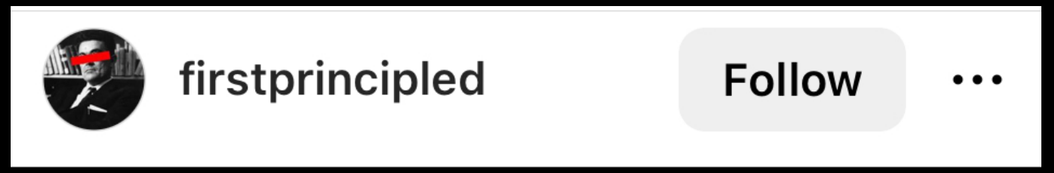
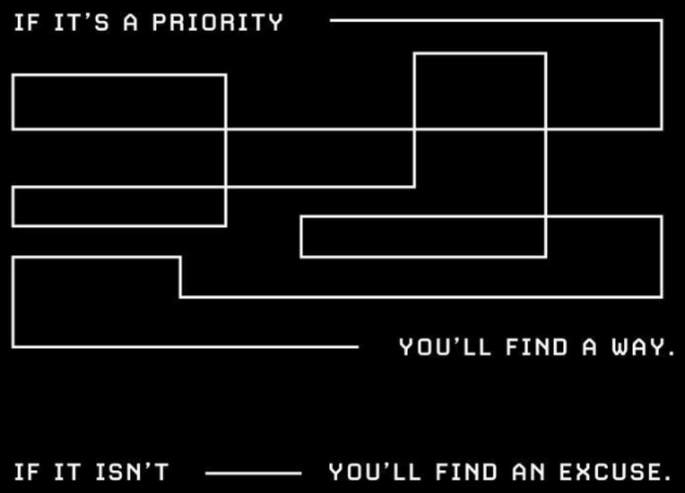
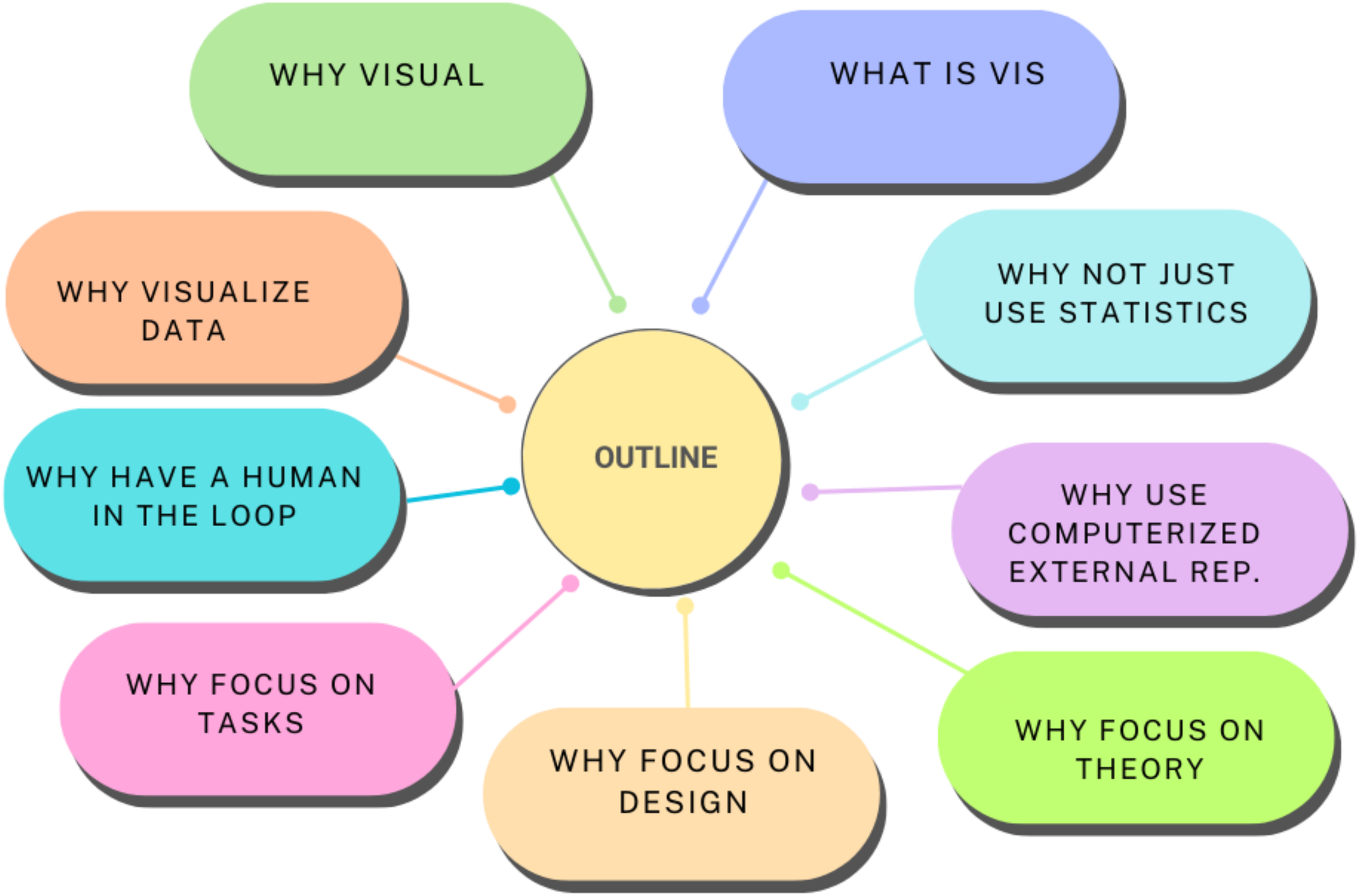

firstprincipled
Follow


IF IT'S A PRIORITY
YOU'LL FIND A WAY.
IF IT ISN'T
YOU'LL FIND AN EXCUSE.

Visualization for Data Science

Data Abstraction

Visualization Grammar


WHY VISUAL
WHAT IS VIS
WHY VISUALIZE DATA
WHY NOT JUST USE STATISTICS
OUTLINE
WHY HAVE A HUMAN IN THE LOOP
WHY USE COMPUTERIZED EXTERNAL REP.
WHY FOCUS ON TASKS
WHY FOCUS ON DESIGN
WHY FOCUS ON THEORY

Clicker Question (Select all that are correct)

A. There are six quizzes in the course, you can drop the lowest two
B. The project labs are required and are worth 5% of your course grade
C. Lectures are optional, attend when you feel like it
D. There are three design challenges, they have nothing to do with design studios.
E. You can use Claudette on your final project

Clicker Question (Select all that are correct)

A. You need to schedule a reservation on PraireTest to book a slot for each quiz.
B. The first quiz is this week and covers Altair Fundamentals
C. Dr. K. will be out of town October 20 – 27 so there is no class ☺
D. The best way to contact Dr. K is by Canvas.
E. Workday is a wonderful tool

Weekly Rhythm

Knowledge Building

- **Prep** activities (readings, videos, coding) ~90 mins/week before lecture
- **Lectures**: interactive, semi-flipped, with clicker Qs + skill practice
- **Weekly quizzes** (50 mins, CBTF) track progress on programming + theory

Explore & Create

- **Design Studios**: hands-on workshops for sketching, critique, design thinking
- **Design Artifacts**: sketches, wireframes, reviews — showcase growth (2–4 hrs follow-up)

Capstone Journey

- **Project Labs**: team-based sessions → prototypes & real datasets
- **Project Milestones**: final project (dashboard + presentation) integrating design, theory & programming

In the first two weeks expect to spend more time on the programming aspects of the course

Clicker Question

Which one should we use for the course icon

A

B

Clicker Question (Select all that are correct)

In Python, what are two key differences between a list and a dictionary?

A. Lists can store multiple data types, while dictionaries can only store one data type.

B. Lists are ordered collections of elements, while dictionaries are unordered collections.

C. Lists use numeric indices to access elements, while dictionaries use keys.

D. Lists are immutable, while dictionaries are mutable.

Clicker Question (Select all that are correct)

```
numbers = [1, 2, 3, 4, 5]


for num in numbers
    if num %% 2 == 0:
        numbers.append(num * 2)
print(numbers)
```

The function is supposed to append the square of values that are even to the list, there are a number of bugs, select all areas that have bugs

A. The for loop's header
B. The if statement condition
C. The operation inside the append statement
D. The loop's behavior during execution
E. The indentation of the append statement

Administrivia: Quiz 1

Quiz One is this week – book on **PraireTest**

- It must occur from Wednesday – Friday
- It covers Python and Pandas you need to have worked through the Jumpstart and Tutorial 1 notebooks.
- Python
  - Multiple choice questions on Python (functions, lists, dictionaries, strings, etc)
  - Coding questions on Python
- Pandas
  - A Jupyter notebook on Pandas wrangling (data overview, filtering, creating new columns, handling missing values, etc)
  - The book takes a while to load (about a minute)

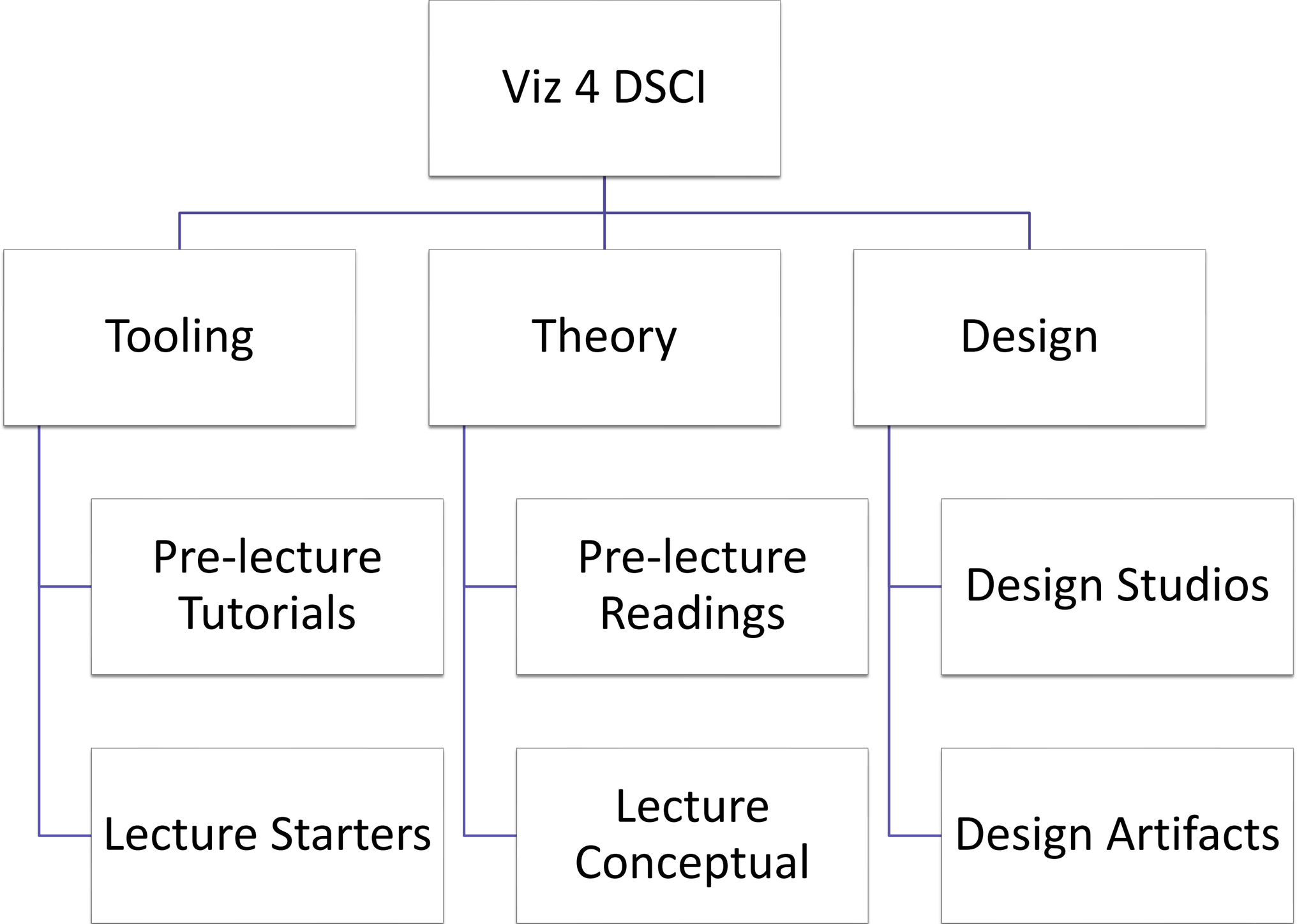
Take Quiz 0 to give you a sense of what it will look like (PraireLearn env.)

If you have CFA accommodations, make sure your time extensions have been configured before you book a slot

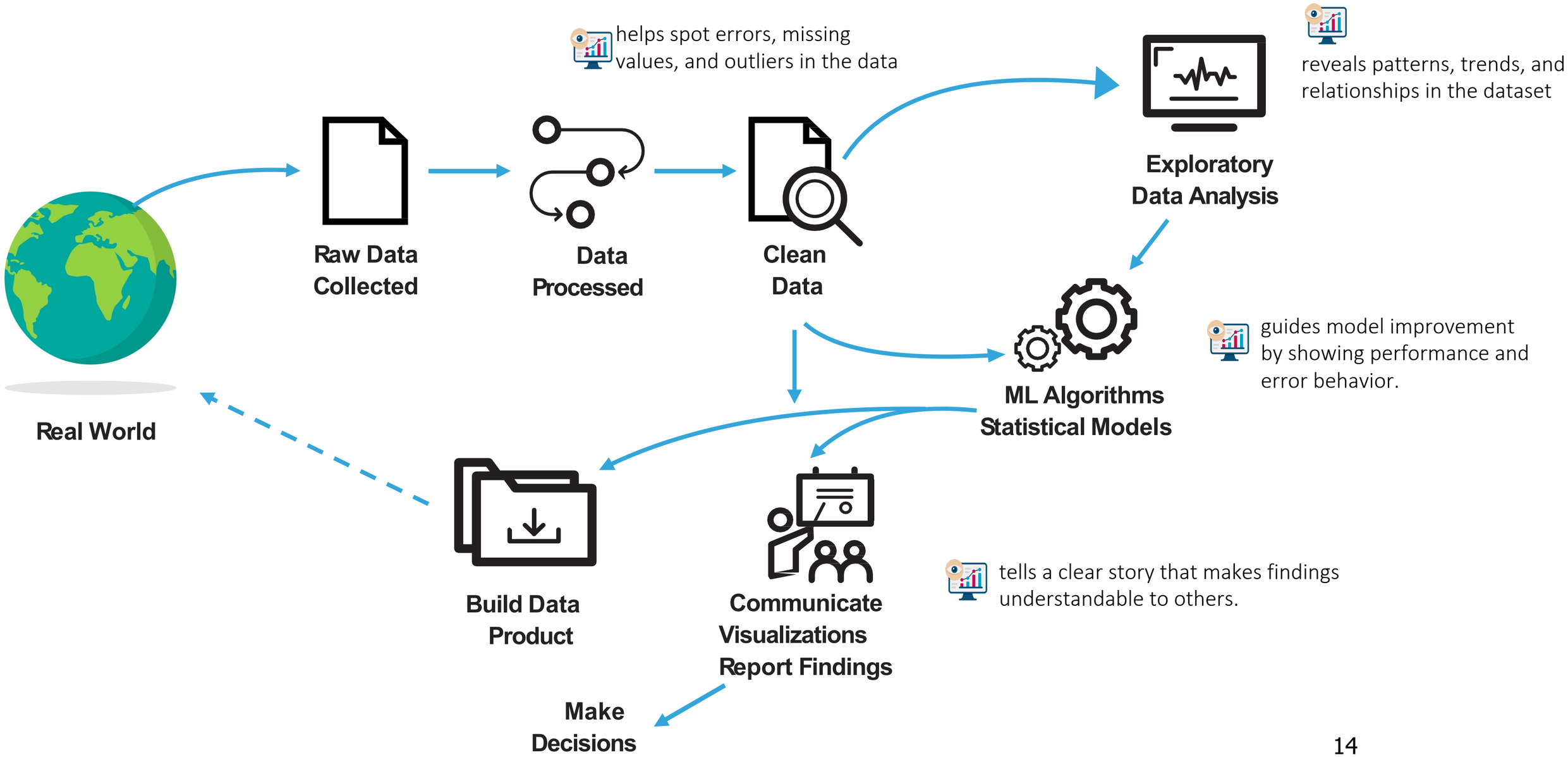
Administrivia: Project Lab 1

- Lab begins this week.
- You must attend the lab section you registered for.
- This week's lab will focus on introductions and providing an overview of the project.
- Attendance is required for your success.

“By the end of this course, my goal is for you to have a toolkit to design, create, critique, and interpret interactive visualizations.


Viz 4 DSCI
Tooling
Theory
Design
Pre-lecture Tutorials
Pre-lecture Readings
Design Studios
Lecture Starters
Lecture Conceptual
Design Artifacts

What's the Role of Visualization in Data Science?



Originally from Alexander Lex. Updated to include viz roles

Learning Outcomes

- Describe the difference between how the phrases “dataset types”, “data types” are used in vis. Literature as opposed to programming
- Describe the characteristics of data
- Differentiate between the different types of data and dataset types
- Describe the basic visual primitives of visualizations (marks and channels)
- Differentiate between a mark and channel

Deconstruct a visualization based on its marks and channels

Data Characterization

Data is characterized by its

- Size (volume)
- Speed at which it generated (velocity)
- Quality (veracity)
- Structure

What does data mean?

Basil, 7, S, Pear

What does data mean?

Basil, 7, S, Pear

What about this data?

- food shipment of produce (basil & pear) arrived in satisfactory condition on 7th day of month
- Basil Point neighborhood of city had 7 inches of snow cleared by the Pear Creek Limited snow removal service
- lab rat Basil made 7 attempts to find way through south section of maze, these trials used pear as reward food

Semantics

semantics: real-world meaning

data types: structural or mathematical interpretation of data

- item, link, attribute, position, (grid)
- **different from data types in programming!**

Name	Age	Shirt Size	Favorite Fruit
Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

Items & Attributes

- item: individual entity, discrete
  - eg patient, car, stock, city
  - "independent variable"

Name	Age	Shirt Size	Favorite Fruit
Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

item: person

Items & Attributes

- item: individual entity, discrete
  - eg patient, car, stock, city
  - "independent variable"
- attribute: property that is measured, observed, logged...
  - eg height, blood pressure for patient
  - eg horsepower, make for car
  - "dependent variable"

Name	Age	Shirt Size	Favorite Fruit
Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

item: person

Items & Attributes

- item: individual entity, discrete
  - eg patient, car, stock, city
  - "independent variable"
- attribute: property that is measured, observed, logged...
  - eg height, blood pressure for patient
  - eg horsepower, make for car
  - "dependent variable"

attributes: name, age, shirt size, fave fruit

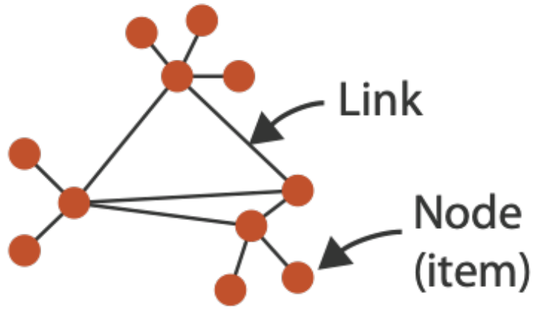
Name	Age	Shirt Size	Favorite Fruit
Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

item: person

Other data types

Links

- express relationship between two items
- e.g. friendship on facebook, interaction between proteins

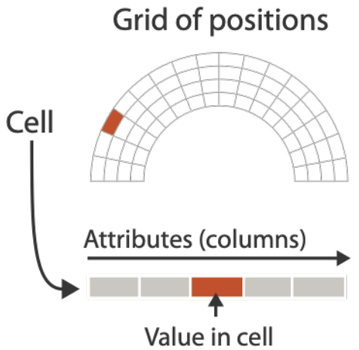


Positions

- spatial data: location in 2D or 3D
- e.g.pixels in photo, voxels in MRI scan, latitude/longitude

Grids

- sampling strategy for continuous data



Data Types are the fundamental units in which observed **phenomena** are **represented**. The structural or mathematical interpretation of data

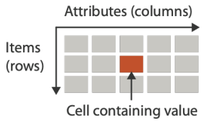
Data Types

➔ Items ➔ Attributes ➔ Links ➔ Positions ➔ Grids

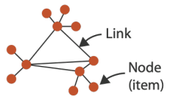
Dataset is any collection of information that is the target of analysis

Dataset Types

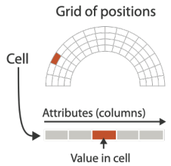
➔ Tables



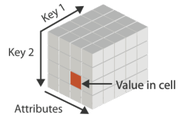
➔ Networks



➔ Fields (Continuous)



➔ *Multidimensional Table*



➔ *Trees*

➔ Geometry (Spatial)

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

Items

Attributes

flat table

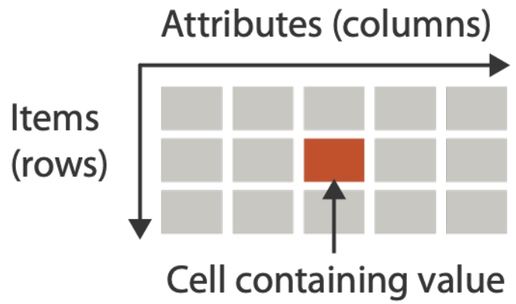
- one item per row
- each column is attribute
- cell holds value for item-attribute pair
- **unique key (could be implicit)**

attributes: name, age, shirt size, fave fruit

ID	Name	Age	Shirt Size	Favorite Fruit
1	Amy	8	S	Apple
2	Basil	7	S	Pear
3	Clara	9	M	Durian
4	Desmond	13	L	Elderberry
5	Ernest	12	L	Peach
6	Fanny	10	S	Lychee
7	George	9	M	Orange
8	Hector	8	L	Loquat
9	Ida	10	M	Pear
10	Amy	12	M	Orange

item: person

➔ Tables



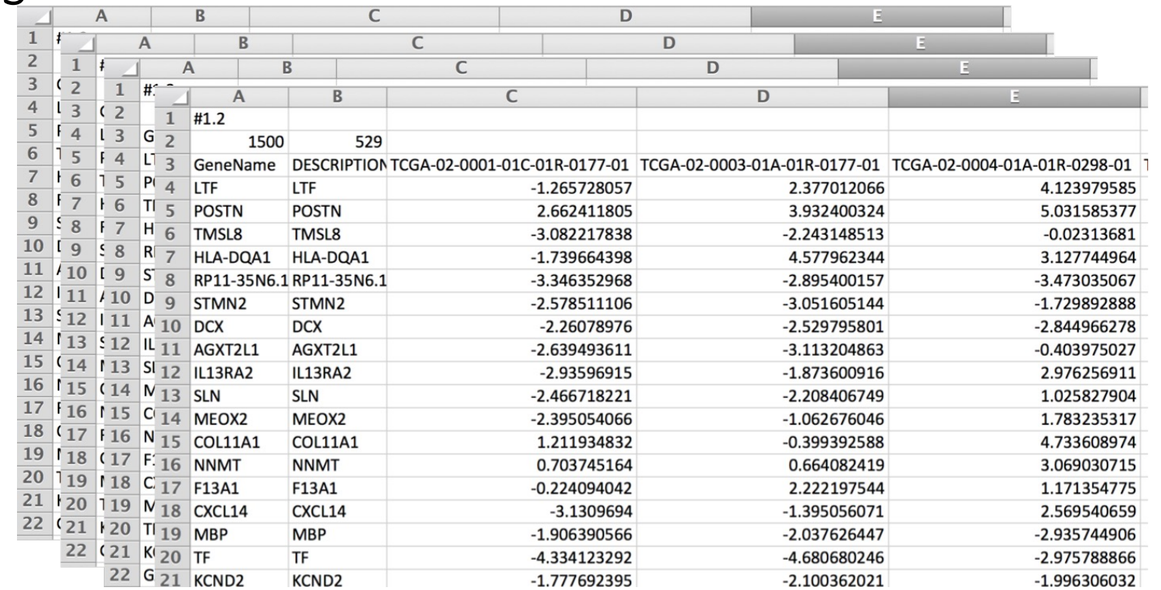
Dataset types

Tables

Items

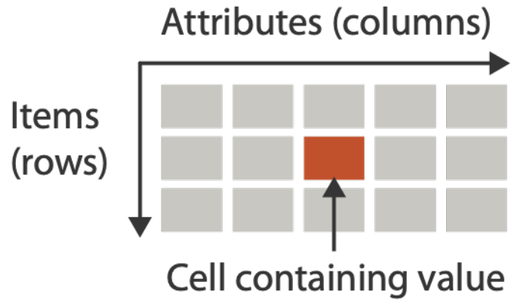
Attributes

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

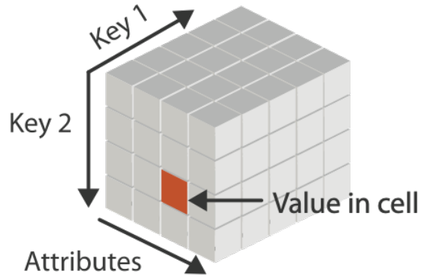


	A	B	C	D	E
1	#1.2				
2	1500	529			
3	GeneName	DESCRIPTION	TCGA-02-0001-01C-01R-0177-01	TCGA-02-0003-01A-01R-0177-01	TCGA-02-0004-01A-01R-0298-01
4	LTF	LTF	-1.265728057	2.377012066	4.123979585
5	POSTN	POSTN	2.662411805	3.932400324	5.031585377
6	TMSL8	TMSL8	-3.082217838	-2.243148513	-0.02313681
7	HLA-DQA1	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

➔ Tables



➔ *Multidimensional Table*



Clicker Question

What is A?

What is B?

What is C?

A – Attribute

B – Item

C – Cell

D – Data type

E – Dataset type

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	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

Attribute Types

➔ Categorical

➔ Ordered

➔ *Ordinal*

➔ *Quantitative*

Data Model vs. Conceptual Model

Data Models

- Formal descriptions (math-like)
- Example: integers with + and ×
- Focus: structure and operations

Conceptual Models

- Mental constructions we build around the data
- Include meaning (semantics) and reasoning

37.0 → data model = float; conceptual model = temperature

[2.0, 5.0, 1.0] → data model = 3 floats, list, array; conceptual model = spatial location

Data Attribute Types: Nominal, Ordinal, Quantitative

N – Nominal or Categorical

- Labels or categories
- Operations: =, ≠
- Example: colors, types of fruit, shapes

O – Ordered (Ordinal)

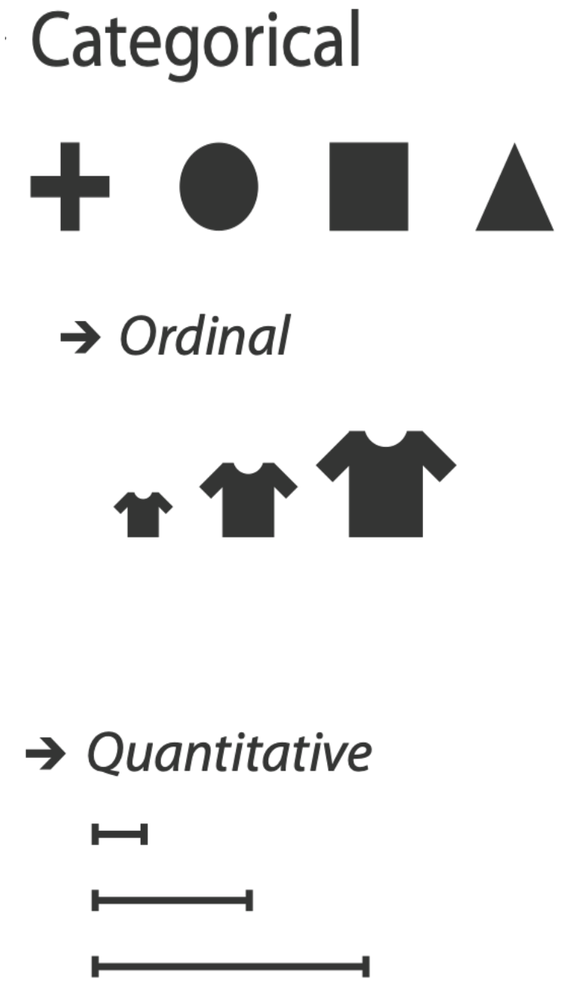
- Ordered categories
- Operations: =, ≠, <, >
- Example: rankings, ratings

Q – Interval

- Numeric, with arbitrary zero
- Operations: =, ≠, <, >, -
- Can measure distances or spans
- Example: temperature in Celsius

Q – Ratio

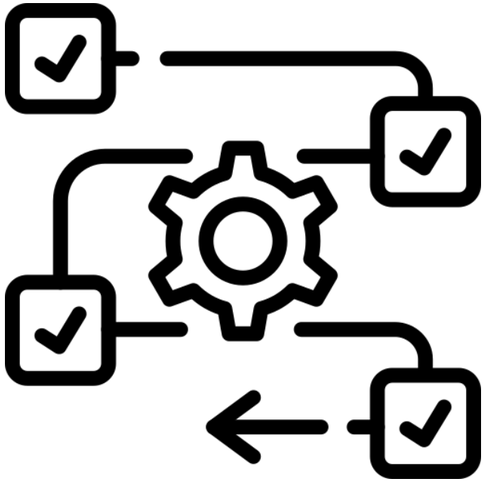
- Numeric, zero is fixed
- Operations: =, ≠, <, >, -, %
- Can measure ratios or proportions
- Example: weight, height, income



1. What is an item
2. What is an attribute
3. What is a feature
4. What is the semantics
5. What is the keys
6. What is attribute type for Order ID
7. What is the attribute type for Order Date
8. What is the attribute type for Order Priority
9. What is the attribute type for Product Container
10. What is the attribute type for Product Base margin
11. What is the attribute type for Ship Date

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	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 abstraction: Three operations

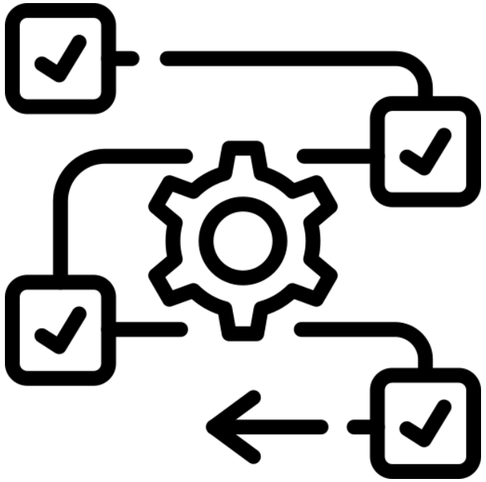


- translate from domain-specific language to generic visualization language

- identify dataset type(s), attribute types
- identify cardinality
  - how many items in the dataset?
  - what is cardinality of each attribute?
    - number of levels for categorical data
    - range for quantitative data

- consider whether to transform data
  - guided by understanding of task

Data Model vs. Conceptual Model



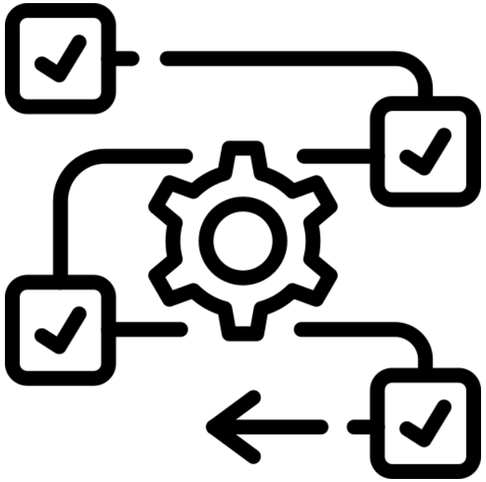
Data Models

- Formal descriptions (math-like)
- Example: integers with + and ×
- Focus: structure and operations

Conceptual Models

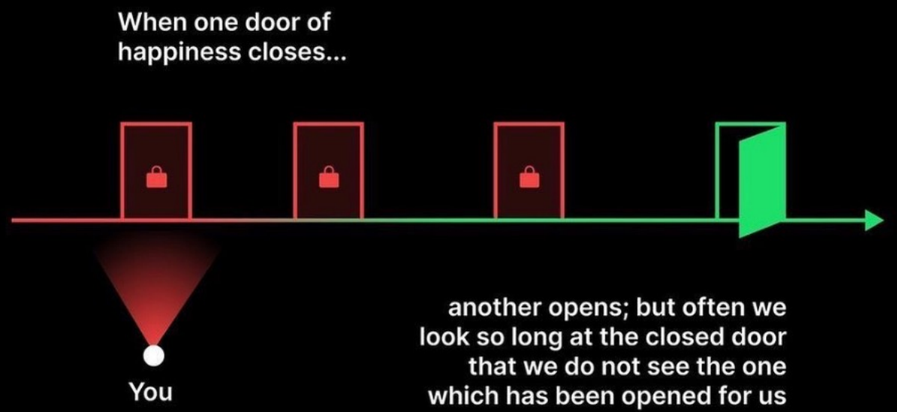
- Mental constructions we build around the data
- Include meaning (semantics)
- supports reasoning
- typically based on understanding of tasks

Data vs conceptual model, example

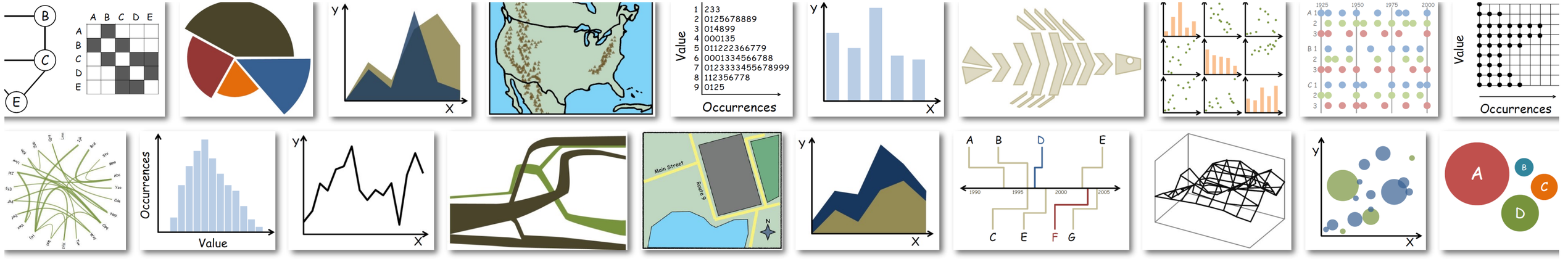


- data model: floats
  - 32.52, 54.06, -14.35, ...
- conceptual model
  - temperature
- multiple possible data abstractions
  - continuous to 2 significant figures: quantitative
    - task: forecasting the weather
  - hot, warm, cold: ordinal
    - task: deciding if bath water is ready
  - above freezing, below freezing: categorical
    - task: decide if I should leave the house today

visualhustles Following •••



Helen Keller • • • @golimitlesss

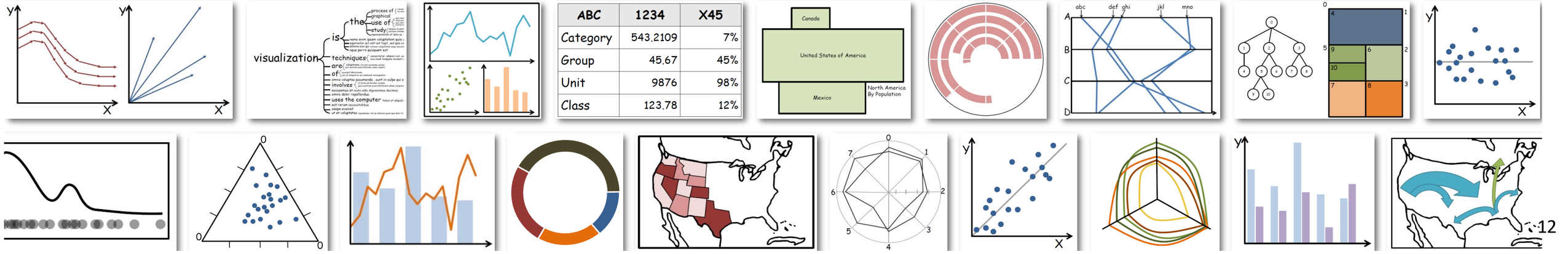


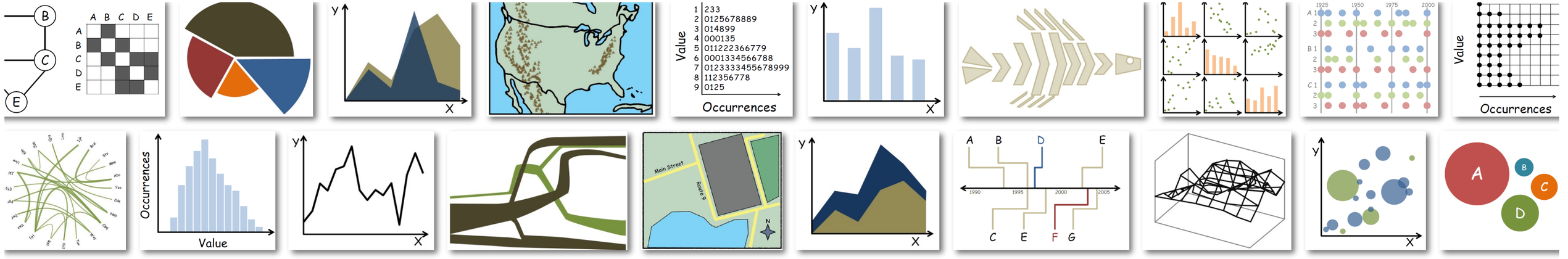
static or interactive

abstract or spatial

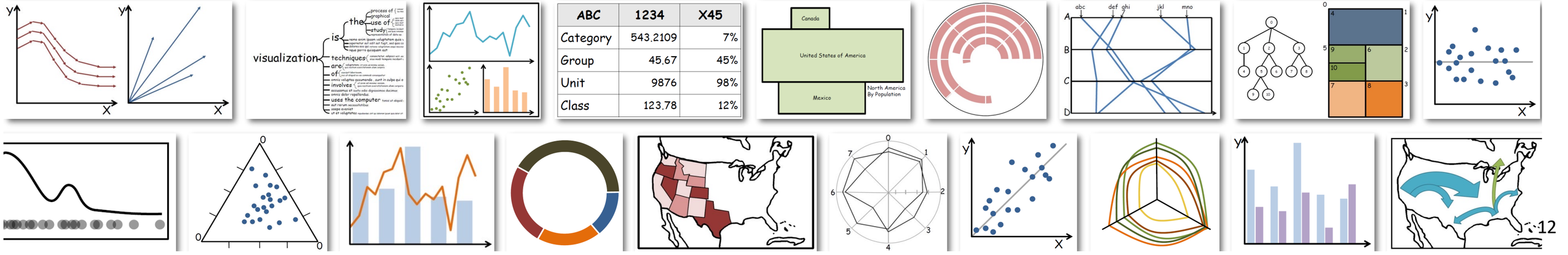
The visual representation of data to enhance cognitive tasks

problem solving, decision making, planning, sensemaking, storytelling





A visualization is a combination of **marks**, where data values are encoded using the marks' **visual channels** to make patterns and relationships perceptible.

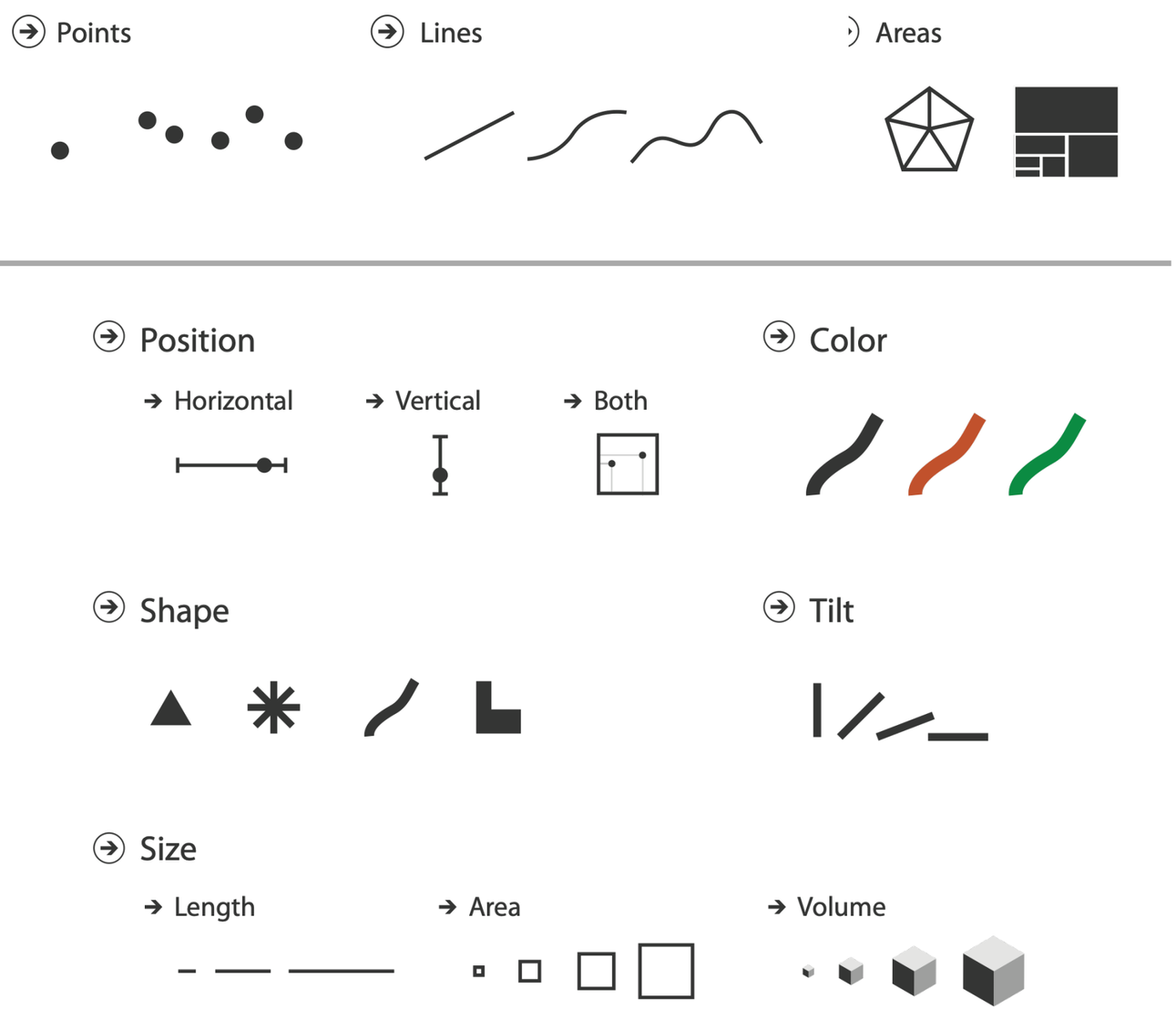


Definitions: Marks and channels

- Marks: what we see, the basic geometric primitives

- Channels: the way to control the appearance of marks, independent of the dimensionality of the geometric primitive

- Channel properties differ: type & amount of information that can be conveyed to human perceptual system



Semiology of Graphics – Alphabet of Viz

Jacques Bertin - French cartographer [1918-2010]

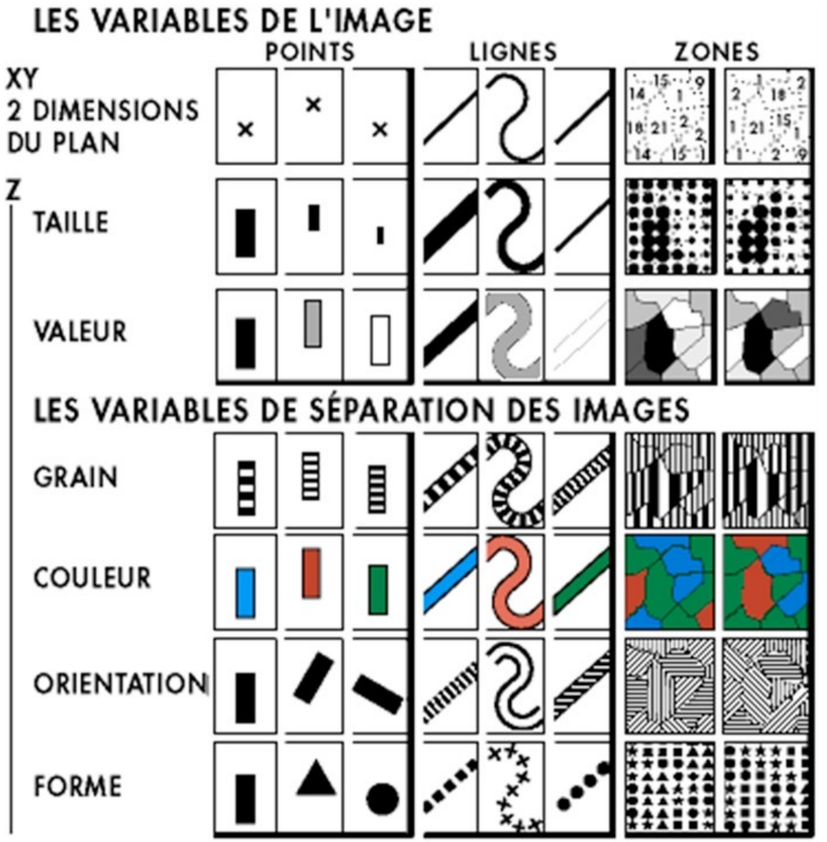
Semiology of Graphics [1967]

Theoretical principles for visual encodings

Marks: Points Lines Areas

Position Size
(Grey)Value

Texture
Color
Orientation
Shape



Semiology of Graphics [J. Bertin, 67]

Channels

Position on common scale

Position on unaligned scale

Length (1D size)

Tilt/angle

Area (2D size)

Depth (3D position)

Color luminance

Color saturation

Curvature

Volume (3D size)

Same

Same

Spatial region

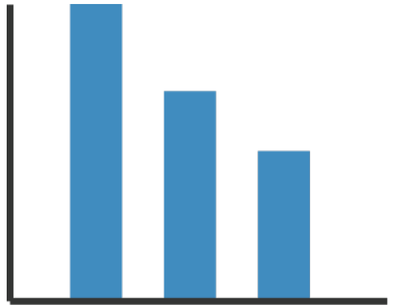
Color hue

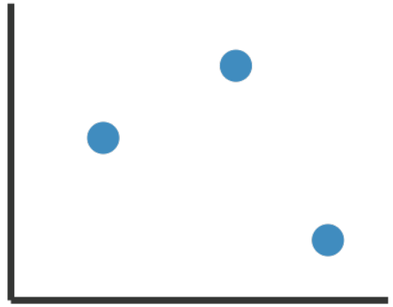
Motion

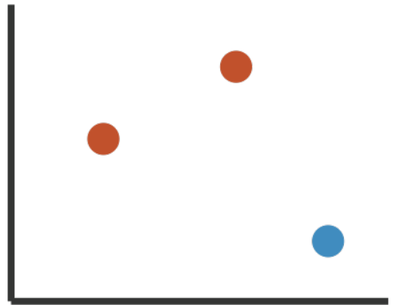
Shape

Visual encoding

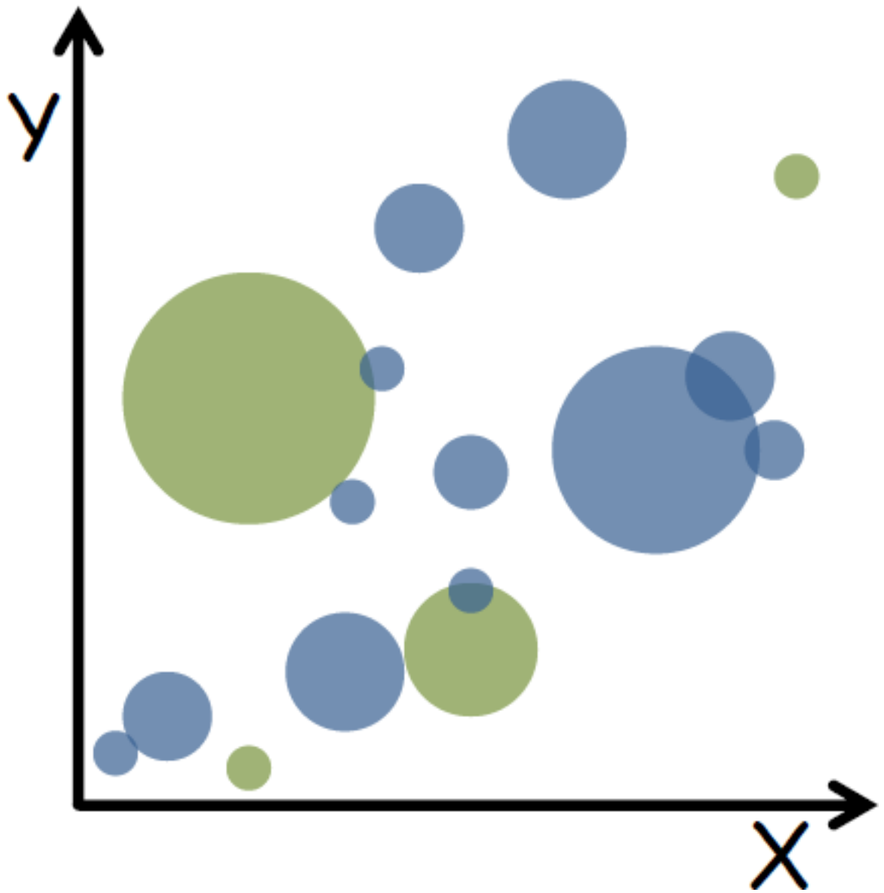
analyze idiom structure as combination of marks and channels







How many attributes are encoded?
What is the mark? What are the channels?



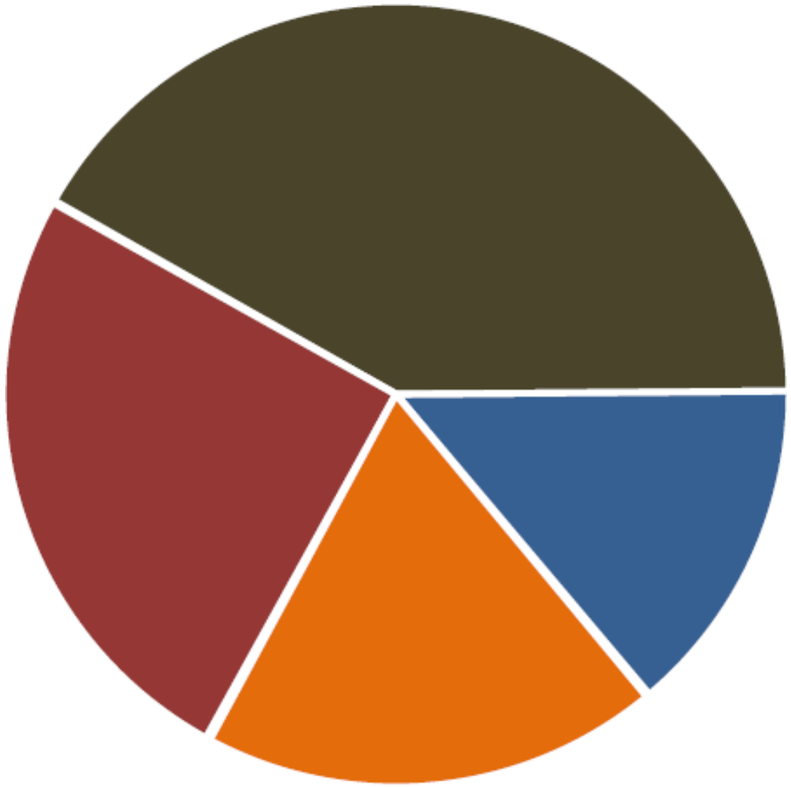
Mark:

- Points
- Lines
- Areas

Channel:

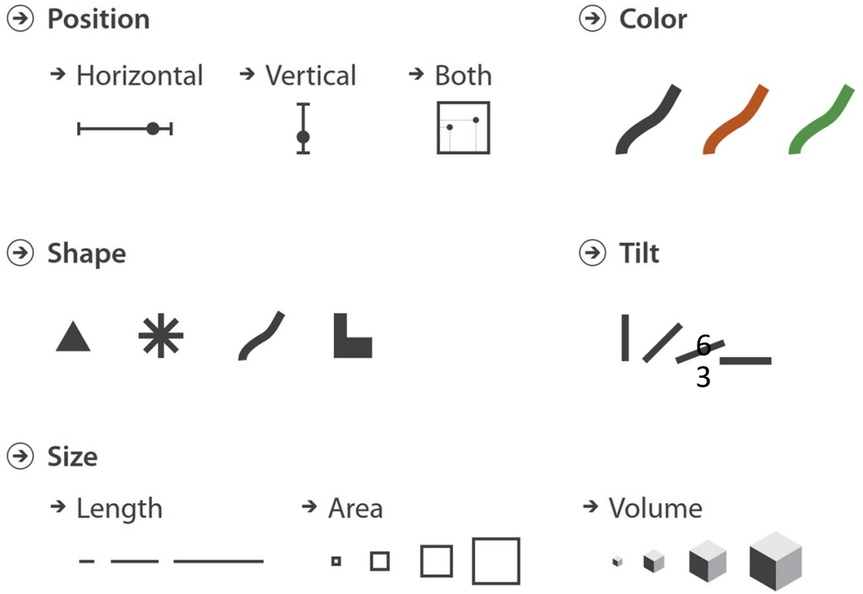
- Position
  - Horizontal
  - Vertical
  - Both
- Color
- Shape
- Tilt
- Size
  - Length
  - Area
  - Volume

How many attributes are encoded?
What is the mark? What are the channels?

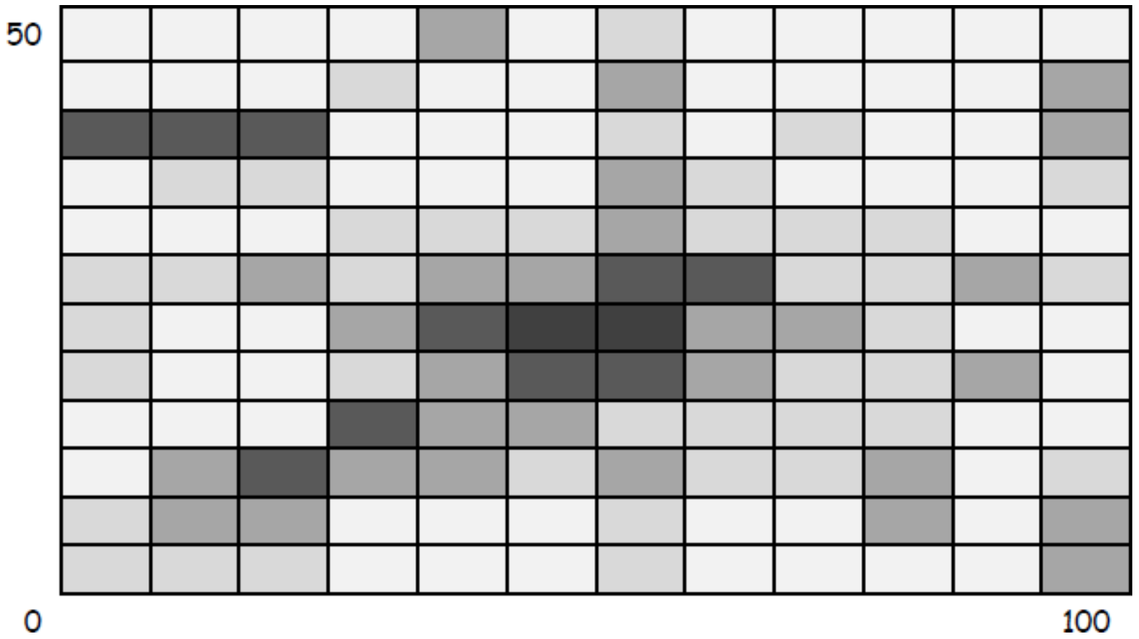


Mark:

Channel :

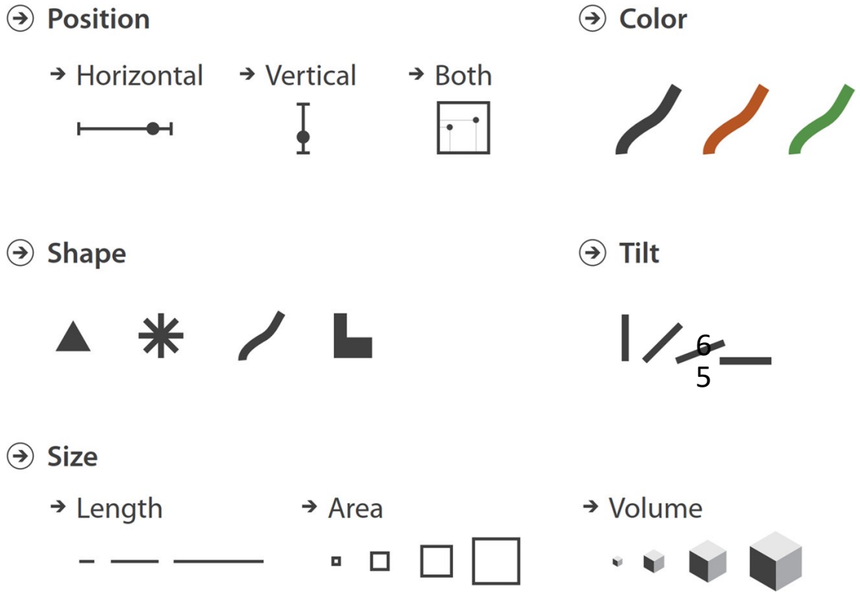


How many attributes are encoded?
What is the mark? What are the channels?



Mark:

Channel:



Clicker Question

Shooting Media Coverage

What is the mark?

A: points

B: lines

C: interlocking areas

What are the channels?

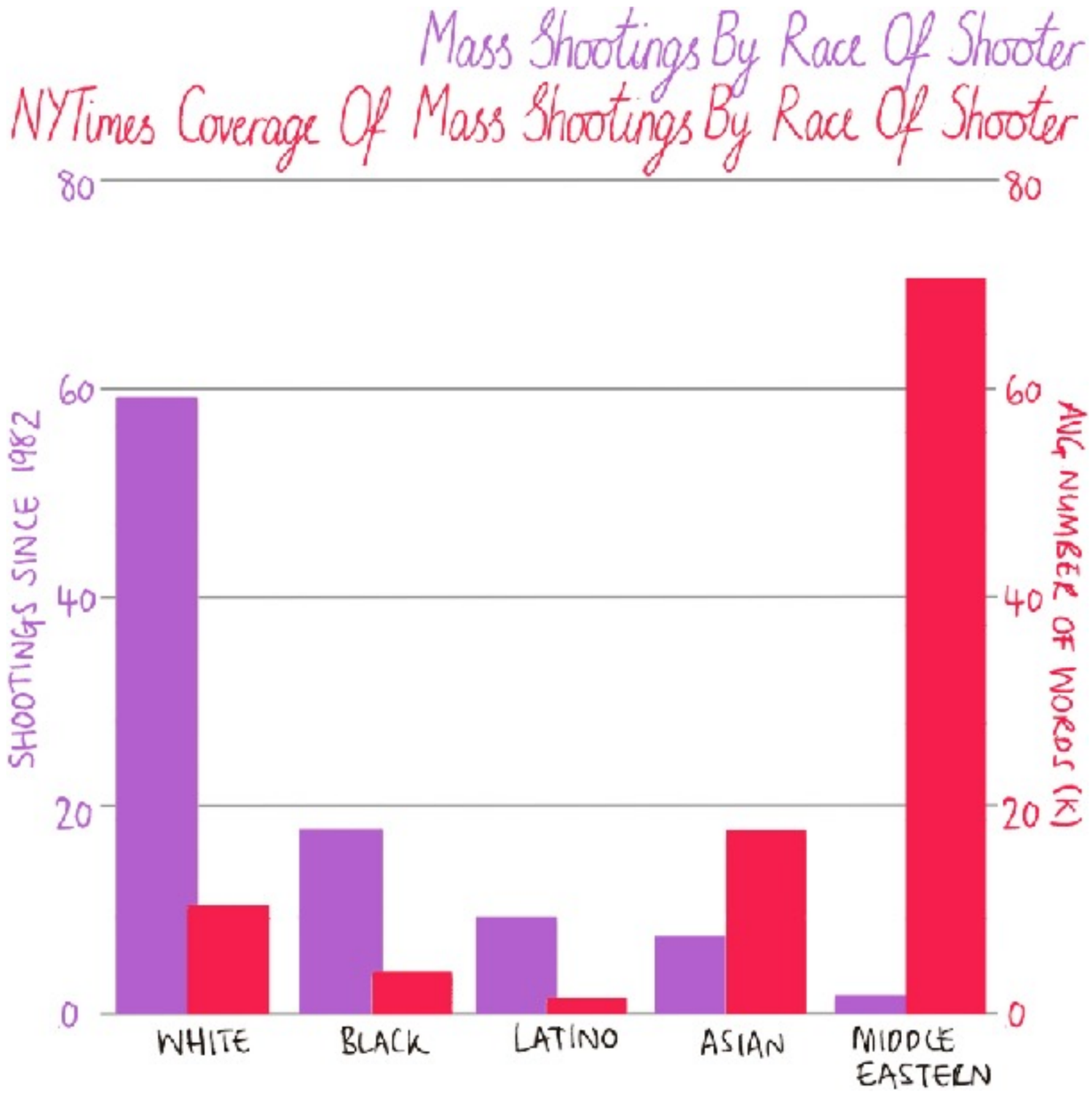
A: vertical position

B: color

C: horizontal position

D: area

E: angle



https://twitter.com/MonaChalabi/status/1158779046693679106?s=20

Clicker Question

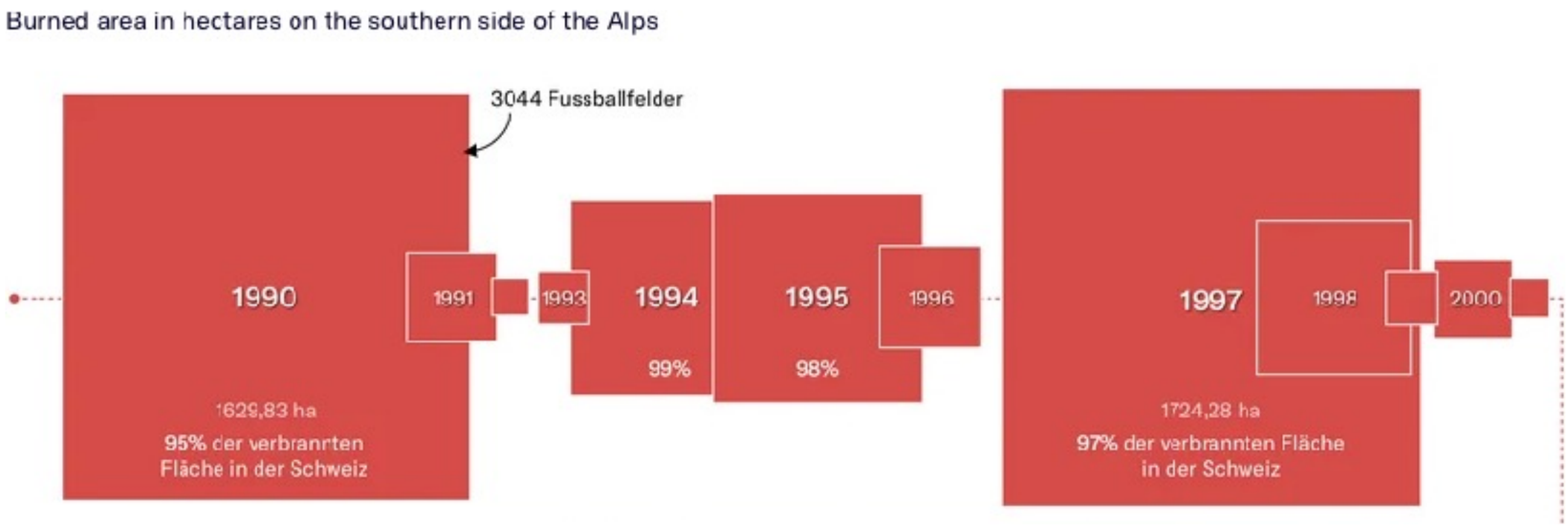
Alpen Forest Fires

What are the marks

- A: points
- B: lines
- C: interlocking areas

What are the channel(s)

- A: vertical position
- B: color
- C: horizontal position
- D: area
- E: angle



https://www.nzz.ch/wissenschaft/waldbraende-erklaert-in-der-schweiz-und-in-europa-ld.1483688

Clicker Question

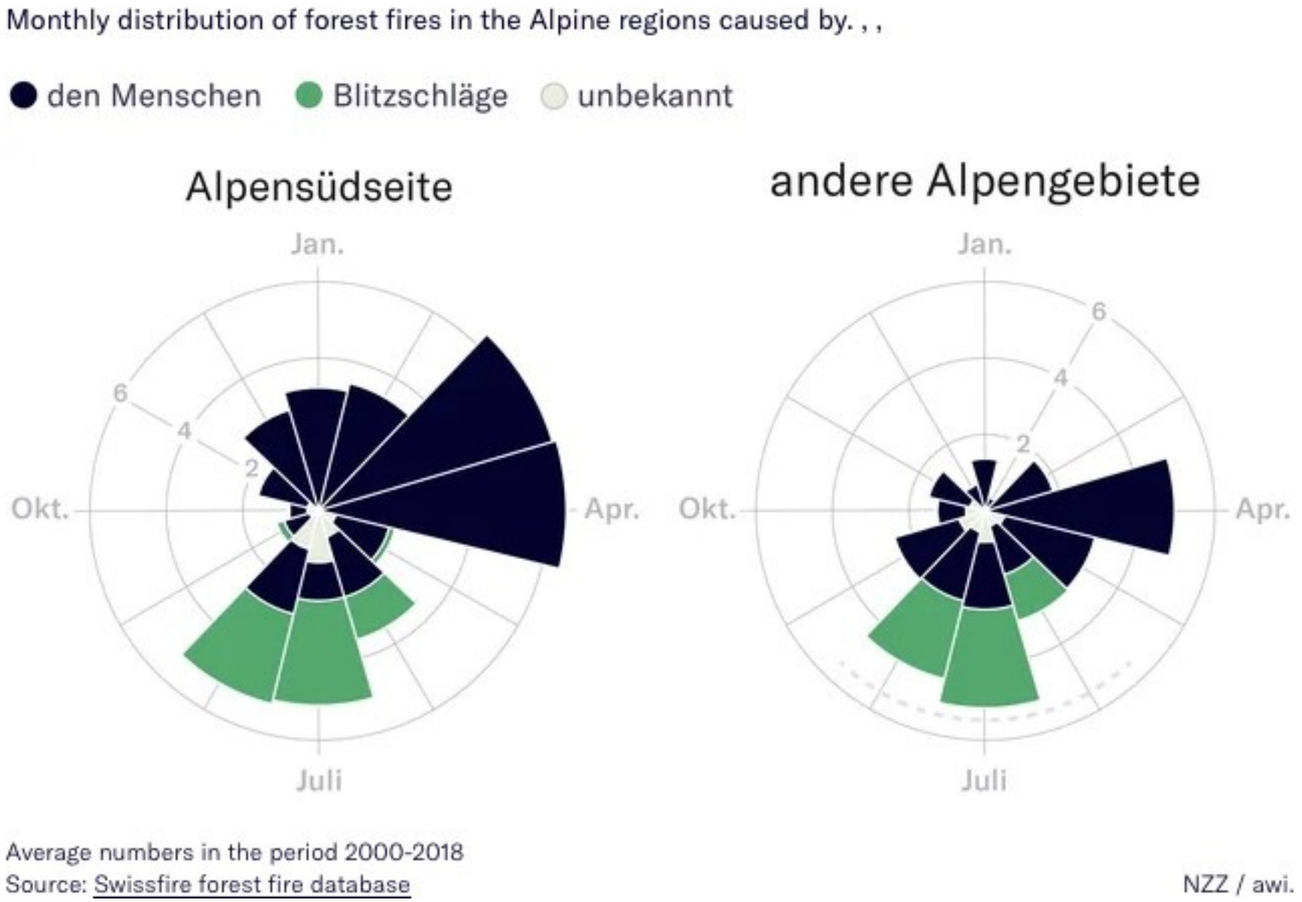
More Alpen Forest Fires

What are the mark(s)

- A: points
- B: lines
- C: interlocking areas

What are the channel(s)

- A: position
- B: color
- C: length
- D: area
- E: angle



https://www.nzz.ch/wissenschaft/waldbraende-erklaert-in-der-schweiz-und-in-europa-ld.1483688

Learning Outcomes

Describe the difference between how the phrases “dataset types”, “data types” are used in vis. Literature as opposed to programming

Describe the characteristics of data

Differentiate between the different types of data and dataset types

Describe the basic visual primitives of visualizations (marks and channels)

Differentiate between a mark and channel

Deconstruct a visualization based on its marks and channels

Get Stepping

- Complete the JumpStart and Tutorial 1 Modules (2 – 5 hrs)
- Book and take your quiz by Friday
- Go to lab on Tuesday or Thursday
- Complete Tutorial 2 before the lecture on Wednesday
- Eat ice cream at Rain or Shine (winter is coming)