

Jubatus

第3回 データ変換とフィルタ

おさらい

- 生のデータのまま機械学習することはできない
 - テキストなどは直接扱うことができない
 - 特徴抽出を行い特徴ベクトルに変換
 - Jubatus
- 前回：一般的なデータ変換
- 今回：

概要

- データ変換
 - フィルター：本日紹介するところ
 - 特徴抽出

生データ



加工データ



特徴ベクトル

フィルター

特徴抽出

```
1 <!DOCTYPE HTML PUBLIC "-//  
2 "http://www.w3.org/TR/ht  
3 <html>  
4 <head>  
5 <title>Example</title>  
6 <link rel="stylesheet" href="s  
7 </head>  
8 <body>  
9 <div id="header">  
10 <h1><a href="." title="Back  
11 </div>  
12 <div id="toolbar">  
13 <span class="left">Today <sp  
14 <span class="right">  
15 <span id="time">&nbsp;</sp  
16 <select id="timezone">  
17 <option value="-12">(GMT  
18 <option value="-11">(GMT
```

局は今現在に収まっており、なん
来しく生きているつもりだ。
五月最後の更新にする予定なので、
分語りでもしようかと思う。
とは言っても、そこら辺はタラタラと書いて
いって、こんな感じで自分の行く先を歩いていた
毎度のことながら結構シビアなこと書いてる
に
人ですっていきま

used terms			
od	apple		web
0.00867	0.04016		0.01075
0.00454	0.00480		0.01003
0.00454	0.04016		0.00970
0.00495	0.00892		0.01890
0.00117	0.00469		0.01075
0.00419	0.00201		0.01003
0.00450	0.00468		0.00947
0.00398	0.04016		0.01816
0.00752	0.00010		0.00182
0.00398	0.00464		0.01890

フィルタ

- フィルタ処理によってデータを整形する
 - HTML
 - 不要な記号列などの除去
- JSON
 - フィルタリングでどのような要素を取り除くか
 - どのように特徴の抽出を行い，重み付けをするか

JSON

- JSON(JavaScript Object Notation)

- 軽量のデータ交換フォーマット
- 人間もマシンも扱いやすい

- JavaScript

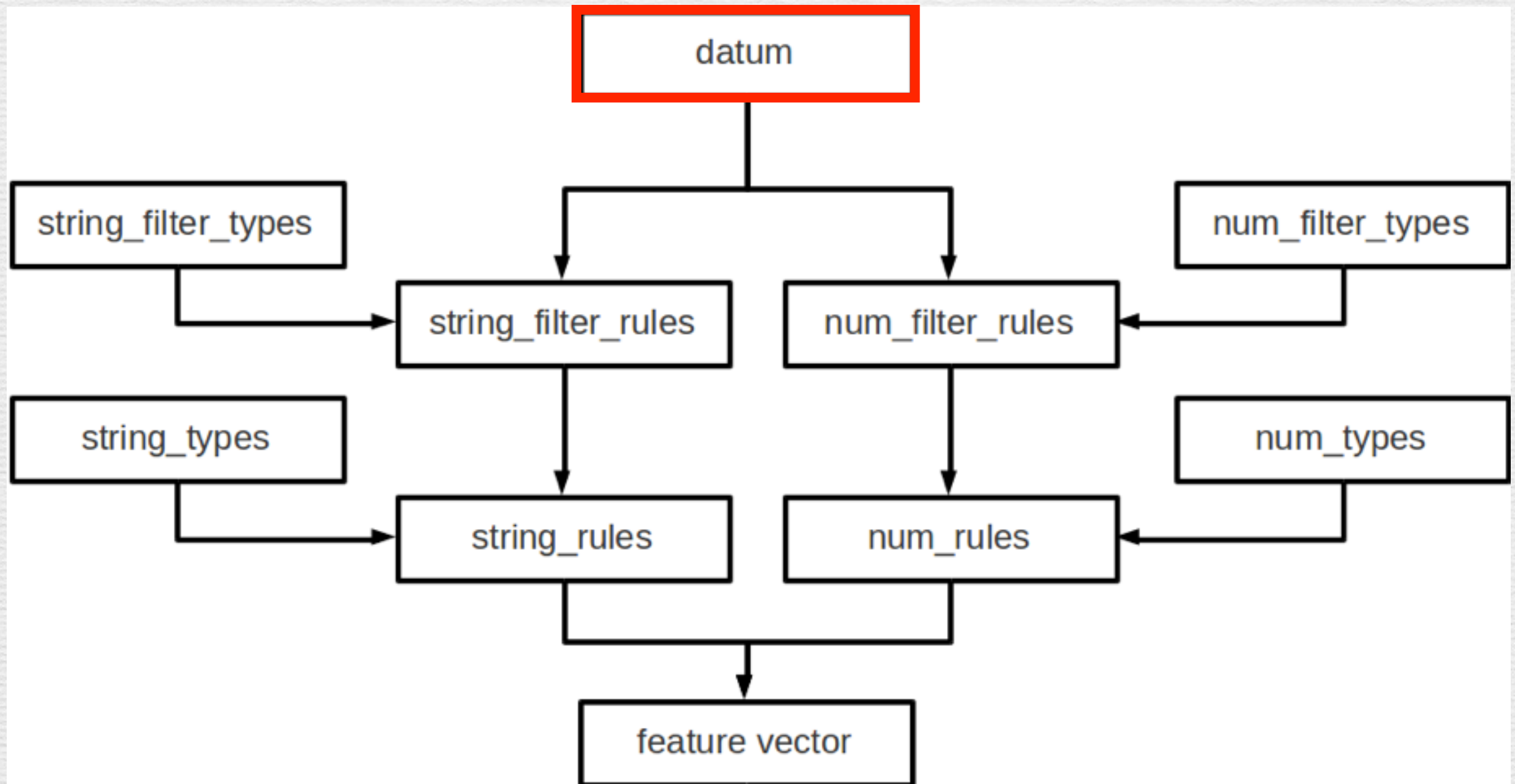
- XML

```
1  {
2  "dict": {
3    "key": ["activities", "katamarize gallery"],
4    "array": {
5      "works": {
6        "client": "KATAMARI INC.",
7        "name": "JSON Editor",
8        "path": "jsoneditor",
9        "releaseDate": "2008/05",
10       "type": "gallery",
11       "url": "http://jsoneditor.net",
12       "credit": [{ "company": "KATAMARI INC.",
13                   "url": "http://www.katamari.co.jp",
14                   "name": "GA4",
15                   "role": "DESIGNER"}, { "company": "KATAMARI INC.",
16                   "name": "MARK",
17                   "role": "JAVASCRIPT PROGRAMMER"}, { "company": "KATAMARI INC.",
18                   "name": "MASARU",
19                   "role": "PROJECT LEADER / JAVASCRIPT PROGRAMMER"}],
20       "color": ["FF3262", "0079CE", "FFFF00"]
21     }
22   }
23 }
24 }
```

Jubatus

```
{
  "string_filter_types": {},
  "string_filter_rules": [],
  "num_filter_types": {},
  "num_filter_rules": [],
  "string_types": {},
  "string_rules": [
    { "key": "*", "type": "space", "sample_weight": "bin", "global_weight": "bin" }
  ],
  "num_types": {},
  "num_rules": [
    { "key": "*", "type": "num" }
  ],
  "binary_types": {},
  "binary_rules": []
}
```

データ変換の流れ

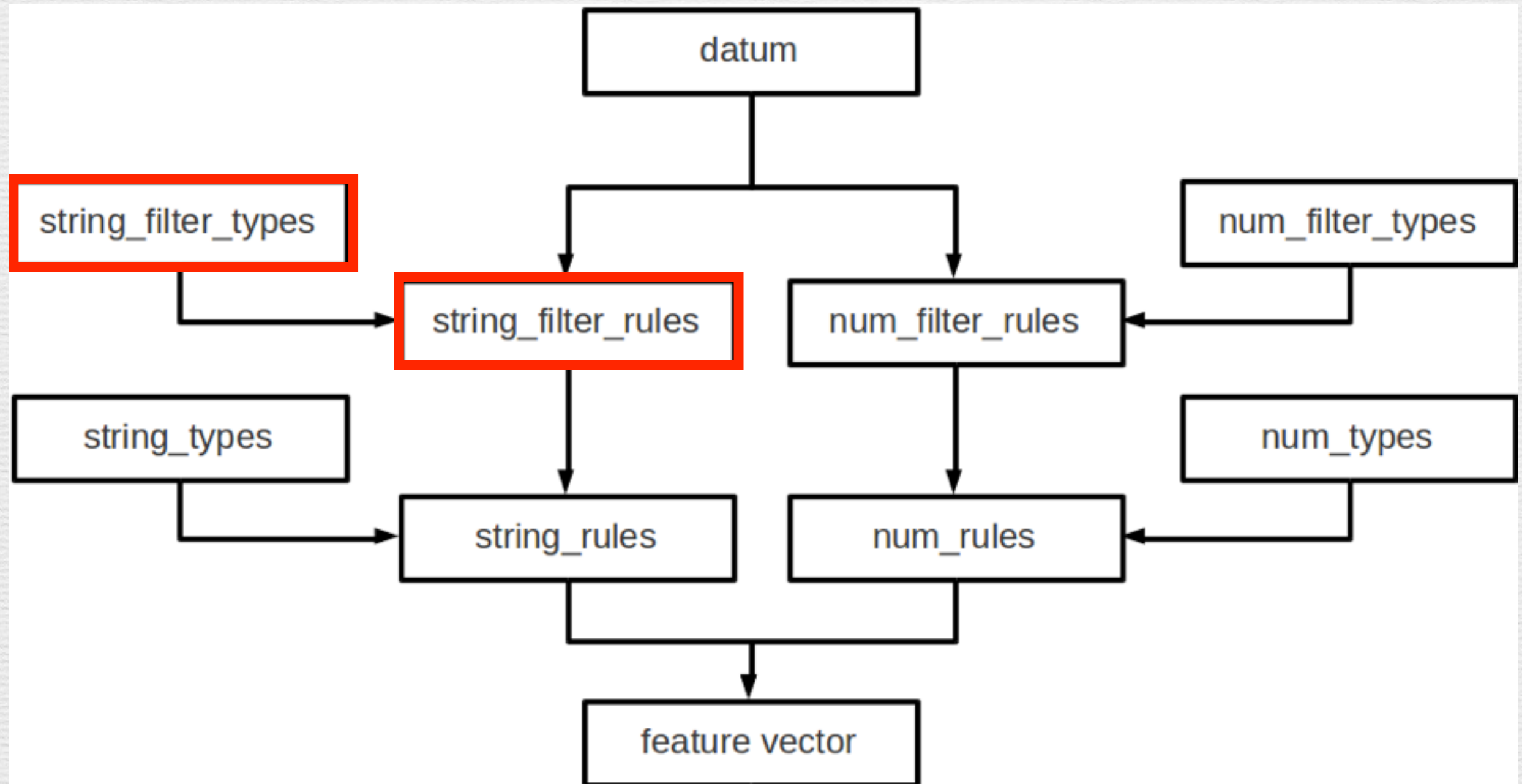


datum

- Jubatus
 - シンプルな
 - string-values
 - num_values
 - binary_values

```
(  
  [  
    ("user/id", "ippy"),  
    ("user/name", "Loren Ipsum"),  
    ("message", "<H>Hello World</H>")  
  ],  
  [  
    ("user/age", 29.0),  
    ("user/income", 100000.0)  
  ],  
  [  
    ("user/image", "xxxxxxxxx")  
  ]  
)
```

データ変換の流れ



string_filter

- HTML

```
"string_filter_types": {  
  "detag": { "method": "regex", "pattern": "<[^>]*>", "replace": "" }  
},  
"string_filter_rules": [  
  { "key": "message", "type": "detag", "suffix": "-detagged" }  
]
```

- string_filter_types
- string_filter_rules

string_filter_types

- <フィルタ名

```
"string_filter_types": {  
  "detag": { "method": "regexp", "pattern": "<[^>]*>", "replace": "" }  
},
```

- method
 - regexp
 - dynamic

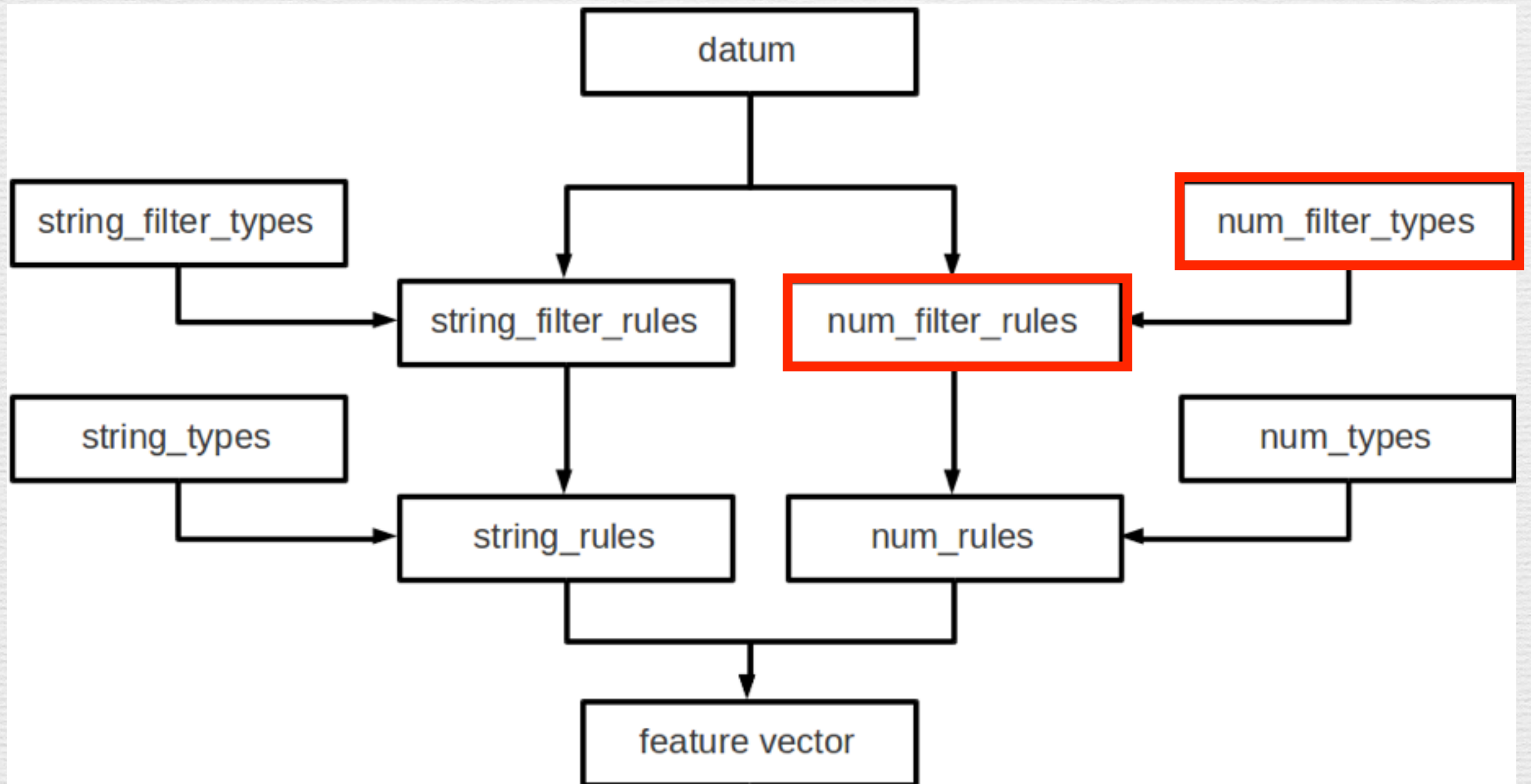
string_filter_rules

- フィルタの適用規則を指定

```
"string_filter_rules": [  
  { "key": "message", "type": "detag", "suffix": "-detagged" }  
]
```

- 引数
 - key
 - except
 - type
 - suffix

データ変換の流れ



num_filter

- 数え年に変換 (

```
"num_filter_types": {  
  "add_1": { "method": "add", "value": "1 " }  
},  
"num_filter_rules": [  
  { "key": "user/age", "type": "add_1", "suffix": "_kazoe" }  
]
```

- num_filter_types
- num_filter_rules

num_filter_types

- <フィルタ名

```
"num_filter_types": {  
  "add_1": { "method": "add", "value": "1 " }  
},
```

- method
 - add
 - dynamic

num_filter_rules

- フィルタの適用規則を指定

```
"num_filter_rules": [  
  { "key": "user/age", "type": "add_1", "suffix": "_kazoe" }  
]
```

- 引数
 - key
 - except
 - type
 - suffix

filter

```
(  
  [  
    ("user/id", "ippy"),  
    ("user/name", "Loren Ipsum"),  
    ("message", "<H>Hello World</H>"),  
    ("message-detagged", "Hello World")  
  ],  
  [  
    ("user/age", 29.0),  
    ("user/age_kazoe", 30.0),  
    ("user/income", 100000.0)  
  ]  
)
```