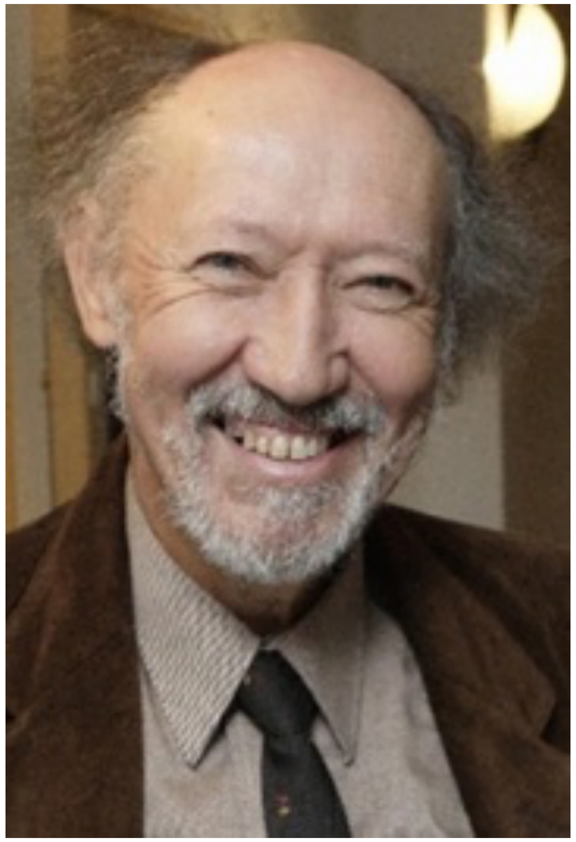
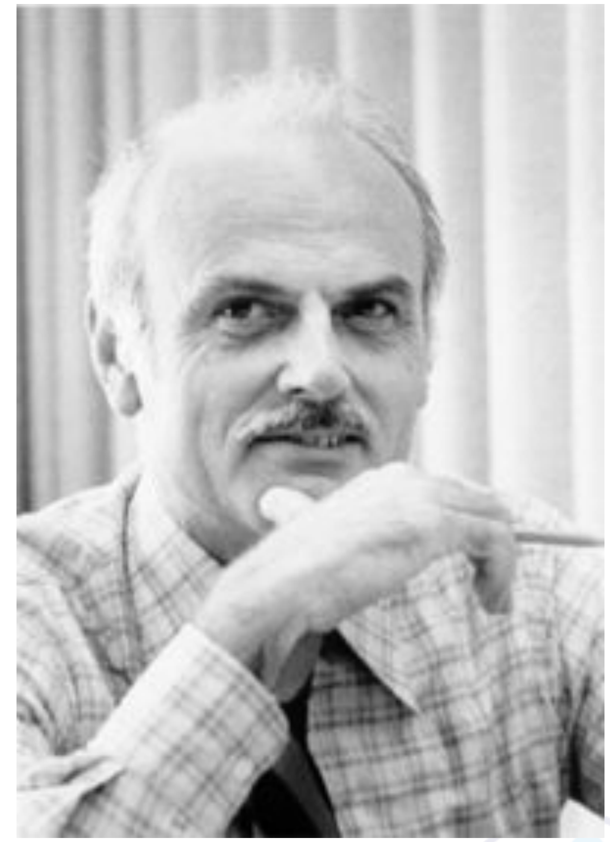
Relational Databases

Lecturer: Азат Якупов (Azat Yakupov)
https://datalaboratory.one



Christopher J.Date



Edgar Frank Codd

Ronald Fagin

Jeffrey David Ullman

“Excel” Data Sample

	A	B	C	D	E	F
1	**ID**	**Name**	**College**	**CollegeCity**	**Hobby**	**CollegeNew**
2	**123**	**Petr**	**KFU**	**Kazan**	**football**	**MSU**
3	123	Petr	**KAI**	Kazan	**tennis**	MSU
4	123	Petr	KFU	Kazan	tennis	MSU
5	123	Petr	KAI	Kazan	football	MSU

	A	B	C	D	E	F
1	**ID**	**Name**	**College**	**CollegeCity**	**Hobby**	**CollegeNew**
2	**123**	**Petr**	**KFU**	**Kazan**	**football**	**MSU**
3	123	Petr	**KAI**	Kazan	**tennis**	MSU
4	123	Petr	KFU	Kazan	tennis	MSU
5	123	Petr	KAI	Kazan	football	MSU

data redundancy

anomaly of inserting

anomaly of deleting

anomaly of updating

	A	B	C	D	E	F
1	**ID**	**Name**	**College**	**CollegeCity**	**Hobby**	**CollegeNew**
2	**123**	**Petr**	**KFU**	**Kazan**	**football**	**MSU**
3	123	Petr	**KAI**	Kazan	**tennis**	MSU
4	123	Petr	KFU	Kazan	tennis	MSU
5	123	Petr	KAI	Kazan	football	MSU

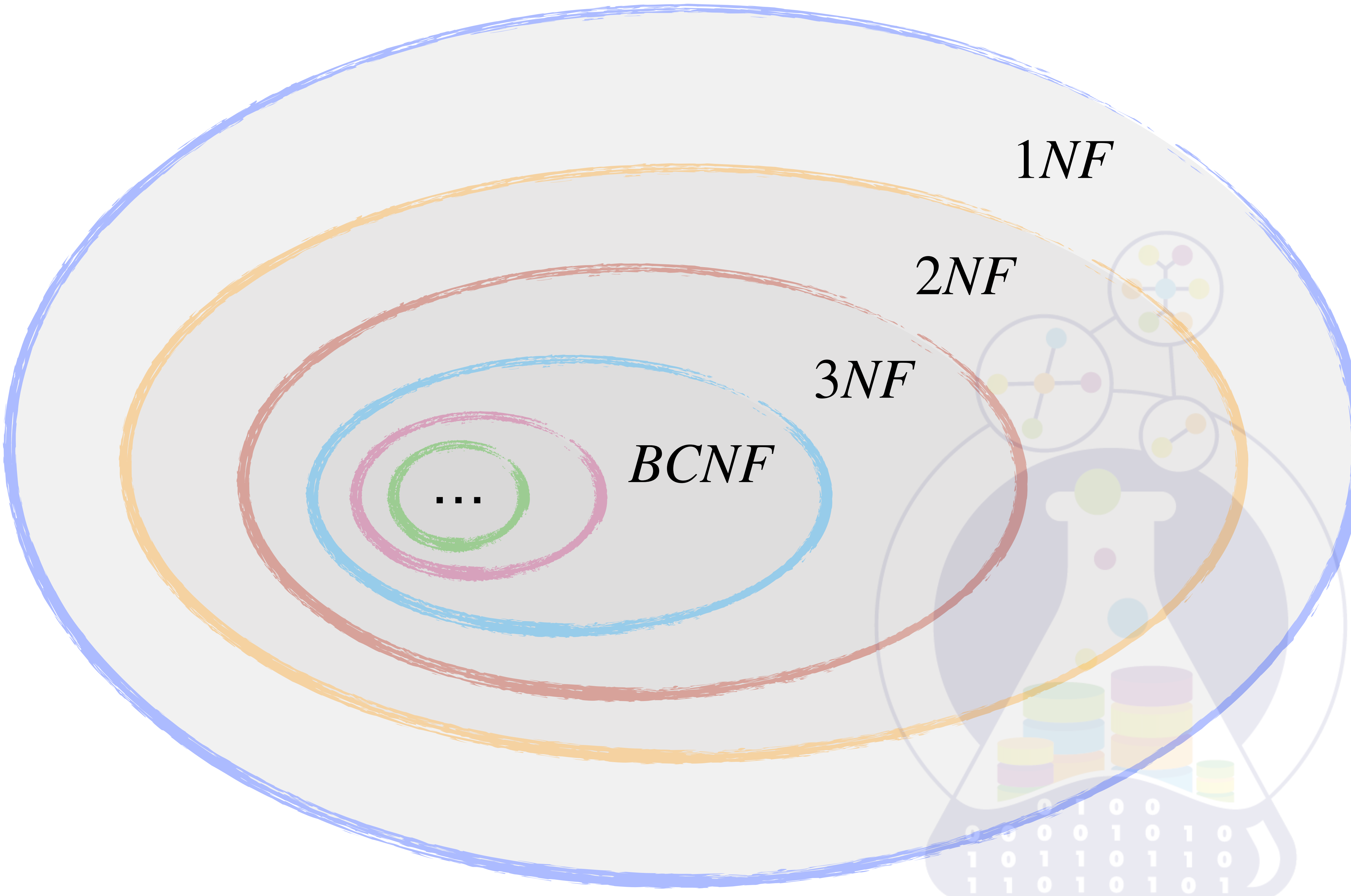
$$Student(ID, Name)$$

$$Apply(ID, CollegeNew)$$

$$EducationHistory(ID, College, CollegeCity)$$

$$Hobby(ID, Hobby)$$

Principle of full **normalization** (PoFN)



1 NF

- Each attribute contains only one value
- No duplicates in table
- All attribute values are atomic

Firm	Model
BMW	M5, X5M, M1
Nissan	GT-R

Firm	Model
BMW	M5
BMW	X5M
BMW	M1
Nissan	GT-R

Firm	AvgPoint	Model	Phone
BMW	5	M5, X5M, M1	123, 4567812
Nissan	5	GT-R	562, 6654332

Firm	AvgPoint	Model	Phone
BMW	5	M5	123
BMW	5	M5	4567812
BMW	5	X5M	123
BMW	5	X5M	4567812
BMW	5	M1	123
BMW	5	M1	4567812
Nissan	5	GT-R	562
Nissan	5	GT-R	6654332

Firm	AvgPoint	Model1	Model2	Model3	Phone1	Phone2
BMW	5	M5	***null***	***null***	123	4567812
BMW	5	***null***	X5M	***null***	123	4567812
BMW	5	***null***	***null***	M1	123	4567812
Nissan	5	GT-R	***null***	***null***	562	6654332

Functional Dependency - a constraint between two sets of attributes in a relation from a database

$X \to Y$, X functionally determines Y

$$1 \to 2$$

$$3 \to 4$$

$$5 \to 6$$

X	Y
1	2
3	4
5	6
1	2
5	6

Functional Dependency - a constraint between two sets of attributes in a relation from a database

$X \nrightarrow Y$, X doesn't functionally determine Y

$$1 \rightarrow 2$$

$$3 \rightarrow 4$$

$$5 \nrightarrow 6$$

X	Y
1	2
3	4
5	6
1	2
5	3

$$R(Model, Shop, Phone)$$

Model	Shop	Phone
BMW	De-Shop	87-33-44
Nissan	Nissan-Auto	15-34-21
Audi	De-Shop	87-33-44
Porsche	De-Shop	87-33-45

$$Model \rightarrow Shop$$

$$Model, Shop \rightarrow Phone$$

$$Shop \nrightarrow Phone$$

Armstrong axioms

Axiom of reflexivity

$$Y \subseteq X \;\Rightarrow\; X \to Y$$

Axiom of augmentation

$$X \to Y, \forall Z \;\Rightarrow\; XZ \to YZ$$

Axiom of transitivity

$$X \to Y, Y \to Z \Rightarrow X \to Z$$

Rule of decomposition

$$X \to YZ \Rightarrow X \to Y, X \to Z$$

Proof

$YZ \to Y$, based on **reflexivity** $Y \subseteq YZ$

$X \to Y$, based on **transitivity**

$YZ \to Z$, based on **reflexivity** $Z \subseteq YZ$

$X \to Z$, based on **transitivity**

#

Rule of composition

$$A \to B, C \to D \Rightarrow AC \to BD$$

Proof

$A \to B \Rightarrow AC \to BC$, based on **augmentation**

$C \to D \Rightarrow BC \to BD$, based on **augmentation**

$AC \to BD$, based on **transitivity**

#

Rule of union (notation)

$$A \to B, A \to C \Rightarrow A \to BC$$

Proof

$$A \to B \Rightarrow AA \to AB \Rightarrow A \to AB$$

$$A \to C \Rightarrow AB \to BC$$

$$A \to BC$$

#

Rule of pseudo transitivity

$$X \to Y, YZ \to W \Rightarrow XZ \to W$$

Proof

$$X \to Y \Rightarrow XZ \to YZ$$

$$XZ \to W$$

#

Rule of self determination

$$\forall X, X \to X$$

Proof

$X \to X$, based on **reflexivity** $X \subseteq X$

#

Functional Dependency

$$\{A_1, ..., A_n\} \rightarrow \{B_1, ..., B_m\}$$

trivial $\Rightarrow \{B_1, ..., B_m\} \subseteq \{A_1, ..., A_n\}$

nontrivial $\Rightarrow \exists\, B_i \notin \{A_1, ..., A_n\}$

completely nontrivial $\Rightarrow \forall i = \overline{1, m};\ B_i \notin \{A_1, ..., A_n\}$

Model	Shop	Phone
BMW	De-Shop	87-33-44
Nissan	Nissan-Auto	15-34-21
Audi	De-Shop	87-33-44
Porsche	De-Shop	87-33-45

trivial:

$$Model, Phone \rightarrow Model; \ \{Model\} \subseteq \{Model, Phone\}$$

nontrivial:

$$Model, Phone \rightarrow Shop, Phone; \ Phone \in \{Model, Phone\}$$

completely nontrivial

$$Model, Shop \rightarrow Phone; \ Phone \notin \{Model, Shop\}$$

Algorithm of Closure

$F = \{F_1, \ldots, F_k\}$ is a set of **FDs** of relation *R*

$(A \to B_i) \in F,$

$$\forall i \;\Rightarrow\; (A \to Y) \in F^+, \forall Y \subseteq \{B_1, \ldots, B_n\}$$

$$X^{(0)} = X$$

$$X^{(i+1)} = X^{(i)} \cup \{A \,|\, (Y \to Z) \in F, Y \subseteq X^{(i)}, A \subseteq Z\}$$

if $F^+ = G^+ \Rightarrow F, G$ are **equivalent sets**

$$R = (A, B, C, D, E, F)$$

$$F = (AB \rightarrow C, BC \rightarrow AD, D \rightarrow E, CF \rightarrow B)$$

$$\{A, B\}^+ = ?$$

$$\{A, B\}^+ = \{A, B\}$$

$$\{A, B\}^+ = \{A, B, C\} \Leftarrow AB \rightarrow C$$

$$\{A, B\}^+ = \{A, B, C, D\} \Leftarrow BC \rightarrow AD$$

$$\{A, B\}^+ = \{A, B, C, D, E\} \Leftarrow D \rightarrow E$$

Model	Shop	Phone
BMW	De-Shop	87-33-44
Nissan	Nissan-Auto	15-34-21
Audi	De-Shop	87-33-44
Porsche	De-Shop	87-33-45

$$F = (Model \rightarrow Shop;\ Model, Shop \rightarrow Phone)$$

$$\{Model\}^+ = ?$$

$$\{Model\}^+ = \{Model\}$$

$$\{Model\}^+ = \{Model, Shop\}$$

$$\{Model\}^+ = \{Model, Shop, Phone\}$$

$$R = (A, B, C, D, E, F)$$

$$F = (AB \to C, BC \to AD, D \to E, CF \to B)$$

$$AB \to D \ ?$$

$$\{A, B\}^{+} = \{A, B, C, D, E\} \Rightarrow AB \to D$$

$$D \to A \ ?$$

$$\{D\}^{+} = \{D, E\} \Rightarrow D \nrightarrow A$$

Relation $R = (A_1, \ldots, A_n)$

$F^R = (F_1, \ldots, F_m)$

Relation $S = (A_1, \ldots, A_{k-1}, A_{k+1}, \ldots, A_n)$

$F^S = ?$

$$R = (A, B, C, D)$$

$$F^R = (A \to B, B \to C, C \to D)$$

$$S = (A, C, D)$$

$$\{A\}^+ = \{A, \cancel{B}, C, D\}$$

$$A \to A$$

$$A \to C$$

$$A \to D$$

$$\{C\}^+ = \{C, D\}$$

$$C \to C$$

$$C \to D$$

$$\{D\}^+ = \{D\}$$

$$D \to D$$

$$A \to C$$

$$A \to D$$

$$C \to D$$

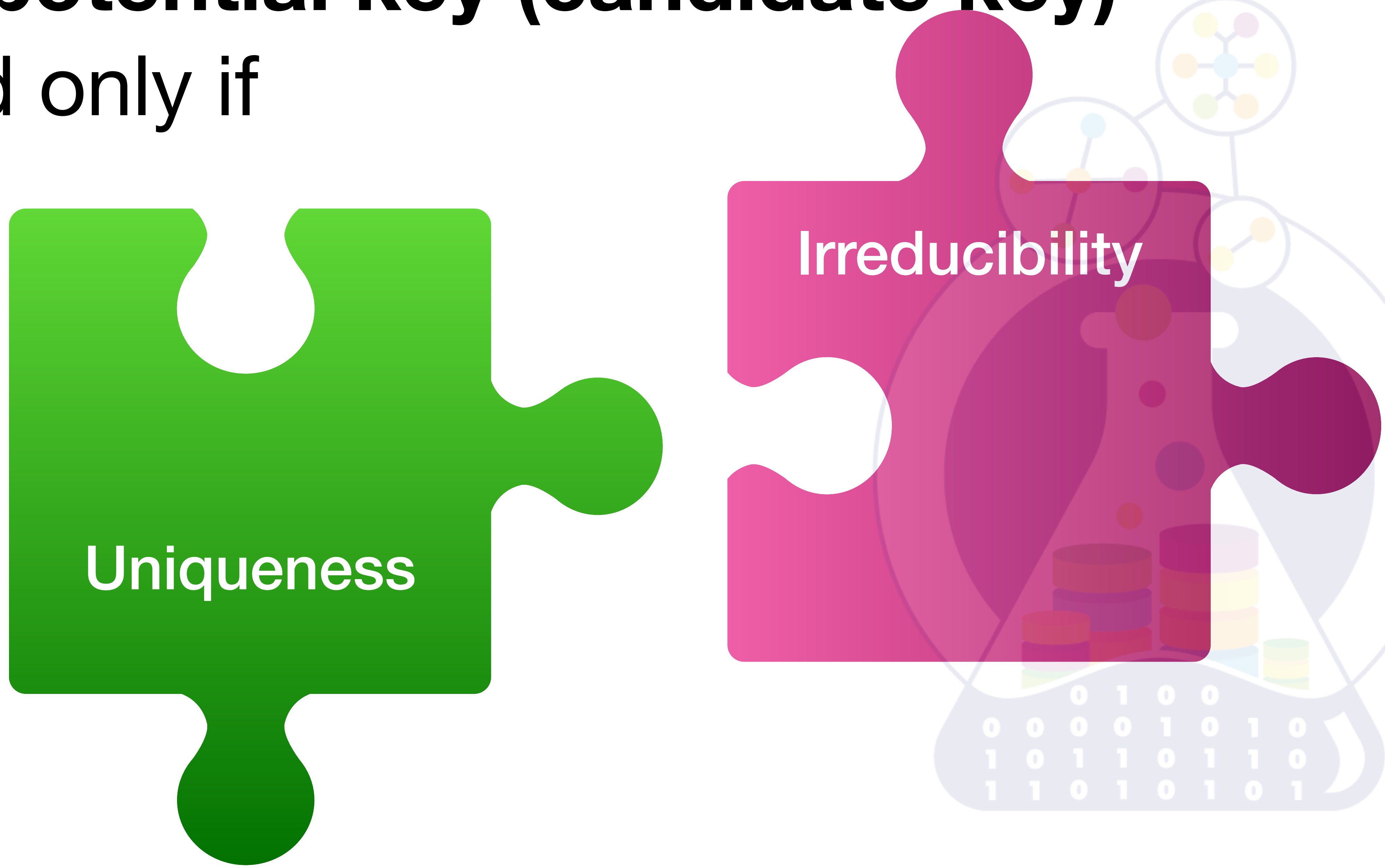
$$A \to B$$

which a role does *A* “play”

$$R(A_1, A_2, \ldots, A_n)$$

$$K = \{A_1, A_2, \ldots, A_m\}, m \leq n$$

K is **potential key (candidate key)**
if and only if

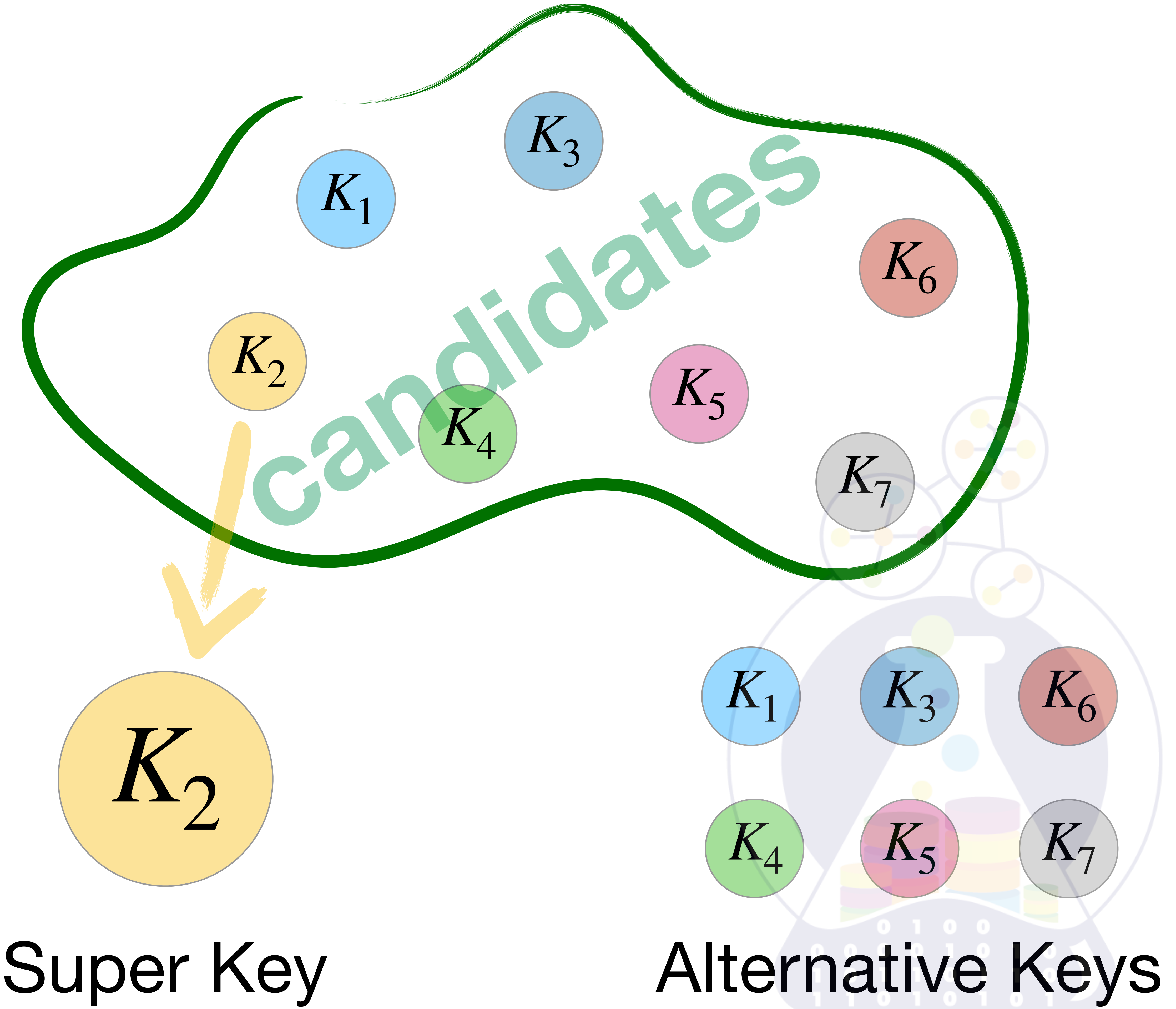


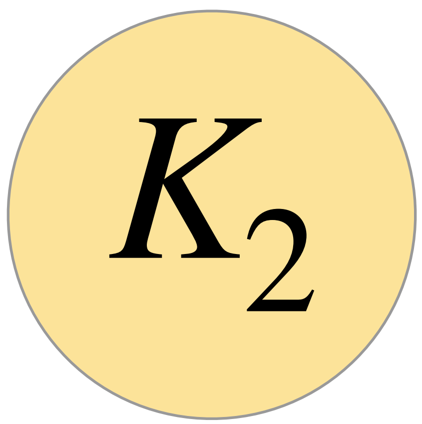
K is **simple potential key (simple candidate key)** if

$K = \{A_j\}$ has only one attribute

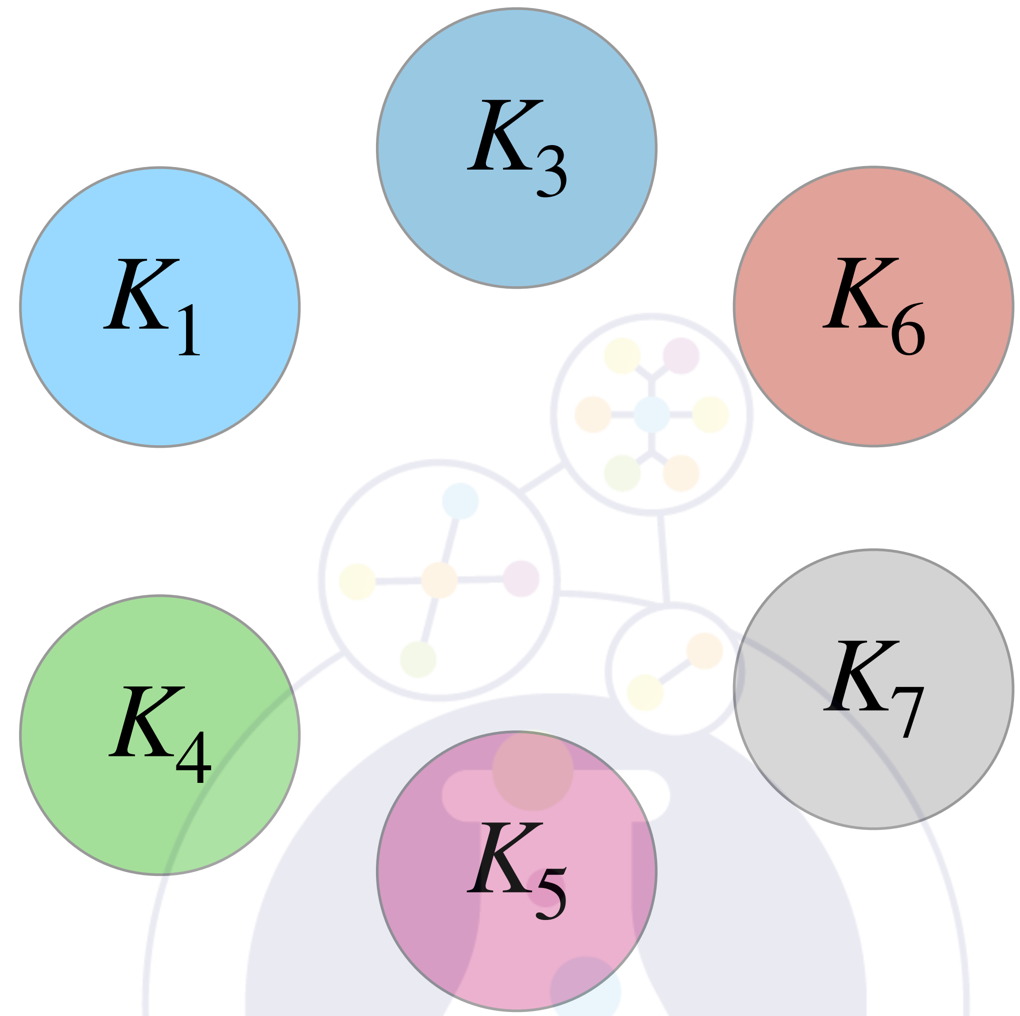
K is **compound potential key (compound candidate key)** if

$K = \{A_1, A_2, \dots\}$ has more then one


candidates
K1
K3
K6
K2
K5
K4
K7
K2
K1
K3
K6
K4
K5
K7
Super Key
Alternative Keys



Primary Key



Unique Keys

Model	Shop	Phone
BMW	De-Shop	87-33-44
Nissan	Nissan-Auto	15-34-21
Audi	De-Shop	87-33-44
Porsche	De-Shop	87-33-45

$$\{Model\}^+ = \{Model, Shop, Phone\}$$

$(Model)$

$(Model, Shop, Phone)$

$(Model, Shop)$

Keys

$$R(A_1, \ldots, A_k)$$

$$\{A_1, \ldots, A_n\} \rightarrow \{A_{n+1}, \ldots, A_k\};$$

$(A_1, \ldots, A_n)$ is a **SuperKey** of R if there is a **closure**

$$\{A_1, \ldots, A_n\}^+ = \{A_1, \ldots, A_n, A_{n+1}, \ldots, A_k\}$$

2 NF

- 1 NF
- Every non-key attribute is functionally dependent on a subset of the candidate key(s)

student	course	grade	course fee
Petr	Database	A	100
Ivan	Data Mining	B	150
Anna	Database	B+	100

$(student, course)$ is a **Candidate** key

$grade, course\ fee$ are **Non-primes**

$$student, course \rightarrow grade$$

student	course	grade	course fee
Petr	Database	A	100
Ivan	Data Mining	B	150
Anna	Database	B+	100

$(student, course)$ is a **Candidate** key

$grade, course\ fee$ are **Non-primes**

$course \rightarrow course\ fee$

student	course	grade	course fee
Petr	Database	A	100
Ivan	Data Mining	B	150
Anna	Database	B+	100

=

student	course	grade
Petr	Database	A
Ivan	Data Mining	B
Anna	Database	B+

$\bowtie$

course	course fee
Database	100
Data Mining	150

3 NF

- 2 NF
- **All** non-prime attributes are directly (non-transitively) dependent on the entire candidate key

order_id	date	customer	email
1/2020	2020-01-15	Ivan	ivan@mail.ru
2/2020	2020-01-16	Anna	anna@mail.ru
3/3030	2020-01-17	Petr	petr@mail.ru
4/2020	2020-01-18	Sergey	s@mail.ru

$(order\_id)$ is a **Candidate** key

$date, customer, email$ are **Non-primes**

$$order\_id \rightarrow date$$

$$order\_id \rightarrow customer$$

order_id	date	customer	email
1/2020	2020-01-15	Ivan	ivan@mail.ru
2/2020	2020-01-16	Anna	anna@mail.ru
3/3030	2020-01-17	Petr	petr@mail.ru
4/2020	2020-01-18	Sergey	s@mail.ru

$(order\_id)$ is a **Candidate** key

$date, customer, email$ are **Non-primes**

$$customer \rightarrow email$$

order_id	date	customer	email
1/2020	2020-01-15	Ivan	ivan@mail.ru
2/2020	2020-01-16	Anna	anna@mail.ru
3/3030	2020-01-17	Petr	petr@mail.ru
4/2020	2020-01-18	Sergey	s@mail.ru

=

order_id	date	customer
1/2020	2020-01-15	Ivan
2/2020	2020-01-16	Anna
3/3030	2020-01-17	Petr
4/2020	2020-01-18	Sergey

$\bowtie$

customer	email
Ivan	ivan@mail.ru
Anna	anna@mail.ru
Petr	petr@mail.ru
Sergey	s@mail.ru

BCNF

- $X \rightarrow Y$ is a trivial FD ($Y \subseteq X$)
- X is a a Super Key for relation R

Author	Nationality	Book	Genre	pages
W.Shakespeare	English	The Comedy of Errors	Comedy	100
M.Winand	Austrian	SQL Performance Explained	Textbook	200
J.Ullman	American	A First Course in Database Systems	Textbook	500
J.Widom	American	A First Course in Database Systems	Textbook	500

$(author, book)$ is **Key**

$$author \rightarrow nationality$$

$$book \rightarrow genre, pages$$

$(author)$ is **Key**

Author	Nationality
W.Shakespear	English
M.Winand	Austrian
J.Ullman	American
J.Widom	American

$\bowtie$

$(book)$ is **Key**

Book	Genre	pages
The Comedy of Errors	Comedy	100
SQL Performance Explained	Textbook	200
A First Course in Database Systems	Textbook	500

$\bowtie$

Author	Book
W.Shakespeare	The Comedy of Errors
M.Winand	SQL Performance Explained
J.Ullman	A First Course in Database Systems
J.Widom	A First Course in Database Systems

$(author, book)$ is **Key**

$$R = \{A_1, \ldots, A_n\}$$

$$S = \{B_1, \ldots, B_m\}$$

$$T = \{C_1, \ldots, C_k\}$$

S, T are **decomposing** of R, if

$$\{A_1, \ldots, A_n\} = \{B_1, \ldots, B_m\} \cup \{C_1, \ldots, B_k\}$$

$$R = S \bowtie T$$

Hez's Theorem

$$R = (A, B, C)$$

$$\exists\, A \rightarrow B \Rightarrow R = \pi_{A,B} R \bowtie \pi_{A,C} R$$

Dependency-preserving decomposition

$$R = S \bowtie T$$

$$F_R = (F_S \cup F_T)$$

$$R = (A, B, C, D)$$

$$F_R = (AB \to C, C \to D, D \to A)$$

$$R = R_1(A, B, C) \bowtie R_2(C, D)$$

R_1 : $\{A\}^+ = \{A\}$

$\{B\}^+ = \{B\}$

$\{C\}^+ = \{C, A, \cancel{D}\}$

$\{A, B\}^+ = \{A, B, C, \cancel{D}\}$

$\{B, C\}^+ = \{B, C, \cancel{D}, A\}$

$\{A, C\}^+ = \{A, C, \cancel{D}\}$

F_{R_1} :

$\Rightarrow C \to A$

$\Rightarrow A, B \to C$

$\Rightarrow B, C \to A$

$$R = (A, B, C, D)$$

$$F_R = (AB \rightarrow C, C \rightarrow D, D \rightarrow A)$$

$$R = R_1(A, B, C) \bowtie R_2(C, D)$$

$$R_2 : \quad \{D\}^+ = \{D, \cancel{A}\} \qquad\qquad F_{R_2} :$$

$$\{C\}^+ = \{C, \cancel{A}, D\}$$

$$\{C, D\}^+ = \{C, D, \cancel{A}\} \qquad\qquad \Rightarrow C \rightarrow D$$

$$R = (A, B, C, D)$$

$$F_R = (AB \rightarrow C, C \rightarrow D, D \rightarrow A)$$

$$R = R_1(A, B, C) \bowtie R_2(C, D)$$

$$F_{R_1} = (C \rightarrow A, AB \rightarrow C, BC \rightarrow A)$$

$$F_{R_2} = (C \rightarrow D)$$

$$F_{R_1} \cup F_{R_2} = (C \rightarrow A, AB \rightarrow C, BC \rightarrow A, C \rightarrow D)$$

$F_R \neq F_{R_1} \cup F_{R_2} \Rightarrow$ **not dependency-preserving**

$$R = Employee(id, nationality, dept, dept\_type, dept\_phone)$$

$$F_R = (id \rightarrow nationality; dept \rightarrow dept\_type, dept\_phone)$$

1st iteration

$$R_1(id, nationality)$$

$$R_2(\underline{id}, dept, dept\_type, dept\_phone)$$

2nd iteration

$$R_1(id, nationality)$$

$$R_3(dept, dept\_type, dept\_phone)$$

$$R_4(\underline{dept}, id)$$

$$R(id, nationality, dept, dept\_type, dept\_phone)$$

$$R_1(id, nationality)$$

$$R_3(dept, dept\_type, dept\_phone)$$

$$R_4(dept, id)$$

$$R = R_1 \bowtie R_3 \bowtie R_4$$

Multivalued Dependency - a describes the relationship between two subsets inside a relation

$X \twoheadrightarrow Y$, X multivalued determines Y

X	Y	Z
1	1	3
2	1	2
2	2	3
1	2	3
2	1	3
2	2	2
3	1	2

Multivalued Dependency - a describes the relationship between two subsets inside a relation

$X \not\twoheadrightarrow Y,\ X$ doesn't multivalued determine Y

X	Y	Z
1	1	3
2	1	2
2	2	3
1	2	1
2	1	3
2	2	2
3	1	2

Armstrong axioms

$$X = \{X_1, \ldots, X_n\};\ Y = \{Y_1, \ldots, Y_k\};\ X \subseteq R;\ Y \subseteq R$$

Rule of complementation

$$X \twoheadrightarrow Y \Rightarrow X \twoheadrightarrow R - Y$$

Rule of augmentation

$$X \twoheadrightarrow Y;\ Z \subseteq W \Rightarrow XW \twoheadrightarrow YZ$$

Rule of transitivity

$$X \twoheadrightarrow Y;\ Y \twoheadrightarrow W \Rightarrow X \twoheadrightarrow W - Y$$

$$X = \{X_1, \ldots, X_n\};\ Y = \{Y_1, \ldots, Y_k\};\ X \subseteq R;\ Y \subseteq R$$

Rule of replication

$$X \rightarrow Y \Rightarrow X \twoheadrightarrow Y$$

Rule of coalescence

$$X \twoheadrightarrow Y; \exists W : W \cap Y = \emptyset; W \rightarrow Z; Z \subseteq Y \Rightarrow X \rightarrow Y$$

Rule of transitivity

$$X \twoheadrightarrow Y;\ Y \rightarrow W \Rightarrow X \twoheadrightarrow W - Y$$

Multivalued Dependency

$$\{A_1, \ldots, A_n\} \twoheadrightarrow \{B_1, \ldots, B_m\}$$

trivial $\Rightarrow \{B_1, \ldots, B_m\} \subseteq \{A_1, \ldots, A_n\}$

trivial $\Rightarrow R = \{B_1, \ldots, B_m\} \cup \{A_1, \ldots, A_n\}$

Fagin Theorem

$R(A, B, C)$ can be **decomposing** into

$$R_1(A, B)$$

$$R_2(A, C)$$

if and only if there is a MVD

$$A \twoheadrightarrow B$$

OR

$$A \twoheadrightarrow C$$

4 NF

- $X \twoheadrightarrow Y$ is a nontrivial MVD
- X is a a Super Key for relation R

Student	Course	Hobby
Ivan	Database	football
Ivan	Database	basketball
Ivan	Java	football
Ivan	Java	basketball
Petr	Database	tennis

$$Student \twoheadrightarrow Course$$

$$Student \twoheadrightarrow Hobby$$

Student	Course	Hobby
Ivan	Database	football
Ivan	Database	basketball
Ivan	Java	football
Ivan	Java	basketball
Petr	Database	tennis

$$=$$

Student	Course
Ivan	Database
Ivan	Java
Petr	Database

$$\bowtie$$

Student	Hobby
Ivan	football
Ivan	basketball
Petr	tennis

5 NF

~ **PJ/NF** (Projection-Join NF)

- For every non-trivial ***join dependency*** is implied by the candidate keys

Join Dependency

$* (X, Y, \ldots, Z)$ on relation $R = (X, Y, \ldots, Z)$

if and only if R can be restored without lose data by join all projections for every $X, Y, \ldots, Z$

Employee	**Dept**	**Project**
Ivanov	IT	Data analyse
Ivanov	Finance	Java integration
Sidorov	IT	Java integration
Ivanov	IT	Java integration

=

Employee	**Dept**
Ivanov	IT
Ivanov	Finance
Sidorov	IT

$\bowtie$

Dept	**Project**
Finance	Java integration
IT	Data analyse
IT	Java integration

$\bowtie$

Employee	**Project**
Ivanov	Data analyse
Ivanov	Java
Sidorov	Java

Fagin Theorem

$R(A, B, C)$ satisfies $*\{AB, AC\}$

$$A \twoheadrightarrow B \mid C \equiv *\{AB, AC\}$$

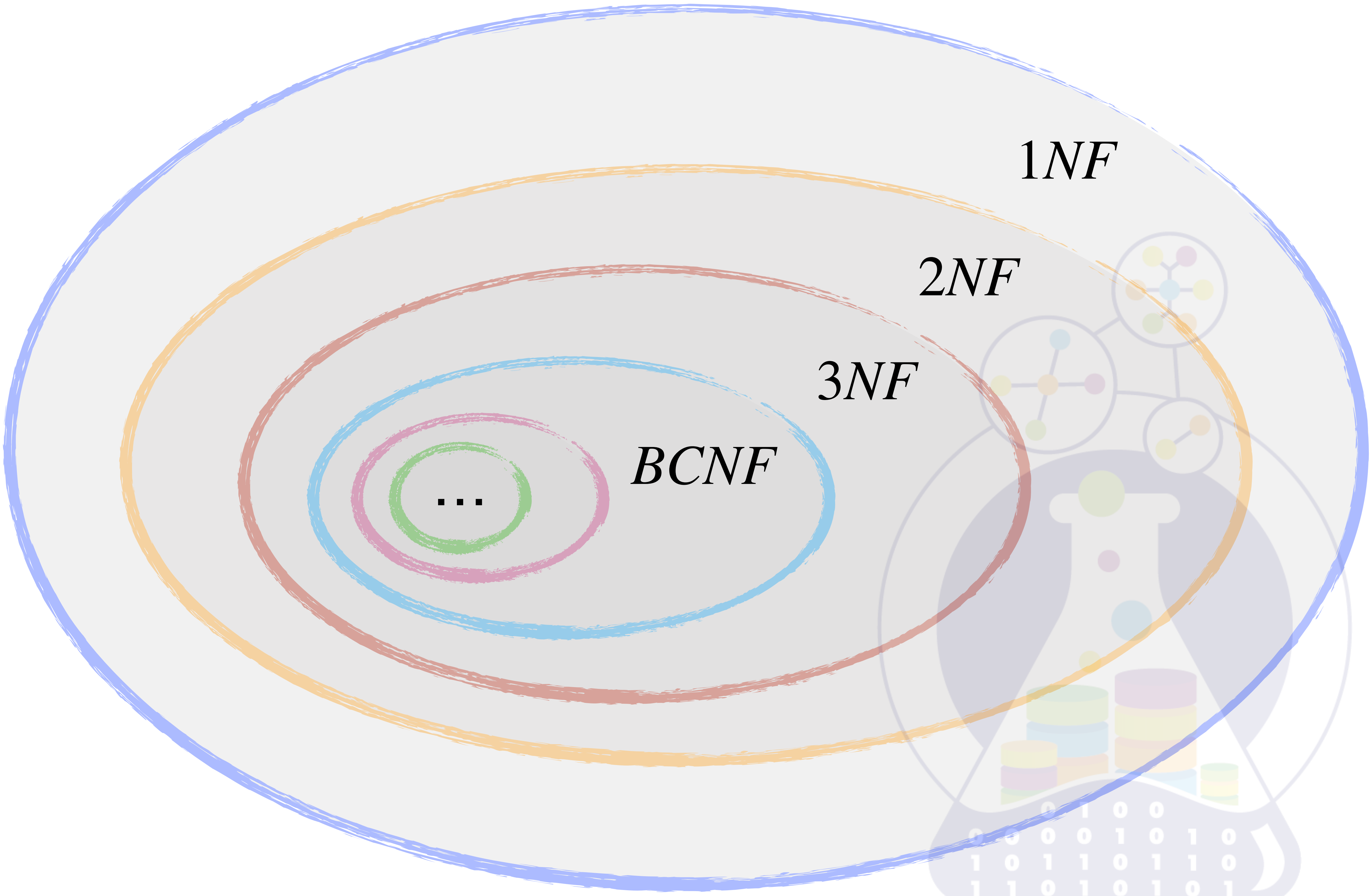
if and only if there is a MVD

$$A \twoheadrightarrow B$$

OR

$$A \twoheadrightarrow C$$

Principle of full **normalization** (PoFN)



$$R = College(\underline{Name}, City, cntStudents, avgPoint)$$

$$R_1 = CollegeCity(\underline{Name}, City)$$

$$R_2 = CollegeStudents(\underline{Name}, cntStudents)$$

$$R_3 = CollegePoint(\underline{Name}, avgPoint)$$

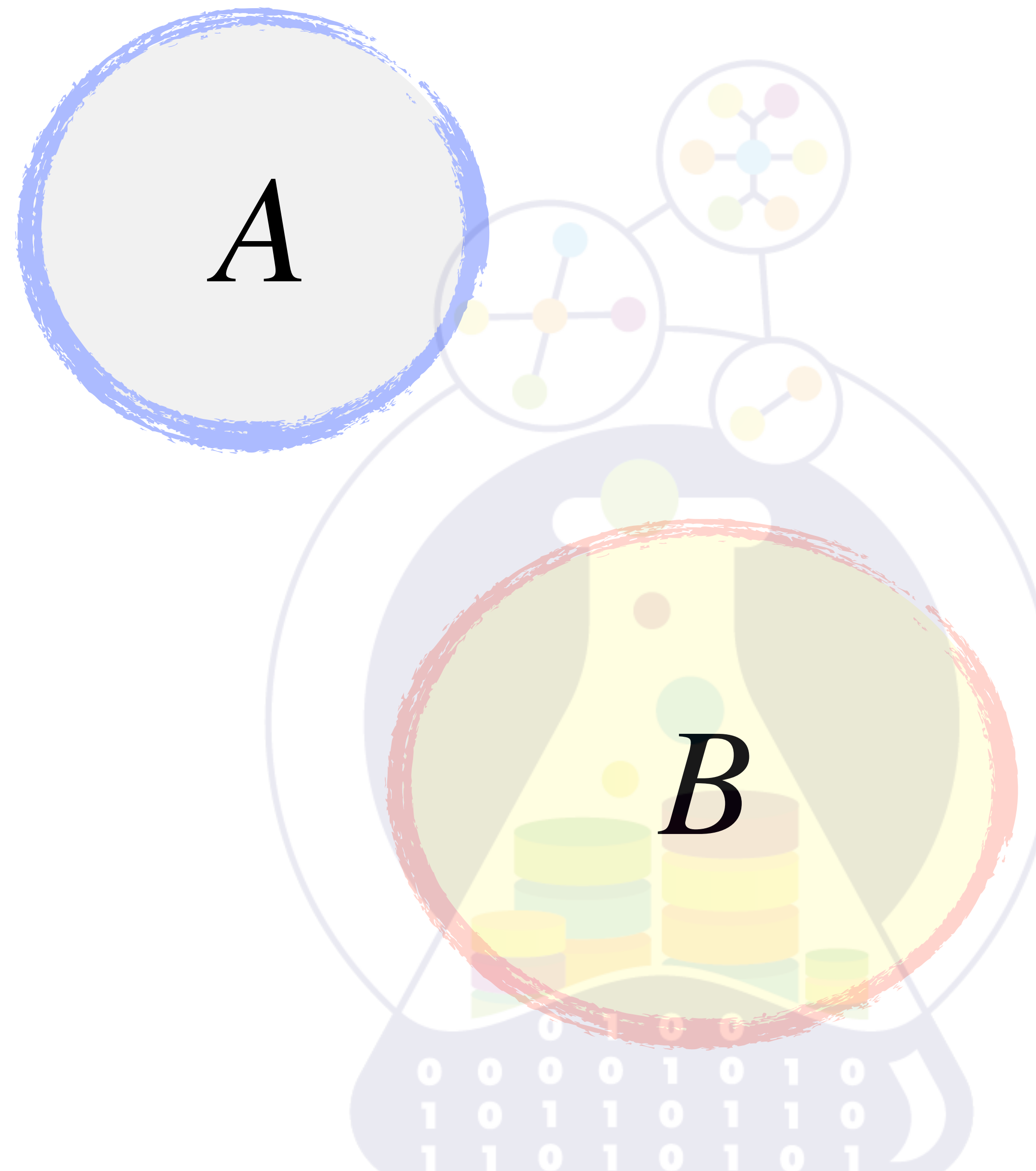
“too decomposed”

Principle of full **orthogonal design** (POOD)

$$A \cup B = \{A, B\}$$

$$A \cap B = \emptyset$$

$$A - B = A$$

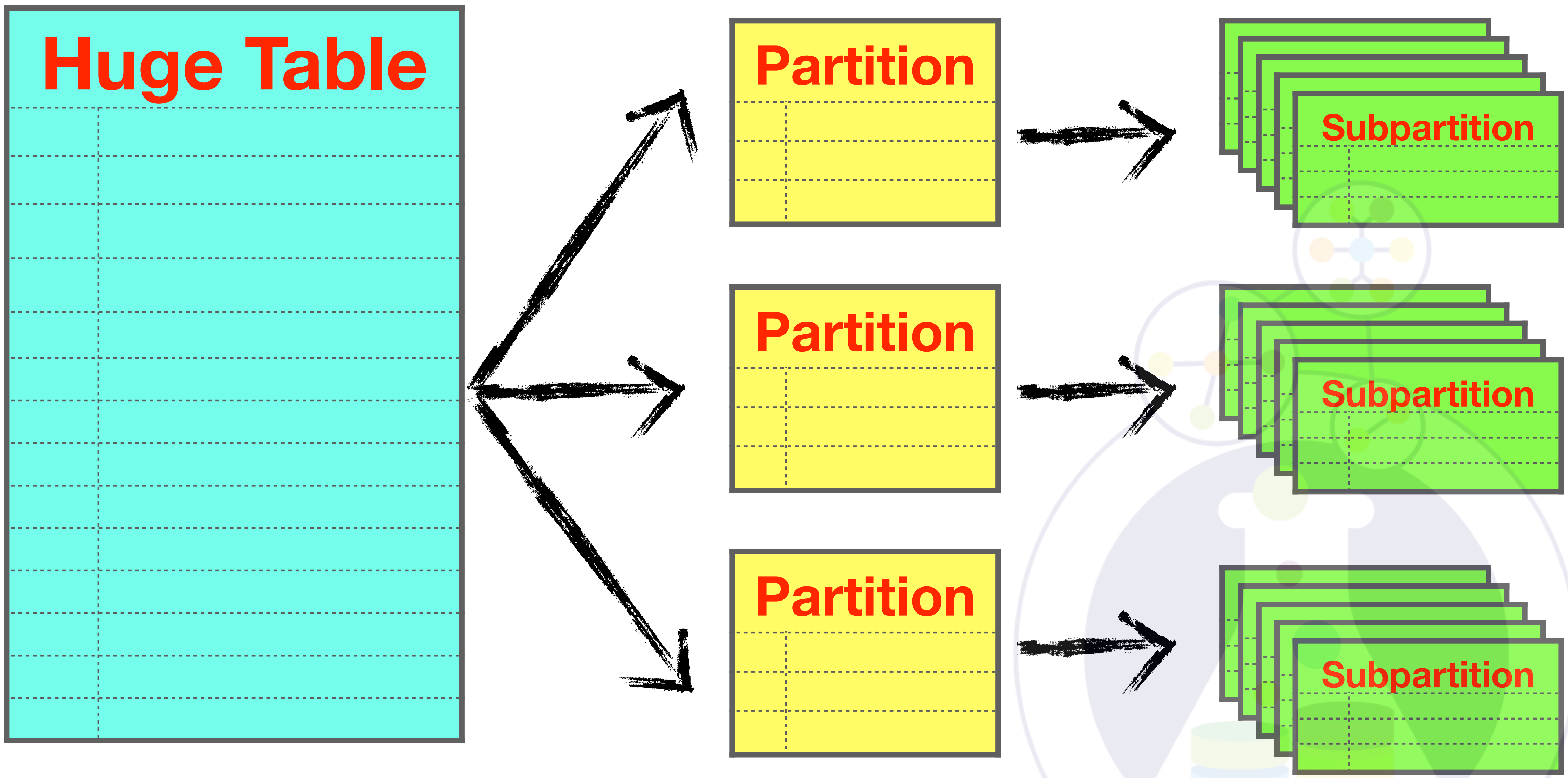


$R_1 =$

ID	Company	Status	City
123	Company #1	10	Kazan
456	Company #2	30	Kazan

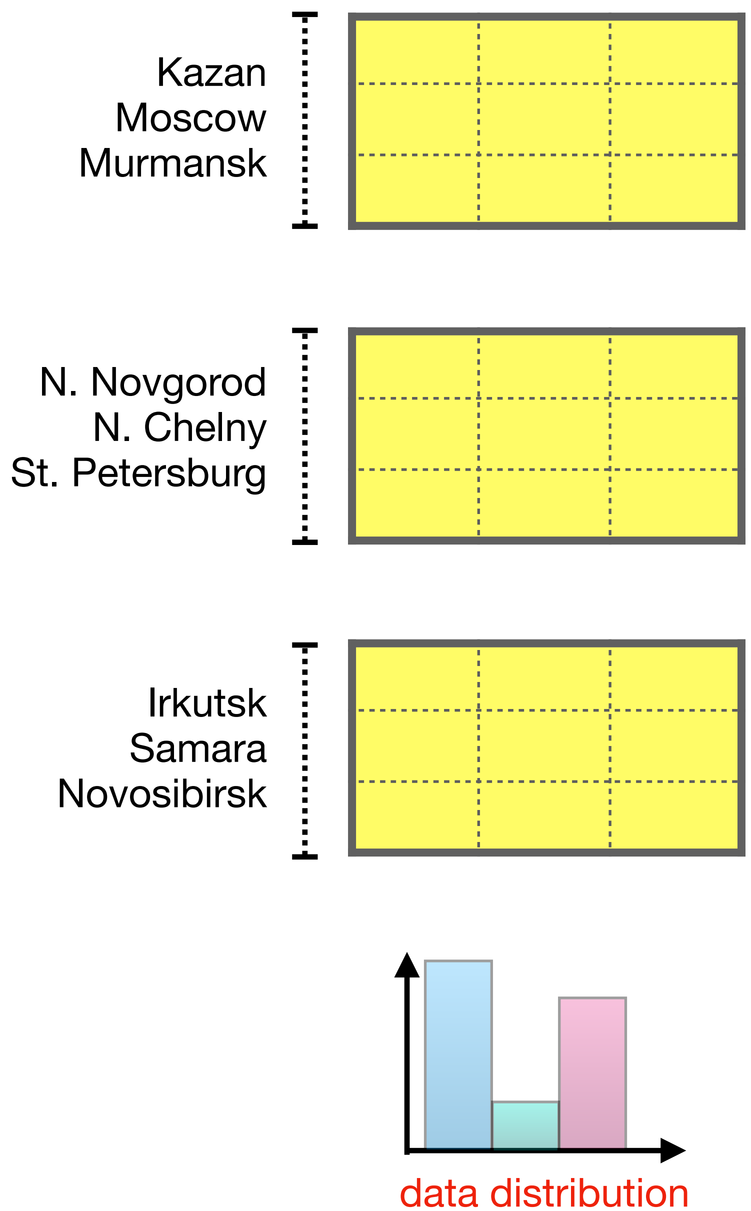
$R_2 =$

ID	Company	Status	City
1	Company #3	10	Kursk
789	Company #4	20	Moscow
321	Company #5	5	Moscow

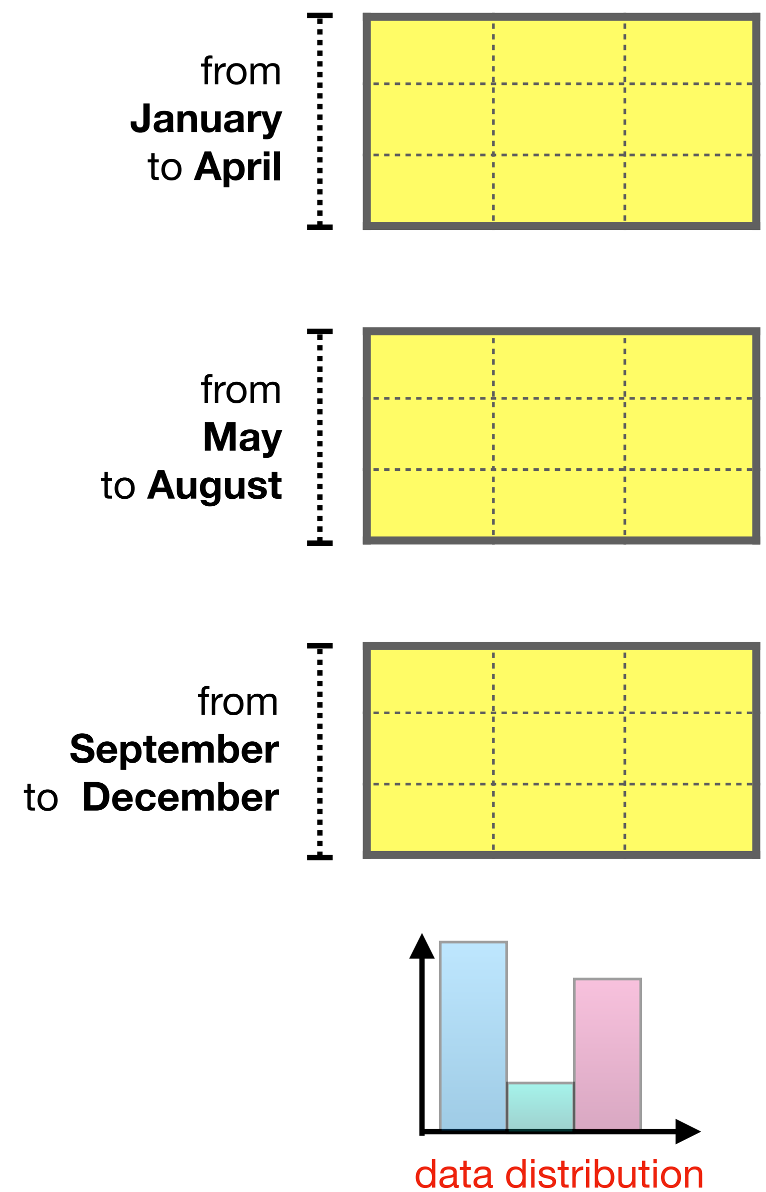


Azat Yakupov

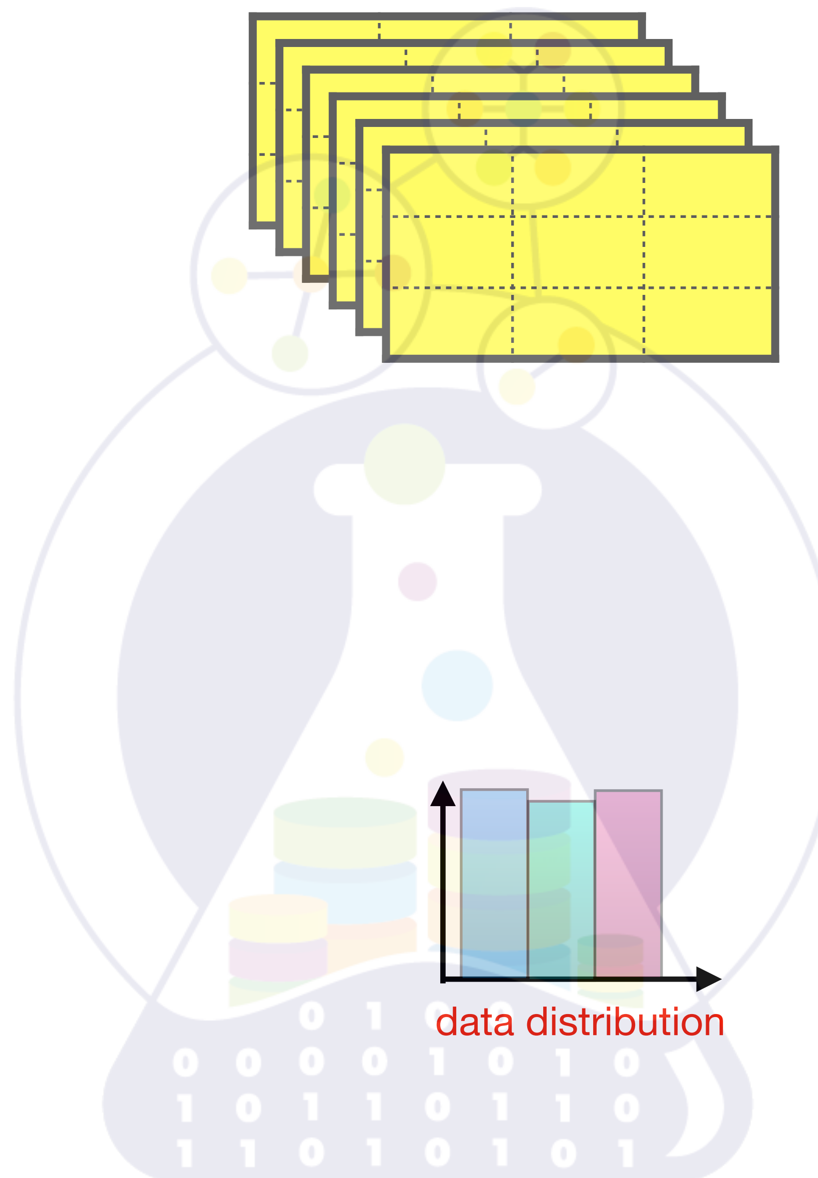
by List



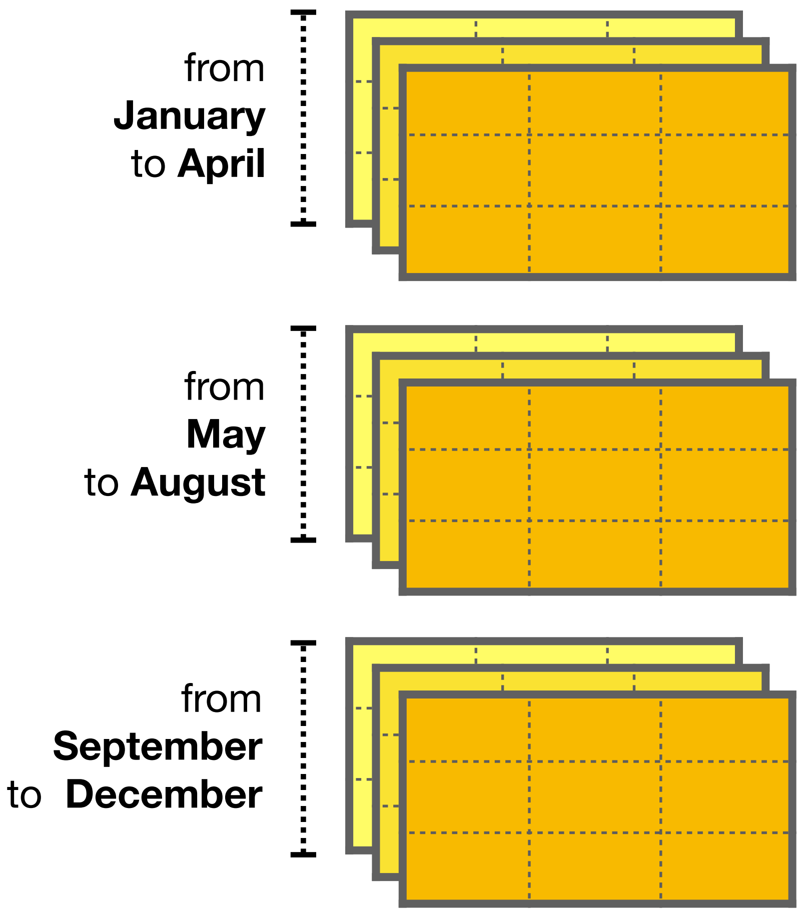
by Range



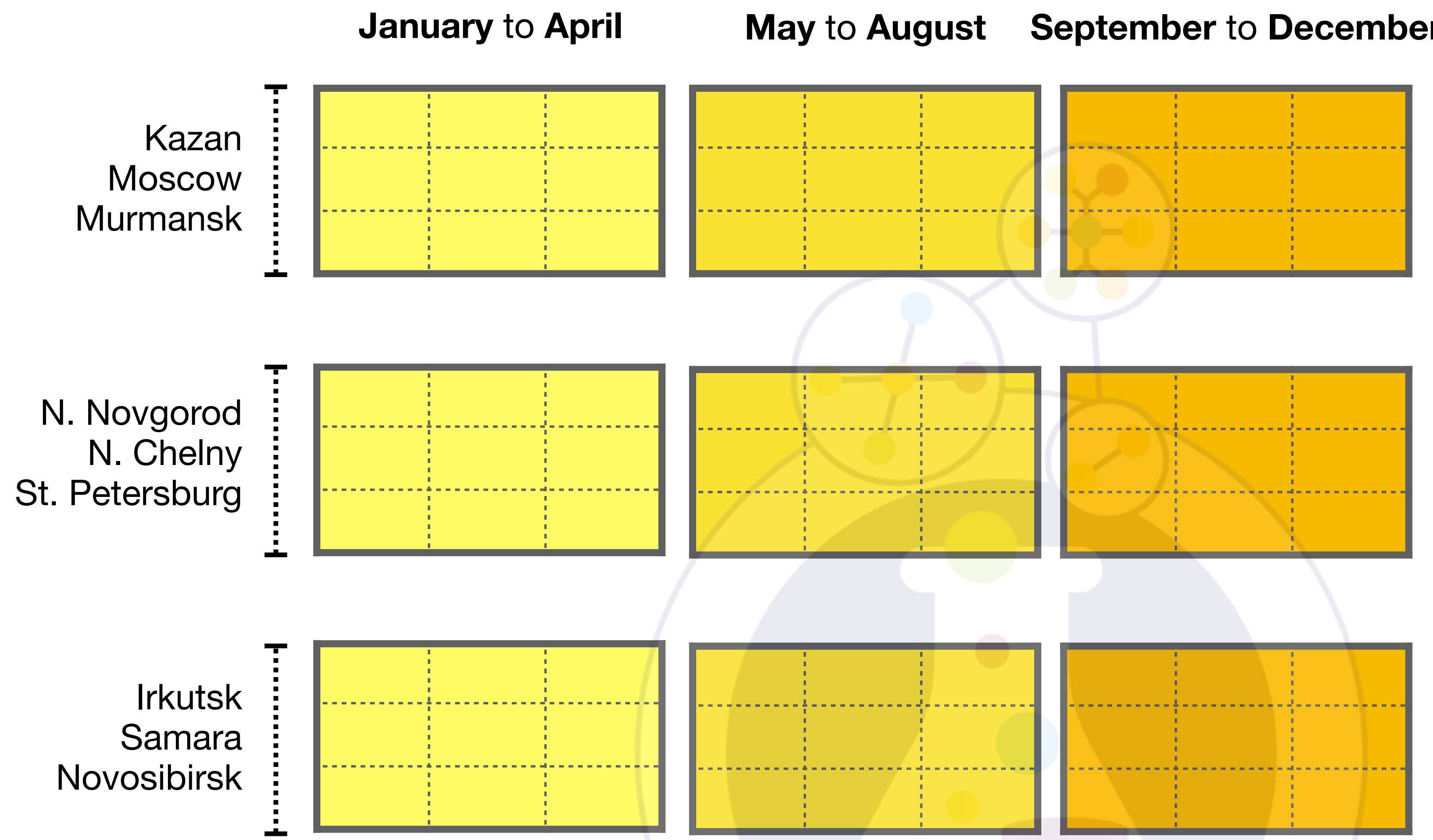
by Hash



by Range-Hash



by List-Range



```
CREATE TABLE test.list
(id BIGINT,
 name VARCHAR)
PARTITION BY LIST(id);
```

```
CREATE TABLE test.list_1
PARTITION OF test.list
FOR VALUES IN (1,2,3,4,5);
```

```
CREATE TABLE test.list_2
PARTITION OF test.list
FOR VALUES IN (6,7,8,9,10);
```

$$R_1 =$$

ID	Company	Status	City
123	Company #1	10	Kazan
456	Company #2	30	Kazan

$$R_2 =$$

ID	Company	Status	City
1	Company #3	10	Kursk
789	Company #4	20	Moscow
321	Company #5	5	Moscow
456	Company #2	30	Kazan

Many designs possible

Some of them are more **better** than others

Which design is more **convenient** for you?

COMMIT;

Azat Yakupov