Databases, SQL, and Pandas

cs109, Fall 2015 (#cs109)

Rahul Dave

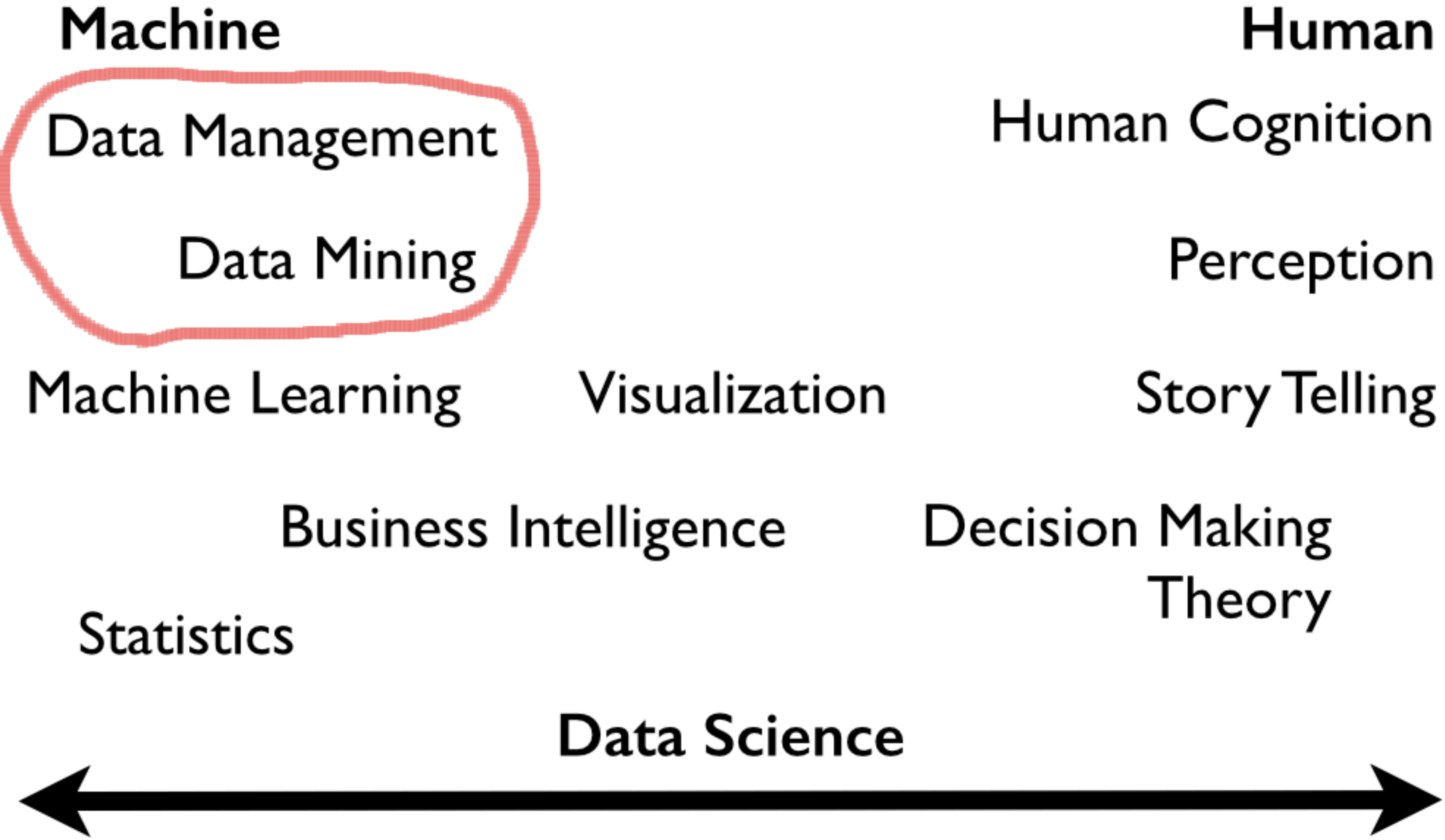
`rahuldave@gmail.com`, @rahuldave, `staff@cs109.org`

ANNOUNCEMENTS

Class in in Science Center B starting THIS thursday, 17th Sep, 2015!

It took about three years before the BellKor's Pragmatic Chaos team managed to win the prize ... The winning algorithm was ... so complex that it was never implemented by Netflix. [1]

[1] https://hbr.org/2012/10/big-data-hype-and-reality



Machine

Human

Data Management

Human Cognition

Data Mining

Perception

Machine Learning

Visualization

Story Telling

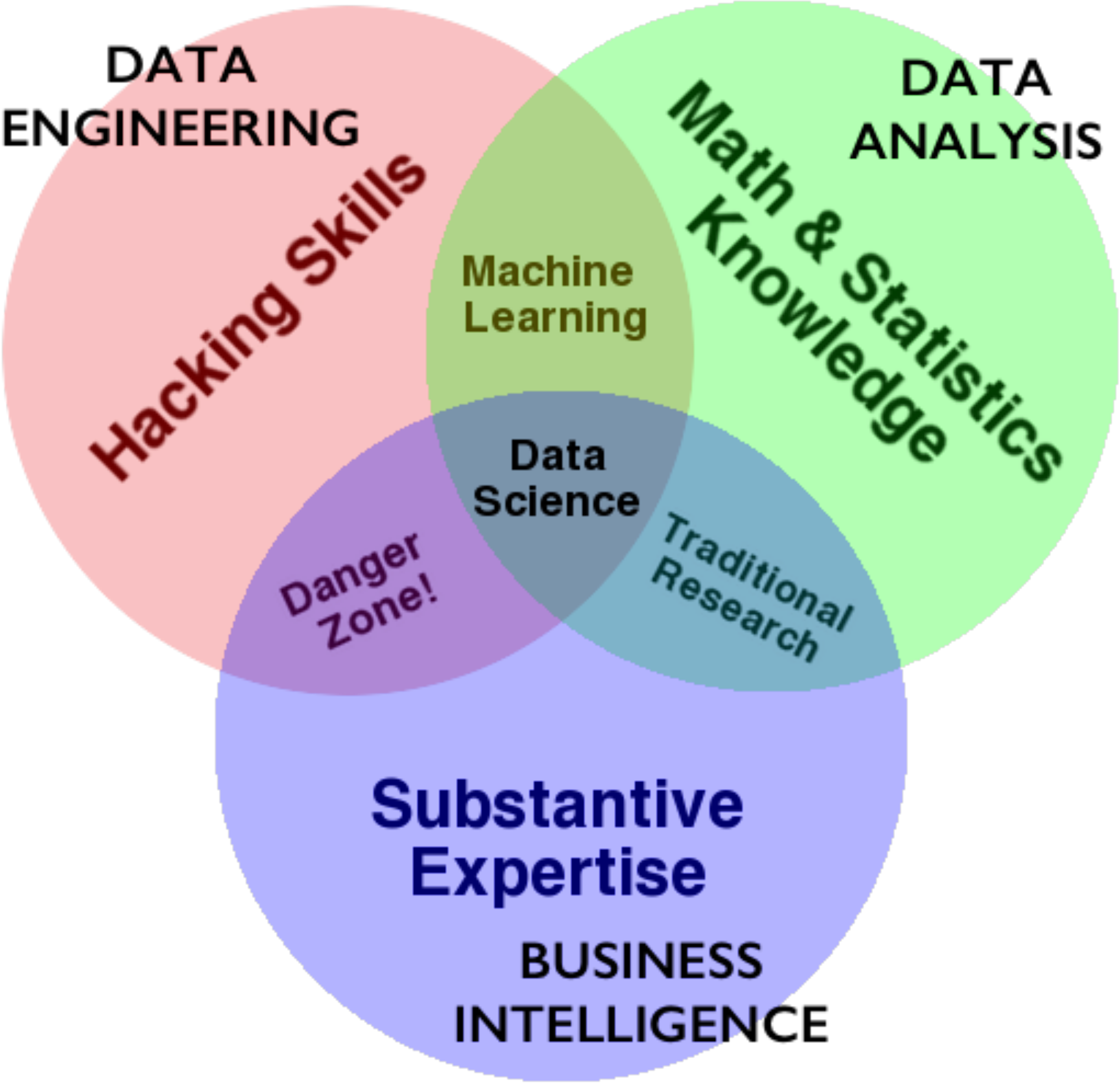
Business Intelligence

Decision Making Theory

Statistics

Data Science

Inspired by Daniel Keim, "Visual Analytics: Definition, Process, and Challenges"


DATA ENGINEERING
Hacking Skills
Machine Learning
Math & Statistics Knowledge
DATA ANALYSIS
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
BUSINESS INTELLIGENCE

Data Scientist: Sexiest Job of the 21st Century

It’s important that our data team wasn’t comprised solely of mathematicians
and other “data people.” It’s a fully integrated product group that includes
people working in design, web development, engineering, product
marketing,
and operations. They all understand and work with data, and I consider them
all data scientists... Often, an engineer can have the insight that makes it
>clear how the product’s design should work, or vice-versa — a designer can
have
the insight that helps the engineers understand how to better use the data.
Or
it may take someone from marketing to understand what a customer really
wants to accomplish.[2]

[2] D. J. Patil, U.S. Chief Data Scientist, Building data science teams. " O'Reilly Media, Inc.", 2011.

DATA ENGINEERING

- **compute**: code, python, R, julia, spark, hadoop
- **storage/database**: git, SQL, NoSQL, HBase, disk, memory
- **devops**: AWS, docker, mesos, repeatability
- **product**: database, web, API, viz, UI, story

Different at different scales....

What kind of data storage do you need?

- **memory**
- **disk**: what if we do not fit?
- **cluster**: what if we still do not fit?
- **cluster**: what if we need/can use parts?
- What if we MUST bring compute to disk?

What kind of data access do you need?

relational most used: use as default for more help

Saves file to disk, fast

- **relational**: pandas, SQL: Postgres, sqlite, Hbase, VoltDB

Stores file after file in jason

- **document oriented**: MongoDB, CouchDB
- **key-value**: Riak, Redis, Memcached

Super fast: give them a key, give a result

Store networks -> faster to find neighbors in a graph

- **graph oriented**: Neo4J

Today we'll focus on relational

- What is a relational Database?
- What Grammar of Data does it follow?

  What a database system can do: so you can find same actions in other database
- How is this grammar implemented in Pandas?
- How is this grammar implemented in SQL

Relational Database

Dont say: seek 20 bytes onto disk and pick up from there. The next row is 50 bytes hence

Say: select data from a set. I dont care where it is, just get the row to me.

ie, select data based on conditionals

Relational Database(contd)

- A collection of tables related to each other through common data values. common data value = keys

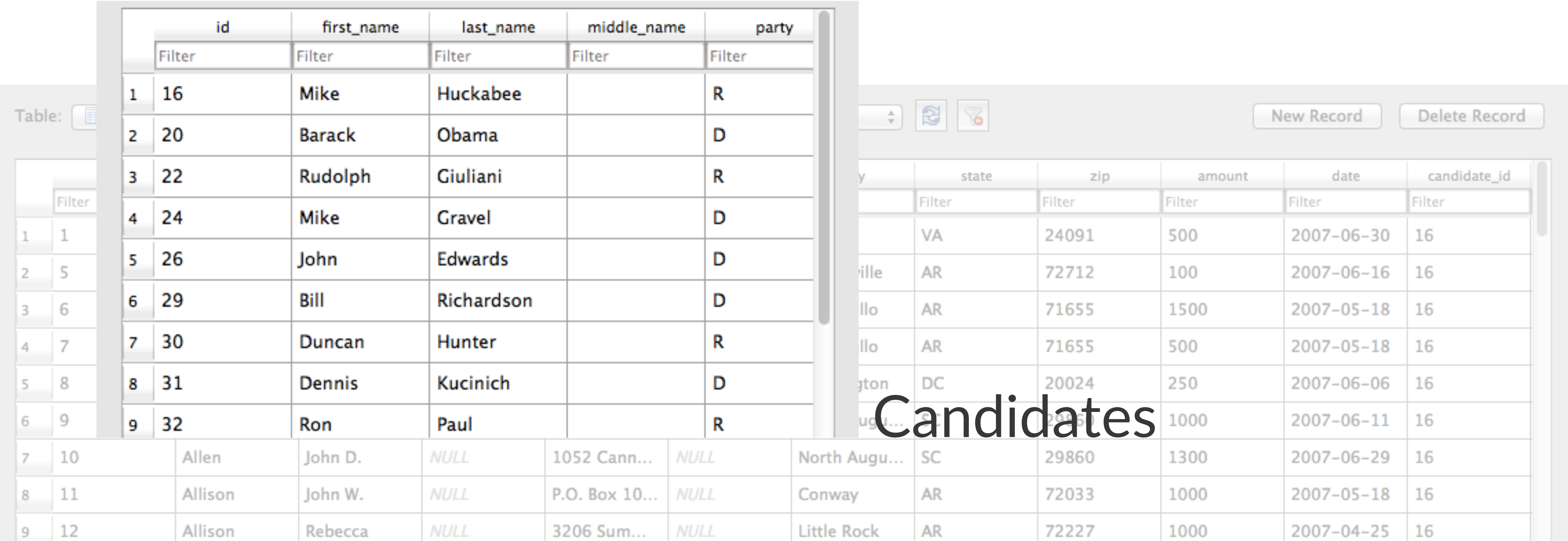
- Rows represent attributes of something

- Everything in a column is values of *one* attributes

- A cell is expected to be atomic atomic -> 1 value (not a list)

- Tables are related to each other if they have columns called keys which represent the same values

Contributors



	id	first_name	last_name	middle_name	party
	Filter	Filter	Filter	Filter	Filter
1	16	Mike	Huckabee		R
2	20	Barack	Obama		D
3	22	Rudolph	Giuliani		R
4	24	Mike	Gravel		D
5	26	John	Edwards		D
6	29	Bill	Richardson		D
7	30	Duncan	Hunter		R
8	31	Dennis	Kucinich		D
9	32	Ron	Paul		R

							y	state	zip	amount	date	candidate_id
	Filter							Filter	Filter	Filter	Filter	Filter
1	1							VA	24091	500	2007-06-30	16
2	5						ille	AR	72712	100	2007-06-16	16
3	6						llo	AR	71655	1500	2007-05-18	16
4	7						llo	AR	71655	500	2007-05-18	16
5	8						gton	DC	20024	250	2007-06-06	16
6	9						ugu...	SC	29860	1000	2007-06-11	16
7	10	Allen	John D.	NULL	1052 Cann...	NULL	North Augu...	SC	29860	1300	2007-06-29	16
8	11	Allison	John W.	NULL	P.O. Box 10...	NULL	Conway	AR	72033	1000	2007-05-18	16
9	12	Allison	Rebecca	NULL	3206 Sum...	NULL	Little Rock	AR	72227	1000	2007-04-25	16

Candidates

Two tables connected by 1 column

Connected by id column -> our keys.
We want keys that are unique

Scales of Measurement

- Quantitative (Interval and Ratio)

Integers or Flaots

- Ordinal

Flexible: characters, integer, etc
My choice

- Nominal[3]

Test-type or character type
A string

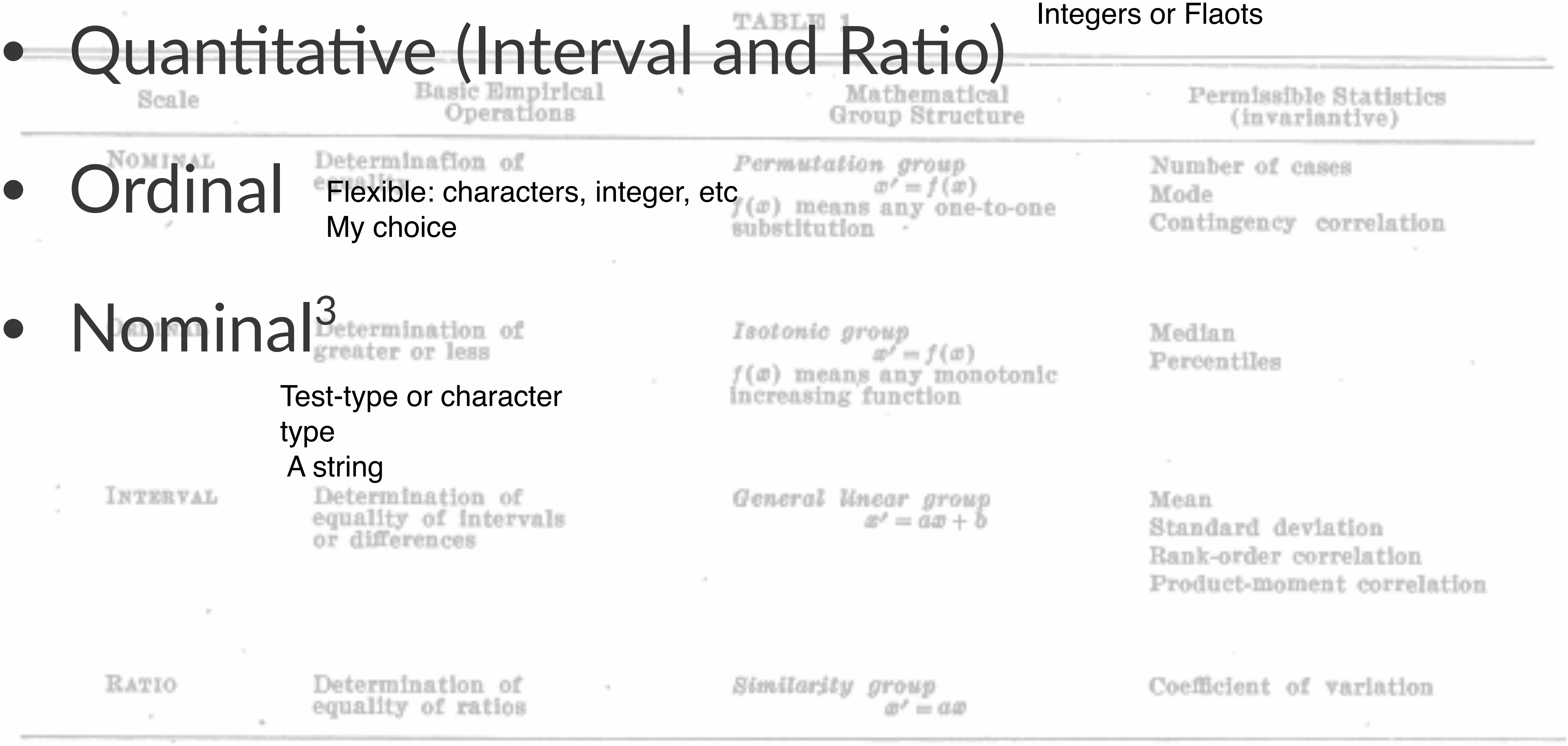


TABLE 1

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

[3] S. S. Stevens, Science, New Series, Vol. 103, No. 2684 (Jun. 7, 1946), pp. 677-680

Grammar of Data

Been there for a while (SQL, Pandas), formalized in `dplyr`[4].

- provide simple verbs for simple things. These are functions corresponding to common data manipulation tasks

Select column, row, insert, delete

- second idea is that backend does not matter. Here we constrain ourselves to Pandas and sqlite
- multiple backends implemented in Pandas, Spark, Impala, Pig, dplyr, ibis, blaze

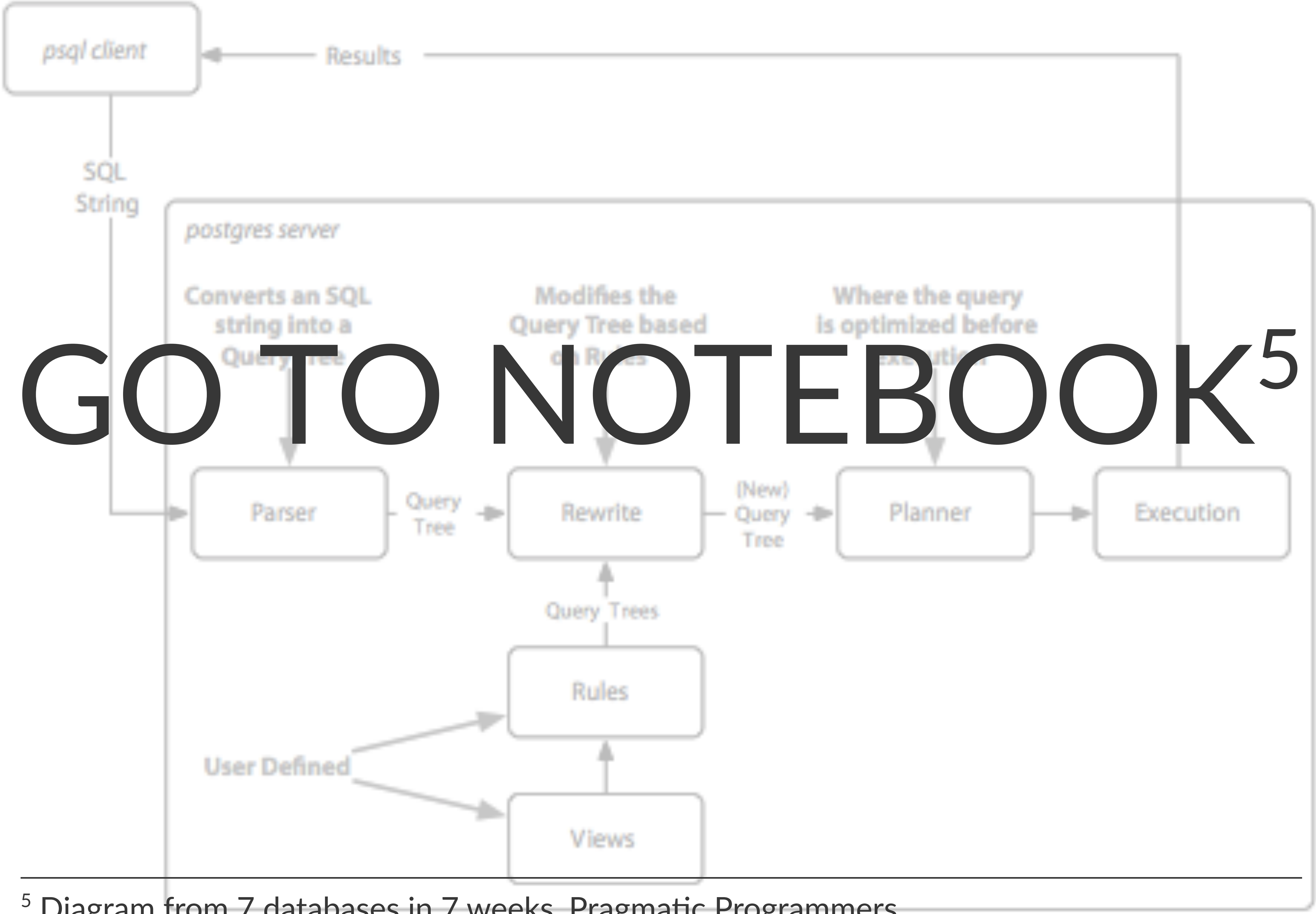
[4] Hadley Wickham: https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html

Why bother

- learn hot to do core data manipulations, no matter what the system
- relational databases critical for mon-memory fits. Big installed base.

  Too much data -> does not fit on local machine.
- one off questions: google, stack-overflow, http://chrisalbon.com

GO TO NOTEBOOK[5]



[5] Diagram from 7 databases in 7 weeks, Pragmatic Programmers

RDBMS when:

- data structure regularity is known
- transactions are required
- benefit from years of tuning
- not good for deep hierarchy
- which kind depends on use case: pandas, hbase, columnar, postgres,…

FIN