Vytěžování dat, cvičení 2: Úvod do RapidMineru

Miroslav Čepek

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

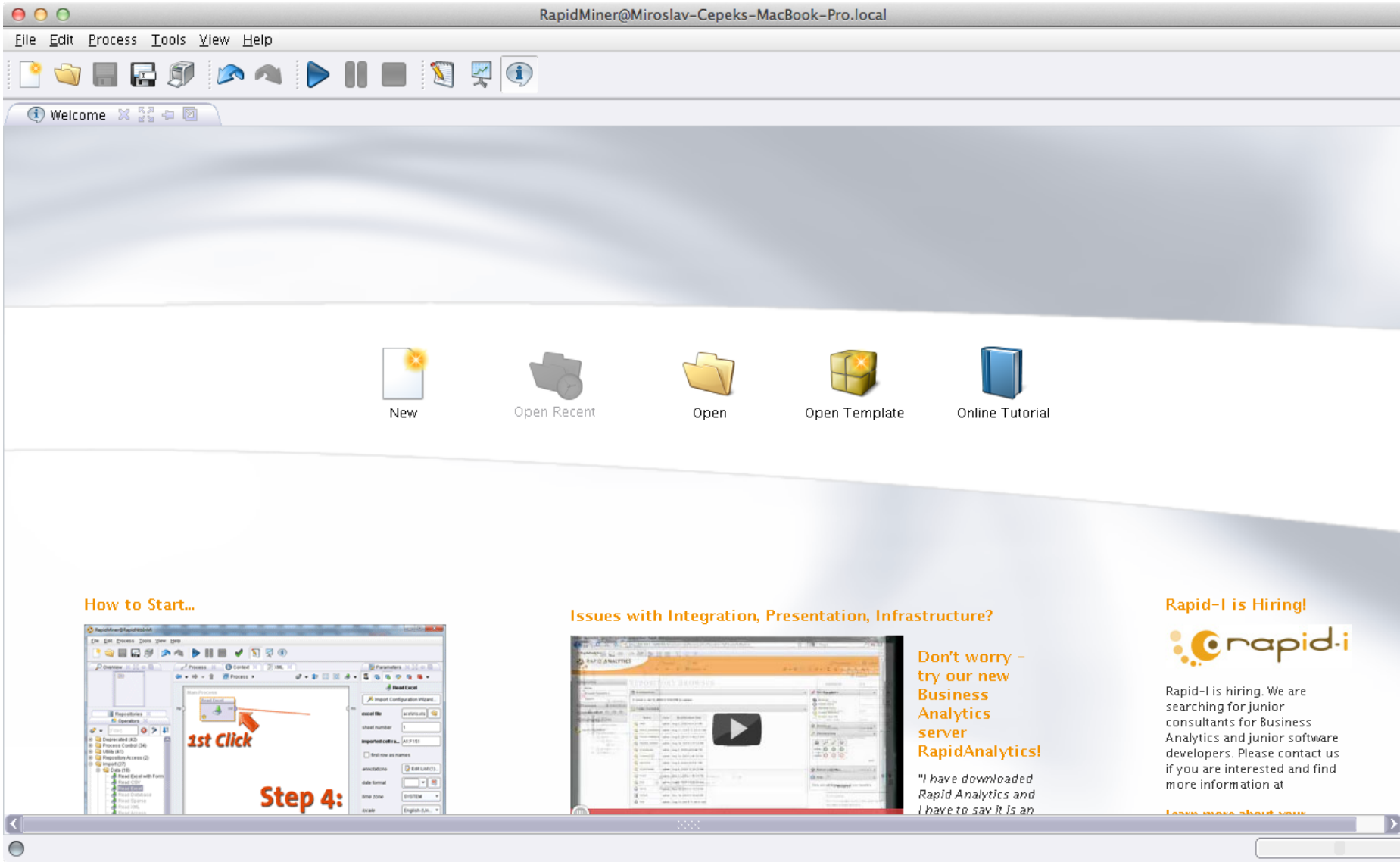
Fakulta elektrotechnická, ČVUT

- Dnes vám ukážeme jeden z mnoha grafických nástrojů pro data mining (aneb nejen příkazovou řádkou živ je člověk).
- Existuje několik open-source nástrojů pro datamining.
- A ten, se kterým si budeme hrát se jmenuje RapidMiner.

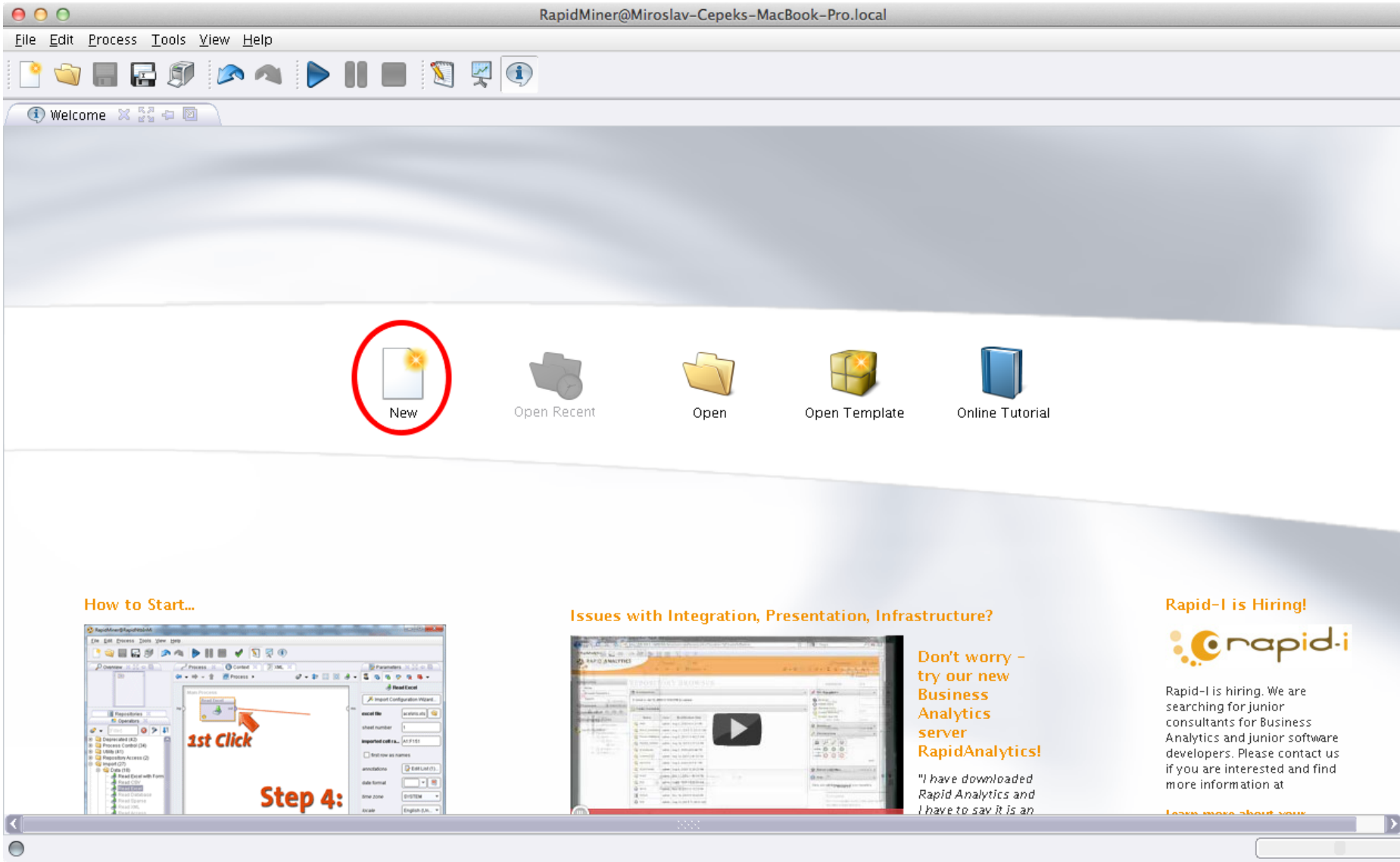
Webové odkazy na Rapidminer

- Rapidminer najdete na adrese `http://rapid-i.com`.
- Stahovat jej můžete z adresy `http://rapid-i.com/content/view/26/84/lang,en/`.
- Stáhněte RapidMiner ze stránek výrobce a nainstalujte jej (můžete do školních počítčů nebo, pokud máte, do vašich notebooků – alespoň to nebudete muset dělat doma znovu :))
- Spusťte jej.

Rapidminer, úvodní obrazovka



Rapidminer, úvodní obrazovka



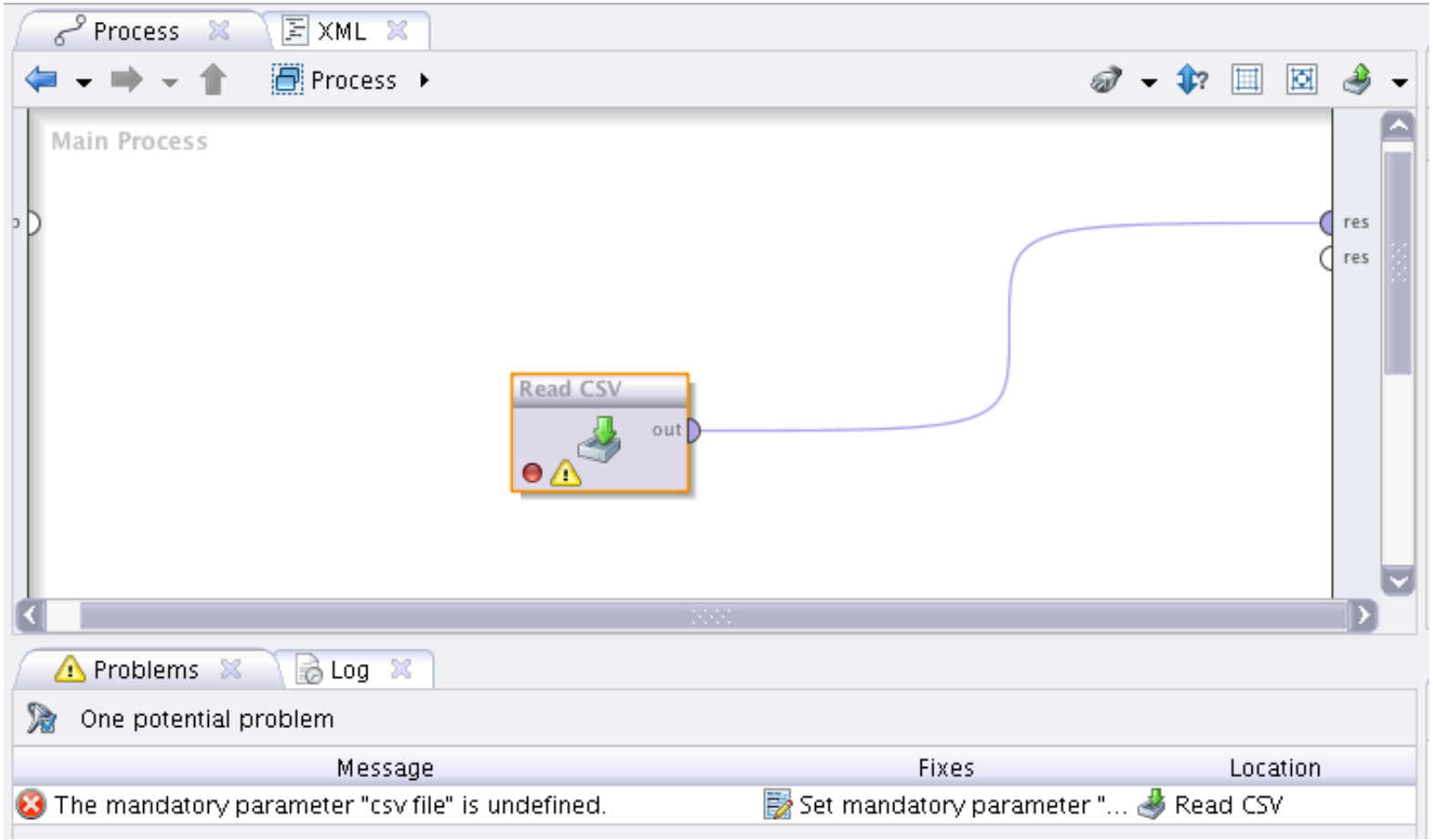
Načtení dat do Rapidmineru

Prvním úkolem je načíst data do RapidMineru.

- Ze stránek cvičení stáhněte soubor iris.csv.
- V seznamu operátorů vyberte uzel (operátor) *Import* $>$ *Data* $>$ *Read CSV* a přetáhněte jej do procesu.
- Měl by se vám automaticky napojit na výstupní vizualizace.

Načtení dat do Rapidmineru (2)

- Výsledek by měl vypadat přibližně tak, jak ukazuje obrázek:



- Všimněte si také, že Rapidminer vám říká, že jste nezadali povinný parametr (název souboru, který chcete načíst).

Načtení dat do Rapidmineru (3)

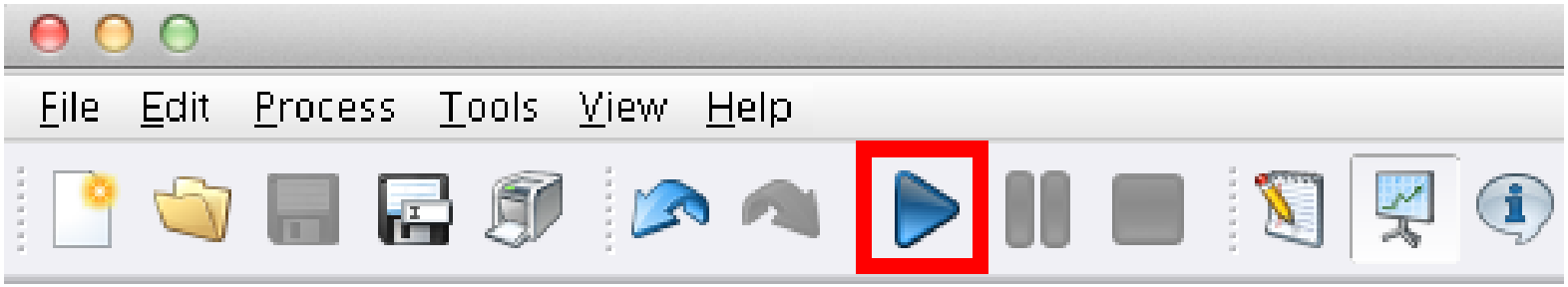
- Napravit tuto chybu můžete buď:
  - kliknutím na *Set mandatory parameter* v dolní záložce se zprávou o chyě nebo
  - doplňením hodnoty v záložce *Paramateres* nebo
  - klinutím na tlačítko *Import Configuration Wizard* v záložce *Paramateres*.
- My si vybreme poslední možnost. Takže klikněte :).

Průvodce načtením dat

- Na první obrazovce průvodce vyberte stažený soubor.
- Druhý krok průvodce se ptá na vlastnosti dat v souboru. Zde nastavte oddělovač sloupců na čárku (comma).
- Ve třetím kroku vám průvodce jen nabídne náhled na data.
- A v posledním kroku můžete (zde musíte) vybrat správný typ dat v jednotlivých sloupcích a jejich roli.
  - U posledního sloupce (class) nastavte typ na *Polynomial* a roli na *label*.
- A dokončete průvodce.

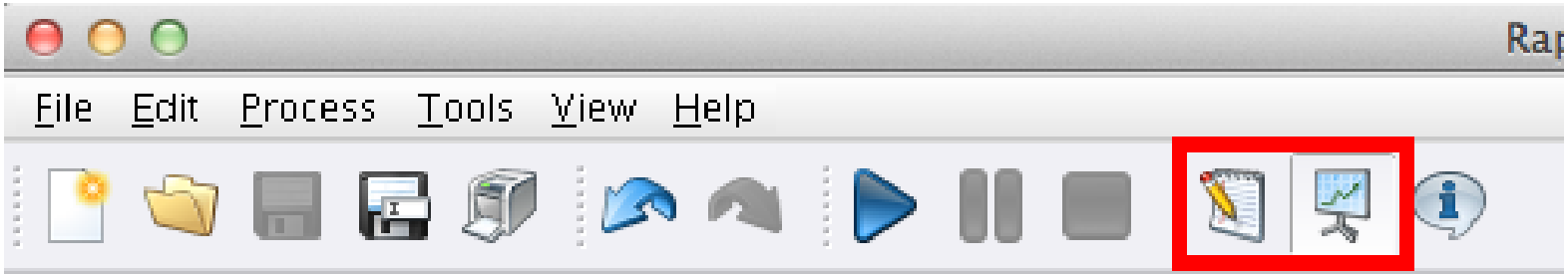
Spuštění proudu

- Nyní je již proud v pořádku a můžeme jej spustit pomocí ikonky.



Zobrazení výsledků

- Po dokončení zpracování proudu se vás RapidMiner dotáže, zda chcete přepnout do *Result Perspective*.
- Mezi perspektivami se můžete přepínat pomocí ikonek v panelu nástrojů.



- Případně z menu *View* > *Perspectives*.

Zobrazení výsledků (2)

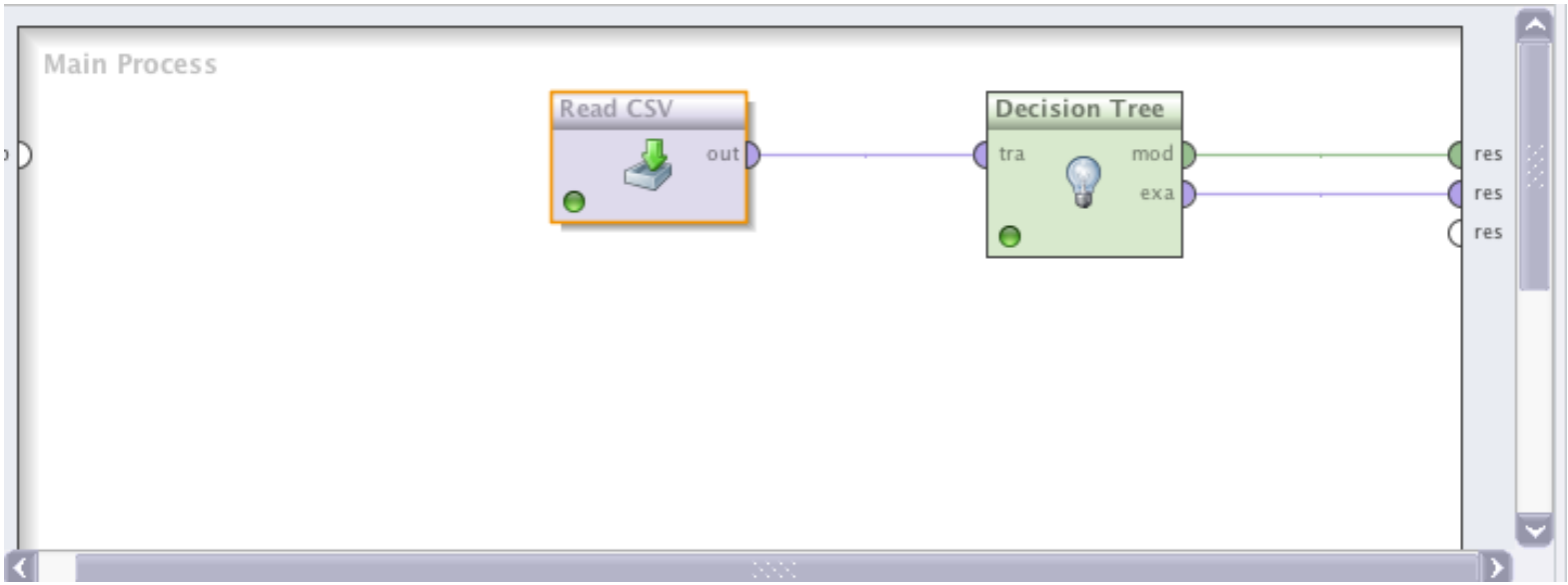
- V *Result Perspective* získáte jednu záložku za každý výstup připojený připojený na pravý okraj plochy.
- Aktuálně si zde můžete prohlédnout jednoduchou popisnou statistiku a různé grafické znázornění načtených dat. Vyzkoušejte!
- Při zkoumání grafického výstupu se podívejte také na bodové grafy (*Scatter plot*) a matici bodových grafů (*scatter plot matrix*).

Vložení modelu do proudu a jeho zobrazení

- Jedna z jednoduchých modelovacích metod pro učení s učitelem je rozhodovací strom.
- V principu je to posloupnost otázek, které navigují k rozhodnutí (třídě do které instanci/objekt zařadíte). Pokud jste někdy viděli klíč k určování rostlin, tak to je přesně ono.
- Podle jakých pravidel se rozhodovací stromy vytváří necháme teď stranou a necháme Rapidminer ať nám nějaký strom vytvoří.
- Najděte uzel *Modelling* > *Classification and Regression* > *Tree Induction* > *Decision Tree*, přetáhněte jej do proudu.

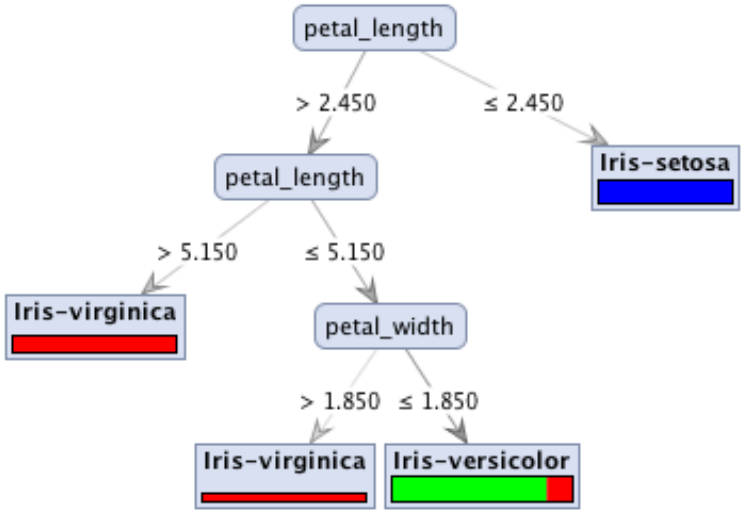
Vložení modelu do proudu

- Na vstup připojte výstup uzlu *Read CSV* a výstup připojte na pravou stranu vizualizací.
- Výsledný proud by pak měl vypadat zhruba takto:



Zobrazení výsledného modelu

- Spusťte proud a přepněte se do *Result Perspective*. V záložce věnované modelu byste měli uvidět podobný obrázek:



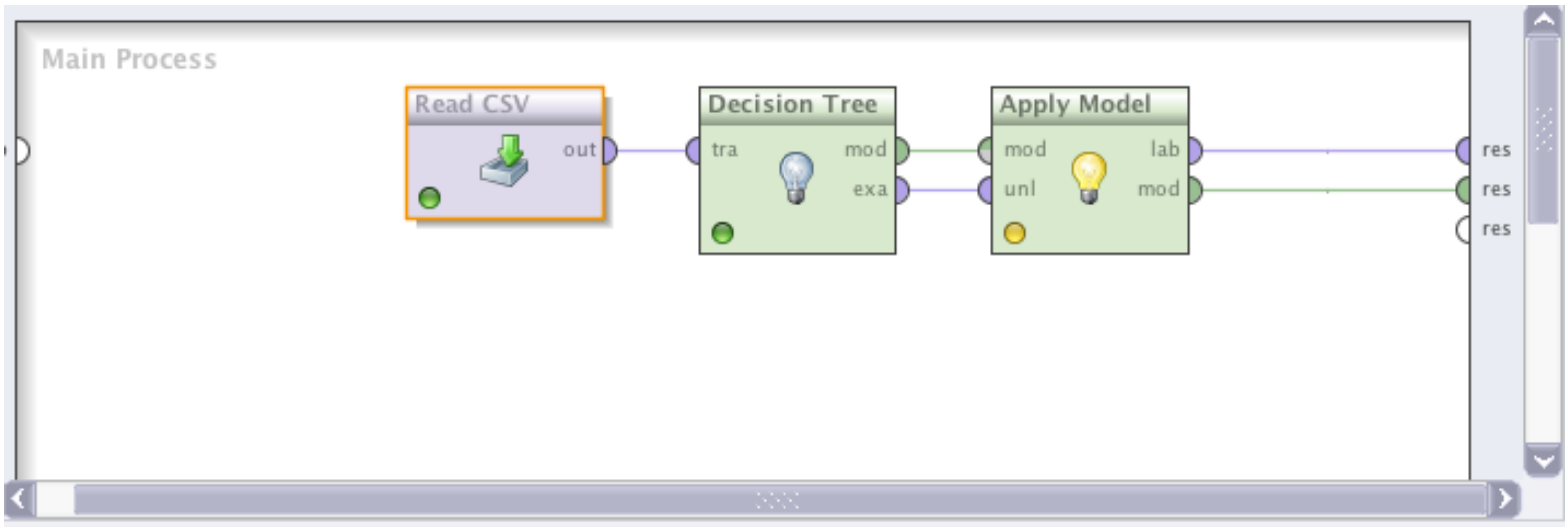
- Jaké podmínky použil strom pro rozhodování? Dokázali byste tyto podmínky vyjádřit třeba v Javě?
- Co znamenají ty červené, zelené a modré pruhy v listech?

Aplikace modelu na data

- Co teď s hotovým modelem :)?
- Můžeme jej aplikovat na data
  - se známou klasifikací a získat tak představu, jak si model vede,
  - s neznámou klasifikací a získat tak předpokládaný výstup.
- My zkusíme nejprve aplikovat model na trénovací data.

Aplikace modelu na data (2)

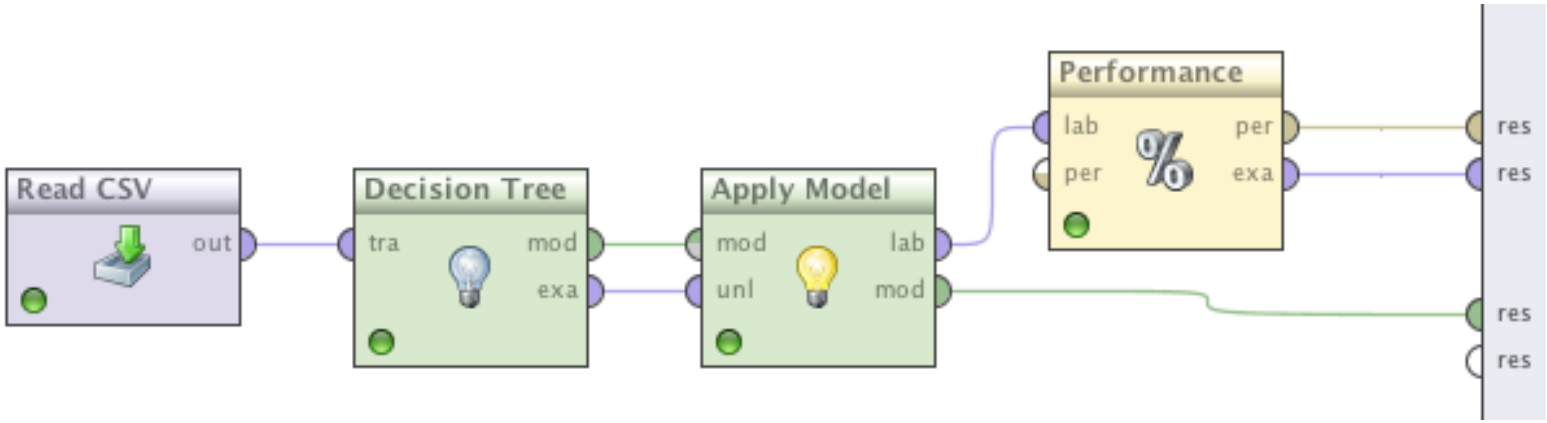
- Najděte uzel *Modeling* > *Model application* > *Apply model* a vložte jej mezi rozhodovaní strom a pravý "vizualizační" okraj.



- Spusťte a zjistěte jak vypadá výstup?
- Které kosatce klasifikuje model špatně?

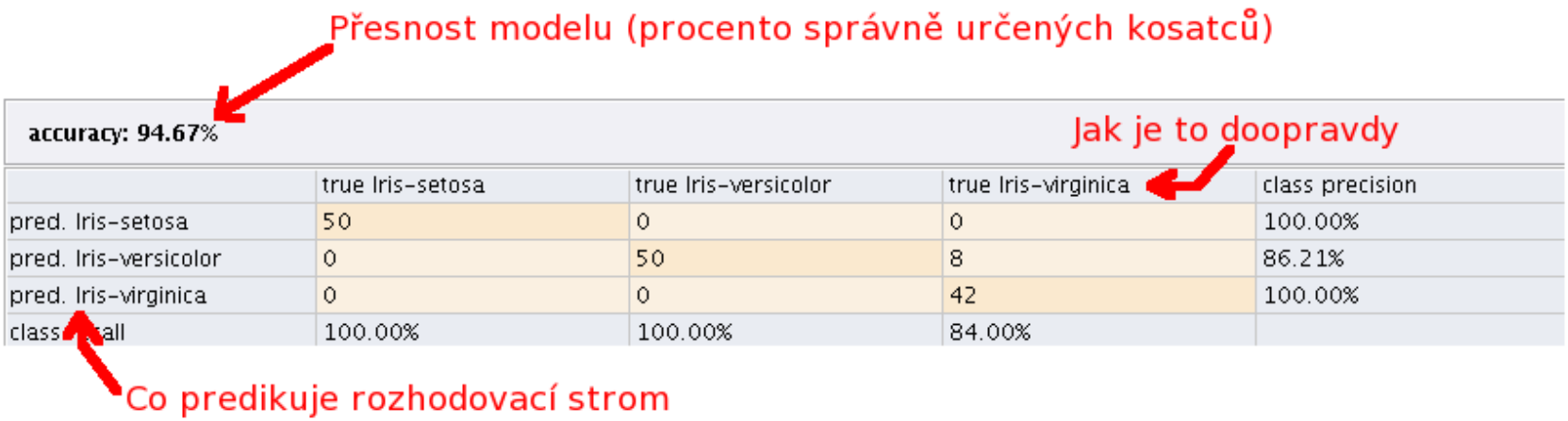
Zobrazení úspěšnosti modelu na trénovacích datech

- Kolik procent kosatců je špatně klasifikováno?
- Abyste je nemuseli počítat ručně, Rapidminer nabízí uzel *Performance*.
- Zapojte ji tedy na konec proudu a jeho výstup opět přiveďte k pravému okraji.



Přesnost a matice záměn

- Ve výstupu uzlu performance můžete uvidět jednak procento správně klasifikovaných kosatců a druhak matici záměn.
- Procento správně oklasifikovaných vzorů říká (celkem nepřekvapivě), v kolika procentech se povedlo modelu zařadit vzor (kosatec) do správné třídy.
- Matice záměn říká do kterých tříd model klasifikoval jaké množství instancí a jak je to s nimi ve skutečnosti.



DŮLEŽITÉ UPOZORNĚNÍ

- POZOR – procento úspěšně klasifikovaných vzorů (v tomto případě kosatců) vůbec neříká, jak moc je model dobrý. Ještě navíc, když testujeme model na datech, na kterých jsme jej učili!
  - Je to jako kdybychom vám dali otázky dopředu a pak se na ně ptali. Tím bychom vás nezkoušeli z látky, ale z toho, jak si umíte zapamatovat správné odpovědi.
  - A pak vás i jen trochu jinak formulovaná otázka dokonale zmate, stejně tak může model zmást trošku jiný vzor (kosatec).
- Tomu se lze čelit například tím, že vytvořený modelu necháte oklasifikovat data, která doposud neviděl – tzv. testovací množinu.
- Tím získáte trochu lepší představu, jak je model dobrý. (Ale existují i další techniky, o kterých si řekneme v průběhu semestru).

2. zápočtová úloha - zadání

- Na stránkách předmětu si vyberete data, která budete zpracovávat v programu Rapidminer a z výstupů Rapidmineru vytvoříte krátký report.
- Váš proud by měl dělat zhruba následující:
  - Načte vaše vybraná a stažená data.
  - Rozdělí data na trénovací a testovací množinu v poměru 2:1 (buď pomocí uzlu Rapidmineru nebo vytvořte v matlabu skript, který to za vás udělá. Pak jen načtete do Rapidmineru 2 množiny).
  - Vytvoříte rozhodovací strom z trénovacích dat.
  - Zjistěte chybu vytvořeného stromu na trénovacích a testovacích datech.

2. zápočtová úloha - obsah reportu

Váš report by měl obsahovat následující výstupy z Rapidmineru:

- Základní statistiku vstupních dat (pro každý sloupec průměr, rozptyl pro číselné atributy, počty hodnot pro nominální). Můžete přidat i bodové grafy nebo matici bodových grafů (scatter plot nebo scatter plot matrix), případně jiné grafy, pokud se vám budou zdát užitečné.
- Vizualizaci rozhodovacího stromu (obrázek) a tento strom přepsaný do formy if-then podmínek (použijte Javovskou nebo Matlabovskou syntaxi).
- Matici záměn (confusion matrix) pro trénovací a testovací data, přesnost klasifikace a krátký komentář, jestli se vám zdá přesnost (accuracy) dostatečná, případně která třída přesnost kazí.

Další zdroje

- Stránky Rapidmineru: `http://rapid-i.com`
- Video tutoriály, základní práce s RapidMinerem: `http://rapid-i.com/content/view/189/212/lang,en`
- Uživatelský manuál v PDF `http://sourceforge.net/projects/rapidminer/files/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf/download`