# Machbarkeitsstudie zur Automatisierten Preiserhebung im Internet

# Das Projekt MAPI

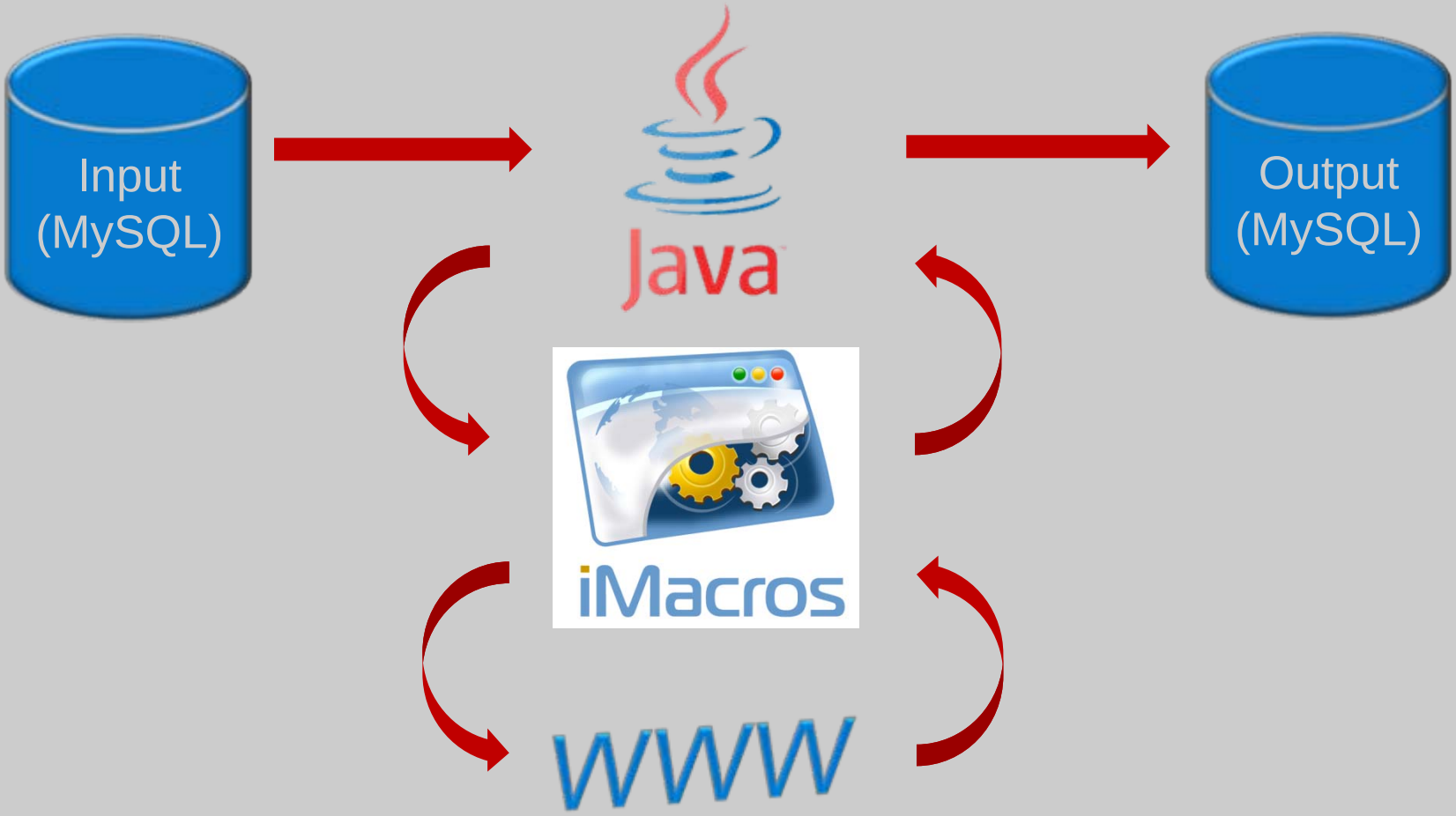
Machbarkeitsstudie zur Automatisierten Preiserhebung im Internet

- **gehört zum Themengebiet Multipurpose Price Statistics**
- **Teilfinanzierung von EUROSTAT zu 90%**
- **Projektdauer: 01.01.2012 bis 31.03.2014**

# Zielsetzung des Projekts

- **Es soll untersucht werden, inwieweit Preiserhebungen im Internet mithilfe automatisierter Verfahren (sogenanntes web scraping oder screen scraping) möglich und sinnvoll sind.**
- **Aufwand und Fehleranfälligkeit manueller Preiserhebungen im Internet können gesenkt werden**
- **Ersparnis ist hauptsächlich bei Erhebungen zu erwarten, die wiederholt in kürzeren Zeitabständen durchgeführt werden**
  - **Verbraucherpreisstatistik**
  - **Kaufkraftparitäten nur bei bestimmten Erhebungen (Flugreisen, Hotels)**

# Verwendete Software



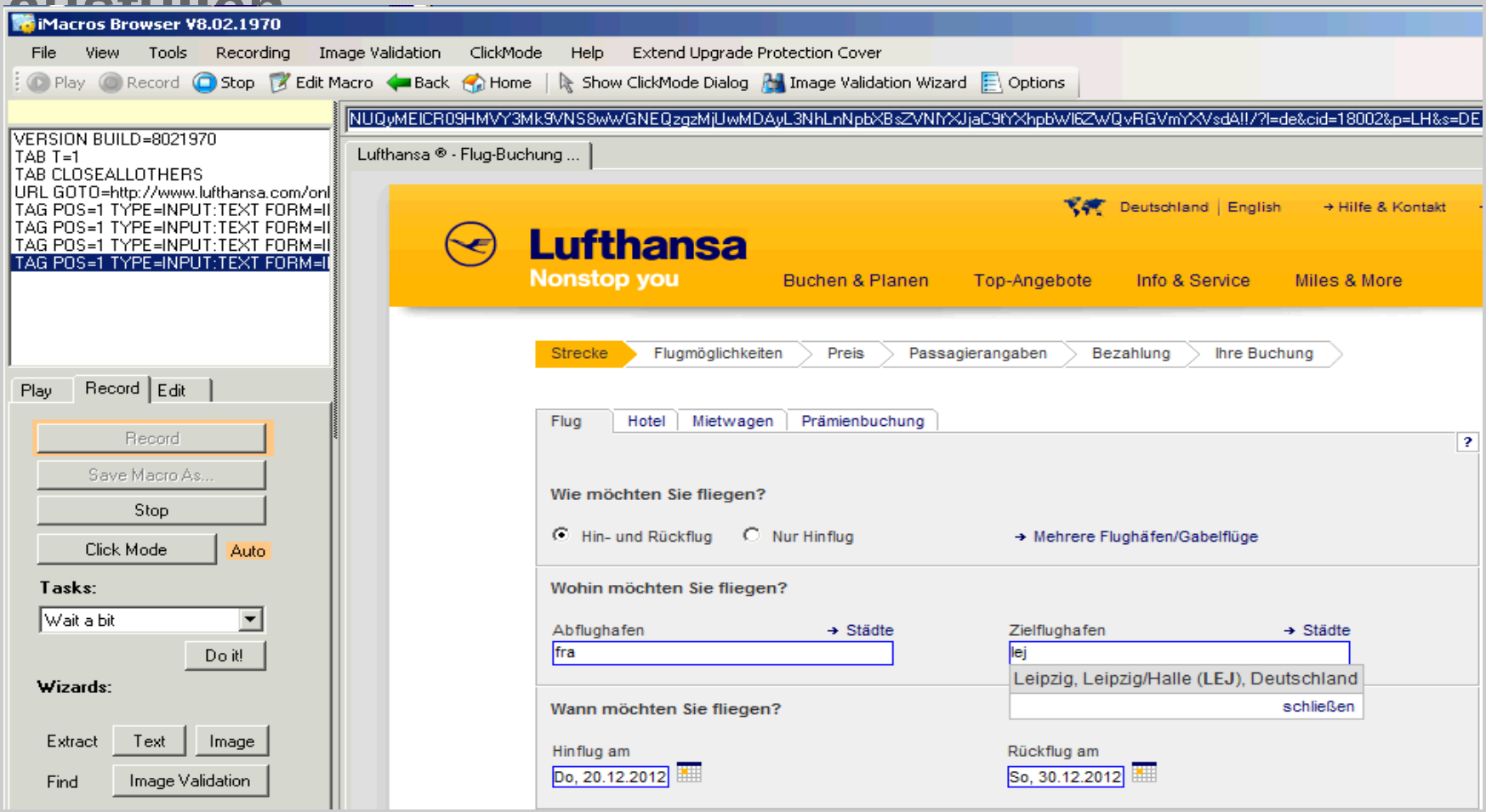
# iMacros

- Spezielles Aufzeichnungstool für Web Scraping
- Einfache Aufzeichnung, ähnlich wie Excel Makros
- Es wird Code generiert, welcher nachträglich bearbeitet werden kann
- Ausfüllen von Formularen, Anklicken von Links, Extrahieren von Inhalten

# Java

- Abbildung von Programmlogik
- Bereinigung von Daten
- Auslesen von Daten aus der Datenbank
- Daten in Datenbank ablegen

# iMacros – Skripte aufzeichnen, Formular ausfüllen



# iMacros – Beispielcode

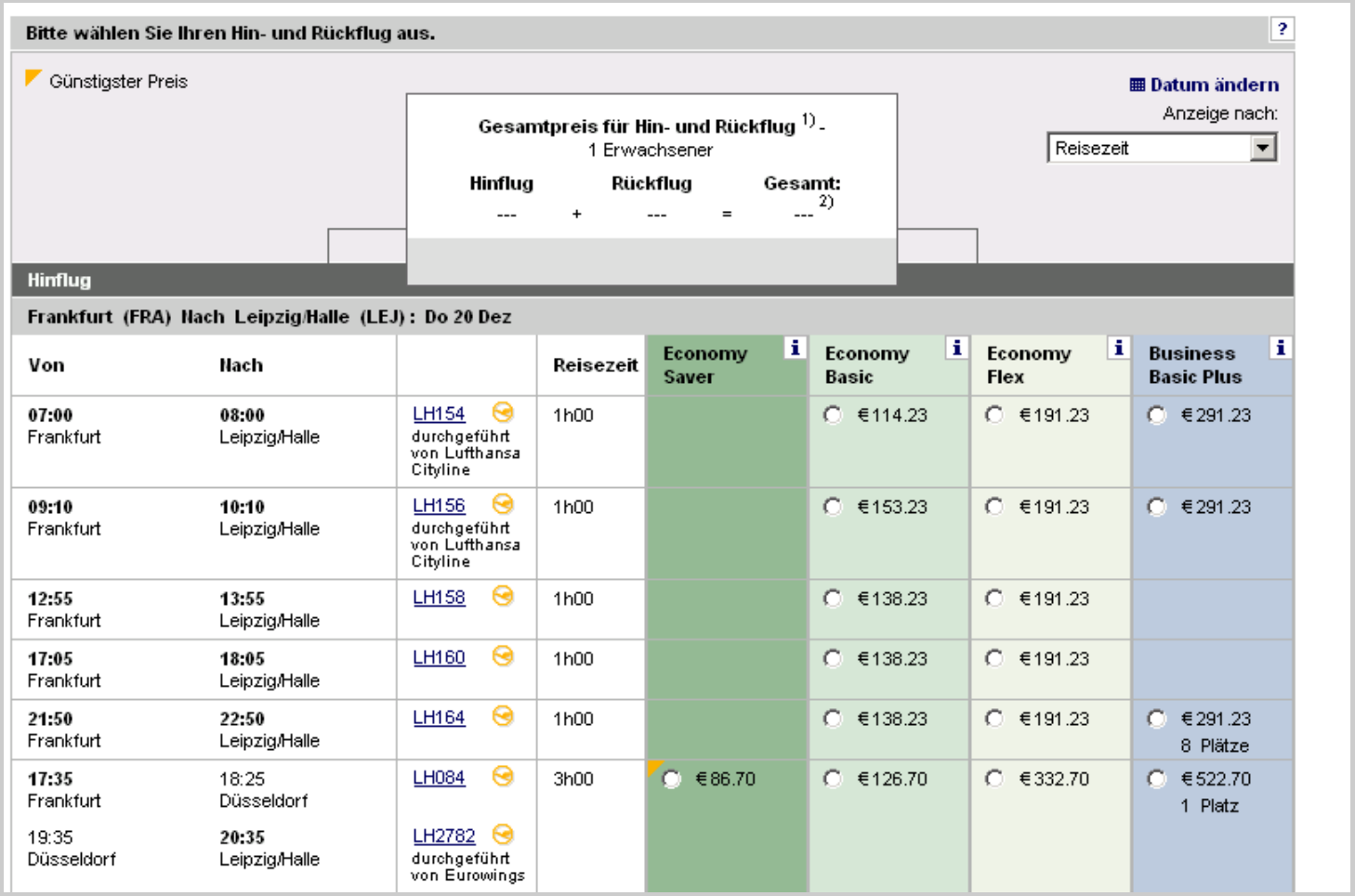
**Aufgezeichnet:**

```
URL GOTO=http://www.lufthansa.com/
TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
ATTR=ID:fld20 CONTENT=fra
TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
ATTR=ID:fld22 CONTENT=lej
TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
ATTR=ID:fdepdateDisplay CONTENT=Do,<SP>20.12.2012
```

**Bearbeitet:**

```
URL GOTO=http://www.lufthansa.com/
TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
ATTR=ID:fld20 CONTENT={{AIRP1}}
TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
ATTR=ID:fld22 CONTENT={{AIRP2}}
TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:simpleSearch_form_name
ATTR=ID:fdepdateDisplay ={{DATE1}}
```

# iMacros – eine Auswahl treffen



Bitte wählen Sie Ihren Hin- und Rückflug aus.

Günstigster Preis

Datum ändern

Anzeige nach: Reisezeit

Gesamtpreis für Hin- und Rückflug 1) -
1 Erwachsener

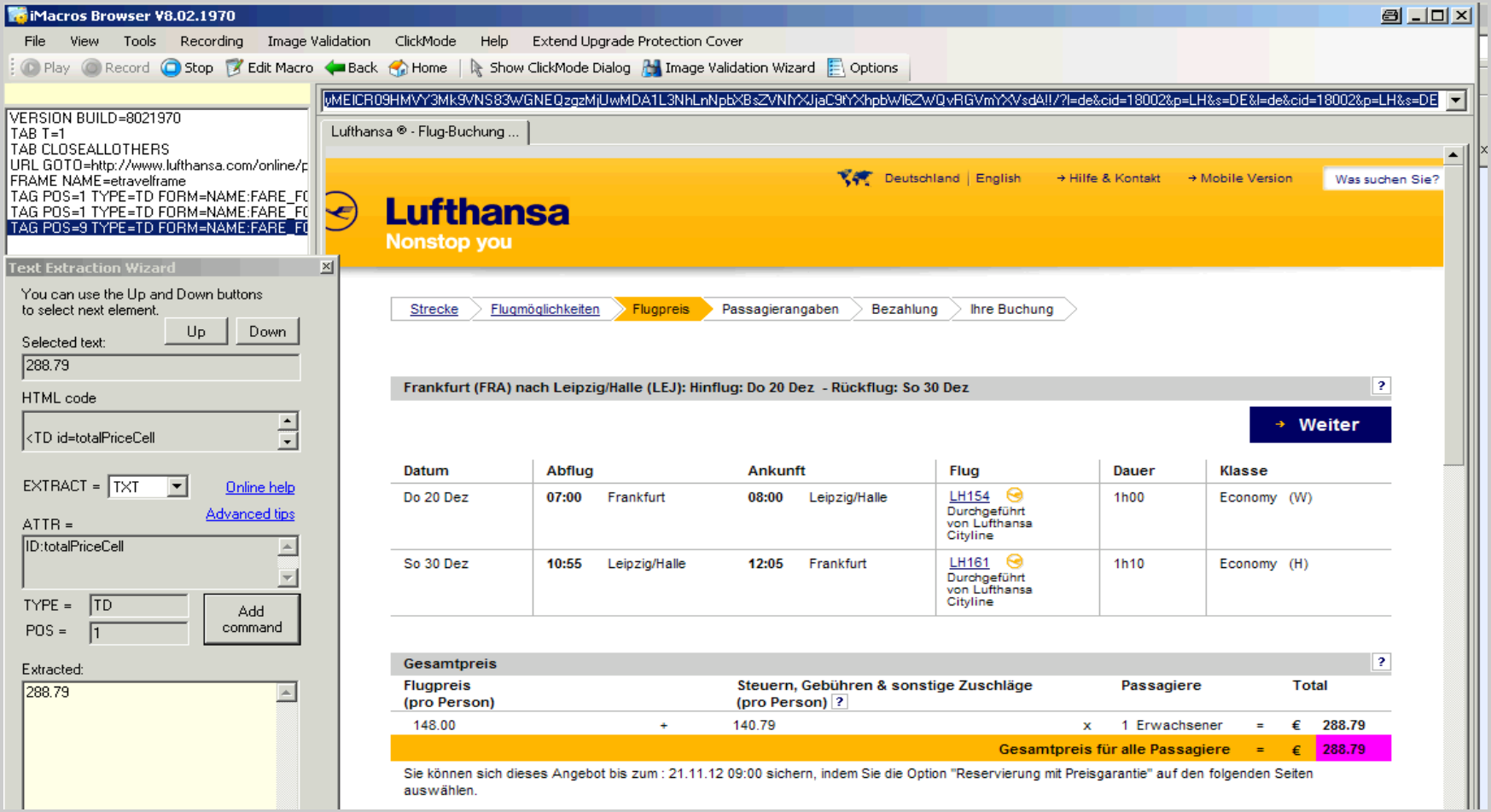
| Hinflug | | Rückflug | | Gesamt: 2) |
|---|---|---|---|---|
| --- | + | --- | = | --- |

**Hinflug**

**Frankfurt (FRA) Nach Leipzig/Halle (LEJ): Do 20 Dez**

| Von | Nach | | Reisezeit | Economy Saver | Economy Basic | Economy Flex | Business Basic Plus |
|---|---|---|---|---|---|---|---|
| **07:00** Frankfurt | **08:00** Leipzig/Halle | LH154 durchgeführt von Lufthansa Cityline | 1h00 | | € 114.23 | € 191.23 | € 291.23 |
| **09:10** Frankfurt | **10:10** Leipzig/Halle | LH156 durchgeführt von Lufthansa Cityline | 1h00 | | € 153.23 | € 191.23 | € 291.23 |
| **12:55** Frankfurt | **13:55** Leipzig/Halle | LH158 | 1h00 | | € 138.23 | € 191.23 | |
| **17:05** Frankfurt | **18:05** Leipzig/Halle | LH160 | 1h00 | | € 138.23 | € 191.23 | |
| **21:50** Frankfurt | **22:50** Leipzig/Halle | LH164 | 1h00 | | € 138.23 | € 191.23 | € 291.23 8 Plätze |
| **17:35** Frankfurt<br>19:35 Düsseldorf | 18:25 Düsseldorf<br>**20:35** Leipzig/Halle | LH084<br>LH2782 durchgeführt von Eurowings | 3h00 | € 86.70 | € 126.70 | € 332.70 | € 522.70 1 Platz |

# iMacros – Datenextraktion

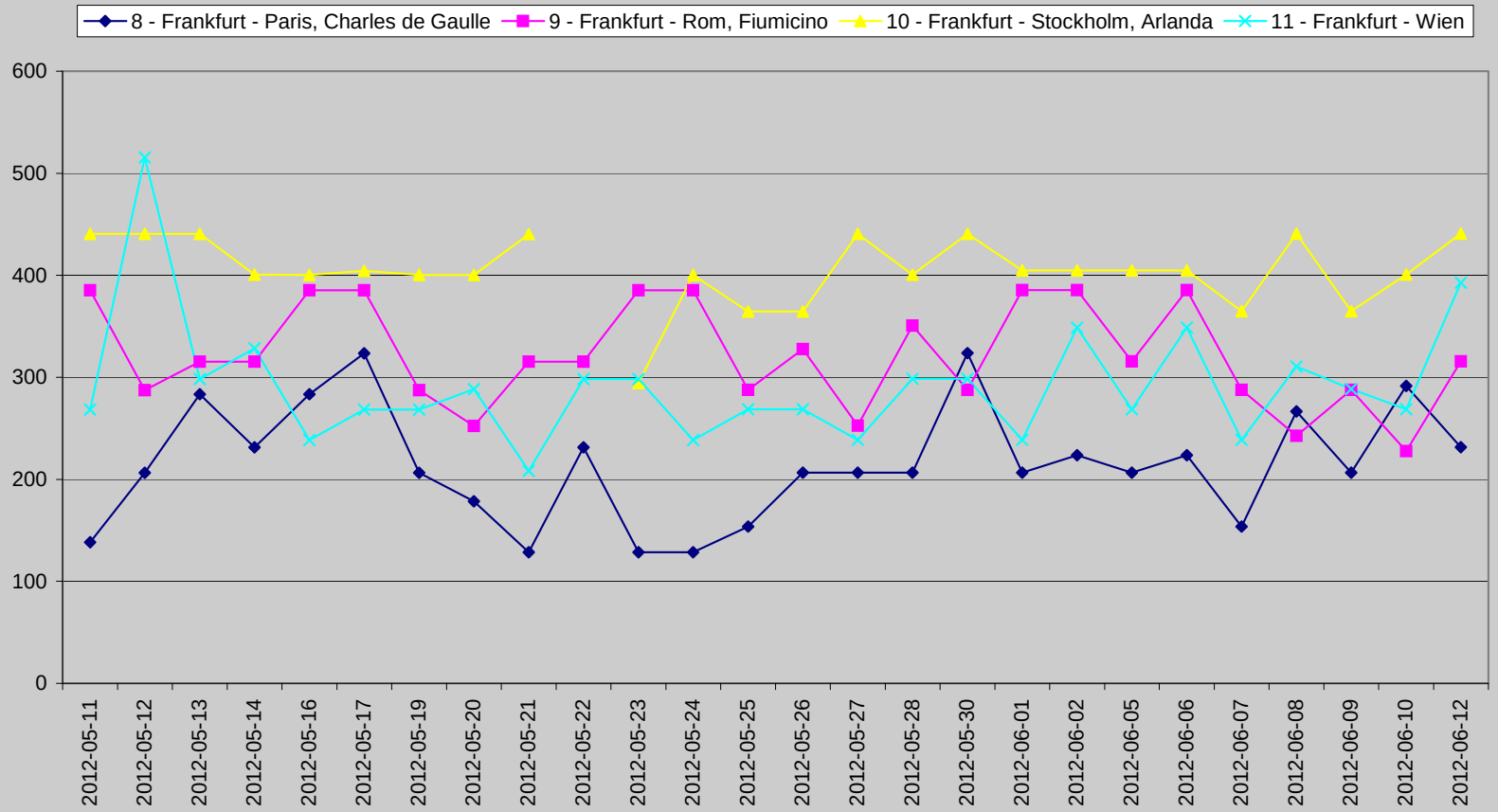


DESTATIS
wissen.nutzen.

# MySQL Datenbank

| airp_from | airp_to | date1 | date2 | airline1 | airline2 | price | i_erh_dat | Time1 | Time2 | rate |
|---|---|---|---|---|---|---|---|---|---|---|
| Frankfurt | Barcelona | Do 25 Okt | So 28 Okt | LH1128 | LH1125 | 325.91 | 26.08.2012 | 13:15 | 10:00 | Economy (S)<br>Economy (S) |
| Frankfurt | Barcelona | Do 25 Okt | So 28 Okt | LH1128 | LH1127 | 325.91 | 26.08.2012 | 13:15 | 12:30 | Economy (S)<br>Economy (S) |
| Frankfurt | Berlin, Tegel | Do 25 Okt | So 28 Okt | LH182 | LH185 | 183.42 | 26.08.2012 | 12:55 | 11:45 | Economy (W)<br>Economy (S) |
| Frankfurt | Berlin, Tegel | Do 25 Okt | So 28 Okt | LH182 | LH187 | 183.42 | 26.08.2012 | 12:55 | 12:45 | Economy (W)<br>Economy (S) |
| Frankfurt | Berlin, Tegel | Do 25 Okt | So 28 Okt | LH178 | LH185 | 207.42 | 26.08.2012 | 10:15 | 11:45 | Economy (V)<br>Economy (S) |
| Frankfurt | London, Heathrow | Do 25 Okt | So 28 Okt | LH906 | LH905 | 208.17 | 26.08.2012 | 12:45 | 11:50 | Economy (E)<br>Economy (T) |
| Frankfurt | London, Heathrow | Do 25 Okt | So 28 Okt | LH906 | LH907 | 208.17 | 26.08.2012 | 12:45 | 13:40 | Economy (E)<br>Economy (T) |
| Frankfurt | London, Heathrow | Do 25 Okt | So 28 Okt | LH908 | LH905 | 208.17 | 26.08.2012 | 13:45 | 11:50 | Economy (E)<br>Economy (T) |
| Frankfurt | Paris, Charles de Gaulle | Do 25 Okt | So 28 Okt | LH1036 | LH1029 | 286.65 | 26.08.2012 | 13:35 | 10:40 | Economy (K)<br>Economy (Q) |
| Frankfurt | Paris, Charles de Gaulle | Do 25 Okt | So 28 Okt | LH1034 | LH1029 | 291.65 | 26.08.2012 | 12:00 | 10:40 | Economy (L)<br>Economy (Q) |
| Frankfurt | Rom, Fiumicino | Do 25 Okt | So 28 Okt | LH232 | LH233 | 393.6 | 26.08.2012 | 10:45 | 13:30 | Economy (T)<br>Economy (Q) |
| Frankfurt | Rom, Fiumicino | Do 25 Okt | So 28 Okt | LH232 | LH231 | 643.6 | 26.08.2012 | 10:45 | 10:15 | Economy (T)<br>Economy (Y) |

## Niedrigster Flugpreis (Lufthansa)



# Bisherige Arbeiten

## Kaufkraftparitäten

- Flüge bei Expedia und Lufthansa
- Hotels bei HRS
- Internationale Versandhändler (Esprit, Zalando, H&M, Zara, Deichmann und Adidas) in mehreren europäische Länder

## Verbraucherpreisindex

- Versandhändler: Otto
- Online Apotheken
- Mietwagen
- Bahn
- Städtereisen

# Was ist zu beachten

## Ausgangsdaten

- **Vorgaben müssen teilweise für maschinelle Zwecke „übersetzt“ werden, z.B. Preise für Flüge um die Mittagszeit ermitteln → Flüge mit Abflugzeit von 10.00 – 14.00.**
- **Für die automatisierte Verarbeitung angelegte Daten liefern bessere Ergebnisse als Weiternutzung von Daten, die für die manuelle Erhebung verwendet werden.**

## Datenextraktion

- **Extraktion von Einzelinformationen schwierig, wenn diese nicht systematisch im HTML Code abgelegt sind.**
- **Bei Preisen gab es hier bisher keine Schwierigkeiten.**
- **Mengenangaben, Merkmale können abhängig von der Website nicht immer sauber extrahiert werden.**

# Was ist zu beachten

## Technische Probleme

- Abwehrmaßnahmen gegen Web Scraping sind möglich.
- Bei Änderungen auf den Websites muss die Software angepasst werden.

# Rechtliche Aspekte

- **Auch die Inhalte von Datenbanken sind urheberrechtlich geschützt sofern wesentliche Teile betroffen sind.**
- **Das zielgerichtete Auslesen von Preisen ist unproblematisch.**
- **AGBs die Web Scraping untersagen sind nur wirksam wenn Sie vorher ausdrücklich akzeptiert wurden.**
- **Es werden nur Daten ausgelesen die ohne die Akzeptierung von AGBs erreichbar sind.**
- **Es dürfen keine technischen Hürden umgangen werden (z.B.: CAPTCHAS)**
- **Web Scraping ist aus diesen Gründen z.B. bei Ryanair nicht möglich**