CYBERTEC
The PostgreSQL Database Company

Machine Learning mit PostgreSQL

Hans-Jürgen Schönig
www.cybertec-postgresql.com

ABOUT CYBERTEC

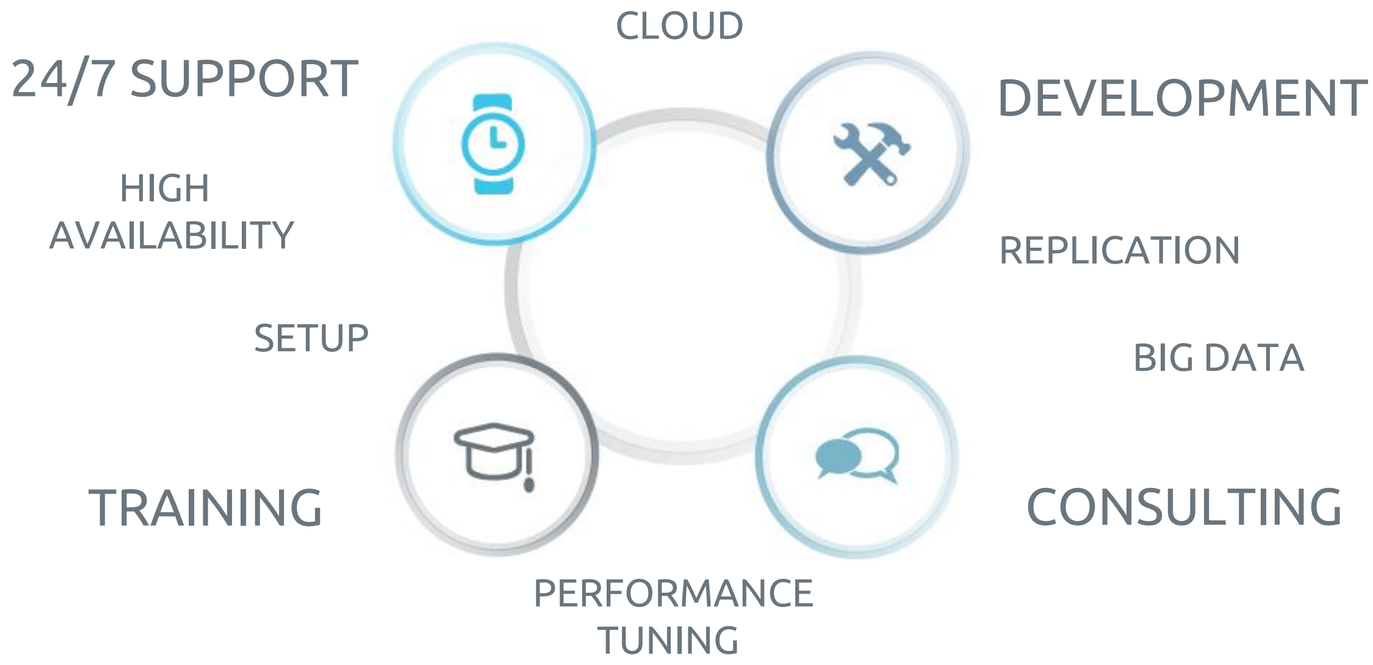
SPECIALIZED IN POSTGRESQL SERVICES

OWNED SINCE 2000

INTERNATIONAL TEAM OF DEVELOPERS

IN-HOUSE DEVELOPMENT

OUR SERVICES



CYBERTEC WORLDWIDE

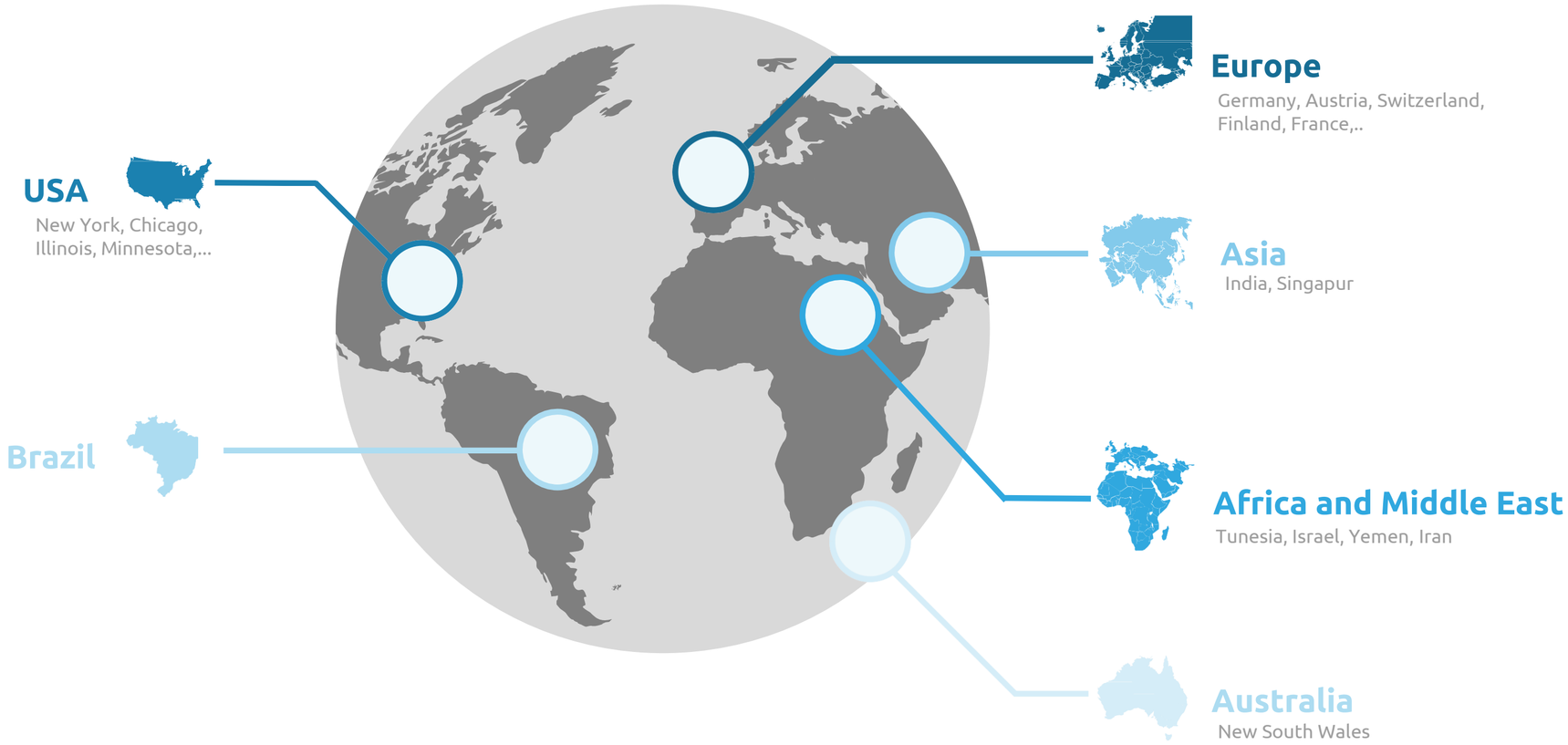
WIENER NEUSTADT,
AUSTRIA

TALLINN,
ESTONIA

ZURICH,
SWITZERLAND

MONTEVIDEO,
URUGUAY

WORLDWIDE CUSTOMERS



MACHINE LEARNING: NEUE WELTEN

- Machine Learning ist fast überall
- McKinsey sagt:
  - 8 - 12 Milliarden USD „external investments“ alleine 2016
  - Summen werden noch massiv steigen

WARUM MACHINE LEARNING UND POSTGRESQL

- AI Algorithmen sind vorhanden und verfügbar
- Die große Herausforderung: Daten, Daten, Daten, Daten, …
  - Speicherung
  - Aufbereitung
  - Transformation
  - Sampling
- Daten sind DER Rohstoff

DATENAUFBEREITUNG

- Oft wird relativ viel „programmiert“
  - Mit Aufwand verbunden
  - Mühsam, wenn man noch nicht sicher ist, was man benötigen wird
  - Kostet massiv Zeit
- Idee: SQL ist zur Aufbereitung von Daten ziemlich gut
  - Wieso also AI nicht in PostgreSQL?

WIE MAN DATEN AUFBEREITET

- Für Machine Learning teilt man Daten in ein
  - Trainingsset (um das Modell zu trainieren)
  - Verification Set
  - Test Set
- Oft wird auch nur in zwei Teile gesplittet
- Die Aufbereitung via SQL ist einfach

SAMPLING: TRAINING VS. VERIFICATION

> PostgreSQL erlaubt Sampling:

```
test=# SELECT *
     FROM t_data
          TABLESAMPLE BERNOULLI (20)
          REPEATABLE (4712);
 id
----
  1
  9
 18
 19
(4 rows)
```

SOPHISTICATED SAMPLING

- PostgreSQL ermöglich customized Sampling
  - Man kann seine eigene Sampling Methode hinterlegen
  - Beispielsweise zur „Stratification“ praktisch
- Meistens reichen aber die eingebauten Methoden

SAMPLING: WEITERE EINFACHE METHODEN (1)

```
test=# SELECT setseed(0.4711);
  setseed
 --------

(1 row)
```

SAMPLING: WEITERE EINFACHE METHODEN (1)

```
test=# SELECT CASE WHEN random() < 0.2
        THEN 'training'
        ELSE 'test' END, *
FROM t_data LIMIT 5;
    case    | id
 ----------+----
test        | 1
training    | 2
training    | 3
test        | 4
test        | 5
(5 rows)
```

DATEN NORMALISIEREN (1)

- Viele Modelle wollen „normalisierte“ Daten
  - Bedeutet: Der Wertebereich sollte zwischen 0 und 1
  - Mit SQL einfach zu bewerkstelligen

DATEN NORMALISIEREN (2)

```
SELECT id,
    id::numeric / max(id)
     OVER () AS normalized
    FROM    t_data TABLESAMPLE BERNOULLI (20)
                REPEATABLE (4712);
 id |   normalized
----+------------------------
  1 | 0.052631578947368421O5
  9 | 0.473684210526315789 47
 18 | 0.947368421052631578 95
 19 | 1.000000000000000000 00
(4 rows)
```

MODELLE IMPLEMENTIEREN

- Fertige Modelle sind für viele verschiedene Sprachen verfügbar
- PL/Python und PL/R sind die beste Methode
- R und Python stellen viele Machine Learning Libraries zur Verfügung

EIN BEISPIEL

- Fehlerhafte Logins erkennen
- Wo gibt es Anomalien?
  - Ohne zu wissen, was „normal“ bedeutet
  - „Normal“ ist vom Kontext abhängig
- Lösungsansatz: Lineare Separierung mittels SVM (Support Vector Machine)

TABELLEN ERZEUGEN

```
CREATE SCHEMA anomaly;

CREATE TABLE anomaly.login_event
(
    login_date      timestamp,
    location        numeric,
    user_agent      numeric,
    login_user      int8
);
```

DEMO TIME ☺

SENTIMENT: WAS IST DAS?

> Wie ist die aktuelle Stimmungslage?

> Wie stehen User zu einem bestimmten Thema?

CODE BEISPIEL

```
SELECT sentiment('we love this very much, excellent');
sentiment
-----------------------------------------
Sentiment(polarity=0.75, subjectivity=0.8)
(1 row)
```

EINE EINFACHE METHODE

```
CREATE FUNCTION sentiment(story text)
RETURNS text
AS
$$
    from textblob import TextBlob
    testimonial = TextBlob(story)
    return testimonial.sentiment
$$ LANGUAGE 'plpython2u';
```

MACHINE LEARNING IN POSTGRESQL

- Sampling und Teilen der Daten ist super einfach
- Verschiedene AI Modelle schnell integrieren
- Daten flexibel zusammen stellen
- Rapid prototyping ist einfach

Hans-Jürgen
SCHÖNIG
CEO

hs@cybertec.at

+43 2622 930 22-2