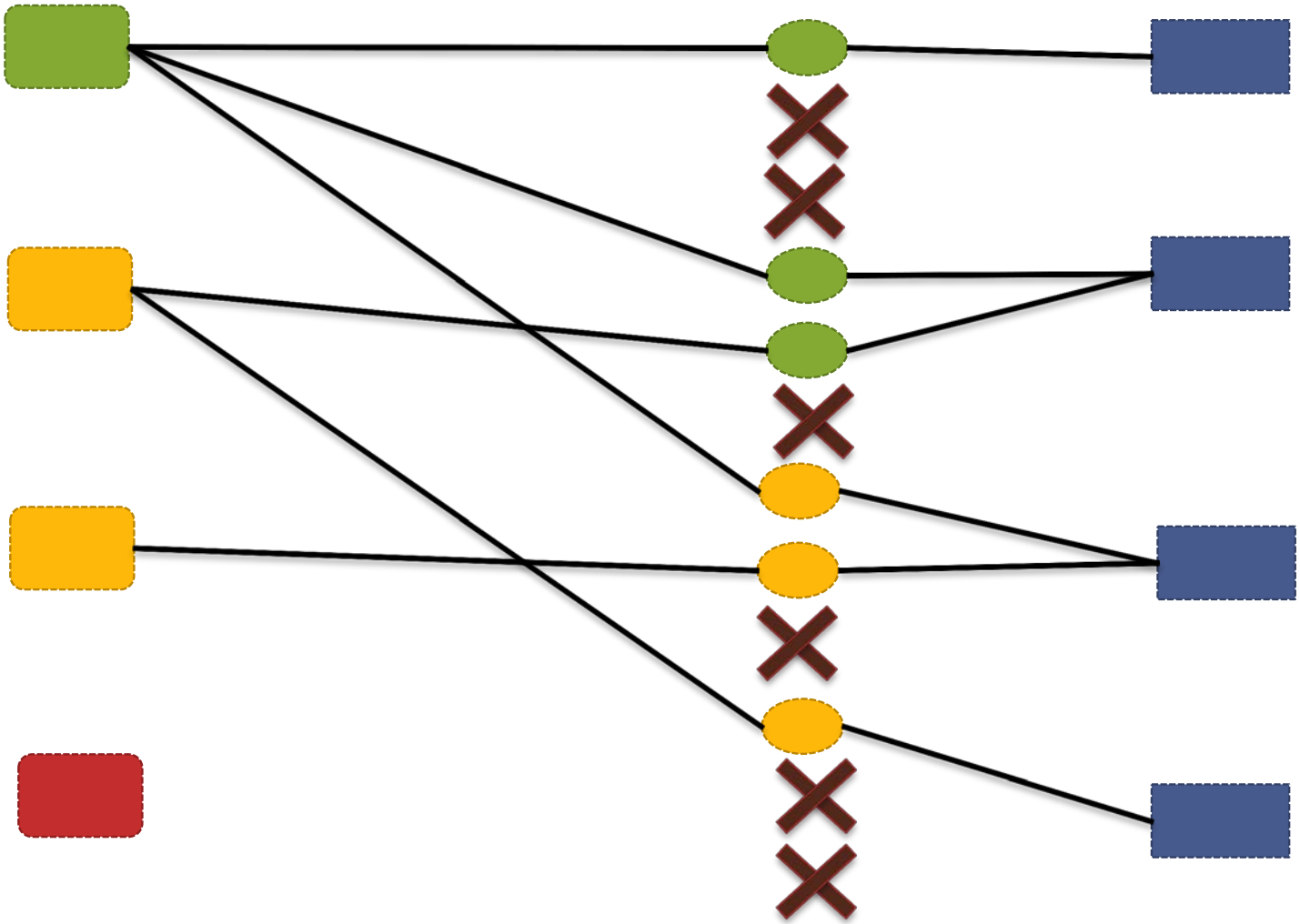


Evoluce

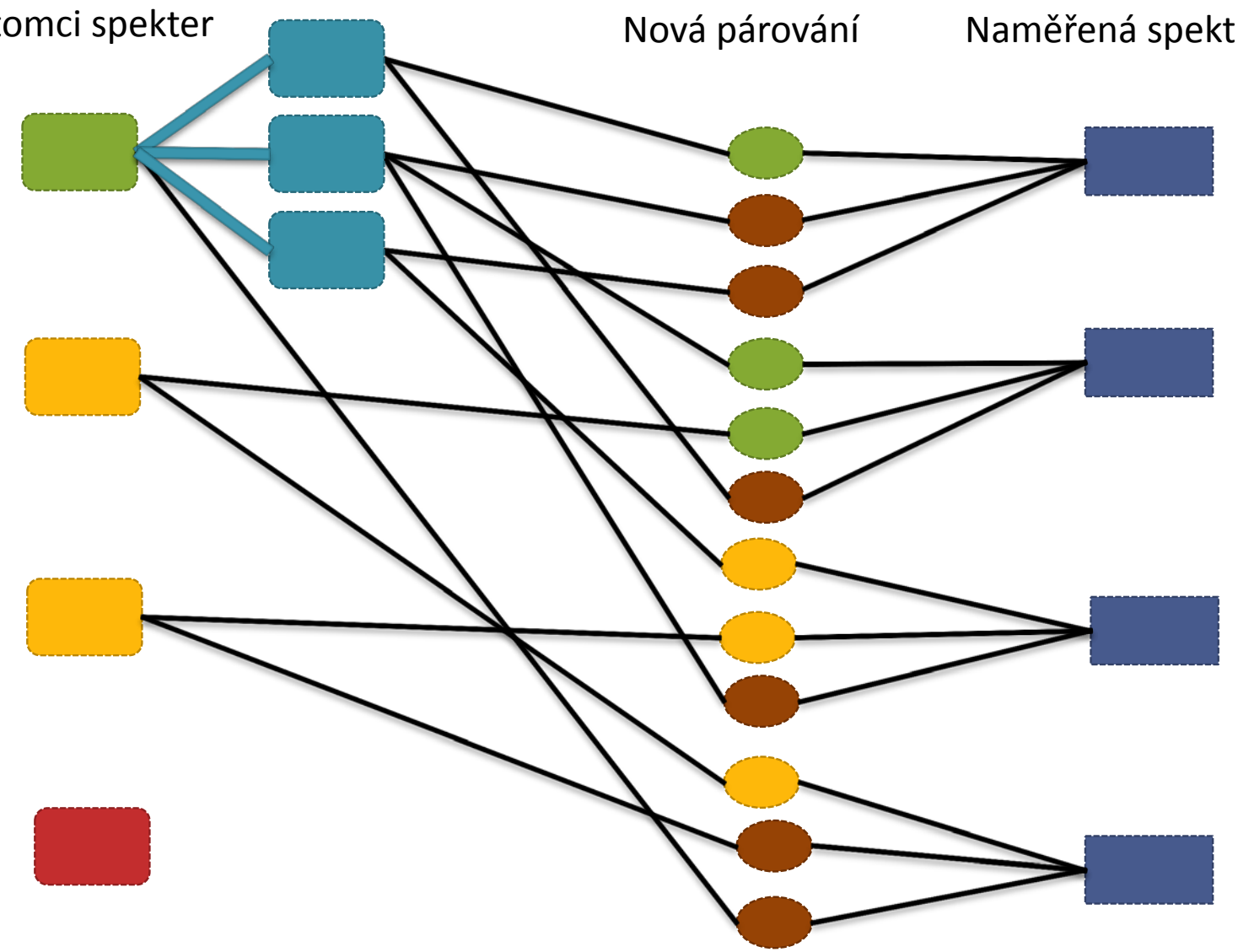
Fitness spekter

Eliminace párování

Naměřená spektra



Evoluce



Algoritmus evoluce

Require: O the observations ; N, t, $M_{\mathtt{M}}$, $M_{\mathtt{X}}$ evolution parameters

1: S := random population of spectra
2: P := random population of pairings on $S \times O$
3: **while** not satisfied **do**
4: compute the fitness $q(P_k)$ for every $P_k \in P$
5: for every $O_j \in O$, determine P'_j = the top N associated P members according to q
6: $P' := \bigcup_j P'_j$
7: for every $S_i \in S$, determine $n(S_i)$ = the number of P' members hosted by S_i
8: $S' := \{S_i \in S \mid n(S_i) > t\}$
9: for every $S_i \in S'$ generate $M_{\mathtt{M}}$ children of S_i by random mutation
10: randomly select $M_{\mathtt{X}}$ pairs from S' for crossover
11: relocate every $P_k \in P'$ to a randomly selected child of the linked S member
12: relocate every $P_k \in P \setminus P'$ to a randomly selected sibling of the linked S member
13: **end while**

Výpočetní náročnost

- Paralelní algoritmus pro SMP a NUMA
  - Naměřená spektra uložena na disku
  - Populace uloženy v paměti
    - Syntetická spektra jsou reprezentována seznamy čar
- Kritické součásti
  - Rychlý přístup k disku
  - Cache-awareness
  - SIMD instrukce
- Výkonnost
  - Kritickým místem je generování transformovaného spektra pro každý pár
    - Cca 25 000 párů za sekundu per core (Intel i7)
  - Rychlost diskového přístupu může být omezujícím faktorem
    - Čtení spekter z disku je 20 až 50-krát pomalejší než generování syntetických
    - Pro méně než 50 párů na naměřené spektrum zdržuje disk

Výpočetní náročnost

- Výkonnost
  - Kritickým místem je generování transformovaného spektra pro každý pár
    - Cca 25 000 párů za sekundu per core (Intel i7)
  - Rychlost diskového přístupu může být omezujícím faktorem
    - Pro méně než 50 párů na naměřené spektrum
    - Spektra musejí být předzpracována do vhodné podoby
- Extrémní test
  - Kompletní ročník SDSS = 800 000 naměřených spekter
  - 100 párů pro každé naměřené spektrum = 80 000 000 párů
  - 6 * 4 core NUMA machine
  - 1 generace za méně než 3 minuty
    - Je to dost nebo málo?

Závěr (srpen 2014)

- Stav
  - Máme funkční paralelní implementaci evolučního algoritmu
  - Výkon dostačuje k provádění experimentů na středně velkých kolekcích
    - To je nejdůležitější výsledek dosavadní práce
    - Pro původní cíl (Be hvězdy) není třeba řešit celou SDSS
  - Výsledky evolučního algoritmu?
    - Chybí nám metodika hodnocení – nemáme průnik kolekcí ani další klasifikátor
- Work in progress
  - Spektra s větším počtem čar konvergují pomalu nebo vůbec
    - Zrychlení pomocí propagace fitness až k jednotlivým čarám
    - Místo náhodné inicializace použití spekter z Ondřejovské kolekce
      - Selection bias?
- Future work
  - Aplikace na celou SDSS
  - Zobecnění evoluční metody

Novější výsledky

- Syntetické spektrum má nevyhovující spojité pozadí
  - Počítalo se s křivkou záření absolutně černého tělesa
    - Tomu ale odpovídají jen některé hvězdy
  - Je nutné dovolit obecnější křivku
    - Příliš mnoho parametrů - další komplikace pro evoluci
    - Lepší nápad: Úprava výpočtu fitness tak, aby ignorovala spojité pozadí
      - Složitější matematika - stihneme to spočítat?
- Spektra s větším počtem čar konvergují pomalu nebo vůbec
  - Zrychlení pomocí propagace fitness až k jednotlivým čarám
    - Velký zásah do algoritmu evoluce - třetí druh organismu
    - Složitější matematika - stihneme to spočítat?
  - Místo náhodné inicializace použití spekter z Ondřejovské kolekce
    - To nepomohlo, problém není v nalezení *přibližně* odpovídajícího spektra

Novější výsledky

- Úprava výpočtu fitness tak, aby ignorovala spojité pozadí
  - Problém: Normalizace samotného měření nefunguje
    - Normalizace měření vede ke křivce se střední hodnotou 1
    - Syntetická spektra jsou křivky s hodnotami menšími než 1
    - Bez přizpůsobení multiplikativní konstantou nikdy nenajdeme shodu
    - Potřebná multiplikativní konstanta je v různých částech spektra různá
  - Je nutné normalizovat pro každou dvojici měřené-syntetické spektrum
    - Hledáme multiplikativní konstantu metodou nejmenších čtverců
    - Konstanta má být v různých částech spektra jiná
    - Hledá se vždy pro lokální výřez ze spektra (cca 60 bodů)
    - Konstanty musí spojitě navazovat
    - Hledáme M konstant, každou spočtenou z K bodů v okolí
    - Umíme to v čase úměrném celkovému počtu bodů spektra, nezávisle na K

Paralelní a vektorová implementace

- Nejdůležitější triky pro dosažení vysokého výkonu
  - Použití SIMD instrukcí
    - Nutnost zarovnání dat
  - Použití předvypočtených průběhů křivek
    - Úspora CPU ale zvýšení počtu paměťových přístupů
  - Náhrada interpolace zvýšeným počtem vzorků
    - Použito i k vyřešení zarovnání SIMD dat
  - Cache-aware algoritmus pro výpočet syntetických spekter
    - Obtížné nalezení optimálních parametrů
  - Používání pseudonáhodných postupů namísto přesných
    - Náhodný výběr s váhami namísto Top-K
    - Snadnější paralelizace
  - Lineární algoritmus pro interpolaci pozadí
    - Inspirace: FFT

Výpočet fitness

Require: O the observations ; A line curves ; S synthetic spectra line lists ; P pairings and transformation parameters

Ensure: fitness value $q(P_k)$ for every $P_k \in P$

for each group $G_j^O \subseteq O$ of observed spectra **do**
  read the group G_j^O into memory
  $L' := \emptyset$
  for each pairing $P_k \in P$ associated to a spectrum from G_j^O [in parallel] **do**
    allocate and initialize buffer C_k for the transformed spectrum
    compute transformed line list L'_k from the base line list and the transformation parameters
    $L' := L' \cup L'_k$
  end for
  sort L' by the index of the referenced line curve
  for each group $G_n^A \subseteq A$ of the line curves **do**
    determine the range $L'' \subseteq L'$ corresponding to G_n^A
    sort L'' by the index of the observed spectrum
    for each group $G_m^L \subseteq L''$ [in parallel] **do**
      for each transformed line $\langle k, l, w, d \rangle \in G_m^L$ **do**
        multiply the line curve $A_{l,w,d}$ to the buffer C_k
      end for
    end for
  end for
  for each pairing $P_k \in P$ associated to a spectrum from G_j^O [in parallel] **do**
    compute fitness $q(P_k)$ of C_k w.r.t. the associated spectrum
  end for
end for

Paralelní a vektorová implementace

- Softwarově inženýrské problémy
  - Vhodné technologie?
    - C++
    - xmmintrin.h - intrinsic functions for SSE/AVX
    - Intel TBB - paralelizace
    - Intel MKL - paralelizovatelné náhodné generátory
    - Asynchronous I/O pro rychlý diskový přístup
  - Technologie nejsou příliš kompatibilní
    - C vs. C++, závislost na OS, nekompatibilita s překladači
  - Nutná znalost fyzických parametrů výpočetního prostředí
    - Cache-aware algoritmus (cache-oblivious verze neexistuje)
    - Stupeň vektorizace (SSE: 4, AVX: 8)

- Je šance, že by tohle všechno mohlo být řešeno automaticky?
  - Cíl: Knihovna pro méně znalé programátory
  - Evoluční algoritmy vyžadují experimentování
    - Existují lidé, kteří rozumějí všem potřebným doménám?
    - Spektroskopie + numerická matematika + evoluční algoritmy + paralelní programování