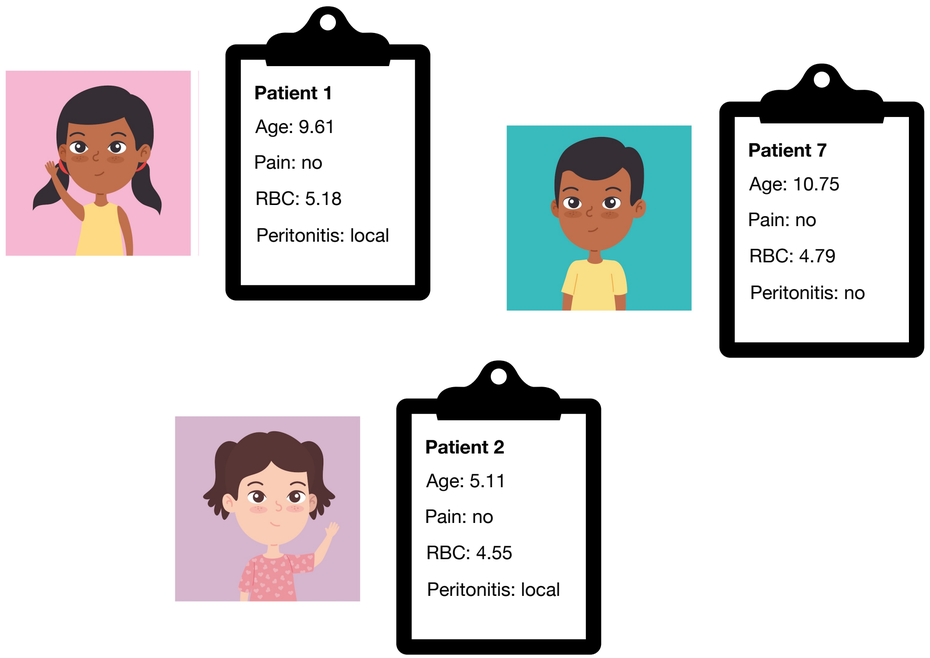
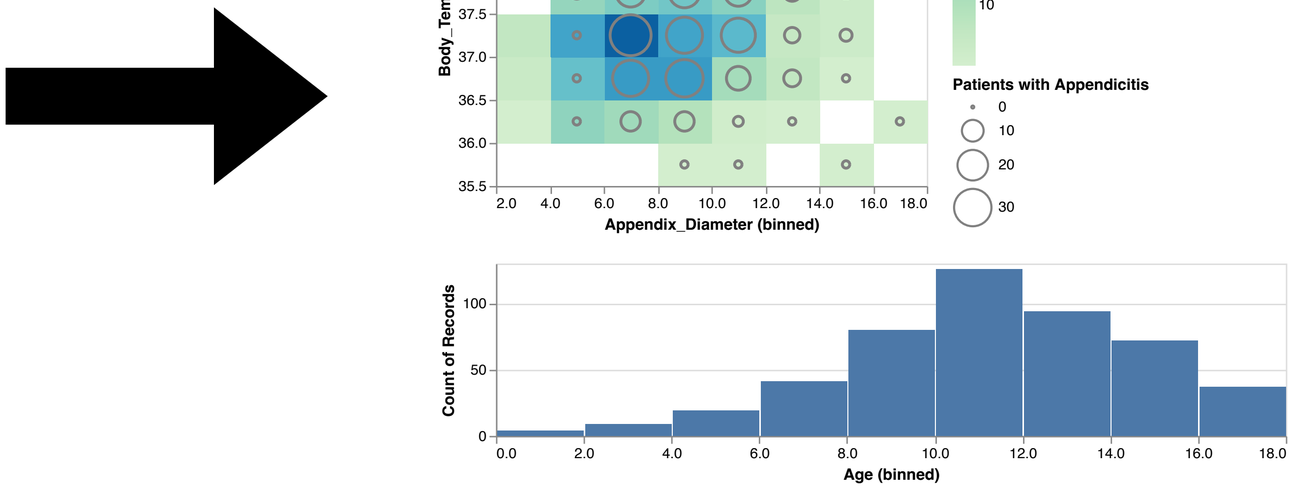
A framework for Data Visualization

Based on slides from Tamara Munzner

Nested model for visualization analysis

Goal: Real world to visualization



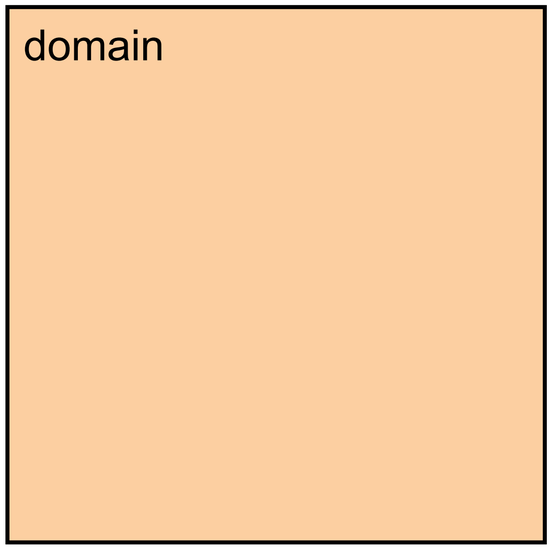


Appendicitis patients

Visualization

Analysis framework: Four levels, three questions

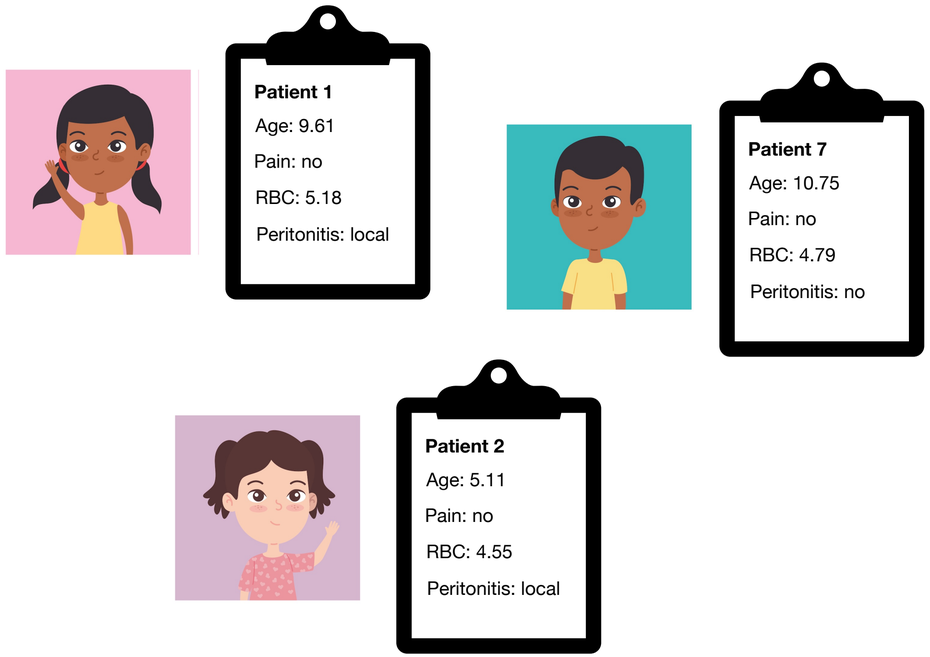
- *domain situation*
  - who are the target users? How do we collect our data?



[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

Analysis framework: Four levels, three questions

- *domain situation*
  - who are the target users? How do we collect our data?



domain

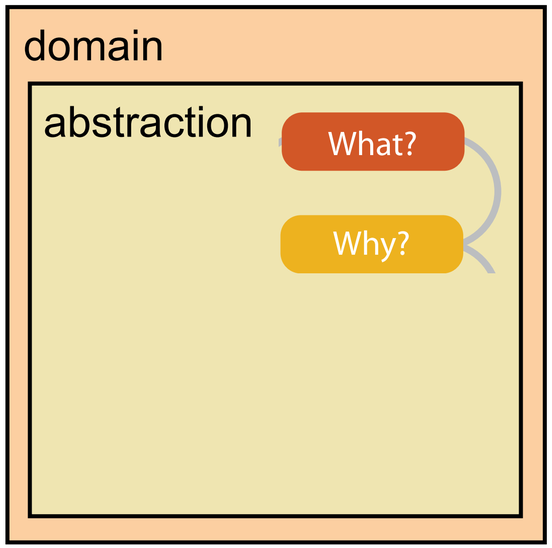
Data: patients

Audience: medical professionals

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

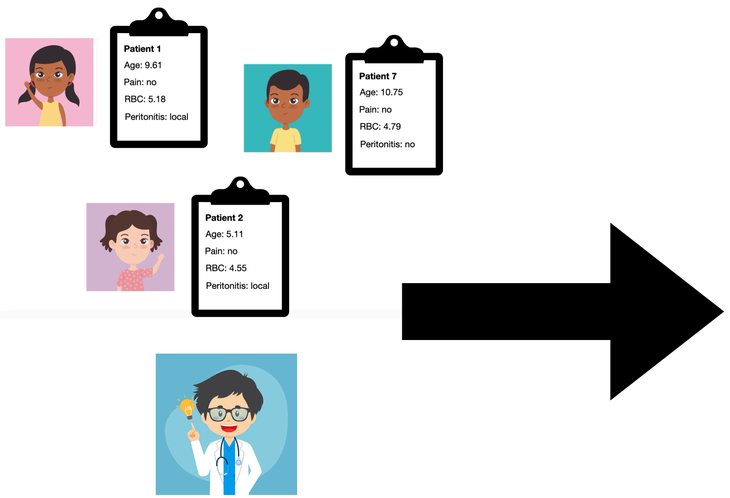
Analysis framework: Four levels, three questions

- *domain situation*
  - who are the target users? How do we collect our data?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data** abstraction
    - **why** is the user looking at it? **task** abstraction



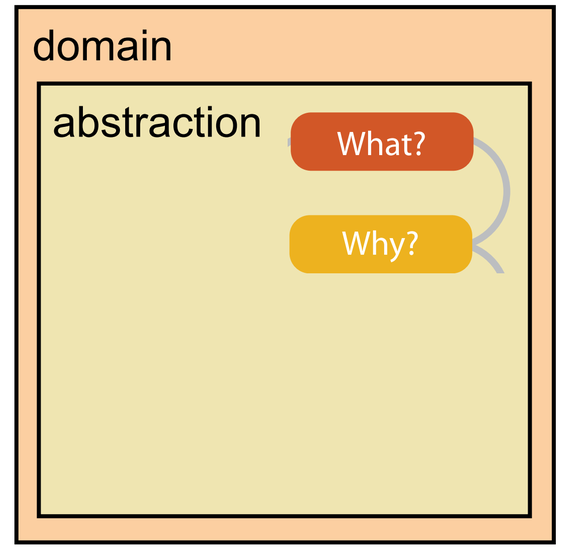
Analysis framework: Four levels, three questions

- *domain situation*
  - who are the target users? How do we collect our data?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data** abstraction
    - **why** is the user looking at it? **task** abstraction



Patients with abdominal pain

	Age	Appendix Size	Migratory Pain	RBC Count	RBC Urine	Peritonitis
0	9.61	9.0	no	5.18	high	local
1	5.11	7.0	no	4.55	medium	local
2	10.75	9.0	no	4.79	none	no
3	10.51	9.0	no	5.03	none	local
4	7.3	6.2	yes	4.64	low	no
5	15.21	8.5	yes	4.62	low	no
6	15.83	12.0	yes	4.33	high	no
7	9.58	7.0	yes	5.04	low	generalized
8	10.37	5.5	no	4.8	none	no
9	16.66	9.0	yes	5.31	none	no
10	14.52	4.5	yes	4.5	none	no
11	10.74	9.0	no	5.66	none	local
12	12.41	3.7	no	5.49	none	no
13	6.67	3.5	no	5.27	none	no
14	14.36	9.0	yes	4.84	low	local
15	9.04	5.3	yes	4.92	low	no
16	12.43	12.0	yes	4.62	none	generalized

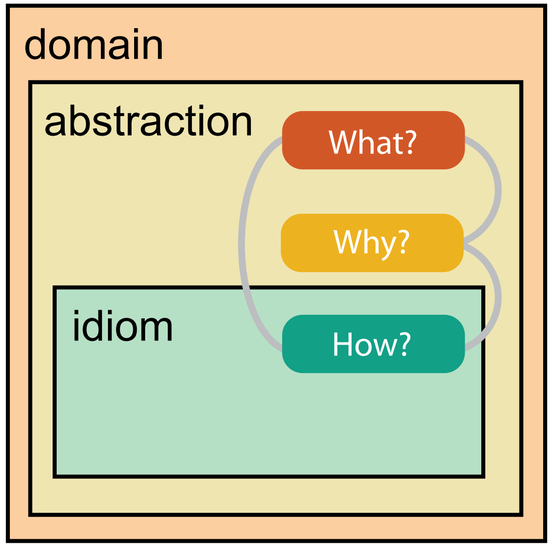


Task: Determine if these variables are related

Domain: Medicine

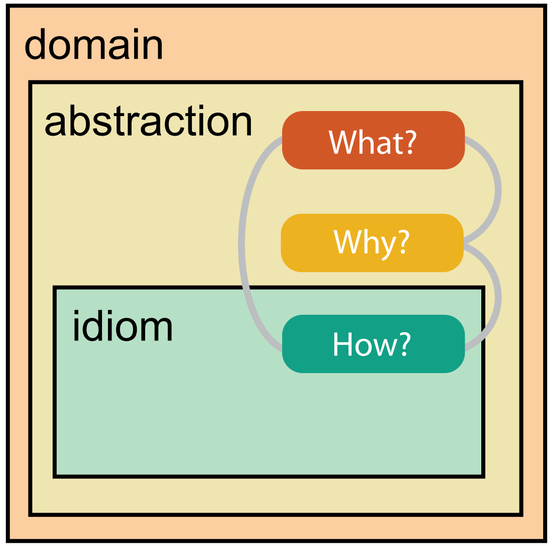
Dataset: Table

Analysis framework: Four levels, three questions

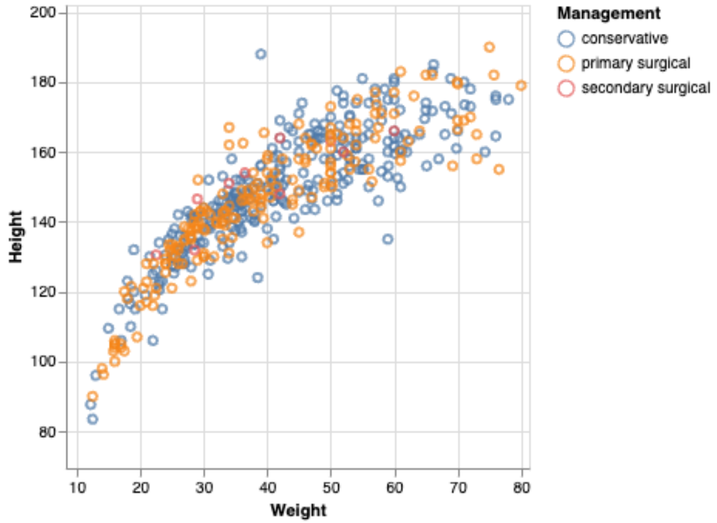


- *domain situation*
  - who are the target users? How do we collect our data?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data** abstraction
    - **why** is the user looking at it? **task** abstraction
- *idiom*
  - **how** is it shown?
    - **visual encoding** idiom: how to draw
    - **interaction** idiom: how to manipulate

Analysis framework: Four levels, three questions

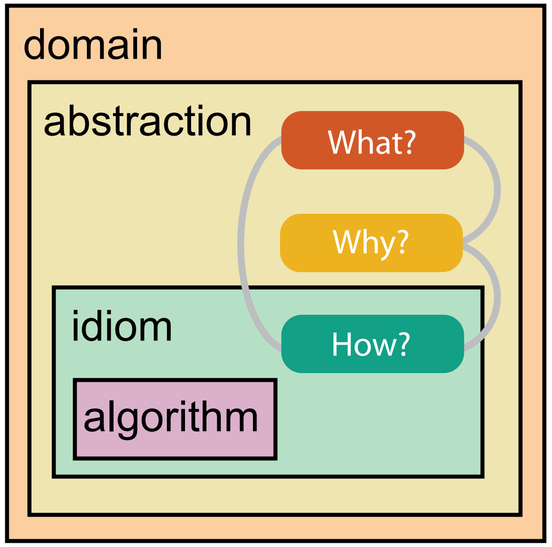


- *domain situation*
  - who are the target users? How do we collect our data?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data** abstraction
    - **why** is the user looking at it? **task** abstraction
- *idiom*
  - **how** is it shown?
    - **visual encoding** idiom: how to draw
    - **interaction** idiom: how to manipulate



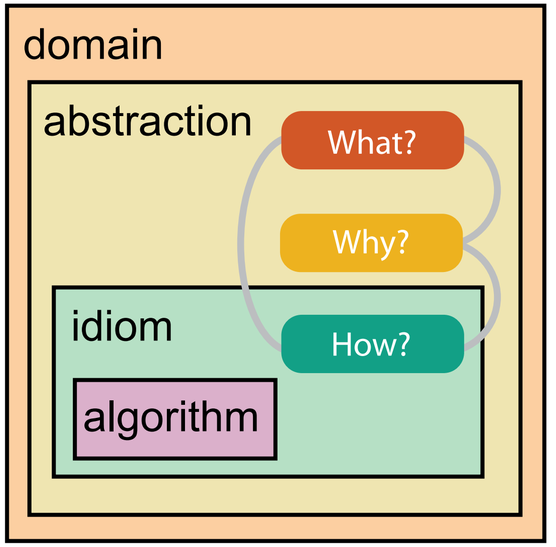
Visual encoding: Scatterplot

Analysis framework: Four levels, three questions



- *domain situation*
  - who are the target users? How do we collect our data?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data** abstraction
    - **why** is the user looking at it? **task** abstraction
- *idiom*
  - **how** is it shown?
    - **visual encoding** idiom: how to draw
    - **interaction** idiom: how to manipulate
- *algorithm*
  - efficient implementation

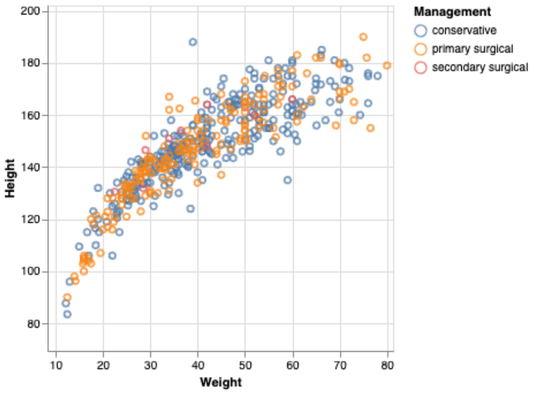
Analysis framework: Four levels, three questions



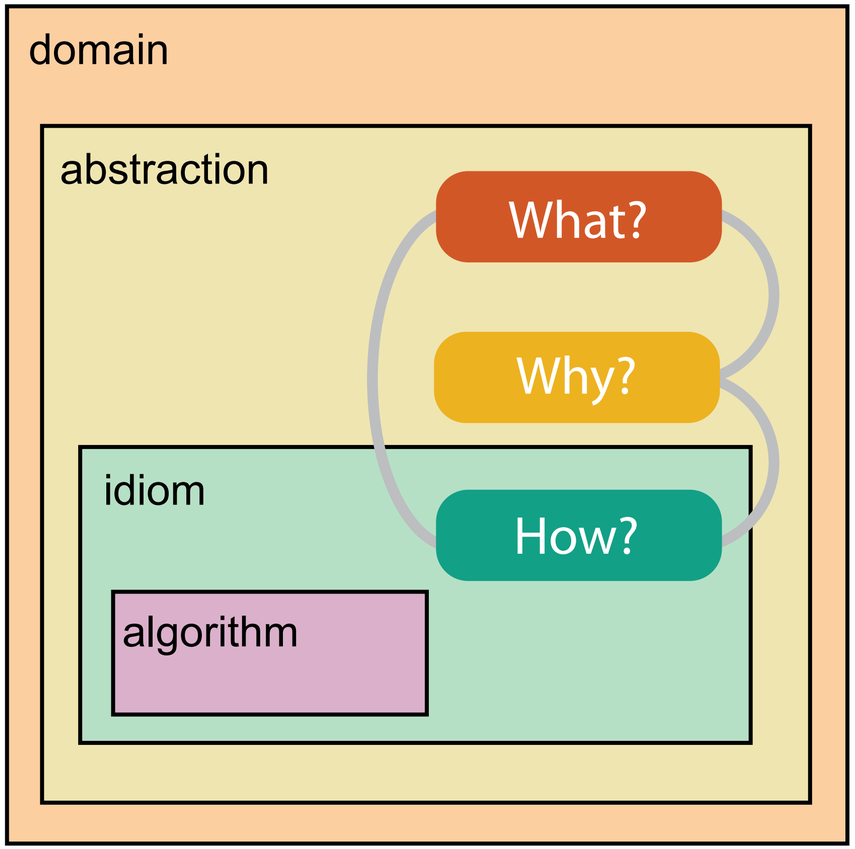
- *domain situation*
  - who are the target users? How do we collect our data?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data** abstraction
    - **why** is the user looking at it? **task** abstraction
- *idiom*
  - **how** is it shown?
    - **visual encoding** idiom: how to draw
    - **interaction** idiom: how to manipulate
- *algorithm*
  - efficient implementation

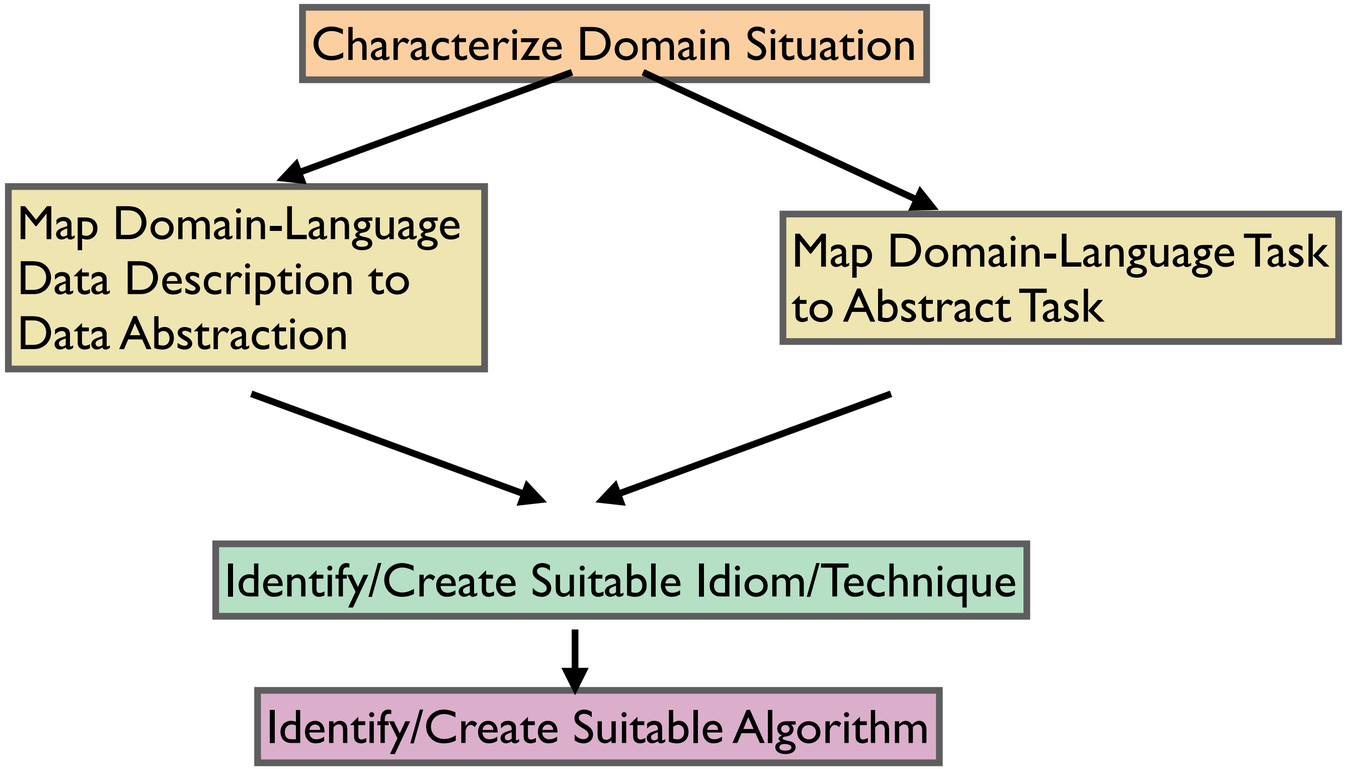
```
alt.Chart(patients)
    .mark_point()
    .encode(
        x='Weight',
        y='Height',
        color='Management'
        ).interactive()
```

Implementation

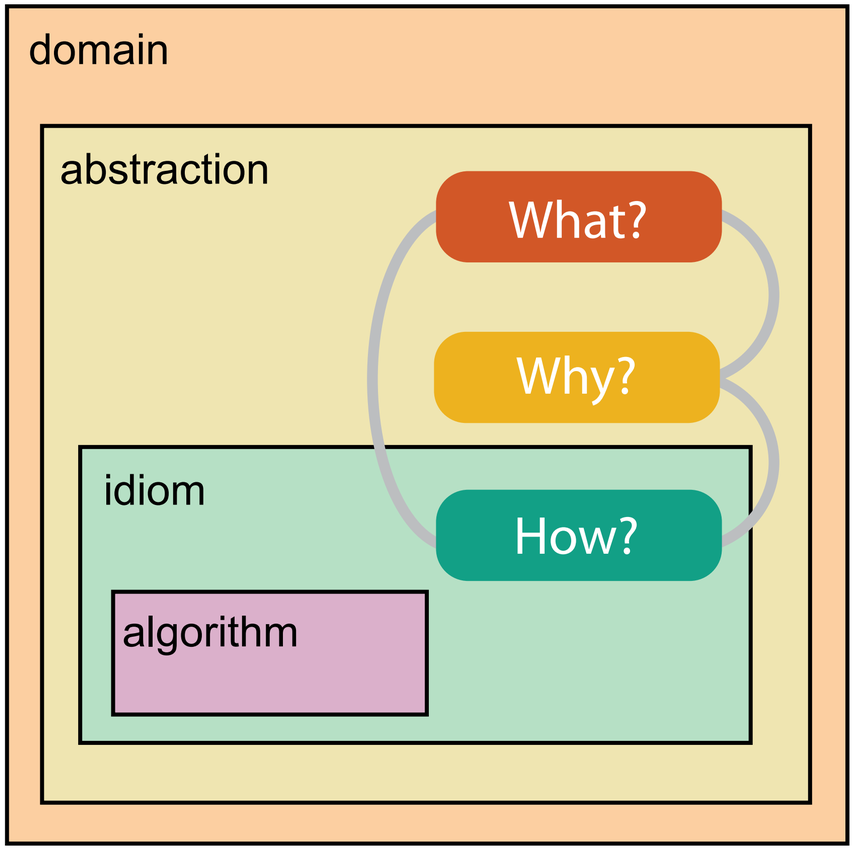


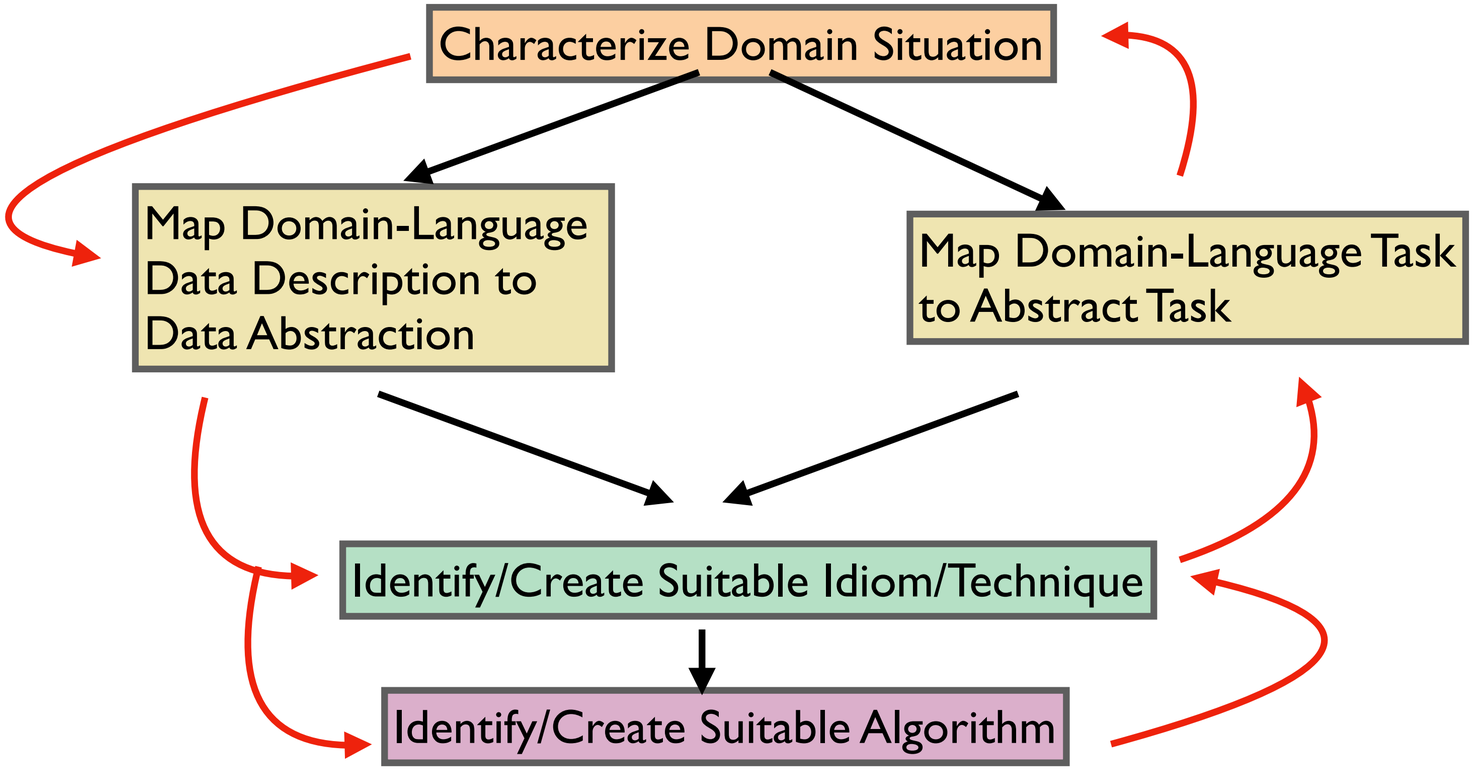
As a process


domain
abstraction
What?
Why?
idiom
How?
algorithm

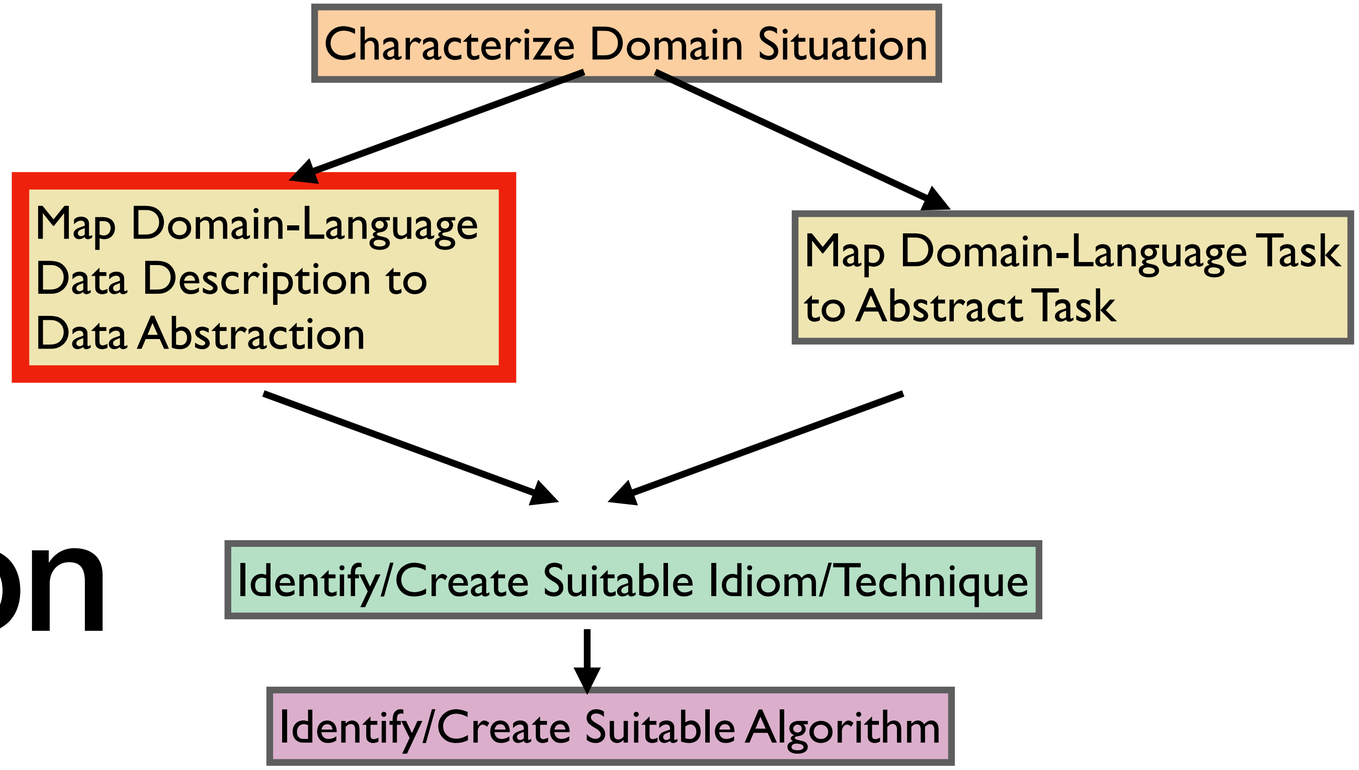

Characterize Domain Situation
Map Domain-Language Data Description to Data Abstraction
Map Domain-Language Task to Abstract Task
Identify/Create Suitable Idiom/Technique
Identify/Create Suitable Algorithm

As a process, May require iteration!

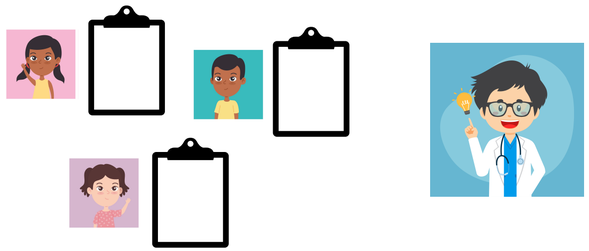




Data Abstraction

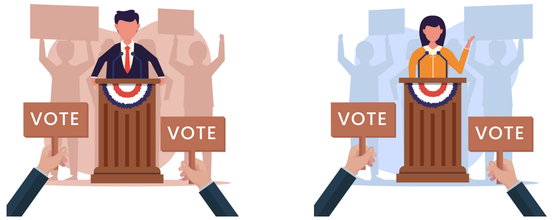


Why do we need abstraction?



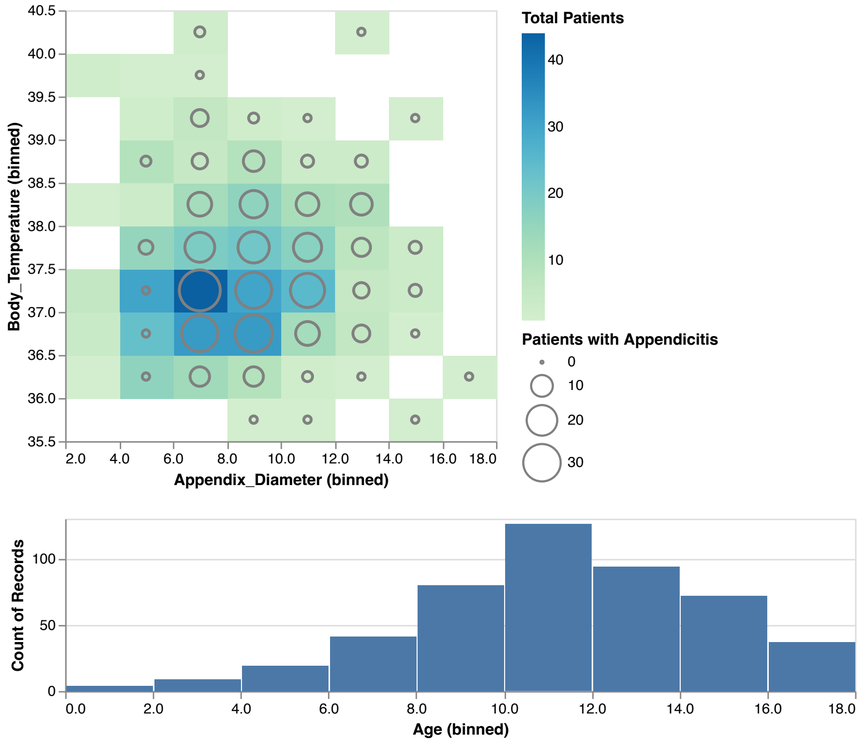
Medicine

Weather



Politics

- Many different domains
- Want a consistent set of visualization tools



Targets of abstraction

Datasets

Data Types

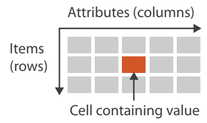
→ Items → Attributes → Links → Positions → Grids

Data and Dataset Types

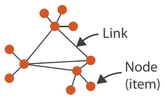
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

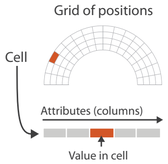
→ Tables



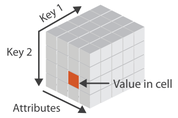
→ Networks



→ Fields (Continuous)



→ *Multidimensional Table*



→ *Trees*

→ Geometry (Spatial)

Attributes

Attribute Types

→ Categorical

→ Ordered

→ *Ordinal*

→ *Quantitative*

Ordering Direction

→ Sequential

→ Diverging

→ Cyclic

Dataset (structure abstraction)

Attribute (type abstraction)

Dataset (structure abstraction)

Recap: Tabular & Tidy Data

Tabular data

Patient 1
Age: 9.61
Pain: no
RBC: 5.18
Peritonitis: local

Patient 2
Age: 5.11
Pain: no
RBC: 4.55
Peritonitis: local

Patient 7
Age: 10.75
Pain: no
RBC: 4.79
Peritonitis: no

Patients

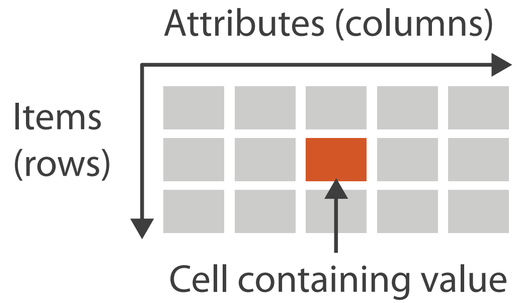
Patients with abdominal pain

	Age	Appendix Size	Migratory Pain	RBC Count	RBC Urine	Peritonitis
0	9.61	9.0	no	5.18	high	local
1	5.11	7.0	no	4.55	medium	local
2	10.75	5.0	no	4.79	none	no
3	10.51	9.0	no	5.03	none	local
4	7.3	6.2	yes	4.64	low	no
5	15.21	8.5	yes	4.62	low	no
6	15.83	12.0	yes	4.33	high	no
7	9.58	7.0	yes	5.04	low	generalized
8	10.37	5.5	no	4.8	none	no
9	16.66	9.0	yes	5.31	none	no
10	14.52	4.5	yes	4.9	none	no
11	10.74	9.0	no	5.66	none	local
12	12.41	3.7	no	5.49	none	no
13	6.67	3.5	no	5.27	none	no
14	14.36	9.0	yes	4.84	low	local
15	9.04	5.3	yes	4.92	low	no
16	12.43	12.0	yes	4.62	none	generalized

Table

- ***Structure:*** *Stored as a table*

➔ Tables



Tabular data

Patient 1
Age: 9.61
Pain: no
RBC: 5.18
Peritonitis: local

Patient 2
Age: 5.11
Pain: no
RBC: 4.55
Peritonitis: local

Patient 7
Age: 10.75
Pain: no
RBC: 4.79
Peritonitis: no

Observation

Patients with abdominal pain

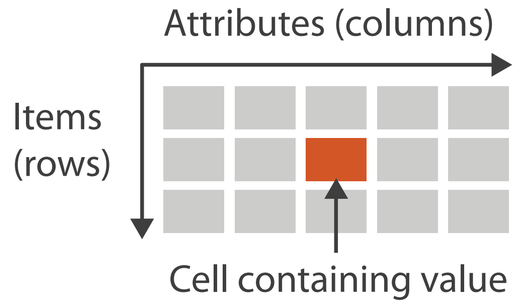
	Age	Appendix Size	Migratory Pain	RBC Count	RBC Urine	Peritonitis
0	9.61	9.0	no	5.18	high	local
1	5.11	7.0	no	4.55	medium	local
2	10.75	5.0	no	4.79	none	no
3	10.51	9.0	no	5.03	none	local
4	7.3	6.2	yes	4.64	low	no
5	15.21	8.5	yes	4.62	low	no
6	15.83	12.0	yes	4.33	high	no
7	9.58	7.0	yes	5.04	low	generalized
8	10.37	5.5	no	4.8	none	no
9	16.66	9.0	yes	5.31	none	no
10	14.52	4.5	yes	4.9	none	no
11	10.74	9.0	no	5.66	none	local
12	12.41	3.7	no	5.49	none	no
13	6.67	3.5	no	5.27	none	no
14	14.36	9.0	yes	4.84	low	local
15	9.04	5.3	yes	4.92	low	no
16	12.43	12.0	yes	4.62	none	generalized

Patients

Table

- ***Structure:*** *Stored as a table*
- *Table outline*
  - Rows correspond to **observations**
    - Items, entities, etc.
  -

➔ Tables



Tabular data

Patient 1
Age: 9.61
Pain: no
RBC: 5.18
Peritonitis: local

Patient 2
Age: 5.11
Pain: no
RBC: 4.55
Peritonitis: local

Patient 7
Age: 10.75
Pain: no
RBC: 4.79
Peritonitis: no

Observation

Patients with abdominal pain

	Age	Appendix Size	Migratory Pain	RBC Count	RBC Urine	Peritonitis
0	9.61	9.0	no	5.18	high	local
1	5.11	7.0	no	4.55	medium	local
2	10.75	5.0	no	4.79	none	no
3	10.51	9.0	no	5.03	none	local
4	7.3	6.2	yes	4.64	low	no
5	15.21	8.5	yes	4.62	low	no
6	15.83	12.0	yes	4.33	high	no
7	9.58	7.0	yes	5.04	low	generalized
8	10.37	5.5	no	4.8	none	no
9	16.66	9.0	yes	5.31	none	no
10	14.52	4.5	yes	4.9	none	no
11	10.74	9.0	no	5.66	none	local
12	12.41	3.7	no	5.49	none	no
13	6.67	3.5	no	5.27	none	no
14	14.36	9.0	yes	4.84	low	local
15	9.04	5.3	yes	4.92	low	no
16	12.43	12.0	yes	4.62	none	generalized

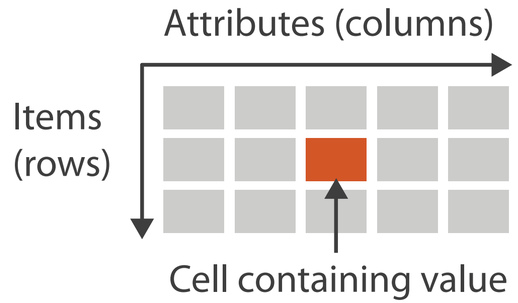
Attribute

Patients

Table

- ***Structure:*** *Stored as a table*
- *Table outline*
  - Rows correspond to **observations**
    - Items, entities, etc.
  - Columns correspond to **attributes (variables)**
    - Measurable properties
  -

➔ Tables



Tabular data

Patient 1
Age: 9.61
Pain: no
RBC: 5.18
Peritonitis: local

Patient 2
Age: 5.11
Pain: no
RBC: 4.55
Peritonitis: local

Patient 7
Age: 10.75
Pain: no
RBC: 4.79
Peritonitis: no

Patients

Observation

Patients with abdominal pain

	Age	Appendix Size	Migratory Pain	RBC Count	RBC Urine	Peritonitis
0	9.61	9.0	no	5.18	high	local
1	5.11	7.0	no	4.55	medium	local
2	10.75	5.0	no	4.79	none	no
3	10.51	9.0	no	5.03	none	local
4	7.3	6.2	yes	4.64	low	no
5	15.21	8.5	yes	4.62	low	no
6	15.83	12.0	yes	4.33	high	no
7	9.58	7.0	yes	5.04	low	generalized
8	10.37	5.5	no	4.8	none	no
9	16.66	9.0	yes	5.31	none	no
10	14.52	4.5	yes	4.9	none	no
11	10.74	9.0	no	5.66	none	local
12	12.41	3.7	no	5.49	none	no
13	6.67	3.5	no	5.27	none	no
14	14.36	9.0	yes	4.84	low	local
15	9.04	5.3	yes	4.92	low	no
16	12.43	12.0	yes	4.62	none	generalized

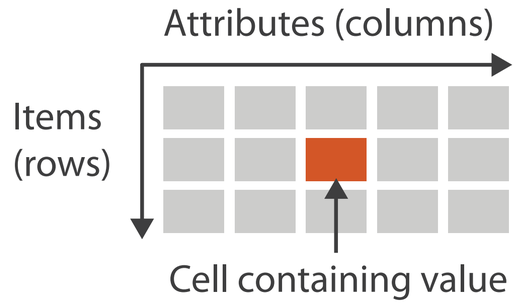
Cell

Attribute

Table

- ***Structure:*** *Stored as a table*
- *Table outline*
  - Rows correspond to **observations**
    - Items, entities, etc.
  - Columns correspond to **attributes (variables)**
    - Measurable properties
  - Cells correspond to **values**
    - Item-attribute pair

➔ Tables



Tabular data

Patient 1
Age: 9.61
Pain: no
RBC: 5.18
Peritonitis: local

Patient 2
Age: 5.11
Pain: no
RBC: 4.55
Peritonitis: local

Patient 7
Age: 10.75
Pain: no
RBC: 4.79
Peritonitis: no

Patients

Observation

Patients with abdominal pain

	Age	Appendix Size	Migratory Pain	RBC Count	RBC Urine	Peritonitis
0	9.61	9.0	no	5.18	high	local
1	5.11	7.0	no	4.55	medium	local
2	10.75	5.0	no	4.79	none	no
3	10.51	9.0	no	5.03	none	local
4	7.3	6.2	yes	4.64	low	no
5	15.21	8.5	yes	4.62	low	no
6	15.83	12.0	yes	4.33	high	no
7	9.58	7.0	yes	5.04	low	generalized
8	10.37	5.5	no	4.8	none	no
9	16.66	9.0	yes	5.31	none	no
10	14.52	4.5	yes	4.9	none	no
11	10.74	9.0	no	5.66	none	local
12	12.41	3.7	no	5.49	none	no
13	6.67	3.5	no	5.27	none	no
14	14.36	9.0	yes	4.84	low	local
15	9.04	5.3	yes	4.92	low	no
16	12.43	12.0	yes	4.62	none	generalized

Cell

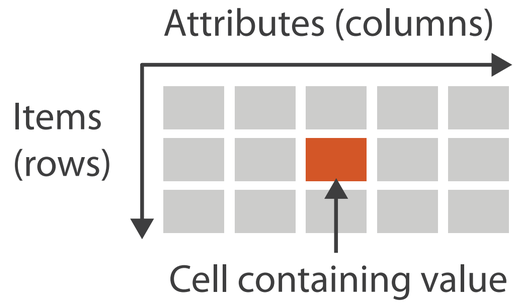
Unique key (could be implicit)

Attribute

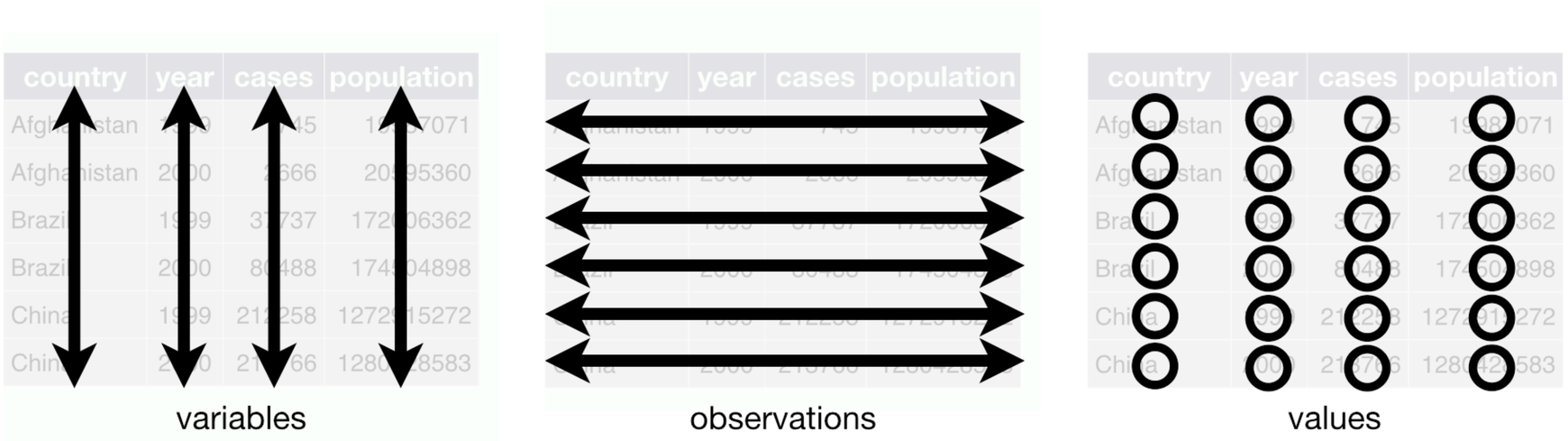
Table

- ***Structure:*** *Stored as a table*
- *Table outline*
  - Rows correspond to **observations**
    - Items, entities, etc.
  - Columns correspond to **attributes (variables)**
    - Measurable properties
  - Cells correspond to **values**
    - Item-attribute pair

➔ Tables



Wickham's tidy data principles



- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table (DataFrame)
- Each value has its own cell.

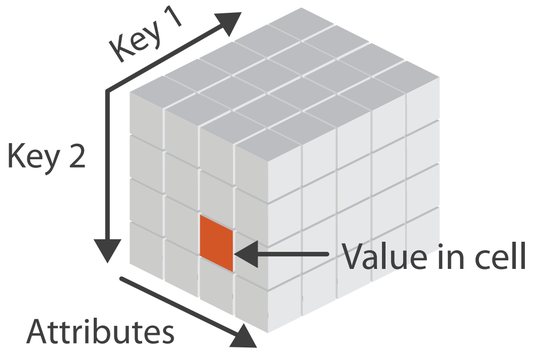
Other data types

- links
  - express relationship between two items
  - eg friendship on facebook, interaction between proteins
- positions
  - spatial data: location in 2D or 3D
  - pixels in photo, voxels in MRI scan, latitude/longitude
- grids
  - sampling strategy for continuous data

Dataset types

- multidimensional tables
  - indexing based on multiple keys
    - eg genes, patients

➔ *Multidimensional Table*

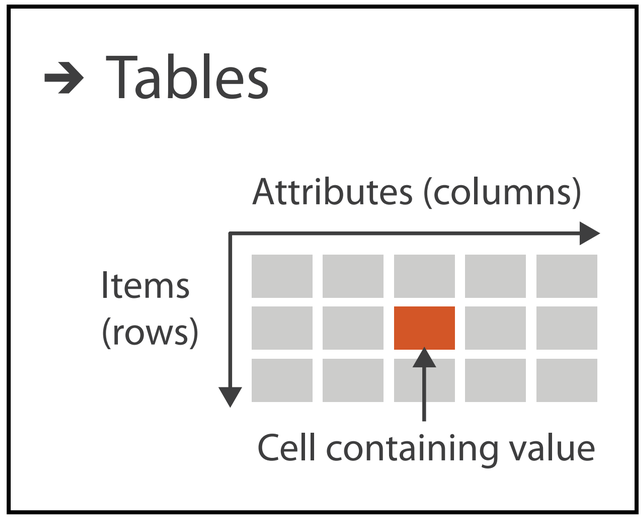


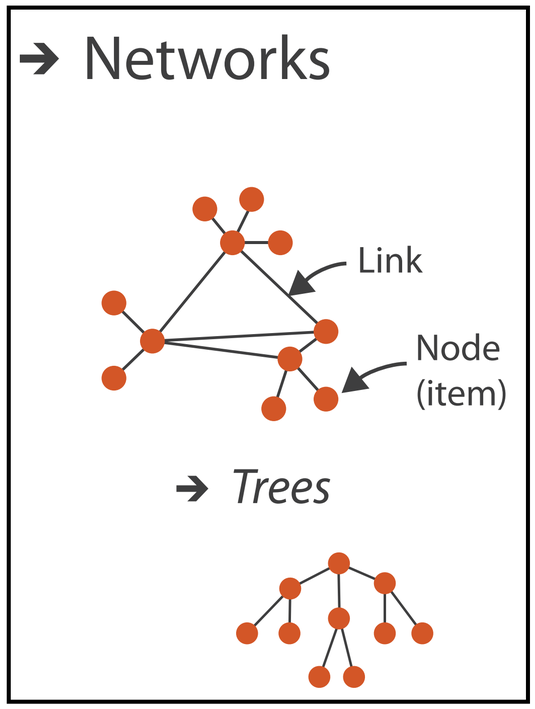
	A	B	C	D	E
1	#1.2				
2	1500	529			
3	GeneName	DESCRIPTION	TCGA-02-0001-01C-01R-0177-01	TCGA-02-0003-01A-01R-0177-01	TCGA-02-0004-01A-01R-0298-01
4	LTF	LTF	-1.265728057	2.377012066	4.123979585
5	POSTN	POSTN	2.662411805	3.932400324	5.031585377
6	TMSL8	TMSL8	-3.082217838	-2.243148513	-0.02313681
7	HLA-DQA1	HLA-DQA1	-1.739664398	4.577962344	3.127744964
8	RP11-35N6.1	RP11-35N6.1	-3.346352968	-2.895400157	-3.473035067
9	STMN2	STMN2	-2.578511106	-3.051605144	-1.729892888
10	DCX	DCX	-2.26078976	-2.529795801	-2.844966278
11	AGXT2L1	AGXT2L1	-2.639493611	-3.113204863	-0.403975027
12	IL13RA2	IL13RA2	-2.93596915	-1.873600916	2.976256911
13	SLN	SLN	-2.466718221	-2.208406749	1.025827904
14	MEOX2	MEOX2	-2.395054066	-1.062676046	1.783235317
15	COL11A1	COL11A1	1.211934832	-0.399392588	4.733608974
16	NNMT	NNMT	0.703745164	0.664082419	3.069030715
17	F13A1	F13A1	-0.224094042	2.222197544	1.171354775
18	CXCL14	CXCL14	-3.1309694	-1.395056071	2.569540659
19	MBP	MBP	-1.906390566	-2.037626447	-2.935744906
20	TF	TF	-4.334123292	-4.680680246	-2.975788866
21	KCND2	KCND2	-1.777692395	-2.100362021	-1.996306032

Dataset types

Tables	Networks & Trees
Items	Items (nodes)
Attributes	Links
	Attributes

- network/graph
  - –nodes (vertices) connected by links (edges)
  - –tree is special case: no cycles
    - often have roots and are directed

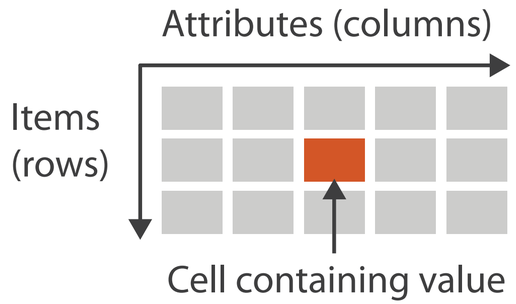




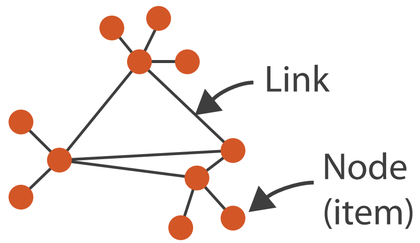
Dataset types

Tables	Networks & Trees	Fields
Items	Items (nodes)	Grids
Attributes	Links	Positions
	Attributes	Attributes

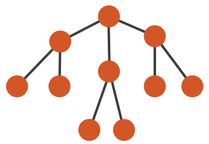
➔ Tables



➔ Networks

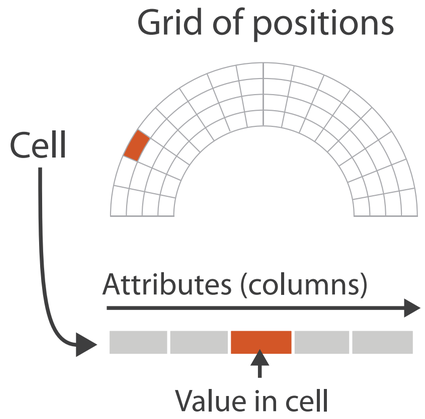


➔ *Trees*



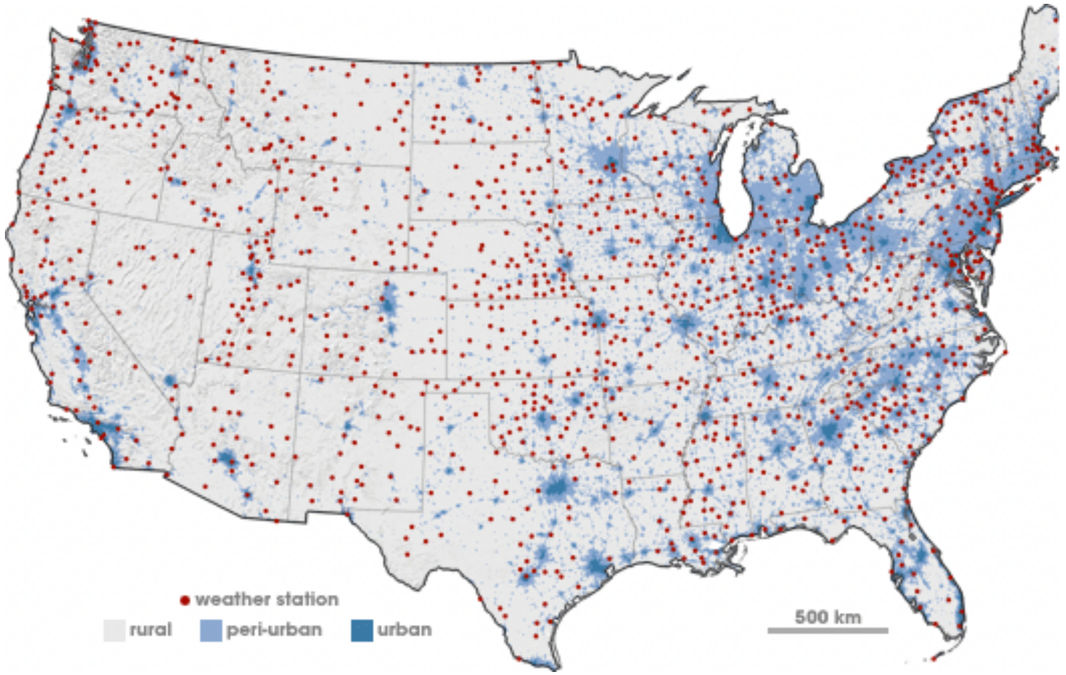
➔ Spatial

➔ Fields (Continuous)



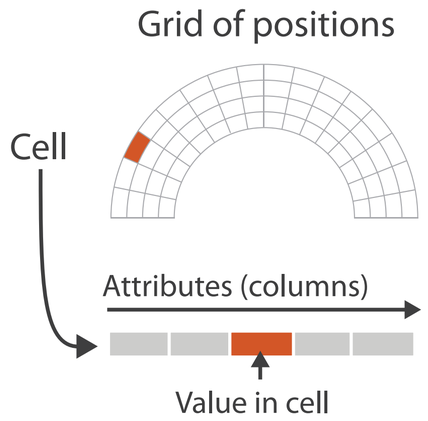
Spatial fields

- attribute values associated w/ cells
- cell contains value from continuous domain
  - eg temperature, pressure, wind velocity
- measured or simulated



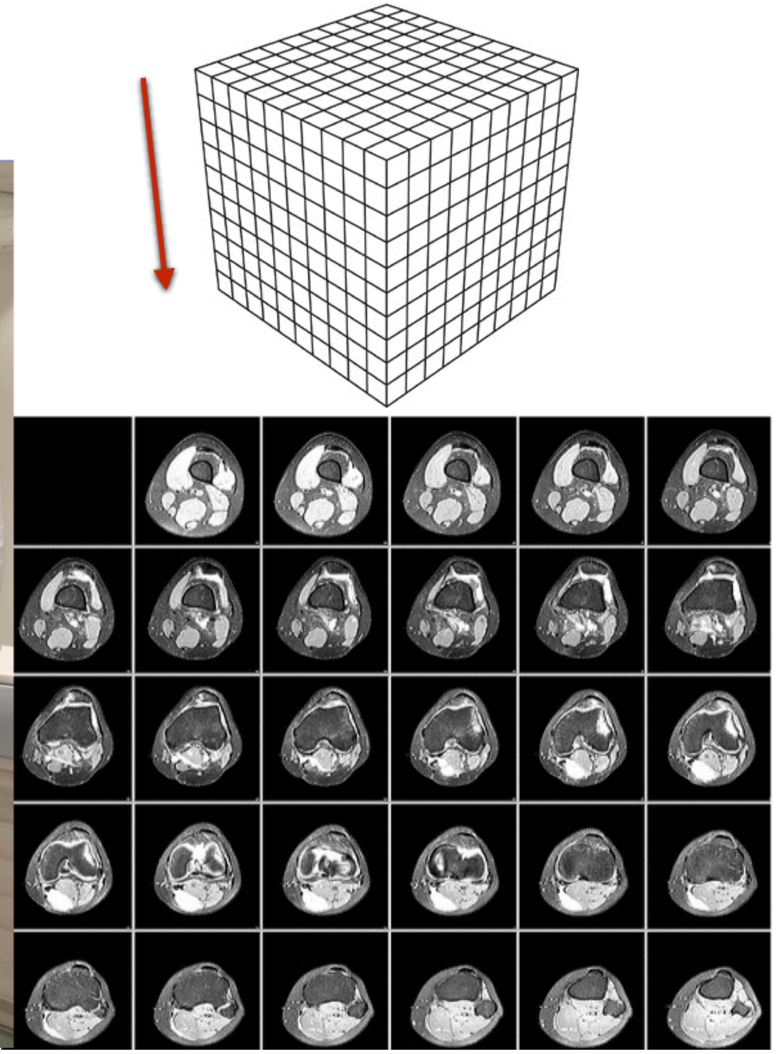
➔ Spatial

➔ Fields (Continuous)



Spatial fields

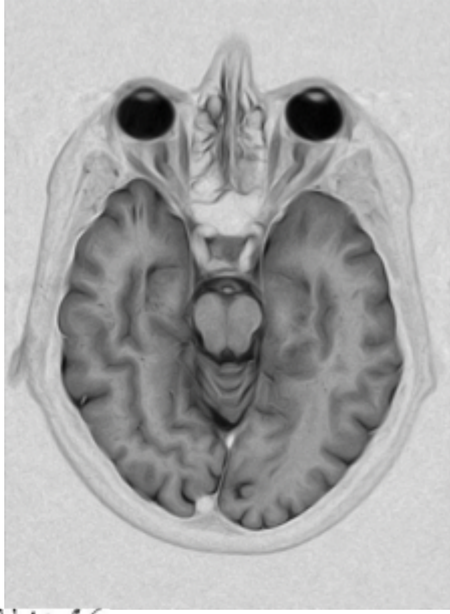
- attribute values associated w/ cells
- cell contains value from continuous domain
  - eg temperature, pressure, wind velocity
- measured or simulated
- major concerns
  - sampling:
    where attributes are measured
  - interpolation:
    how to model attributes elsewhere
  - grid types



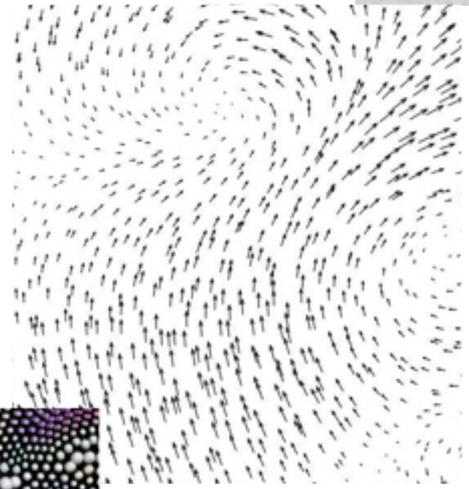
Spatial fields

- attribute values associated w/ cells
- cell contains value from continuous domain
  - eg temperature, pressure, wind velocity
- measured or simulated
- major concerns
  - sampling:
    where attributes are measured
  - interpolation:
    how to model attributes elsewhere
  - grid types
- major divisions
  - attributes per cell:
    scalar (1), vector (2), tensor (many)

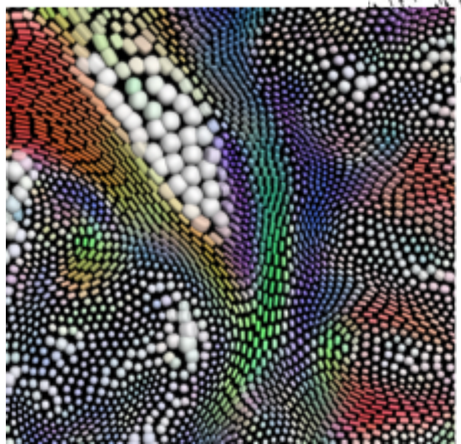
scalar



vector



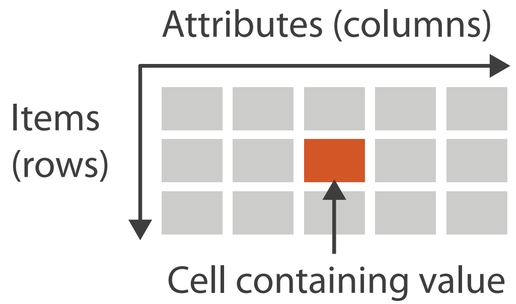
tensor



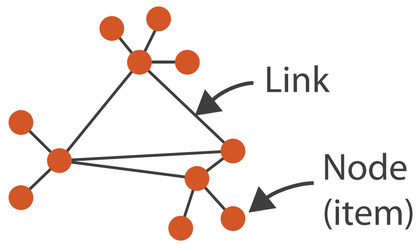
Dataset types

Tables	Networks & Trees	Fields	Geometry
Items	Items (nodes)	Grids	Items
Attributes	Links	Positions	Positions
	Attributes	Attributes	

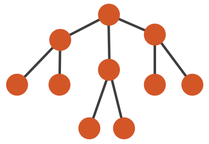
➔ Tables



➔ Networks

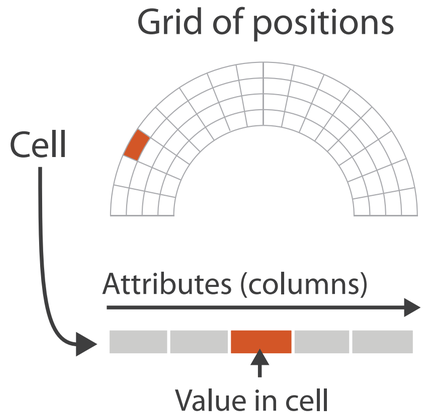


➔ *Trees*

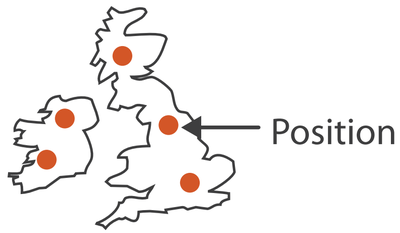


➔ Spatial

➔ Fields (Continuous)

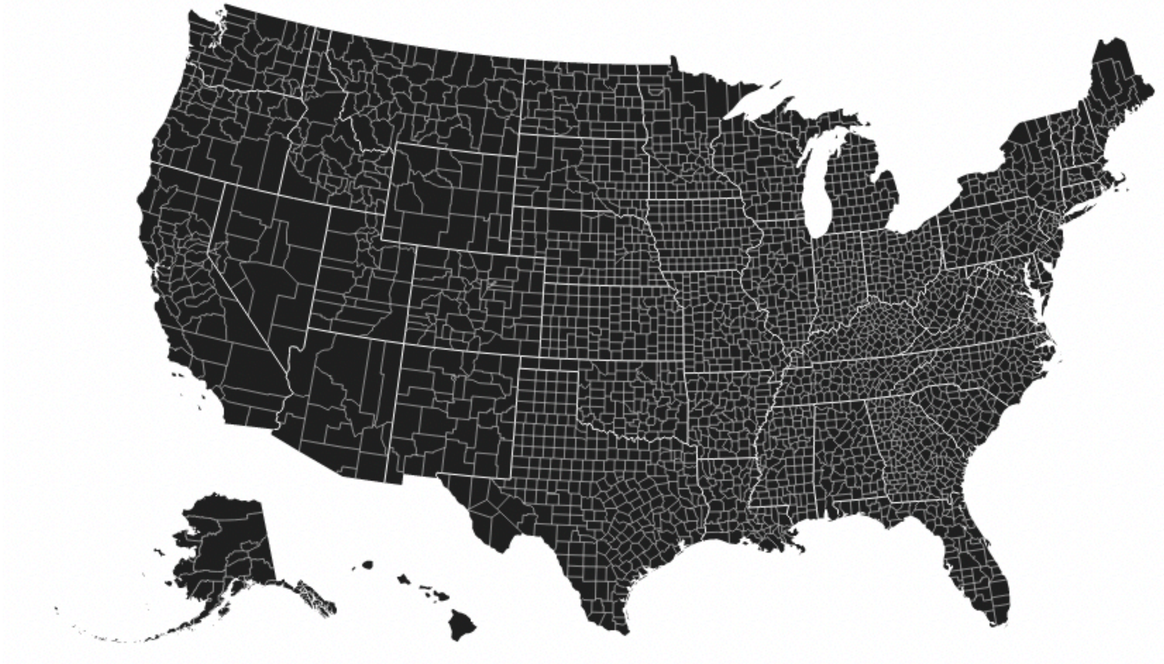


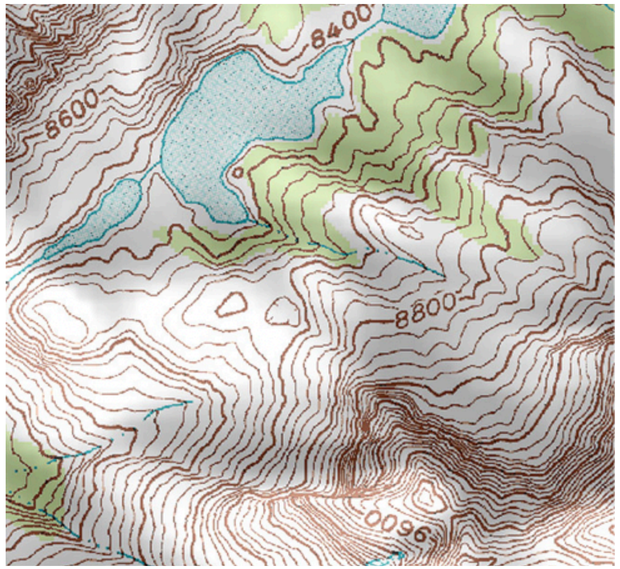
➔ Geometry (Spatial)



Geometry

- shape of items
- explicit spatial positions / regions
  - points, lines, curves, surfaces, volumes
- boundary between computer graphics and visualization
  - graphics: geometry taken as given
  - vis: geometry is result of a design decision





Type abstraction

Variable types

A defining characteristic of a variable is its **type**

```
1 nations['population'].dtype
```

```
dtype('int64')
```

- This should be familar to computer scientists!

The **type** of a variable determines what kinds of values it can take and how we should interpret them

What is an appropriate level of type abstraction?

Patients with abdominal pain

	Age	Appendix Size	Height	Weight	RBC Count	Temperature	WBC Count
0	16.66	9.0	174.0	65.0	5.31	36.6	6.6
1	10.74	9.0	146.0	57.5	5.66	37.3	10.2
2	9.04	5.3	134.0	29.4	4.92	36.0	5.1
3	10.75	5.0	155.0	54.5	4.79	37.7	10.3
4	7.3	6.2	123.0	23.5	4.64	37.4	21.1
5	5.11	7.0	116.0	22.0	4.55	40.2	19.4
6	14.36	9.0	163.0	50.0	4.84	37.5	14.3
7	9.61	9.0	140.0	29.2	5.18	38.7	14.3
8	15.83	12.0	153.0	59.0	4.33	36.7	12.8
9	9.58	7.0	132.0	24.7	5.04	38.4	13.5
10	10.37	5.5	156.0	39.0	4.8	37.4	5.6
11	14.52	4.5	181.0	55.0	4.9	37.0	9.0
12	12.41	3.7	150.5	42.5	5.49	37.2	9.1
13	6.67	3.5	124.0	38.5	5.27	39.6	16.8
14	15.21	8.5	155.0	85.0	4.62	36.8	12.4
15	12.43	12.0	157.0	46.0	4.62	37.1	16.4
16	10.51	9.0	134.5	27.0	5.03	37.4	12.8

- **Semantics (conceptual model)**
  - e.g. Temperature
    - Too domain specific!
- **Concrete type (data model)**
  - e.g. Float
    - Too implementation specific!
- **Data type abstraction**

- Attribute Types
  - Categorical

  - Ordered
    - *Ordinal*

    - *Quantitative*

- Ordering Direction
  - Sequential

  - Diverging

  - Cyclic

Variable types

- **Nominal**
  - Categorical
  - Arbitrary
- Ordinal
- Quantitative
  - Interval
  - Ratio

Nominal variables

An *unordered* set of non-numeric values, representing labels or categories

Categorical - Finite

Possible values are finite and *known*

- Colors: *{red, green, blue}*
- Regions: *{South Asia, Europe & Central Asia, …}*

Arbitrary - Infinite

Possible values are unbounded

- Addresses: *{"12 Main St. Boston MA", "45 Wall St. New York NY", …}*
- Names: *{"John Smith", "Jane Doe", …}*

- Attribute Types
  - Categorical
  - Ordered
    - *Ordinal*
    - *Quantitative*
- Ordering Direction
  - Sequential
  - Diverging
  - Cyclic

Variable types

- Nominal
  - Categorical
  - Arbitrary
- **Ordinal**
- Quantitative
  - Interval
  - Ratio

Ordinal variables

An *ordered* set of (usually) non-numeric values, representing levels

- Grades: *{A, B, C, D, F}*
- Ratings: *{G, PG, PG-13, R}*

Attributes

- **Attribute Types**
  - Categorical

  - Ordered
    - *Ordinal*

    - *Quantitative*

- **Ordering Direction**
  - Sequential

  - Diverging

  - Cyclic

Variable types

- Nominal
  - Categorical
  - Arbitrary
- Ordinal
- Quantitative
  - Interval
  - Ratio

Quantitative variables

Numeric data that we can perform mathematical operations on

Interval

Location of 0 is arbitrary, only differences/intervals can be compared

- Date-Times: *Jan, 19, 2006*

Ratio

0 is well-defined and meaningful. Can compare values as *ratios*

- Height: *{5'3", 6'1", 5'9",…}*
- Population: *23M, 350k, 1.4B, …*

Attributes

- **Attribute Types**
  - Categorical

  - Ordered
    - *Ordinal*

    - *Quantitative*

- **Ordering Direction**
  - Sequential

  - Diverging

  - Cyclic

Variable types

- Nominal
  - Categorical
  - Arbitrary
- Ordinal
- Quantitative
  - Interval
  - Ratio

Can you think of a cyclic variable?

Quantitative variables

Numeric data that we can perform mathematical operations on

Interval

Location of 0 is arbitrary, only differences/intervals can be compared

- Date-Times: *Jan, 19, 2006*

Ratio

0 is well-defined and meaningful. Can compare values as *ratios*

- Height: *{5'3", 6'1", 5'9",…}*
- Population: *23M, 350k, 1.4B, …*

Attributes

- Attribute Types
  - Categorical

  - Ordered
    - *Ordinal*

    - *Quantitative*

- Ordering Direction
  - Sequential

  - Diverging

  - Cyclic

Operations by type

Nominal (Labels or categories)

- Concrete types: `str`, `bool`, `category`
- Operations: =, ≠

Ordinal (Ordered, non-numeric)

- Concrete types: `str`, `int`, `category`
- Operations: =, ≠, <, >

Interval (Arbitrary 0)

- Concrete types: `float`, `int`, `datetime`
- Operations: =, ≠, <, >, -

Ratio (Well-defined 0)

- Concrete types: `float`, `int`, `timedelta`
- Operations: =, ≠, <, >, -, /

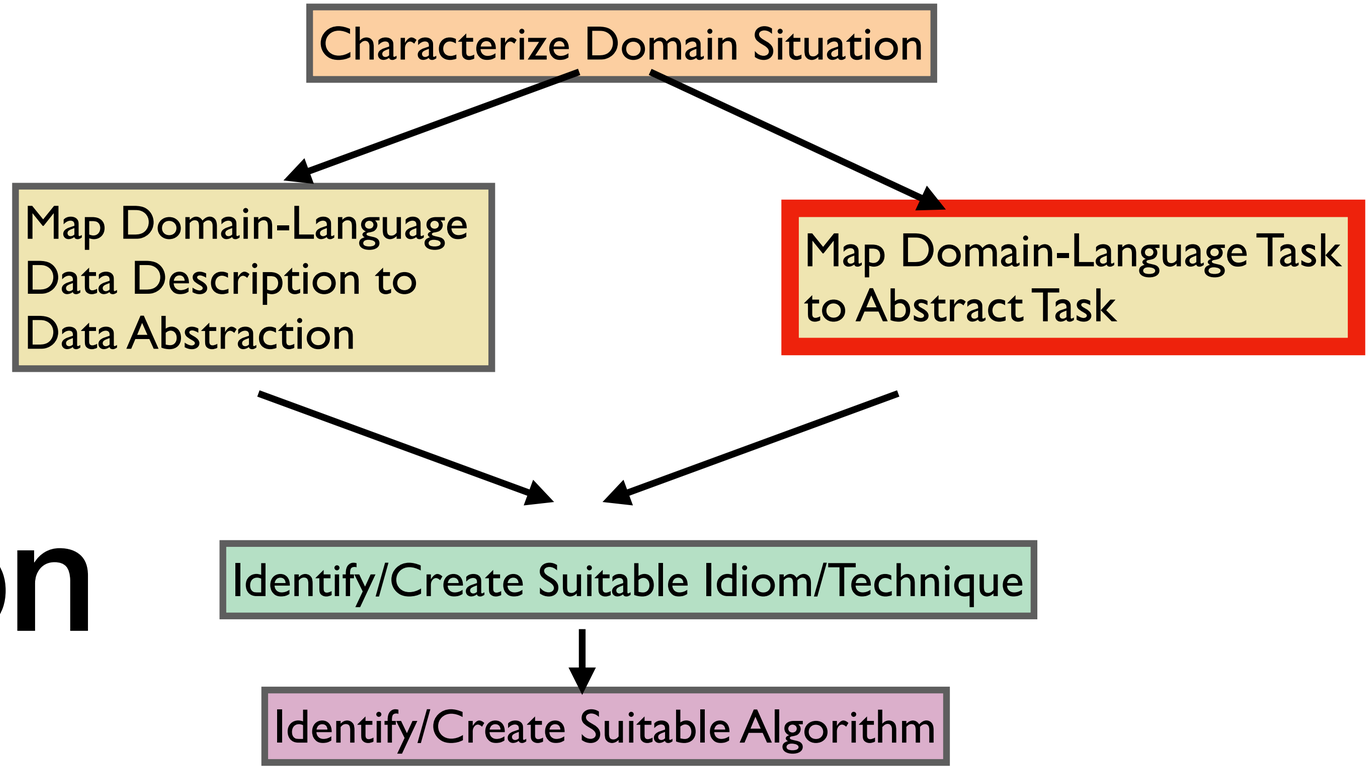
What is an appropriate level of type abstraction?

Patients with abdominal pain

	Age	Appendix Size	Height	Weight	RBC Count	Temperature	WBC Count
0	16.66	9.0	174.0	65.0	5.31	36.6	6.6
1	10.74	9.0	146.0	57.5	5.66	37.3	10.2
2	9.04	5.3	134.0	29.4	4.92	36.0	5.1
3	10.75	5.0	155.0	54.5	4.79	37.7	10.3
4	7.3	6.2	123.0	23.5	4.64	37.4	21.1
5	5.11	7.0	116.0	22.0	4.55	40.2	19.4
6	14.36	9.0	163.0	50.0	4.84	37.5	14.3
7	9.61	9.0	140.0	29.2	5.18	38.7	14.3
8	15.83	12.0	153.0	59.0	4.33	36.7	12.8
9	9.58	7.0	132.0	24.7	5.04	38.4	13.5
10	10.37	5.5	156.0	39.0	4.8	37.4	5.6
11	14.52	4.5	181.0	55.0	4.9	37.0	9.0
12	12.41	3.7	150.5	42.5	5.49	37.2	9.1
13	6.67	3.5	124.0	38.5	5.27	39.6	16.8
14	15.21	8.5	155.0	85.0	4.62	36.8	12.4
15	12.43	12.0	157.0	46.0	4.62	37.1	16.4
16	10.51	9.0	134.5	27.0	5.03	37.4	12.8

- **Semantics (conceptual model)**
  - e.g. Temperature
    - Too domain specific!
- **Concrete type (data model)**
  - e.g. Float
    - Too implementation specific!
- **Data type abstraction**
  - Multiple possible abstractions
    - Quantitative: Temp. in deg. C
    - Categorical: {fever, normal}
    - Ordinal: {normal, low fever, high fever}

Task Abstraction



Task abstraction: Actions and targets

- very high-level pattern

- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

Task abstraction: Actions and targets

- very high-level pattern

- actions
  - analyze
    - high-level choices
  - search
    - find a known/unknown item
  - query
    - find out about characteristics of item

- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

Task abstraction: Actions and targets

- very high-level pattern

- actions
  - analyze
    - high-level choices
  - search
    - find a known/unknown item
  - query
    - find out about characteristics of item

- targets
  - what is being acted on

- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

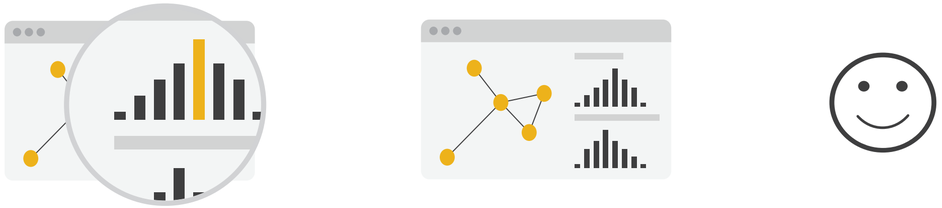
Actions: Analyze

- consume
  - discover vs present
    - classic split
    - aka explore vs explain
  - enjoy
    - newcomer
    - aka casual, social
- produce
  - annotate, record
  - derive
    - crucial design choice

➔ **Analyze**

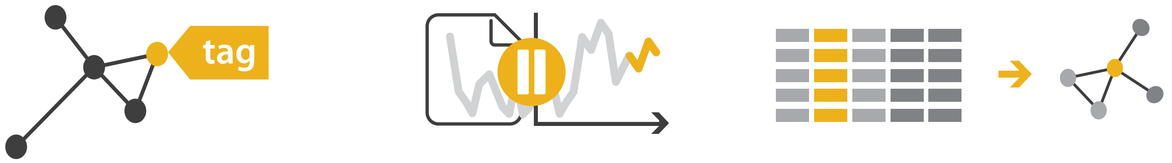
➔ Consume

➔ *Discover* ➔ *Present* ➔ *Enjoy*



➔ Produce

➔ *Annotate* ➔ *Record* ➔ *Derive*



Actions: Search

Actions: Search

- what does user know?
  - target, location

Search

	Target known	Target unknown
Location known	 *Lookup*	 *Browse*
Location unknown	 *Locate*	 *Explore*

Actions: Search

- what does user know?
  - target, location
- lookup
  - ex: word in dictionary
    - alphabetical order

Search

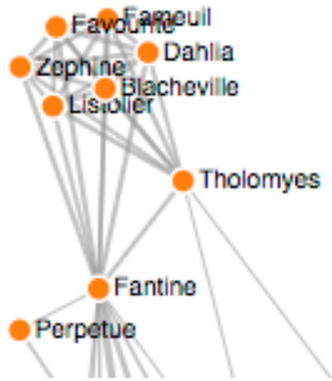
	Target known	Target unknown
Location known	 *Lookup*	 *Browse*
Location unknown	 *Locate*	 *Explore*

Actions: Search

- what does user know?
  - target, location
- lookup
  - ex: word in dictionary
    - alphabetical order
- locate
  - ex: keys in your house
  - ex: node in network

Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*



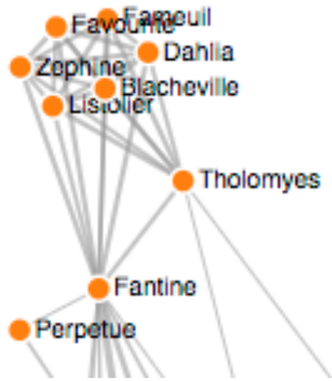
https://bl.ocks.org/heybignick/3faf257bbbbc7743bb72310d03b86ee8

Actions: Search

- what does user know?
  - target, location
- lookup
  - ex: word in dictionary
    - alphabetical order
- locate
  - ex: keys in your house
  - ex: node in network
- browse
  - ex: books in bookstore

Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*



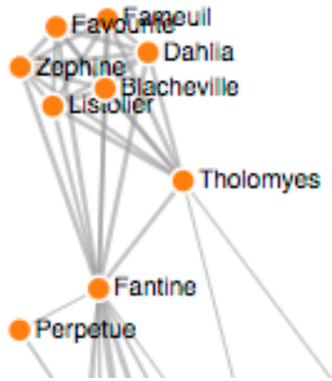
https://bl.ocks.org/heybignick/3faf257bbbbc7743bb72310d03b86ee8

Actions: Search

- what does user know?
  - target, location
- lookup
  - ex: word in dictionary
    - alphabetical order
- locate
  - ex: keys in your house
  - ex: node in network
- browse
  - ex: books in bookstore
- explore
  - ex: find cool neighborhood in new city

Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*



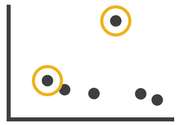
https://bl.ocks.org/heybignick/3faf257bbbbc7743bb72310d03b86ee8

Actions: Query

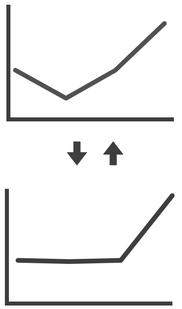
- how much of the data matters?
  - one: identify
  - some: compare
  - all: summarize

Query

→ Identify



→ Compare



→ Summarize

Actions

- independent choices for each of these three levels
  - analyze, search, query
  - mix and match

Actions

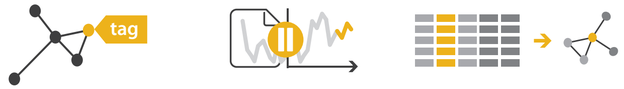
Analyze

→ Consume

→ *Discover* → *Present* → *Enjoy*

→ Produce

→ *Annotate* → *Record* → *Derive*

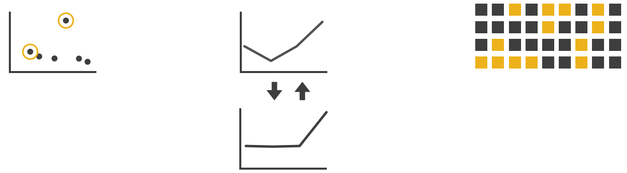


Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

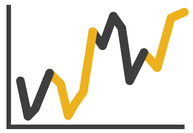
Query

→ Identify → Compare → Summarize



Task abstraction: Targets

Task abstraction: Targets

- **All Data**
  - Trends
    
  - Outliers
    
  - Features
    

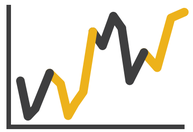
Task abstraction: Targets

All Data

- Trends

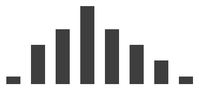
- Outliers

- Features

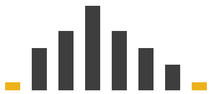


Attributes

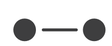
- One
  - *Distribution*



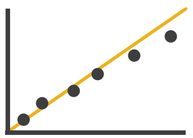
  - *Extremes*



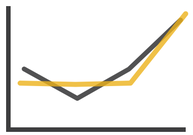
- Many
  - *Dependency*



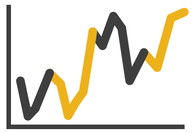
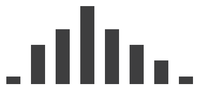
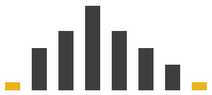
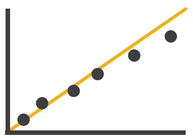
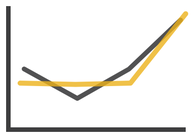
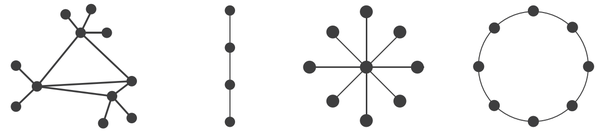
  - *Correlation*



  - *Similarity*



Task abstraction: Targets

- **All Data**
  - Trends 
  - Outliers 
  - Features 
- **Attributes**
  - One
    - *Distribution* 
    - *Extremes* 
  - Many
    - *Dependency* 
    - *Correlation* 
    - *Similarity* 
- **Network Data**
  - Topology 
  - *Paths* 

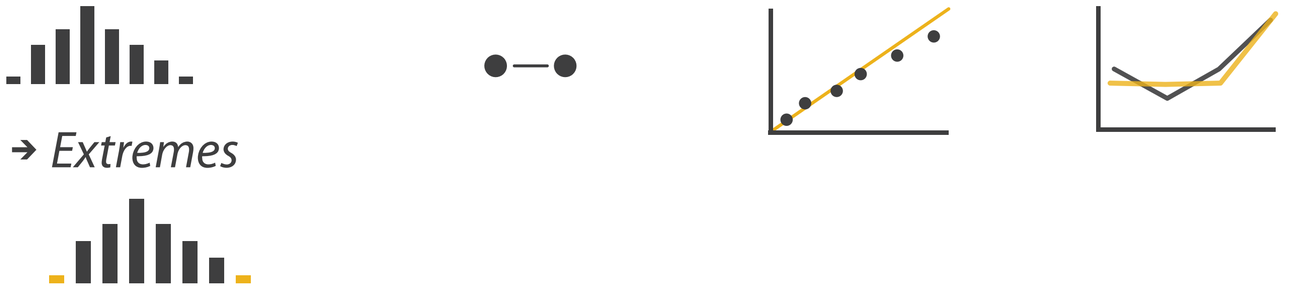
Task abstraction: Targets

All Data

- Trends
- Outliers
- Features

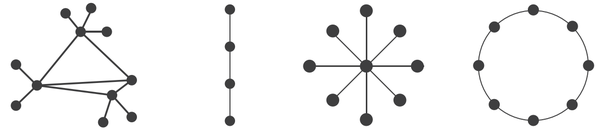
Attributes

- One
  - *Distribution*
  - *Extremes*
- Many
  - *Dependency*
  - *Correlation*
  - *Similarity*



Network Data

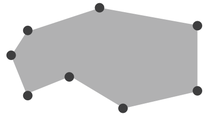
- Topology



- *Paths*

Spatial Data

- Shape

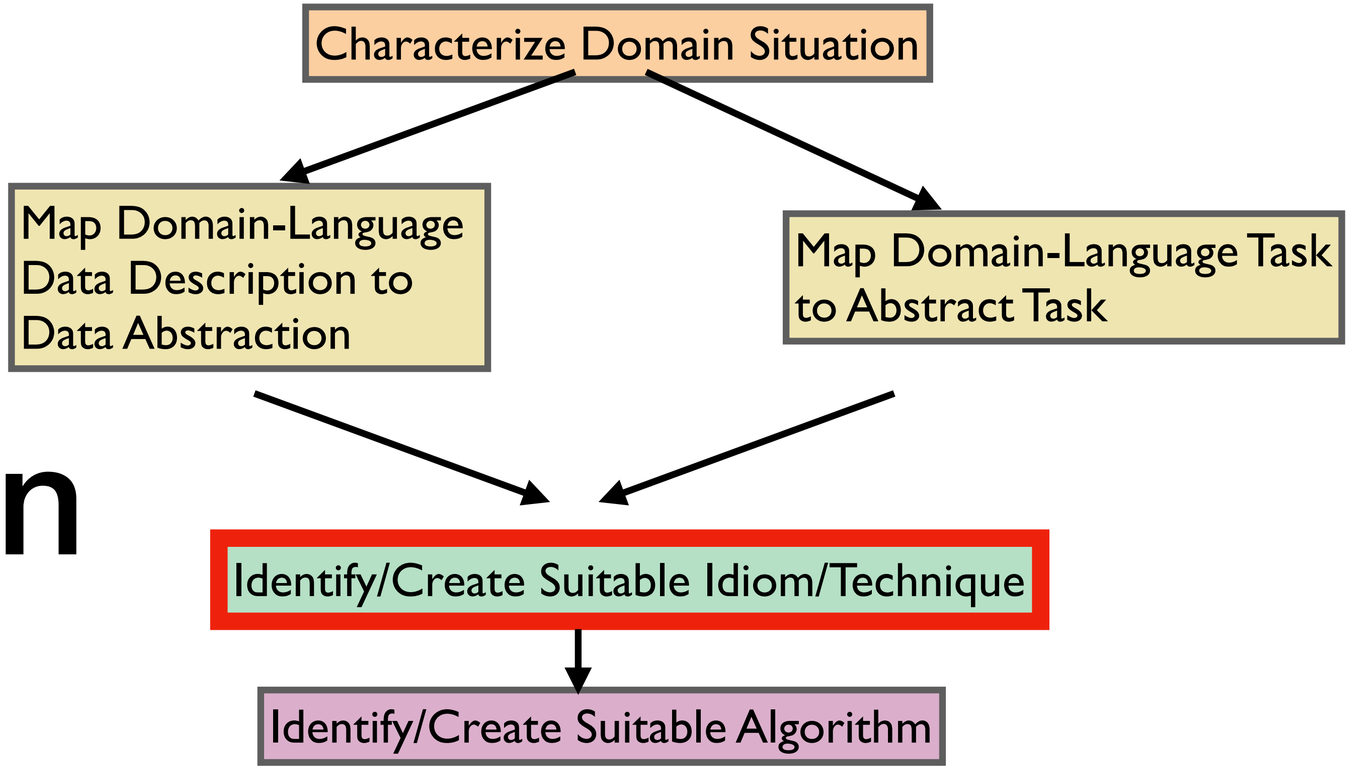


Abstraction

- these {action, target} pairs are good starting point for vocabulary
  - but sometimes you'll need more precision!
- rule of thumb
  - systematically remove all domain jargon

- interplay: task and data abstraction
  - need to use data abstraction within task abstraction
    - to specify your targets!
    - but task abstraction can lead you to transform the data
  - iterate back and forth
    - first pass data, first pass task, second pass data, ...

The visualization language



Bertin: A Semiology of Graphics

Images perceived as a set of signs

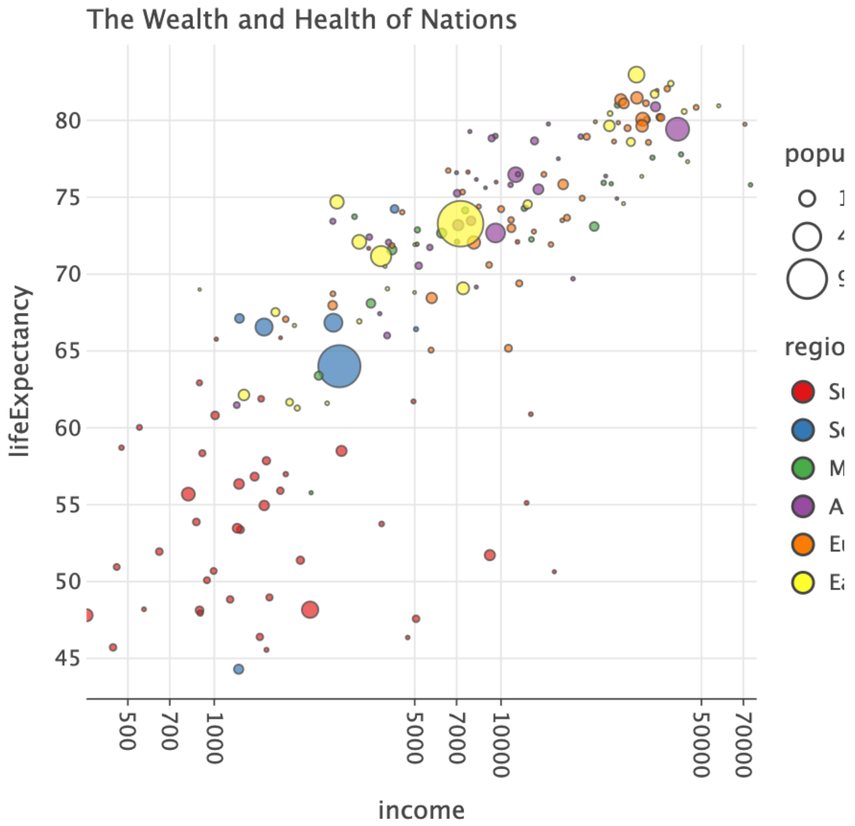
Sender encodes information in signs

Receiver decodes information from signs

- *Jacques Bertin - Sémiologie Graphique, 1967*

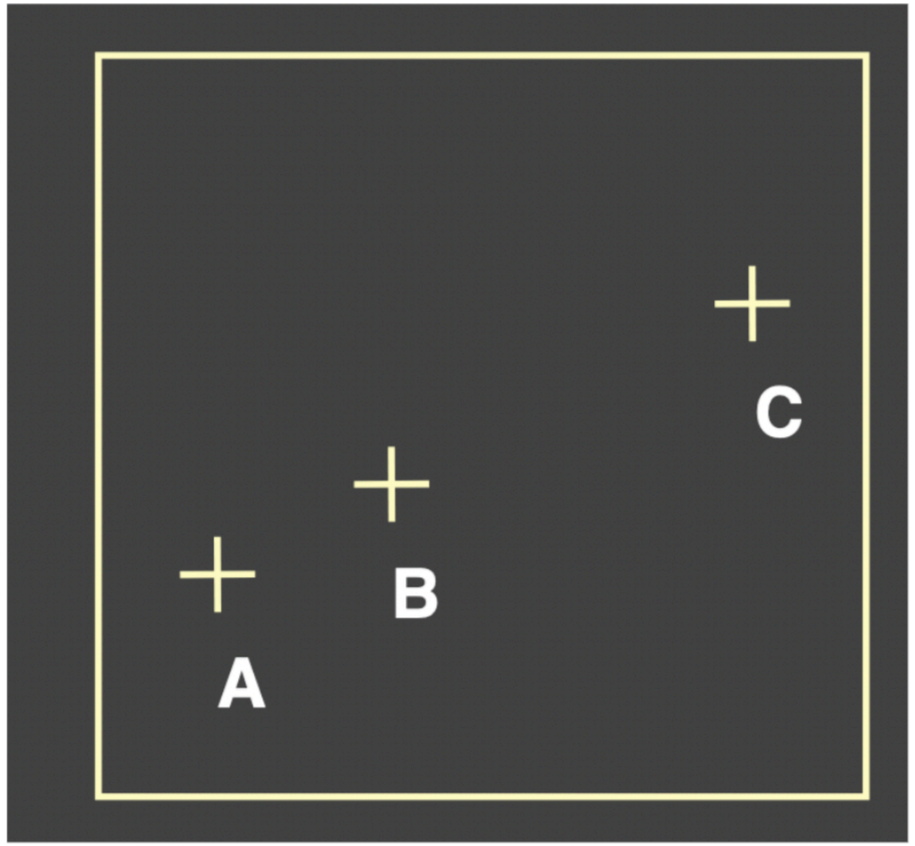
Visualization communicates information *efficiently*

	name	region	income	population	lifeExpectancy
0	Angola	Sub-Saharan Africa	5055.59	12707546	47.58
1	Benin	Sub-Saharan Africa	1457.57	8294941	61.89
2	Botswana	Sub-Saharan Africa	12282.28	1638393	55.12
3	Burkina Faso	Sub-Saharan Africa	1234.42	14761339	53.38
4	Burundi	Sub-Saharan Africa	457.07	8691005	50.95
5	Cameroon	Sub-Saharan Africa	1997.18	18054929	51.39
6	Cape Verde	Sub-Saharan Africa	3456.14	426113	71.68
7	Chad	Sub-Saharan Africa	1557.83	10541156	48.97
8	Comoros	Sub-Saharan Africa	1016.42	731281	65.77
9	Congo, Dem. Rep.	Sub-Saharan Africa	358.80	66604314	47.81
10	Congo, Rep.	Sub-Saharan Africa	3834.67	3903318	53.75
11	Cote d'Ivoire	Sub-Saharan Africa	1520.23	18373060	57.86
12	Equatorial Guinea	Sub-Saharan Africa	15342.20	562339	50.64
13	Eritrea	Sub-Saharan Africa	548.37	5028475	60.03
14	Ethiopia	Sub-Saharan Africa	812.16	78254090	55.69
15	Gabon	Sub-Saharan Africa	12704.99	1484149	60.89
16	Ghana	Sub-Saharan Africa	1382.95	23336661	56.83
17	Guinea	Sub-Saharan Africa	908.86	10211437	58.35
18	Guinea-Bissau	Sub-Saharan Africa	568.94	1502442	48.20
19	Kenya	Sub-Saharan Africa	1493.53	36529155	54.95



We can process much more information from the visualization than the table.

Visual semantics



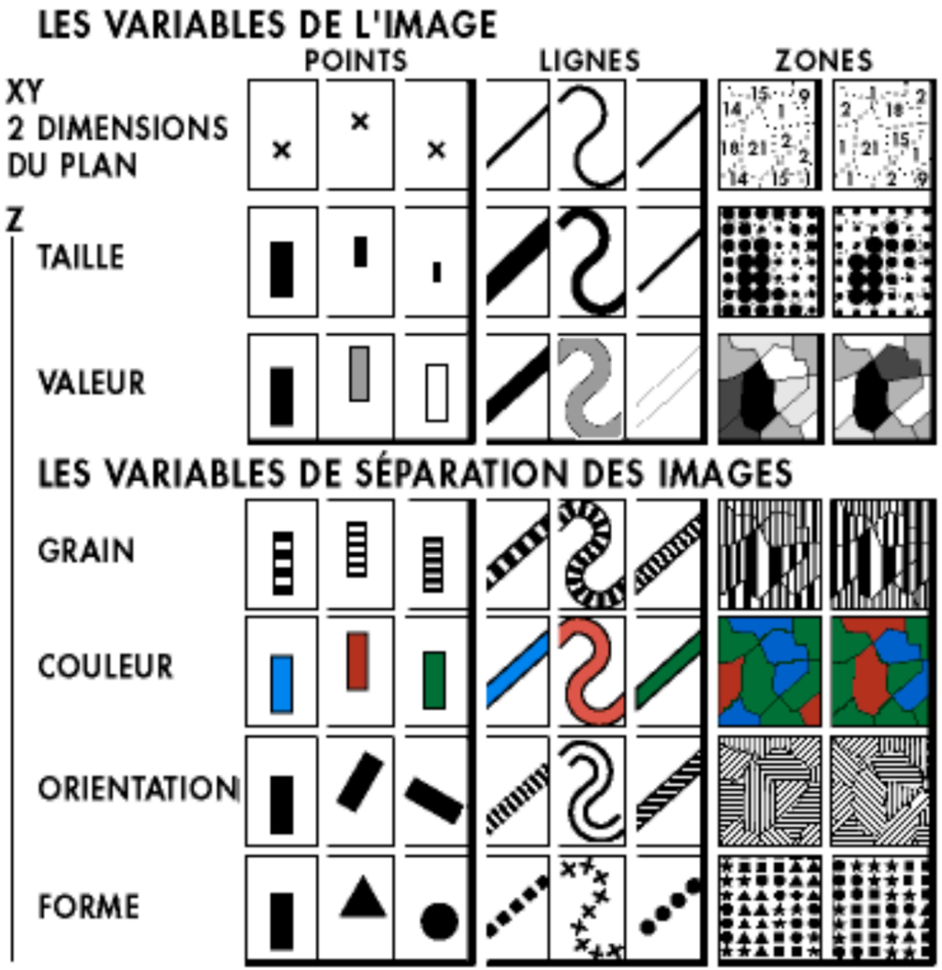
What can we understand from this plot?

From Jeffery Heer

What tools do we have to communicate through visualization?

What is our vocabulary?

Bertin’s visual variables



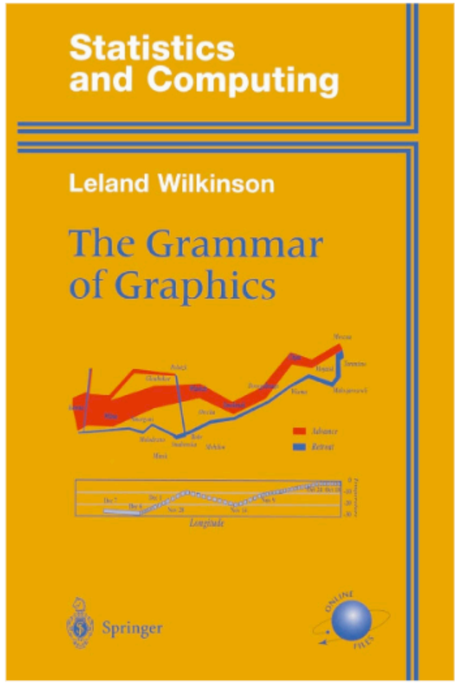
Exercise!

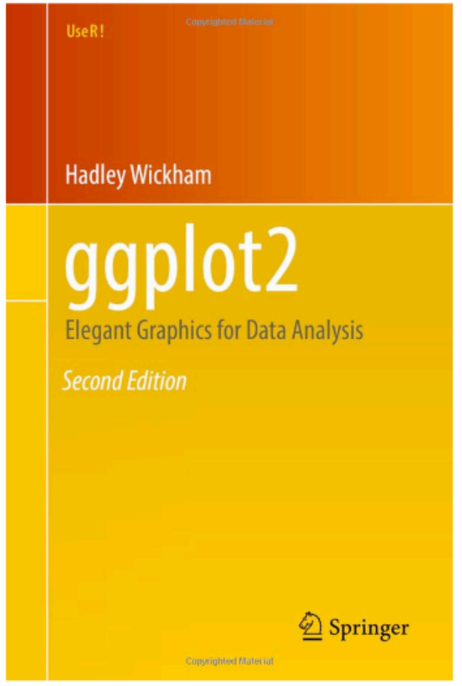
Sketch a visualization of this small dataset

A	*B*	*C*	*D*
2	3	4	a
1	2	1	a
4	5	15	b
9	10	80	b

Grammar of graphics

Grammar of graphics





Components of the grammar of graphics

Stretching the metaphor: The *parts of speech* for our visual language

- Data
- Aesthetic mappings
- Geometries
- Transforms
- Scales
- Coordinate systems
- Faceting systems
- Annotations

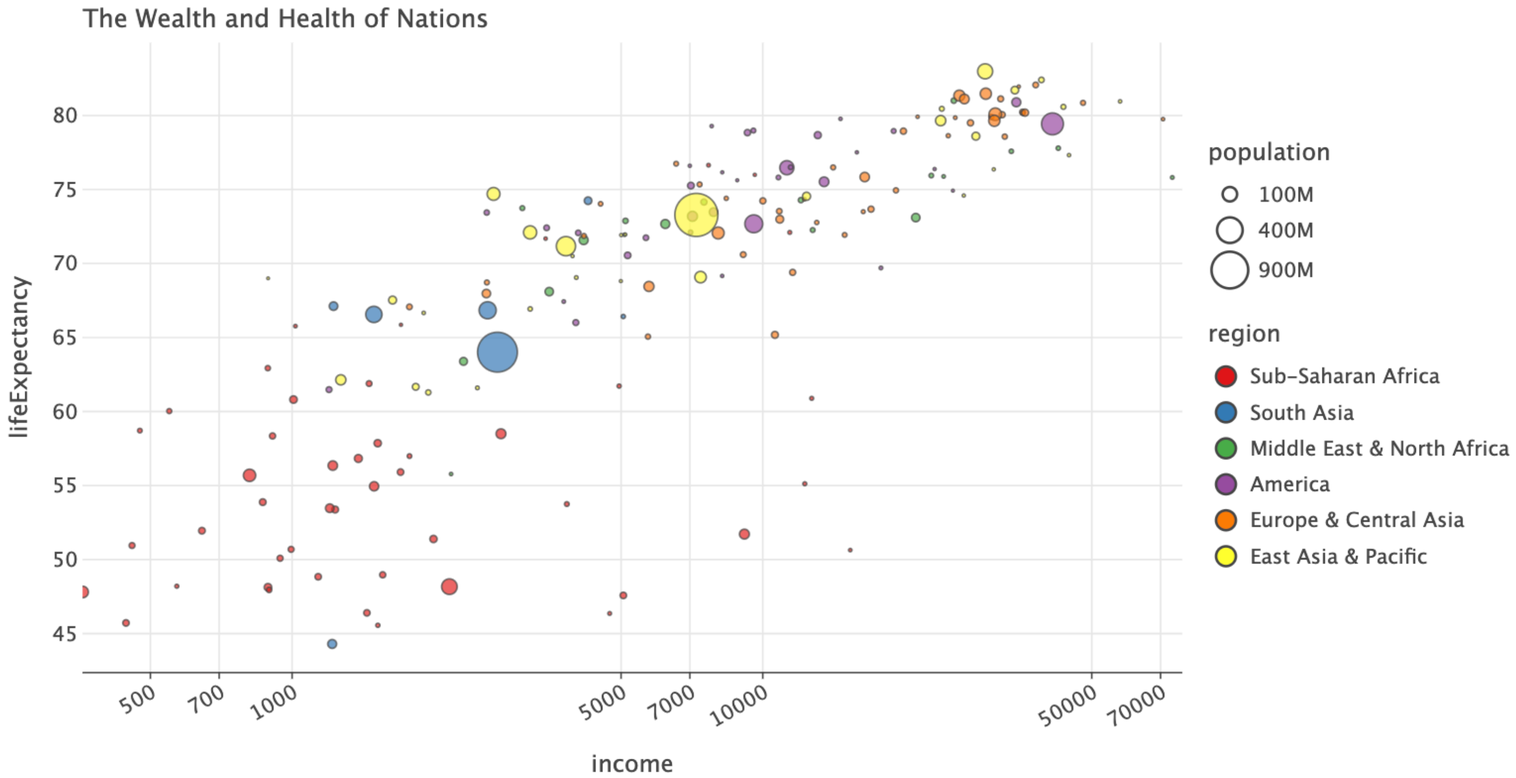
Start with tidy data

Clearly define the observations and variables in our dataset

	name	region	income	population	lifeExpectancy
0	Angola	Sub-Saharan Africa	5055.59	12707546	47.58
1	Benin	Sub-Saharan Africa	1457.57	8294941	61.89
2	Botswana	Sub-Saharan Africa	12282.28	1638393	55.12
3	Burkina Faso	Sub-Saharan Africa	1234.42	14761339	53.38
4	Burundi	Sub-Saharan Africa	457.07	8691005	50.95
5	Cameroon	Sub-Saharan Africa	1997.18	18054929	51.39
6	Cape Verde	Sub-Saharan Africa	3456.14	426113	71.68
7	Chad	Sub-Saharan Africa	1557.83	10541156	48.97
8	Comoros	Sub-Saharan Africa	1016.42	731281	65.77
9	Congo, Dem. Rep.	Sub-Saharan Africa	358.80	66604314	47.81
10	Congo, Rep.	Sub-Saharan Africa	3834.67	3903318	53.75
11	Cote d'Ivoire	Sub-Saharan Africa	1520.23	18373060	57.86
12	Equatorial Guinea	Sub-Saharan Africa	15342.20	562339	50.64
13	Eritrea	Sub-Saharan Africa	548.37	5028475	60.03
14	Ethiopia	Sub-Saharan Africa	812.16	78254090	55.69

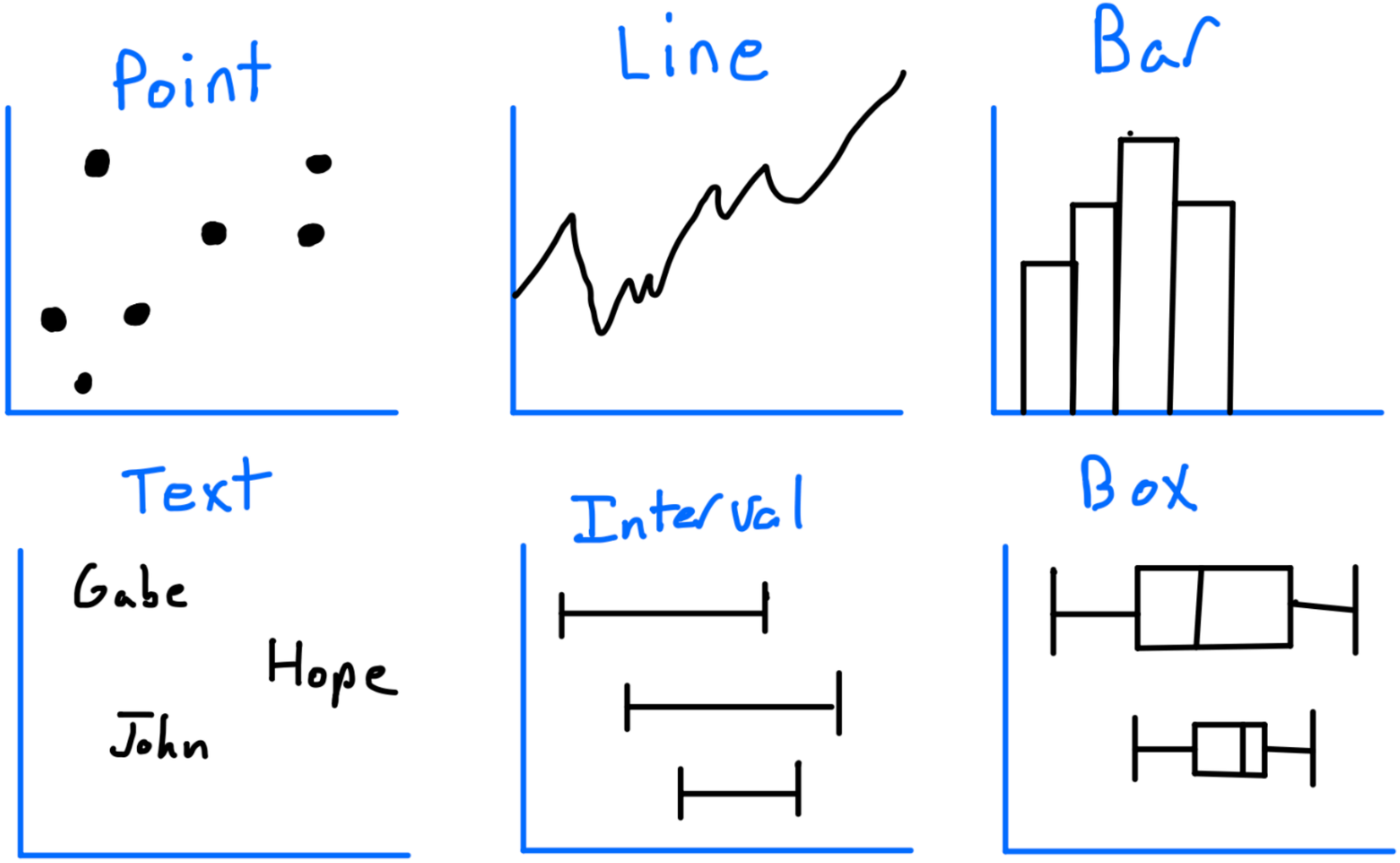
Choose a geometry

How do we represent each observation?



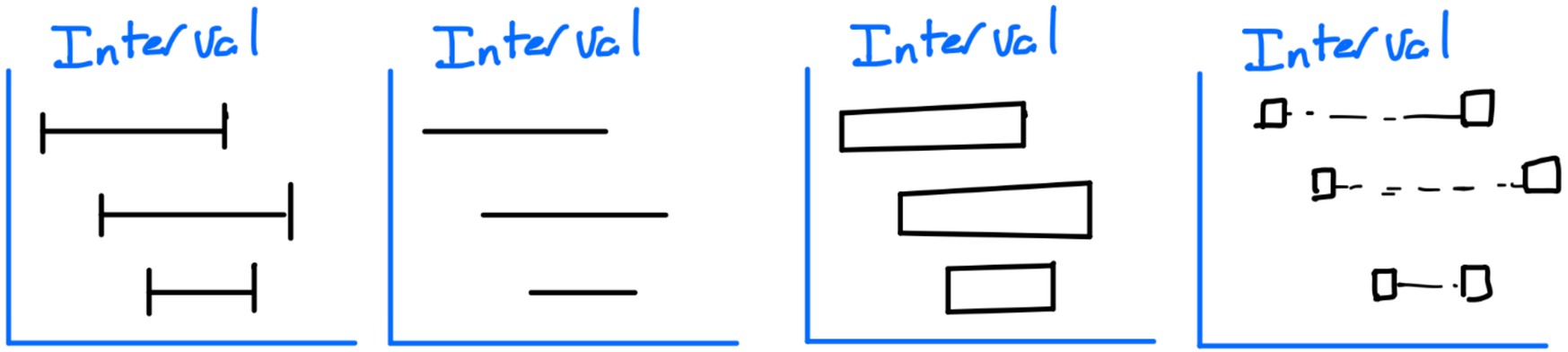
Here we are representing each observation with a distinct **point** in 2-d space (a scatterplot).

Many different geometries



Geometries are abstract

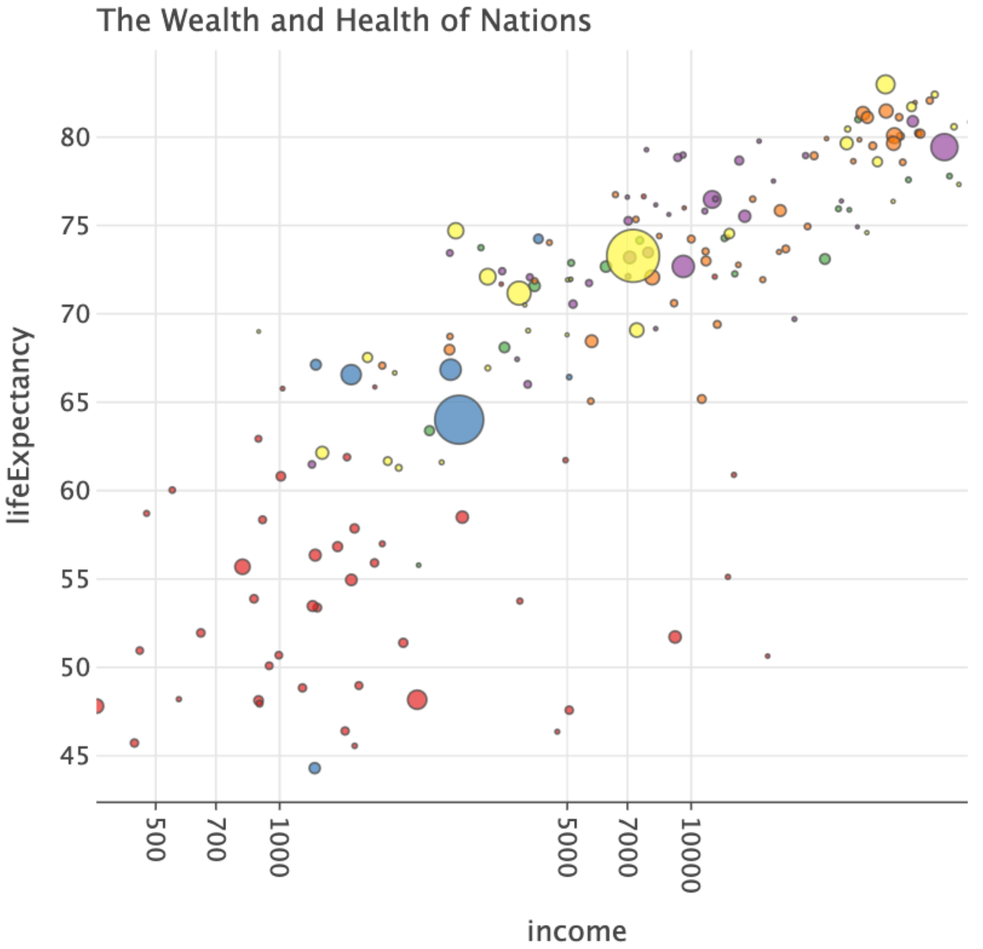
The same geometry might be rendered in different ways



Define an aesthetic mapping

Our aesthetic mapping defines how we'll map each **data dimension** (variable) to a corresponding **visual dimension** (aesthetic)

	name	region	income	population	lifeExpectancy
0	Angola	Sub-Saharan Africa	5055.59	12707546	47.58
1	Benin	Sub-Saharan Africa	1457.57	8294941	61.89
2	Botswana	Sub-Saharan Africa	12282.28	1638393	55.12
3	Burkina Faso	Sub-Saharan Africa	1234.42	14761339	53.38
4	Burundi	Sub-Saharan Africa	457.07	8691005	50.95
5	Cameroon	Sub-Saharan Africa	1997.18	18054929	51.39
6	Cape Verde	Sub-Saharan Africa	3456.14	426113	71.68
7	Chad	Sub-Saharan Africa	1557.83	10541156	48.97
8	Comoros	Sub-Saharan Africa	1016.42	731281	65.77
9	Congo, Dem. Rep.	Sub-Saharan Africa	358.80	66604314	47.81
10	Congo, Rep.	Sub-Saharan Africa	3834.67	3903318	53.75
11	Cote d'Ivoire	Sub-Saharan Africa	1520.23	18373060	57.86
12	Equatorial Guinea	Sub-Saharan Africa	15342.20	562339	50.64
13	Eritrea	Sub-Saharan Africa	548.37	5028475	60.03
14	Ethiopia	Sub-Saharan Africa	812.16	78254090	55.69
15	Gabon	Sub-Saharan Africa	12704.99	1484149	60.89
16	Ghana	Sub-Saharan Africa	1382.95	23336661	56.83
17	Guinea	Sub-Saharan Africa	908.86	10211437	58.35
18	Guinea-Bissau	Sub-Saharan Africa	568.94	1502442	48.20
19	Kenya	Sub-Saharan Africa	1493.53	36529155	54.95



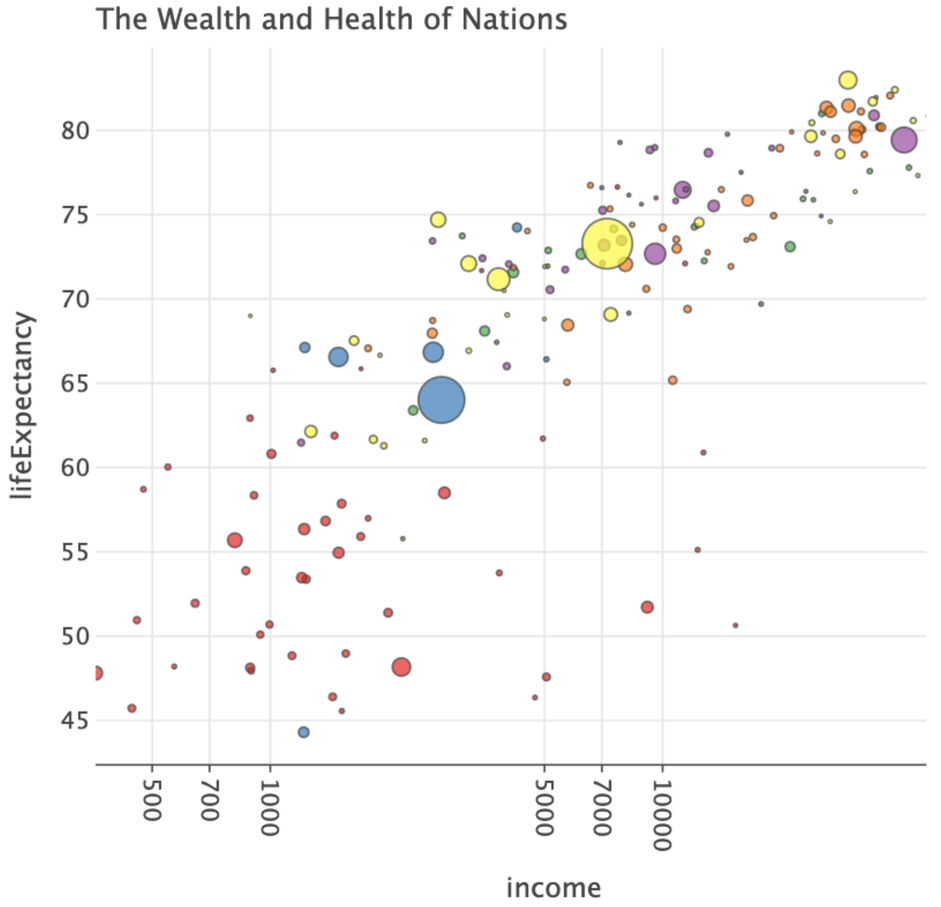
Define an aesthetic mapping

Our aesthetic mapping defines how we'll map each **data dimension** (variable) to a corresponding **visual dimension** (aesthetic)

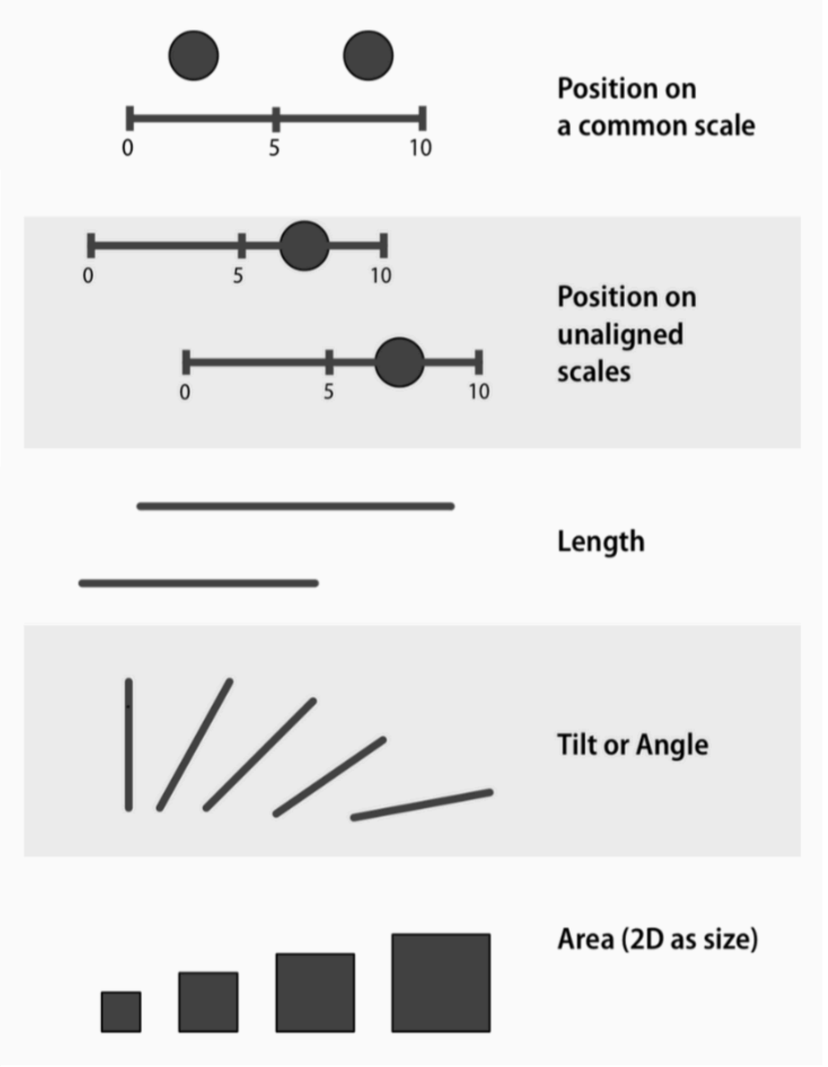
	name	region	income	population	lifeExpectancy
0	Angola	Sub-Saharan Africa	5055.59	12707546	47.58
1	Benin	Sub-Saharan Africa	1457.57	8294941	61.89
2	Botswana	Sub-Saharan Africa	12282.28	1638393	55.12
3	Burkina Faso	Sub-Saharan Africa	1234.42	14761339	53.38
4	Burundi	Sub-Saharan Africa	457.07	8691005	50.95

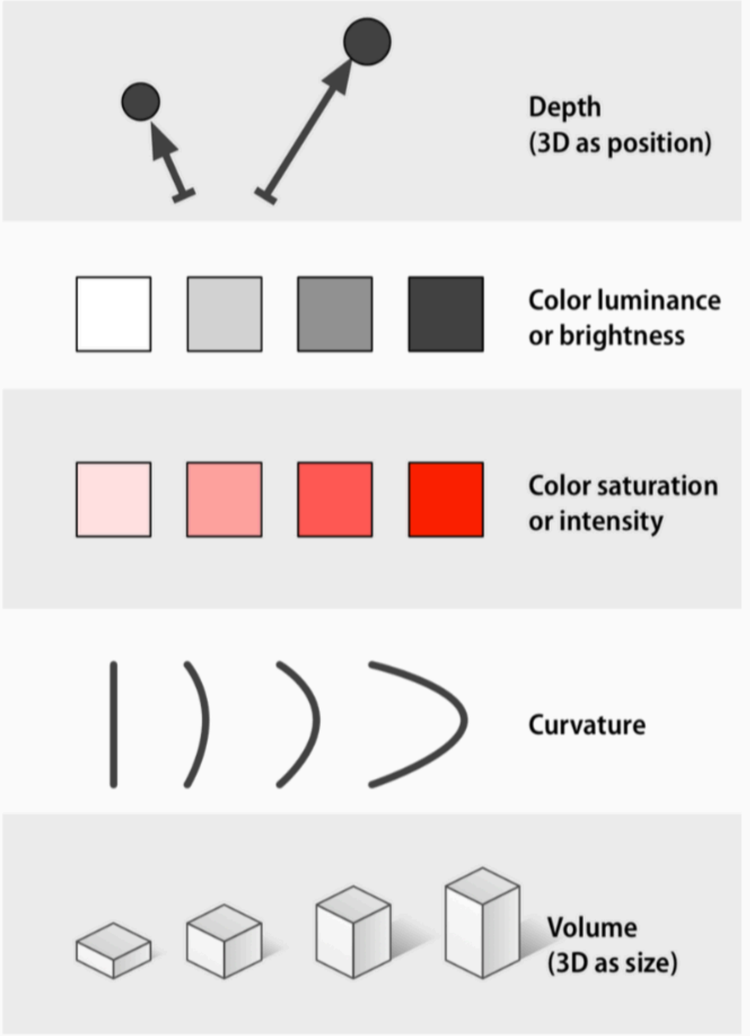
In our example:

- Income -> x position
- Life exp. -> y position
- Population -> size
- Region -> color



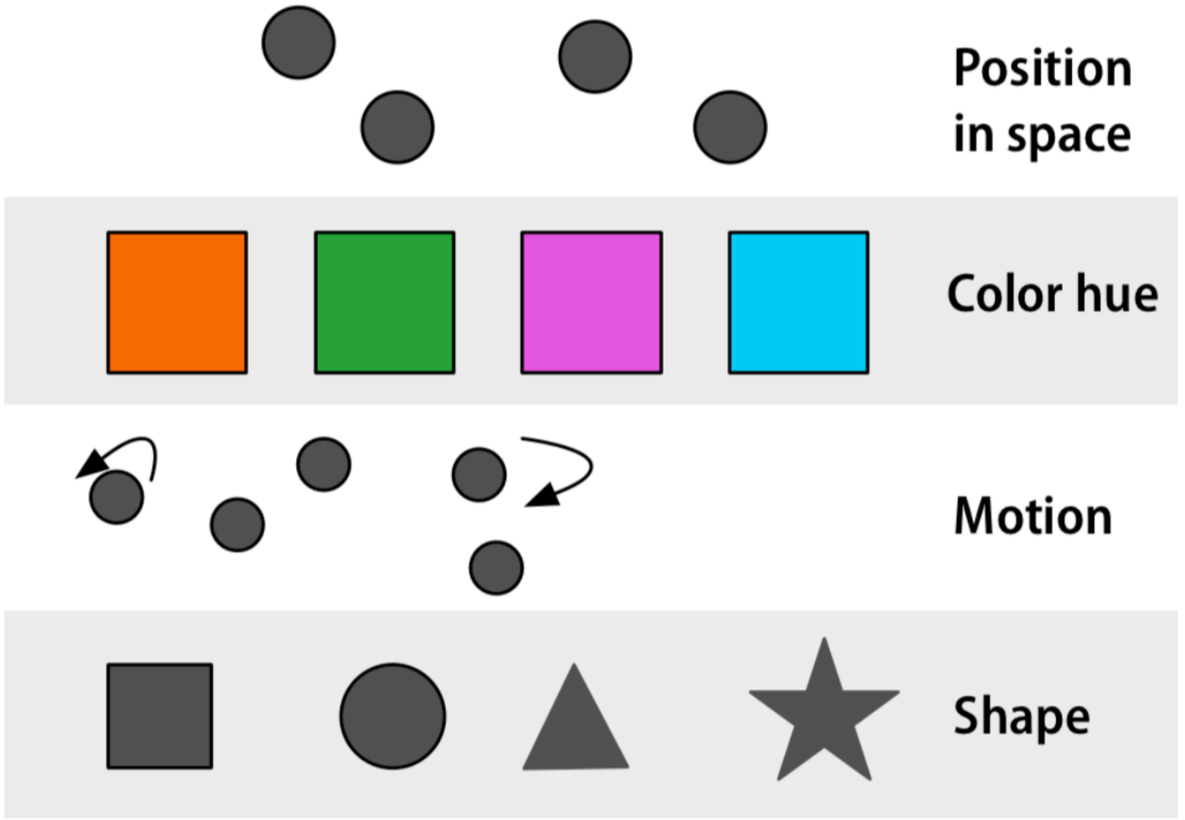
Examples of visual encoding dimensions





The set of possible encodings depends on the variable type

Categorical variables have a different space of possibilities



Geometries are abstract

Diffent geometries may have different requirements (and options) for aesthetics:

Point: x and y coordinates

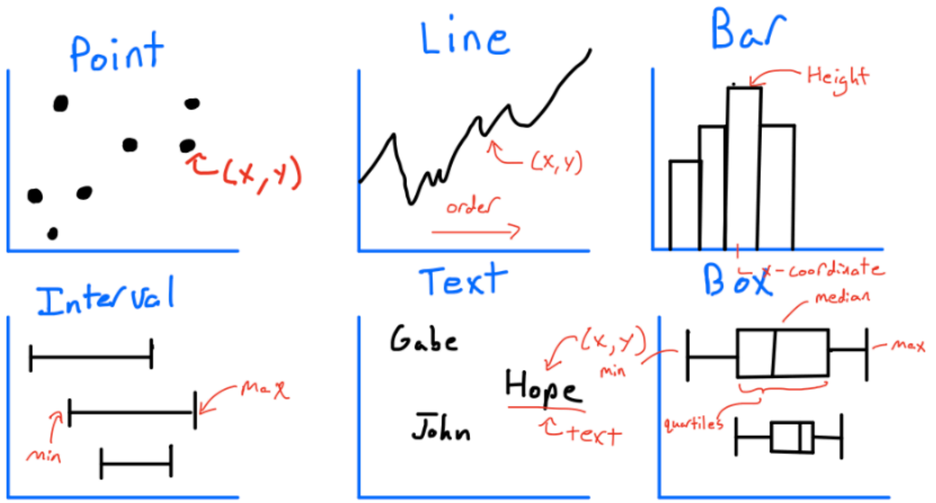
Line: x and y coordinates, order

Bar: x or y coordinate, length

Text: x and y coordinates, text

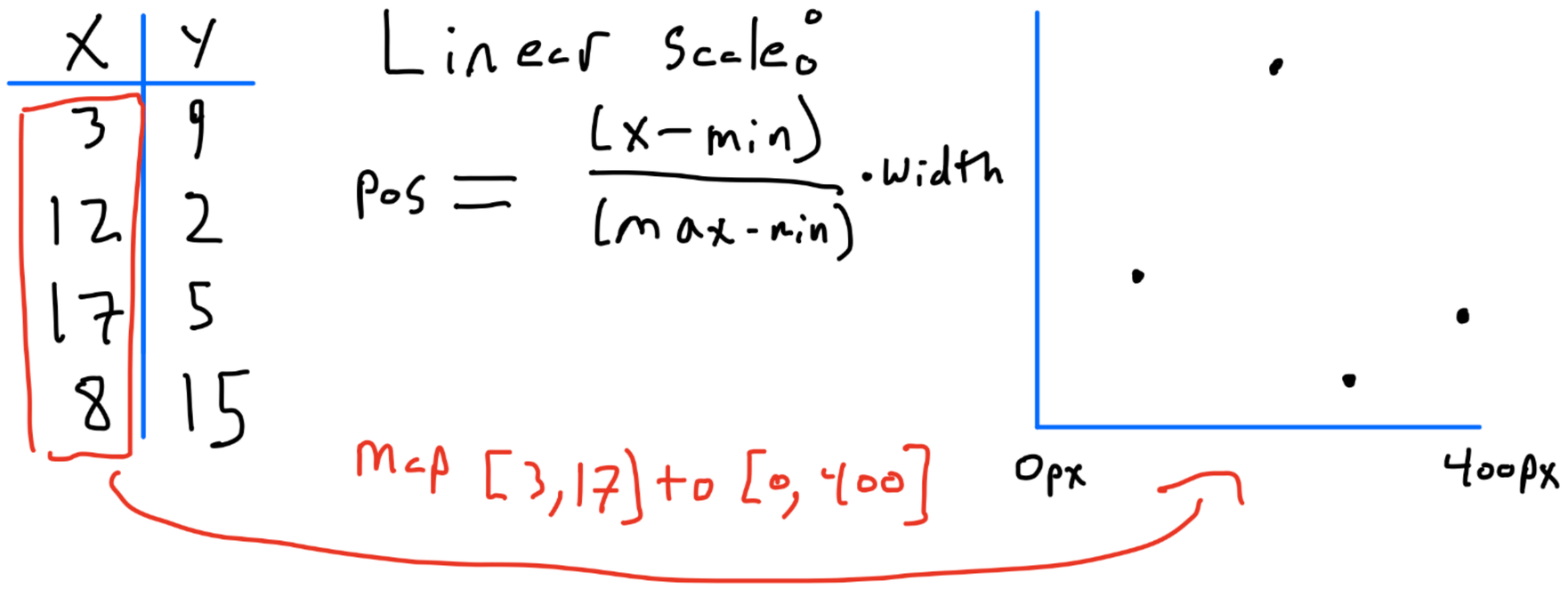
Interval: x coordinate, min, max

Boxplot: Median, quantiles



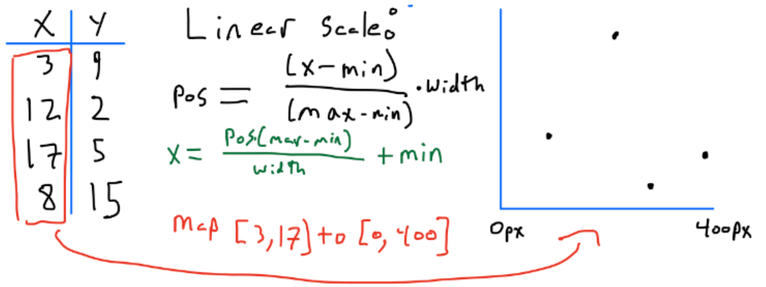
Define a scale for each mapping

A scale determines how we translate the **domain** of a variable to the **range** of a visual dimension

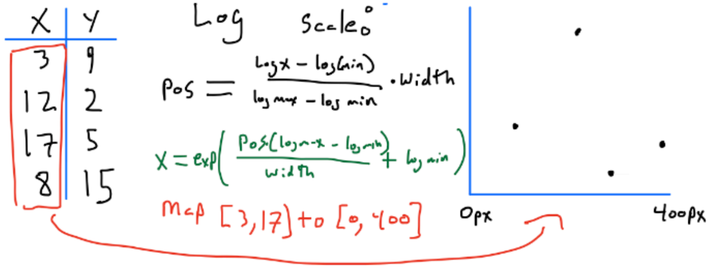


Scales are (usually) invertable functions

Most commonly a **linear** function



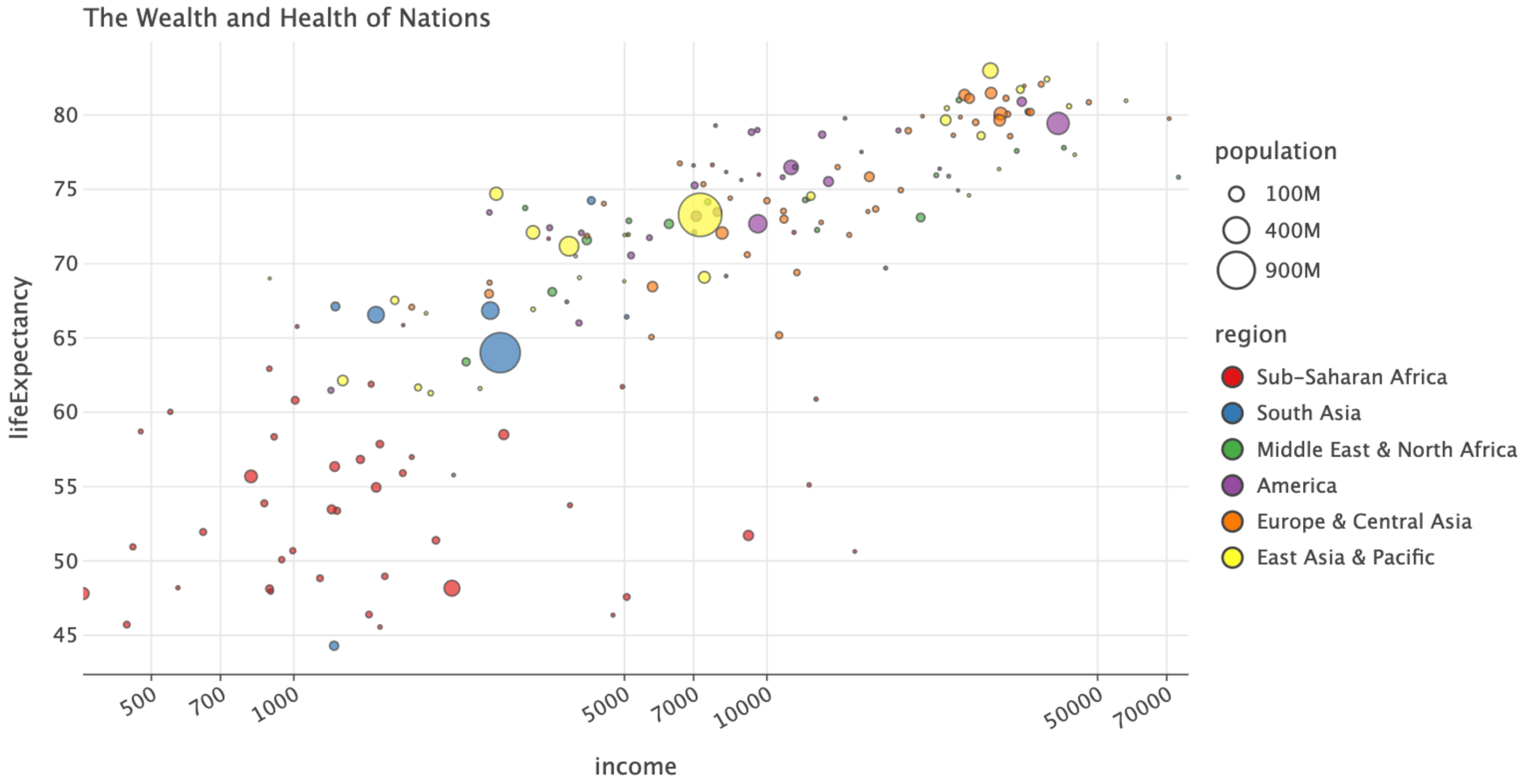
But can be **non-linear**, e.g. logarithmic



Show scales through axes and legends

Axes show the scale for each *position* dimension.

Legends show the scale for other dimensions (color, shape, size, etc.)



Some scales are complicated!

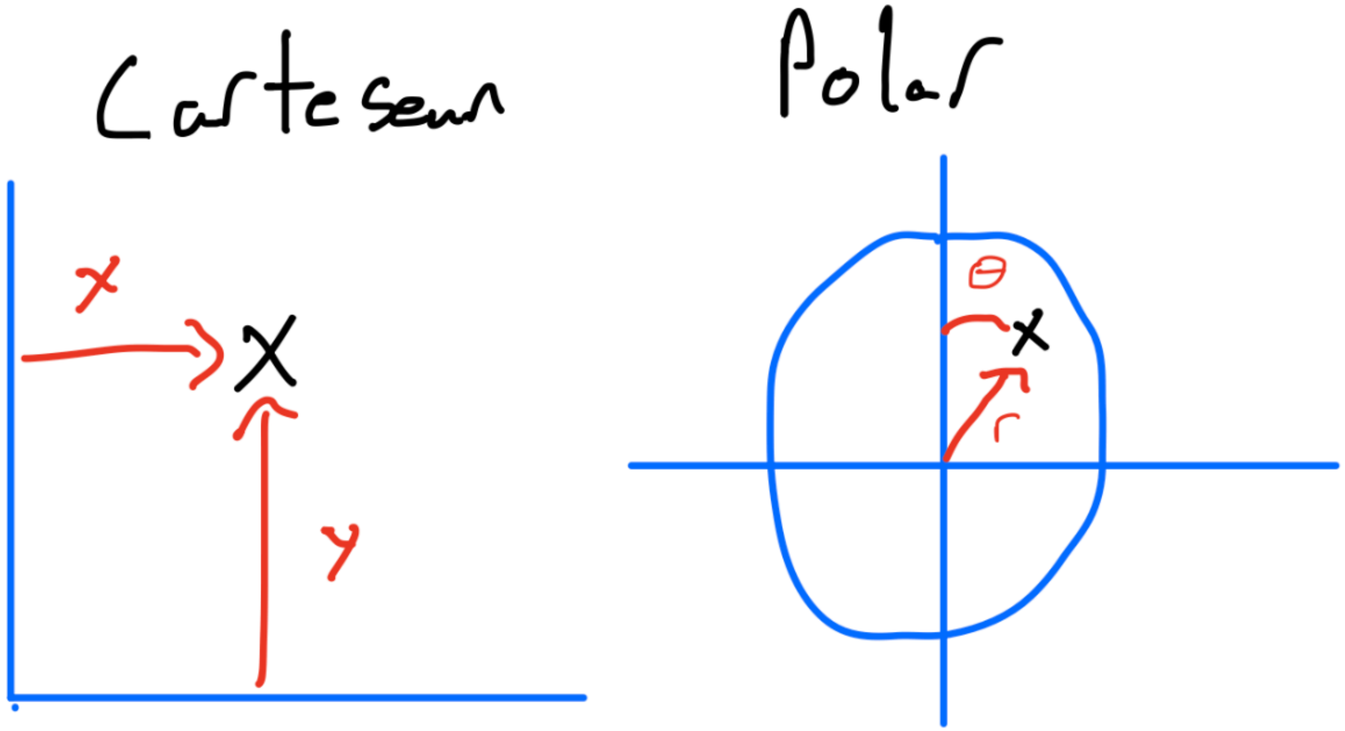
Color is especially tricky.

- Humans percieve color non-linearly
- Not all humans percieve the same set of colors
- One color can encode multiple channels

We'll cover this in a later lecture!

(Optional) Define a coordinate system

A **coordinate system** defines how the range of each scale is represented. Usually Cartesian coordinate system, but others are possible, such as *polar*

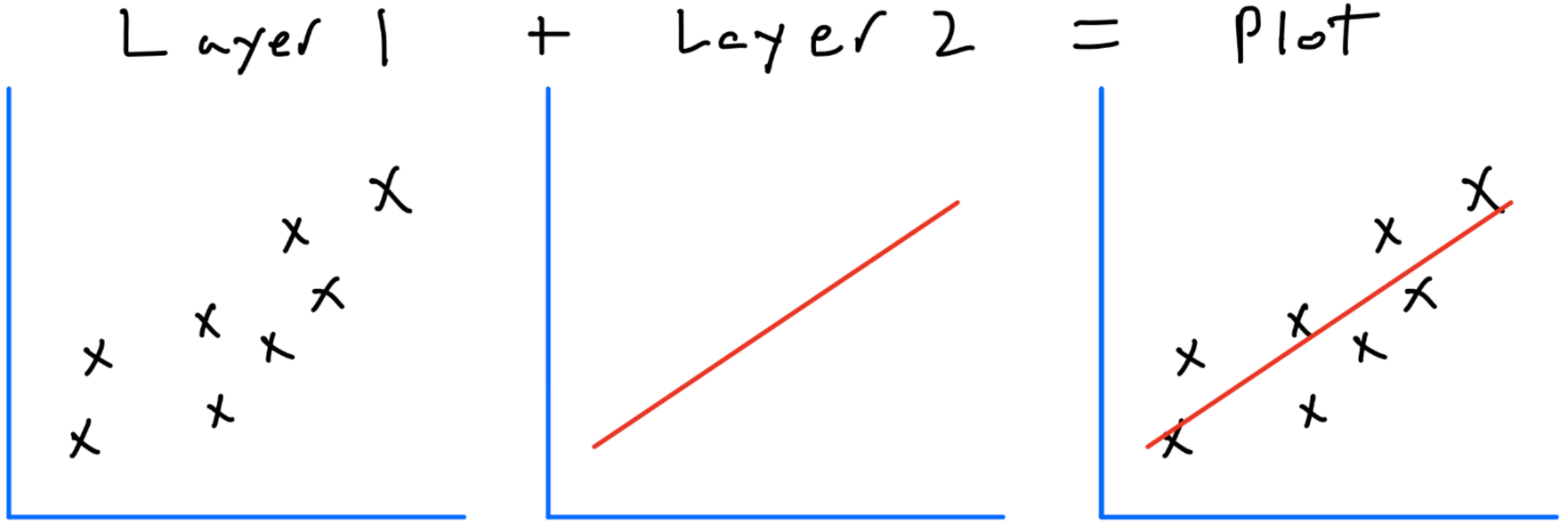


A Layered Grammar of Graphics

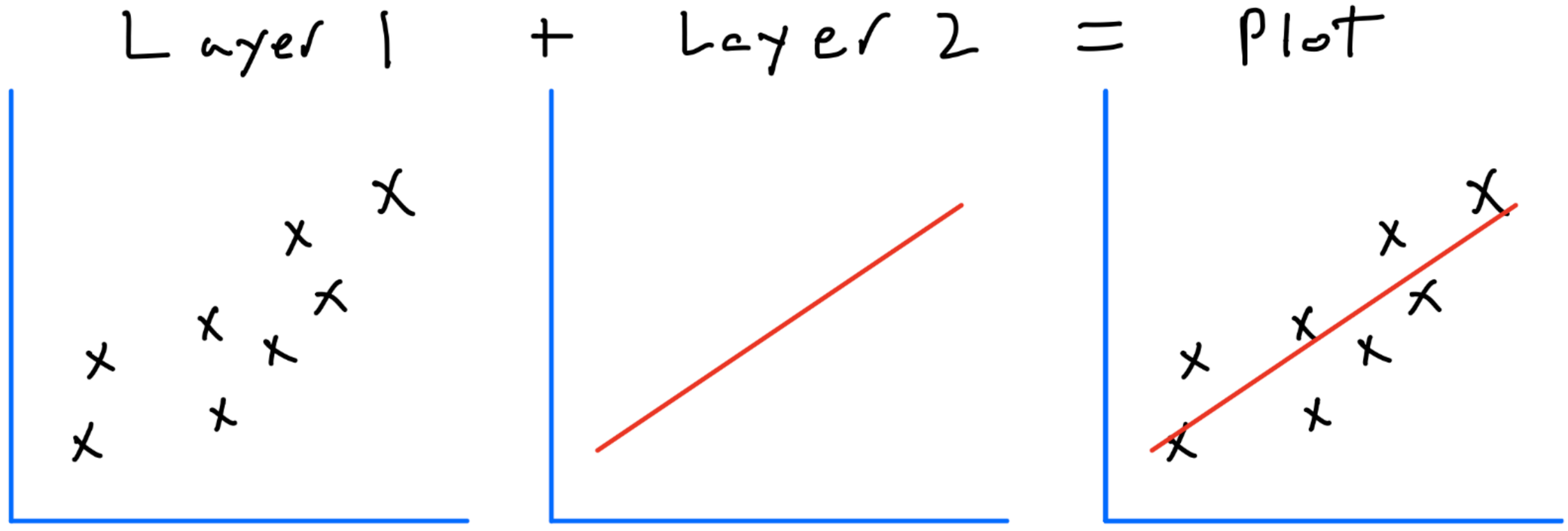
Hadley Wickham

A layer is a single view of the data

Encompasses data + mapping + geometry. A plot may have more than one layer to show multiple properties of a dataset or multiple datasets.



Scales are (generally) shared across layers



Finally: add annotations

Annotations, such as titles captions and highlights, give additional context and information to the viewer.