Managing Structured Collections of Community Data

Wolfgang Gatterbauer, Dan Suciu

University of Washington, Seattle

1: Flashcards

1: Flashcards

407
frequency rank

radical

character (線)

tradition (if different)

线

乡 纟纩线线

stroke order
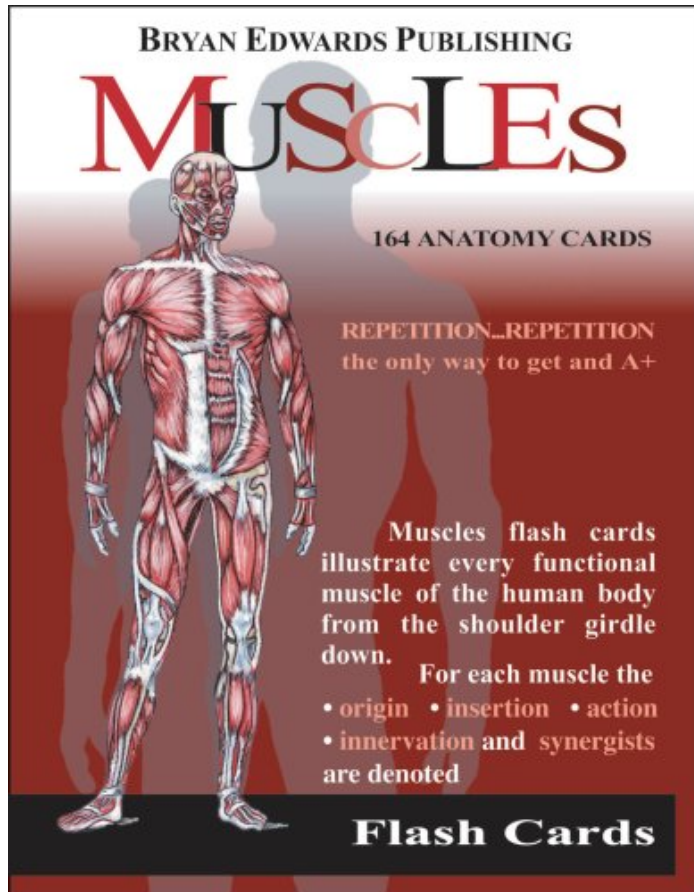
真

zhēn real, true, genuine

МОЛОД

nish

Hebrew Flash Cards

Front side

Back side

English with Ancient Hebrew meaning

Modern Hebrew

Alef
A

Strong, Power, Leader

Ancient Hebrew   Paleo Hebrew

1: Flashcards

1: Flashcards

Computer Science Abbreviations:

- 4NF
- ACID
- MVD
- RAID
- SQL
- FPGA
- FTL
- ...

Computer Science Concepts:

- Merge Sort
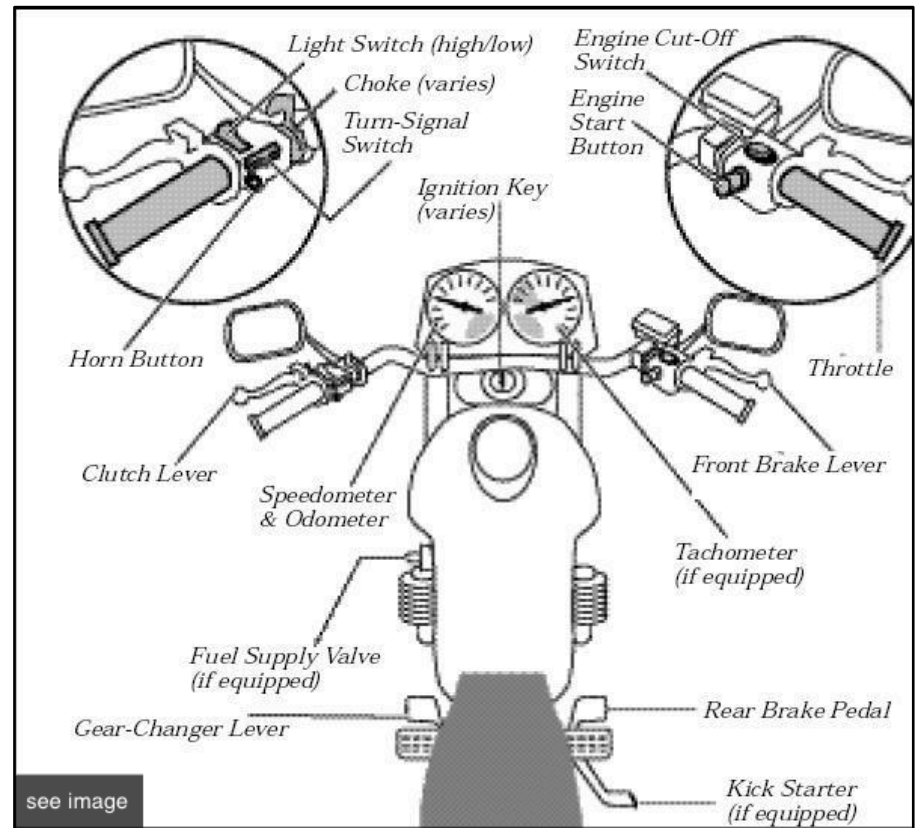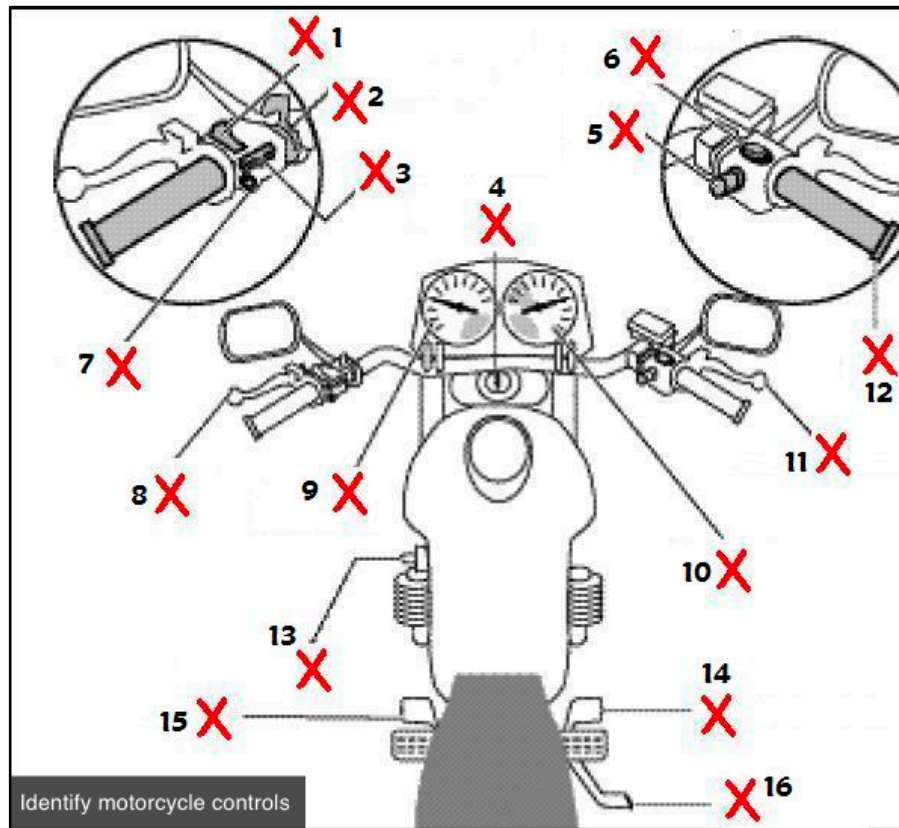- Two-phase locking
- ...

1: Flashcards

What is **height**

16mm

The fitting height for a progressive addition lens (FPD) is measured from the lowest point on the lens, or lens opening, to the center of the wearer's pupil.
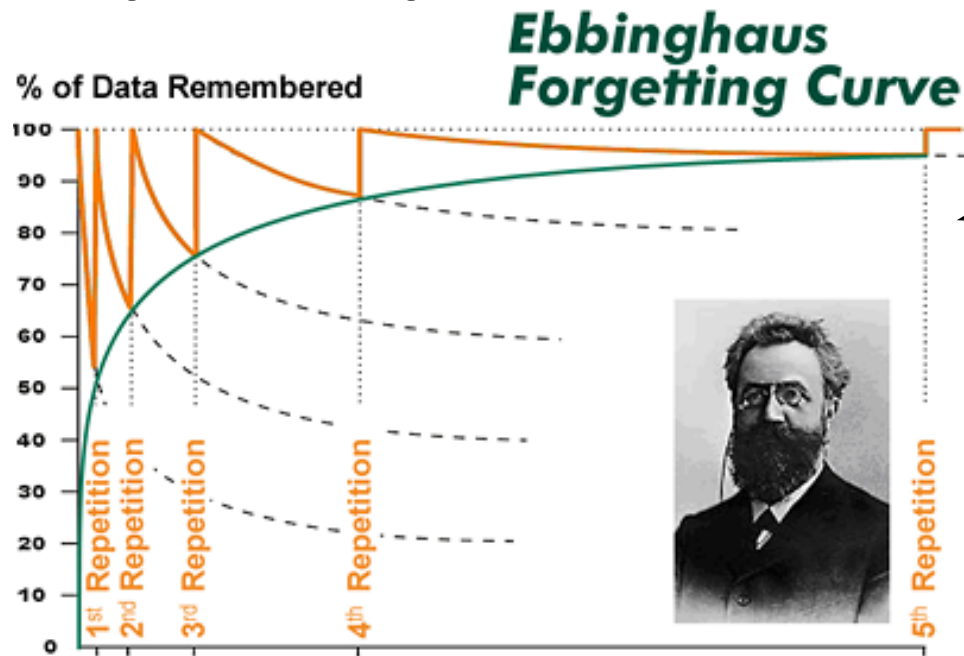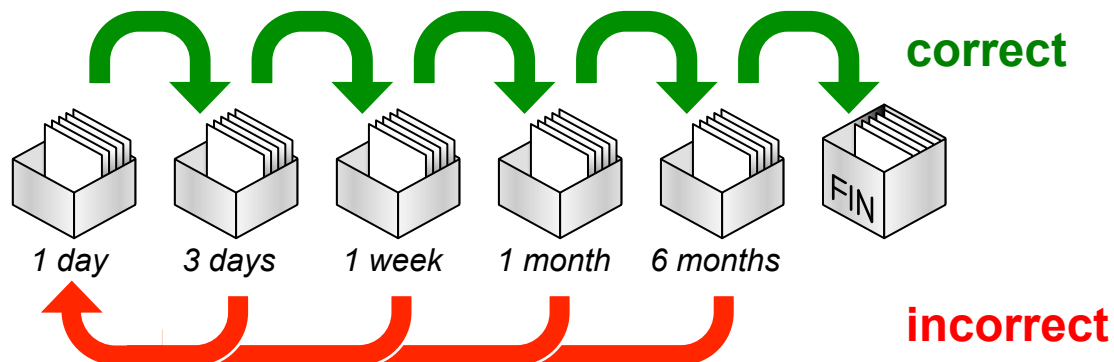

16mm

1: Flashcards

Texas DPS Motorcycle Operators Manual

2: Spaced Repetition

% of Data Remembered

Ebbinghaus Forgetting Curve

100
90
80
70
60
50
40
30
20
10
0

1st Repetition
2nd Repetition
3rd Repetition
4th Repetition
5th Repetition

Ebbinghaus Forgetting Curve

Leitner System (Pimsleur's graduated interval recall)

correct

1 day   3 days   1 week   1 month   6 months   FIN

incorrect

2: Spaced Repetition

2: Spaced Repetition

Specialized Software
- used by 3.000 schools
- sold 500.000 times

3: A Community

myPairSpace.com
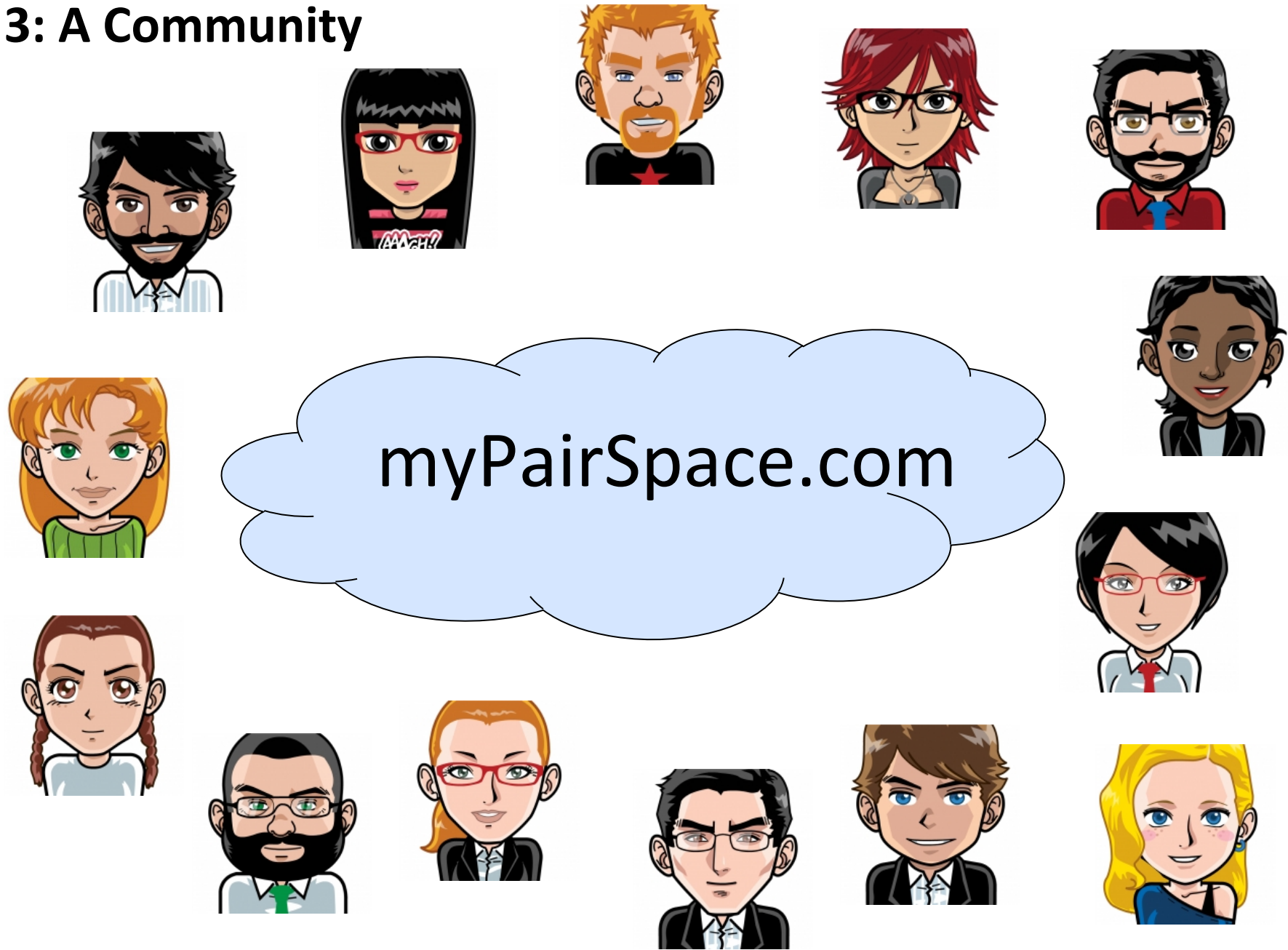
An example PairSpace scenario
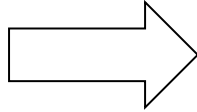
Spanish 1

1. go/ir
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Alice

Bob
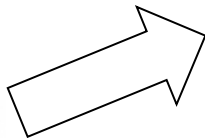
Spanish 1

1. go/andar
2. pay/pagar
3. come/venir
... ...
100. hear/oir
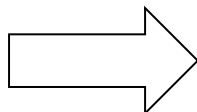
Charlie **?**

A. Alice inserts her first Spanish lesson

B. Bob searches and finds Alice's lesson

C. Bob adapts his copy of her original lesson

D. Charlie comes and searches for Spanish lessons

What to return, how to present, how to query, and how to rank?

Challenge 1

Alice

Spanish 1
1. go/ir
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Bob

Spanish 1
1. go/andar
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Charlie ?

1: What to return?

- Alice's (original)

- Bob's (most recent)

- their intersection

- their union

- presenting the one *conflicting* tuple

How to inform the user about the structural variation in collections?

Challenge 2

Alice

Spanish 1
1. go/ir
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Bob

Spanish 1
1. go/andar
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Charlie ?

2: How to present?

- lists of tuples ☹

- lists lessons & example tuples

- majority vs diversity

- cluster collections into meta-collections

What are optimal "return structures" and their visual representation?

Challenge 3

Alice → **Spanish 1**
1. go/ir
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Bob → **Spanish 1**
1. go/andar
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Charlie → ?

3: How to search?

- Keyword-based
- Form-based
- Language-based

  – varying trust
  – given we search for collections

How to best (fast, easy) allow users to to express their search needs?

Challenge 4

Alice

Spanish 1
1. go/ir
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Bob

Spanish 1
1. go/andar
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Charlie ?

4: How to rank?

- Syntactic & semantic similarity (across languages)

- Structure (items vs collection)

- Trust (vote- vs rule-based

- Provenance (on collections)

- Learning/Adjustment over time

Overview of Challenges

Spanish 1
1. go/ir
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Alice

Spanish 1
1. go/andar
2. pay/pagar
3. come/venir
... ...
100. hear/oir

Bob

Charlie ?

- **New Challenges**
  - Representation
  - Interface
  - Relevance measures

- **Cross-Cutting Challenges**
  - inconsistency/trust
  - non-monotonicy (dynamic evolution)
  - uncertainty
  - provenance

Some promising solutions

- New Challenges
  - Representation
  - Interface
  - Relevance measures

- Cross-Cutting Challenges

VLDB 2009

Sigmod 2010
  - inconsistency/trust
  - non-monotonicy (dynamic evolution)

MUD 2010
  - uncertainty

(VLDB 2011)
  - provenance

Managing the human genome

1: ACCGCAACGTTATAGGCACGCTATATCG

2: ACCGCAACGTATTATAGGCACGCTATATCG

3: ACCGCAACGTATTATAGGCACGATATCTCG

4: ACCGCAACGTATTAGGCACGATATCTCG

5: ACCGCAATTAGGCACGTACGATATCTCG

…

1B: ACCGCAATTAGGGACGTACGATATCTCG

Managing the human genome

1: ACCGCAACGTTATAGGCACGCTATATCG

insertion

2: ACCGCAACGTATTATAGGCACGCTATATCG   *large-scale structural variations*

inversion

3: ACCGCAACGTATTATAGGCACGATATCTCG

deletion

4: ACCGCAACGTATTAGGCACGATATCTCG

translocation

5: ACCGCAATTAGGCACGTACGATATCTCG   *single nucleotide polymorphism*

...

SNP

1B: ACCGCAATTAGGGACGTACGATATCTCG

20
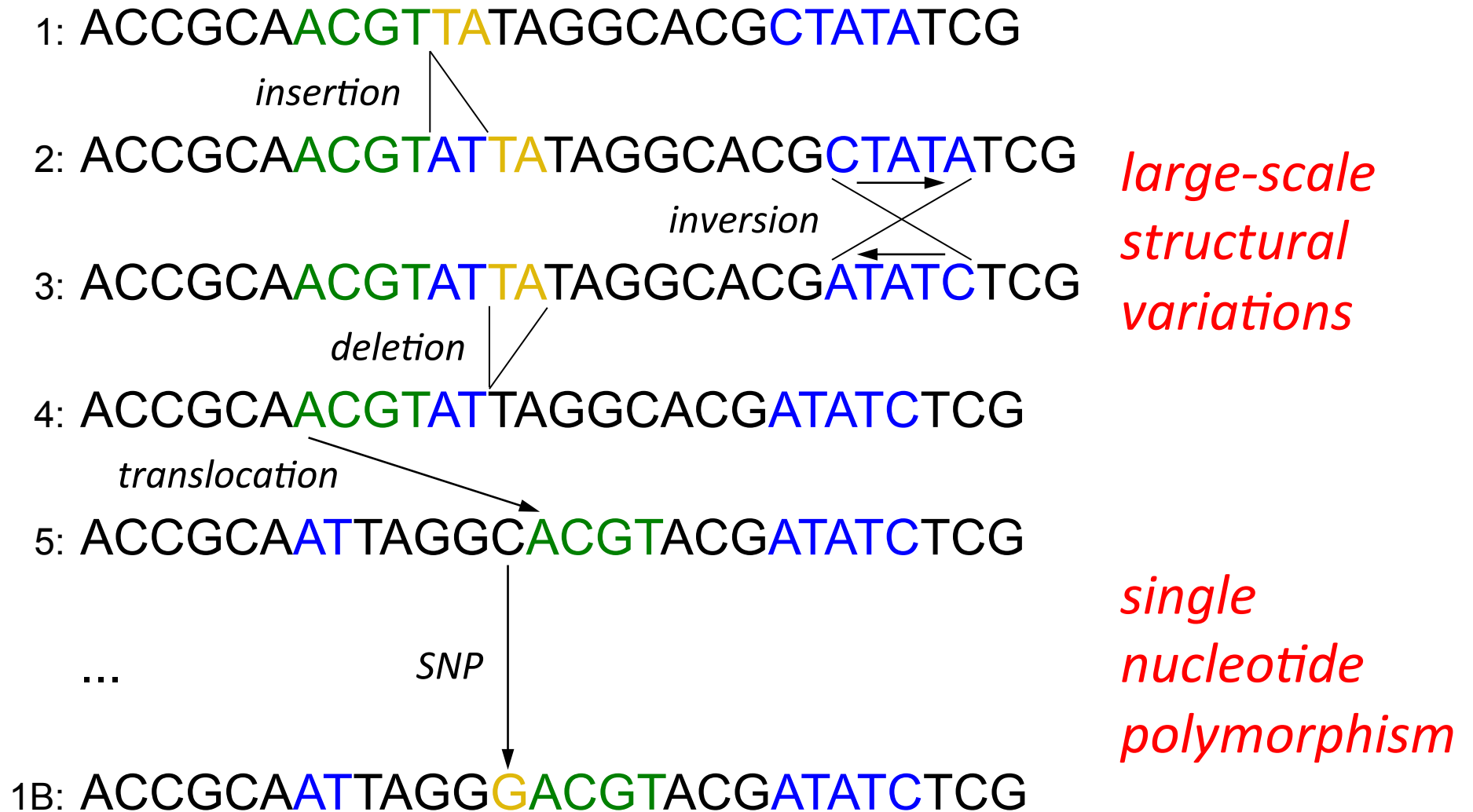
The Vision

- myPairSpace.com
  - one massive central repository for ce-learning needs
  - has the typical DM challenges of any community DB
  - new: management of *collections* and their evolution

- Then abstract and apply learned principles
  - data determines the structure
  - management of the human genome ("management" versus "scientific management")