Data and Image Models

Maneesh Agrawala
Jessica Hullman

CS 294-10: Visualization
Fall 2014

Last Time: The Purpose of Visualization

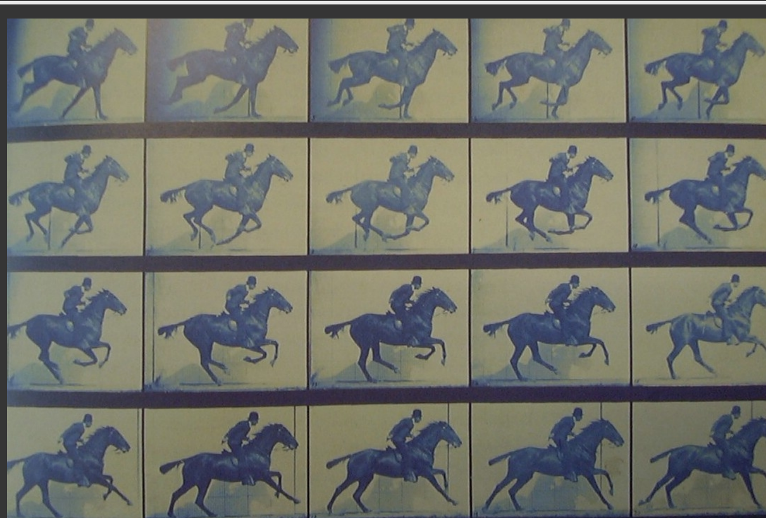
Three functions of visualizations

Record information
- Photographs, blueprints, …

Support reasoning about information (analyze)
- Process and calculate
- Reason about data
- Feedback and interaction

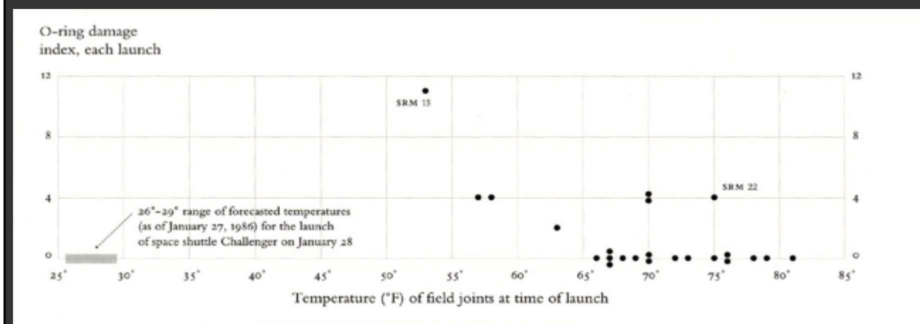
Convey information to others (present)
- Share and persuade
- Collaborate and revise
- Emphasize important aspects of data

Record information



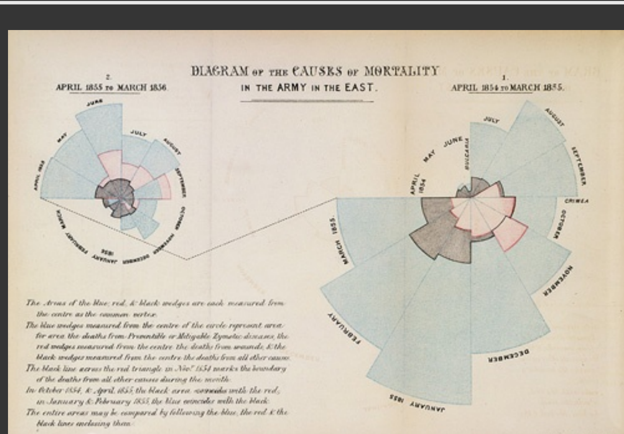
Gallop, Bay Horse "Daisy" [Muybridge 1884-86]

Analysis: Challenger



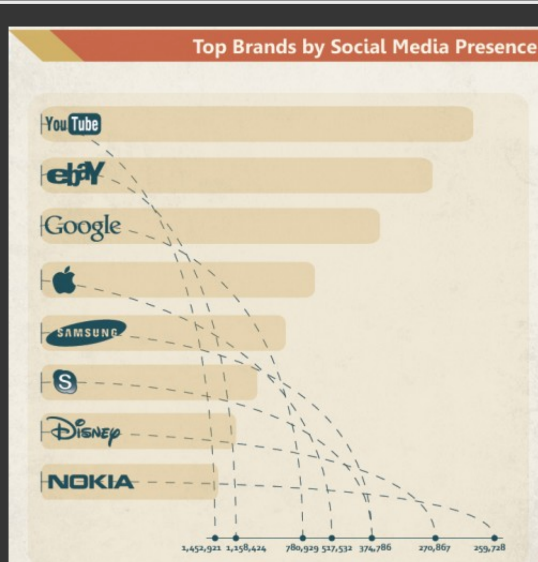
Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]

Communicate: War Deaths



Crimean War Deaths [Nightingale 1858]

Confuse: Top Brands



from wtfviz.net

Announcements

Auditors, *please* enroll in the class (1 unit, P/NP)

- Requirements: Come to class and participate (online as well)
- Requirements: Assignment 1

Class participation requirements

- Complete readings before class
- In-class discussion
- Post at least 1 discussion substantive comment/question by 11am on day of lecture

All, add yourself to participants page on the wiki

Class wiki
http://vis.berkeley.edu/courses/cs294-10-fa14/wiki/

Assignment 1: Visualization Design

1	Full_date_start	Year_start	Month_start	Day_start	Full_date_end	Year_end	Month_end	Day_end	Country	Location	Type	Sub_Type	Name	Killed	Cost_dollars	Affected	Id
45	31051970	1970	5	31	31051970	1970	5	31	Peru	Chimbote	Earthquake (seismic	Earthquake (ground shaking)		66794	3216240	530	1970-0032
46	101942	1942	10		101942	1942	10		Bangladesh	West Sundar	Storm	Tropical cyclone		61000			1942-0008
47	1910	1910			1910	1910			China P Rep	Manchuria	Epidemic	Bacterial Infectio	Pneumonic	60000			1910-0001
48	1935	1935			1935	1935			India		Storm	Tropical cyclone		60000			1935-0015
49	31051935	1935	5	31	31051935	1935	5	31	Pakistan	Quetta (Balu	Earthquake (seismic	Earthquake (ground shaking)		60000			1935-0005
50	71949	1949	7		71949	1949	7		China P Rep		Flood			57000			1949-0025
51	11918	1918	1		11918	1918	1		Canada	Natiuonwide	Epidemic	Viral Infectious D	Spanish Influ	50000	2000000		1918-0015
52	81912	1912	8		81912	1912	8		China P Rep	Wenchou	Storm	Tropical cyclone		50000			1912-0005
53	101949	1949	10		101949	1949	10		Guatemala	East	Flood			40000		15	1949-0012
54	14101942	1942	10	14	14101942	1942	10	14	India	Orissa, West	Storm	Tropical cyclone		40000			1942-0009
55	21061990	1990	6	21	21061990	1990	6	21	Iran Islam Rep	Rasht, Astara	Earthquake (seismic	Earthquake (ground shaking)		40000	710000	8000	1990-0034
56	11051965	1965	5	11	11051965	1965	5	11	Bangladesh	Barisal distric	Storm	Tropical cyclone		36000	15600000	57.7	1965-0028
57	26122004	2004	12	26	26122004	2004	12	26	Sri Lanka		Earthquake (seismic	Tsunami		35399	1019306	1316.5	2004-0659
58	26121939	1939	12	26	26121939	1939	12	26	Turkey	Erzincan (Ana	Earthquake (seismic	Earthquake (ground shaking)		32962		20	1939-0010
59	1946	1946			1946	1946			Cape Verde Is	Countrywide	Drought	Drought		30000			1946-9004
60	24011939	1939	1	24	24011939	1939	1	24	Chile	Chillan regio	Earthquake (seismic	Earthquake (ground shaking)		30000	58500	920	1939-0001
61	81954	1954	8		81954	1954	8		China P Rep	Hopeh, Wuh	Flood			30000			1954-0014
62	8051902	1902	5	80	8051902	1902	5	80	Martinique		Volcano	Volcanic eruptior	Mount Pelee	30000			1902-0006
63	15121999	1999	12	15	20121999	1999	12	20	Venezuela	Federal distr	Flood	Flash flood		30000	483635	3160	1999-0547
64	13011915	1915	1	13	13011915	1915	1	13	Italy	Avezzano	Earthquake (seismic	Earthquake (ground shaking)		29980		60	1915-0002
65	71974	1974	7		71974	1974	7		Bangladesh		Flood			28700	38000000	579.2	1974-0034
66	26122003	2003	12	26	26122003	2003	12	26	Iran Islam Rep	Bam (Kerran	Earthquake (seismic	Earthquake (ground shaking)		26796	267628	500	2003-0630
67	16091978	1978	9	16	16091978	1978	9	16	Iran Islam Rep	Tabas (Khora	Earthquake (seismic	Earthquake (ground shaking)		25000	40000	50	1978-0115
68	7121988	1988	12	71	7121988	1988	12	71	Soviet Union	Armenia, Ler	Earthquake (seismic	Earthquake (ground shaking)		25000	1642000	14000	1988-0528
69	1920	1920			1920	1920			Cape Verde Is	Countrywide	Drought	Drought		24000			1920-9004

Worldwide Disasters 1900-2008

Due by 11:59pm on Sep 9

Data and Image Models

The big picture

task

data
- physical type
  - int, float, etc.
- abstract type
  - nominal, ordinal, etc.

domain
- metadata
- semantics
- conceptual model

processing algorithms

→

mapping
- visual encoding
- visual metaphor

image
- visual channel
- retinal variables

Topics

Properties of data or information

Properties of the image

Mapping data to images

Data

Data models vs. Conceptual models

Data models: low level descriptions of the data

- Math: Sets with operations on them
- Example: integers with + and × operators

Conceptual models: mental constructions

- Include semantics and support reasoning

Examples (data vs. conceptual)

- (1D floats) vs. Temperature
- (3D vector of floats) vs. Space

Taxonomy

- 1D (sets and sequences)
- Temporal
- 2D (maps)
- 3D (shapes)
- nD (relational)
- Trees (hierarchies)
- Networks (graphs)

Are there others?

The eyes have it: A task by data type taxonomy for information visualization [Schneiderman 96]

Types of variables

Physical types

- Characterized by storage format
- Characterized by machine operations

Example:

bool, short, int32, float, double, string, …

Abstract types

- Provide descriptions of the data
- May be characterized by methods/attributes
- May be organized into a hierarchy

Example:

plants, animals, metazoans, …

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

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

Nominal, ordinal and quantitative

N - Nominal (labels)

- Operations: =, ≠

O - Ordered

- Operations: =, ≠, <, >, ≤, ≥

Q - Interval (Location of zero arbitrary)

- Operations: =, ≠, <, >, ≤, ≥, -
- Can measure distances or spans

Q - Ratio (zero fixed)

- Operations: =, ≠, <, >, ≤, ≥, -, ÷
- Can measure ratios or proportions

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

From data model to N,O,Q data type

Data model

- 32.5, 54.0, -17.3, …
- floats

Conceptual model

- Temperature

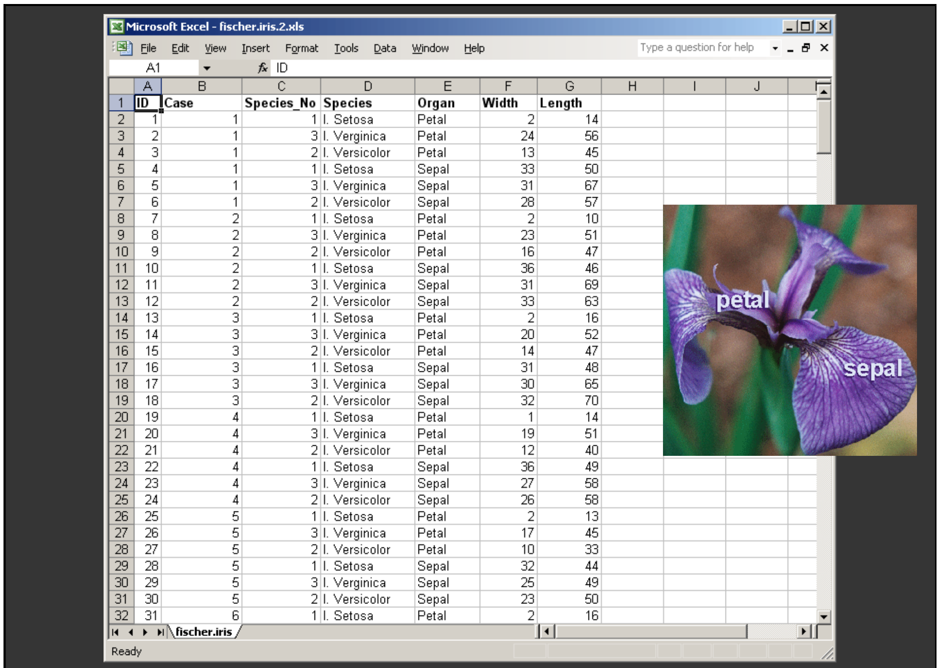
Data type

- Burned vs. Not burned (N)
- Hot, warm, cold (O)
- Continuous range of values (Q)

Iris Setosa

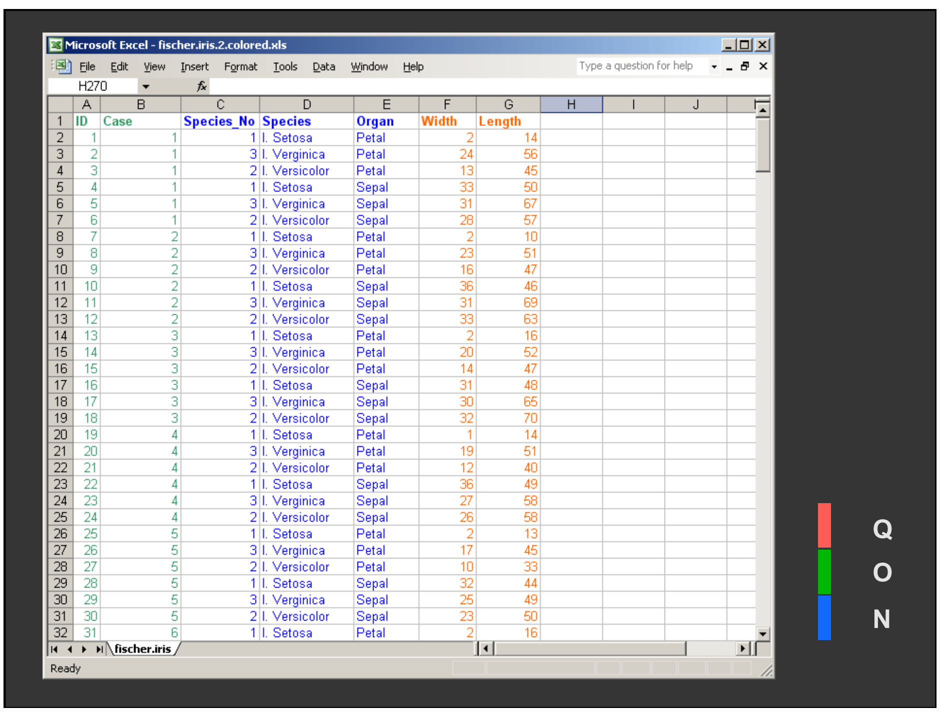
Iris Versicolor

Iris Virginica



ID	Case	Species_No	Species	Organ	Width	Length
1	1	1	I. Setosa	Petal	2	14
2	1	3	I. Verginica	Petal	24	56
3	1	2	I. Versicolor	Petal	13	45
4	1	1	I. Setosa	Sepal	33	50
5	1	3	I. Verginica	Sepal	31	67
6	1	2	I. Versicolor	Sepal	28	57
7	2	1	I. Setosa	Petal	2	10
8	2	3	I. Verginica	Petal	23	51
9	2	2	I. Versicolor	Petal	16	47
10	2	1	I. Setosa	Sepal	36	46
11	2	3	I. Verginica	Sepal	31	69
12	2	2	I. Versicolor	Sepal	33	63
13	3	1	I. Setosa	Petal	2	16
14	3	3	I. Verginica	Petal	20	52
15	3	2	I. Versicolor	Petal	14	47
16	3	1	I. Setosa	Sepal	31	48
17	3	3	I. Verginica	Sepal	30	65
18	3	2	I. Versicolor	Sepal	32	70
19	4	1	I. Setosa	Petal	1	14
20	4	3	I. Verginica	Petal	19	51
21	4	2	I. Versicolor	Petal	12	40
22	4	1	I. Setosa	Sepal	36	49
23	4	3	I. Verginica	Sepal	27	58
24	4	2	I. Versicolor	Sepal	26	58
25	5	1	I. Setosa	Petal	2	13
26	5	3	I. Verginica	Petal	17	45
27	5	2	I. Versicolor	Petal	10	33
28	5	1	I. Setosa	Sepal	32	44
29	5	3	I. Verginica	Sepal	25	49
30	5	2	I. Versicolor	Sepal	23	50
31	6	1	I. Setosa	Petal	2	16

Sepal and petal lengths and widths for three species of iris [Fisher 1936].



ID	Case	Species_No	Species	Organ	Width	Length
1	1	1	I. Setosa	Petal	2	14
2	1	3	I. Verginica	Petal	24	56
3	1	2	I. Versicolor	Petal	13	45
4	1	1	I. Setosa	Sepal	33	50
5	1	3	I. Verginica	Sepal	31	67
6	1	2	I. Versicolor	Sepal	28	57
7	2	1	I. Setosa	Petal	2	10
8	2	3	I. Verginica	Petal	23	51
9	2	2	I. Versicolor	Petal	16	47
10	2	1	I. Setosa	Sepal	36	46
11	2	3	I. Verginica	Sepal	31	69
12	2	2	I. Versicolor	Sepal	33	63
13	3	1	I. Setosa	Petal	2	16
14	3	3	I. Verginica	Petal	20	52
15	3	2	I. Versicolor	Petal	14	47
16	3	1	I. Setosa	Sepal	31	48
17	3	3	I. Verginica	Sepal	30	65
18	3	2	I. Versicolor	Sepal	32	70
19	4	1	I. Setosa	Petal	1	14
20	4	3	I. Verginica	Petal	19	51
21	4	2	I. Versicolor	Petal	12	40
22	4	1	I. Setosa	Sepal	36	49
23	4	3	I. Verginica	Sepal	27	58
24	4	2	I. Versicolor	Sepal	26	58
25	5	1	I. Setosa	Petal	2	13
26	5	3	I. Verginica	Petal	17	45
27	5	2	I. Versicolor	Petal	10	33
28	5	1	I. Setosa	Sepal	32	44
29	5	3	I. Verginica	Sepal	25	49
30	5	2	I. Versicolor	Sepal	23	50
31	6	1	I. Setosa	Petal	2	16

Relational data model

Represent data as a **table** (*relation*)

Each **row** (*tuple*) represents a single record

Each record is a fixed-length tuple

Each **column** (*attribute*) represents a single *variable*

Each attribute has a *name* and a *data type*

A table's **schema** is the set of names and data types

A **database** is a collection of tables (relations)

Relational algebra [Codd]

Data transformations (SQL)

- **Selection (WHERE) – restrict values**
- **Projection (SELECT) – choose subset of attributes**
- **Sorting (ORDER BY)**
- **Aggregation (GROUP BY, SUM, MIN, …)**
- **Set operations (UNION, …)**
- **Combine (INNNER JOIN, OUTER JOIN, …)**

Statistical data model

Variables or measurements

Categories or factors or dimensions

Observations or cases

Statistical data model

Variables or measurements

Categories or factors or dimensions

Observations or cases

Month	Control	Placebo	300 mg	450 mg
March	165	163	166	168
April	162	159	161	163
May	164	158	161	153
June	162	161	158	160
July	166	158	160	148
August	163	158	157	150

Blood Pressure Study (4 treatments, 6 months)

Dimensions and measures

Dimensions: Discrete variables describing data

Dates, categories of values (independent vars)

Measures: Data values that can be aggregated

Numbers to be analyzed (dependent vars)

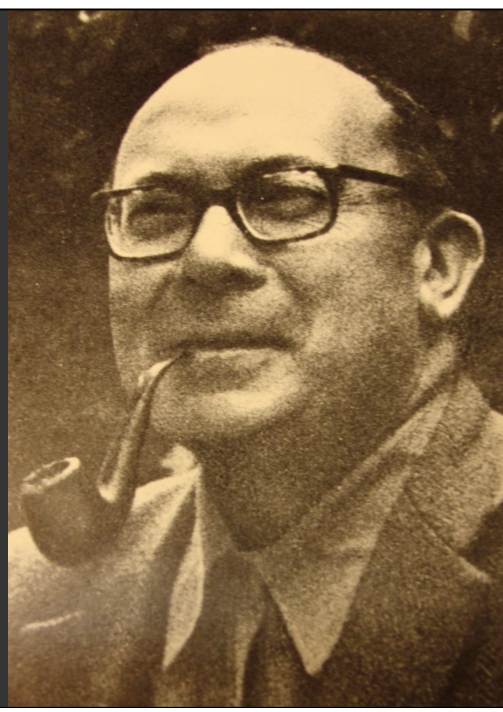
Aggregate as sum, count, average, std. deviation

Dimensions and measures

Independent vs. dependent variables

- **Example: $y = f(x,a)$**
- **Dimensions: Domain(x) × Domain(a)**
- **Measures: Range(y)**

Image



Visual language is a sign system

Jacques Bertin

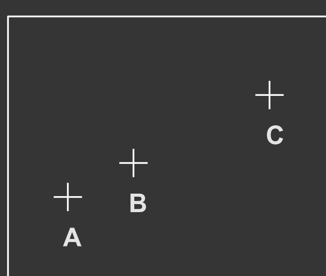
Images perceived as a set of signs

Sender encodes information in signs

Receiver decodes information from signs

Semiology of Graphics, 1967

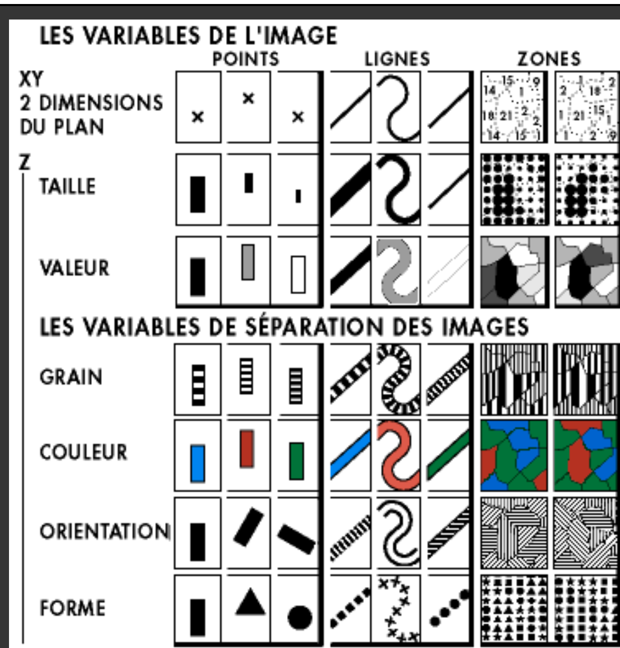
Information in position



1. A, B, C are distinguishable
2. B is between A and C.
3. BC is twice as long as AB.

∴ Encode quantitative variables

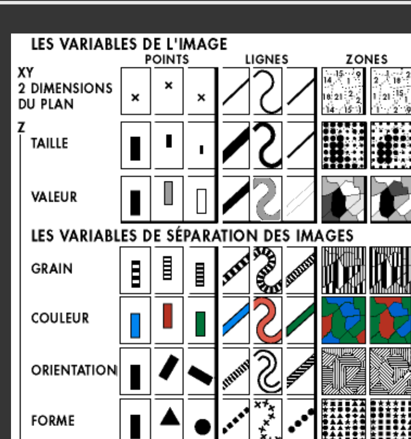
"Resemblance, order and proportional are the three signfields in graphics." - Bertin



[Bertin, Semiology of Graphics, 1983]

Visual variables

- Position (x 2)
- Size
- Value
- Texture
- Color
- Orientation
- Shape



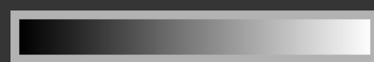
Note: Bertin does not consider 3D or time
Note: Card and Mackinlay extend the number of vars.

Information in color and value

Value is perceived as ordered

∴ Encode ordinal variables (O)

∴ Encode continuous variables (Q) [not as well]



Hue is normally perceived as unordered

∴ Encode nominal variables (N) using color

Bertins' "Levels of Organization"

Position	N	O	Q
Size	N	O	Q
Value	N	O	Q
Texture	N	O	
Color	N		
Orientation	N		
Shape	N		

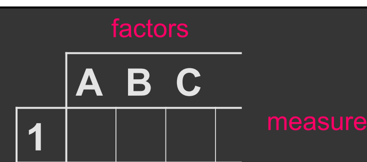
N Nominal
O Ordered
Q Quantitative

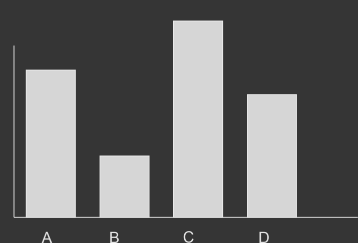
Note: Q < O < N

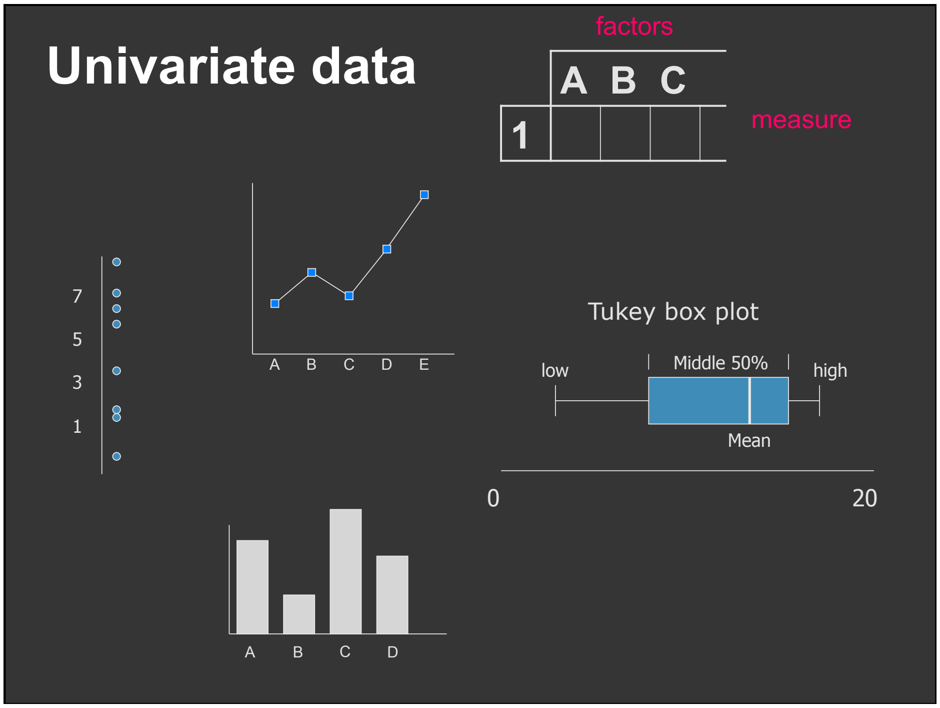
Note: Bertin actually breaks visual variables down into differentiating (≠) and associating (≡)

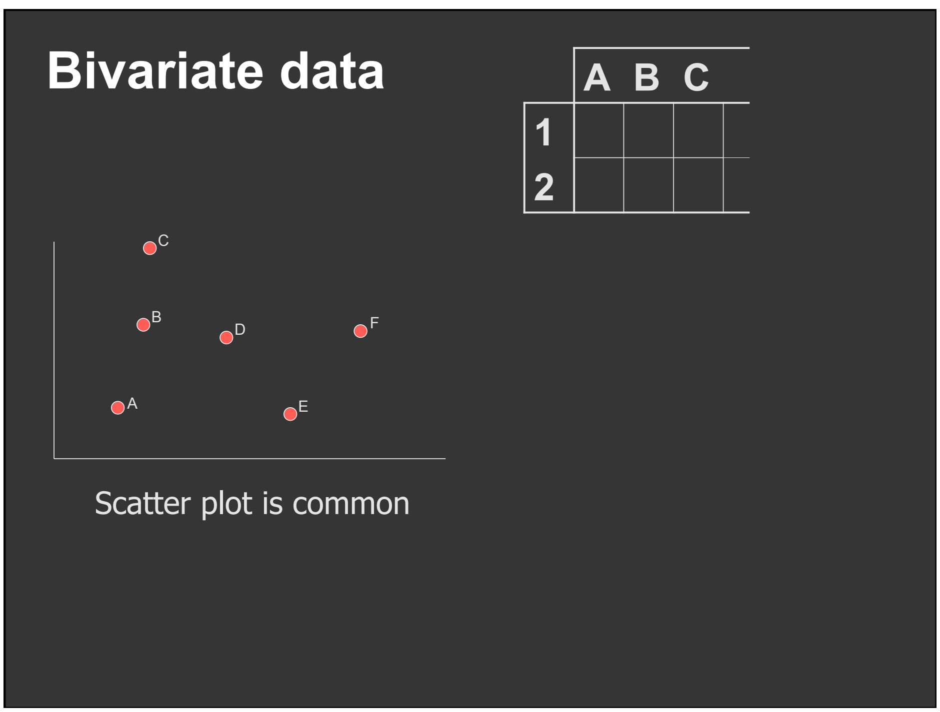
Encoding rules

Univariate data



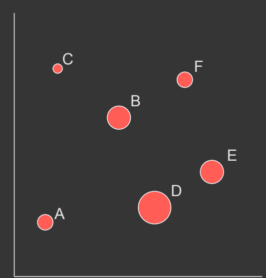



Univariate data
factors
A B C
1
measure
7
5
3
1
A B C D E
Tukey box plot
low
Middle 50%
high
Mean
0
20
A B C D

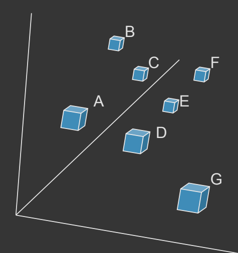

Bivariate data
A B C
1
2
C
B
D
F
A
E
Scatter plot is common

Trivariate data

	A	B	C
1			
2			
3			



3D scatter plot is possible



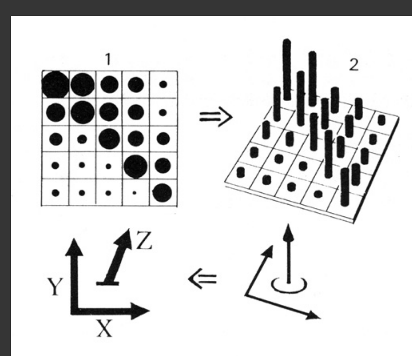
Three variables

Two variables [x,y] can map to points

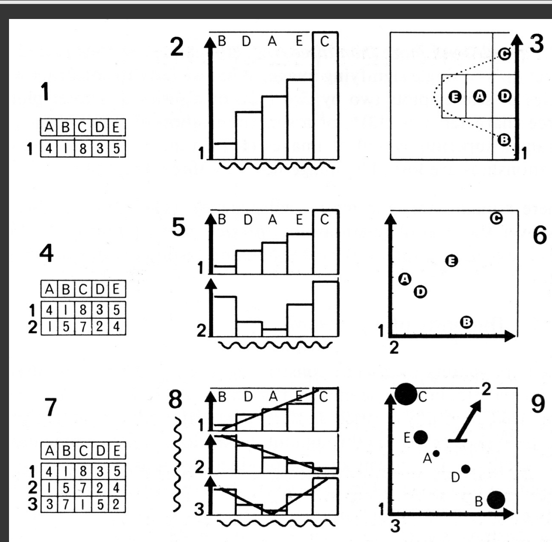
- Scatterplots, maps, …

Third variable [z] must use …

- Color, size, shape, …



Large design space (visual metaphors)



[Bertin, Graphics and Graphic Info. Processing, 1981]

Multidimensional data

How many variables can be depicted in an image?

	A	B	C	
1				
2				
3				
4				
5				
6				
7				
8				

Multidimensional data

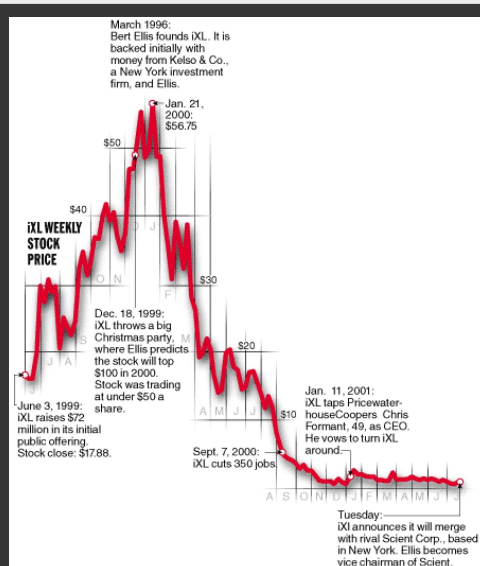
How many variables can be depicted in an image?

	A	B	C	
1				
2				
3				
4				
5				
6				
7				
8				

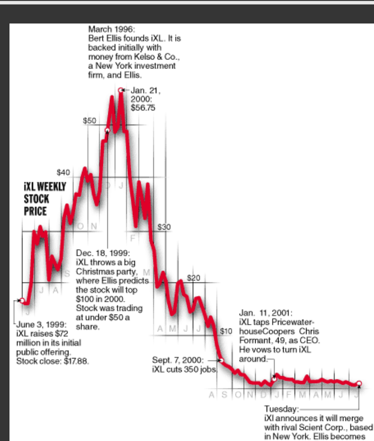
"With up to three rows, a data table can be constructed directly as a single image … However, an image has only three dimensions. And this barrier is impassible." Bertin

Deconstructions

Stock chart from the late 90s

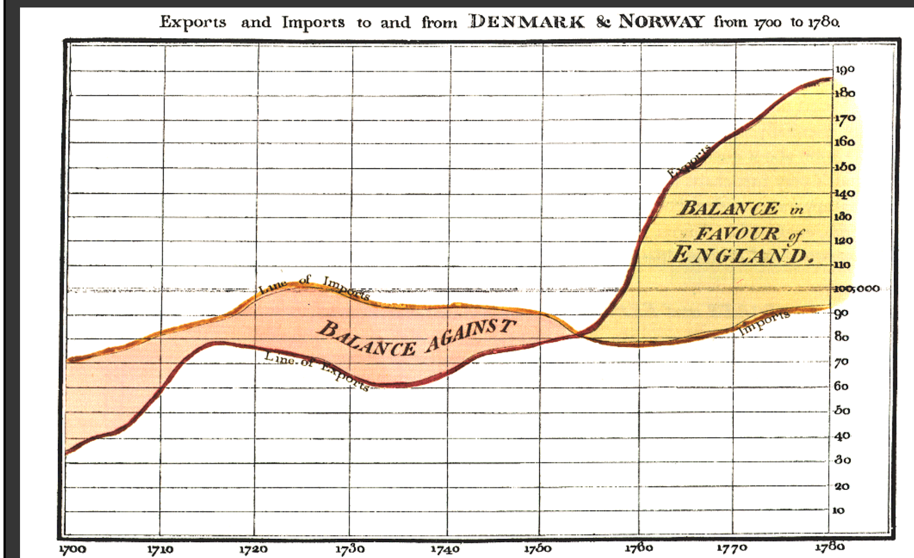


Stock chart from the late 90s

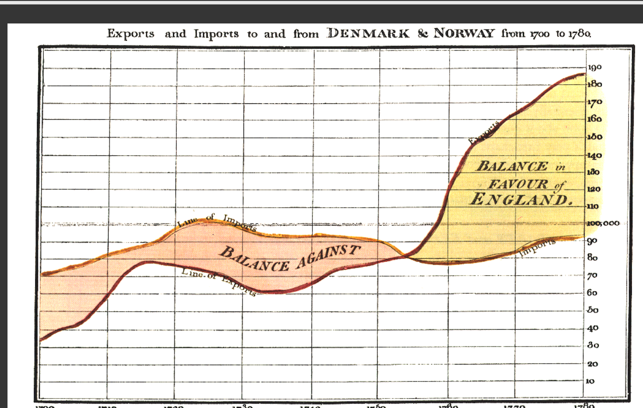


- Time → x-position (Q, linear)
- Price → y-position (Q, linear)

Playfair 1786

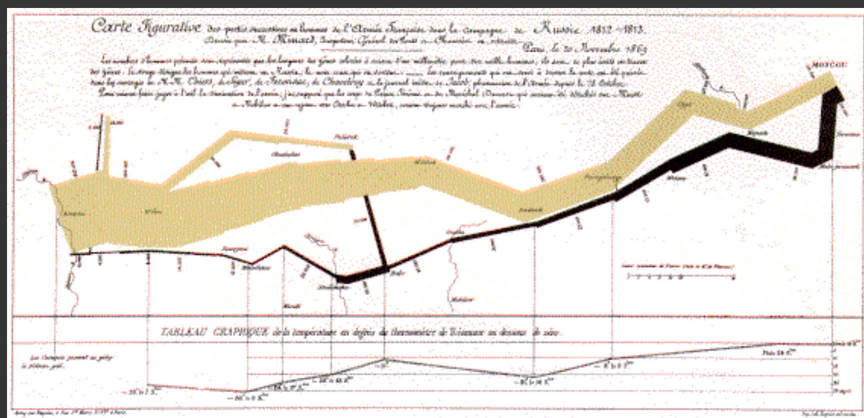


Playfair 1786

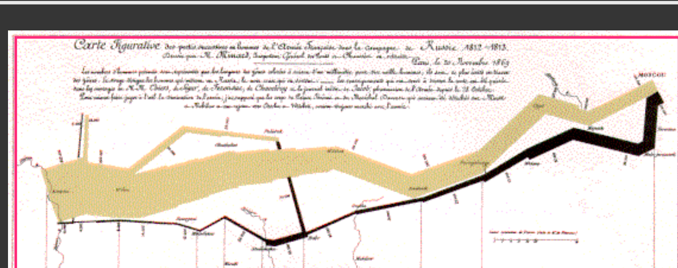


- Time → x-position (Q, linear)
- Exports/Imports Values → y-position (Q, linear)
- Exports/Imports → color (N, O, nominal)
- Balance for/against → area (maybe length??) (Q, linear)
- Balance for/against → color (N, O, nominal)

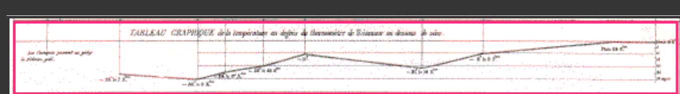
Minard 1869: Napoleon’s march



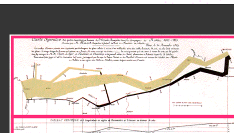
Single axis composition



+



=



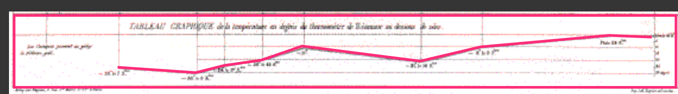
[based on slide from Mackinlay]

Mark composition

Temperature → y-position (Q, linear)

\+ longitude → x-position (Q, linear)

=



temp over longitude (Q x Q)

[based on slide from Mackinlay]

Mark composition

Longitude → y-position (Q, linear)

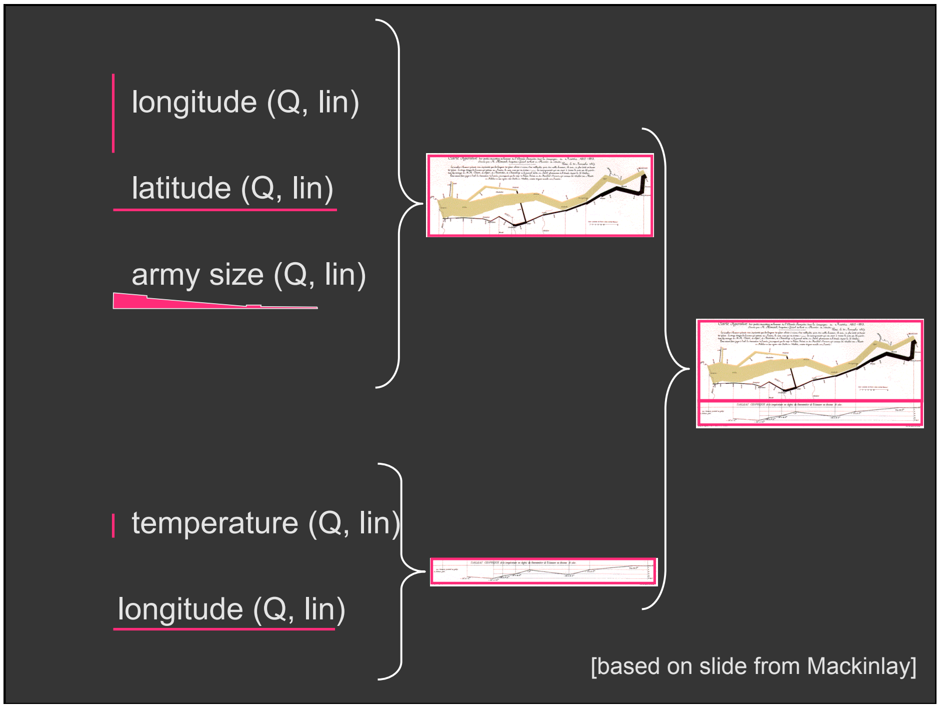
\+ Latitude → x-position (Q, linear)

\+ army size → width (Q, linear)

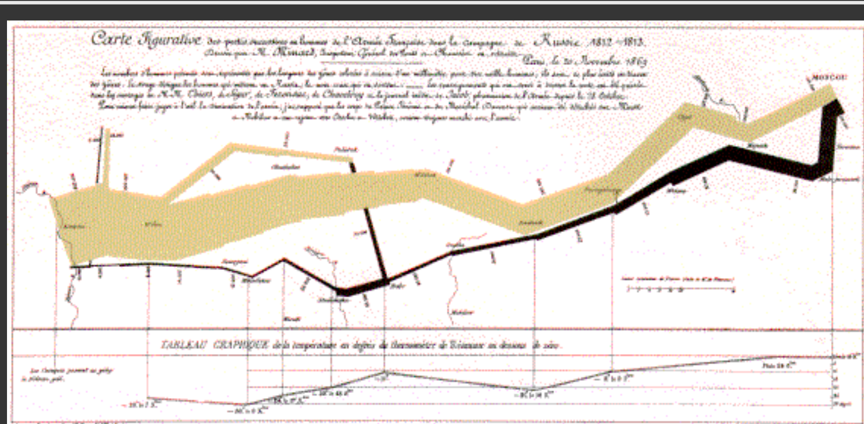
=

army position (Q x Q) and army size (Q)

[based on slide from Mackinlay]



Minard 1869: Napoleon's march

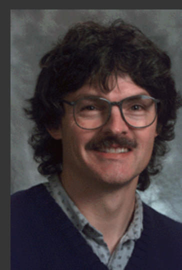


Depicts at least 5 quantitative variables

Any others?

Automated design

Jock Mackinlay's APT 86



Combinatorics of encodings

Challenge:

Assume 8 visual encodings and n data attributes

Pick the best encoding from the exponential number of possibilities $(n+1)^8$

Principle of Consistency:

The properties of the image (visual variables) should match the properties of the data

Principle of Importance Ordering:

Encode the most important information in the most effective way

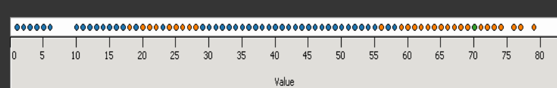
Mackinlay's expressiveness criteria

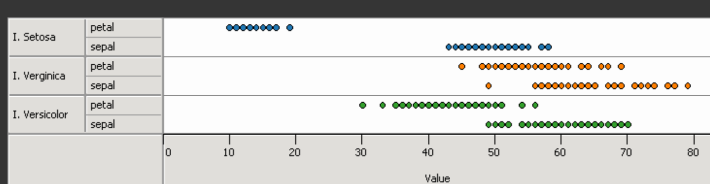
Expressiveness

A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express *all* the facts in the set of data, and *only* the facts in the data.

Cannot express the facts

A one-to-many (1 → N) relation cannot be expressed in a single horizontal dot plot because multiple tuples are mapped to the same position





Expresses facts not in the data

A length is interpreted as a quantitative value;
∴ Length of bar says something untrue about N data

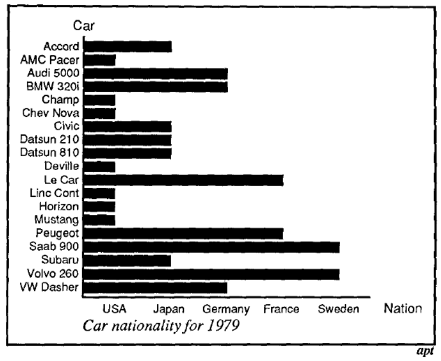


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

[Mackinlay, APT, 1986]

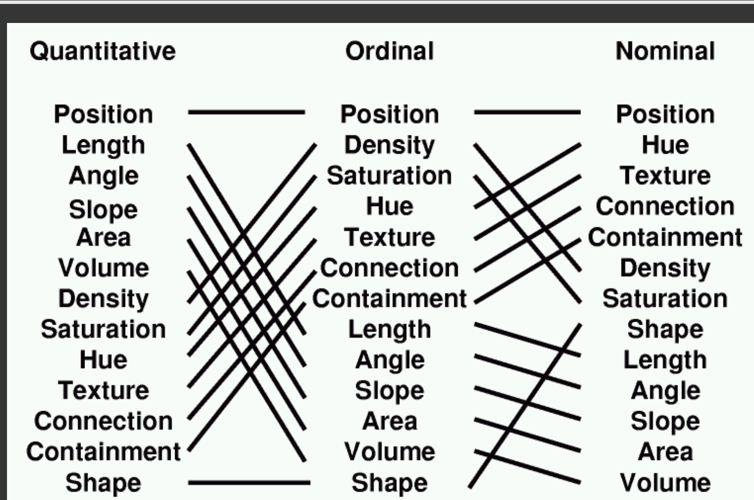
Mackinlay's effectiveness criteria

Effectiveness

A visualization is more effective than another visualization if the information conveyed by one visualization is more readily *perceived* than the information in the other visualization.

Subject of perception lecture

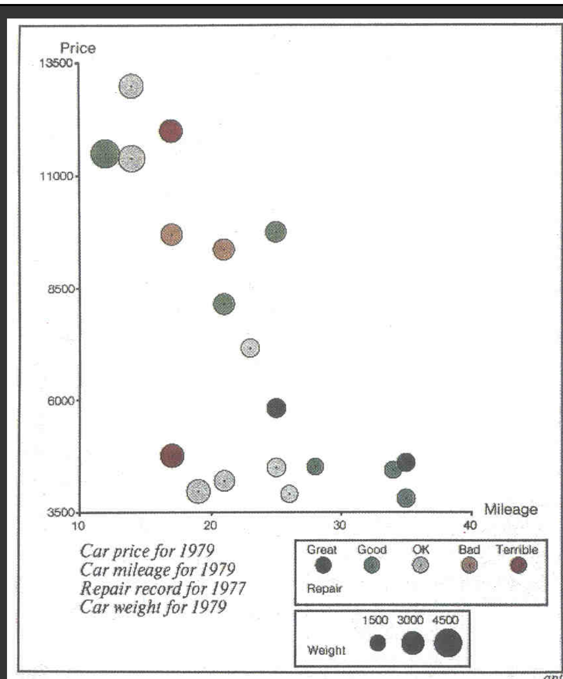
Mackinlay’s ranking



Conjectured *effectiveness* of the encoding

Mackinlay’s design algorithm

- User formally specifies data model and type
- APT searches over design space
  - Tests expressiveness of each visual encoding
  - Generates image for encodings that pass test
  - Tests perceptual effectiveness of resulting image
- Outputs most effective visualization



[Mackinlay, APT, 1986]

Limitations

Does not cover many visualization techniques

- Bertin and others discuss networks, maps, diagrams
- They do not consider 3D, animation, illustration, photography, …

Does not model interaction

Summary

Formal specification

- Data model
- Image model
- Encodings mapping data to image

Choose expressive and effective encodings

- Formal test of expressiveness
- Experimental tests of perceptual effectiveness