

Before we start...

Save some time by downloading these files

- Download Orange <https://orange.biolab.si/>
- Download Datasets <http://bit.ly/44con-ml>

Orange is the new Hack

***AN INTRODUCTION TO MACHINE LEARNING
WITH ORANGE***

44CON

Presented by Philippe Arteau

Who am I ?

- Philippe Arteau
- Security Researcher at CounterTack GoSecure
- Open-source developer
 - Find Security Bugs (SpotBugs - Static Analysis for Java)
 - Security Code Scan (Roslyn – Static Analysis for .NET)
 - Burp and ZAP Plugins (Retire.js, CSP Auditor)
- Machine Learning Enthousiast



Agenda

- Machine Learning introduction
 - Definition
 - Supervised vs Unsupervised
 - ...
- Hands on exercises
 - Data visualization exercises
 - Classification exercises
 - Using/Building custom plugin
- Conclusion

This workshop is for you

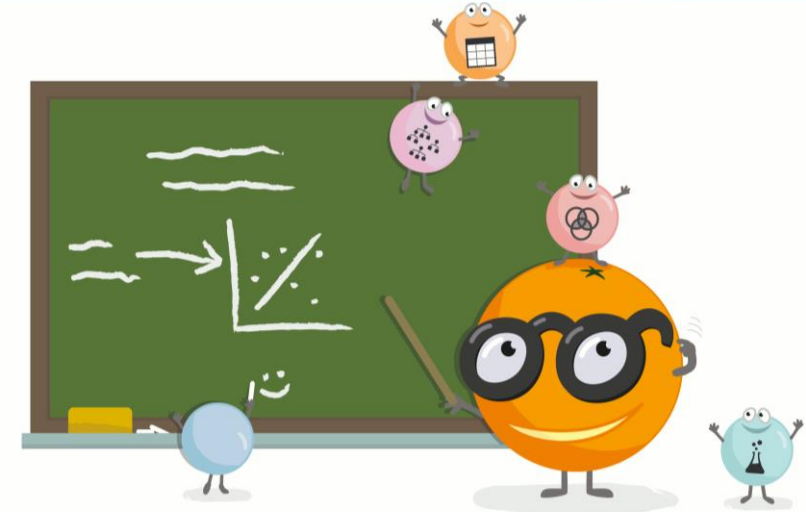
Knowledge requirement

- You don't need prior knowledge of Machine Learning
 - I have no statistics or machine learning specialization myself !
- Machine Learning is a technique that can be applied to many fields
- Aside from learning the basic principles, the workshop might give you some ideas for future applications



Tool used in this workshop

- Orange is a python Machine Learning library
- Orange has a powerful UI
 - No programming knowledge required
- Many framework can be used to reproduce the exercise
 - I highly encourage you to try other tools in your projects
- I find Orange extremely useful for prototyping, visualization and **training**
- It may not scale for datasets that can't fit in RAM

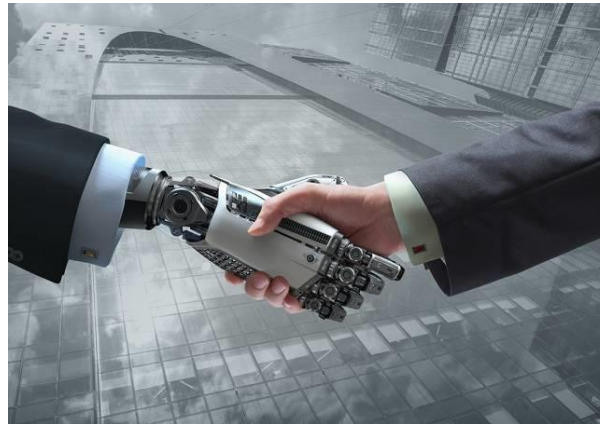


The background of the slide features a complex network diagram. It consists of numerous nodes, represented by small circles and larger hexagons, connected by a web of thin red lines. Some nodes are highlighted with larger, more prominent hexagonal shapes. The network is dense and spans the entire width of the slide, with a horizontal white band in the center containing the text.

Machine Learning?

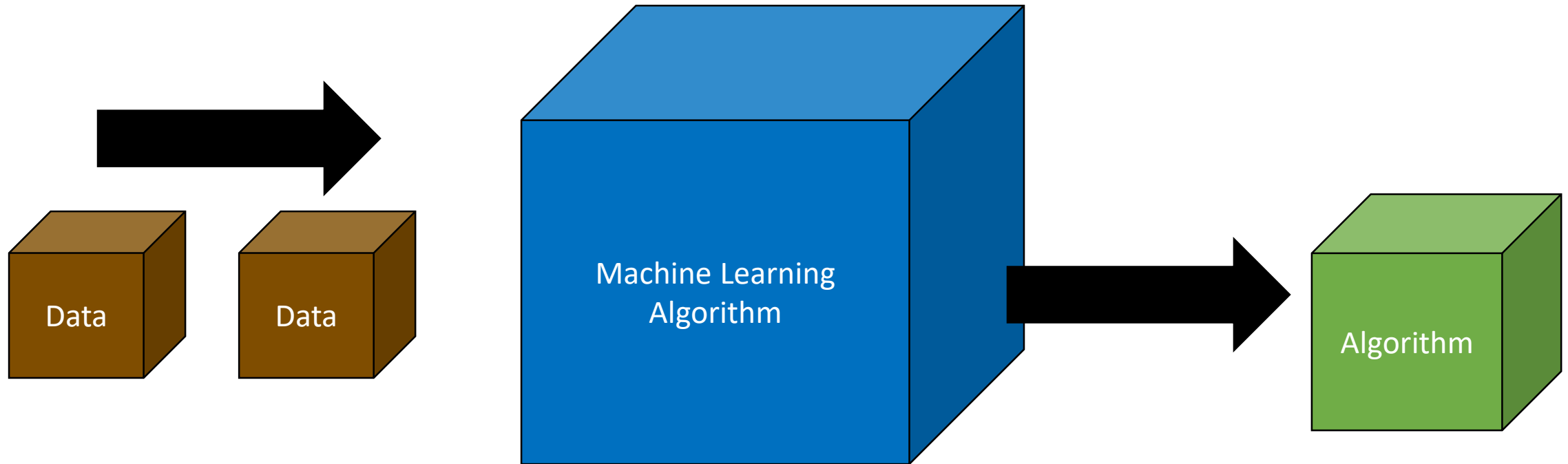
What is Machine Learning?

“Machine learning is a field of computer science that uses **statistical techniques** to give computer systems the ability to “**learn**” with data, without being **explicitly programmed**.”



- https://en.wikipedia.org/wiki/Machine_learning

“without being explicitly programmed”



Data-driven algorithm

```
def calc_price_house(year,nb_rooms, ...):
```

```
    price = nb_rooms * 50000
```

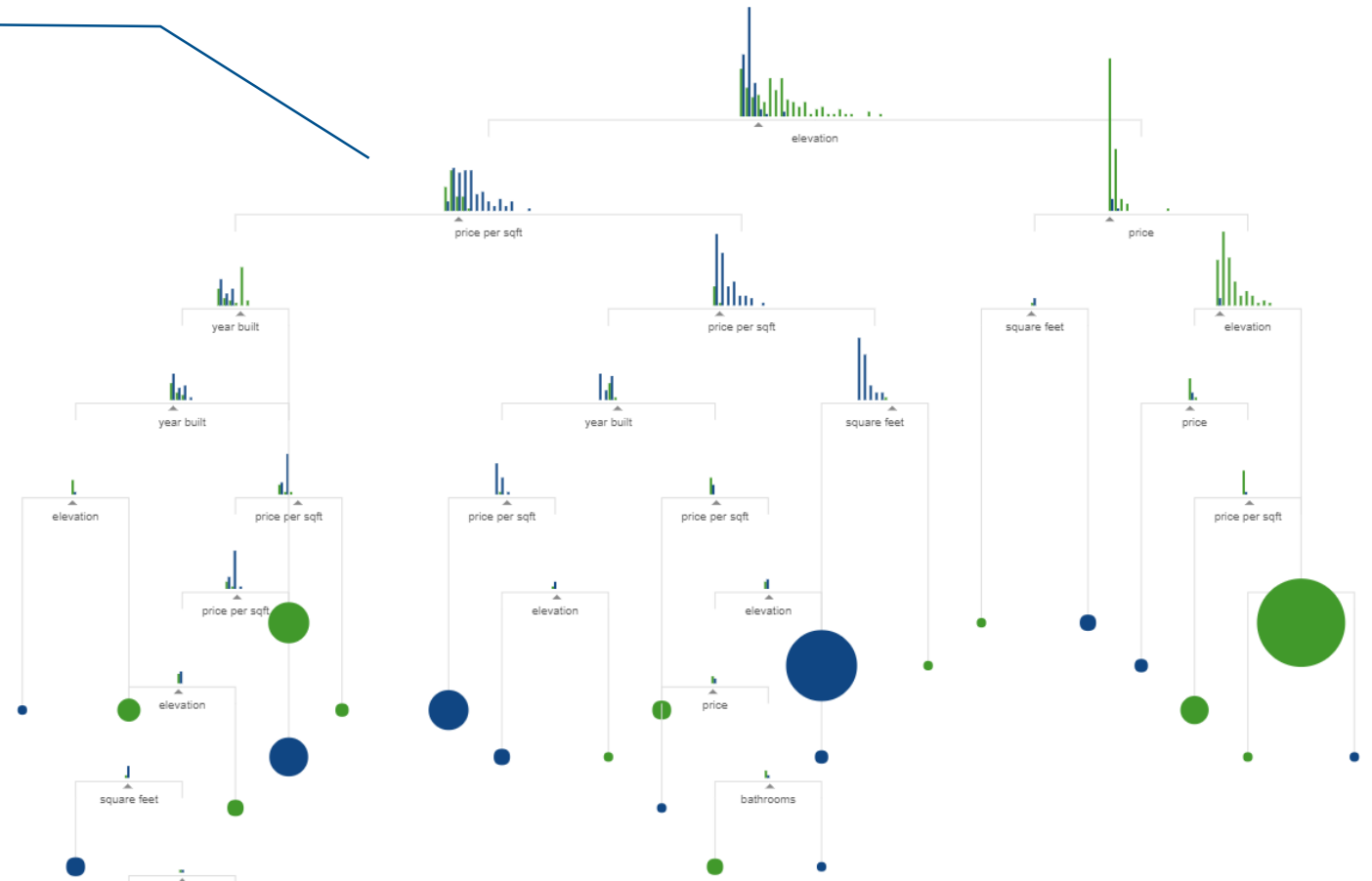
```
    if(year <= 1950):
```

```
        price += 40000
```

```
    price -= year * 500
```

```
    [...]
```

```
    return price
```



<http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>

Supervised vs Unsupervised



Supervised Classification vs Regression

Supervised learning has two variations

Supervised classification

Attempts to predict the **right answer** from a discrete number of possibilities

Prediction Features ->

- Recommend or not to user
- Types of species

Supervised regression

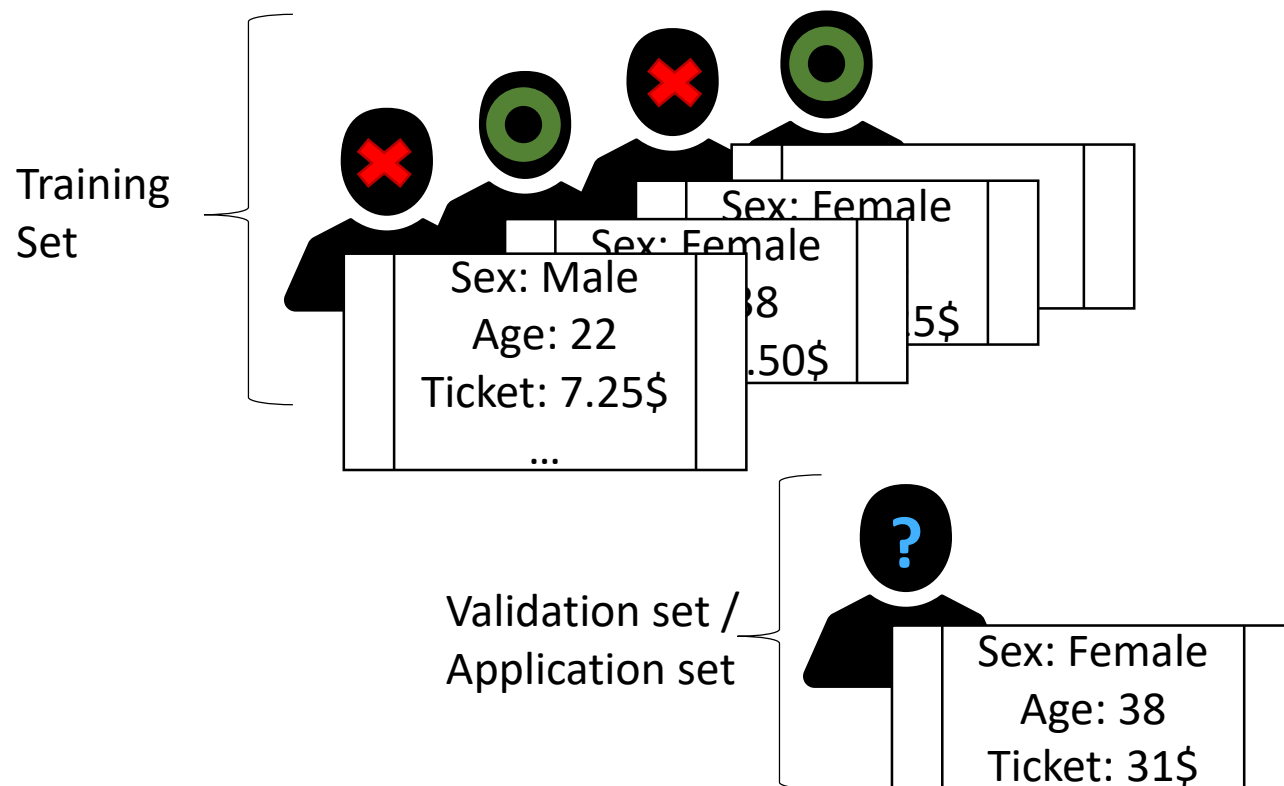
Attempts to predict a **continuous value**

Prediction Features ->

- House pricing (10k to 10M)
- Time of recovery from an incident

Supervised Classification

■ Example of prediction



Data Table

Info

891 instances
9 features (19.4% missing values)
Discrete class with 2 values (no missing values)
3 meta attributes (0.1% missing values)

Variables

☐ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection

☒ Select full rows

Restore Original Order

☒ Send Automatically

	Name	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
1	Braund, Mr. Owen Harris	Died	3	male	22.00	1	0	7.2500
2	Cummings, Mrs. John Bradley (Florence Briggs)	Survived	1	female	38.00	1	0	71.2833
3	Heikkinen, Miss. Laina	Survived	3	female	26.00	0	0	7.9250
4	Futrelle, Mrs. Jacques Heath (Lily May Peel)	Survived	1	female	35.00	1	0	53.1000
5	Allen, Mr. William Henry	Died	3	male	35.00	0	0	8.0500
6	Moran, Mr. James	Died	3	male	?	0	0	8.4583
7	McCarthy, Mr. Timothy J	Died	1	male	54.00	0	0	51.8625
8	Palsson, Master. Gosta Leonard	Died	3	male	2.00	3	1	21.0750
9	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina)	Survived	3	female	27.00	0	2	11.1333
10	Nasser, Mrs. Nicholas (Adele Achem)	Survived	2	female	14.00	1	0	30.0708
11	Sandstrom, Miss. Marguerite Rut	Survived	3	female	4.00	1	1	16.7000
12	Bonnell, Miss. Elizabeth	Survived	1	female	58.00	0	0	26.5500
13	Saunderscock, Mr. William Henry	Died	3	male	20.00	0	0	8.0500
14	Andersson, Mr. Anders Johan	Died	3	male	39.00	1	5	31.2750
15	Vestrom, Miss. Hilda Amanda Adolfin	Died	3	female	14.00	0	0	7.8542
16	Hewlett, Mrs. (Mary D Kingcome)	Survived	2	female	55.00	0	0	16.0000
17	Rice, Master. Eugene	Died	3	male	2.00	4	1	29.1250
18	Williams, Mr. Charles Eugene	Survived	2	male	?	0	0	13.0000
19	Vander Planke, Mrs. Julius (Emelia Maria Van)	Died	3	female	31.00	1	0	18.0000
20	Massemani, Mrs. Fatima	Survived	3	female	?	0	0	7.2250
21	Fynney, Mr. Joseph J	Died	2	male	35.00	0	0	26.0000
22	Quinn, Mr. James	Died	3	male	24.00	0	0	10.0000



Unsupervised Classification: Text

Use case : Classification of related articles

- One potential implementation would be based on the **top rare keywords from the content**
- Titles of article don't need to be similar



The Atlantic
"The bottom line: Healthcare.gov on December 1st is night and day from where it was on October 1st."

Welcome to the Marketplace
The Health Insurance Marketplace is now open for enrollment. It's a place where you can compare plans, get help choosing one, and enroll. And if you qualify for help, you can get it.

How the Marketplace works

The Biggest News in the Healthcare.gov Report: Just How Bad It Was
www.theatlantic.com
In order to boast how far the federal site has come, HHS laid out how bad it used to be.

Like · Comment · Share · 141 18 6 · 12 hours ago ·

MORE FROM THE ATLANTIC AND OTHER SITES

In rural Kentucky, health-care debate takes back seat as the long-uninsured line...
The Washington Post
As many who don't have or have never had insurance get coverage, the national debate takes a

Like · Share 50,199 people shared this.

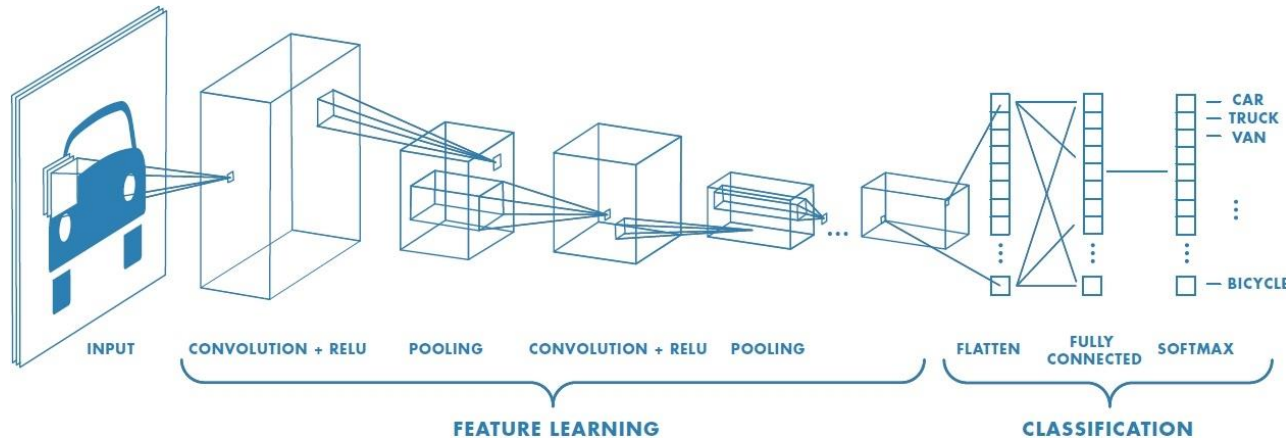
The Crucial Fix the Healthcare.gov Progress Report Ignores
The Atlantic
The Obama Administration touts an improved customer experience. But there's reason to fear that

Like · Share 17,014 people shared this.

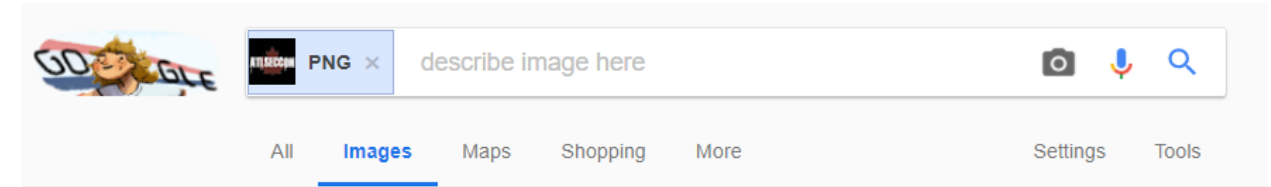
Unsupervised Classification: Images

Use case : Classification of **similar** images (mostly unlabeled)

- Convolutional neural network (CNN) could be applied



Demo of supervised CNN : <http://cs231n.stanford.edu/>



About 5,680,000 results (1.38 seconds)



Image size:
400 x 400

Find other sizes of this image:
[All sizes](#) - [Medium](#)

Pages that include matching images

[Ulrike Bahr-Gedalia - President & CEO - Digital Nova Scotia | LinkedIn](#)



<https://ca.linkedin.com/in/ubahrged>

400 x 400 - **View** Ulrike Bahr-Gedalia's profile on LinkedIn, the world's largest professional community. Ulrike has 6 jobs listed on their profile. ... make plans and discuss industry trends. Towards the end of the cycle, Mentees are encouraged to **pay** their experience forward, by mentoring an undergraduate student for a short period.

[Digital Nova Scotia – AtlSecCon 2018](#)



<https://www.digitalnovascotia.com/events/atlseccon-2018/>

150 x 95 - Established in 2011, our goal is to provide quality information security education and training at an affordable cost. ... With **over** 600 flights a week, you can travel **by air** to Halifax on direct flights from most Canadian cities, along with Boston, Newark, Detroit, Bangor, and overseas via London, Frankfurt, Munich and ...

[Atlantic Security Conference: Full Schedule](#)



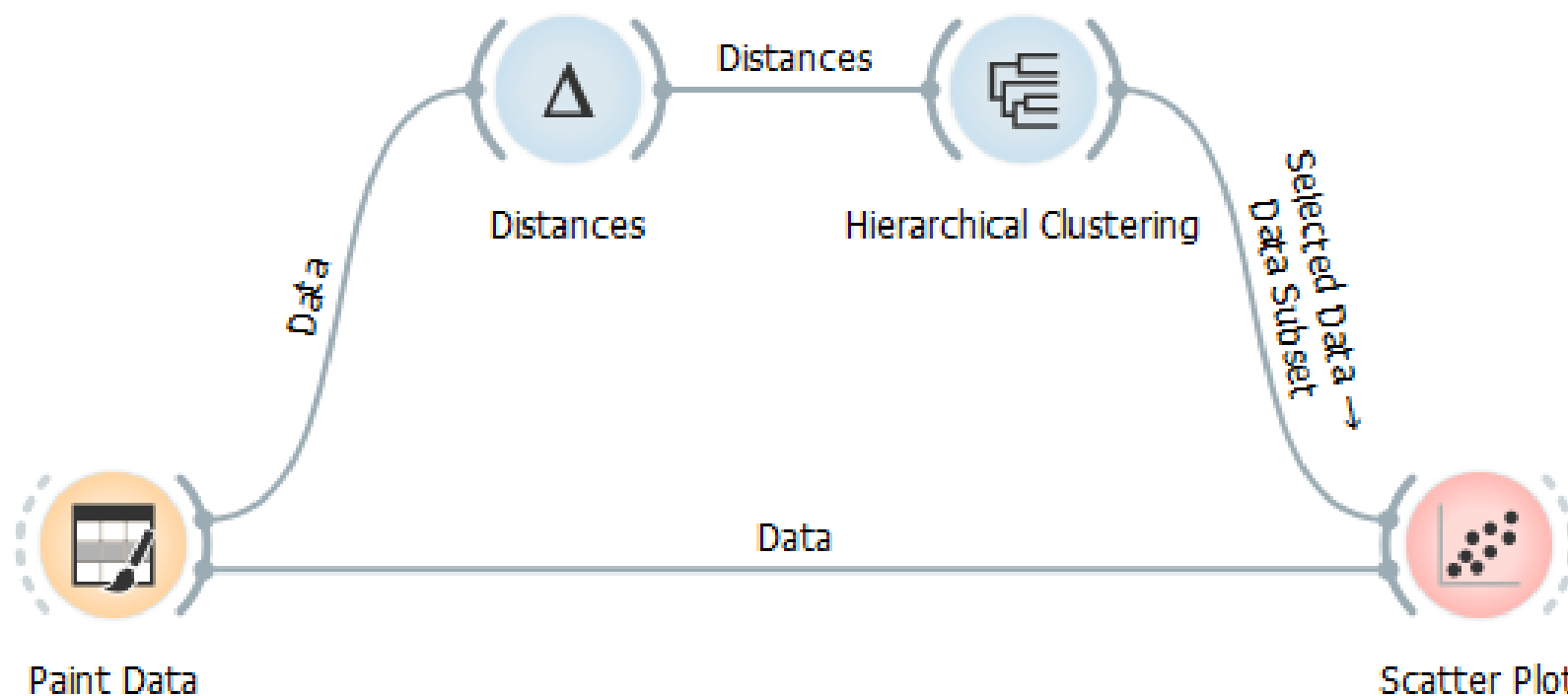
<https://atlanticsecurityconference2018.sched.com/list/descriptions/>

2400 x 500 - Then, we are going to follow the code in the MBR and show how a simple malicious kernel code can take control of the boot process until you **pay** the ransom. I will show a demo on how to debug the MBR to see how the actual native code executes without any API. We are also going to see how we can use a combination of ...



Clustering Warm-up with Orange

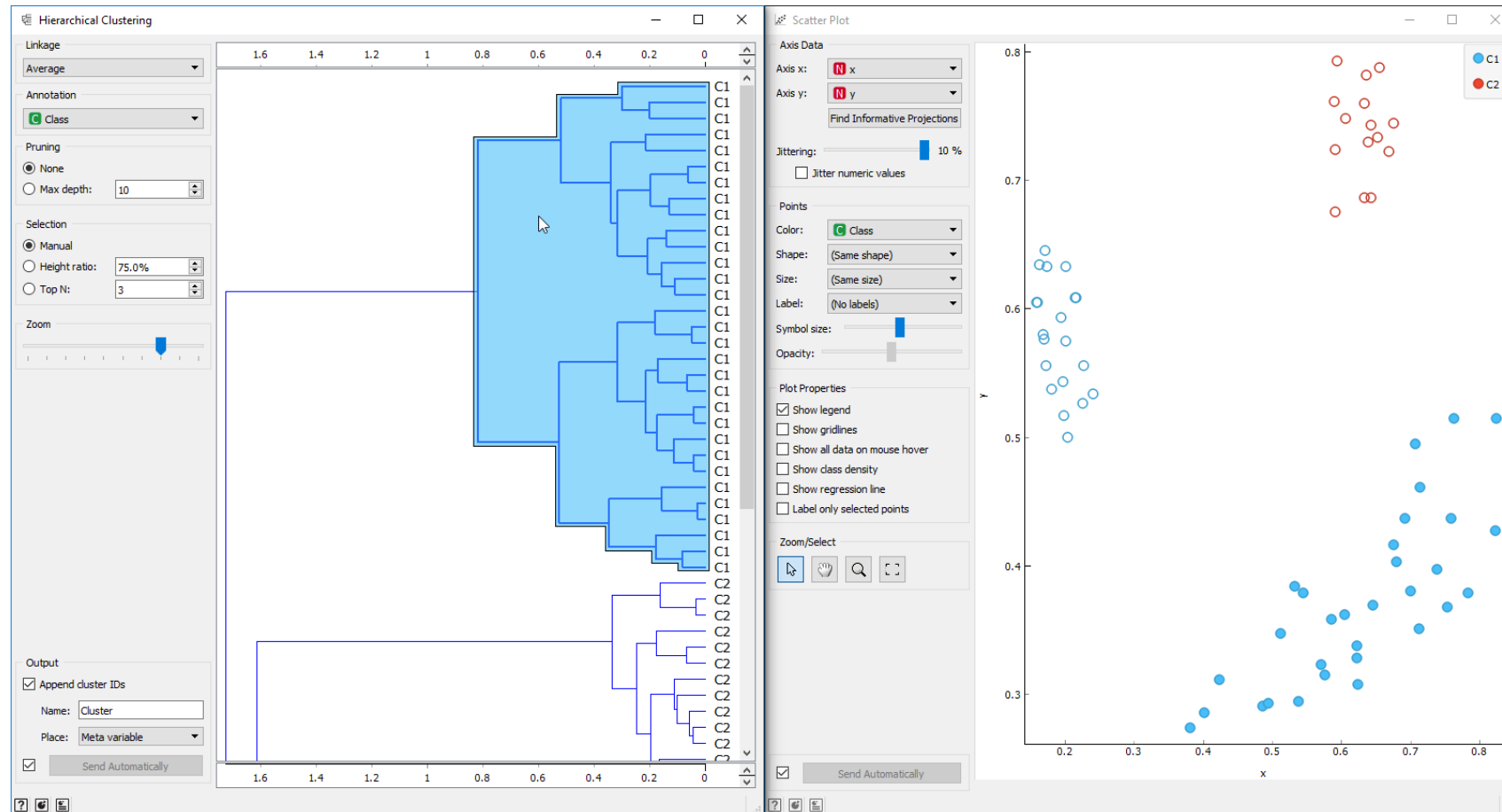
In order to warmup for Orange workflow, we are going to create the following clustering classifier



Hand-on !



Interactive Visualization (Selection)



A complex network diagram with red nodes and lines, some nodes highlighted with larger circles, serving as a background for the slide.

Data Visualization (Titanic Dataset)

The Titanic Dataset

- **survival:** Survival (0 = No; 1 = Yes)
- **pclass:** Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- **name:** Name
- **sex:** Sex
- **age:** Age
- **sibsp:** Number of Siblings/Spouses Aboard
- **parch:** Number of Parents/Children Aboard
- **ticket:** Ticket Number
- **fare:** Passenger Fare
- **cabin:** Cabin
- **embarked:** Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
- <https://jasonicarter.github.io/survival-analysis-titanic-data/>

Visualization

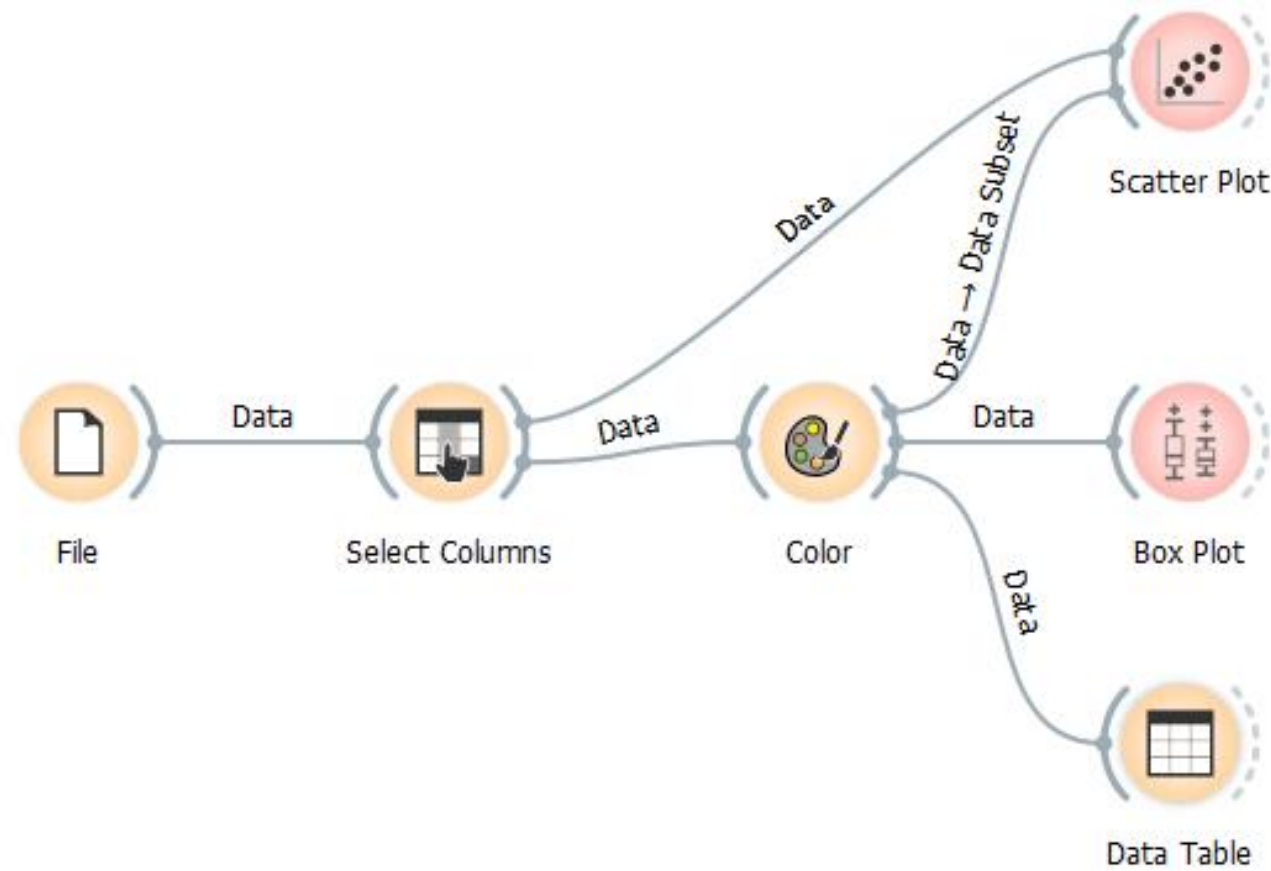


Why visualization is important?

- The choice of algorithm is not always obvious
- Visualization can help greatly the debugging of ML algorithm
 - Quickly identify and review extreme value
 - Review misclassified entries
- Better understand how algorithm behave

Your First Orange workflow

- Before doing any classification, we are going to visualize the dataset

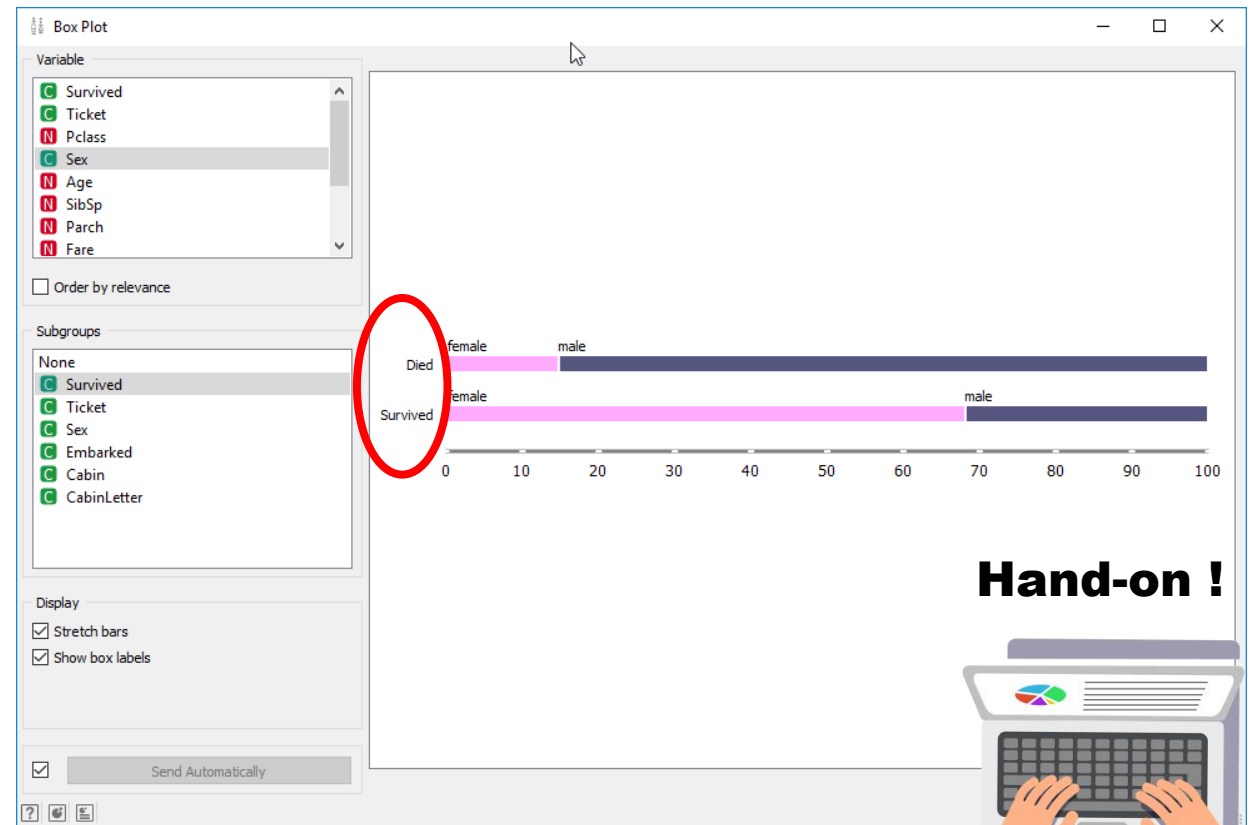


Hand-on !



Color Change and Values Renamed

- The default colors might be counterintuitive (Red for Positive / Green for Negative)
- Some values can be primitive (True/False or 0/1)
- Configure the color component to generate the adjacent view

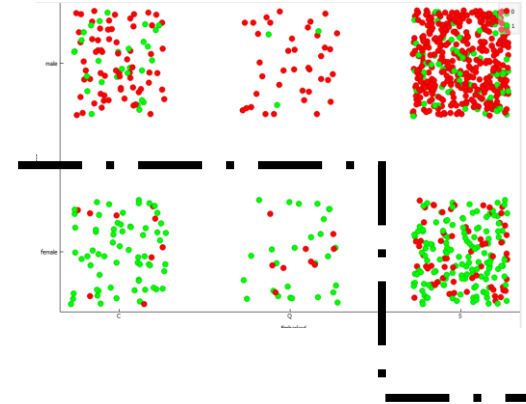


Hand-on !



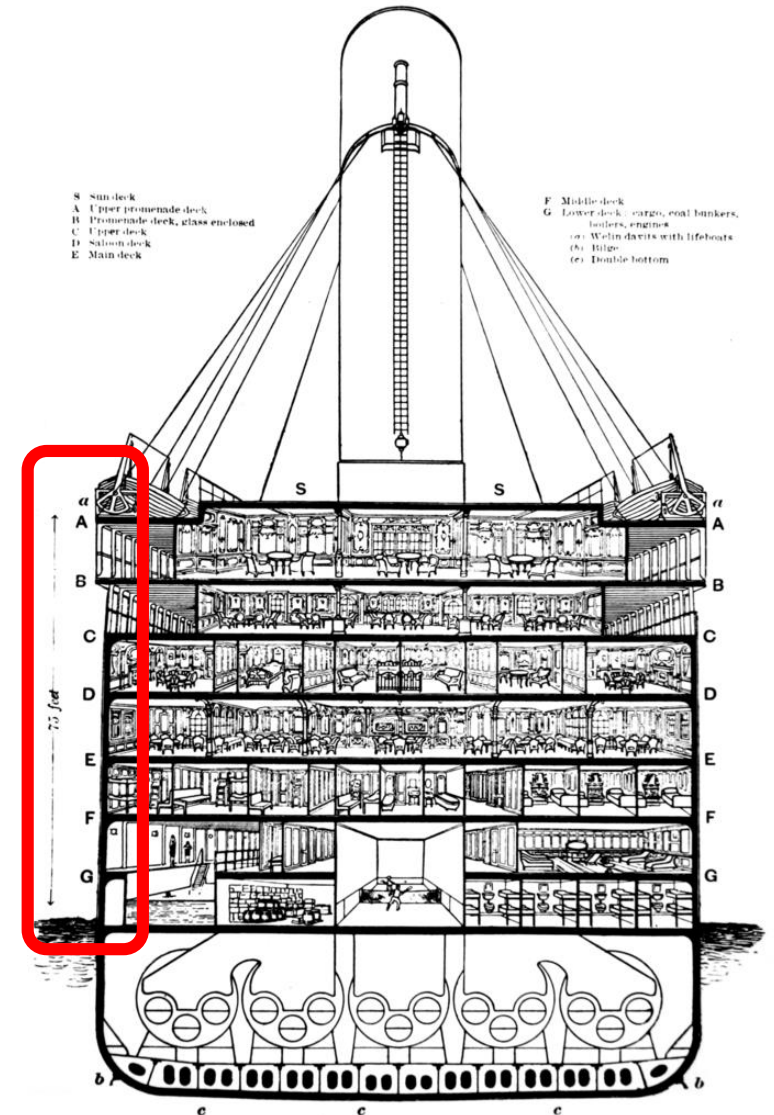
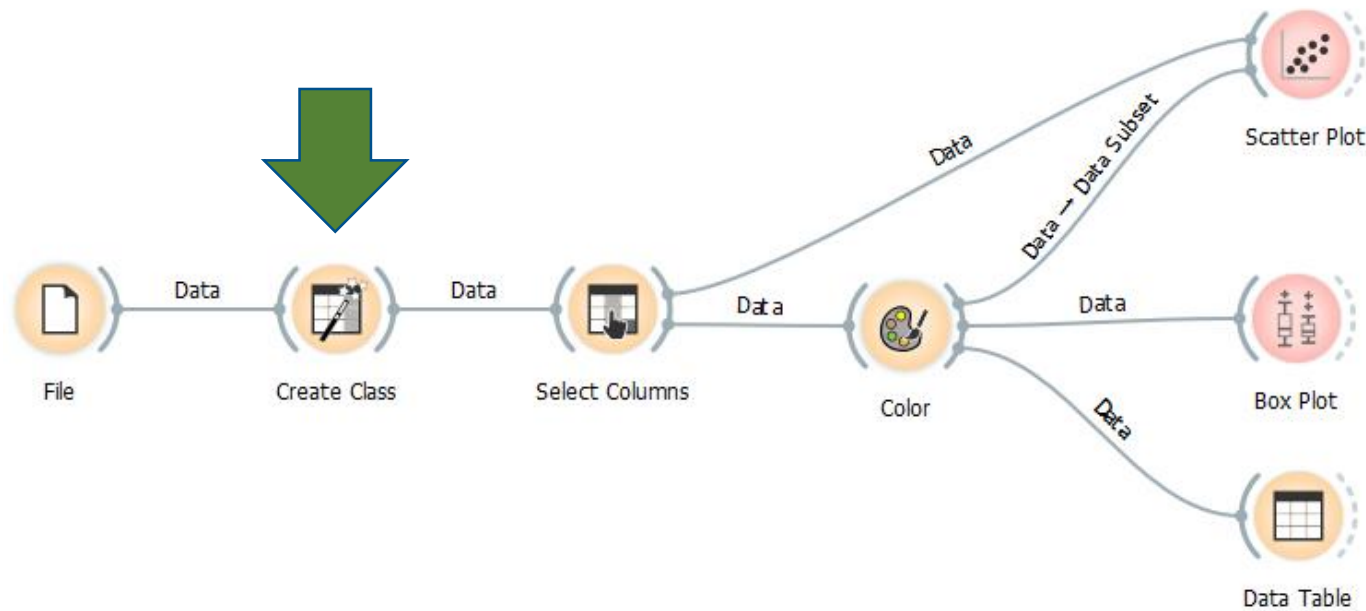
Visualisation Exercise

- In the **Scatter Plot View**, find two attributes that would allow a good projection. You should be able to see a clear separation of class (Survived or not). It doesn't need to be perfect.
- Can you identify an abnormally precise attribute ?
- Which attribute could be improved ?



Cabin Number

- The cabin attribute contains the level (B86, C85, E46 → B,C,E)
 - A transformation is required to allow the future classification to see this information



Creating the “Cabin Letter” Attribute

For simple features creation, two widgets are available

- **Create class (Substring operation)**
- Feature constructor (Numeric transformation)

Create Class

From column: C Cabin

Name	Substring	#Instances
A	A	15
B	B	47
C	C	59
D	D	33
E	E	32
F	F	13
G	G	4
H	H	0

+

Name for the new class: CabinLetter

☒ Match only at the beginning

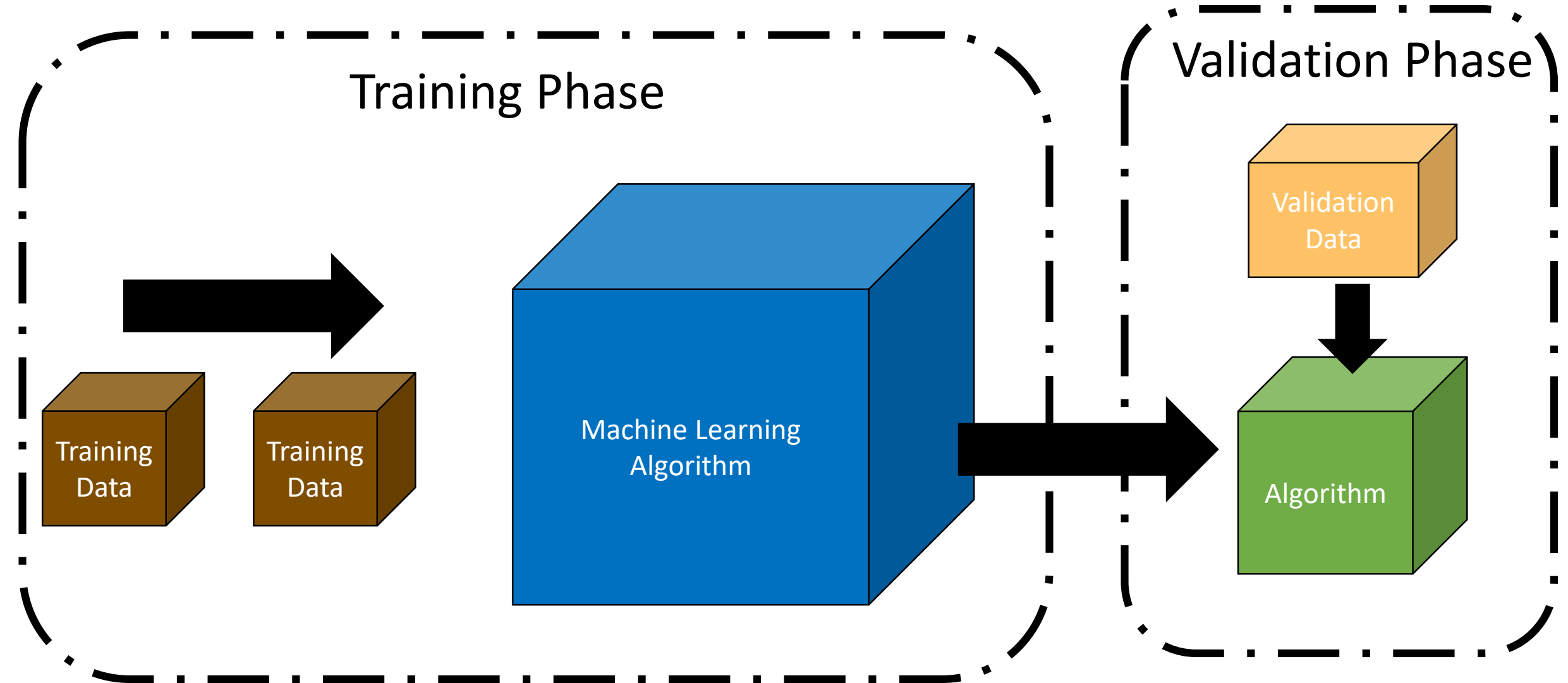
☐ Case sensitive

Hand-on !

A complex network diagram with red nodes and lines, some nodes highlighted with larger circles, serving as a background for the slide.

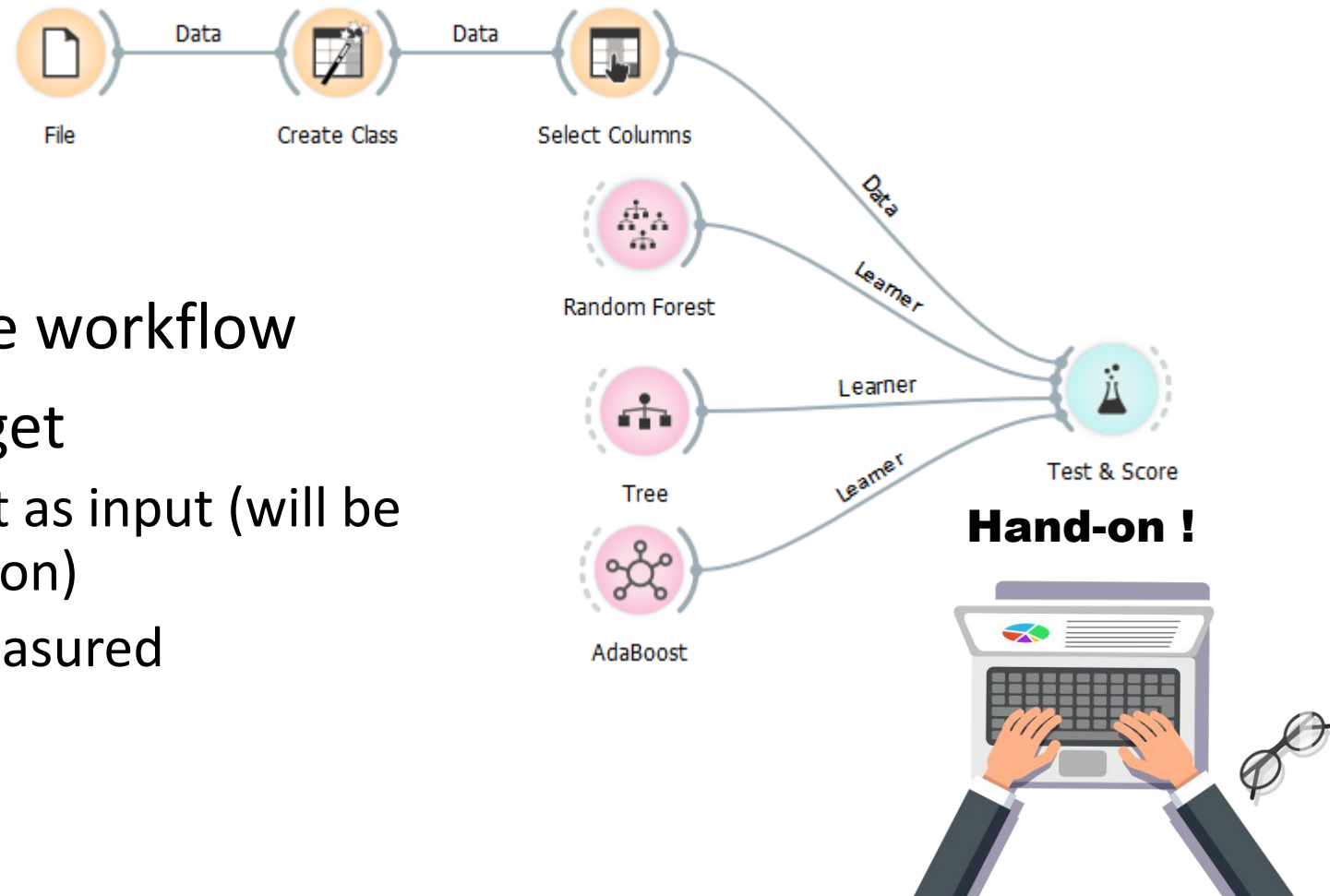
Making Prediction (Titanic Dataset)

Identifying Efficient Algorithm



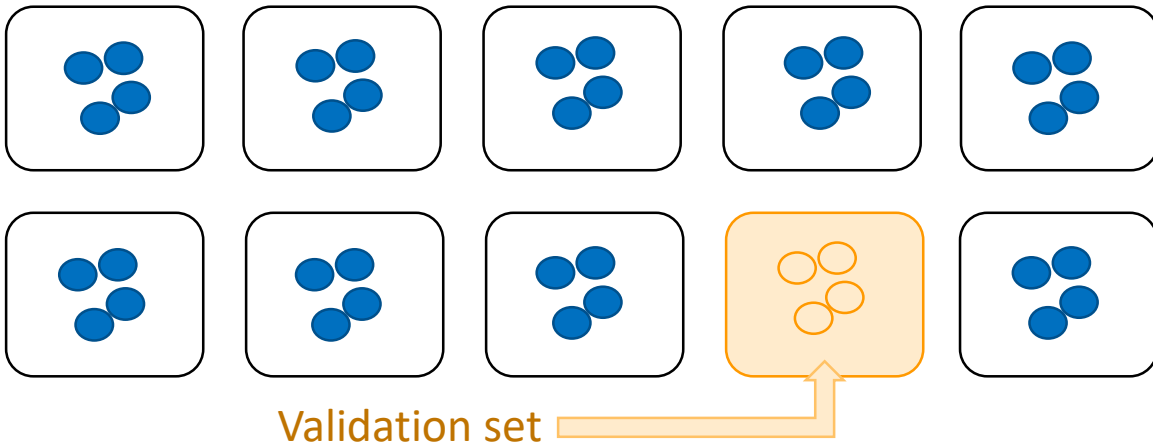
Testing Algorithms

- Drag three algorithms to the workflow
- Use the “Test & Score” widget
 - It takes the complete dataset as input (will be used for training and validation)
 - All algorithms link will be measured



10-fold Cross validation

Technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.



Test & Score

Sampling

- ☒ Cross validation
 - Number of folds: 10
 - ☒ Stratified
- ☐ Cross validation by feature
 - Ticket
- ☐ Random sampling
 - Repeat train/test: 10
 - Training set size: 66 %
 - ☒ Stratified
- ☐ Leave one out
- ☐ Test on train data
- ☐ Test on test data

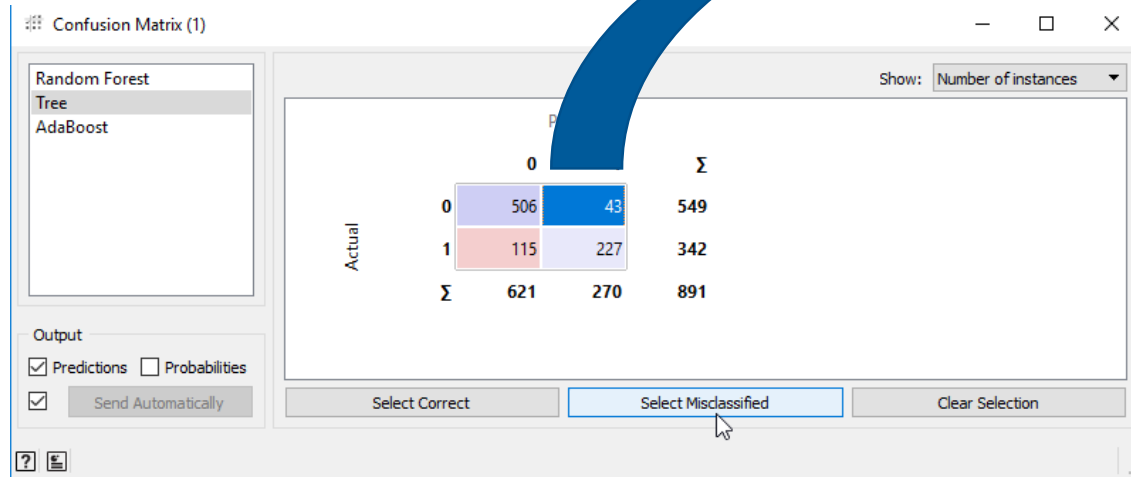
Target Class

(Average over classes)

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Tree	0.827	0.823	0.742	0.841	0.664
Random Forest	0.865	0.816	0.743	0.801	0.693
AdaBoost	0.839	0.799	0.734	0.746	0.722

Analyzing Misclassified Elements



Data Table (1)

Info: 43 instances, 9 features (21.7% missing values), Discrete class with 2 values (no missing values), 4 meta attributes (no missing values)

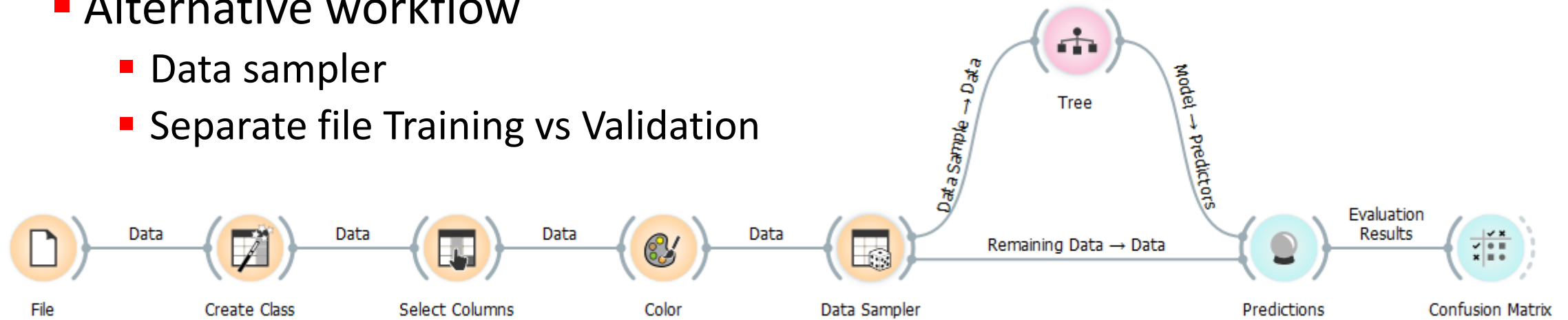
Variables: ☒ Show variable labels (if present), ☒ Visualize numeric values, ☒ Color by instance classes

Selection: ☒ Select full rows

	Survived	PassengerId	Name	Ticket	Survived(Tree)
1	0	236	Harknett, Miss. ...	W./C. 6609	1
2	0	265	Henry, Miss. De...	382649	1
3	0	313	Lahtinen, Mrs. ...	250651	1
4	0	356	Vanden Steen, ...	345783	1
5	0	637	Leinonen, Mr. ...	STON/O 2. 3101...	1
6	0	681	Peters, Miss. Ka...	330935	1
7	0	773	Mack, Mrs. (Ma...	S.O./P.P. 3	1
8	0	19	Vander Planke, ...	345763	1
9	0	25	Palsson, Miss. T...	349909	1
10	0	42	Turpin, Mrs. Wi...	11668	1
11	0	141	Boulos, Mrs. Jo...	2678	1
12	0	375	Palsson, Miss. S...	349909	1
13	0	883	Dahlberg, Miss...	7552	1
14	0	499	Allison, Mrs. H...	113781	1
15	0	504	Laitinen, Miss. ...	4135	1
16	0	618	Lobb, Mrs. Willi...	A/5. 3336	1
17	0	855	Carter, Mrs. Ern...	244252	1
18	0	15	Vestrom, Miss. ...	350406	1
19	0	112	Zabour, Miss. H...	2665	1
20	0	178	Isham, Miss. An...	PC 17595	1
21	0	241	Zabour, Miss. T...	2665	1
22	0	655	Hegarty, Miss. ...	365226	1
23	0	853	Boulos, Miss. N...	2678	1
24	0	358	Funk, Miss. An...	237671	1
25	0	714	Larsson, Mr. Au...	7545	1
26	0	808	Pettersson, Mis...	347087	1
27	0	220	Harris, Mr. Walter	W/C 14208	1
28	0	383	Tikkanen, Mr. J...	STON/O 2. 3101...	1

Data Sampler + Predictions Widget (alternative)

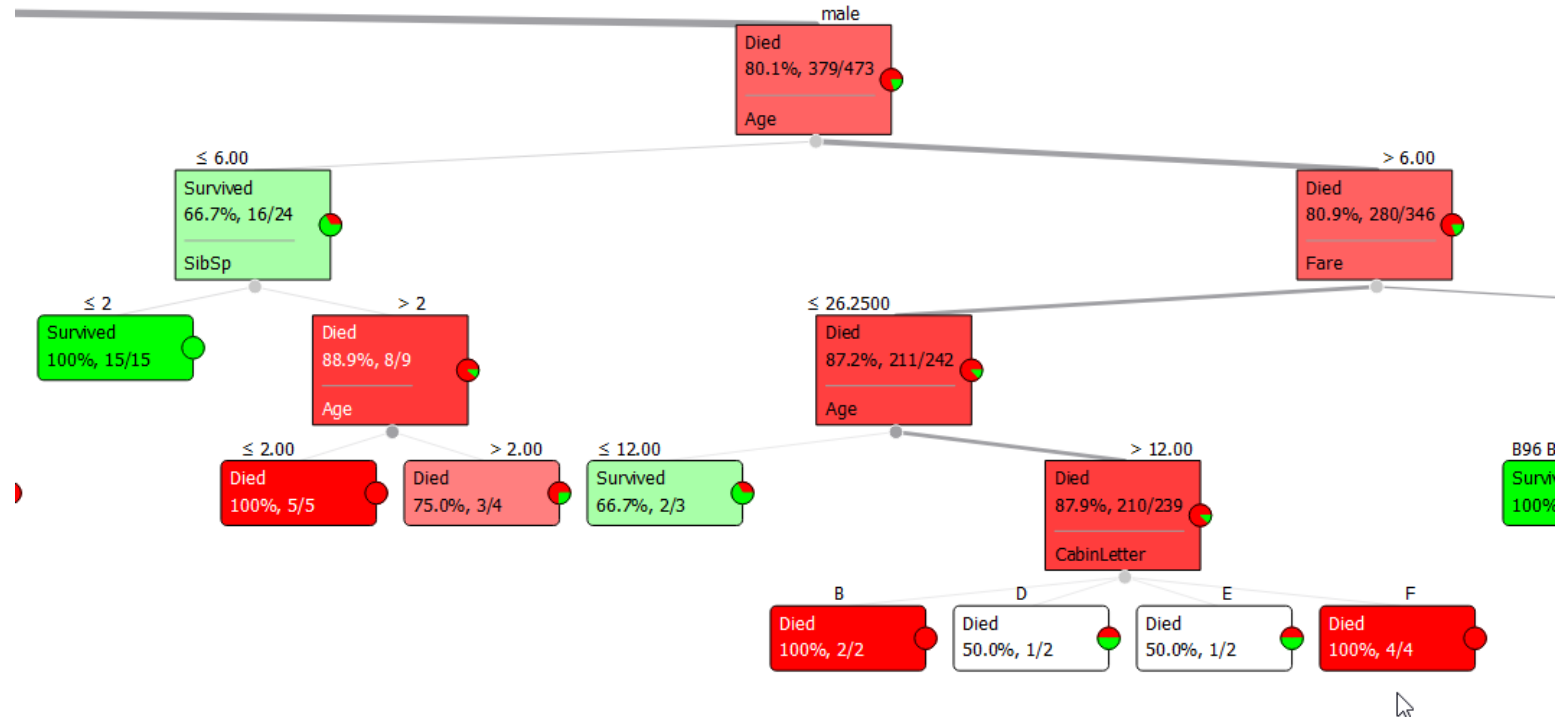
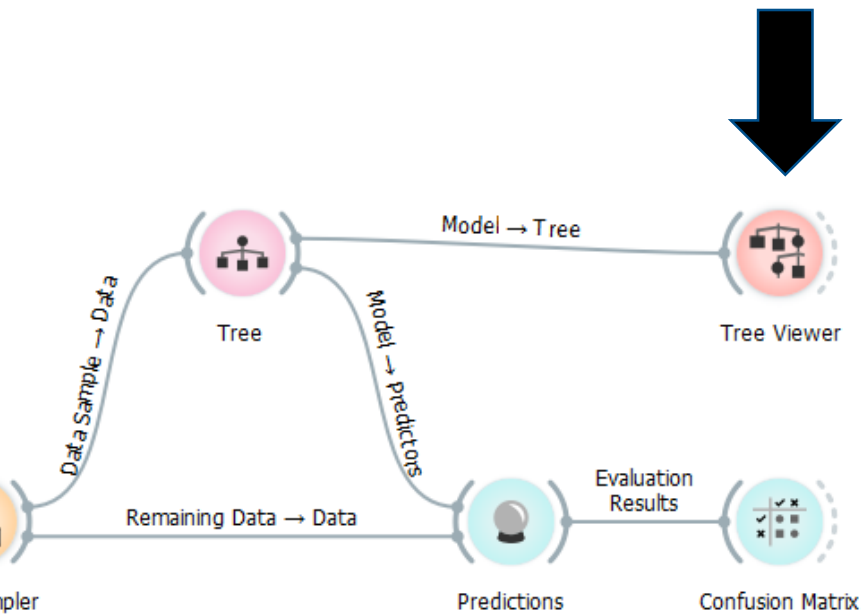
- Alternative workflow
 - Data sampler
 - Separate file Training vs Validation



Predictions													
Info													
Data: 90 instances. Predictors: 1 Task: Classification													
Restore Original Order													
Show													
<input checked="" type="checkbox"/> Predicted class													
<input checked="" type="checkbox"/> Predicted probabilities for:													
<div><div>Died</div><div>Survived</div></div>													
<input checked="" type="checkbox"/> Draw distribution bars													
	Random Forest	Survived	PassengerId	Name	Ticket	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Ci
1	0.96 : 0.04 → Died	Died	1	Braund, Mr. Ow...	A/5 21171	3	male	22.00	1	0	7.2500	S	?
2	0.10 : 0.90 → Survived	Survived	2	Cumings, Mrs. ...	PC 17599	1	female	38.00	1	0	71.2833	C	C85
3	0.19 : 0.81 → Survived	Survived	3	Heikkinen, Miss...	STON/O2. 3101...	3	female	26.00	0	0	7.9250	S	?
4	0.11 : 0.89 → Survived	Survived	4	Futrelle, Mrs. Ja...	113803	1	female	35.00	1	0	53.1000	S	C123
5	0.76 : 0.24 → Died	Died	5	Allen, Mr. Willia...	373450	3	male	35.00	0	0	8.0500	S	?
6	0.91 : 0.09 → Died	Died	6	Moran, Mr. Jam...	330877	3	male	?	0	0	8.4583	Q	?
7	0.79 : 0.21 → Died	Died	7	McCarthy, Mr. ...	17463	1	male	54.00	0	0	51.8625	S	E46
8	1.00 : 0.00 → Died	Died	8	Palsson, Master...	349909	3	male	2.00	3	1	21.0750	S	?
9	0.14 : 0.86 → Survived	Survived	9	Johnson, Mrs. ...	347742	3	female	27.00	0	2	11.1333	S	?
10	0.09 : 0.91 → Survived	Survived	10	Nasser, Mrs. Ni...	237736	2	female	14.00	1	0	30.0708	C	?
11	0.27 : 0.73 → Survived	Survived	11	Sandstrom, Mis...	PP 9549	3	female	4.00	1	1	16.7000	S	G6
12	0.01 : 0.99 → Survived	Survived	12	Bonnell, Miss. E...	113783	1	female	58.00	0	0	26.5500	S	C103

Visualizing the Model Produced

- Select the widget Tree Viewer





Testing Classification Algorithm (Static-Analysis Dataset)

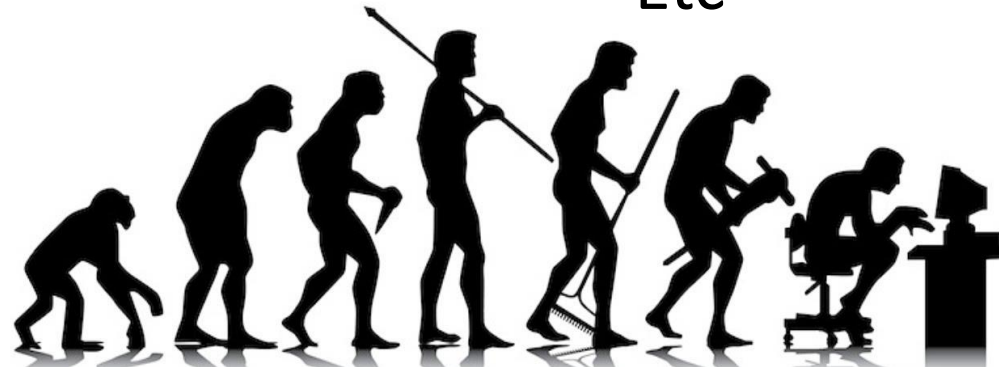
Applying ML to Static Analysis

Typical Static Code Analysis (SCA) report includes

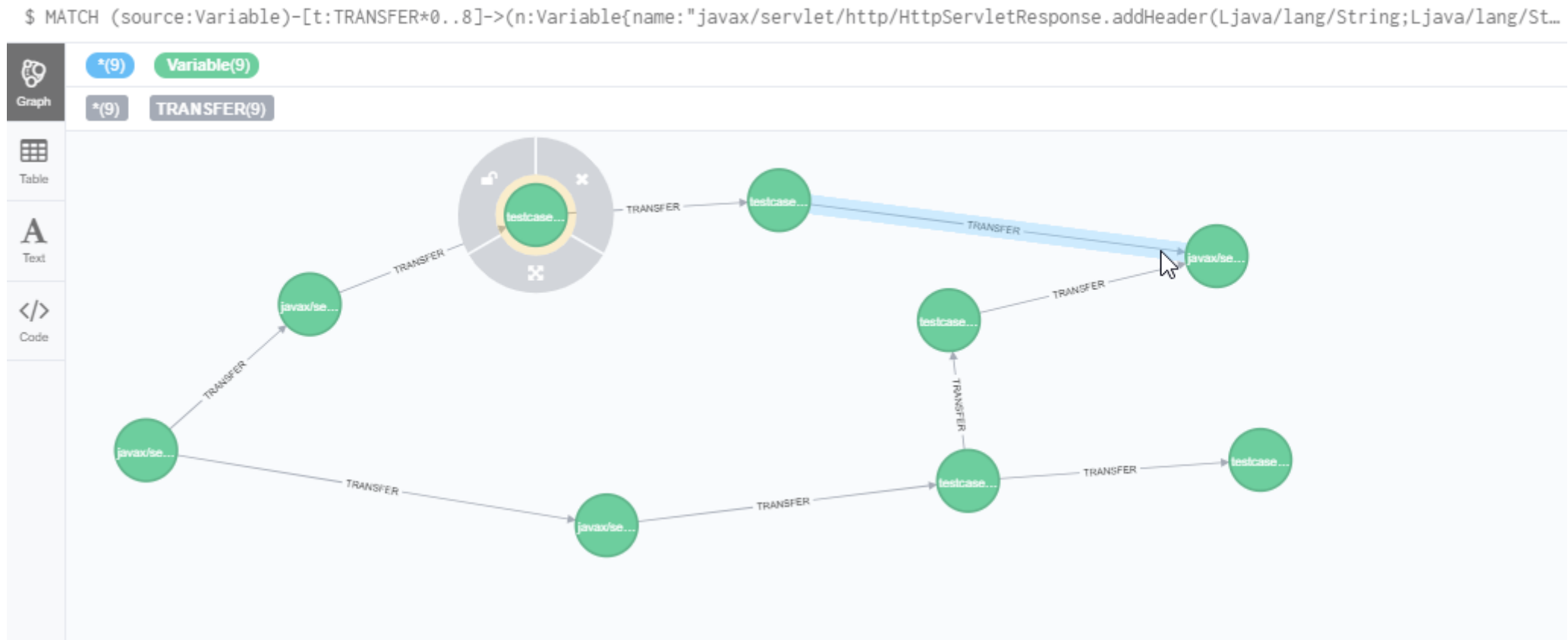
- Bug type
- Source file (or class name)
- Line number
- Description for remediation

“Enrich” Information added to the model

- Presence of Tainted variable
- Presence of Safe source
- Methods called
 - Sources and Sinks
- Etc



Where Do Those New Attributes Comes From?



- In order to do taint analysis accross the entire web application, a graph was built (The CSV is the result of this graph query on such graph).

Preview with a Subset of the Data

Data Table

Info
13 instances
12 features (26.9% missing values)
No target variable.
2 meta attributes (no missing values)

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

	Key	GroupId	ArtifactId	Author	BugType	CWE	MethodSink	UnknownSource	SourceMethod	hasTaintedSource	HasSafeSource	asUnknownSource	Status
1	?	gov.nist.samate	juliet-test-suite...	?	SQL_INJECTIO...	89	java/sql/Statem...	java/lang/System.gete...	testcases/CWE89_SQ...	true	true	false	?
2	?	gov.nist.samate	juliet-test-suite...	?	SQL_INJECTIO...	89	java/sql/Statem...	?	testcases/CWE89_SQ...	true	true	true	?
3	?	gov.nist.samate	juliet-test-suite...	?	LDAP_INJECTION	90	javax/naming/...	javax/servlet/http/Coo...	testcases/CWE90_LD...	true	true	false	?
4	?	gov.nist.samate	juliet-test-suite...	?	LDAP_INJECTION	90	javax/naming/...	javax/servlet/http/Http...	testcases/CWE90_LD...	true	true	false	?
5	?	gov.nist.samate	juliet-test-suite...	?	XSS_SERVLET	79	java/io/PrintWr...	java/lang/Throwable.g...	testcasesupport/Abst...	false	true	true	?
6	?	gov.nist.samate	juliet-test-suite...	?	XSS_SERVLET	79	java/io/PrintWr...	java/lang/Throwable.g...	testcasesupport/Abst...	false	true	true	?
7	?	gov.nist.samate	juliet-test-suite...	?	XSS_SERVLET	79	java/io/PrintWr...	java/lang/Throwable.g...	testcasesupport/Abst...	false	true	true	?
8	?	gov.nist.samate	juliet-test-suite...	?	XSS_SERVLET	79	java/io/PrintWr...	java/lang/Throwable.g...	testcasesupport/Abst...	false	true	true	?
9	?	gov.nist.samate	juliet-test-suite...	?	XSS_SERVLET	79	java/io/PrintWr...	java/lang/Throwable.g...	testcasesupport/Abst...	false	true	true	?
10	?	gov.nist.samate	juliet-test-suite...	?	XSS_SERVLET	79	java/io/PrintWr...	java/lang/Throwable.g...	testcasesupport/Abst...	false	true	true	?
11	?	gov.nist.samate	juliet-test-suite...	?	DMI_EMPTY_D...	259	?	?	testcasesupport/IO.g...	?	?	?	?
12	?	gov.nist.samate	juliet-test-suite...	?	HARD_CODE_P...	259	?	?	testcasesupport/IO.g...	?	?	?	?
13	?	gov.nist.samate	juliet-test-suite...	?	PREDICTABLE_...	330	?	?	testcasesupport/IO.st...	?	?	?	?

Deceptive Criminal Order

Juliet Static Analysis Dataset

Developed by the NIST to test static code analysis tools

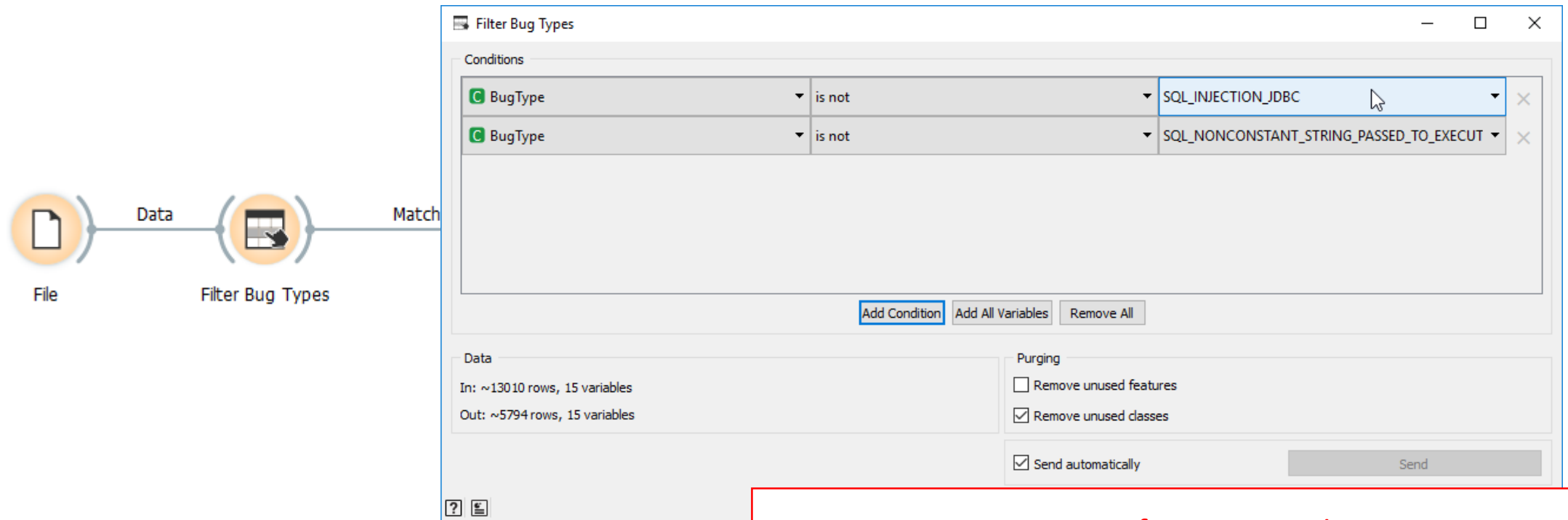
- 28881 individual test cases
- 118 CWE categories
- Language: Java



<https://samate.nist.gov/SRD/testsuite.php>

Filtering data (possibly noise)

- Additional filter needed “Select Rows Widget”



The image shows a workflow diagram on the left and a detailed view of the 'Filter Bug Types' widget on the right.

Workflow Diagram:

- A 'File' icon (document) is connected to a 'Data' label.
- The 'Data' label is connected to a 'Filter Bug Types' widget icon (table with a hand cursor).
- The 'Filter Bug Types' widget is connected to a 'Match' label.

Filter Bug Types Widget Interface:

The widget window has a title bar 'Filter Bug Types' and standard window controls. It contains the following sections:

- Conditions:** A table with two rows of conditions.

Variable	Operator	Value	Action
BugType	is not	SQL_INJECTION_JDBC	X
BugType	is not	SQL_NONCONSTANT_STRING_PASSED_TO_EXECUT	X
- Buttons:** 'Add Condition' (highlighted with a blue border), 'Add All Variables', and 'Remove All'.
- Data:** A section showing 'In: ~13010 rows, 15 variables' and 'Out: ~5794 rows, 15 variables'.
- Purging:** A section with checkboxes for 'Remove unused features' (unchecked), 'Remove unused classes' (checked), and 'Send automatically' (checked). A 'Send' button is at the bottom right.

For your information only:
These categories have been removed from the dataset.

Objectives for the SCA datasets

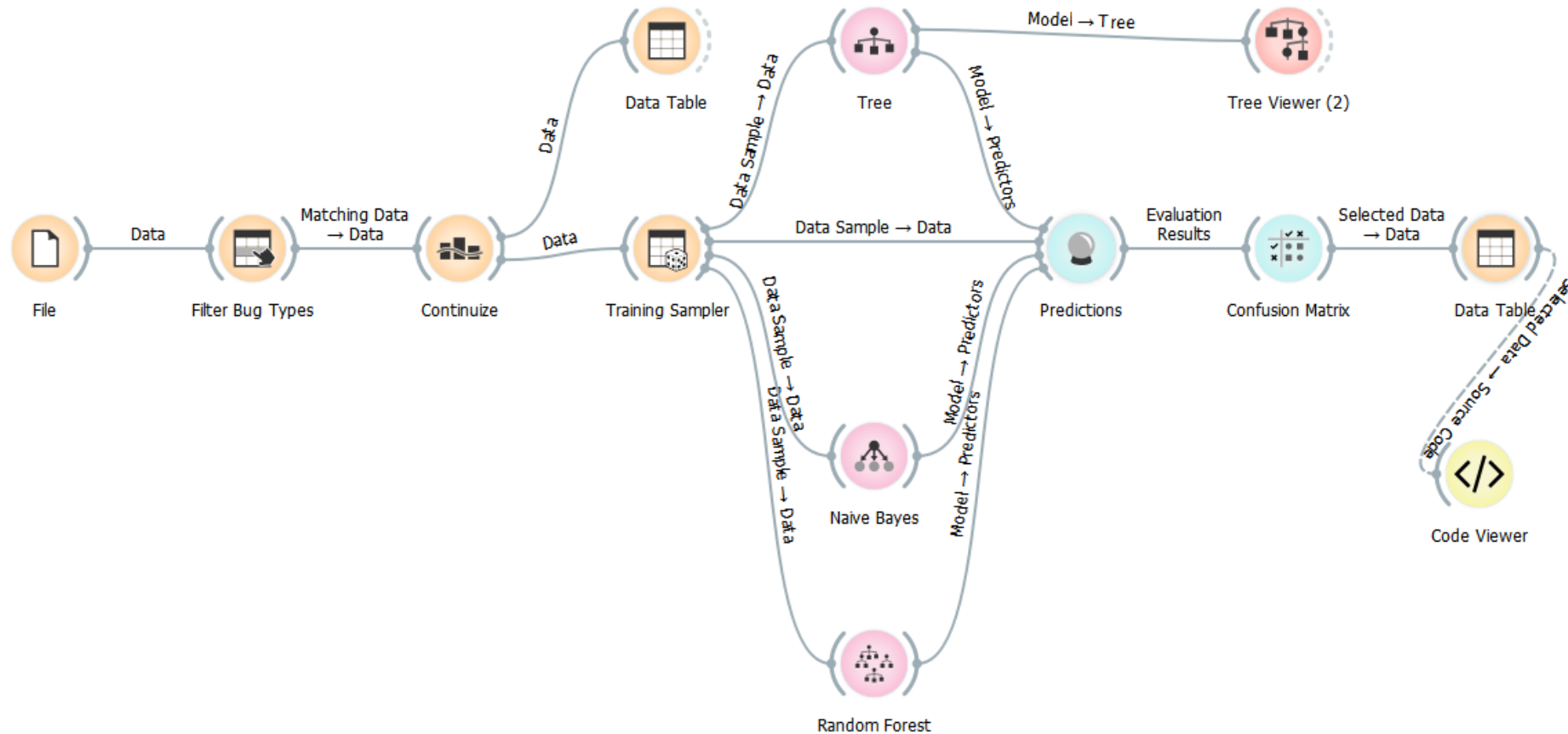
- Find the types of bugs that have low or high false positives
- Which attributes should be transfer to metadata?
 - Attributes that should be ignored
- Reproduce the same prediction workflow to this new dataset
- Which algorithm perform the best?



Hand-on !



Expected Orange Canvas



Code Viewer Widget :

<https://github.com/GoSecure/orange-code-widget>

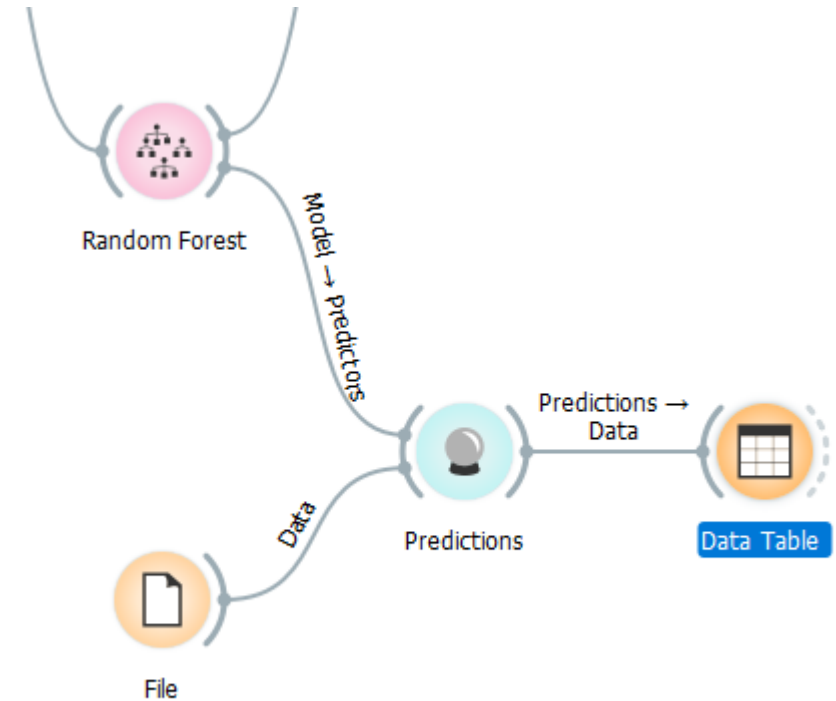
Hand-on !



GoSecure

Can you find the vulnerability?

- Apply the "Random forest" algorithm to the dataset "03_hotel-spotbugs-results.csv"
- Use the Code Widget to explore the potential vulnerability
 - <https://github.com/GoSecure/orange-code-widget>





Conclusion

Potential Extensions to this Workshop

- Repeat the exercise with the **Orange API**

```
>>> import Orange
>>> data = Orange.data.Table("juliet-result.csv")
```

- Develop **custom widgets** for Orange (Python based add-on)
- Start experimenting with your **own dataset**

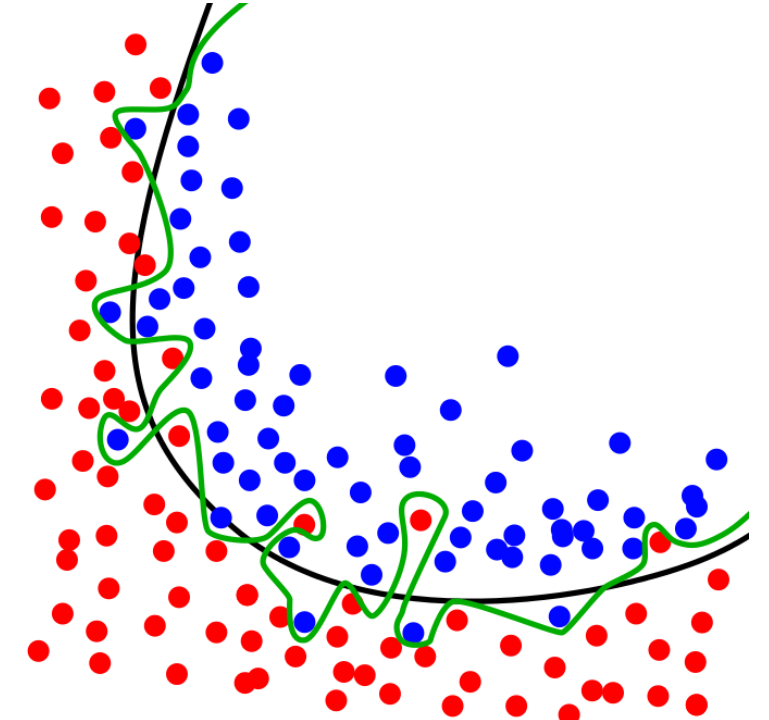
Training ML algorithm




Short documentary on Machine Learning training risks

Be careful ...

- Inadequate training set (noisy, incomplete)
- Overfitting
 - Excellent results with training data
 - ..but inaccurate with application data
- Underfitting
 - Lack of attributes
- Keep skepticism regarding the validity of your data and the performance associated to your algorithm's results



Overfitting example using two dimensions (numeric attributes)

- 
- This workshop touches only a subset of Machine Learning (Supervised-learning)
 - Supervised-learning requires prior classification of a subset of values
 - Machine learning will **not work on any data**
 - The selection of attributes is crucial
 - Selection of training data and validation data is very important

References

Machine Learning Coursera

- <https://www.coursera.org/learn/machine-learning>

Introduction to Machine Learning

- <https://developers.google.com/machine-learning/crash-course/ml-intro>

- Getting Started with Orange : Tutorial Series

<https://www.youtube.com/watch?v=HXjnDIgGDuI&list=PLmNPvQr9Tf-ZSDLwOzxpY-HrE0yv-8Fy>

Additional Datasets

Kaggle

- <https://www.kaggle.com/datasets>

Google BigQuery public datasets

- <https://bigquery.cloud.google.com/publicdatasets/>

OpenML

- <https://www.openml.org/search?type=data>

Data from your systems..

- Log files (WAF, IDS), malware samples (exe, ps, apk), network capture (pcap), etc.