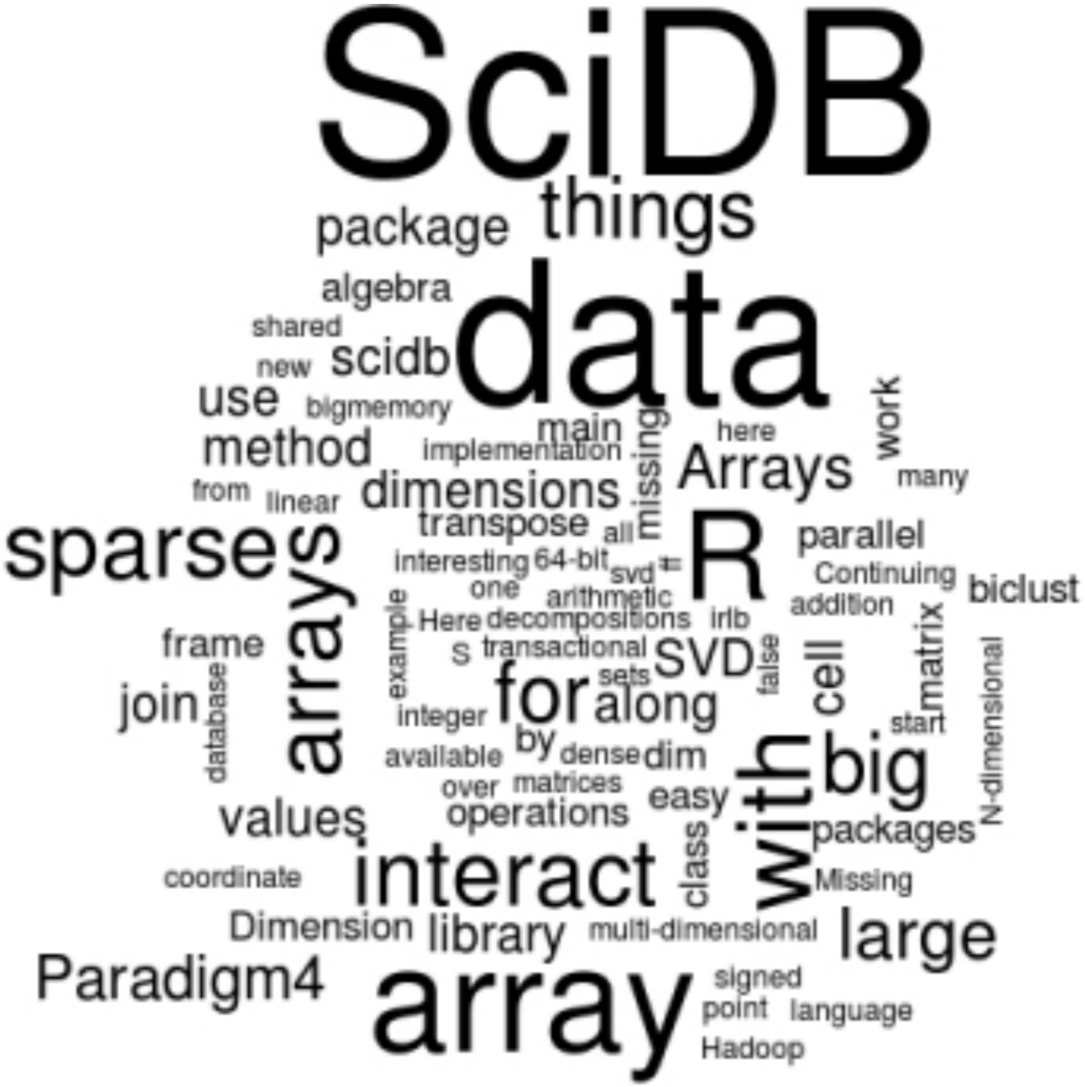


http://goo.gl/wPRSO

Bryan Lewis, Paradigm4
blewis@paradigm4.com

SciDB

Free and open-source array database

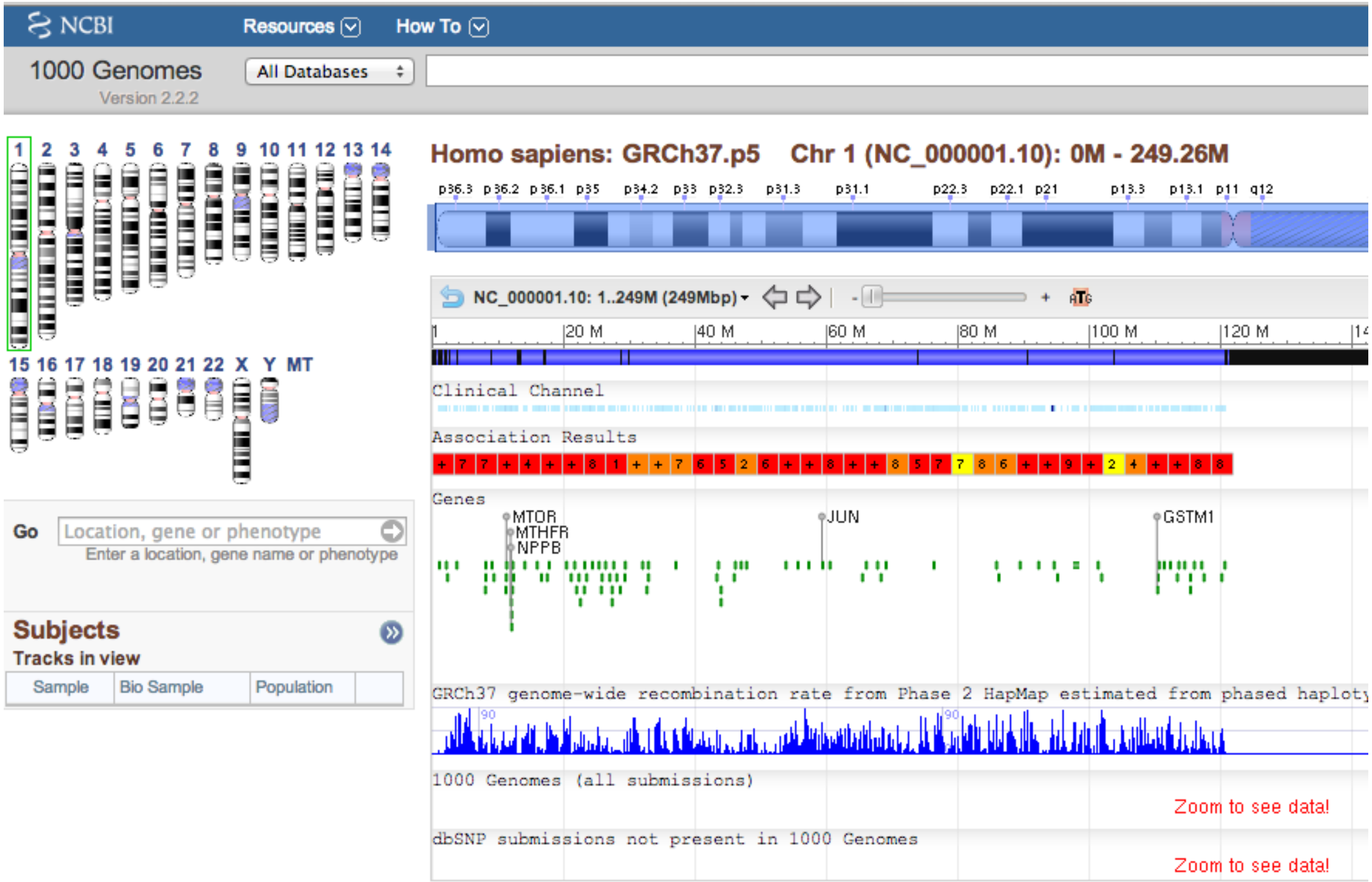
Sparse/dense, multi-dimensional arrays

Distributed storage, parallel processing

Excels at parallel sparse/dense linear algebra

ACID, data replication, versioned data

The NCBI 1K Genome Browser Runs on SciDB


NCBI
Resources
How To
1000 Genomes
Version 2.2.2
All Databases
1 2 3 4 5 6 7 8 9 10 11 12 13 14
15 16 17 18 19 20 21 22 X Y MT
Go
Location, gene or phenotype
Enter a location, gene name or phenotype
Subjects
Tracks in view
Sample
Bio Sample
Population
Homo sapiens: GRCh37.p5 Chr 1 (NC_000001.10): 0M - 249.26M
p36.3 p36.2 p36.1 p35 p34.2 p33 p32.3 p31.3 p31.1 p22.3 p22.1 p21 p13.3 p13.1 p11 q12
NC_000001.10: 1..249M (249Mbp)
1 20 M 40 M 60 M 80 M 100 M 120 M
Clinical Channel
Association Results
+ 7 7 + + + + 8 1 + + 7 6 5 2 6 + + 8 + + 8 5 7 7 8 6 + + 9 + 2 4 + + 8 8
Genes
MTOR
MTHFR
NPPB
JUN
GSTM1
GRCh37 genome-wide recombination rate from Phase 2 HapMap estimated from phased haploty
90
90
1000 Genomes (all submissions)
Zoom to see data!
dbSNP submissions not present in 1000 Genomes
Zoom to see data!

http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/

http://www.ncbi.nlm.nih.gov/variation/tools/get-rm/browse/

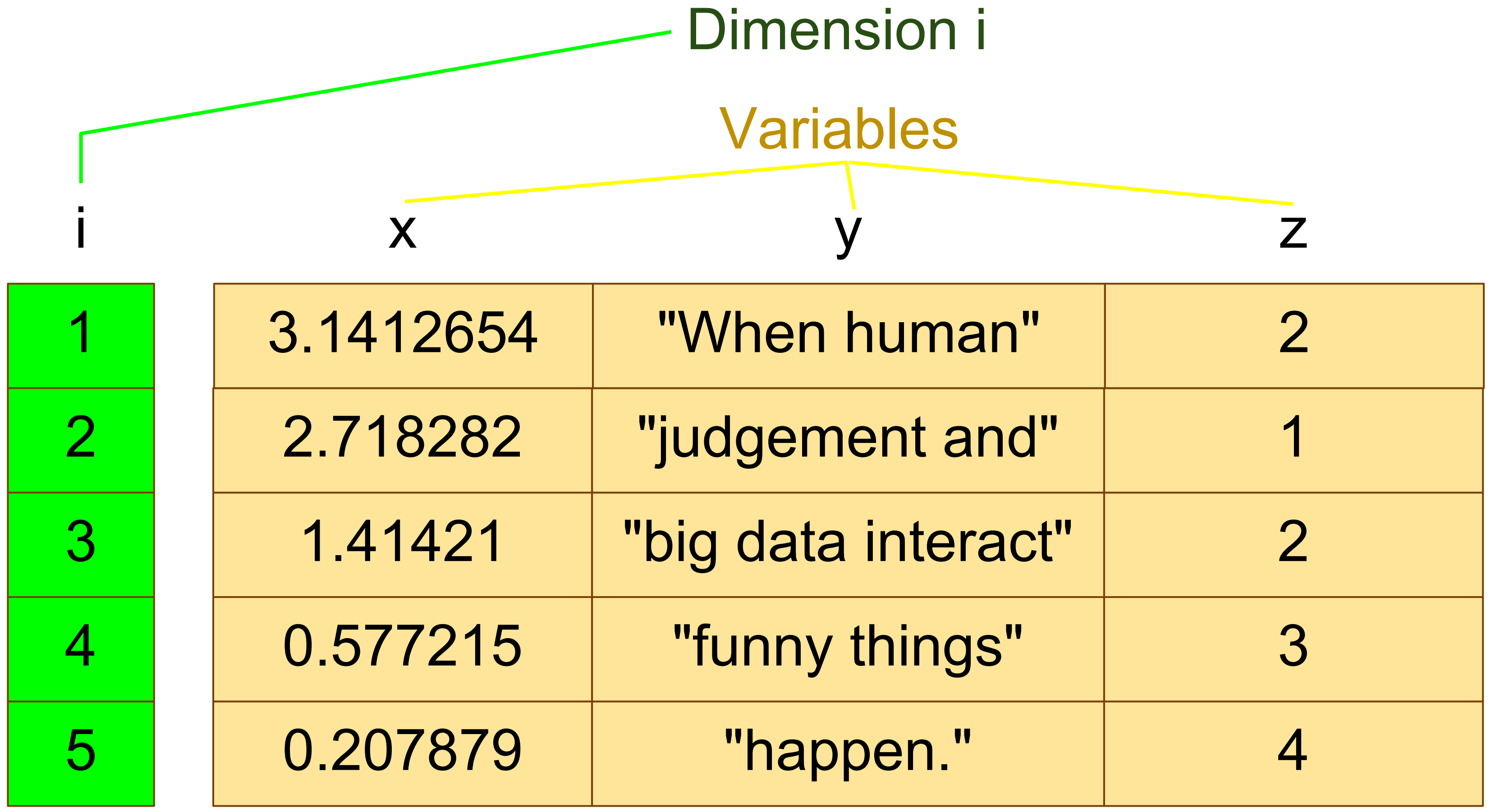
SciDB Arrays

Each cell in a SciDB array consists of a fixed number of typed values.

Here is an example cell:

x	y	z
3.141593	"When human"	2

Cells are ordered along coordinate axes. A 1-D array looks like an R data frame.

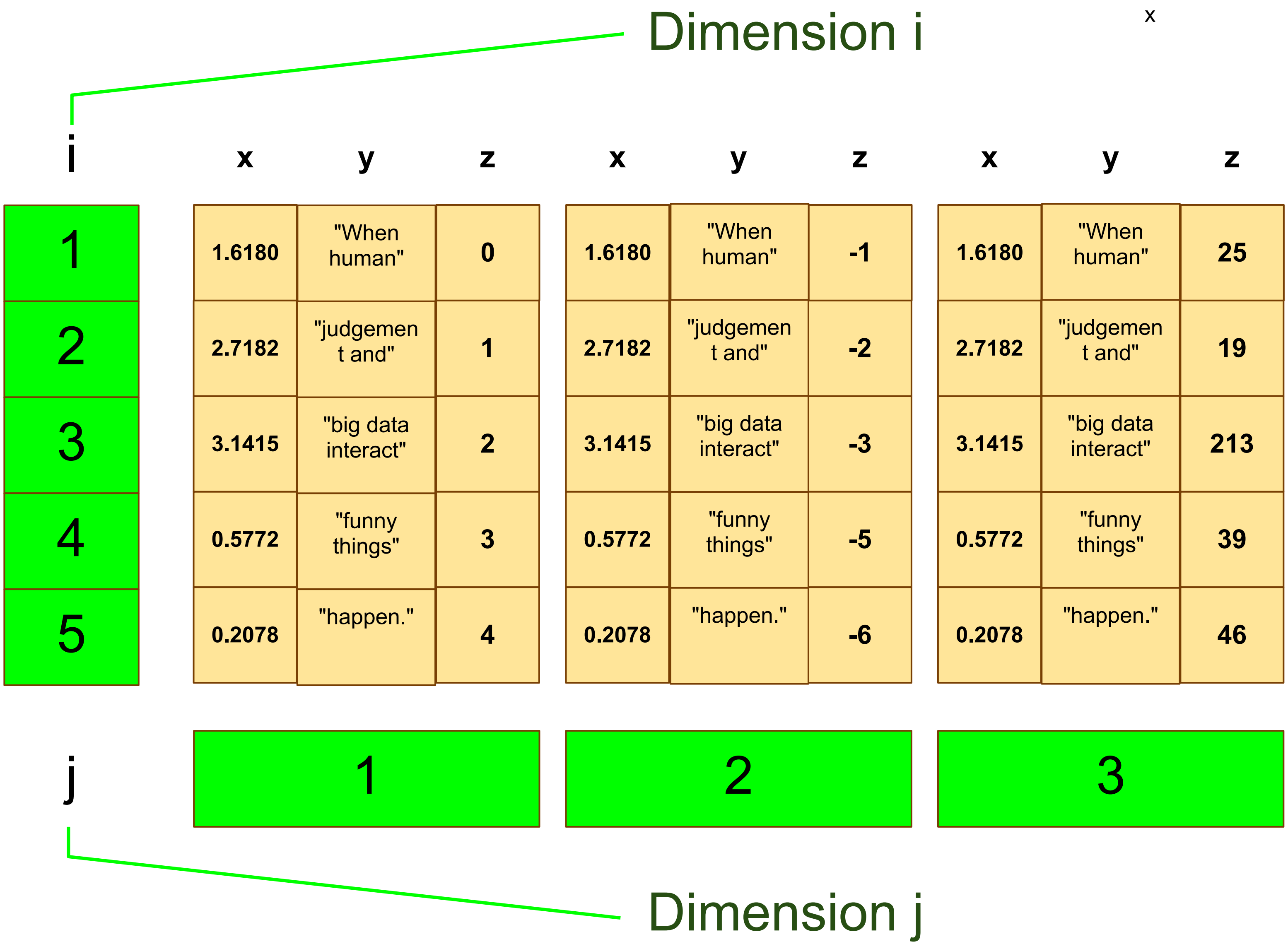


Dimension i

Variables

i	x	y	z
1	3.1412654	"When human"	2
2	2.718282	"judgement and"	1
3	1.41421	"big data interact"	2
4	0.577215	"funny things"	3
5	0.207879	"happen."	4

SciDB arrays can be multi-dimensional



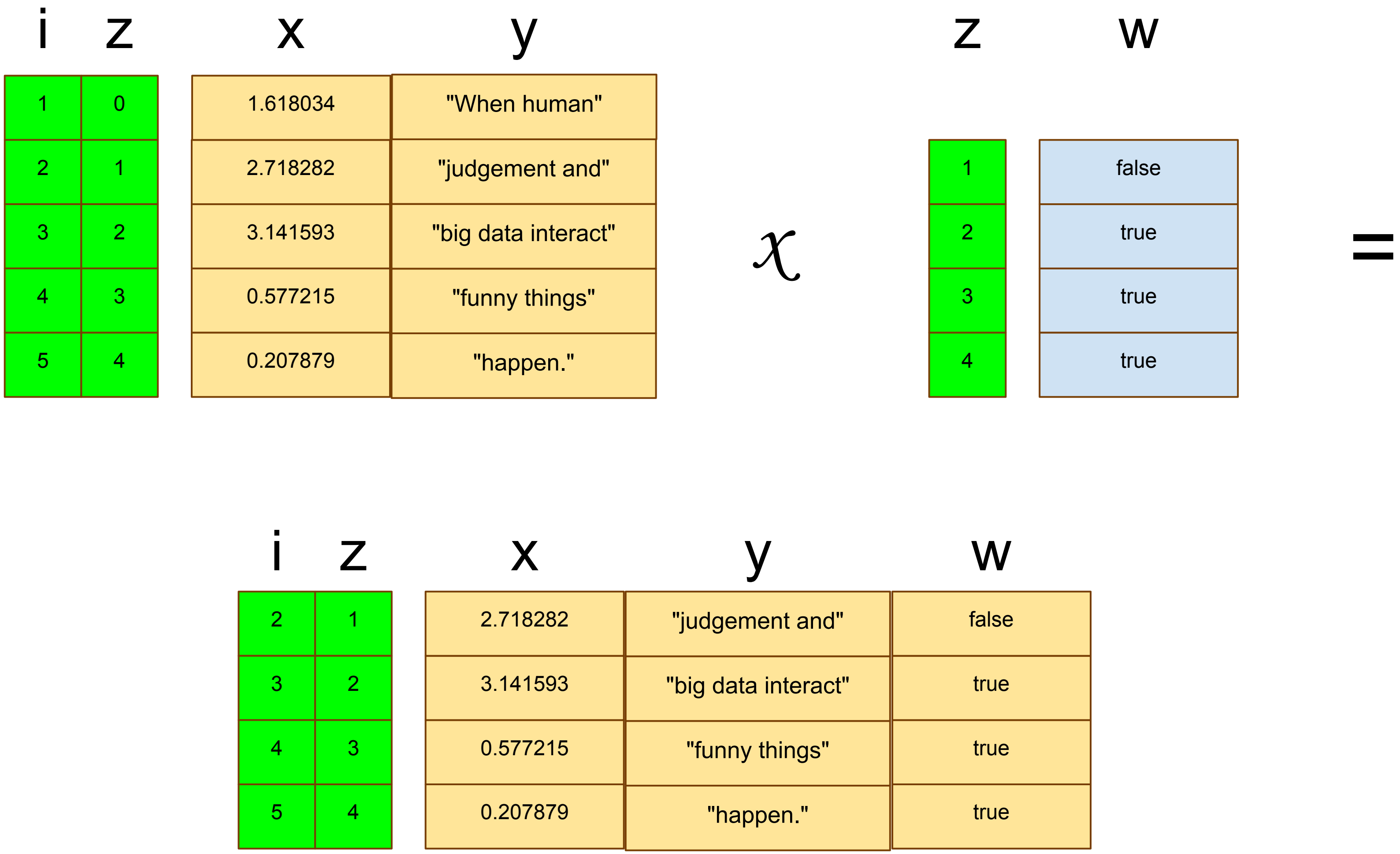
x

i	x	y	z	x	y	z	x	y	z
1	**1.6180**	"When human"	**0**	**1.6180**	"When human"	**-1**	**1.6180**	"When human"	**25**
2	**2.7182**	"judgement and"	**1**	**2.7182**	"judgement and"	**-2**	**2.7182**	"judgement and"	**19**
3	**3.1415**	"big data interact"	**2**	**3.1415**	"big data interact"	**-3**	**3.1415**	"big data interact"	**213**
4	**0.5772**	"funny things"	**3**	**0.5772**	"funny things"	**-5**	**0.5772**	"funny things"	**39**
5	**0.2078**	"happen."	**4**	**0.2078**	"happen."	**-6**	**0.2078**	"happen."	**46**
j	1			2			3		

Arrays can be sparse and values may be explicitly marked missing in several ways.

i	x	y	z
1	NA	"When human"	0
2			
3	Missing(1)	"big data interact"	Missing(7)
4	0.577215	"funny things"	3
5	0.207879	"happen."	4

Arrays can be joined along common dimensions (like R's *merge*):



i	z	x	y
1	0	1.618034	"When human"
2	1	2.718282	"judgement and"
3	2	3.141593	"big data interact"
4	3	0.577215	"funny things"
5	4	0.207879	"happen."

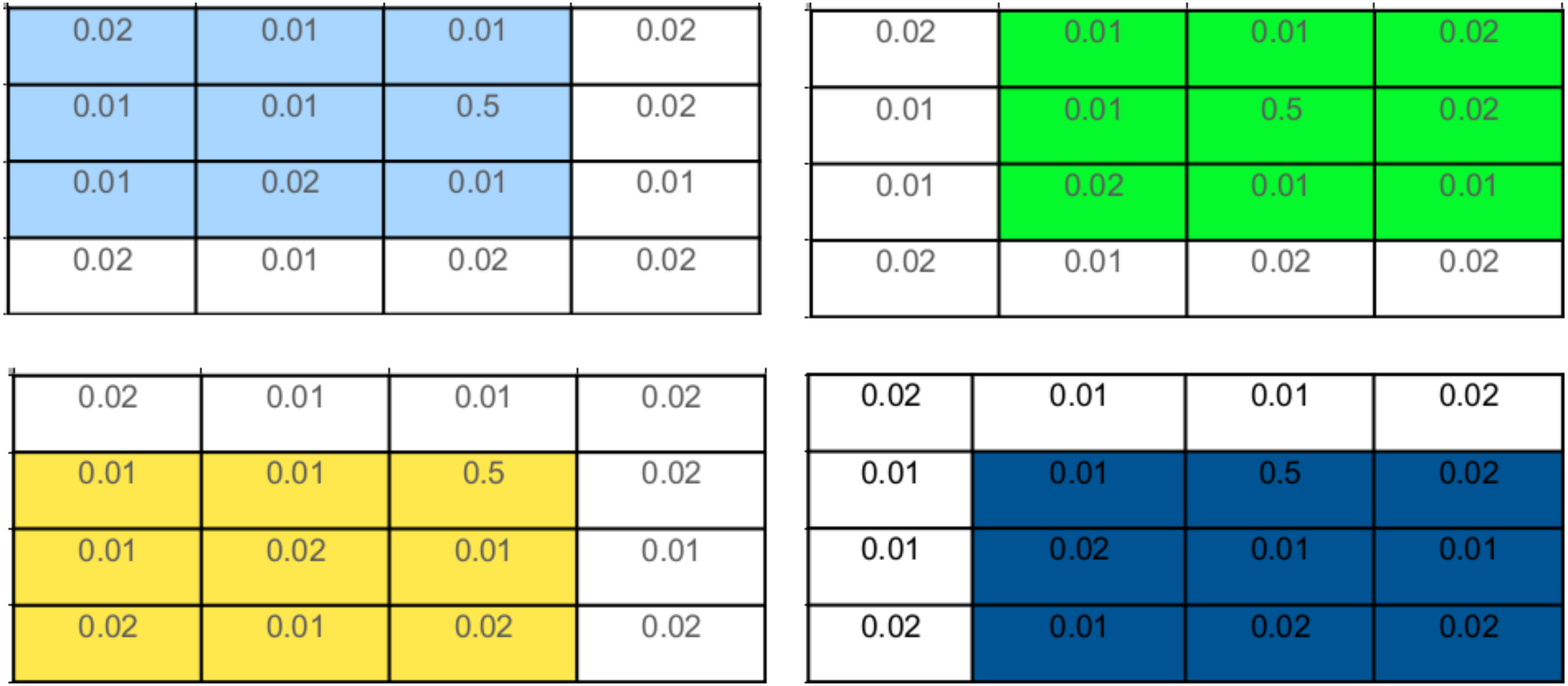
x

z	w
1	false
2	true
3	true
4	true

=

i	z	x	y	w
2	1	2.718282	"judgement and"	false
3	2	3.141593	"big data interact"	true
4	3	0.577215	"funny things"	true
5	4	0.207879	"happen."	true

SciDB array partitioning and overlap



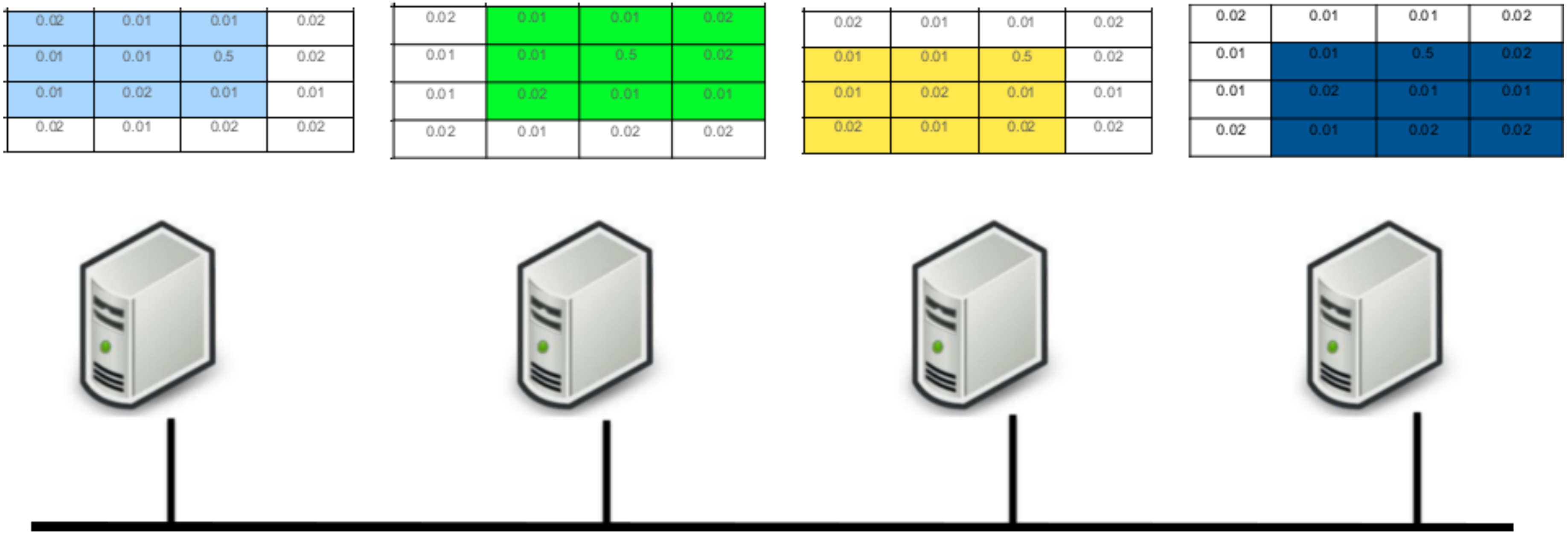
0.02	0.01	0.01	0.02
0.01	0.01	0.5	0.02
0.01	0.02	0.01	0.01
0.02	0.01	0.02	0.02

0.02	0.01	0.01	0.02
0.01	0.01	0.5	0.02
0.01	0.02	0.01	0.01
0.02	0.01	0.02	0.02

0.02	0.01	0.01	0.02
0.01	0.01	0.5	0.02
0.01	0.02	0.01	0.01
0.02	0.01	0.02	0.02

0.02	0.01	0.01	0.02
0.01	0.01	0.5	0.02
0.01	0.02	0.01	0.01
0.02	0.01	0.02	0.02

Array chunks are distributed



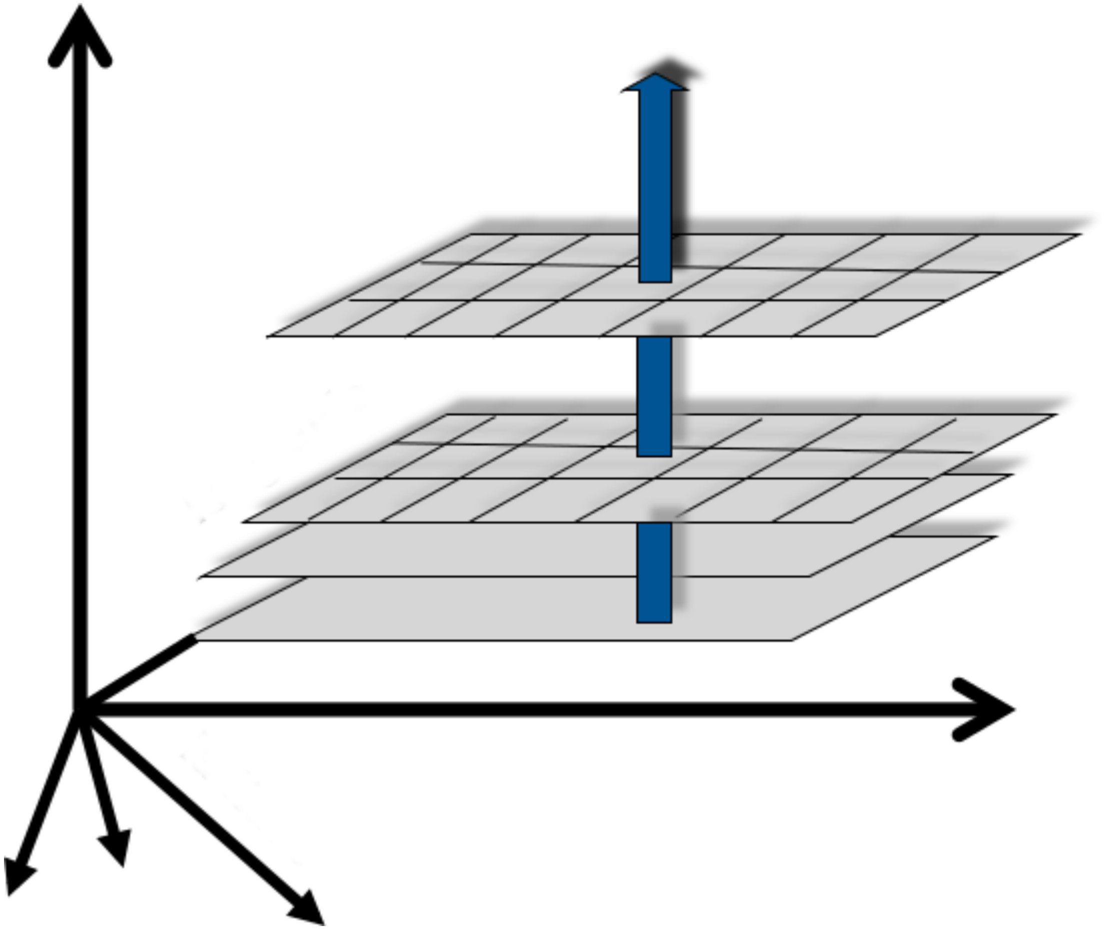
0.02	0.01	0.01	0.02
0.01	0.01	0.5	0.02
0.01	0.02	0.01	0.01
0.02	0.01	0.02	0.02

0.02	0.01	0.01	0.02
0.01	0.01	0.5	0.02
0.01	0.02	0.01	0.01
0.02	0.01	0.02	0.02

0.02	0.01	0.01	0.02
0.01	0.01	0.5	0.02
0.01	0.02	0.01	0.01
0.02	0.01	0.02	0.02

0.02	0.01	0.01	0.02
0.01	0.01	0.5	0.02
0.01	0.02	0.01	0.01
0.02	0.01	0.02	0.02

Regular chunk distribution across arrays = fast n-dimensional join/merge



Values can be **aggregated**, along dimensions optionally over **windows**

Functions can be **applied** over values in arrays

Arrays can be sparse

Linear algebra operations and matrix decompositions are available for matrices and vectors.

The scidb package for R

List/Dataframe-like

RObjectTables, g.data, filehash, ff, DBI and many database interfaces RPgSQL, RMySQL, ROracle, ...), Vertica/R, Netezza/R, rredis, **scidb**, RBerkeley, RCassandra, LaF, lazy.frames

Hadoop

rmr, HadoopStreaming, RHIPE

Array-like

ff, bigmemory, pbdR, **scidb,** Netezza

Other

rdsm, forthcoming from Simon, flexmem

The package defines two main ways to interact with SciDB:

1. Iterable data frame interface using SciDB query language directly

2. **N-dimensional sparse/dense array class for R backed by SciDB arrays**

```
library("scidb")
scidbconnect(host="localhost")


# An example reference to a SciDB matrix:
A <- scidb("A")
dim(A)
[1] 50000 50000
```

Subarrays return new SciDB array objects

```
A[c(0,49000,171), 5:8]
```

```
Reference to a 3x4 SciDB array
```

Use `[]` to materialize data to R

```
A[c(0,49000,171), 5:8][]
```

```
          [,1]       [,2]       [,3]       [,4]
[1,]  0.9820799 -0.4563357 -1.2947495 -0.8085465
[2,] -1.5090126  0.1547963 -0.2435732 -0.1836875
[3,]  1.3296710 -1.5006536 -0.5980172  0.3752186
```

Arithmetic

```
X <- A %*% A[,1:5]
dim(X)
```

```
[1] 50000     5
```

Mixed **SciDB** and **R** object arithmetic

```
Z <- A[c(0,49000,171), 5:7]

(0.5*(Z + t(Z)) %*% rnorm(3)[, drop=FALSE]

         [,1]
[1,]  3.707263
[2,] -2.833560
[3,]  3.518370
```

Basic aggregation (scidbdf class)

```
A <- as.scidb(iris)
Warning message:
In df2scidb :Attribute names have been changed

aggregate(A, Petal_Length ~ Species, "avg
(Petal_Length) as mean")

     Species  mean
1     setosa 1.462
2 versicolor 4.260
3  virginica 5.552
```

SVD and principal components

```
S <- svd(A, nu=3, nv=3)
dim(S)
[1]     4 50000 50000

# Result is a 3-D array containing U,
  S (sparse), and V
```

It is sometimes possible to use SciDB arrays in R packages with little modification.

```
library("biclust")
library("s4vd")
data(lung)
A <- lung
x <- biclust(A, method=BCssvd, K=1)

# Now with SciDB arrays:
library("s4vdp4")
X <- as.scidb(A)
x1 <- biclust(X, method=BCssvd, K=1)

# Compare the results:
sqrt( x@info$res[[1]]$u - x1@info$res[[1]]$u) )
            [,1]
[1,] 5.202109e-16
```

paradigm4

data-driven discovery

Virtual machines and EC2 images ready to roll (including Rstudio) available from:

www.scidb.org

R package on CRAN and development version at:

github.com/Paradigm4