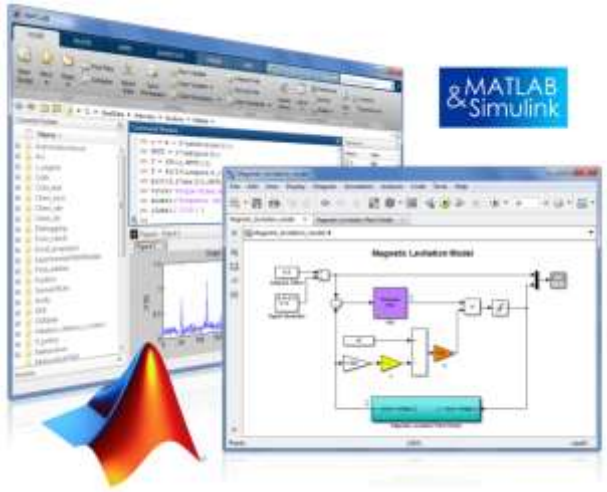
HUMUSOFT®

Datová analytika

Machine Learning a Big Data

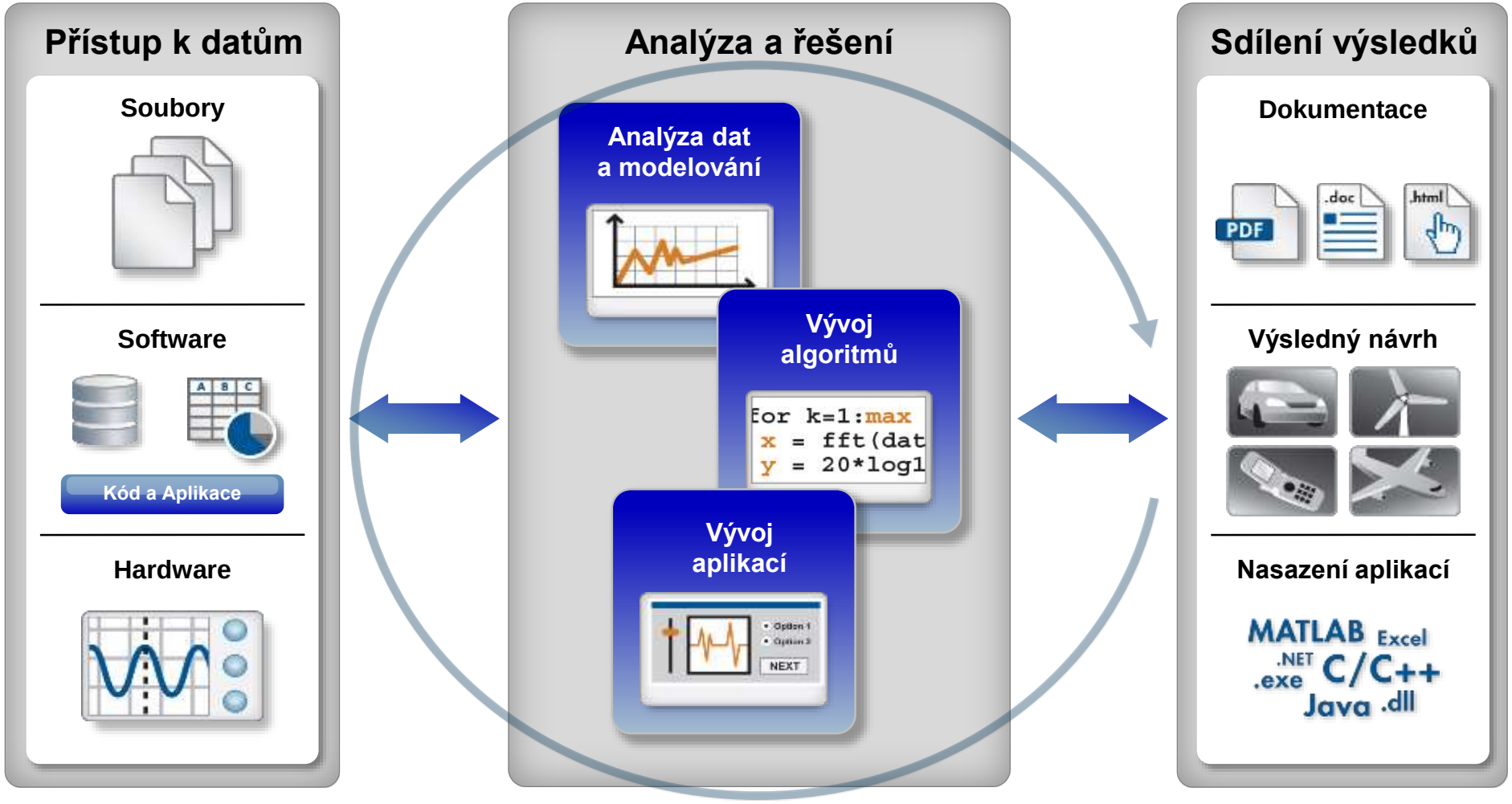


Jan Studnička
studnicka@humusoft.cz

www.humusoft.cz
info@humusoft.cz

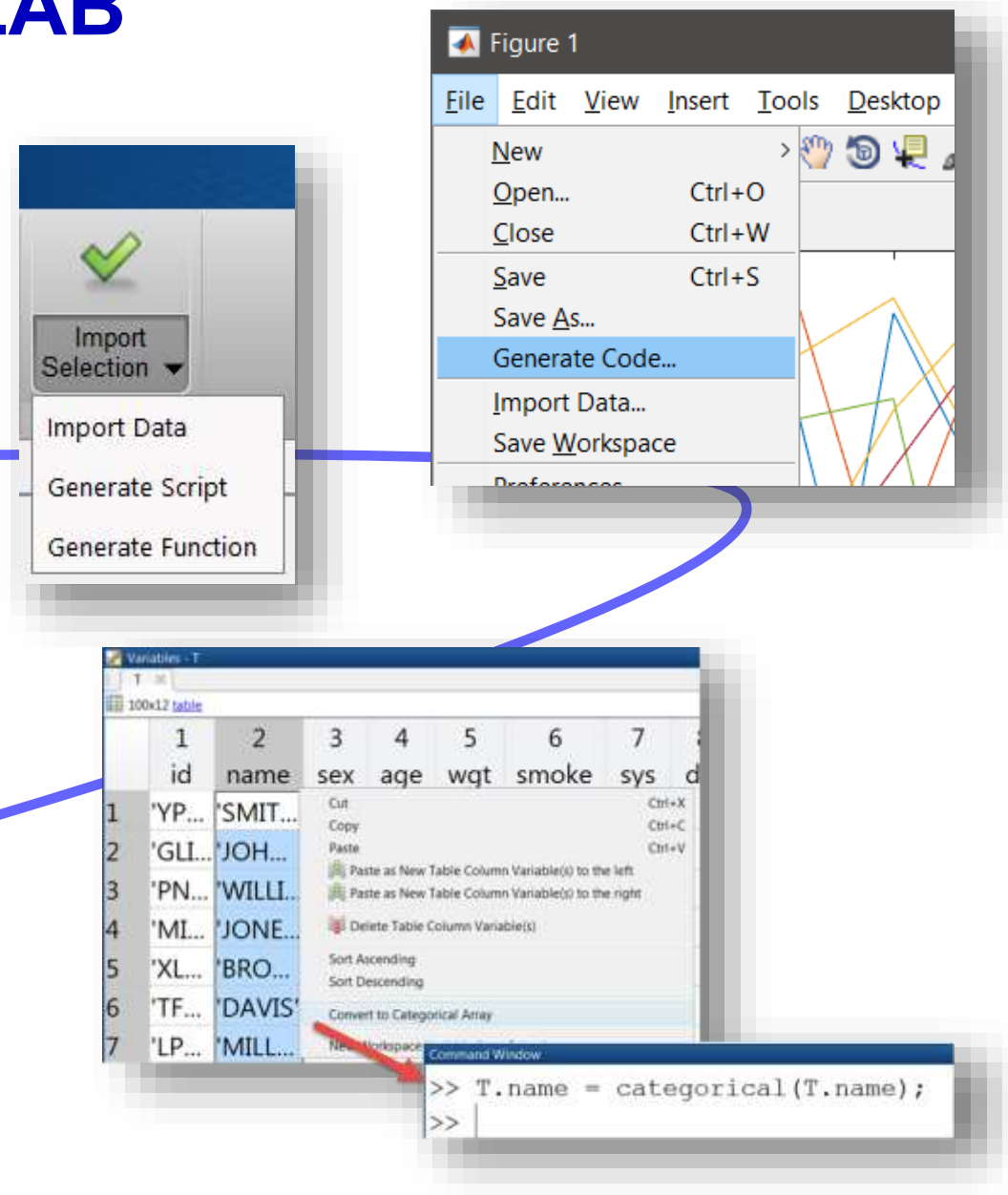
www.mathworks.com

Technické výpočty v MATLABu



Způsoby práce v prostředí MATLAB

- **Zápis kódu**
  - Plnohodnotný programovací jazyk
  - Zadávání příkazů
  - Skripty
  - Funkce
  - Objektově orientované programování
- **Interaktivní přístup**
  - Využití grafických nástrojů
    - Umožnují generovat kód

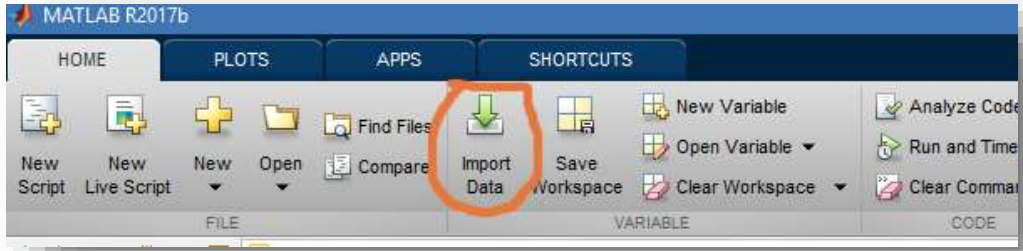


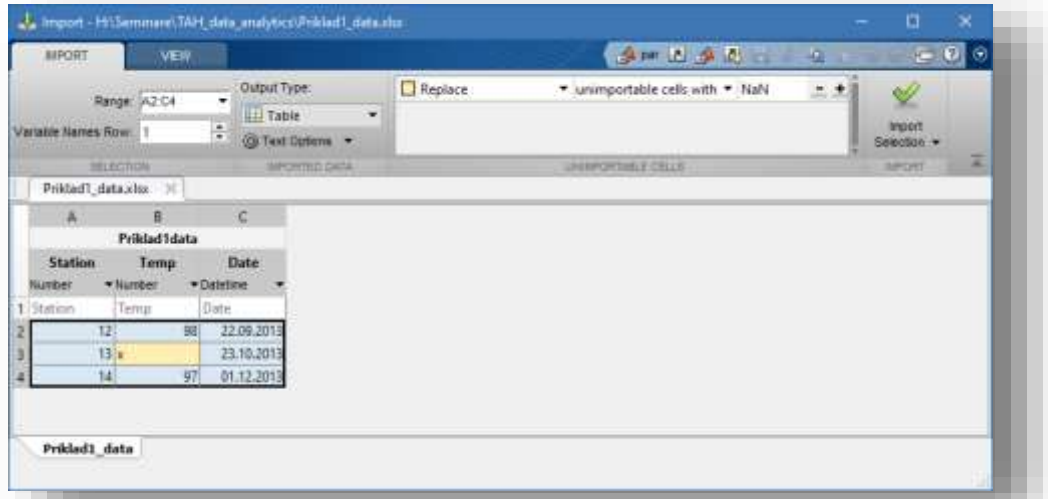
Analýza dat

načítání dat a data v MATLABu

Načítání dat ze souborů

- **Interaktivní nástroj**
  - Import Tool





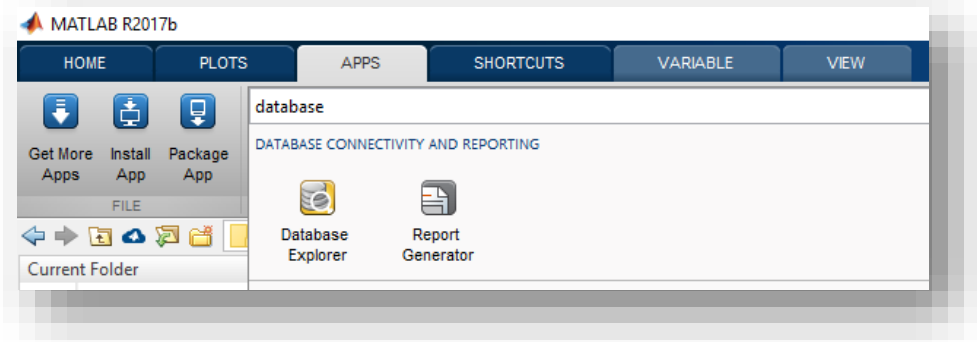
- **Funkce**
  - `readtable`
    - **Načte data ze souboru do tabulky v MATLABu**
    - **Nastavitelné parametry načítání**

R2016b – `detectImportOptions`

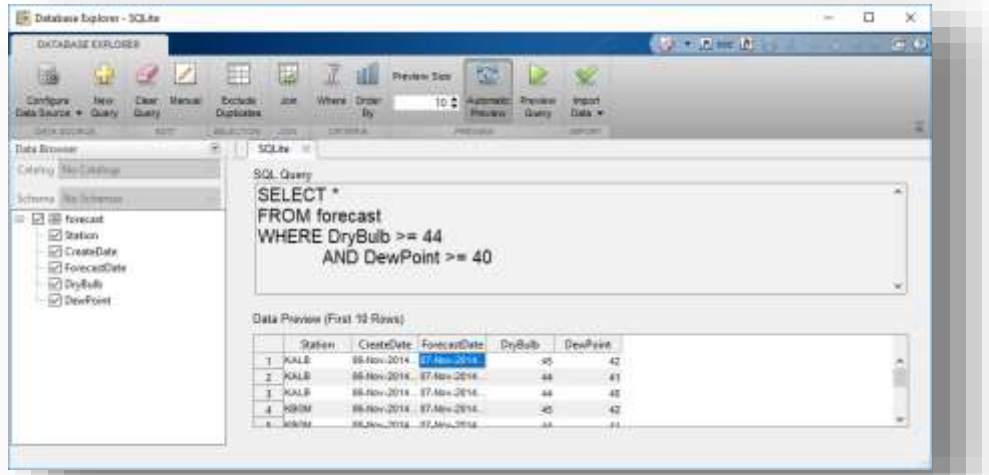
- **Parametry pro funkci `readtable` detekuje z dat**

www.mathworks.com/examples/matlab/mw/matlab-ex22647148

Načítání dat z databází



- **Interaktivní nástroj**
  - Database Explorer
    - **ODBC a JDBC kompatibilní databáze**



- **Funkce**
  - `database`
    - **Vytvoří připojení k databázi**

R2017a
  - `select`
    - **Načte tabulku z databáze do tabulky v MATLABu**
    - **datový typ importovaných numerických dat určen z databáze**

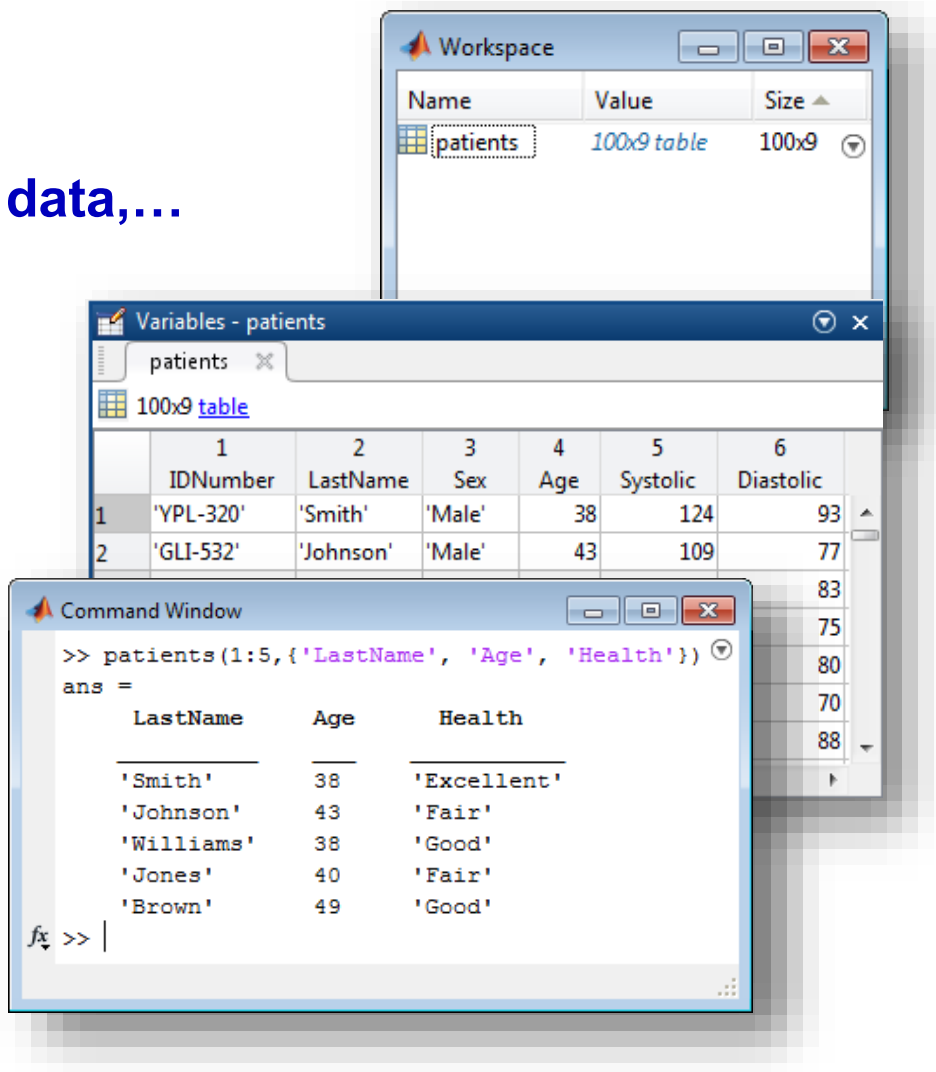
Další Import a Export Dat

- **Přístup k online datům**
  - RESTful, JSON, HTTP, CSV, text, a obrázková data
  - `webread, tcpclient` R2014b
- **Načítání velkých kolekcí dat**
  - `datastore` R2014b
    - Text, spreadsheet, database, image
    - custom format datastores R2017b

Tabulky → `table`

R2013b

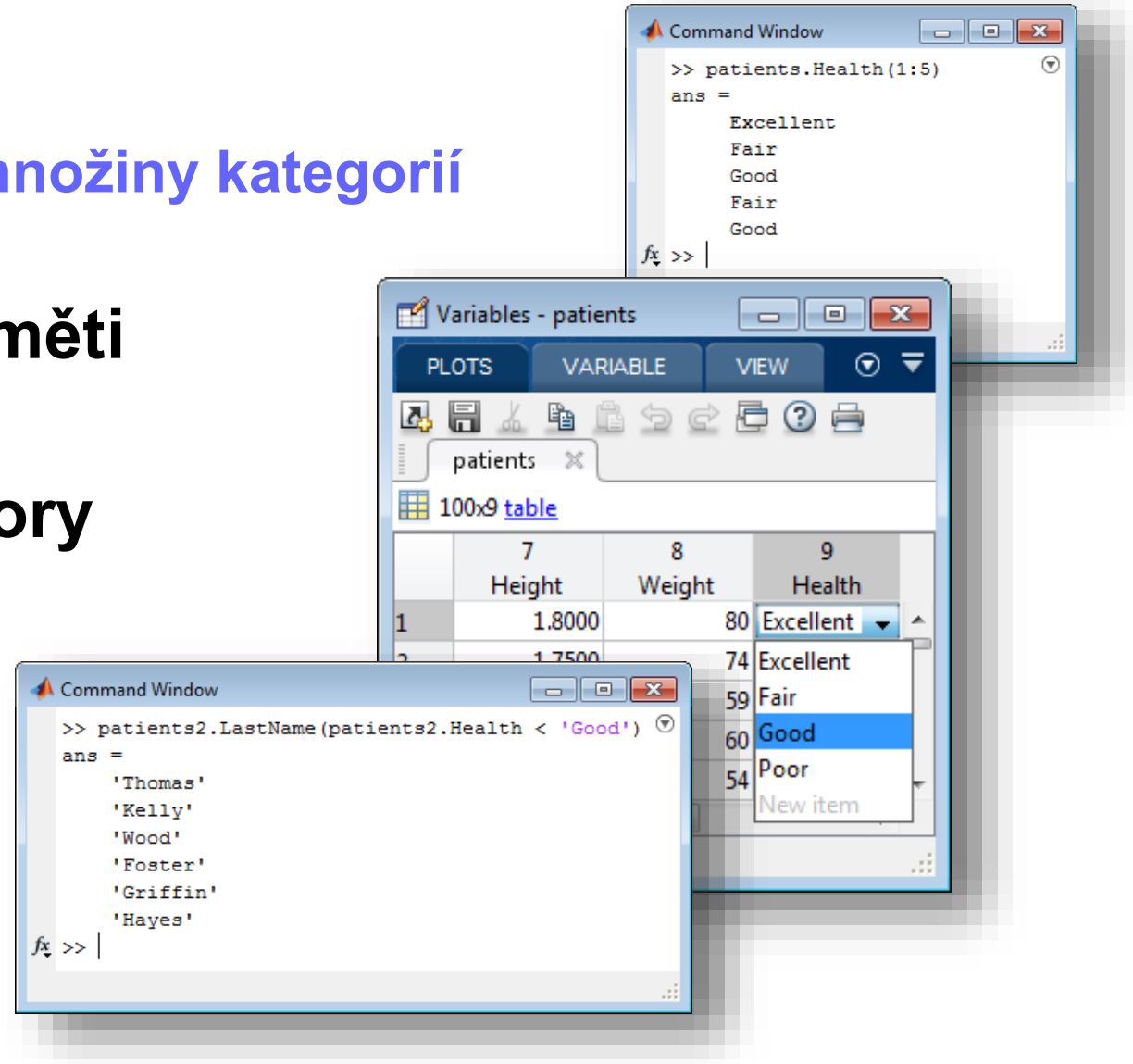
- **Různorodá data v tabulkovém formátu**
  - Různé typy dat v různých proměnných
    - Text, numerická data, časová data, kategoriální data,…
  - Drží data i metadata
- **Zpracování a analýza**
  - Voláním funkcí přímo na tabulky
    - Zpracování chybějících a odlehlých údajů
    - Třídění, přeskládání a spojování tabulek
    - Souhrnné statistiky
    - Tvorba modelu, predikce / klasifikace



Kategoriální data → `categorical`

HUMUSOFT
R2013b

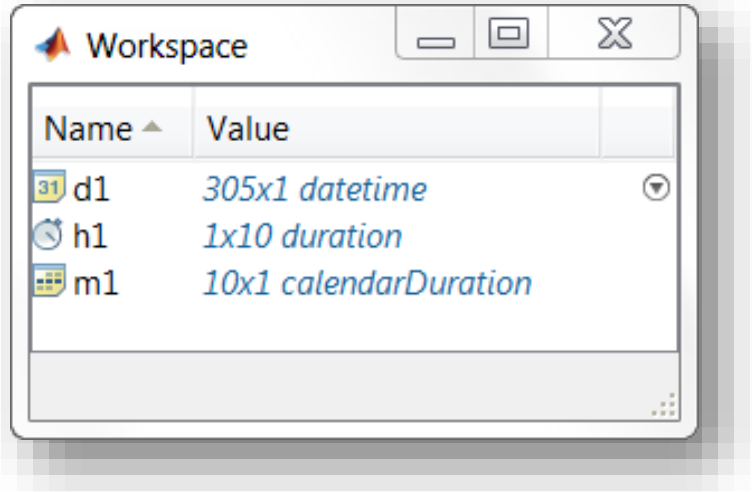
- **Diskrétní nenumerická data**
  - Data nabývají hodnot z konečné množiny kategorií
- **Efektivní z hlediska zabrané paměti**
- **Porovnáváme logickými operátory**
  - ==, ~=
- **Můžeme zavést uspořádání**
  - <, <=, >, >=



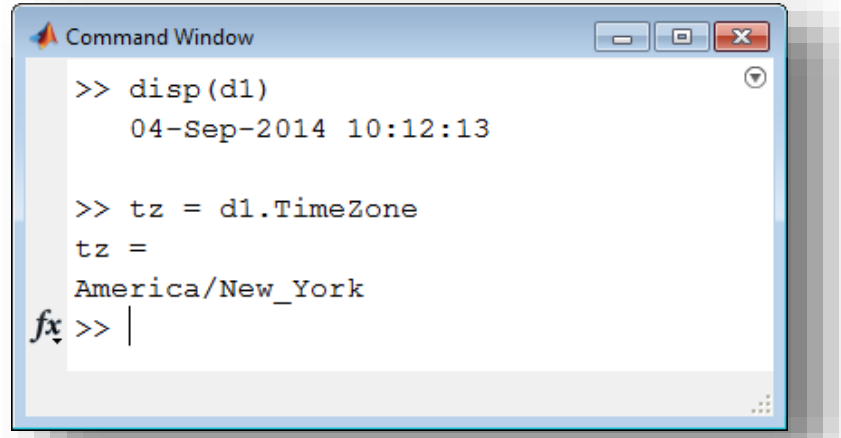
Datum a čas

HUMUSOFT
R2014b

- `datetime`
  - reprezentuje datum a čas (body na časové ose)
- `duration`, `calendarDuration`
  - reprezentují dobu trvání (časové intervaly)



- **Slouží pro výpočty i zobrazení**
  - sčítání, odčítání, seřazení, porovnání, vykreslení
  - nastavitelný formát zobrazení
  - přesnost na nanosekundy
  - časové zóny, přestupné sekundy, letní čas



HUMUSOFT®

R2016b

Časové tabulky → `timetable`

Time	Day	Total	Westbound	Eastbound
06/24/2015 00:00:00	Wednesday	13	9	4
06/24/2015 01:00:00	Wednesday	3	3	0
06/24/2015 02:00:00	Wednesday	1	1	0
06/24/2015 03:00:00	Wednesday	1	1	0
06/24/2015 04:00:00	Wednesday	1	1	0
06/24/2015 05:00:00	Wednesday	7	3	4
06/24/2015 06:00:00	Wednesday	36	6	30
06/24/2015 07:00:00	Wednesday	141	13	128
06/24/2015 08:00:00	Wednesday	327	44	283
06/24/2015 09:00:00	Wednesday	184	32	152

- **Tabulky s časovými značkami pro jednotlivé řádky**
  - indexování dle času
- **Zpracování dat pomocí specializovaných funkcí**
  - reorganizace dat
  - změna časové škály
  - synchronizace a spojení časových tabulek
- **Zpracování dat pomocí funkcí pro `table`**

www.mathworks.com/examples/matlab/mw/matlab-ex28366094

Text → `string`

HUMUSOFT

R2016b

- **Efektivní práce s textovými daty**

```
>> "image" + (1:3) + ".png"
```

R2017a

```
1×3 string array
 "image1.png"   "image2.png"   "image3.png"
```

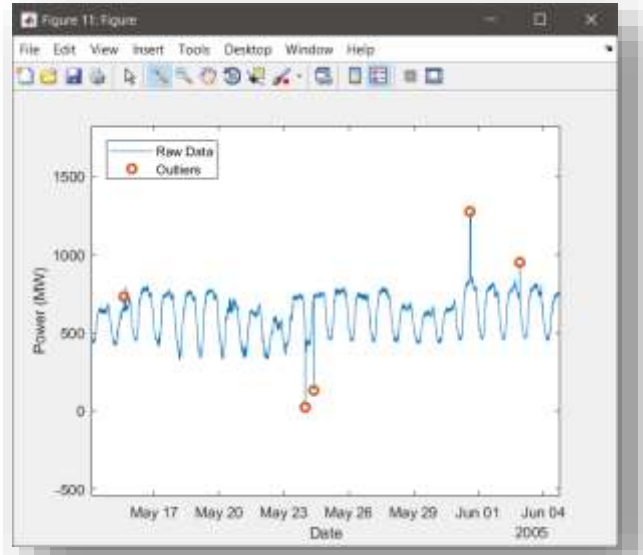
- **Příklad: Ověření, zda je v textovém řetězci obsažen jiný text**
  - Dříve: `if ~isempty(strfind(textdata,"Dog"))`
  - Nyní: `if contains(textdata,"Dog"))`

    - Až 50x rychlejší výpočet s funkcí `contains` a datovým typem `string` než se `strfind` a `cellstr`.
    - Až 2x méně využité paměti s datovým typem `string` oproti `cellstr`.

HUMUSOFT

Předzpracování dat

R2016b

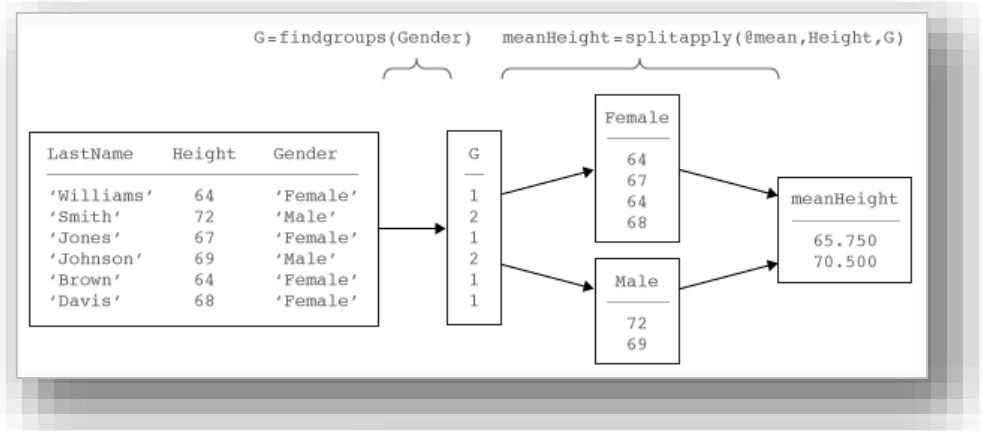


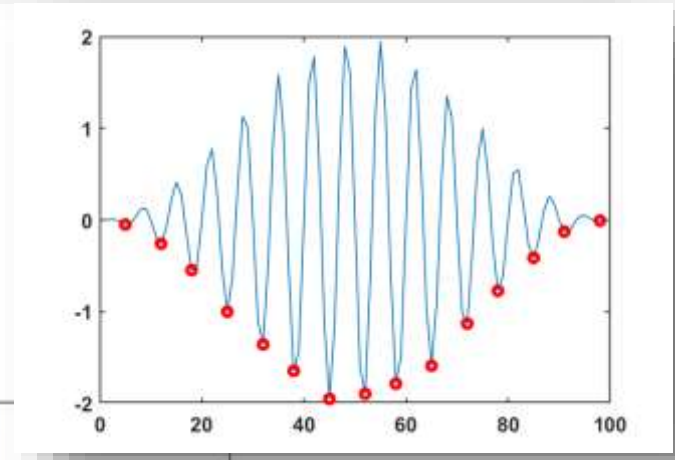
- Zpracování chybějících dat pomocí `*missing` funkcí
- Manipulace s textem pomocí `replace`, `contains`, `endsWith`, a dalších…
- Vyhlazení dat pomocí filtrace nebo lokální regrese
  - `smoothdata` R2017a
- Práce s odlehlými pozorováními
  - `isoutlier`, `filloutliers` R2017a

www.mathworks.com/examples/matlab/mw/matlab-ex52135330

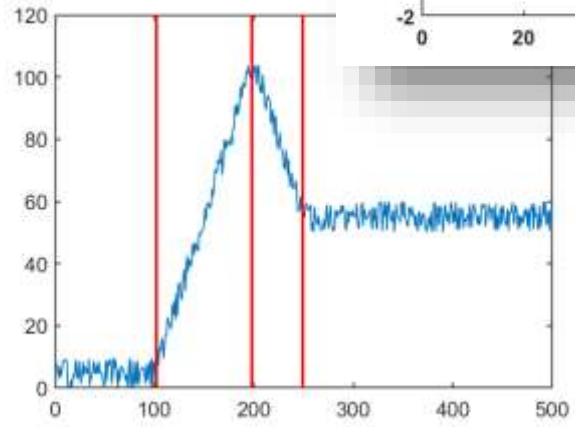
Analýza dat

- Split-Apply-Combine Workflow
  - `findgroups` R2015b
    - rozdělí data do skupin
  - `splitapply`
    - aplikuje funkci na každou skupinu a zkombinuje výsledky





- Detekce lokálních extrémů R2017b `islocalmin` and `islocalmax`

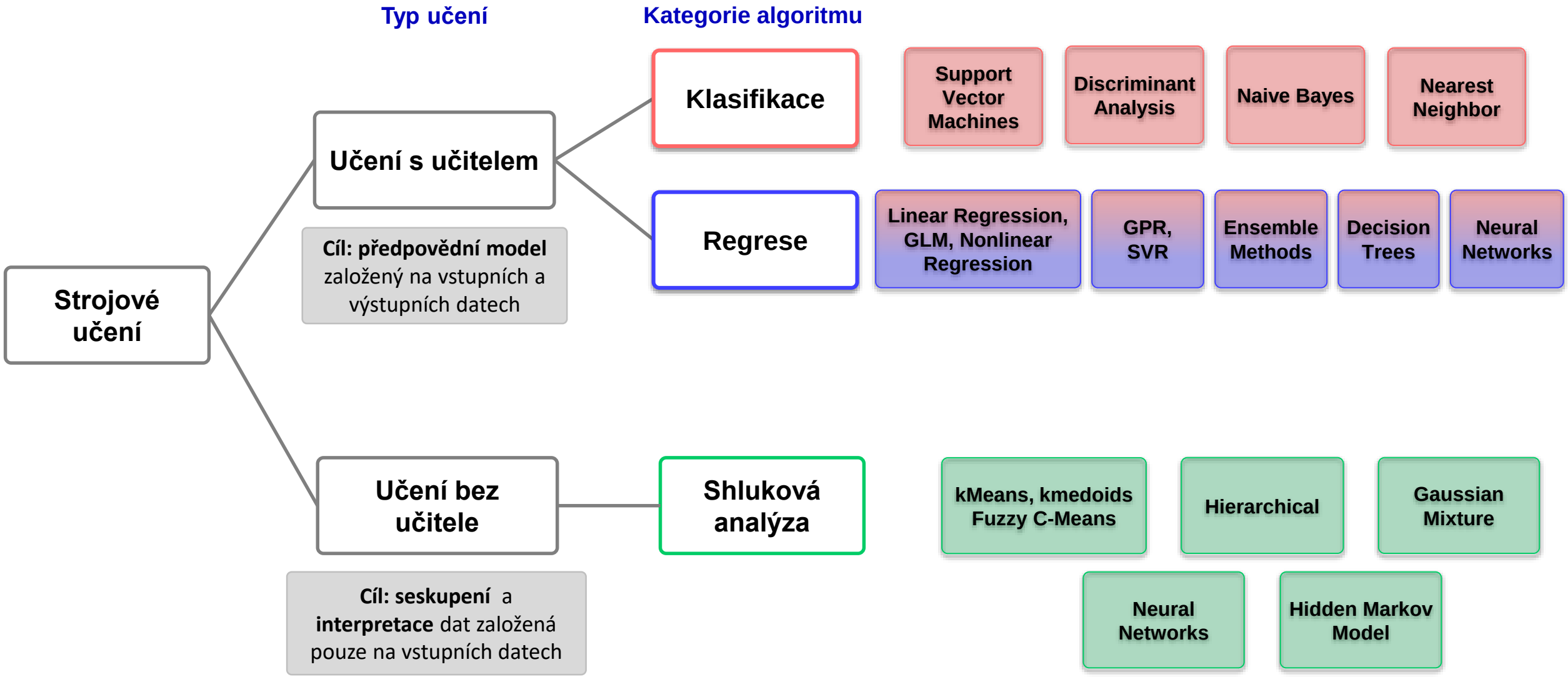


- Detekce náhlých změn v datech `ischange` R2017b

Machine Learning

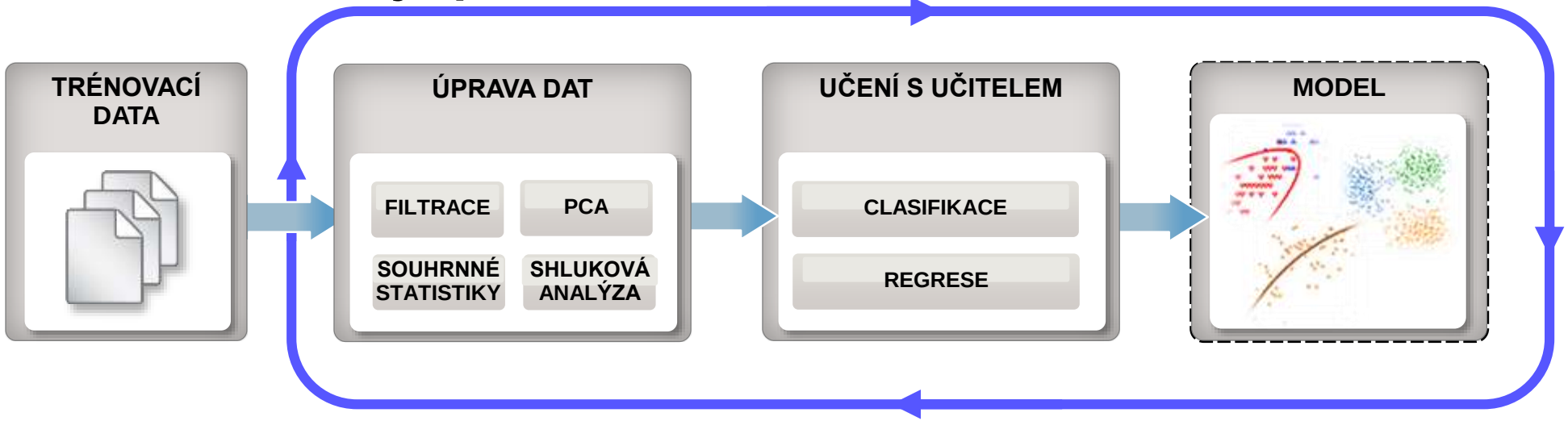
nové algoritmy a přístupy

Strojové učení

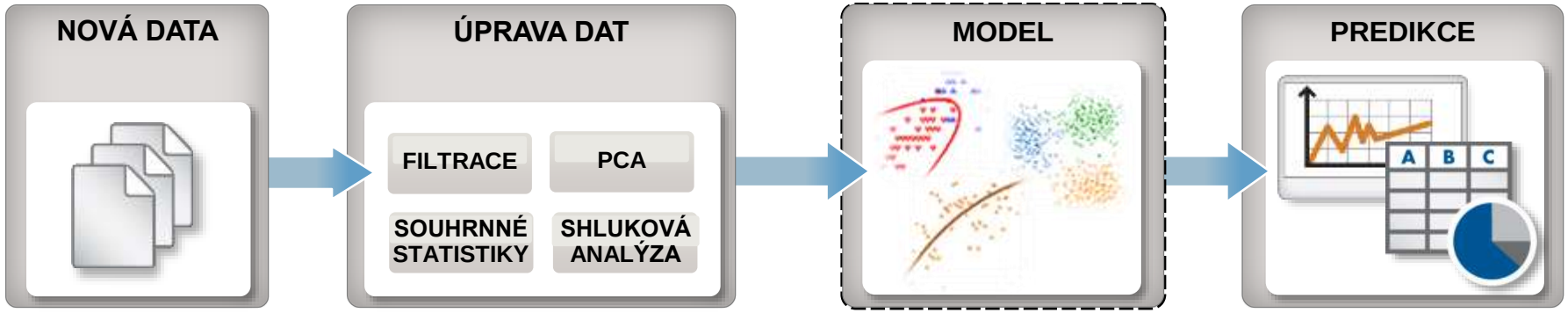


Postup při strojovém učení

Fáze učení: hledání nejlepšího modelu

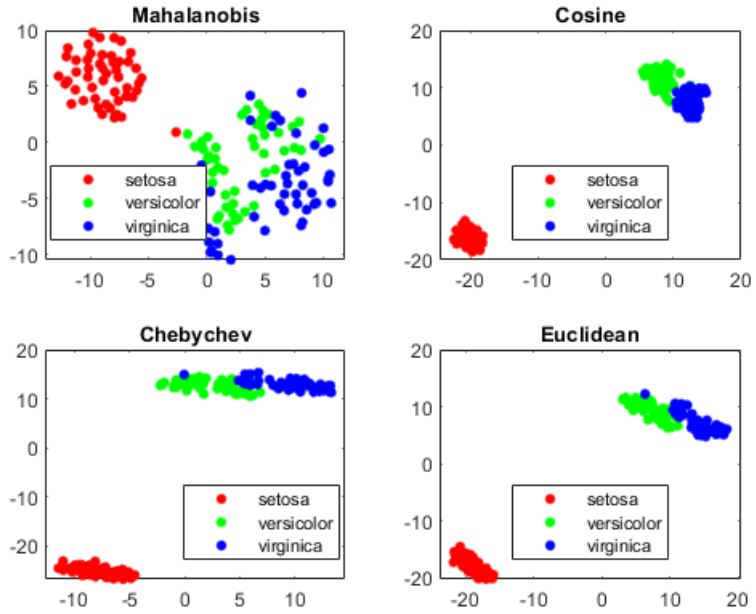


Predikce: začlenění výsledného modelu do koncové aplikace



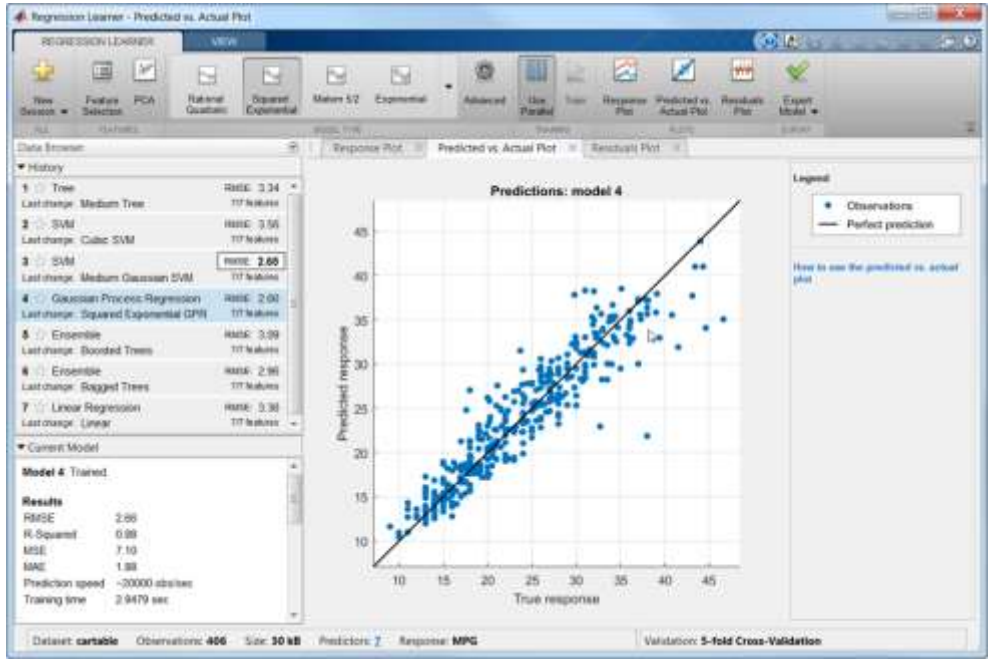
Výběr a extrakce prediktorů

- **Výběr prediktorů**
  - Neighborhood component analysis
    - fscnca, fsrnca
- **Extrakce prediktorů**
  - Sparse filtering, Reconstruction ICA
    - sparsefilt, rica
    - transform
- **Vizualizace vícedimenzionálních dat**
  - t-Distributed Stochastic Neighbor Embedding
    - tsne



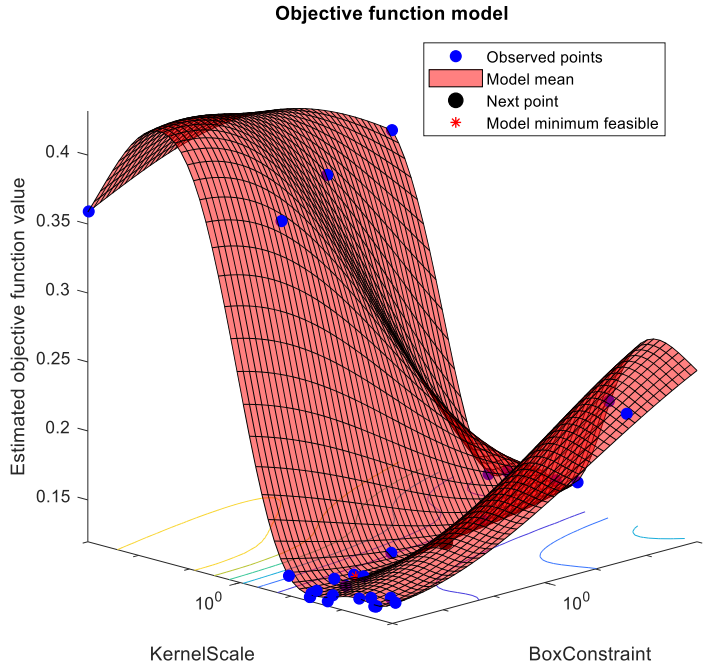
Klasifikace a regrese

- **Trénování modelu**
  - *fit**
    - **SVM, ECOC, Ensemble, regrese,…**
  - *stepwise**
    - **kroková regrese**
- **Predikce a diagnostiky**
  - **vlastnosti a metody modelu**
    - predict …



Nové efektivní algoritmy

- **optimalizace hyperparametrů**
  - bayesovská optimalizace
- **klasifikace a regrese pro Big Data**
  - techniky pro snížení výpočetní náročnosti trénování modelu
    - SVM, ECOC, logistická regrese
      - stochastic gradient descent, LBFGS, aj.
    - Gaussian kernel classification / regression
      - random feature expansion



www.mathworks.com/examples/statistics/mw/stats-ex21115654

Nové interaktivní aplikace

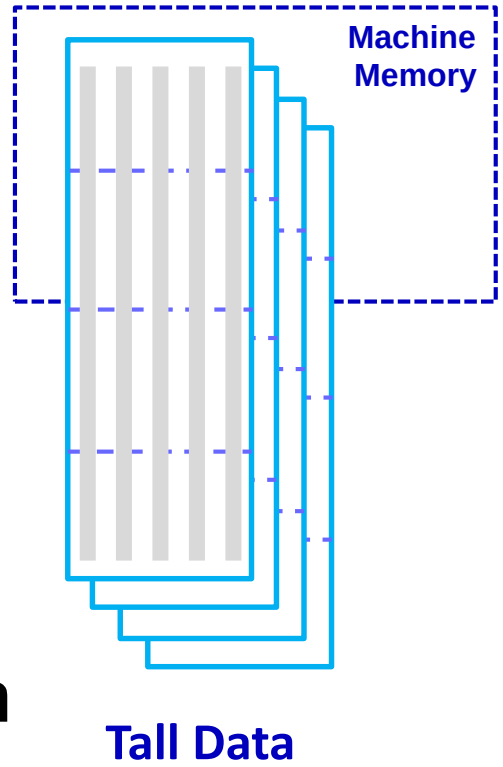
Aplikace	Toolbox	Uvedeno
Classification Learner	Statistics and Machine Learning	R2015a
Signal Analyzer	Signal Processing	R2016b
Regression Learner	Statistics and Machine Learning	R2017a
Wavelet Signal Denoiser	Wavelet	R2017b
Econometric Modeler	Econometrics	R2018a

Big Data

tall array

Big Data → `tall`

- **Případy využití:**
  - **Sloupcová data – s mnoho řádky**
  - **Příliš mnoho dat, aby se vešly do paměti**
  - **Operace jsou povahou statistické**
- **Pro statistické výpočty i Machine Learning**
  - **Stovky funkcí v základním MATLABu a Statistics and Machine Learning Toolbox**
- **Automaticky optimalizuje přistupování k datům**
- **Na desktopu, clusteru, nebo clusteru s nástroji Hadoop a Spark**



www.mathworks.com/examples/statistics/mw/stats_featured-ex42806539

Big Data bez velkých změn

Jeden soubor

Access Data

```
measured = readtable('PumpData.csv');
measured = table2timetable(measured);
```

Preprocess Data

Select data of interest

```
measured = measured(timerange(seconds(1),seconds(2)),:)
```

Work with missing data

```
measured = fillmissing(measured,'linear');
```

Calculate statistics

```
m = mean(measured.Speed);
s = std(measured.Speed);
```

Sto souborů

Access Data

```
measured = datastore('PumpData*.csv');
measured = tall(measured);
measured = table2timetable(measured);
```

Preprocess Data

Select data of interest

```
measured = measured(timerange(seconds(1),seconds(2)),:)
```

Work with missing data

```
measured = fillmissing(measured,'linear');
```

Calculate statistics

```
m = mean(measured.Speed);
s = std(measured.Speed);
```

```
[m,s] = gather(m,s);
```

Big Data Workflow

Access Data

- Text
- Spreadsheet (Excel)
- Database (SQL)
- Custom Reader

Datastores for common types of structured data

Tall Data Types

- table
- cell
- double
- numeric
- cellstr
- datetime
- categorical

Tall versions of commonly used MATLAB data types

Exploration & Pre-processing

- Numeric functions
- Summary stats reductions
- Date/Time capabilities
- Categorical
- String processing
- Table wrangling
- Missing data handling
- Summary visualizations:
- Histogram/histogram2
- Kernel density plot
- Bin-scatter

Hundreds of pre-built functions

Machine Learning

- Linear Model
- Logistic Regression
- Discriminant Analysis
- Decision trees classif.
- K-means
- PCA
- Random data sampling
- Summary statistics
- Validation techniques

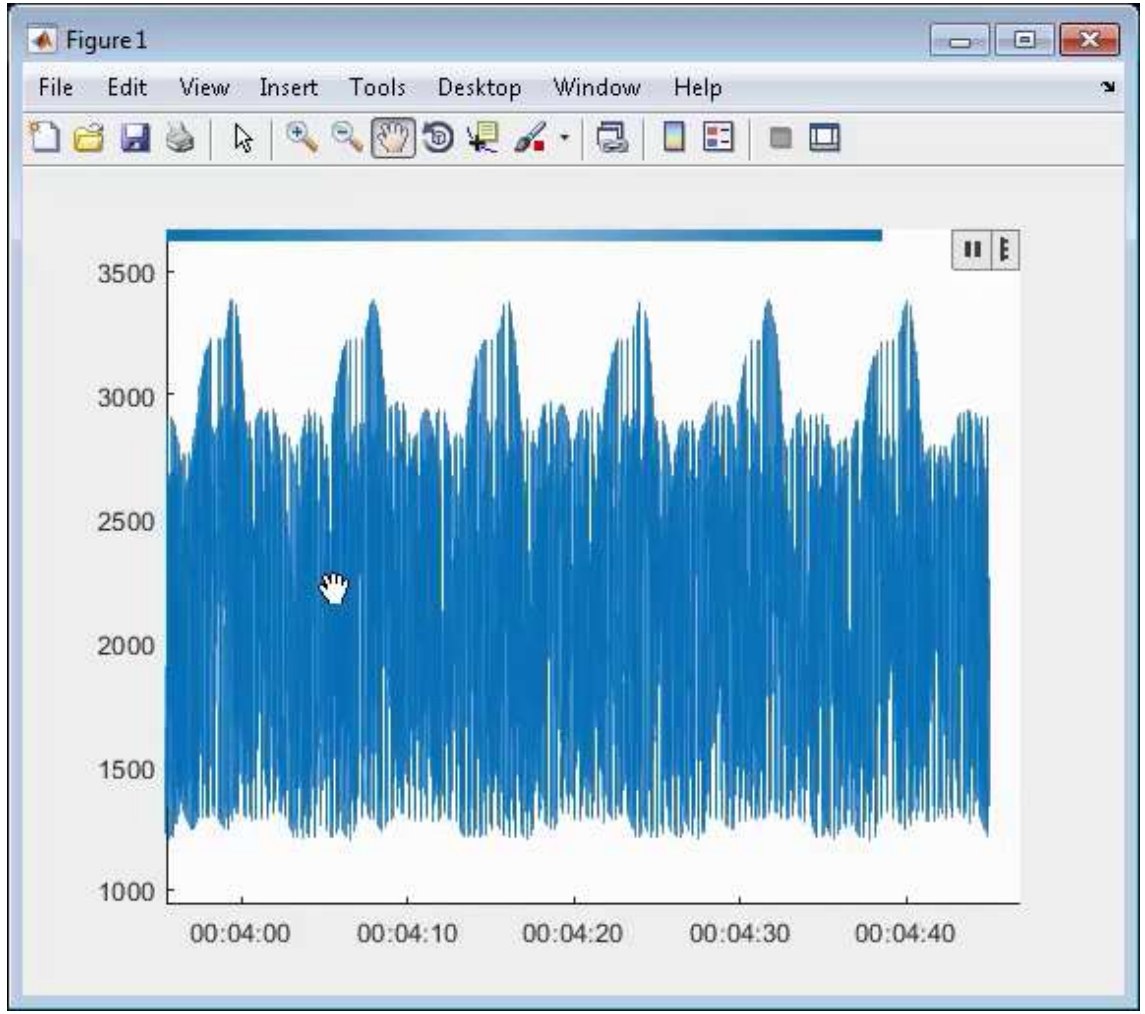
Key statistics and machine learning algorithms

“Tall” data types and functions for use with out-of-memory data

Big Data – vizualizace pomocí `tall`

- **Podpora pro:**
  - `histogram`
  - `histogram2`
  - `ksdensity`
  - `plot`
  - `scatter`
  - `binscatter`
  - `confusionmat`

- **Podpora dále bude růst!**

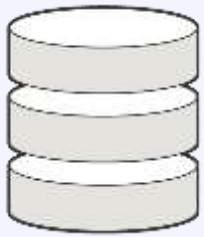


Tall Array

- Run in parallel on compute clusters
  MATLAB Distributed Computing Server

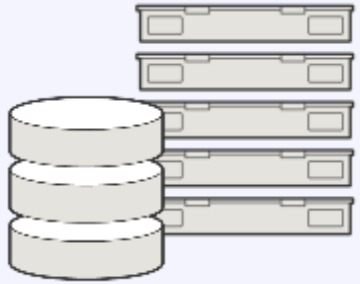
Compute Clusters

Local disk
Shared folders
Databases



- Tall arrays
  MATLAB
- 100's of functions supported
  MATLAB
  Statistics and Machine Learning Toolbox
- Run in parallel
  Parallel Computing Toolbox

Spark + Hadoop



- Run in parallel on Spark clusters
  MATLAB Distributed Computing Server
- Deploy MATLAB applications as standalone applications on Spark clusters
  MATLAB Compiler

Kontakty

Pobřežní 20

186 00 Praha 8

Česká republika

Email: info@humusoft.cz

www.humusoft.cz

Tel.: +420 284 011 720

www.facebook.com/humusoft

www.youtube.com/humusoft

www.twitter.com/humusoft