KNIME and the Web – Extract, Test, Automate

KNIME Spring Summit,
Berlin, 25.02.2016

Philipp Katz,

Our Background

- Three former PhD students at TU Dresden (me, Klemens Muthmann, David Urbansky)
- Computer Science, Information Extraction
- After PhD, each of us founded a startup

SEMKNOX
Smart Semantic Product Search

LINEUPR

CYFACE
(fancy logo under construction)

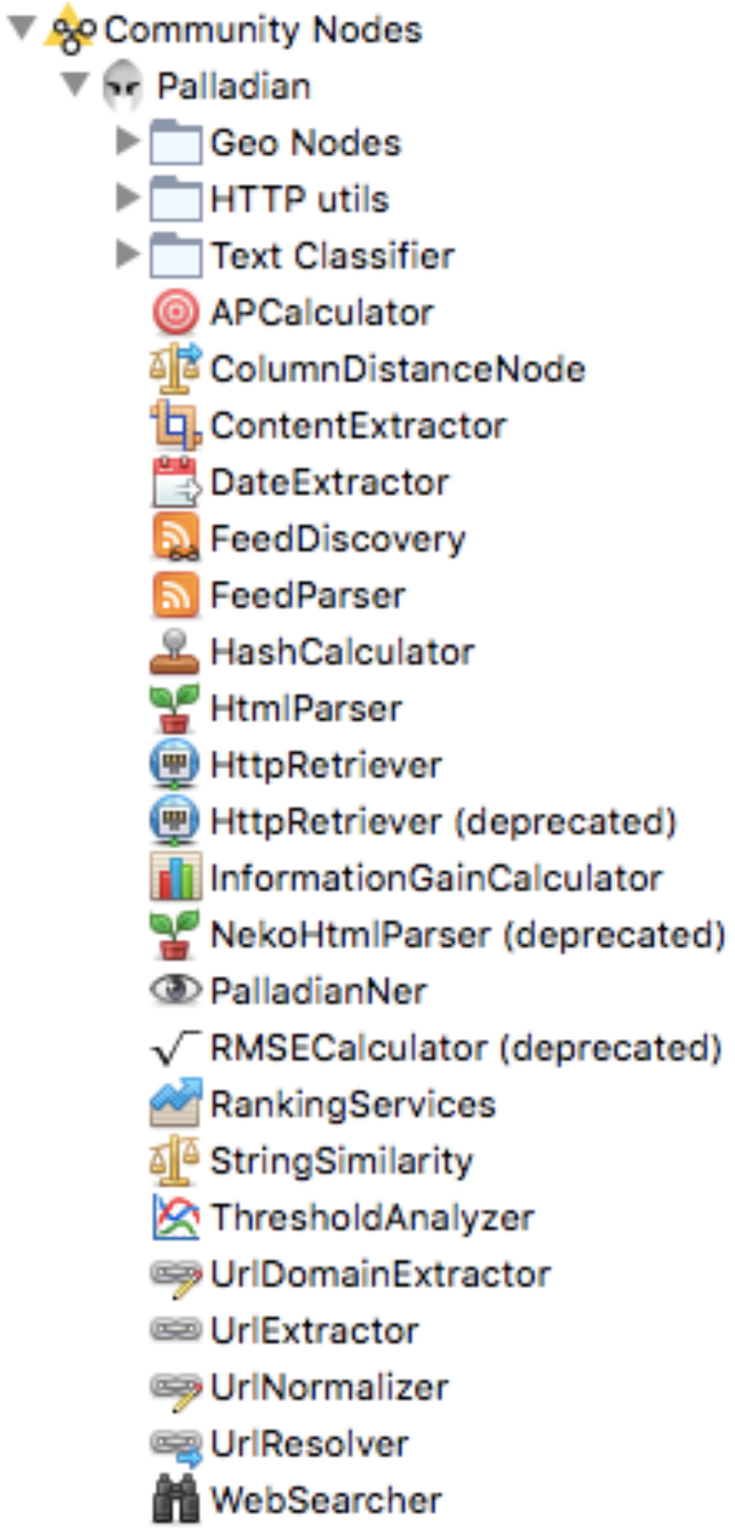
Palladian Nodes

Palladian?

- Java-based toolkit for information retrieval started in 2009
- Palladian KNIME nodes since 2011
- Used in commercial and academic projects
- Available from KNIME Community Contributions download site

The Palladian Nodes

- Text classification
- Content extraction
- Date extraction
- Named entity recognition
- Geo data extraction
- Web page, image, news search
- HTML, RSS, Atom parsing
- Ranking value retrieval
- Evaluation metrics



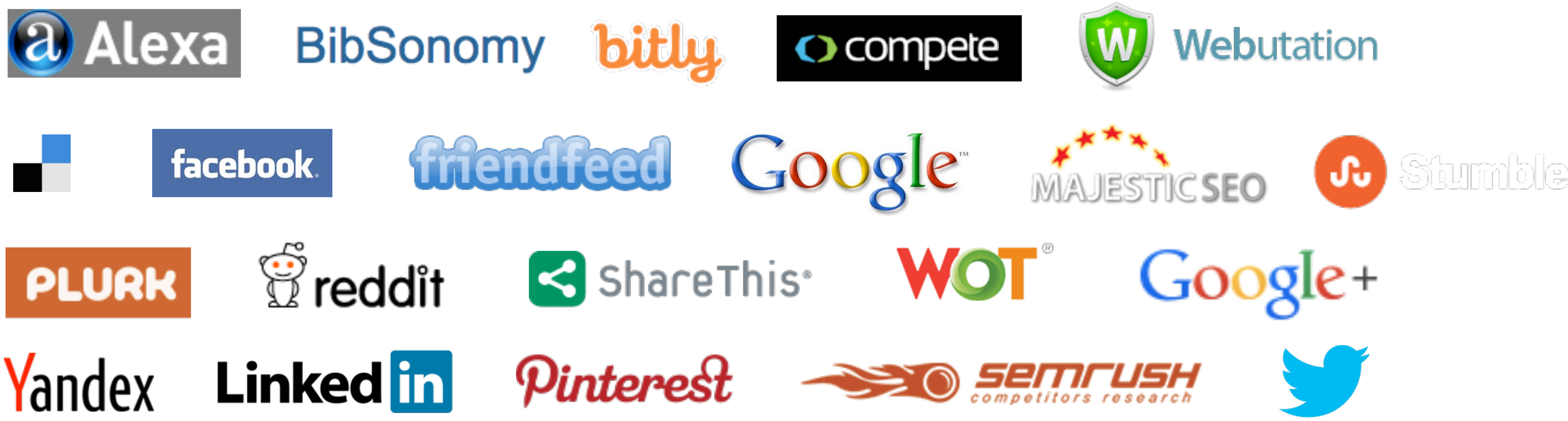
- ▼ Community Nodes
  - ▼ Palladian
    - ▶ Geo Nodes
    - ▶ HTTP utils
    - ▶ Text Classifier
    - APCalculator
    - ColumnDistanceNode
    - ContentExtractor
    - DateExtractor
    - FeedDiscovery
    - FeedParser
    - HashCalculator
    - HtmlParser
    - HttpRetriever
    - HttpRetriever (deprecated)
    - InformationGainCalculator
    - NekoHtmlParser (deprecated)
    - PalladianNer
    - RMSECalculator (deprecated)
    - RankingServices
    - StringSimilarity
    - ThresholdAnalyzer
    - UrlDomainExtractor
    - UrlExtractor
    - UrlNormalizer
    - UrlResolver
    - WebSearcher

Access Web APIs

- Web Searcher

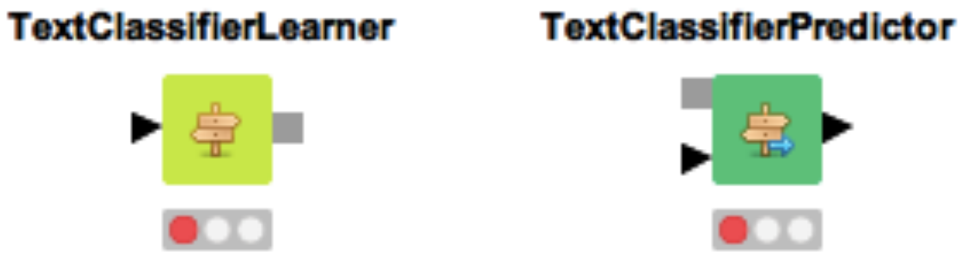
Google
bing
facebook
DuckDuckGo
vimeo
YouTube
Yandex
flickr from YAHOO!

- Ranking Services



Text Classification

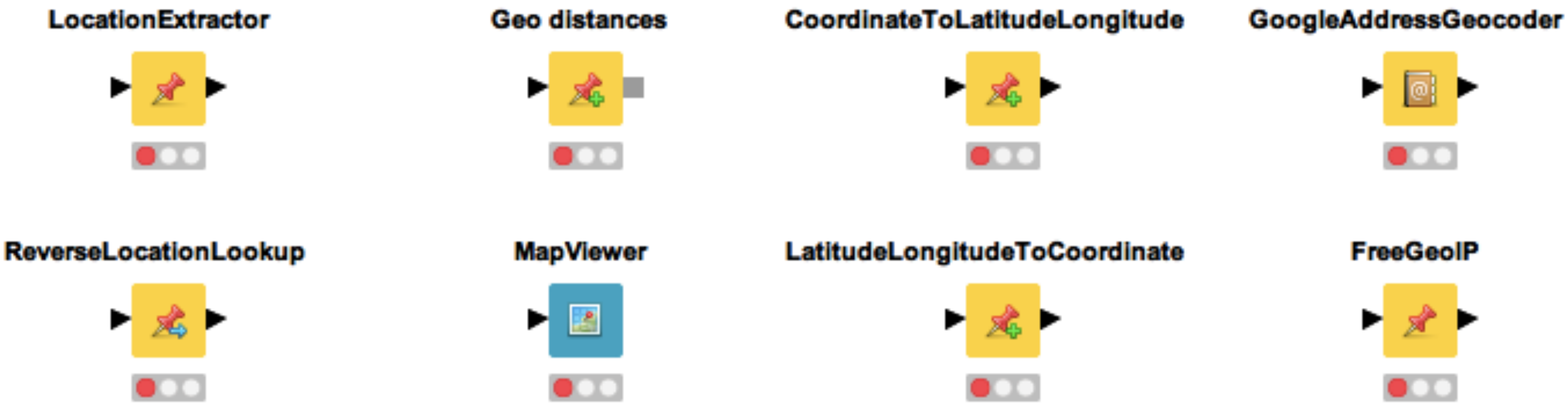
- Very simple, one predictor, one learner



- *n*-gram features and Naïve Bayes scoring
- Optimized for big amounts of training data
- Learner is now *streamable*, Predictor soon
- Competitive accuracy for many use cases

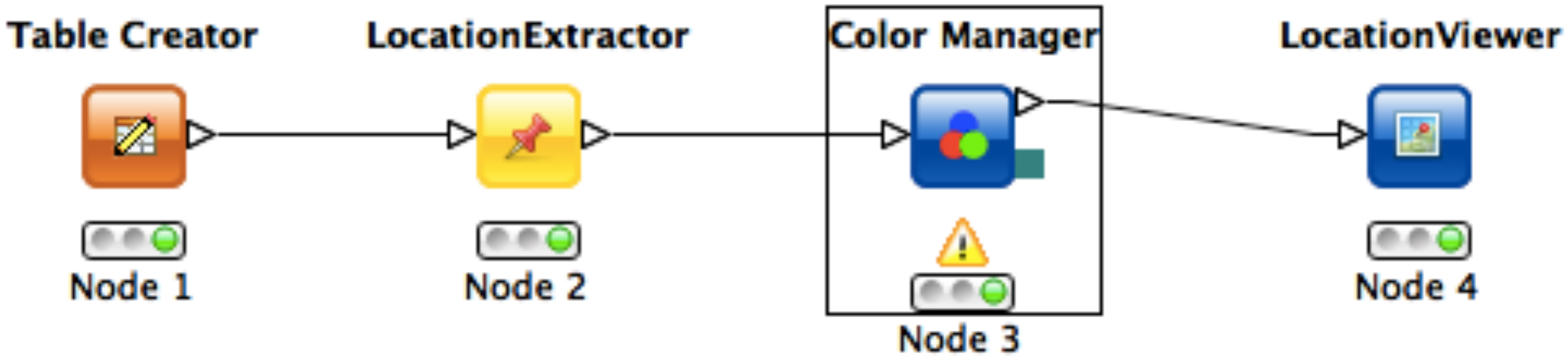
Geographic Data

- Was cooking for a while, added after last year's summit due to popular demand
- **New:** Nodes for IP and address lookup
- **New:** Use local gazetteer as source for location extraction node



Geographic Data

- Extract and disambiguate locations from unstructured text, visualize them on the map



Geographic Data

- Extract and disambiguate locations from unstructured text, visualize them on the map

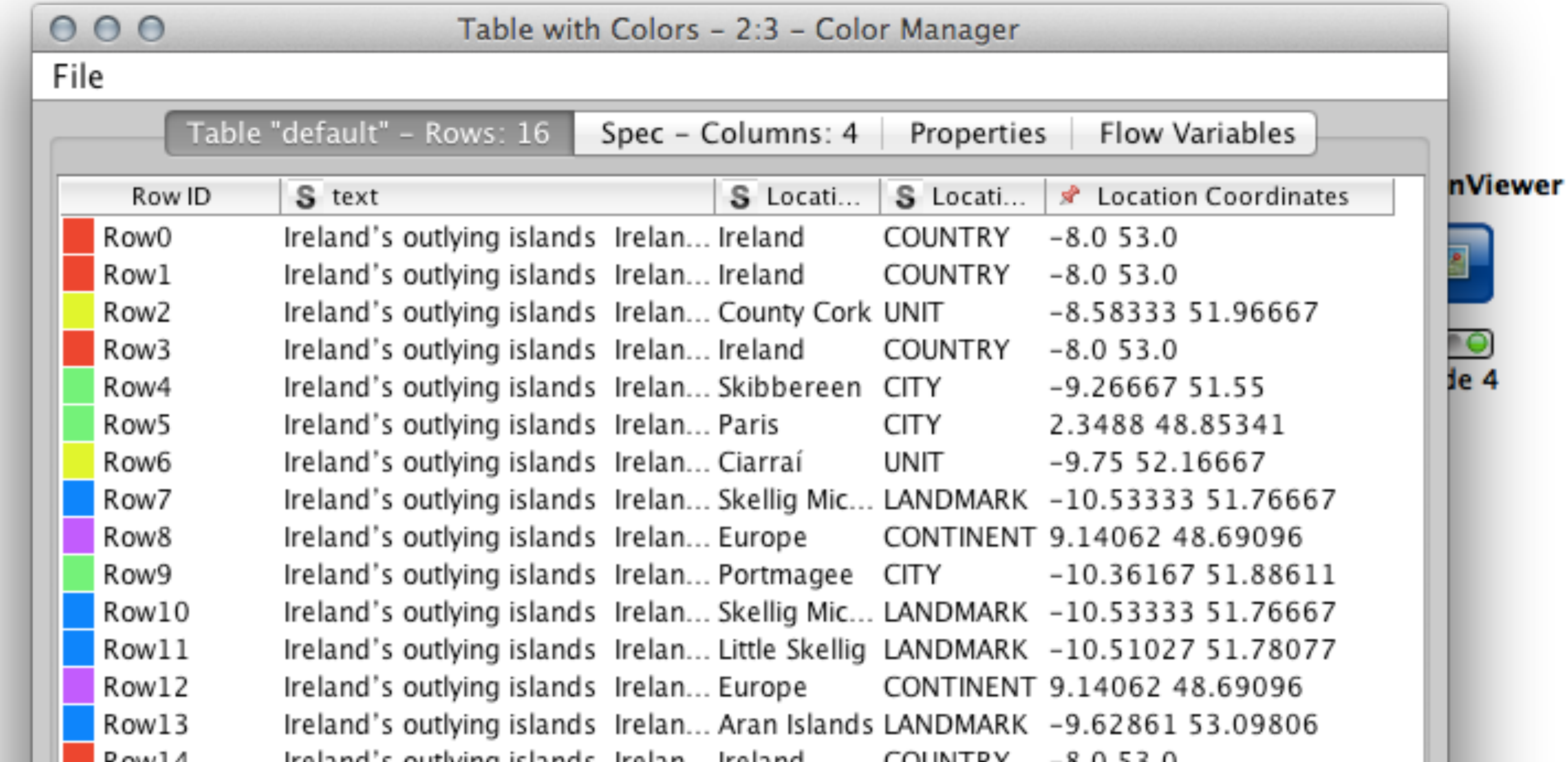


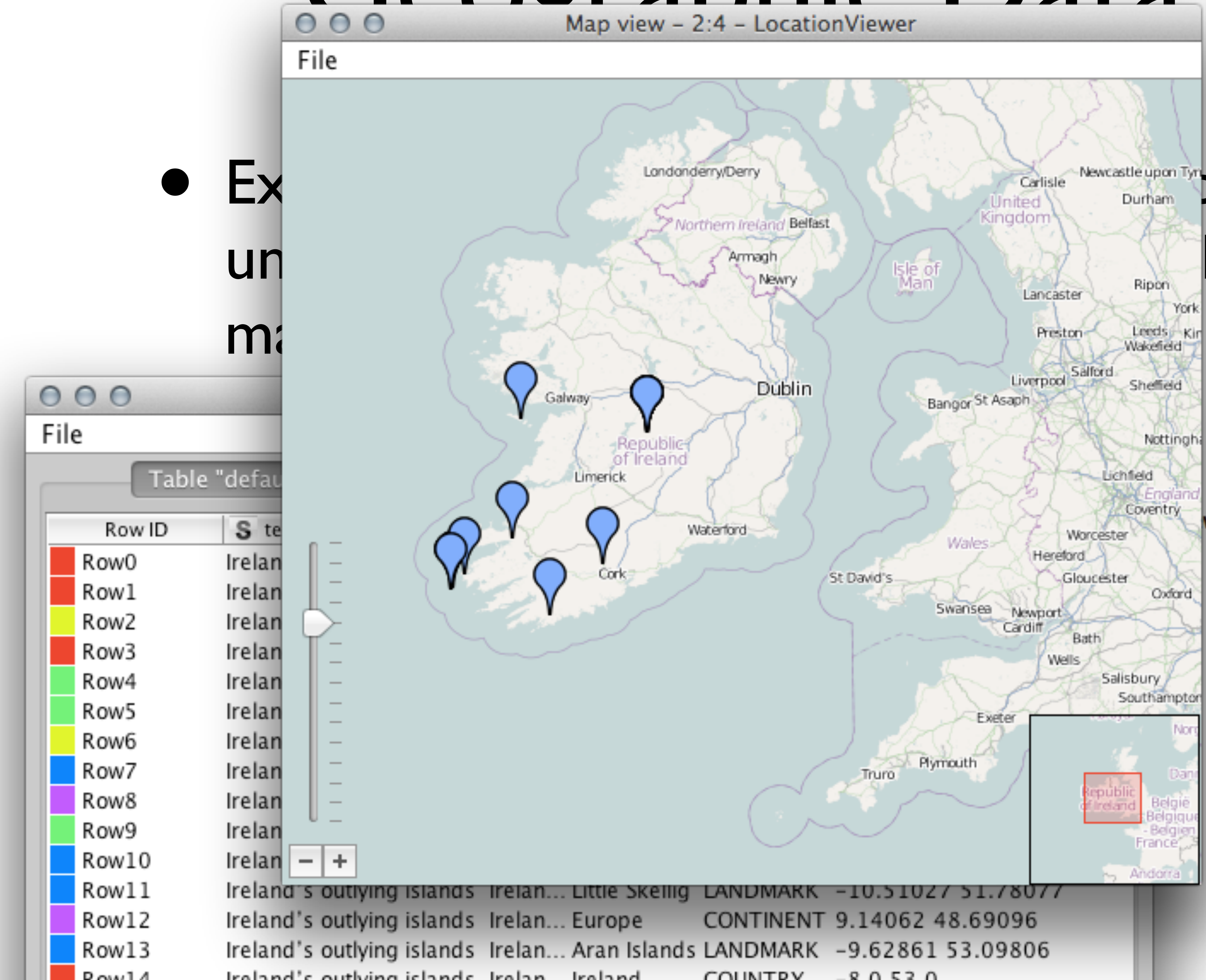
Table with Colors - 2:3 - Color Manager
File
Table "default" - Rows: 16
Spec - Columns: 4
Properties
Flow Variables
nViewer
de 4

Row ID	S text	S Locati...	S Locati...	Location Coordinates
🟥 Row0	Ireland's outlying islands Irelan...	Ireland	COUNTRY	-8.0 53.0
🟥 Row1	Ireland's outlying islands Irelan...	Ireland	COUNTRY	-8.0 53.0
🟨 Row2	Ireland's outlying islands Irelan...	County Cork	UNIT	-8.58333 51.96667
🟥 Row3	Ireland's outlying islands Irelan...	Ireland	COUNTRY	-8.0 53.0
🟩 Row4	Ireland's outlying islands Irelan...	Skibbereen	CITY	-9.26667 51.55
🟩 Row5	Ireland's outlying islands Irelan...	Paris	CITY	2.3488 48.85341
🟨 Row6	Ireland's outlying islands Irelan...	Ciarraí	UNIT	-9.75 52.16667
🟦 Row7	Ireland's outlying islands Irelan...	Skellig Mic...	LANDMARK	-10.53333 51.76667
🟪 Row8	Ireland's outlying islands Irelan...	Europe	CONTINENT	9.14062 48.69096
🟩 Row9	Ireland's outlying islands Irelan...	Portmagee	CITY	-10.36167 51.88611
🟦 Row10	Ireland's outlying islands Irelan...	Skellig Mic...	LANDMARK	-10.53333 51.76667
🟦 Row11	Ireland's outlying islands Irelan...	Little Skellig	LANDMARK	-10.51027 51.78077
🟪 Row12	Ireland's outlying islands Irelan...	Europe	CONTINENT	9.14062 48.69096
🟦 Row13	Ireland's outlying islands Irelan...	Aran Islands	LANDMARK	-9.62861 53.09806
🟥 Row14	Ireland's outlying islands Irelan...	Ireland	COUNTRY	-8.0 53.0

Geographic Data

- Ex… om
  un… he
  ma…

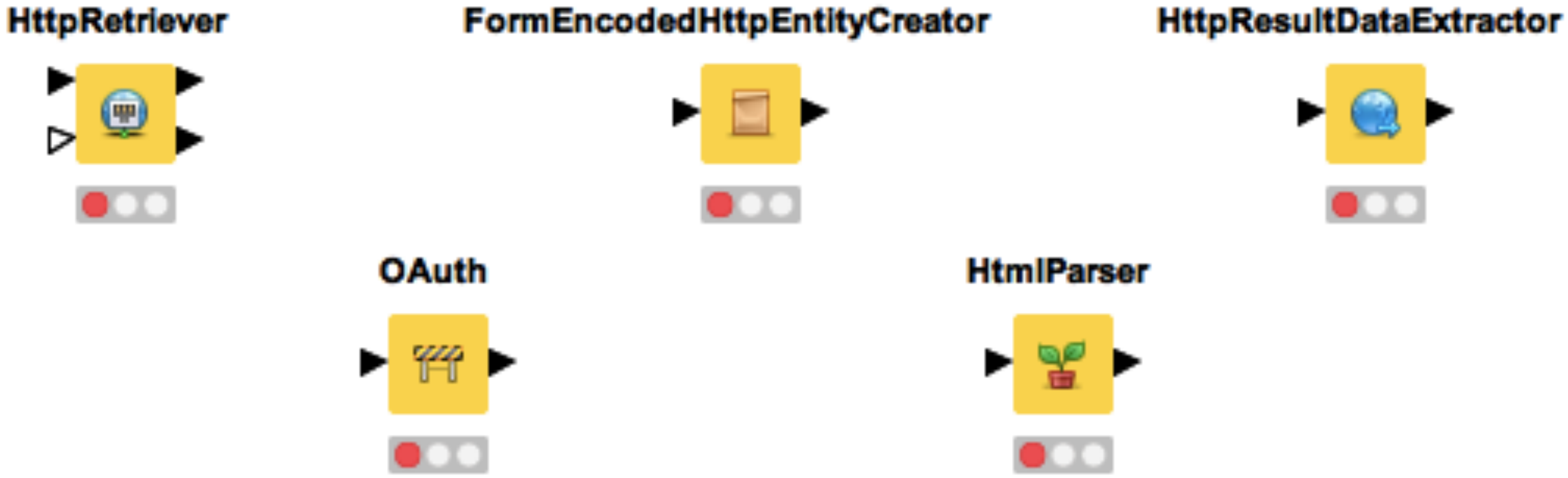
wer

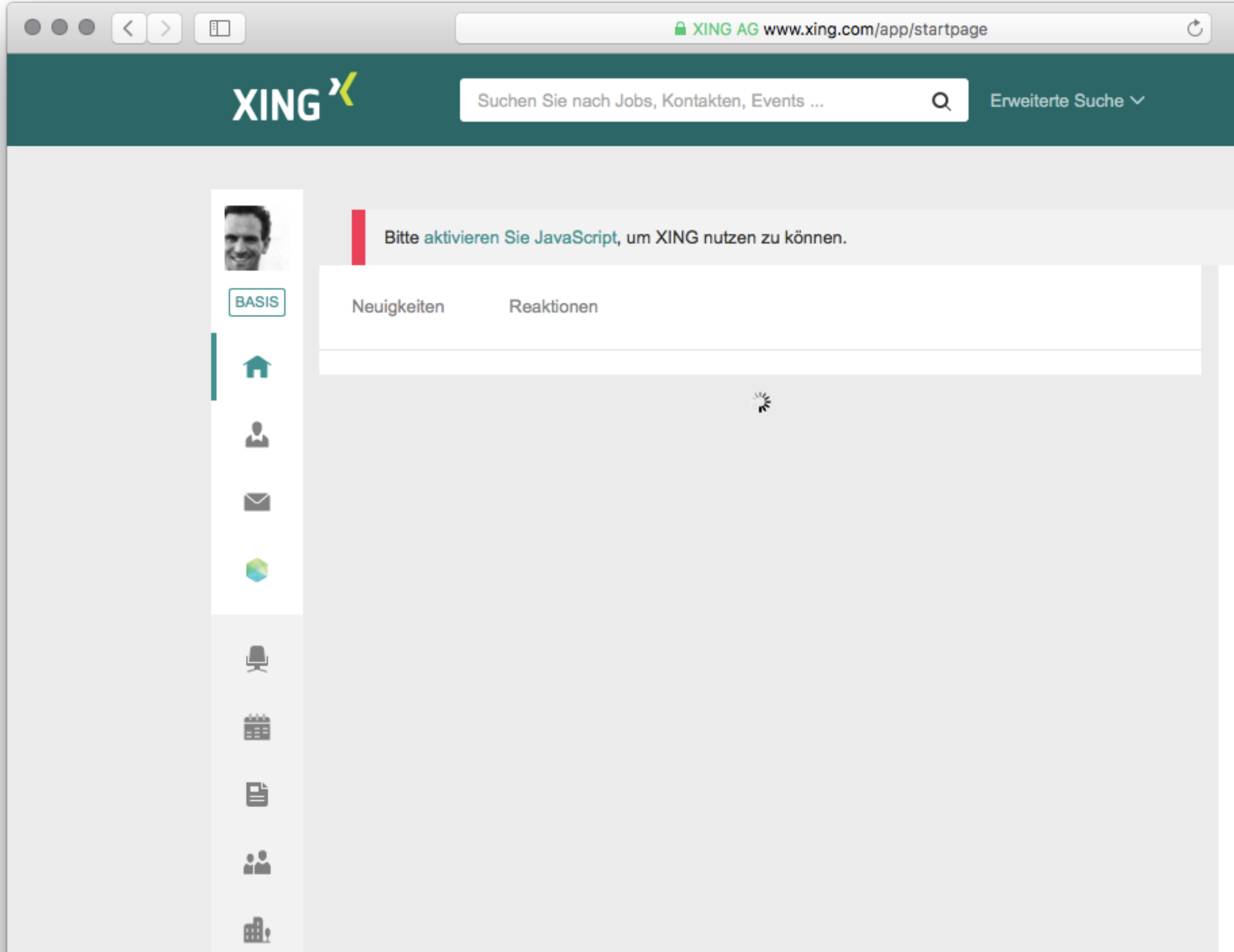


	Row ID	S te					
🟥	Row0	Irelan					
🟥	Row1	Irelan					
🟨	Row2	Irelan					
🟥	Row3	Irelan					
🟩	Row4	Irelan					
🟩	Row5	Irelan					
🟨	Row6	Irelan					
🟦	Row7	Irelan					
🟪	Row8	Irelan					
🟩	Row9	Irelan					
🟦	Row10	Irelan					
🟦	Row11	Ireland's outlying islands	Irelan...	Little Skellig	LANDMARK	-10.51027	51.78077
🟪	Row12	Ireland's outlying islands	Irelan...	Europe	CONTINENT	9.14062	48.69096
🟦	Row13	Ireland's outlying islands	Irelan...	Aran Islands	LANDMARK	-9.62861	53.09806
🟥	Row14	Ireland's outlying islands	Irelan...	Ireland	COUNTRY	-8.0	53.0

HTTP and HTML

- **New:** Support for cookies, headers, and further HTTP methods besides GET
- **New:** Sending arbitrary byte stream content, form-encoding of table data
- **New:** OAuth signing for HTTP requests

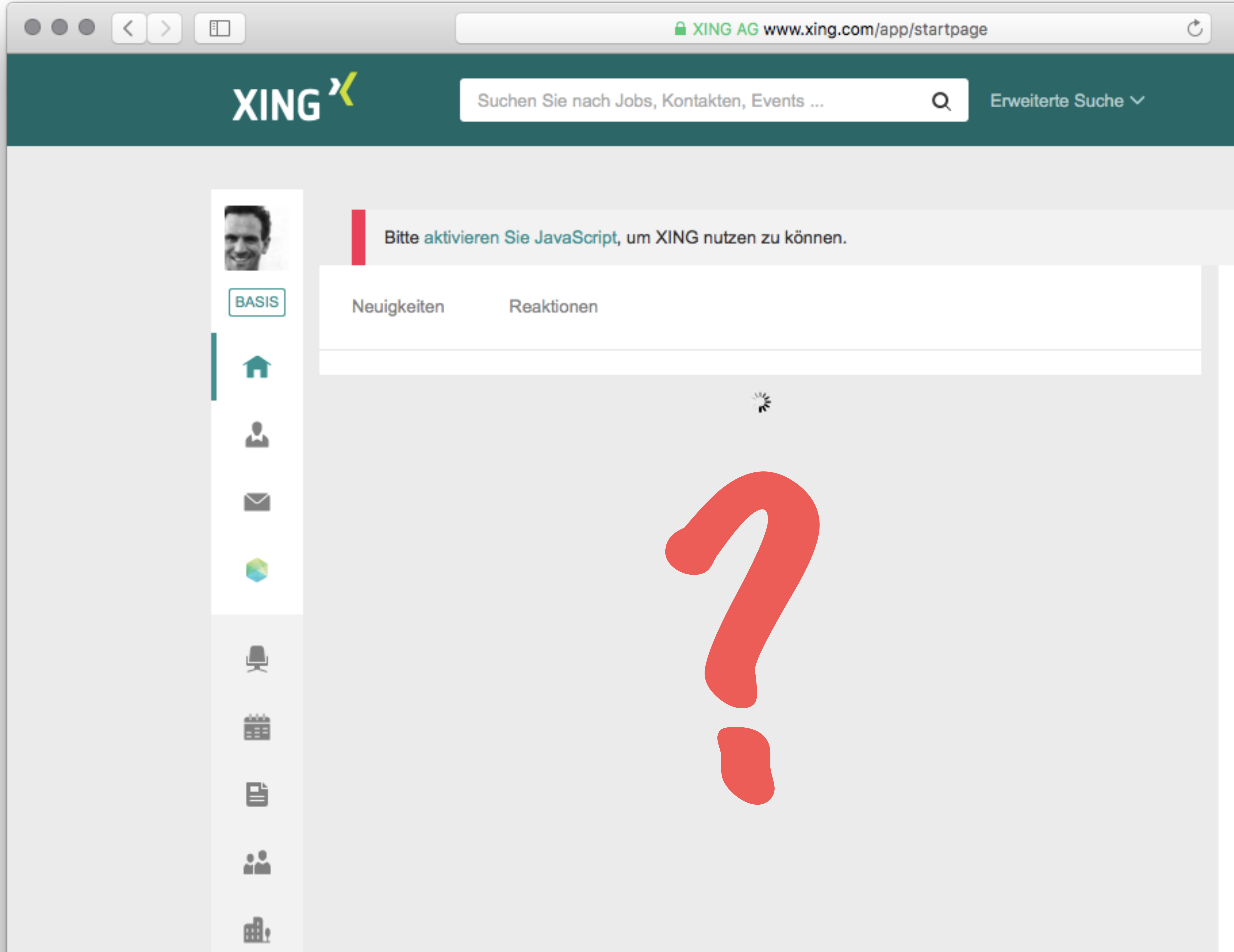




XING AG www.xing.com/app/startpage

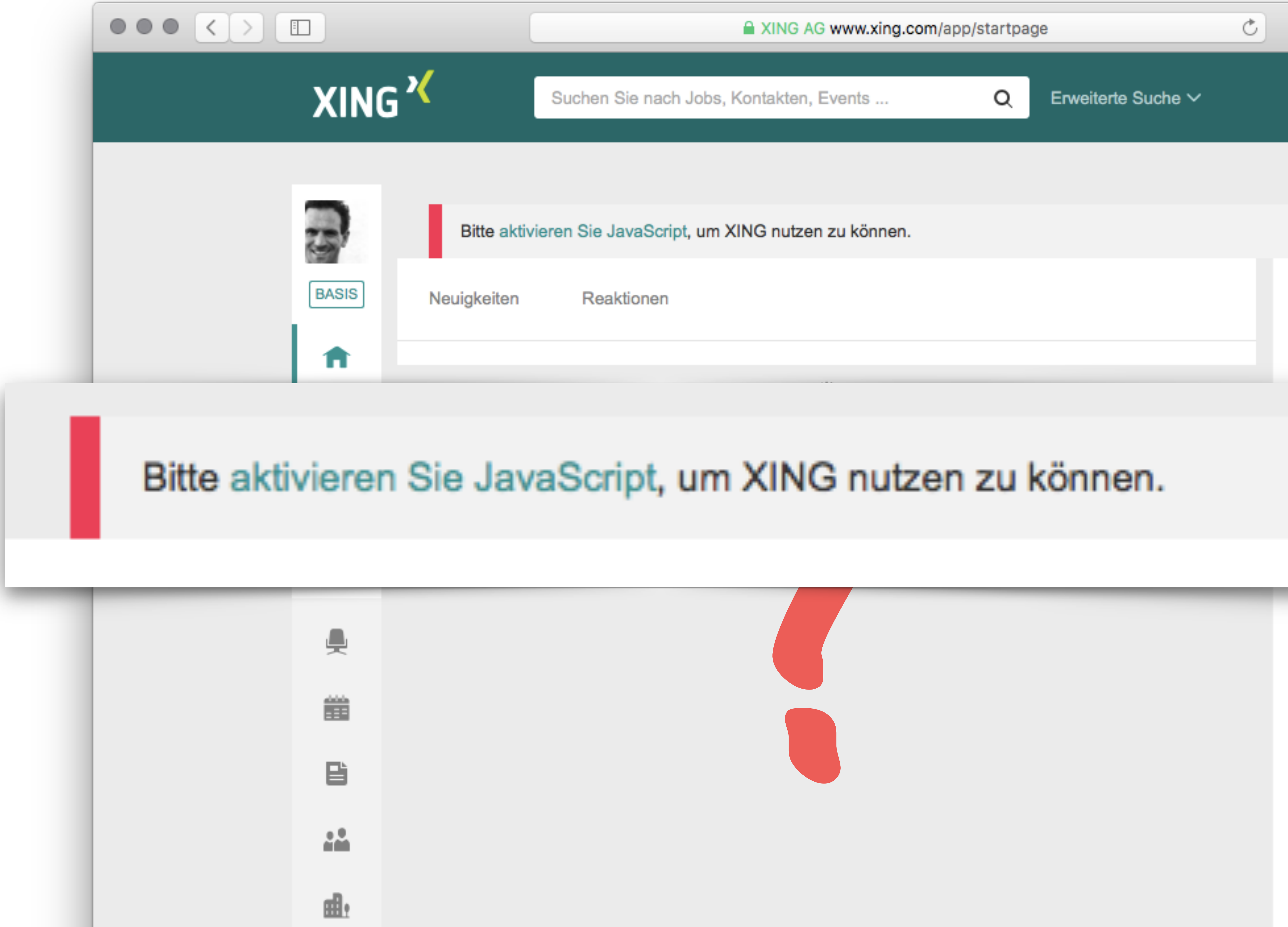
Suchen Sie nach Jobs, Kontakten, Events ...

Bitte aktivieren Sie JavaScript, um XING nutzen zu können.



XING AG www.xing.com/app/startpage
XING
Suchen Sie nach Jobs, Kontakten, Events ...
Erweiterte Suche
BASIS
Neuigkeiten
Reaktionen

Bitte aktivieren Sie JavaScript, um XING nutzen zu können.


XING AG www.xing.com/app/startpage
XING
Suchen Sie nach Jobs, Kontakten, Events ...
Erweiterte Suche
BASIS
Neuigkeiten
Reaktionen

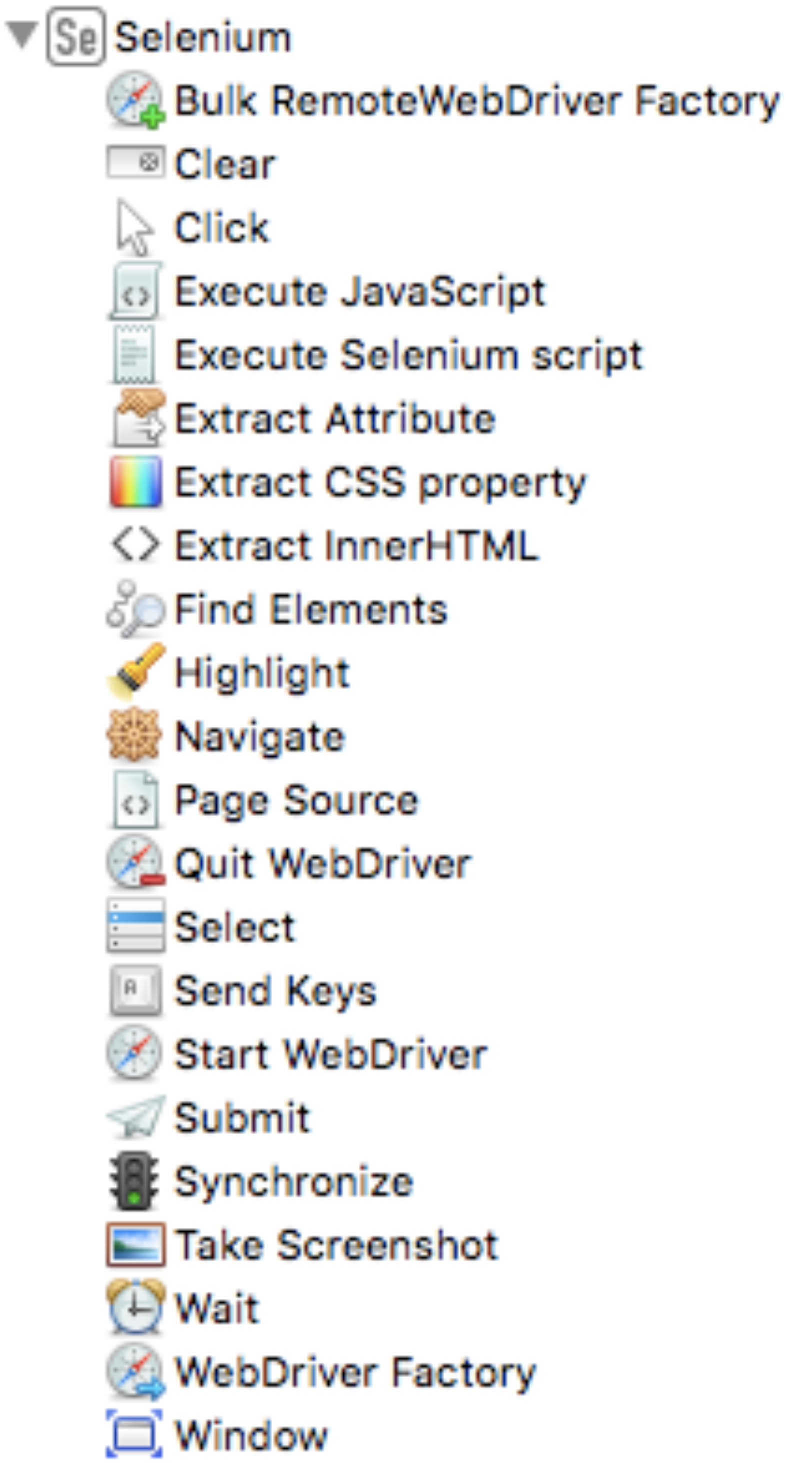
Bitte aktivieren Sie JavaScript, um XING nutzen zu können.

Bitte aktivieren Sie JavaScript, um XING nutzen zu können.

Selenium Nodes

Selenium?

- *"Selenium automates browsers."*
- The **Selenium Nodes** allow to simulate a *real* web browser with KNIME
- Use a KNIME workflow to describe actions and extract all the data you need



- Selenium
  - Bulk RemoteWebDriver Factory
  - Clear
  - Click
  - Execute JavaScript
  - Execute Selenium script
  - Extract Attribute
  - Extract CSS property
  - Extract InnerHTML
  - Find Elements
  - Highlight
  - Navigate
  - Page Source
  - Quit WebDriver
  - Select
  - Send Keys
  - Start WebDriver
  - Submit
  - Synchronize
  - Take Screenshot
  - Wait
  - WebDriver Factory
  - Window

Use Cases

Data extraction

Task automatization

Web application testing

Browser Support

- Local installations

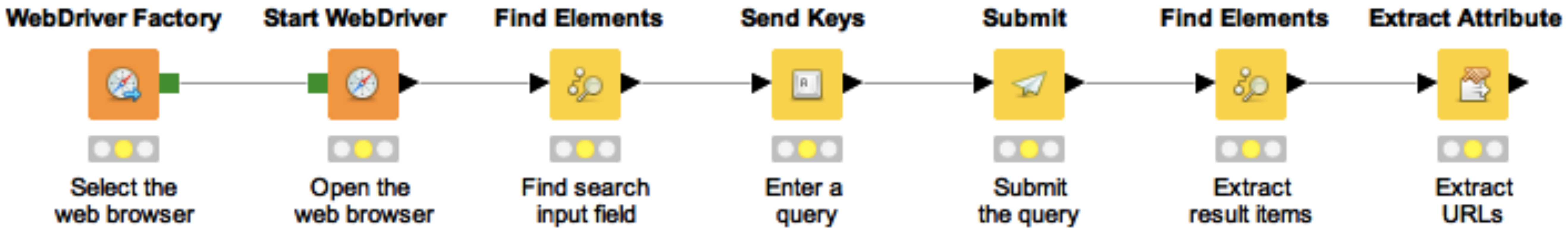
- Headless “browsers”
  - PhantomJS, jBrowserDriver

- Remotely running

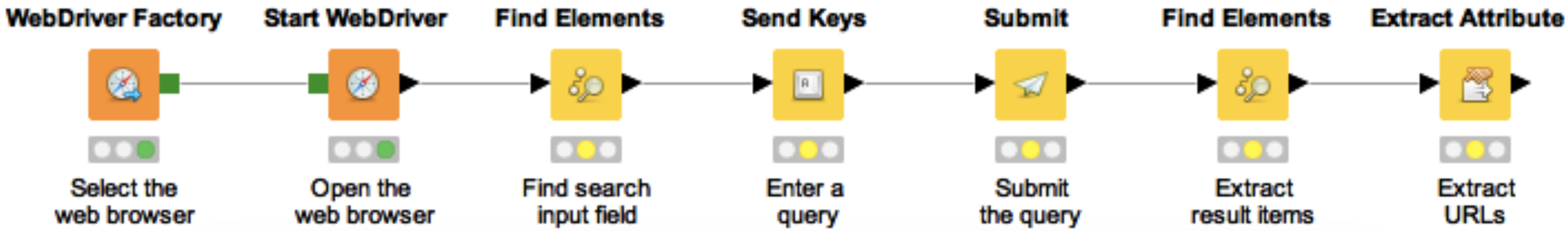
Browser Support

- Remotely running
  - Connect to Selenium servers or VMs on your local network to simulate a variety of operating systems or browsers
  - Use cloud services such as BrowserStack or SauceLabs, which provide ready-to-use Selenium instances (even iOS and Android)

Example Workflow



Example Workflow



WebDriver Factory — Select the web browser

Start WebDriver — Open the web browser

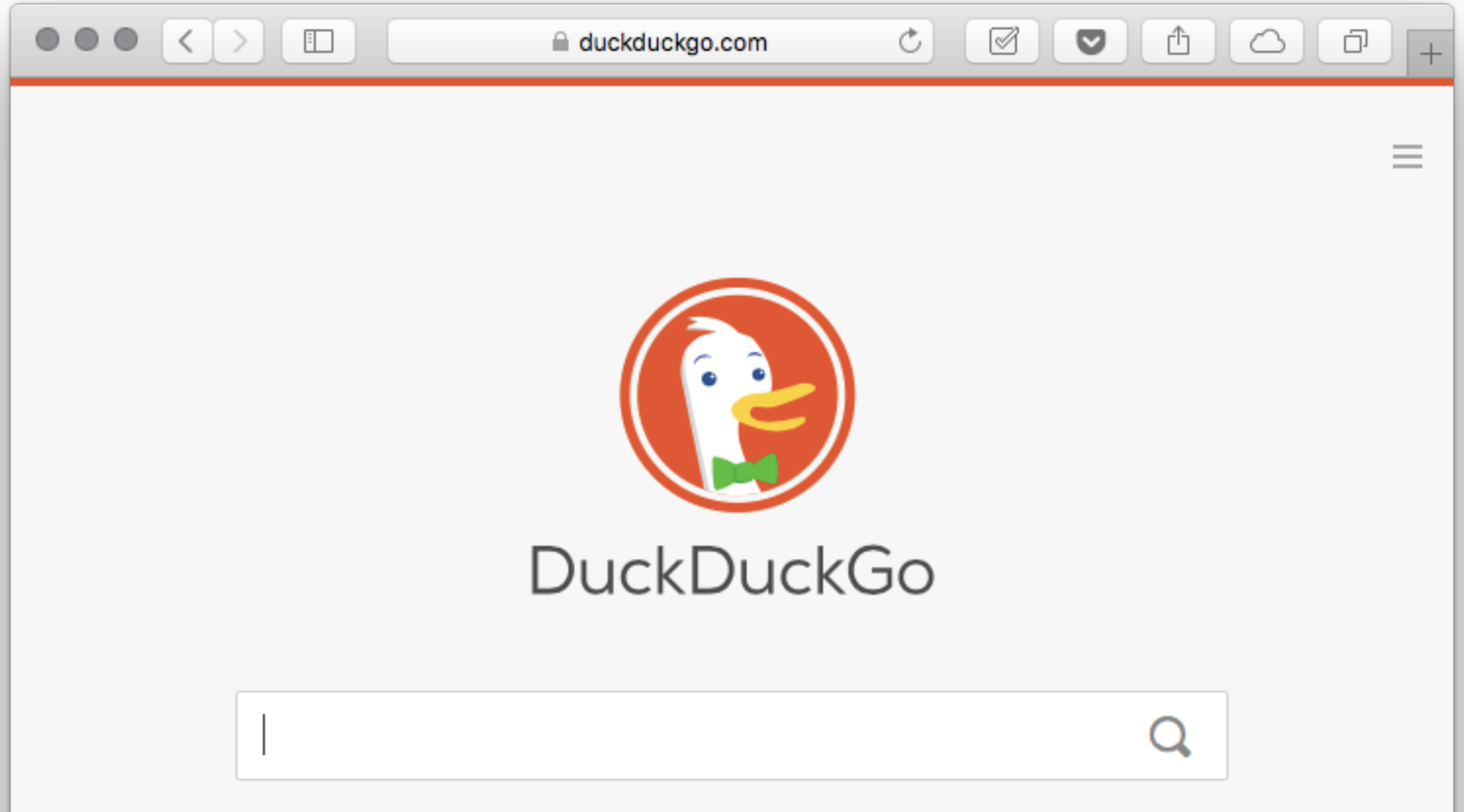
Find Elements — Find search input field

Send Keys — Enter a query

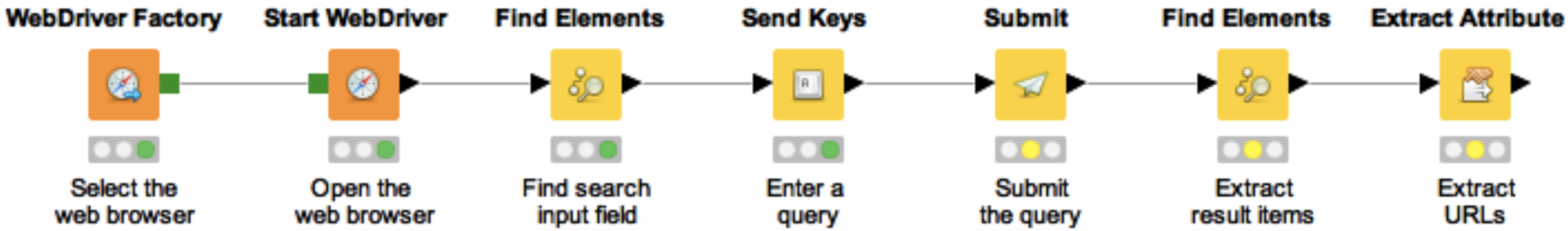
Submit — Submit the query

Find Elements — Extract result items

Extract Attribute — Extract URLs



Example Workflow



WebDriver Factory — Select the web browser

Start WebDriver — Open the web browser

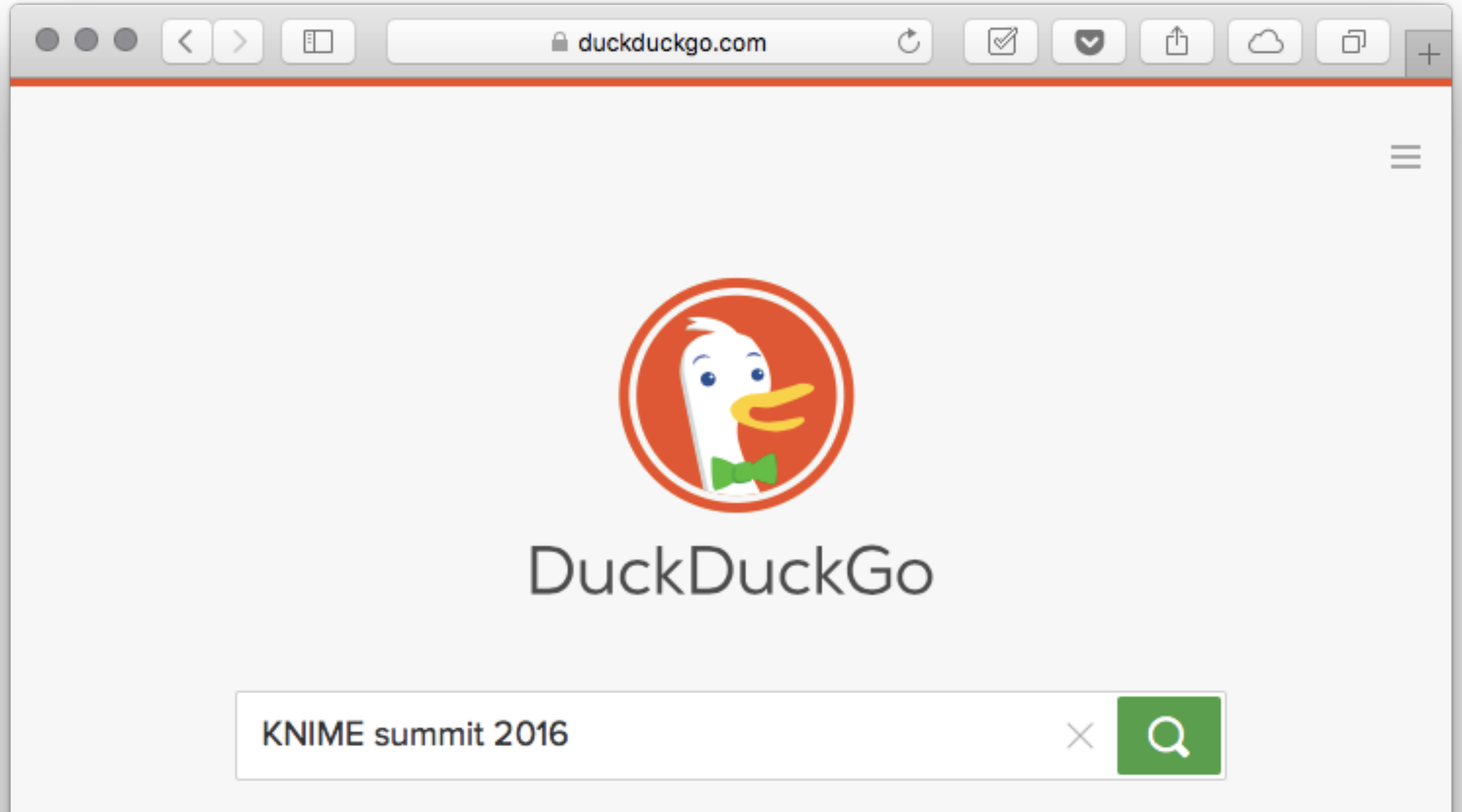
Find Elements — Find search input field

Send Keys — Enter a query

Submit — Submit the query

Find Elements — Extract result items

Extract Attribute — Extract URLs

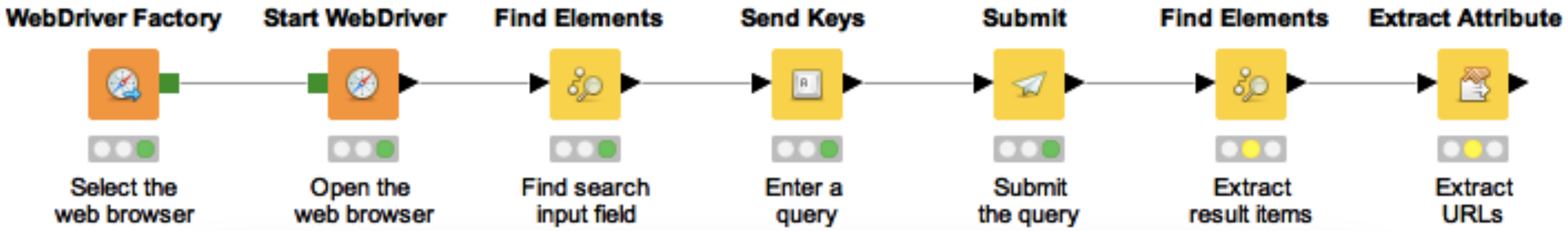


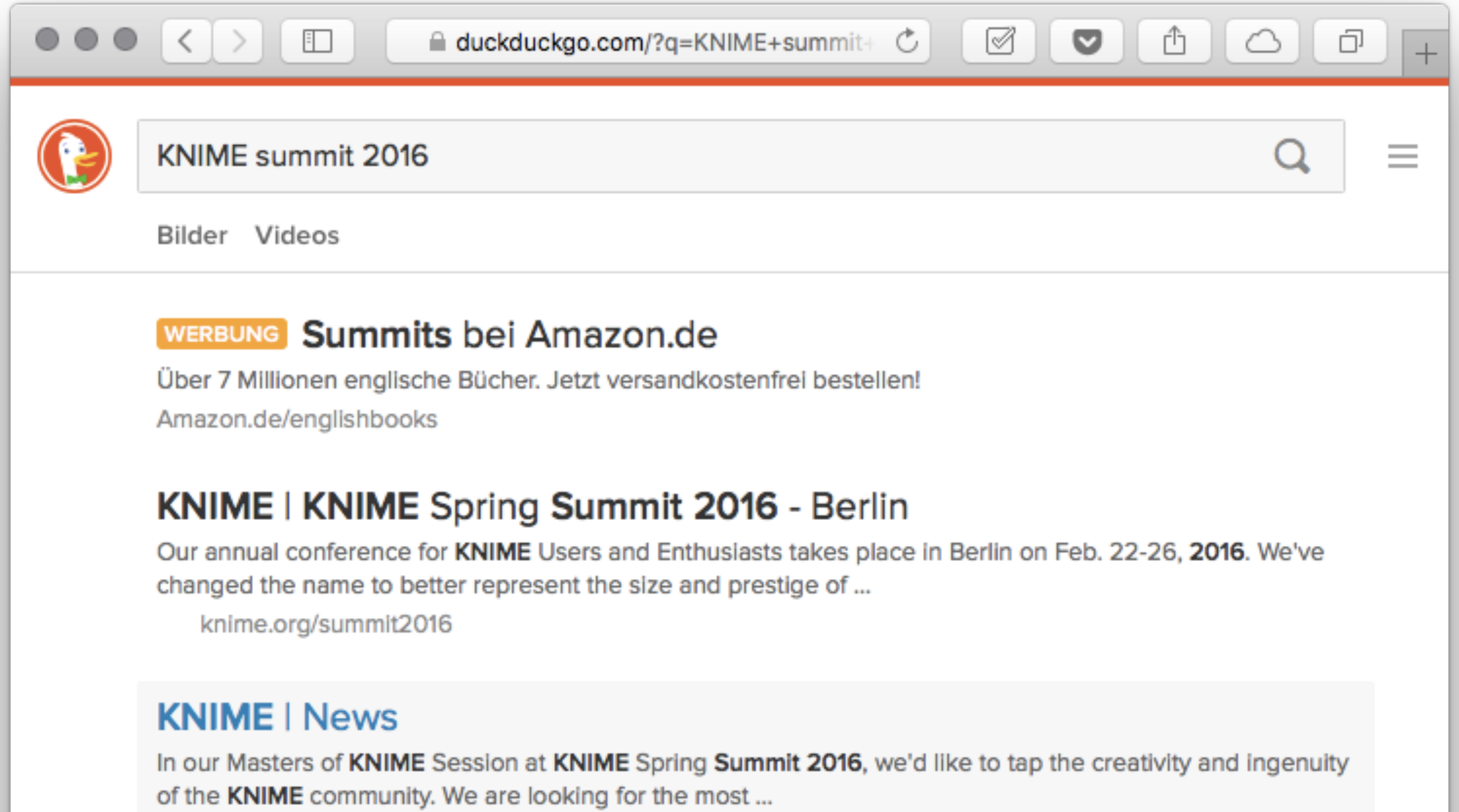
duckduckgo.com

DuckDuckGo

KNIME summit 2016

Example Workflow





KNIME summit 2016

WERBUNG **Summits** bei Amazon.de

Über 7 Millionen englische Bücher. Jetzt versandkostenfrei bestellen!

Amazon.de/englishbooks

KNIME | **KNIME** Spring **Summit 2016** - Berlin

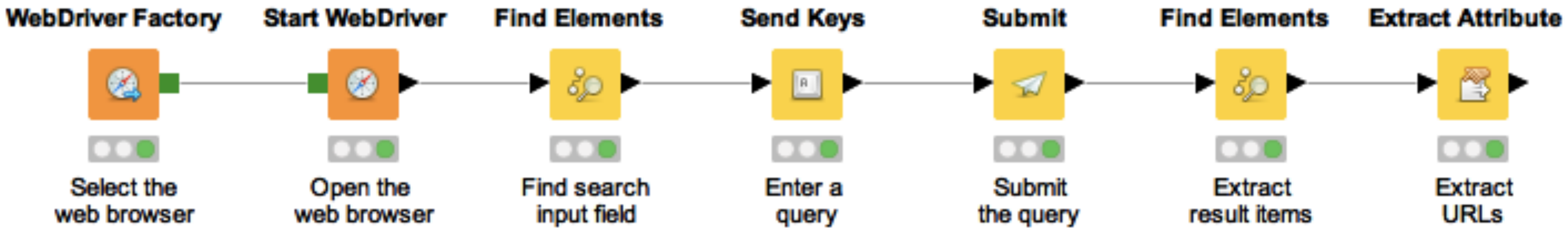
Our annual conference for **KNIME** Users and Enthusiasts takes place in Berlin on Feb. 22-26, **2016**. We've changed the name to better represent the size and prestige of ...

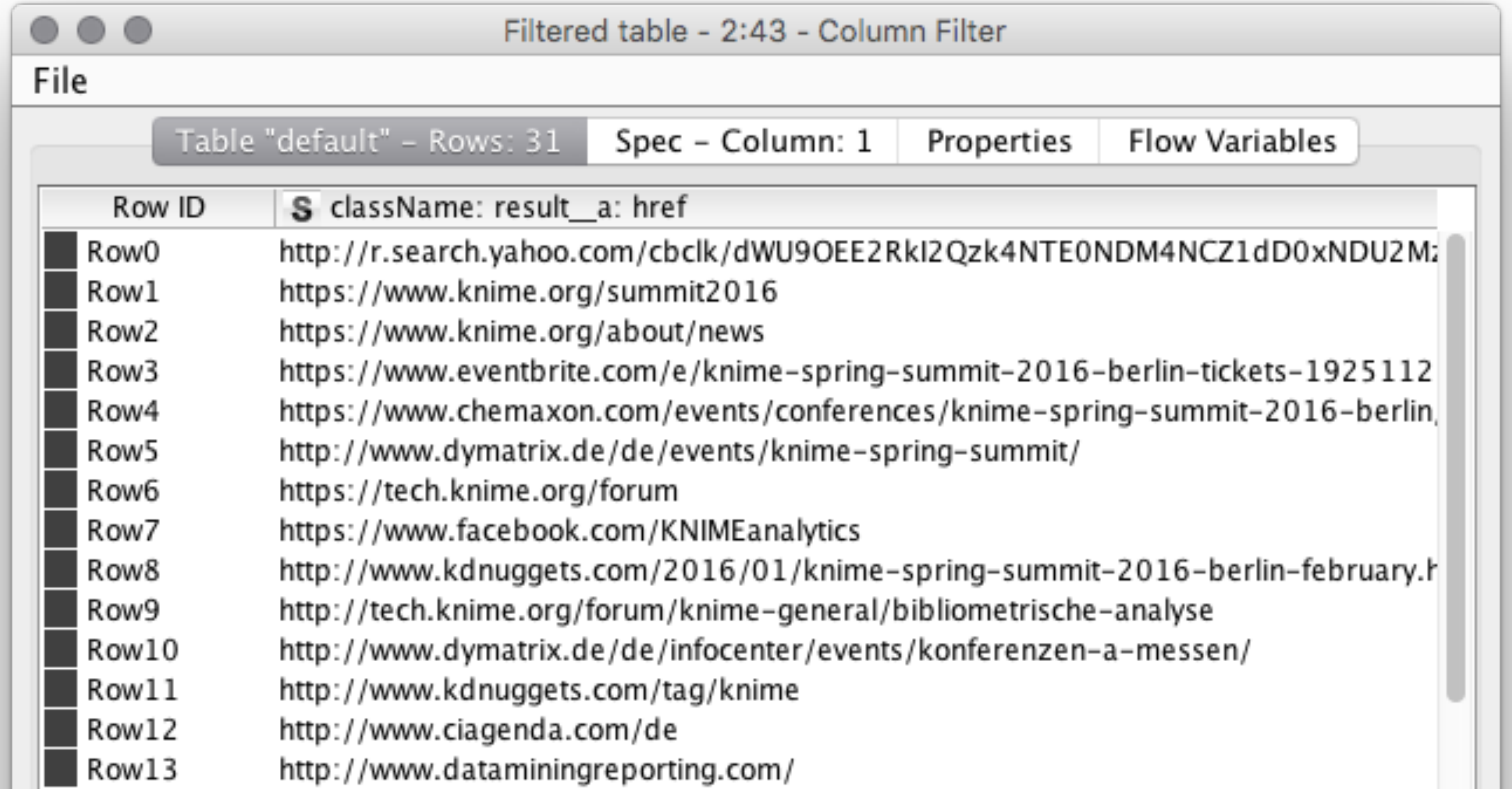
knime.org/summit2016

KNIME | News

In our Masters of **KNIME** Session at **KNIME** Spring **Summit 2016**, we'd like to tap the creativity and ingenuity of the **KNIME** community. We are looking for the most ...

Example Workflow

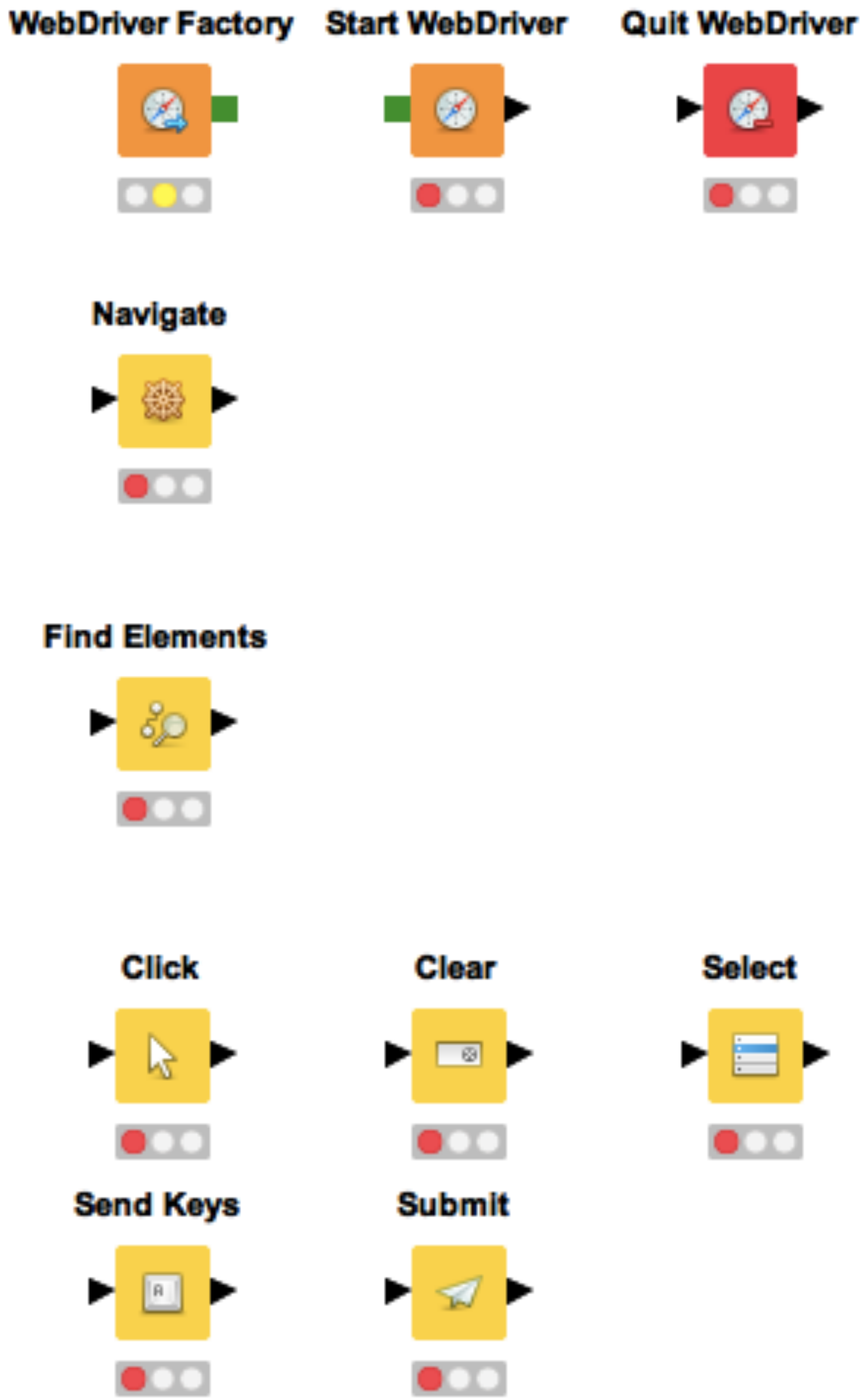



Filtered table - 2:43 - Column Filter
File
Table "default" – Rows: 31
Spec – Column: 1
Properties
Flow Variables

Row ID	className: result__a: href
Row0	http://r.search.yahoo.com/cbclk/dWU9OEE2RkI2Qzk4NTE0NDM4NCZ1dD0xNDU2M
Row1	https://www.knime.org/summit2016
Row2	https://www.knime.org/about/news
Row3	https://www.eventbrite.com/e/knime-spring-summit-2016-berlin-tickets-1925112
Row4	https://www.chemaxon.com/events/conferences/knime-spring-summit-2016-berlin
Row5	http://www.dymatrix.de/de/events/knime-spring-summit/
Row6	https://tech.knime.org/forum
Row7	https://www.facebook.com/KNIMEanalytics
Row8	http://www.kdnuggets.com/2016/01/knime-spring-summit-2016-berlin-february.
Row9	http://tech.knime.org/forum/knime-general/bibliometrische-analyse
Row10	http://www.dymatrix.de/de/infocenter/events/konferenzen-a-messen/
Row11	http://www.kdnuggets.com/tag/knime
Row12	http://www.ciagenda.com/de
Row13	http://www.dataminingreporting.com/

Node Overview

- Configure, start, and quit web browsers
- Navigate
- Locate Elements (using attributes, XPath, or CSS)
- Interact with Elements (click, input text, select, submit, …)



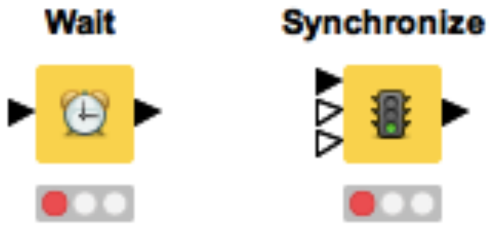
Node Overview

- Highlight elements
- Take screenshots

- Extract data (page source, text content, attributes, …)

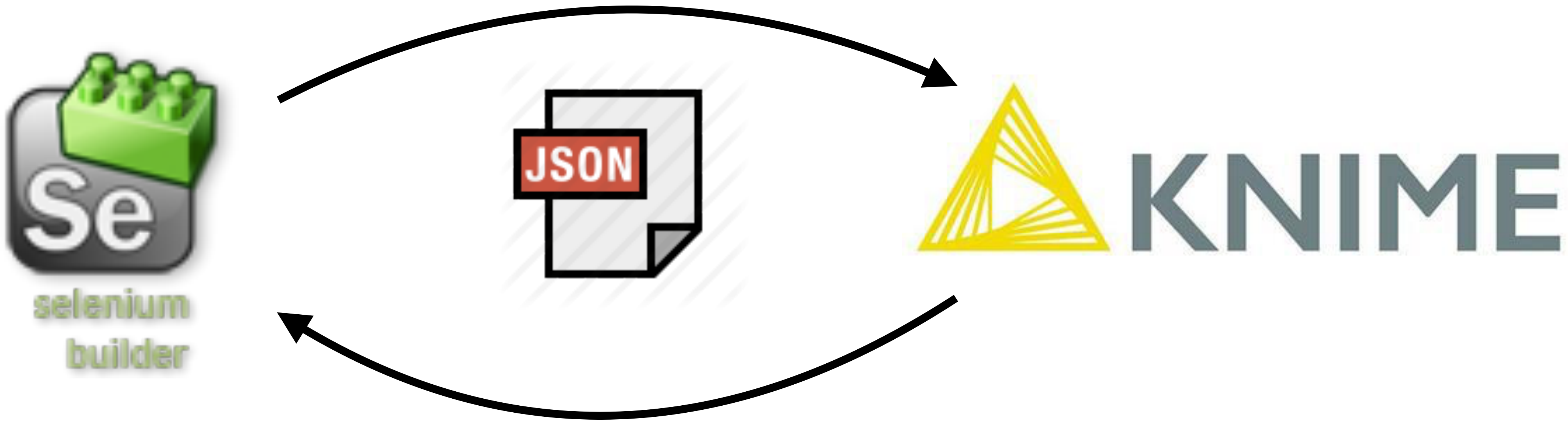
- Execute JavaScript
- Execute Selenium script

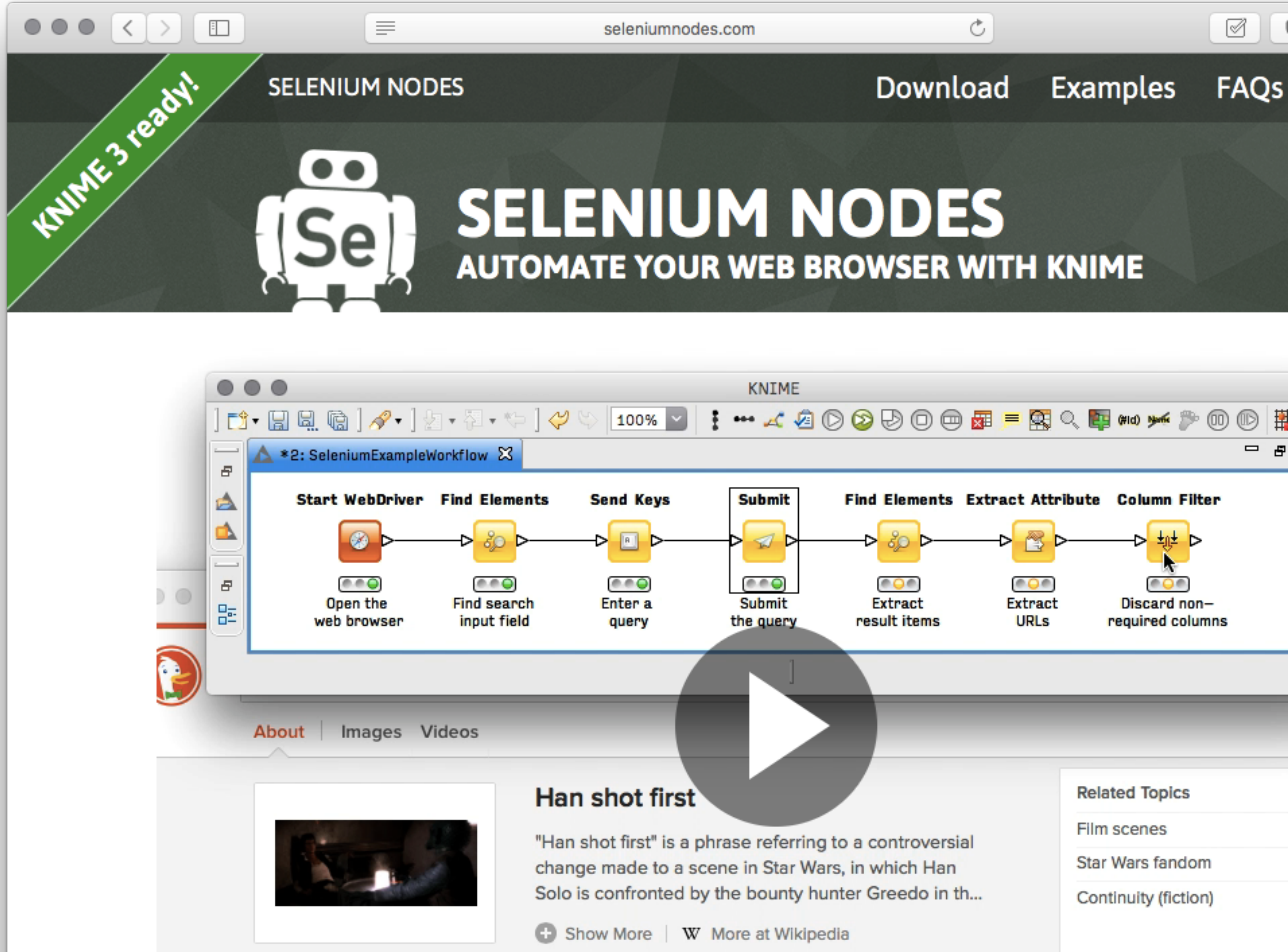
- **Waiting and synchronization**



Outlook

- More sample workflows
- Documentation, how-tos, …
- **Workflow import and export for Selenium Scripts**





SELENIUM NODES

AUTOMATE YOUR WEB BROWSER WITH KNIME

Han shot first

"Han shot first" is a phrase referring to a controversial change made to a scene in Star Wars, in which Han Solo is confronted by the bounty hunter Greedo in th...

Questions?
Get in touch!

mail@seleniumnodes.com
KNIME forum