

Lecture: Data and Data types

DATA ANALYSIS & VISUALIZATION
FALL 2021

Dr. Muhammad Faisal Cheema
FAST-NU

GOALS FOR TODAY

- Learn what are data types and dataset types
- Learn what are attribute types
- Learn what are dimensions and measures

Data

- What is this data?

R011	42ND STREET & 8TH AVENUE	00228985	00008471	00000441	00001455	00000134	00033341	00071255
R170	14TH STREET-UNION SQUARE	00224603	00011051	00000827	00003026	00000660	00089367	00199841
R046	42ND STREET & GRAND CENTRAL	00207758	00007908	00000323	00001183	00003001	00040759	00096613

- Semantics: real-world meaning of the data
- Type: structural or mathematical interpretation
- Both often require metadata
 - Sometimes we can infer some of this information
 - Line between data and metadata isn't always clear

Data

	REMOTE	STATION	FF	SEN/DIS	7-D AFAS UNL	D AFAS/RMF I	JOINT RR TKT	7-D UNL	30-D UNL	
1	R011	42ND STREET & 8TH AVENUE	00228985	00008471	00000441	00001455	00000134	00033341	00071255	
2	R170	14TH STREET-UNION SQUARE	00224603	00011051	00000827	00003026	00000660	00089367	00199841	
3	R046	42ND STREET & GRAND CENTRAL	00207758	00007908	00000323	00001183	00003001	00040759	00096613	
4	R012	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00001279	00003622	00035527	00067483	
5	R293	34TH STREET - PENN STATION	00168768	00006155	00000523	00001065	00005031	00030645	00054376	
6	R033	42ND STREET/TIMES SQUARE	00159382	00005945	00000378	00001205	00000690	00058931	00078644	
7	R022	34TH STREET & 6TH AVENUE	00156008	00006276	00000487	00001543	00000712	00058910	00110466	
8	R084	59TH STREET/COLUMBUS CIRCLE	00155262	00009484	00000589	00002071	00000542	00053397	00113966	
9	R020	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000384	00001159	00000723	00037978	00090745	
10	R179	86TH STREET-LEXINGTON AVE	00142169	00010367	00000470	00001839	00000271	00050328	00125250	
11	R023	34TH STREET & 6TH AVENUE	00134052	00005005	00000348	00001112	00000649	00031531	00075040	
12	R029	PARK PLACE	00121614	00004311	00000287	00000931	00000792	00025404	00065362	
13	R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000704	00001241	00022808	00068216	
14	R031	34TH STREET & 7TH AVENUE	00095076	00003990	00000232	00000727	00001459	00024284	00038671	
15	R017	LEXINGTON AVENUE	00094655	00004688	00000190	00000833	00000754	00020018	00055066	
16	R175	8TH AVENUE-14TH STREET	00094313	00003907	00000286	00001144	00000256	00038272	00074661	
17	R057	BARCLAYS CENTER	00093804	00004204	00000454	00001386	00001491	00039113	00068119	
18	R138	WEST 4TH ST-WASHINGTON SQ	00093562	00004677	00000251	00000965	00000127	00031628	00074458	

Data Types

TYPE = structural or mathematical interpretation of the data

→ Data Types

- Items → Attributes → Links → Positions → Grids
- (row, node)* *(variable,
data dimension)* *(relationship)* *(spatial location)* *(sampling)*

Data Types

- Items
 - An item is an individual discrete entity
 - e.g. row in a table, node in a network
- Attributes
 - An **attribute** is some specific property that can be measured, observed, or logged
 - a.k.a. variable, (data) dimension

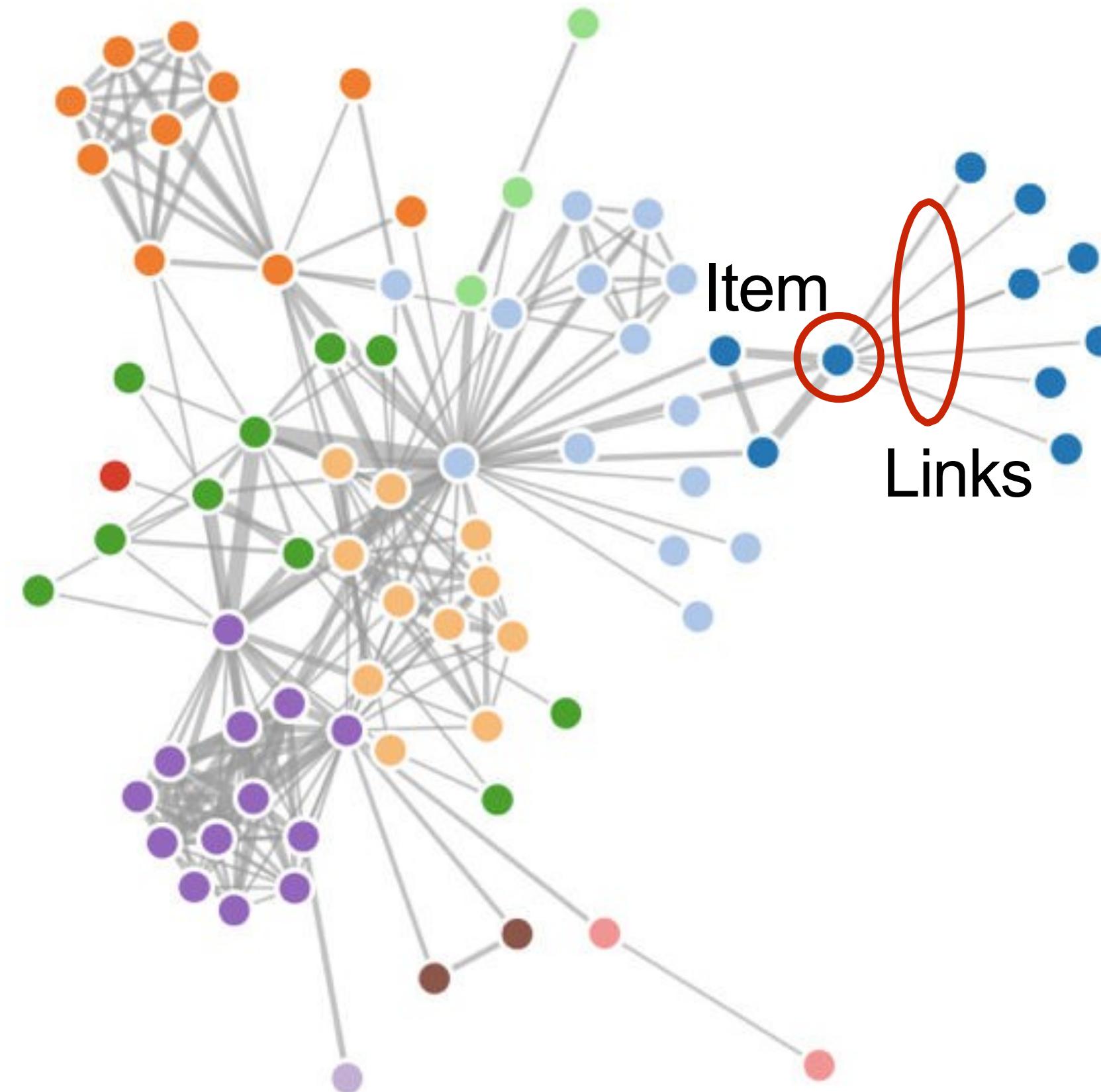
Items & Attributes

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	5	4-Not Specified	Small Pack	0.44	6/6/05
69	5	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Data Types

- Nodes
 - Synonym for item but in the context of networks (graphs)
- Links
 - A link is a relation between two items
 - e.g. social network friends, computer network links

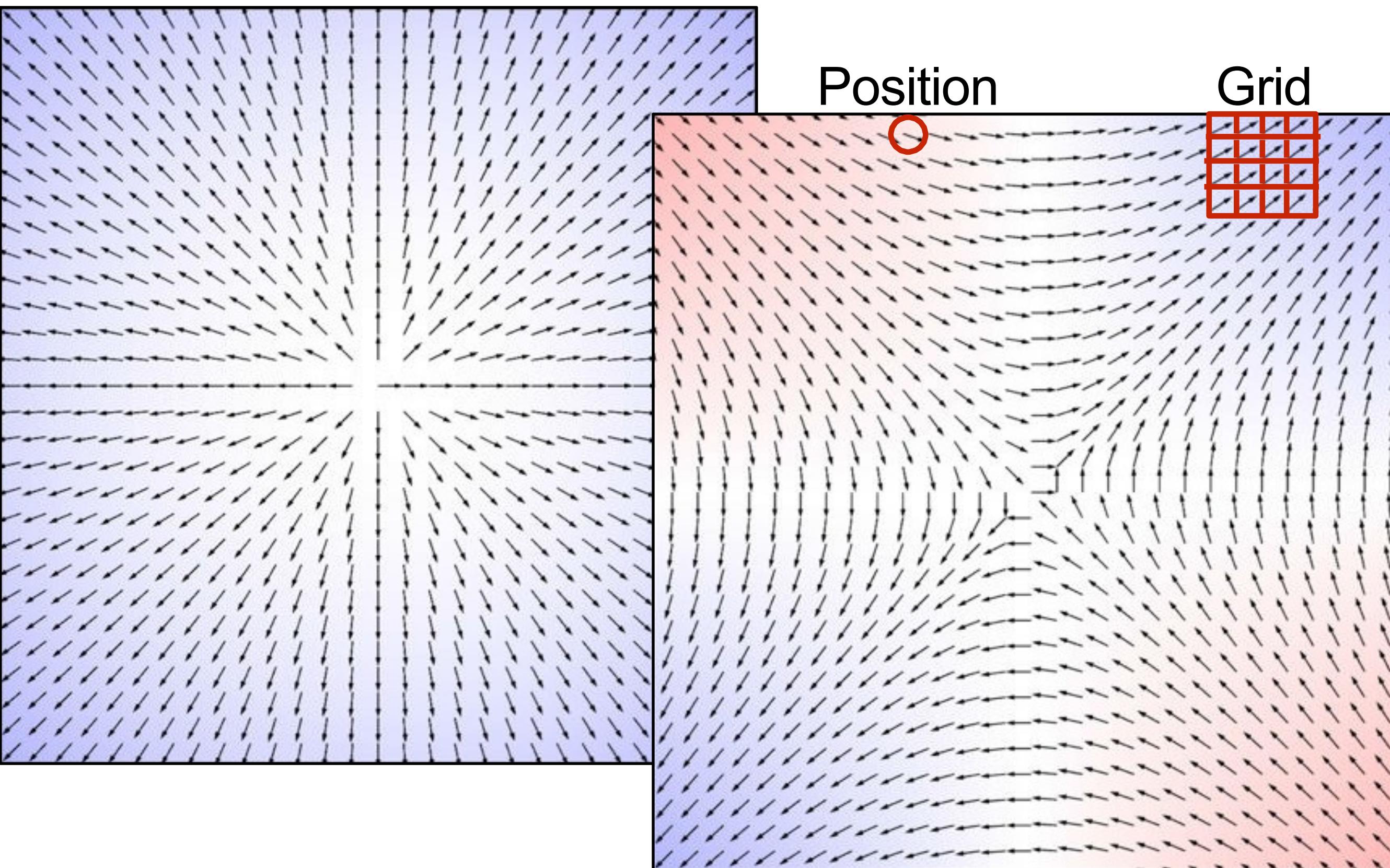
Items & Links



Data Types

- Positions:
 - A position is a location in space (usually 2D or 3D)
 - May be subject to projections
 - e.g. cities on a map, a sampled region in an CT scan
- Grids:
 - A grid specifies how data is sampled both geometrically and topologically
 - e.g. how CT scan data is stored

Positions and Grids

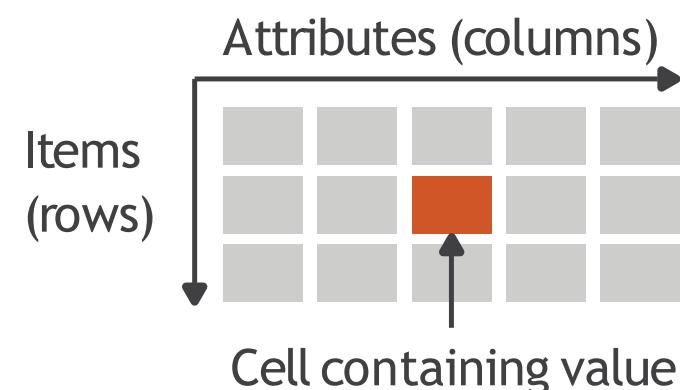


Data Types

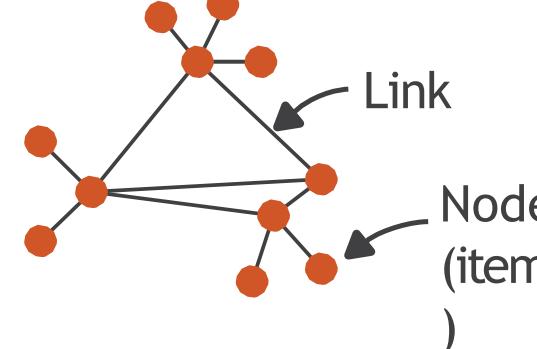
DATASET = collection of information that is the target of analysis

→ Dataset Types

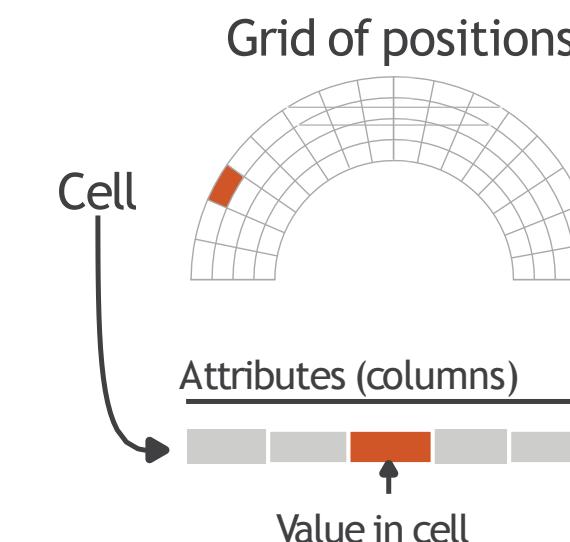
→ Tables



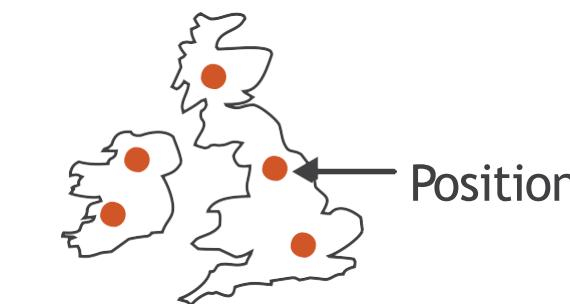
→ Networks



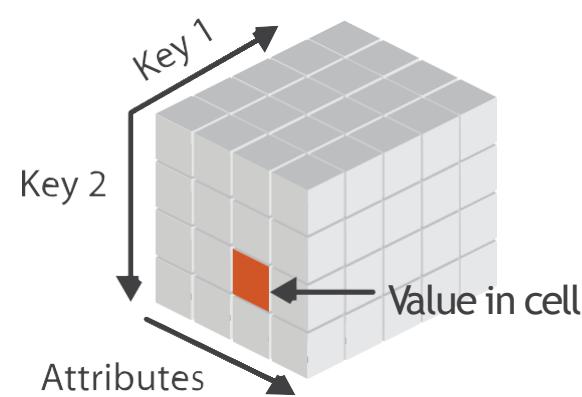
→ Fields (Continuous)



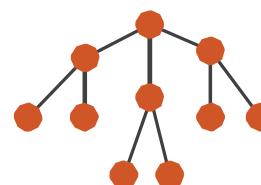
→ Geometry (Spatial)



→ Multidimensional Table



→ Trees



Data Types

DATASET = collection of information that is the target of analysis

→ Data and Dataset Types

Tables

Networks &
Trees

Fields

Geometry

Clusters,
Sets, Lists

Items

Items (nodes)

Grids

Items

Items

Attributes

Links

Positions

Positions

Attributes

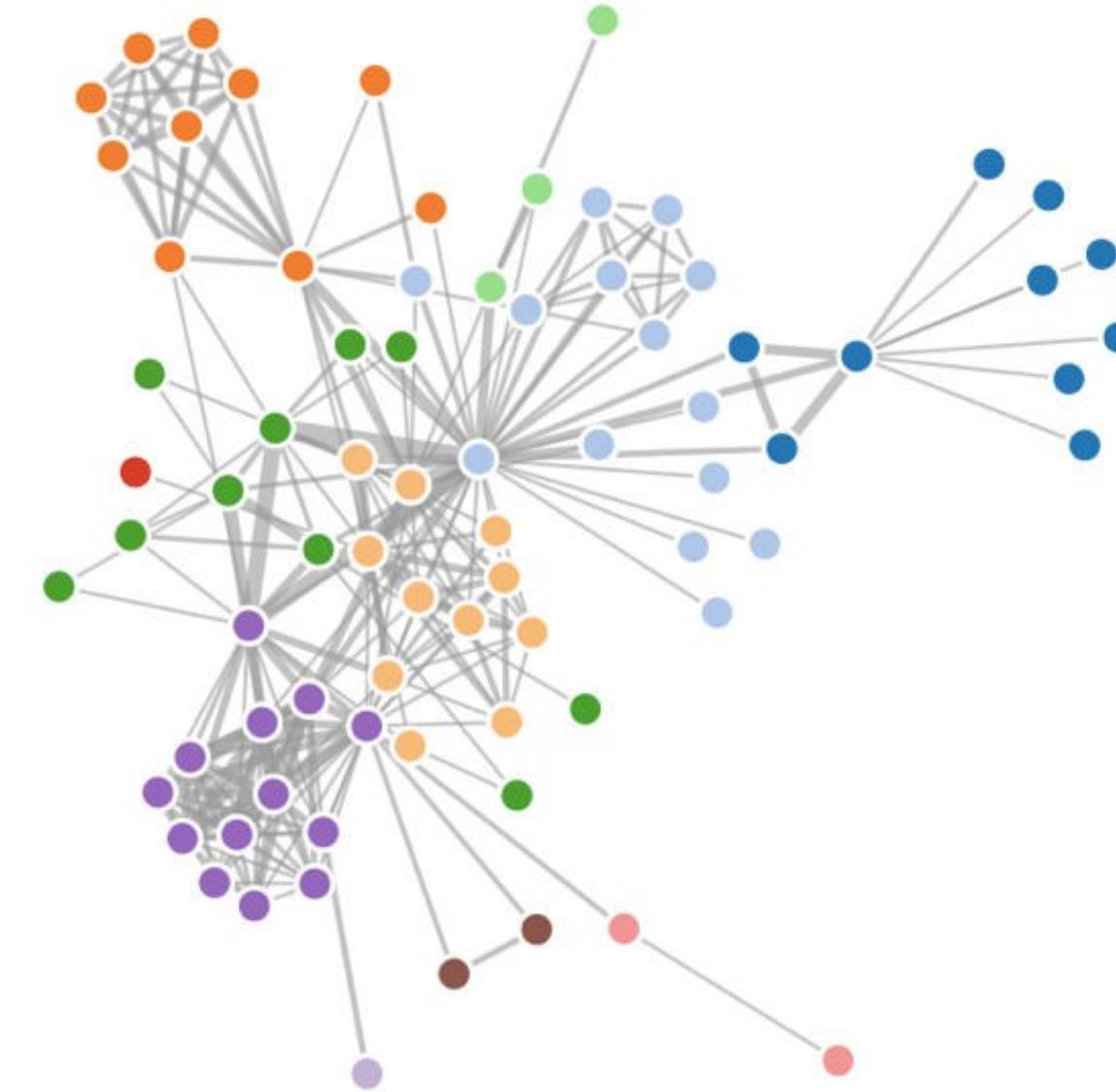
Attributes

Tables

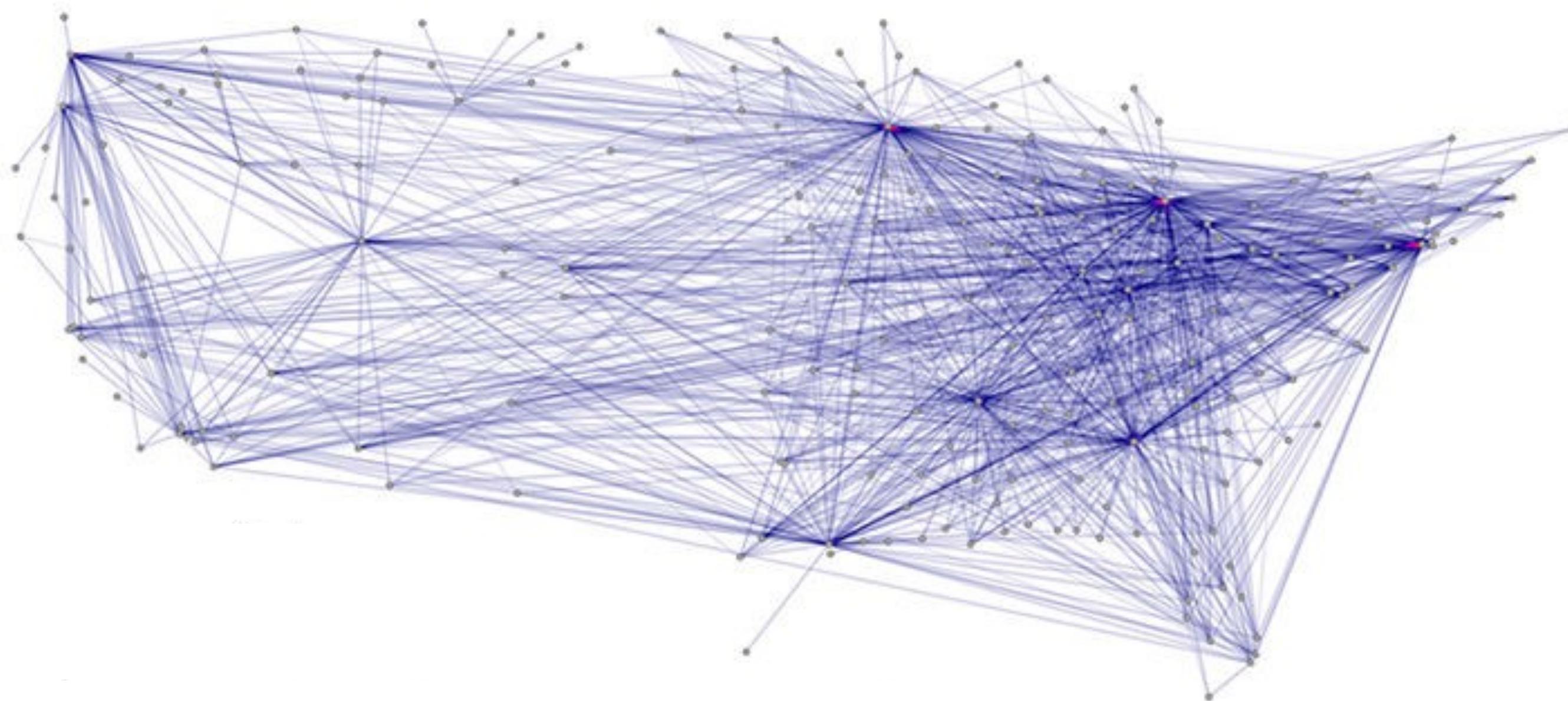
A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	5	4-Not Specified	Small Pack	0.44	6/6/05
69	5	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Networks

- Why networks instead of graphs?
- Tables can represent networks
 - Many-many relationships
 - Also can be stored as specific graph databases or files



Networks



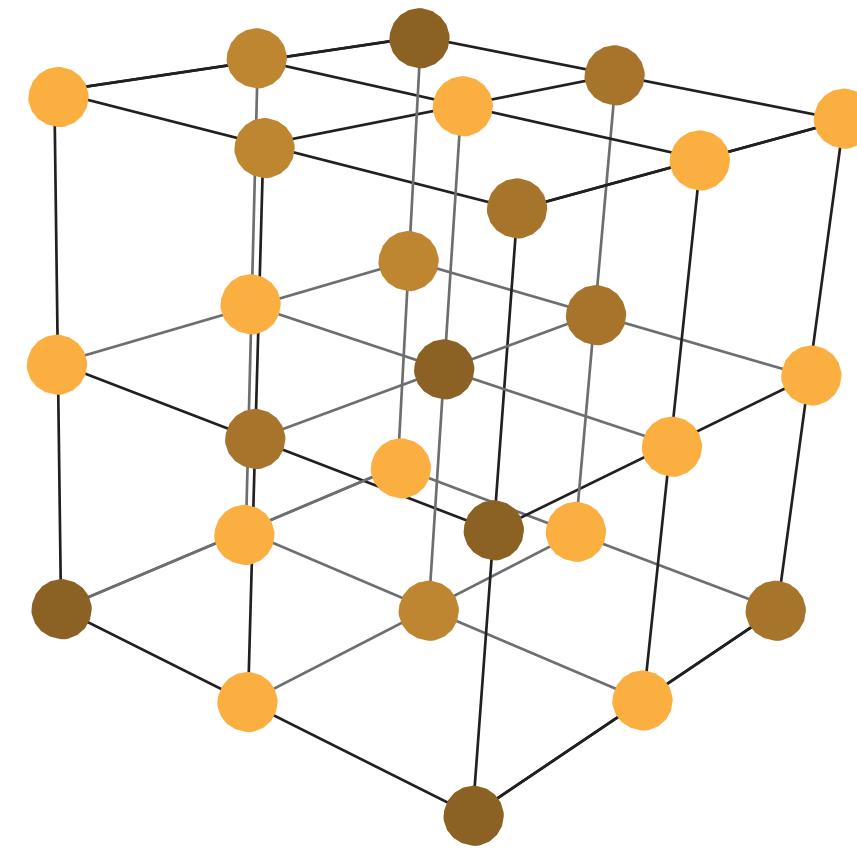
[Holten & van Wijk, 2009]

Networks

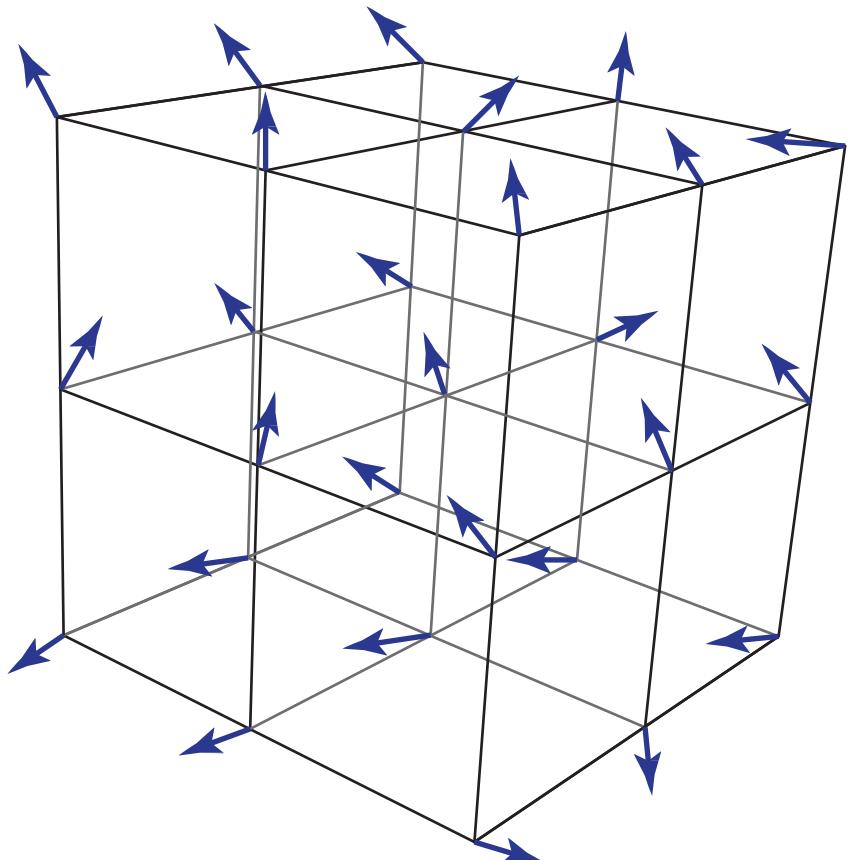


[Holten & van Wijk, 2009]

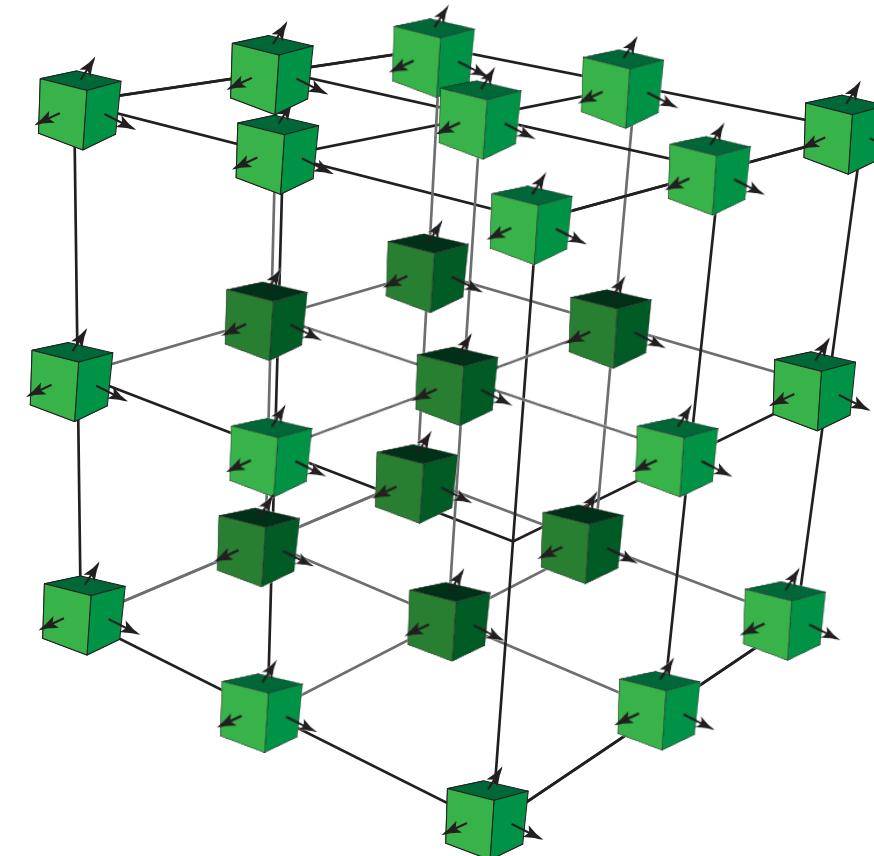
Fields



Scalar Fields



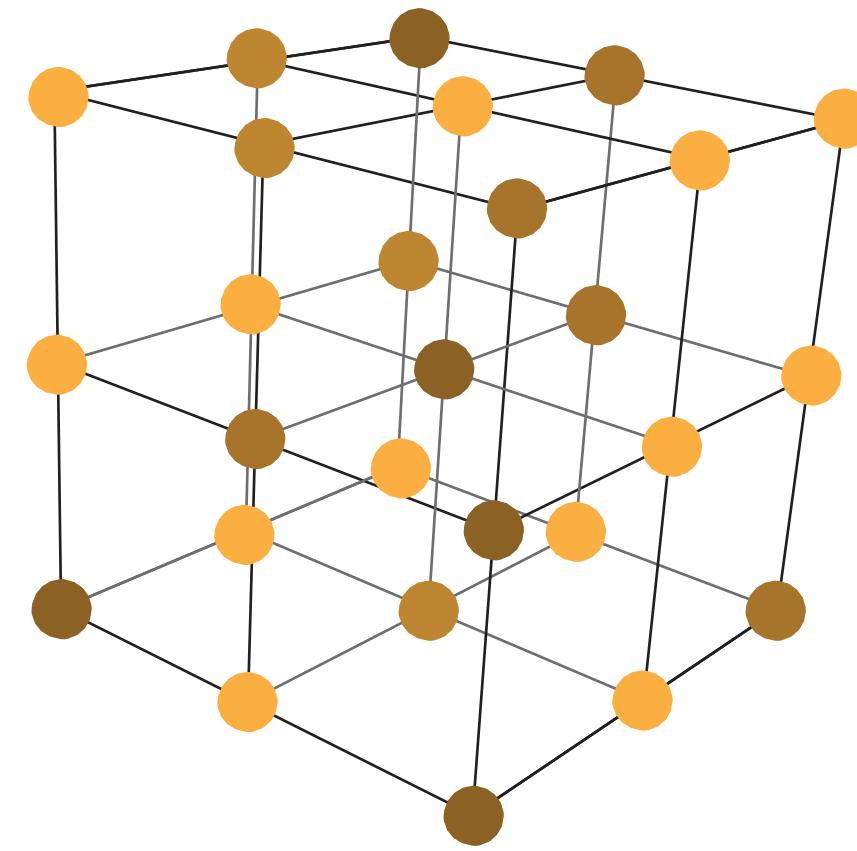
Vector Fields



Tensor Fields

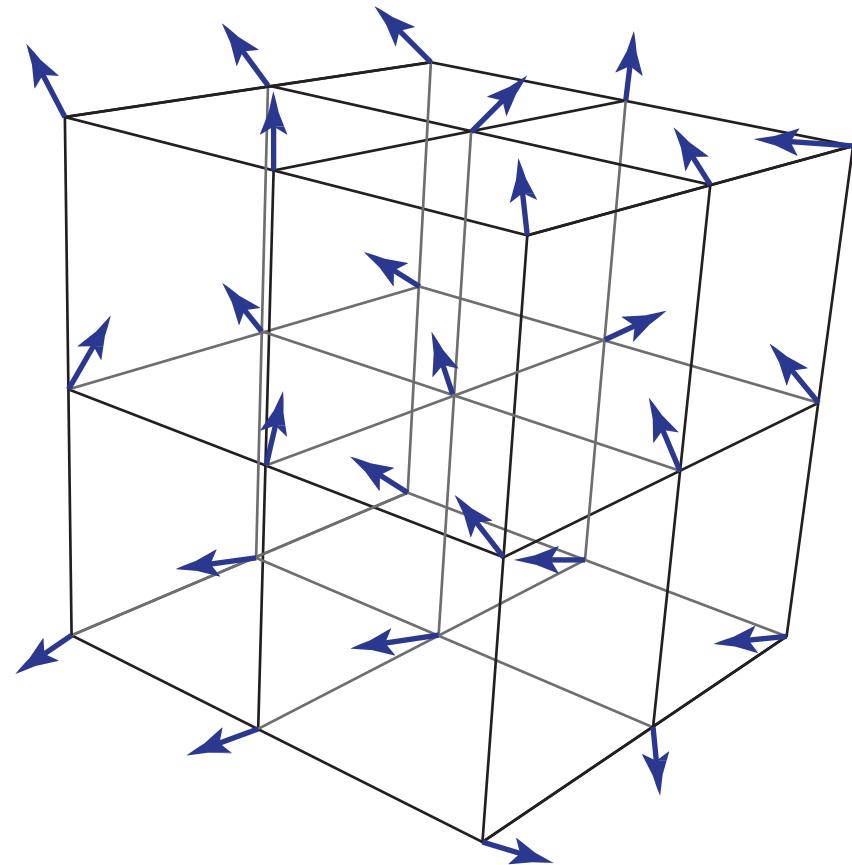
Each point in space has an associated...

Fields



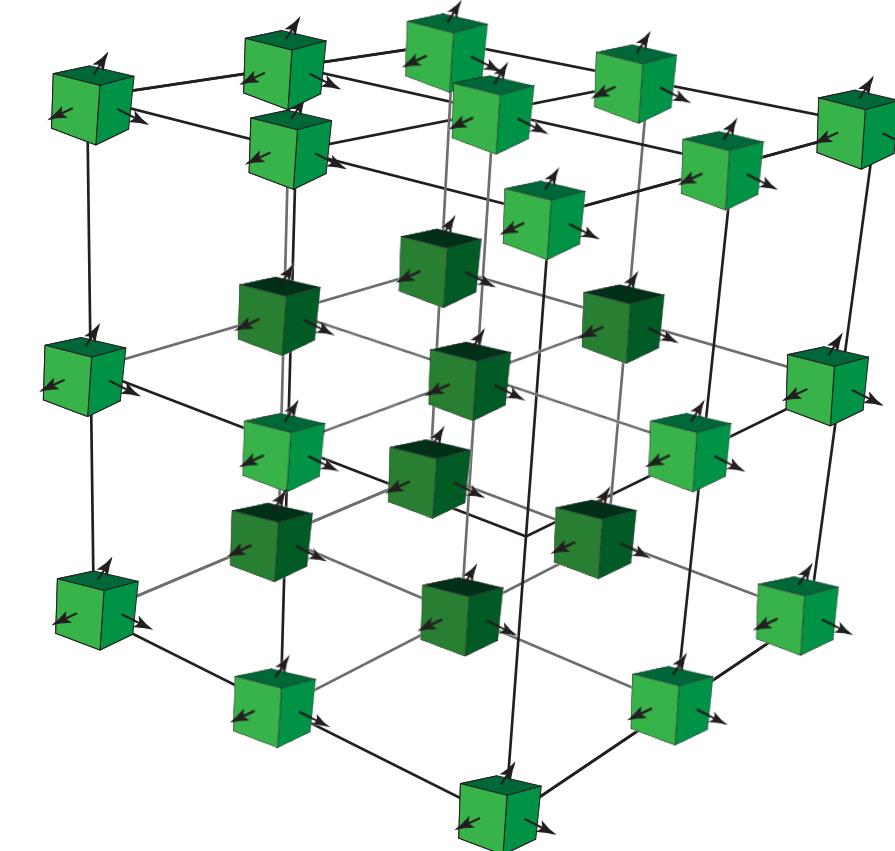
Scalar Fields

(Order-0 Tensor Fields)



Vector Fields

(Order-1 Tensor Fields)



Tensor Fields

(Order-2+)

Each point in space has an associated...

0

Scalar

v_0

v_1

v_2

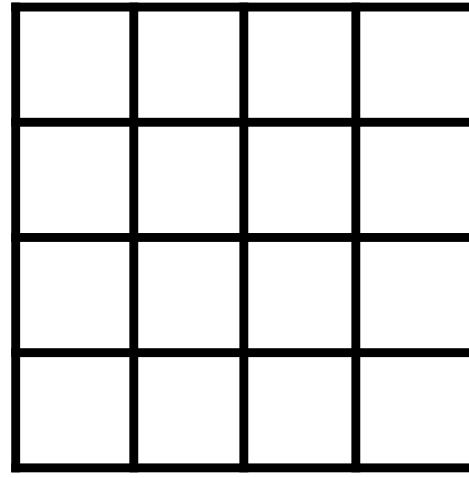
Vector

σ_{00} σ_{01} σ_{02}
 σ_{10} σ_{11} σ_{12}
 σ_{20} σ_{21} σ_{22}

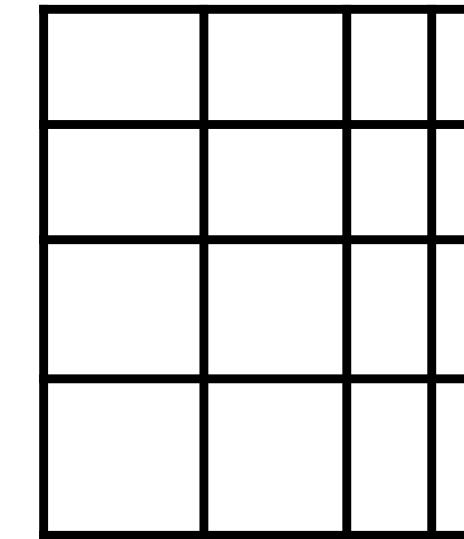
Tensor

Fields

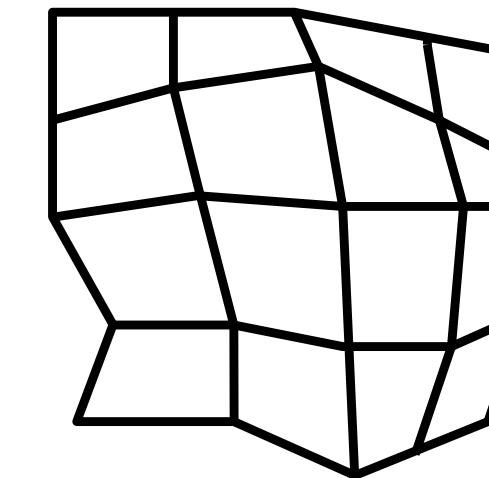
- Difference between continuous and discrete values
- Examples: temperature, pressure, density
- Grids necessary to sample continuous data:



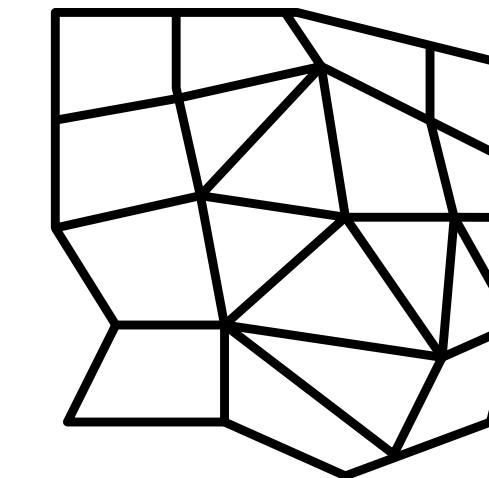
uniform



rectilinear



structured

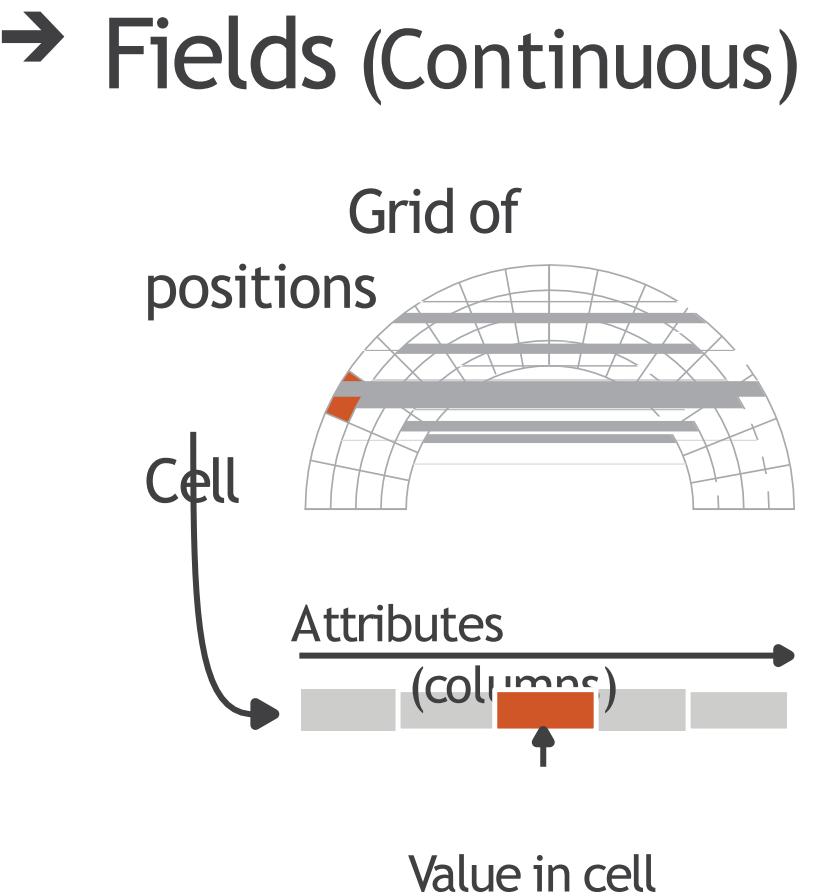


unstructured

[Weiskopf, Machiraju, Möller]

- Interpolation: “how to show values between the sampled points in ways that do not mislead”

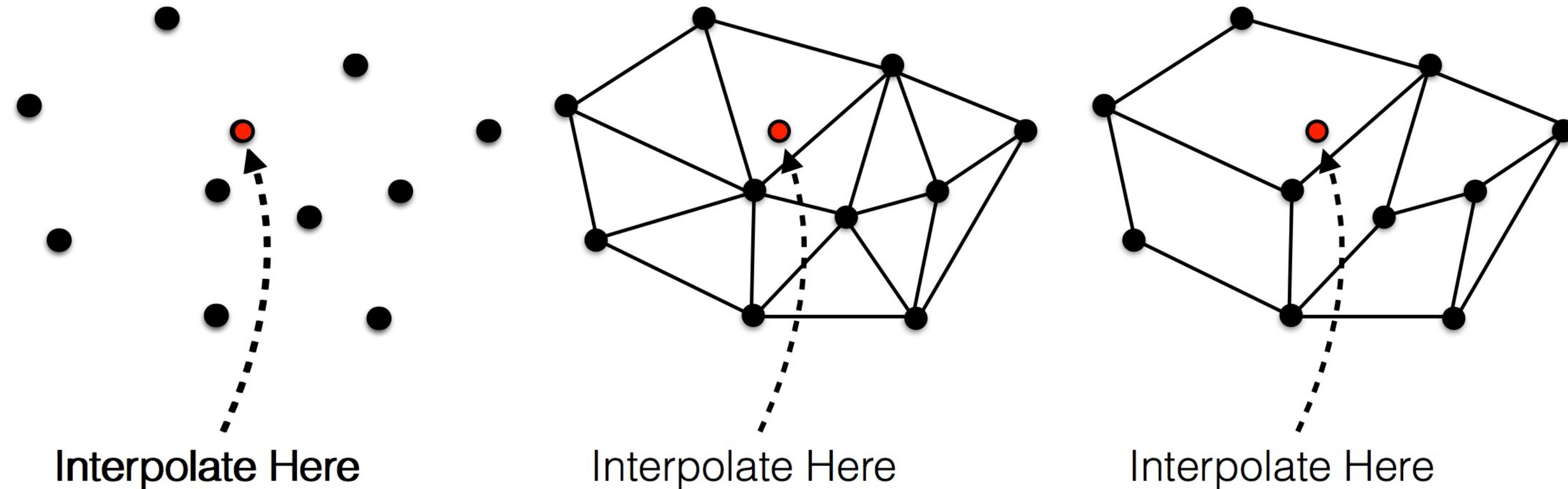
grid choices impact how continuous data is interpreted



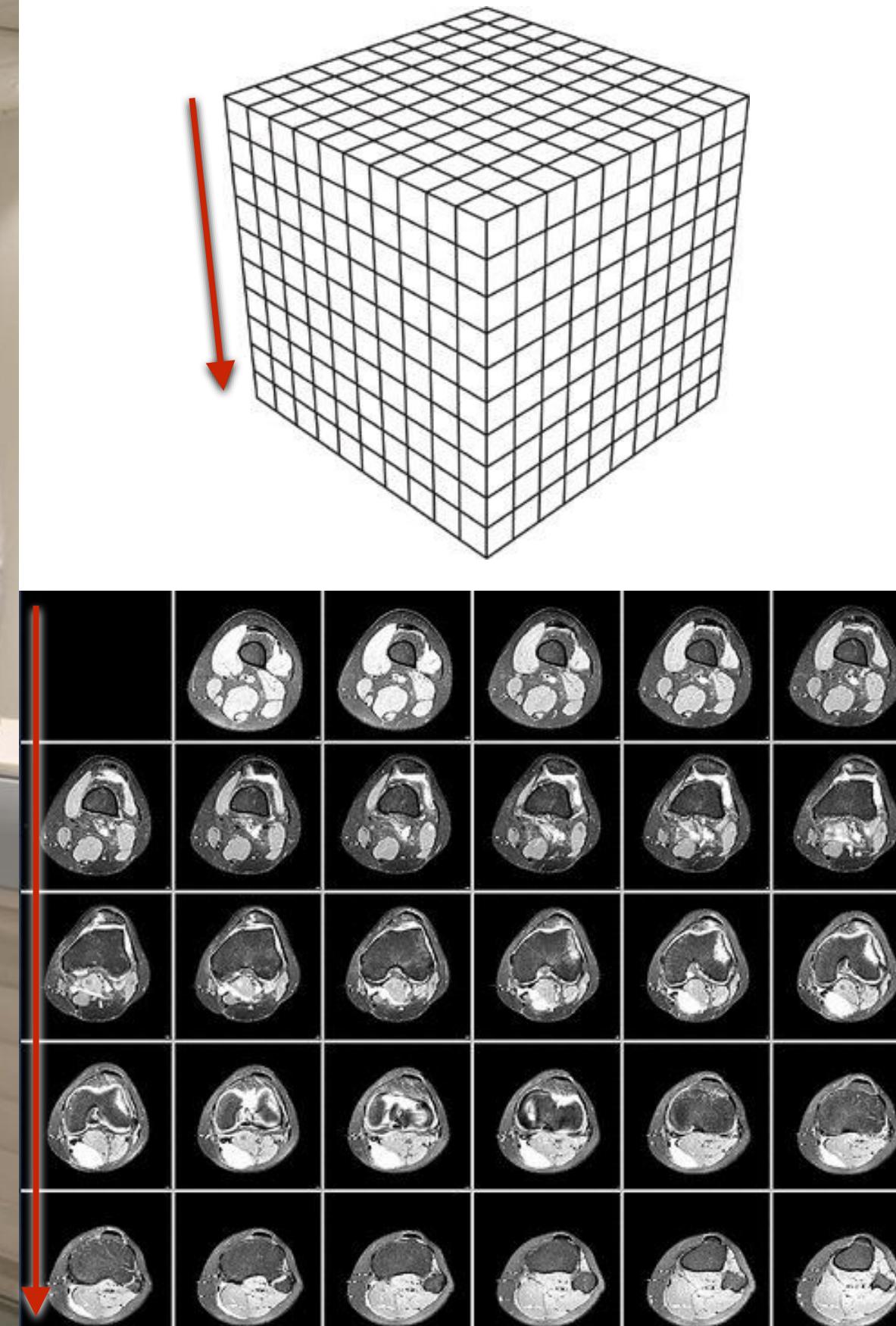
two key considerations:

sampling, or the choice of where attributes are measured

interpolation, or how to model the attributes in the rest of space

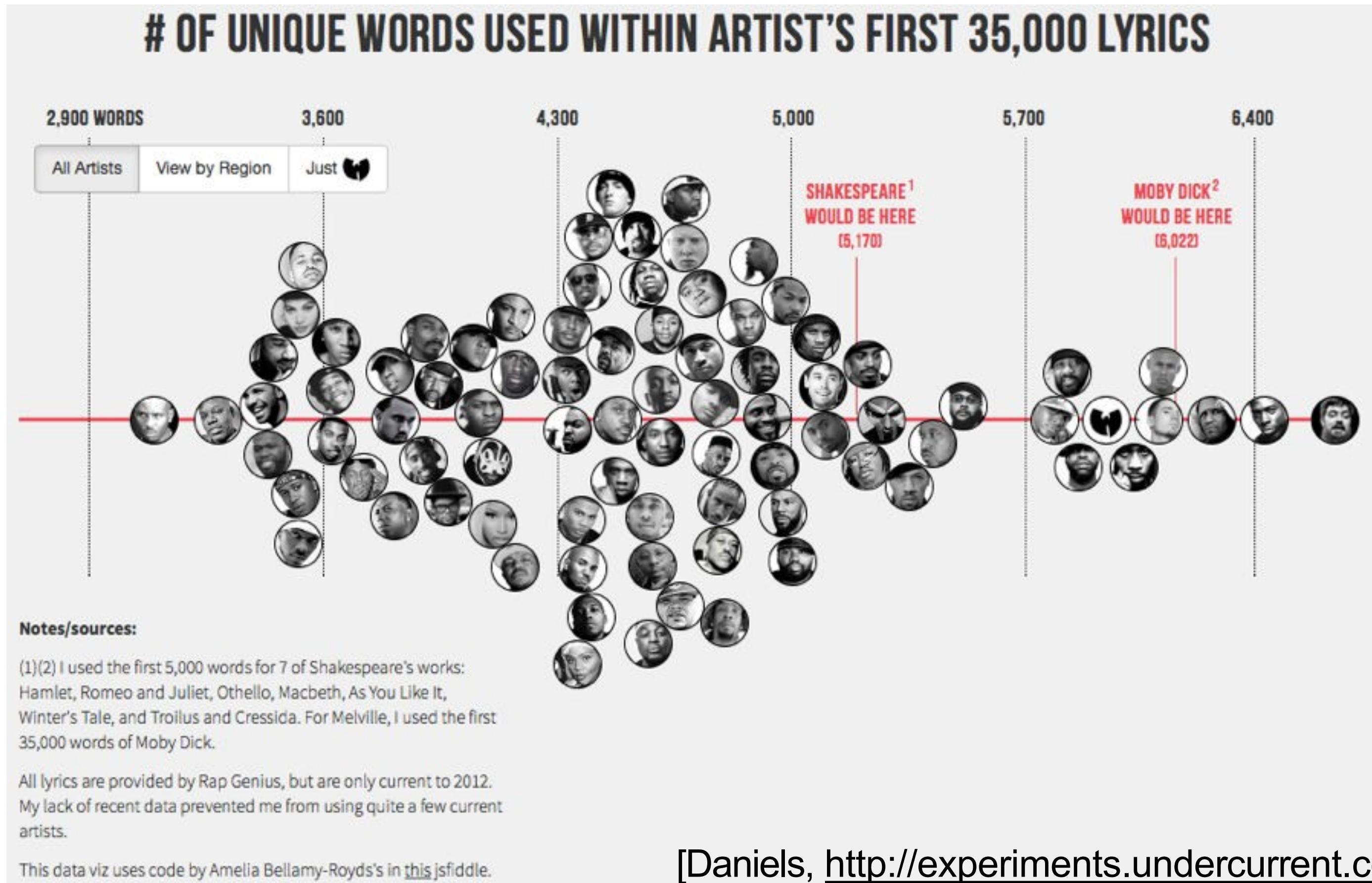


Spatial Data Example: MRI



[slide via Levine, 2014]

Sets & Lists



Data Types

DATASET = collection of information that is the target of analysis

→ Data and Dataset Types

Tables

Items

Attributes

Networks &
Trees

Items (nodes)

Links

Attributes

Fields

Grids

Positions

Attributes

Geometry

Items

Positions

Clusters,
Sets, Lists

Items

Attribute Types

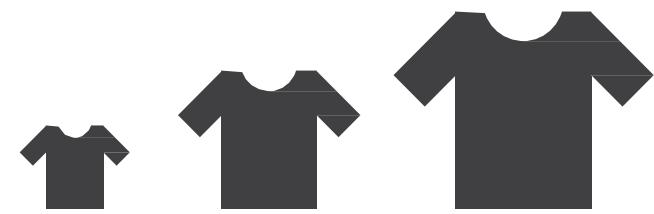
Attribute Types

→ Categorical (*nominal*) → Ordered



e.g., fruit (apple, pear, etc.)
colleges (CCIS, CAMD, etc.)

→ *Ordinal (ordered)*



e.g., months (J,F,M, etc.)
sizes (xs, s, m, l, xl)

→ Quantitative (*continuous*)



e.g., lengths (1", 2.5", 5")
population

Note: time and temperature
are quantitative

Ordering Direction

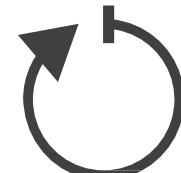
→ Sequential



→ Diverging



→ Cyclic



A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

quantitative
ordinal
categorical

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08
- - -	- - -	- - -	- - -	- - -	- - -

quantitative
ordinal
categorical

Categorial, Ordinal, and Quantitative

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

quantitative
ordinal
categorical

Categorial, Ordinal, and Quantitative

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08
195	4/5/08	3-Medium		0.6	4/7/08

quantitative
ordinal
categorical

Nominal, Ordinal and Quantitative

N - Nominal (labels)

- Fruits: Apples, oranges, ...

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (Location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
- Like a geometric point. Cannot compare directly Only
• differences (i.e. intervals) may be compared

Q - Ratio (zero fixed)

- Physical measurement: Length, Mass, Temp, ...Counts
- and amounts
- Like a geometric vector, origin is meaningful

Nominal, Ordinal and Quantitative

N - Nominal (labels)

- Operations: $=, \neq$

O - Ordered

- Operations: $=, \neq, <, >$

Q - Interval (Location of zero arbitrary)

- Operations: $=, \neq, <, >, -$

Can measure distances or spans

Q - Ratio (zero fixed)

- Operations: $=, \neq, <, >, -, \%$

Can measure ratios or proportions

S. S. Stevens, On the theory of scales of measurements, 1946

From Data Model to N, O, Q

Data Model

32.5, 54.0, -17.3, ...

Floating point numbers

Conceptual Model

Temperature (°C)

Data Type

Burned vs. Not-Burned (N)

Hot, Warm, Cold (O)

Temperature Value (Q)

Dimensions & Measures

Dimensions (~ independent variables)

Often discrete variables describing data (N, O)

Categories, dates, binned quantities

Measures (~ dependent variables)

Data values that can be aggregated (Q)

Numbers to be analyzed

Aggregate as sum, count, avg, std. dev...

Not a strict distinction. The same variable may be treated either way depending on the task.

Example: U.S. Census Data

Example: U.S. Census Data

- **People Count:** # of people in group
- **Year:** 1850 – 2000 (every decade)
- **Age:** 0 – 90+
- **Sex:** Male, Female
- **Marital Status:** Single, Married, Divorced, ...

Example: U.S. Census

People Count

Year

Age

Sex

Marital Status

2,348 data points

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534
30	1850	70	0	1	73677
31	1850	70	0	2	71762
32	1850	75	0	1	40834
33	1850	75	0	2	40229
34	1850	80	0	1	23449
35	1850	80	0	2	22949
36	1850	85	0	1	8186
37	1850	85	0	2	10511
38	1850	90	0	1	5259
39	1850	90	0	2	6569
40	1860	0	0	1	2120846
41	1860	0	0	2	2092162

Census: N, O, Q?

People Count	Q-Ratio
Year	Q-Interval (O)
Age	Q-Ratio (O)
Sex	N
Marital Status	N

Census: Dimension or Measure?

People Count	Measure
Year	Dimension
Age	Depends!
Sex	Dimension
Marital Status	Dimension