Lesson Plan 2/22

Tuesday, February 20, 2018 5:15 AM

Admin
- Midterm I next Tuesday
- Questions, clarifications on HW2

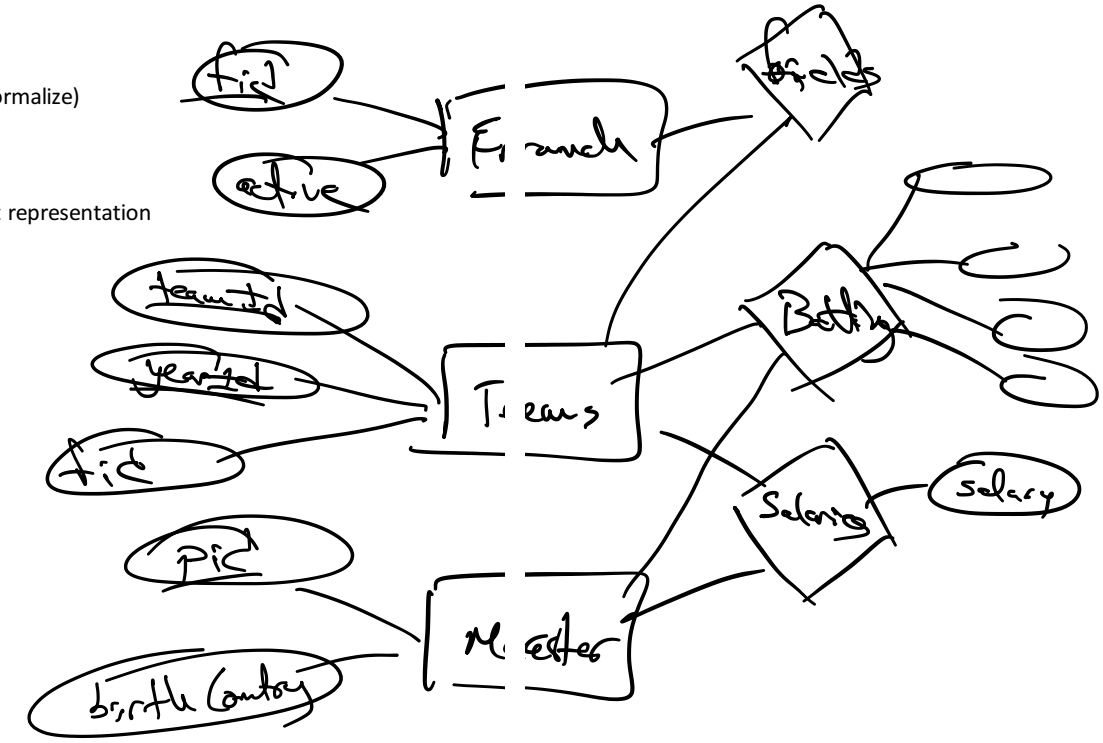
HDSC
- Of topical interest: example paper from Facebook modeling behavior to spot illegitimate use: https://research.fb.com/publications/copycatch-stopping-group-attacks-by-spotting-lockstep-behavior-in-social-networks/

Common Tidying Operations
- Data values as headers (gather)
- More than one entity in a table (normalize)

Cleaning Text
- Regular expressions basics
- Regular expression tools
- The _one_term_per_row_ tidy text representation

Midterm I discussion

Operations exercise



Tidy Data

- One row per specific entity / relationship
- One column per attribute
- One table: per type of entity / relationship

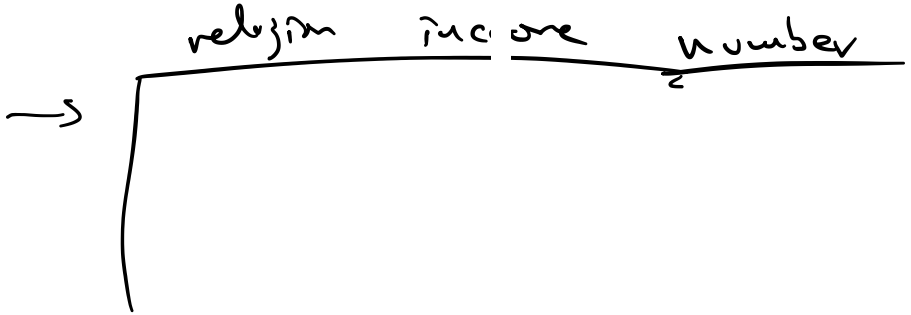
Problems: 1) Headers as values
2) Multiple entity problem

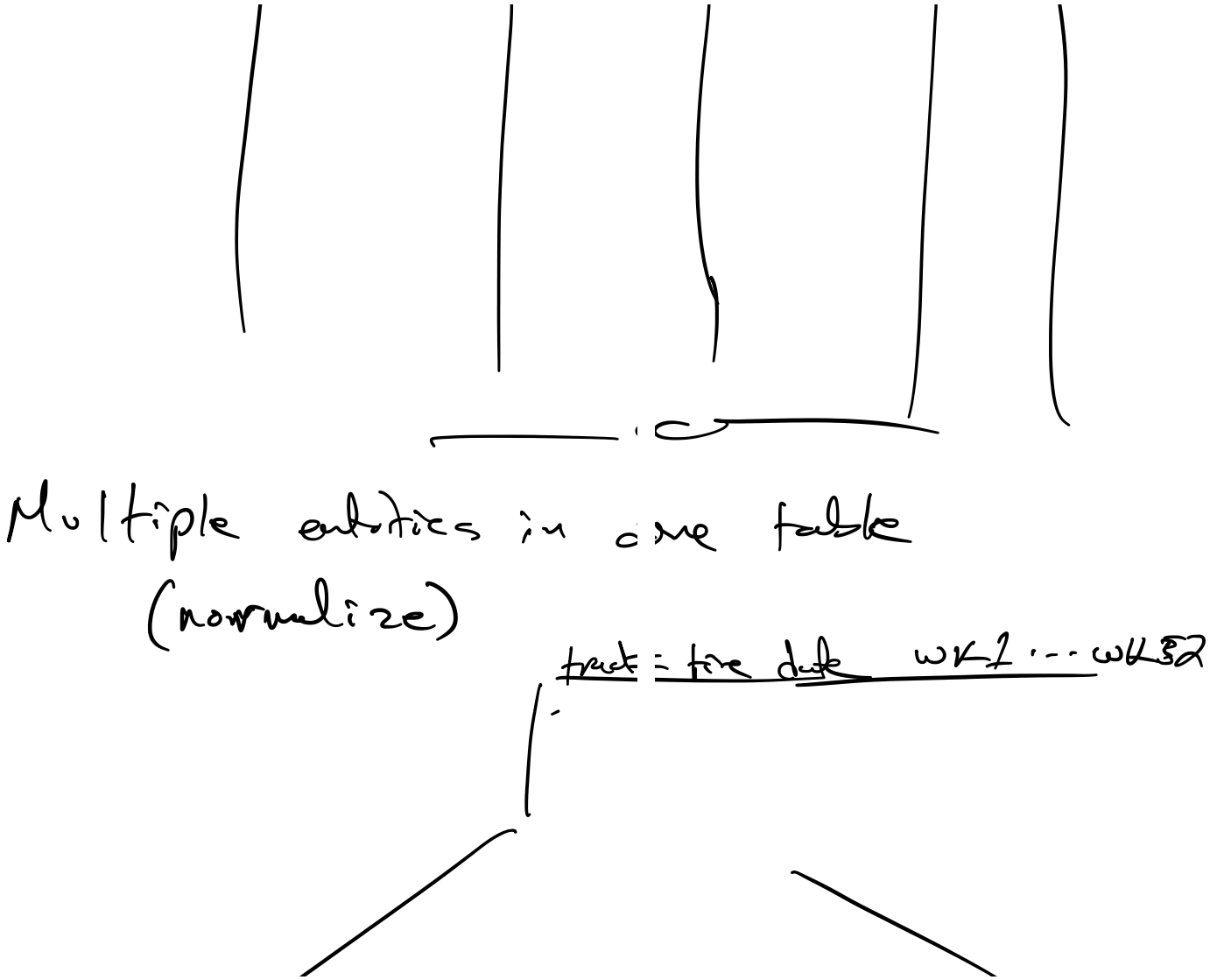
Pew:

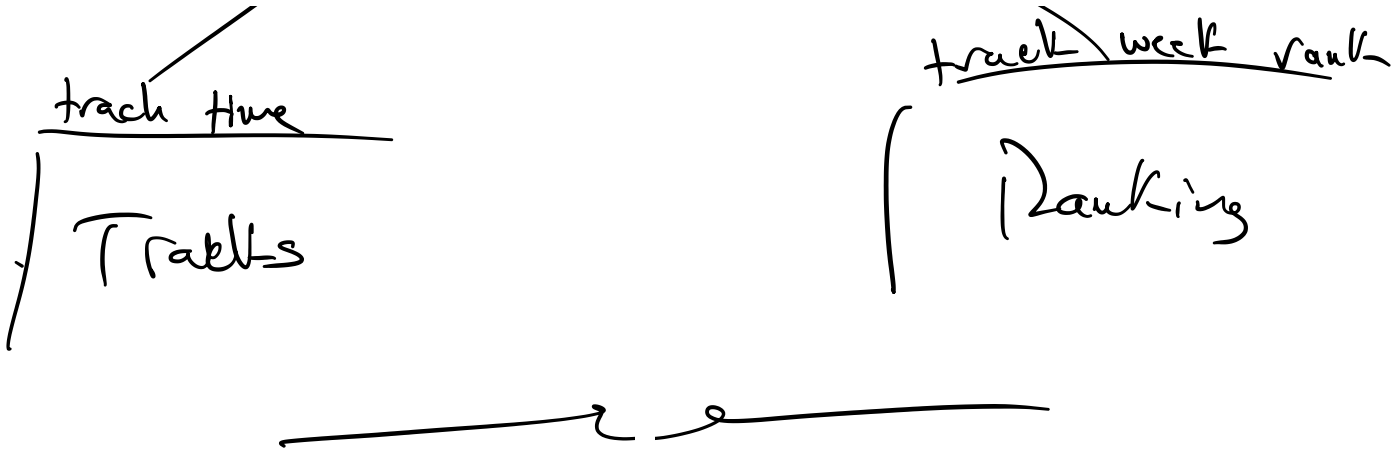
entities: group of people of specific religion & income

attr: religion
income
number

"gather": function performs operation




Multiple entities in one table
(normalize)
track = fire data wk1 ... wk52

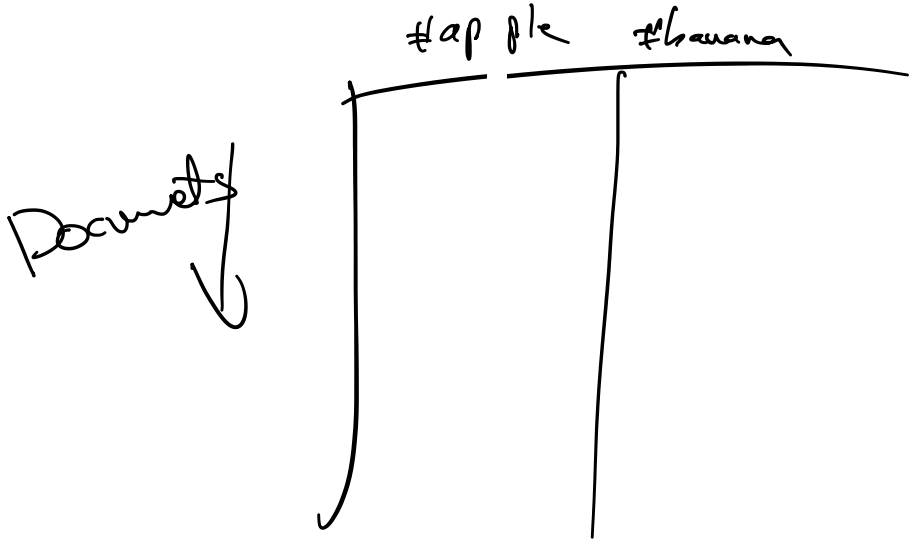


Clearing Text

- Regular Expressions
- Common operations based on RE
- One-token-per row representation

Text:

- Document
- Terms



- one-term-per-row

term	document