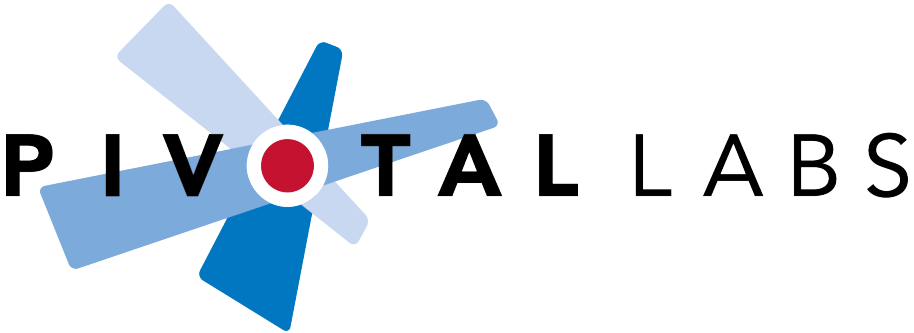
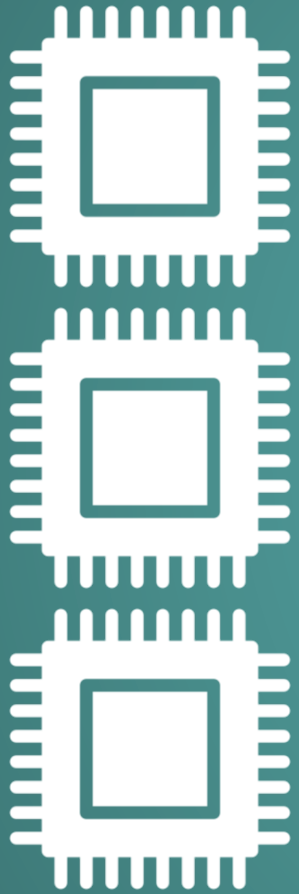
Bring Your Code to Your Data

Ian Huston
@ianhuston

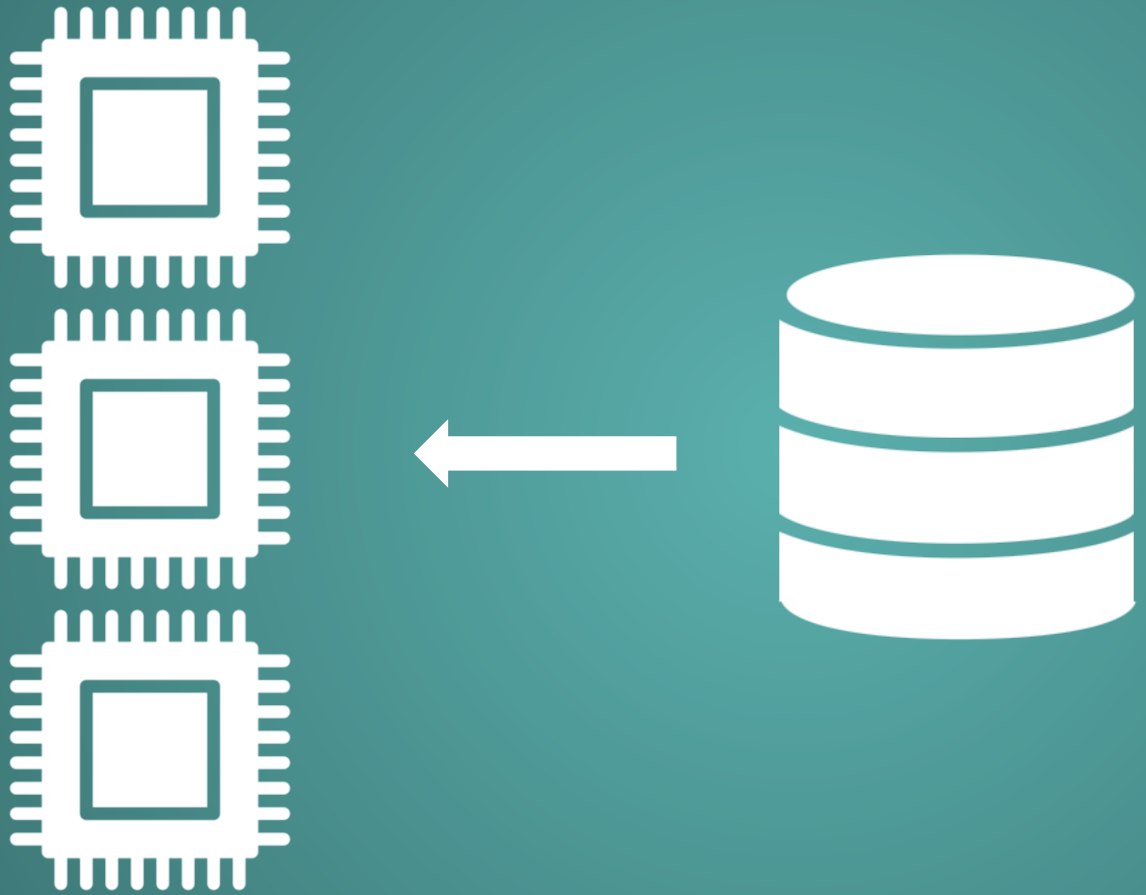




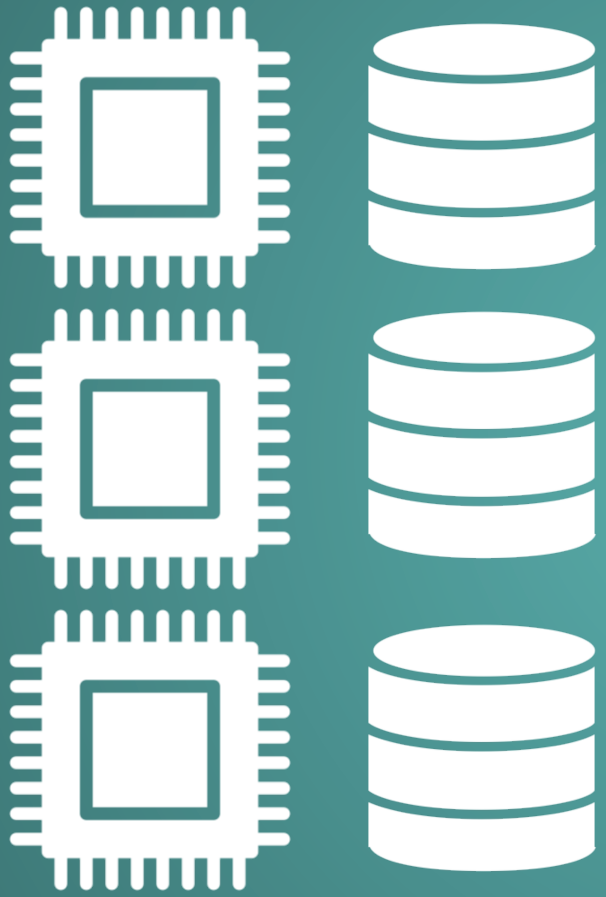
High Performance Computing

HPC ≠ BIG DATA

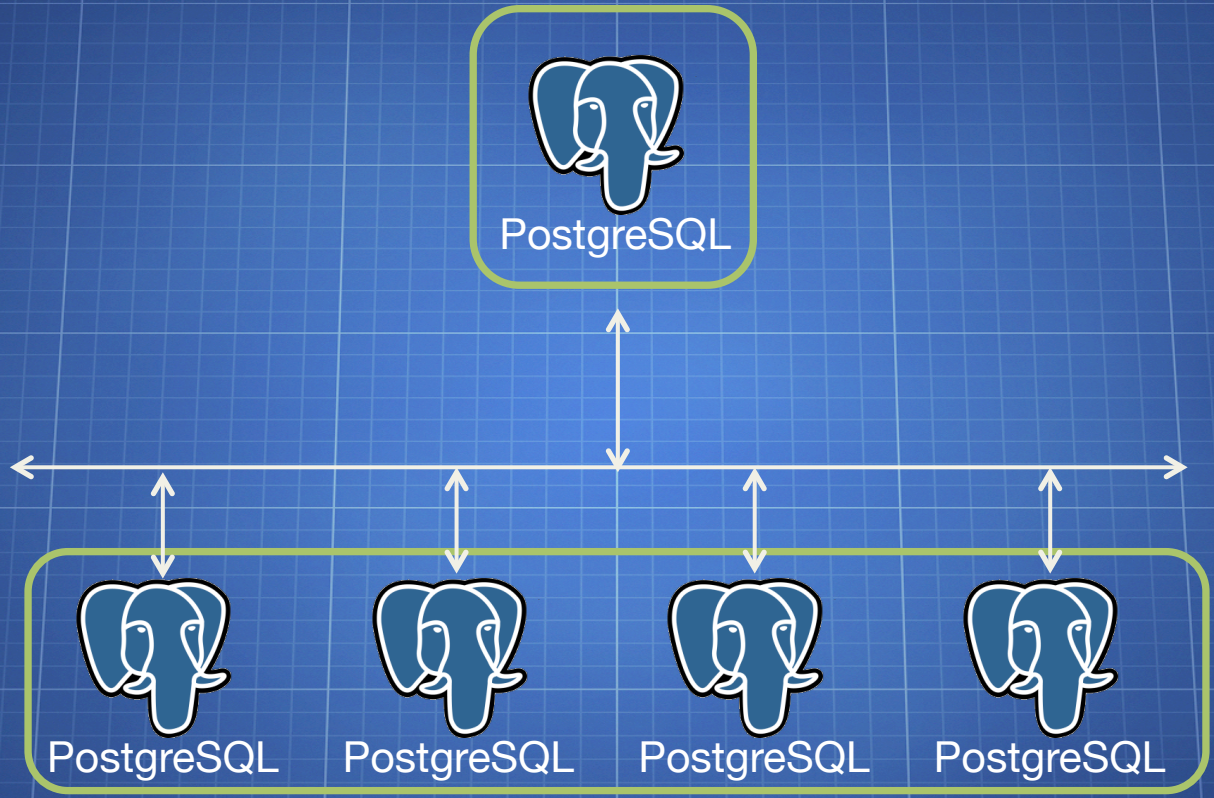
Pivotal



In-Database Computing



In-Database Distributed Computing


PostgreSQL
PostgreSQL
PostgreSQL
PostgreSQL
PostgreSQL

1. SQL for analytics
2. Packaged libraries
3. In-Database Python

1. SQL for analytics

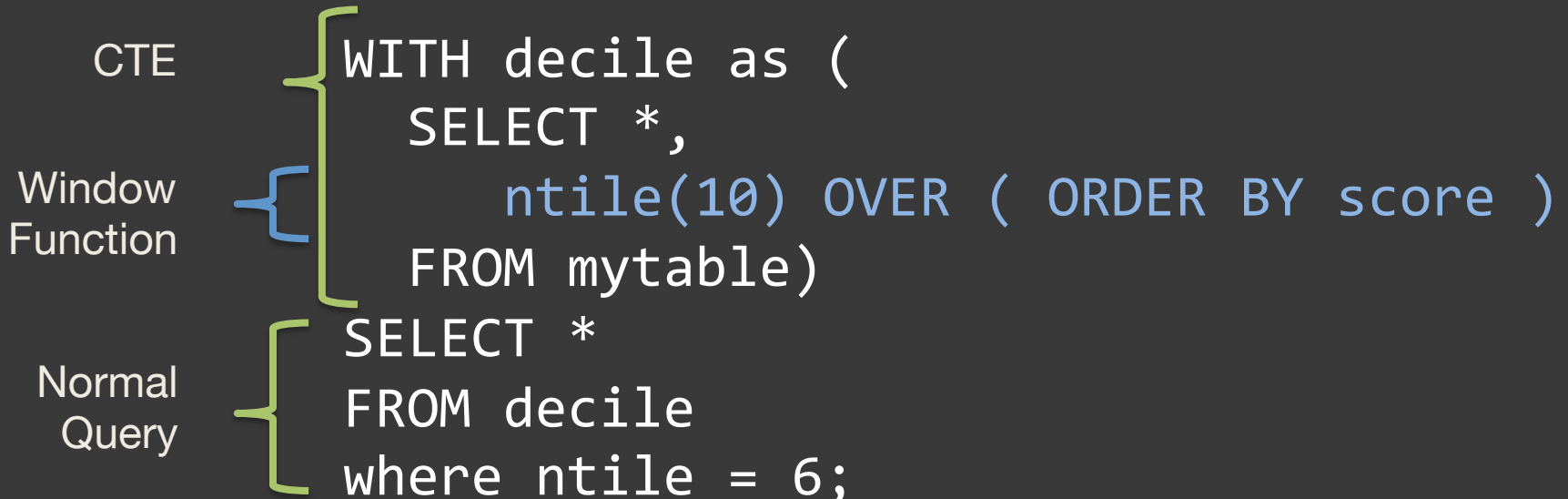
SQL is more than SELECT

- Window Functions
- WITH queries (Common Table Expressions)
- User Defined Aggregations
- User Defined (SQL) Functions

Examples:

- Time Series using Windowing: http://blog.pivotal.io/author/caleb-welton
- Heroku's Postgres bits: http://postgres-bits.herokuapp.com

Get rows in 6th decile of scores



http://postgres-bits.herokuapp.com/#60

User Defined (SQL) Functions

Declaration

SQL Statements

Language

```
CREATE FUNCTION times2(INT)
RETURNS INT
AS $$
    SELECT 2 * $1
$$
LANGUAGE sql;
```

Execution

```
SELECT times2(1);
 times2
--------
      2
```

2. Packaged libraries

IN-DATABASE
MACHINE LEARNING

http://madlib.net

Pivotal

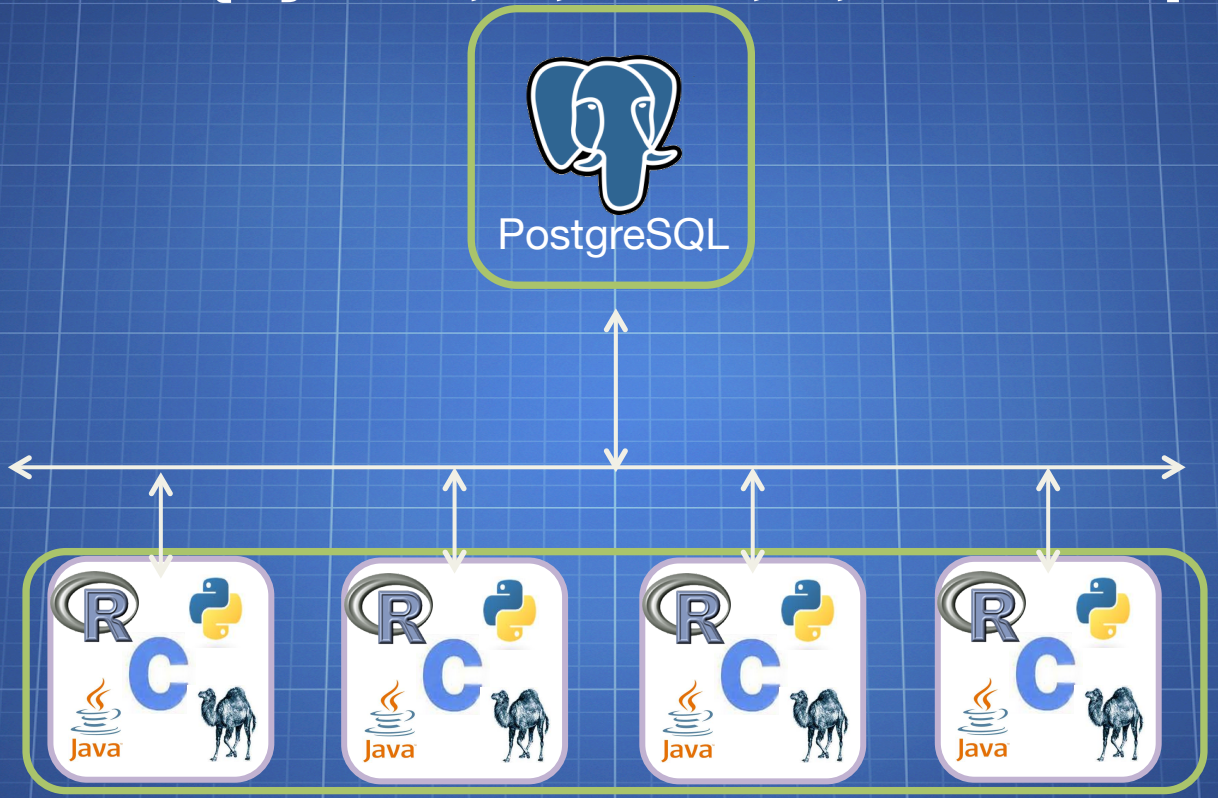
IN-DATABASE
GEOGRAPHIC QUERIES

http://postgis.net

Pivotal

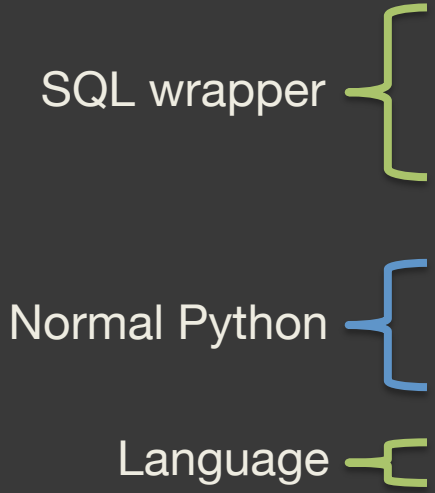
3. In-Database Python (and R, Java, C, etc)

PL/X : X in {Python, R, Java, C, JavaScript, etc.}



Data Parallelism

- Little or no effort is required to break up the problem into a number of parallel tasks, and there exists no dependency (or communication) between those parallel tasks.
- Examples:
  - Measure the height of each student in a classroom (explicitly parallelizable by student)
  - MapReduce
  - map() function in Python




```
CREATE FUNCTION
  pymax (a integer, b integer)
RETURNS integer
AS $$
  if a > b:
    return a
  return b
$$ LANGUAGE plpythonu;
```

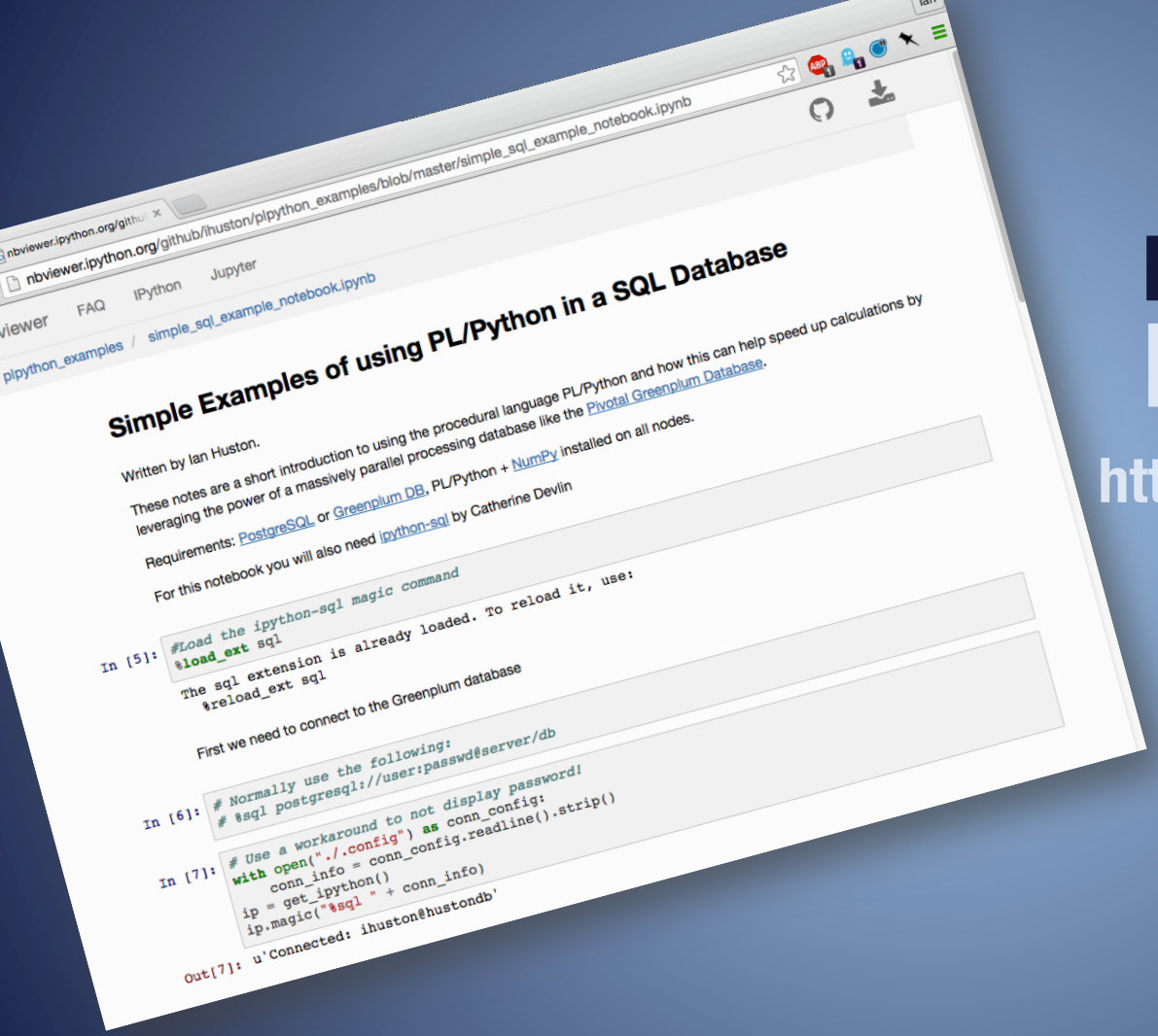
BENEFITS:
Reuse Python & R code
Access Python & R libraries
Implicit parallelism

Financial Examples

Insurance Risk Analysis

Stress Testing

Asset Management and Churn



MORE DETAILS

http://tinyurl.com/ih-plpython

Pivotal

@ianhuston

Pivotal