Handouts and Cheetsheet

First day deliverable

Story-driven:

- Write the draft story
- Leave blanks for numbers, data tables, and charts
- Survey related data set

Data-driven:

- Forget about "news" for now
- Determine your main dataset -- the more dimension, the better potential
- Observe the dimensions and ask questions
- Survey related data set

Workflow and Strategy

Data Journalism

- A crowd-sourced book from practitioners and edited by experienced insiders
- Creative Commons Attribute (CC3.0)
- Concrete Topics:
  - Perspectives & opinions of DJ
  - Real newsroom workflows of DJ
  - Case studies of modern works regarded as DJ
  - Scrape, FOI, open data,
  - Data cleaning
  - Visual exploration
  - NewsApp (recent buzzing counterpart is NewsGame)
  - Google Fusion table, Tableau Public, …
- Good introduction, broad coverage, practical insights. Less "handbook" like.
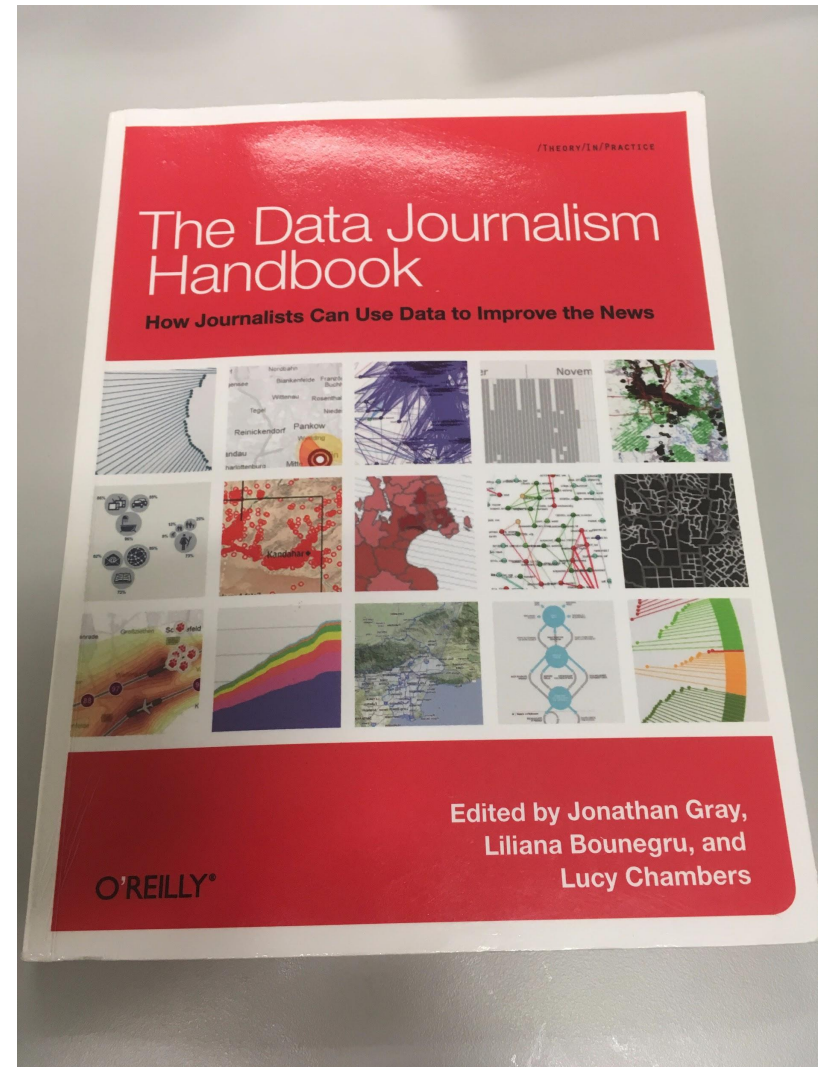
Workflow

- Ideation
- Data collection
- Data analysis
  - Exploratory analysis
  - Hypothesis testing
- Story telling
- Data visualisation
  - Turn story into visuals
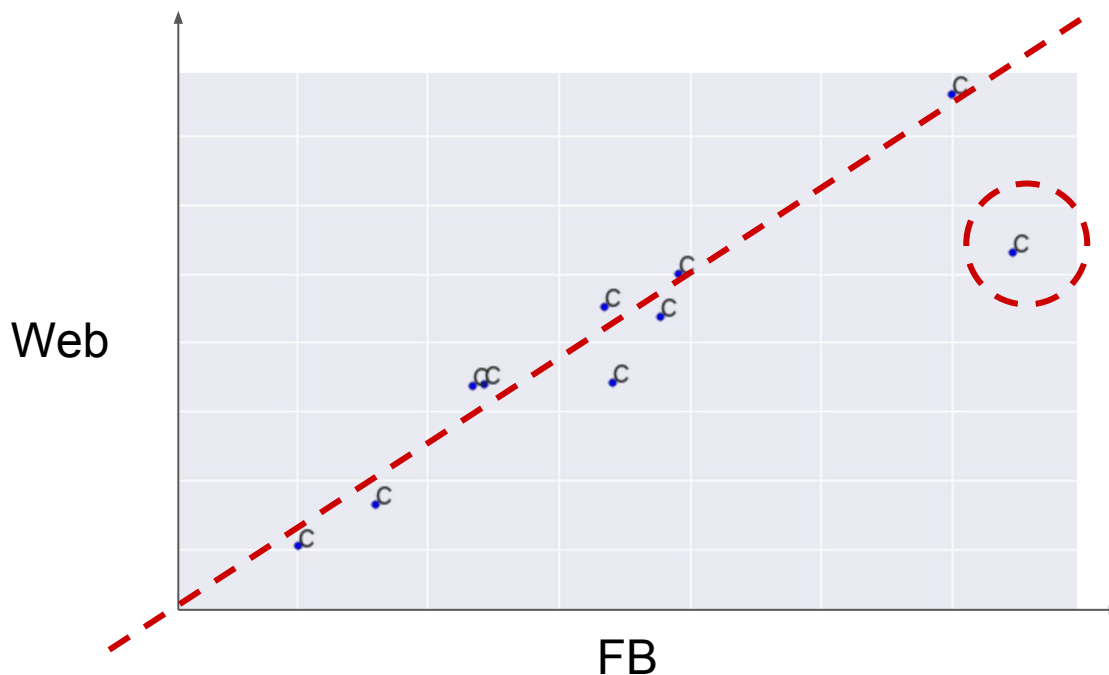 - As part of/ one way of analysis
- Publication

Project timetable

- Plan ahead
- Regularly review & revise
- Tailor the timetable when necessary
  - You can not do everything you initially intended

Item	Deadline	PIC	Dependency
Clean data	20170325	A	
Get GeoJSON of HK	20170405	B	
Visualise on a map	20170410	C	Get GeoJSON of HK
Draft the story	20170405	A	
Assemble webpage	20170415	C	Visualise on a map Draft the story

Pattern recognition & Anomaly Detection

- Prove "common sense":
  - Find patterns
  - Discover trends
  - Interpolate history
  - Predict future
  - "Explanatory News"

- Disprove "common sense":
  - Anomaly detection
  - "News"

Traffic distribution on web and on Facebook by channels. Real data from Initium Media, anonymised. (Pili Hu @ DNN, Dec 2015)

Ideation

Learn from master pieces

Some top medias:

- FiveThirtyEight: http://fivethirtyeight.com/
- TheGuardian: https://www.theguardian.com/data
- Financial Times Chinese. **Google it**: *inurl:www.ftchinese.com/interactive*
- Whashtington Post:
  https://www.washingtonpost.com/news/wonk/wp/category/data-visualization/
- NYT Upshot: https://www.nytimes.com/section/upshot
- Initium: https://theinitium.com/channel/interactive/

Awards:

- Data Journalism Awards (by GEN):
  https://www.datajournalismawards.org/shortlist/
- Information is beautiful awards:
  https://www.informationisbeautifulawards.com/

Useful Links

- [Tools] Carl V. Lewis, DataVis Tools , a curated guide to the best tools, resources and technologies for data visualization, http://dataviz.tools/

- [Templates & Patterns] CodePen, Collection of design patterns, http://codepen.io/patterns

- [Datasets] Xiaming Chen, Awesome Public Datasets, https://github.com/caesar0301/awesome-public-datasets

- [Readings] Chia-ni Liu, Hand-picked data journalism sites to follow, http://initiumlab.com/blog/20170401-data-news/

Communities

- Hacks and Hackers Hong Kong: https://www.meetup.com/Hacks-Hackers-Hong-Kong/
- DNN Society (HKBU): https://dnnsociety.org/

Data sourcing

Nation-grade government data portal in China

The Goto resource in its specific domain:

国家统计局数据门户	http://data.stats.gov.cn/
气象数据开放	http://www.weather.com.cn/
专利数据开放	http://patdata2.sipo.gov.cn
中国林业数据开放共享平台	http://data.forestry.gov.cn
国家地表水水质自动监测实时数据发布系统	http://58.68.130.147/
环保部数据中心	http://datacenter.mep.gov.cn/
全国企业信用信息公示系统	http://gsxt.saic.gov.cn/
中国政府网	http://www.gov.cn/shuju/index.htm
中国环境监测总站	http://106.37.208.233:20035/
中国裁判文书网	http://wenshu.court.gov.cn/
中国国家质检总局信息公开平台目录	http://www.aqsiq.gov.cn/xxgk_13386/xxgkztfl/zcfg/
中国全国法院被执行人信息查询	http://zhixing.court.gov.cn/search/
对外承包工程企业名录	http://wszw.hzs.mofcom.gov.cn/fecp/zsma/corp/corp_ml_list.jsp?
人民检察院案件信息公开网	http://www.ajxxgk.jcy.cn/html/index.html

Major data exchange platforms in China

- Can commission for certain customised data on some platforms

平台名称	链接
贵州大数据交易所	http://www.gbdex.com/website/
长江大数据交易中心	http://www.cjbigdata.com/
武汉东湖大数据交易中心	http://www.chinadatatrading.com/
江苏省大数据交易中心	http://www.bigdatahd.com/
华中数交所	http://www.ccbde.cn/
钱塘数据大数据交易中心	http://www.qtjiaoyi.com/
聚合数据	https://www.juhe.cn/docs
阿凡达数据	http://www.avatardata.cn/
HaoService	http://www.haoservice.com/
极速数据	http://www.jisuapi.com/
API Store	http://apistore.baidu.com/
京东万象	https://wx.jcloud.com/
数粮	http://www.datasl.com/
数据宝	https://www.chinadatapay.com/
数多多	http://www.dataduoduo.com/
大海洋	http://www.dahaiyang.com/
发源地	http://www.finndy.com/
数据堂	http://www.datatang.com/
优易数据	http://www.youedata.com/
通联数据商城	https://mall.datayes.com/
阿里云市场	https://market.aliyun.com/products/

Scraper (Chrome extension)

Source: http://ovt.lcsd.gov.hk/ovt/Listing.aspx?lang=en-US&q=fulllist

Step 1

Step 2

Outcome:
https://docs.google.com/spreadsheets/d/1w1gpnN5bSPHaiEpRC
RdONc0g2xdd-gTXSf9phiKkKtQ/edit#gid=81631626

Table Capture

IMPORTHTML in Google Spreadsheet

=IMPORTHTML("https://www.tbfacts.org/cases-of-tb/", "table",1)

The data will be automatically updated with the data source (HTML)

=IMPORTHTML("https://www.tbfacts.org/cases-of-tb/", "table",1)

	A	B	C	D	E
	Country	New & relapse ca	Mortality	WHO Region	
	Afghanistan	37,001	12,000	EMR	
	Algeria	23,879	3,300	AFR	
	Angola	61,060	18,000	AFR	
	Argentina	10,506	810	AMR	
	Armenia	1,104	110	EUR	
	Australia	1,254	48	WPR	
	Azerbaijan	5,701	58	EUR	
	Bangladesh	209,438	73,000	SEAR	
	Belarus	4,177	530	EUR	
	Benin	4,092	1,300	AFR	
	Bolivia	7,893	510	AMR	
	Bosnia and Herz	1,095	120	EUR	
	Botswana	5,073	2,000	AFR	
	Brazil	81,137	7,700	AMR	
	Bulgaria	1,660	90	EUR	
	Burkina Faso	5,808	1,900	AFR	
	Burundi	6,969	3,300	AFR	

http://www.bazhuayu.com/
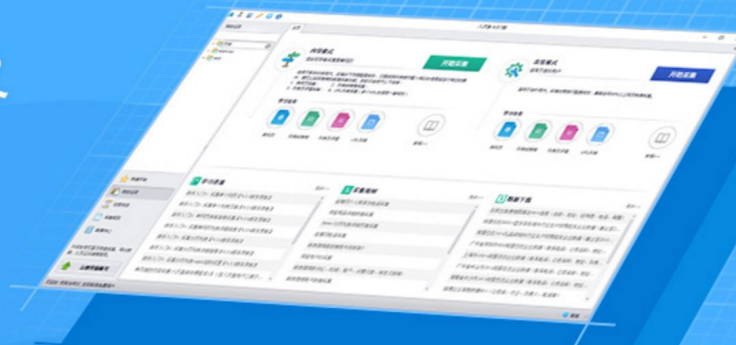
Today's news will be history

Search for the past news to get a picture of the history.

-- The core idea of "Journalism"

Importance of "keep an eye on"

- Every good data journalist has **a database of database (DoD)**
  - What website shall I search when studying topic of this area?
  - What database has my organisation subscribed to?
- Keep on taking notes and organising information
- People do organise their own DoD
  - https://github.com/caesar0301/awesome-public-datasets

Data analysis

Nottingham pupil dataset

Prepare:

[Data Download](#)

[Download Excel Analysis ToolPack](#)

Optional reading:

[Python Notebook](#)

Load the Analysis ToolPak in Excel

Applies To: Excel for Office 365, Excel for Office 365 for Mac, Excel 2016, More...

If you need to develop complex statistical or engineering analyses, you can save steps and time by using the Analysis ToolPak. You provide the data and parameters for each analysis, and the tool uses the appropriate statistical or engineering macro functions to calculate and display the results in an output table. Some tools generate charts in addition to output tables.

The data analysis functions can be used on only one worksheet at a time. When you perform data analysis on grouped worksheets, results will appear on the first worksheet and empty formatted tables will appear on the remaining worksheets. To perform data analysis on the remainder of the worksheets, recalculate the analysis tool for each worksheet.

Data Visualisation

Chart Suggestions—A Thought-Starter

http://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf
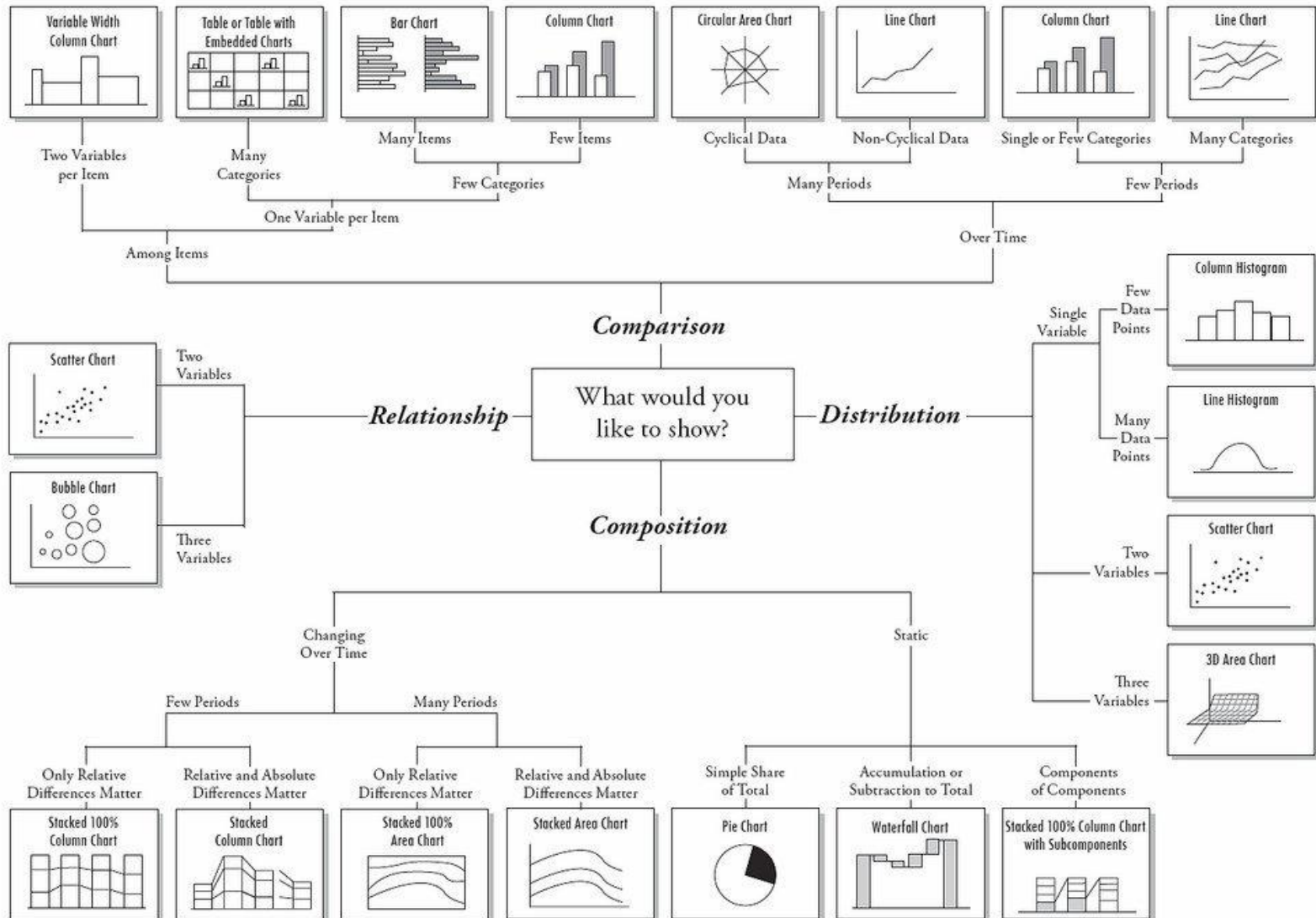
Sankey Diagram
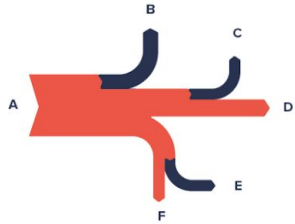
Alluvial Diagram
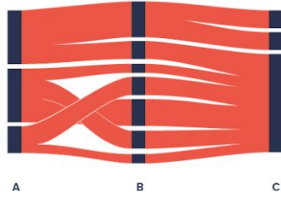
Matrix Diagram
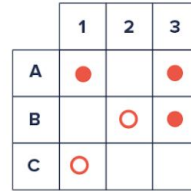
Donut Chart

Matrix Diagram (Roof Shaped)
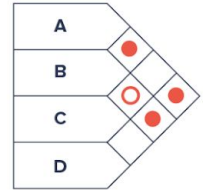
Radial Bar Chart
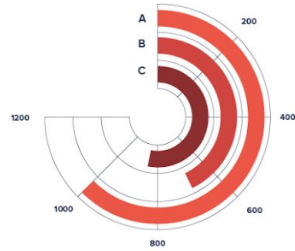
Sorted Stream Graph
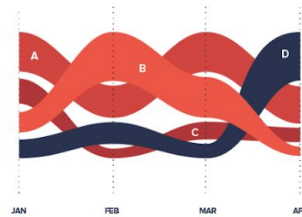
Radial Histogram

Fishbone Diagram
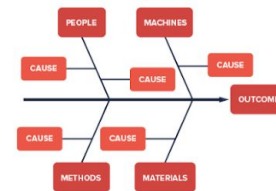
Pictorial fraction chart

Isoline Map

Flow Map

Exploded View Drawing
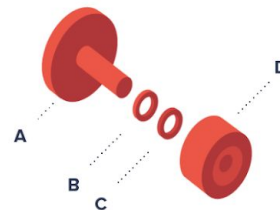
Arc Diagram
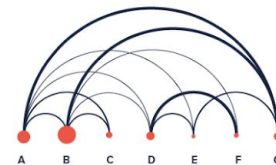
Pictorial Stacked Chart

http://datavizproject.com/
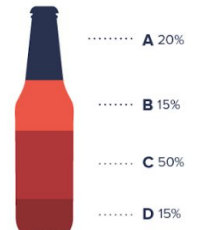
Arc Diagram   Area Graph   Bar Chart   Box & Whisker Plot   Brainstorm   Bubble Chart

Bubble Map   Bullet Graph   Calendar   Candlestick Chart   Chord Diagram   Choropleth Map
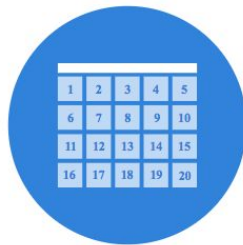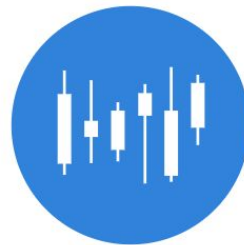
Circle Packing   Connection Map   Density Plot   Donut Chart   Dot Map   Dot Matrix Chart
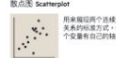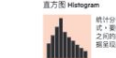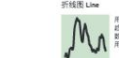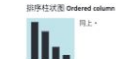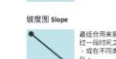
https://datavizcatalogue.com/

可视化辞典 / Visual vocabulary

离差 Deviation
强调相对于一个固定参考值的变化（正／负值），通常参考值为零，但也可能是一个目标数值或是长期平均值。也能用来展现情绪倾向（正／中正／负面）。

FT使用范例：贸易顺差／逆差、气候变化

- 分向条形图 Diverging bar — 一种简单、标准化的条形图，能用来展示正、负值。
- 分向堆叠条形图 Diverging stacked bar — 最适合用来展现问卷调查的（正向／中立／负向）的调查结果。
- 成对条形图 Spine — 将单一数值分成两组相互对比的组成（例如男性／女性）。
- 盈余／赤字填充线图 Surplus/deficit filled line — 把数据分布展示出一种平衡关系。曲线在超过或低于某一基准线时，可标示为两组不同的数值之间。

相关性 Correlation
展示两个或多个变量之间的关系。请注意除非告诉读者，否则多数人会假设你展示的两个变量之间存在因果关系（例如一个变量导致另一个变量变化）。

FT使用范例：通胀与失业率、收入与平均寿命

- 散点图 Scatterplot — 用来描绘两个连续变量关系的标准方式，每一个变量有自己的轴线。
- 折线图+柱状图 Line + Column — 既能显示一个数量（柱状图）和另一个比率（折线图）的关系。
- 链接散点图 Connected scatterplot — 通常用来展示两个变量的关系如何随着时间而改变。
- 泡泡图 Bubble — 与散点图类似，但会用泡泡第三变量的大小来决定其相对大小。从而直观显示数据的大小。
- XY热图 XY heatmap — 适合展示两组不同类别之间关系的好方法。但不适合展示数据的精微差异。

排序 Ranking
当某个项目在排序列表中的位置比其他数值本身更重要时，就可以使用排序。多数读者会从最高值向下看，不妨着重标注你关注的重点。

FT使用范例：财富、贫困、排名表、选区选举结果

- 排序条形图 Ordered bar — 依照顺序排列，标准条形图是最直观的排序。
- 排序柱状图 Ordered column — 同上。
- 排序比例符号 Ordered proportional symbol — 当展示非常大差异的数值、或者有排序但需要显示比例时的选择。
- 点状条形图 Dot strip plot — 圆点使在线条上排列有的标准方式，能节省空间来布局重要数据。
- 坡度图 Slope — 最适合用来展示以排列组合一段时间之后的变化。现在不同类别间的排序关系。
- 棒棒糖图 Lollipop — 棒棒糖图比标准的条形图或柱状图更吸引人们注意图形重点，也有有效展示数据的排序和大小。
- 凹凸线图 Bump — 可以很好地展示排名有多个元素的变化。对于下方的数据集，可通过用颜色来进行分组。

分布 Distribution
显示数据集中的数值及其呈现的频率。分布的形状（或离差或平坦）是一项数据的一项清晰、一致或不平均的方式。

FT使用范例：收入分配、人口（年龄／性别）分布

- 直方图 Histogram — 统计分析的标准呈现方式，要将一条长的基准条以适当的间距放置以显示出"形状"。
- 箱线图 Boxplot — 适合于比较（中值）数据的数据集——用来抵消多个数据分布。
- 小提琴图 Violin plot — 类似于箱线图，但可以更有效地显示复杂的分布。（当数据无法用简单描述的平均数和间隔时）。
- 人口金字塔 Population pyramid — 展示人口年龄和性别分布的标准方式，实际上就是对背靠的条形图。
- 点状条形图 Dot strip plot — 用来呈现单个中值的相对好方式。适合于展示相邻大小的值，别于某一范围。
- 点状图 Dot plot — 呈现不同类别数值的变化范围（最小值／最大值）。
- 二维条码图 Barcode plot — 如同点状条码图，适合在一处展示并呈现有数据，较好的展示高值数据的密度。
- 累积曲线图 Cumulative curve — 呈现数据分布不均等的好方法，Y轴标注出数据累积量，X轴显示变量状态。
- 频率多边图 Frequency polygons — 此同时显示多个数据分布的好方法，最好一次只展现3-4组数据。

随时间的变化 Change over Time
强调趋势的变化，有可能是短周期（一日内），也可能是长周期（几千万年的变化），选择合适的时间间隔很重要。

FT使用范例：股价变动、经济动态时间序列

- 折线图 Line — 用来显示时间序列变化的标准、首选的呈现方式，如果符号没关系示数据点。
- 柱状图 Column — 适合，适用于不显示连续序列的数据，但单个柱之间的空隙会让读者一次只呈现一个数列。
- 折线图+柱状图 Line + column — 把两组不同类型的数据结合（柱状图和折线图），有时间间隔的关系上。
- 股价图 Stock price — 通常着重于数据每日的动态，包含有每日的开盘价、收盘价，以及点状和柱状图。
- 坡度图 Slope — 只要数据是简化为2项或2个点不误关键变量变化，此图便能适当地简约且明了不同。
- 区域图 Area chart — 既小心使用，适合合能显示数值的整体变化，但读者会把区域视为整体。
- 扇形图（预测）Fan chart (projections) — 用来呈现未来预测值的不确定性。通常扇形越越大便意味不确定性越大。
- 链接散点图 Connected scatterplot — 适合同示一个变量随数据相关。只要数字有相对于时间的顺序。
- 日历式热图 Calendar heatmap — 适合展示有规律（每日／每周／每月）、但依照特定数据持续时间的展示。
- 普利斯特利时间轴 Priestley timeline — 适用于以清楚对比长度来呈现年的关键变更事件。（在表示出时间的长久、维度到的数量的变化）。
- 圆圈时间轴 Circle timeline — 适合展示不同类型数据的不连续数值（例如在同周期的地震发生度）。
- 垂直时间轴 Vertical timeline — 以Y轴表示时间、在带有动设备上清楚地呈现时刻，是展示此类详细的好时刻顺序。
- 震波图 Seismogram — 面着时间轴的替代选择，适合展示数据大概率波动的数列。

规模 Magnitude
用来比较数据的规模。有可能是绝对对短规（显示出单一一个比较大的）。有可能显示比较对短规（需要显示出精微的差异）。通常用来比较（例如苹果、人、美元）而不是经计算百分比或率（后分比）。

FT使用范例：大宗商品产量、市值

- 柱状图 Column — 以其简便性展现规的标准方式，精确比较起始值较为0。
- 条形图 Bar — 同上。适用于不那针对排序的数据，以及那列名称较多的数据。
- 成对柱状图 Paired column — 类似于一般柱状图，但却用同样空间表达两到三个数值。但别让一组数据过多，混淆焦点来互相比较。
- 成对条形图 Paired bar — 同上。
- 比例堆叠条形图 Marimekko — 适合用来同时比较两数据集，把两个比率都放一起——既能展示出比率的关系，以同时展示出整体的关系。
- 比例符号图 Proportional symbol — 当数据是大差异的数值时，以及需要显示精确差别时，使用此法图圈。
- 象形符号 Isotype (pictogram) — 特别适用有某例—目让只用数字呈现的。以符号呈现有一个，便可以以一组引导案来显示示意（不一组小数据）。
- 棒棒糖图 Lollipop — 比较棒棒糖有时比标准表把更把人们引入，通过其起始值或数值呈现为0，但适合对为0。
- 雷达图 Radar — 呈现多个变量的一个有空间的方式，但务必设法值都照同同的量度。需要有一定的据量。
- 文氏图 Venn — 适合仅用于简化的数据呈现。
- 瀑布图 Waterfall — 当数据组成部分为负值时，适合用量合来呈现出部分和整体的关系。

部分和整体的关系 Part-to-whole
能显示出一个整体如何被拆解成不同组合。如果组成份只是想要了解的组成部分的大小，不妨注意规模来。

FT使用范例：财政预算、公司架构、全国选举结果

- 堆叠柱状图 Stacked column — 以简单但很直观显示部分与整体。但比对各组成份值困难得较多。
- 比例堆叠条形图 Marimekko — 同上。
- 成对柱状图 Paired column — 类似于一般柱状图，但却用同样空间用两种。
- 饼图 Pie — 用来呈现整体各部组成关系的常见方式，不要让太多组成部分和太小的项模糊，较难以同组中较多的大小。
- 甜甜圈图 Donut — 类似于饼图、但中间的空间能放入更多信息。例如总数，少较之。
- 树状图 Treemap — 适合用来呈现很多层关系的数据。比单只用柱状图来呈现，但要当中子集要注意层大小。
- 沃罗诺伊图 Voronoi — 把点转成或区域的图表，每个区域的中心—个——其距离中间的任何中心点，（沃罗诺伊便产生每种区域图示）。
- 拱形图 Arc — 一个半圈，通常用来呈现议会或改治组成果。
- 网格 Gridplot — 在地图上呈现百分比很好地，最适合比较展组成份的形式或范围图形。
- 文氏图 Venn — 适合用于简化的区分作为有整集合之变量。
- 瀑布图 Waterfall — 当有些组成部分为负值时，适合用瀑布来显示部分和整体的关系。
- 分组符号图 Grouped symbol — 柱状图、条形图的替代选择。特别是当数据组计数或需要靠数字来简单显示出的部呈现。

地理空间 Spatial
当数据中的精确位置和地理分布规律比其他信息对读者来决更重要时，可使用此类图表。

FT使用范例：人口密度、自然资源分布、自然灾害风险／影响、集水区域、投票结果调查

- 基础地理分布图（比率／比值）Basic choropleth (rate/ratio) — 把数值区别到地表上展示有规律对短规，以这让多个地理区域之间有更好的比，同时要密集在一个合理的基准区划上。
- 比例符号地图（数值／规模）Proportional symbol (count/magnitude) — 有比例符号来显示出图所示的大小、差等需的合图的数据，在图此便能将合给成量会产生重叠。
- 流向地图 Flow map — 在地图上描示有明确移动方向的信息。
- 等高线地图 Contour map — 在地理上描示有相同数值的区域—较为方便。但需要读者看懂地形面展示出正、负的值。
- 均等化示意地图 Equalised cartogram — 把地理上的单位区域把化为相同大小的六以小规图框图，组合以规等种大小数据代表义图化区区域部分。
- 缩放统计地图（数值）Scaled cartogram (value) — 放大或缩小地理单位，以呈现出地图于因此缩放地图变—个，（以呈式伊便缩放数据便示义呈统计）。
- 点状密度地图 Dot density — 用来显示一事件的地点、分集相比连接的分布，不妨注意让读者注意的方范围。
- 热度地图 Heatmap — 通过描示区域的色彩来表示不同区域值数据，与此不妨通过地理小之方，各具地比把区域不会同地方呈现以区域地呈现。

流向 Flow
向读者展示两个或两个以上的状态、情境之间的流动量或流动程度。这些图的状态、情境可能是逻辑关系或地理位置等。

FT使用范例：资金、贸易、移民、资讯、诉讼的流向、关系图表。

- 桑基图 Sankey — 呈现从一个情境到另另一个情境的流向变化—以凸各合数据组组最终结果。过程的目标。
- 瀑布图 Waterfall — 通过状态过程来描示增值的过程序化，最终值与起点之间的起伏变化—以凸合正或负的变动累合过程。
- 和弦图 Chord — 一种复杂如有弦力的图表，能在图形中可回流向的箭头，以凸合流间量展现。
- 网络图 Network — 在不同连接的对象之间，展示出关系关系度和内部关系者点。

数据的设计 Designing with data

有许多方法可以把数据可视化，但我们怎么知道要挑选哪一种？使用本表顶部的类别，来决定在你的故事中，哪一种数据关系最重要。接着选择类别下方不同种类的图表，找出最适合表达的图表。这份列表并不是详尽无遗的，也不是一份万能向导，而是一个有用的制作起点，能协助你设计出内容丰富、有意义的数据可视化。

KISS Principle and Don't Make Me Think

Wiki: *KISS (Keep It Simple and Stupid) is an acronym for "Keep it simple, stupid" as a design principle noted by the U.S. Navy in 1960*

https://en.wikipedia.org/wiki/KISS_principle
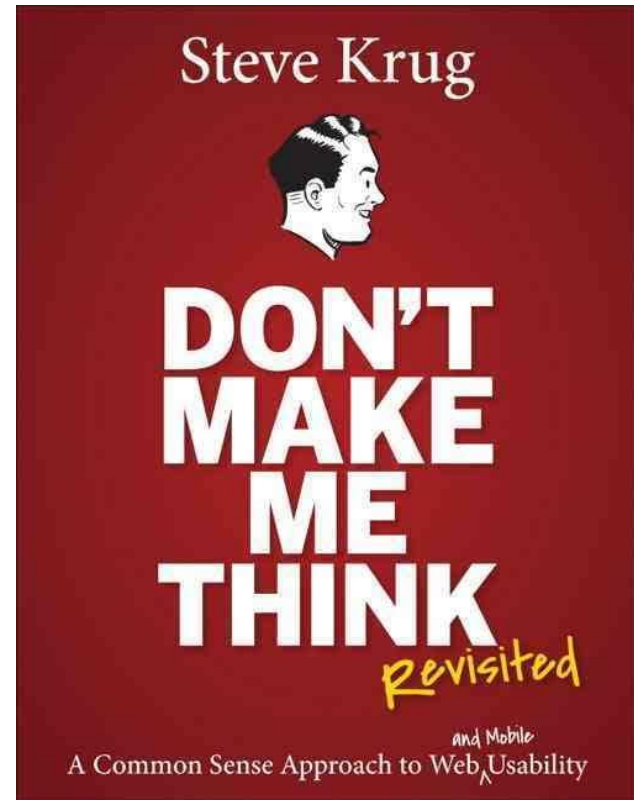
A classical textbook on Web Usability. Still inspiring today

infogram.com

- Charting, mapping
- Templates & Themes
- Quick annotation
- Integrate with cloud data source, e.g. Google Drive
- Multimedia elements
- Clear column layout

magic.piktochart.com

- Charting, mapping
- Infographics layout
- Rich template library
- Rich icon library
- Slick modern design
- Export interactive/ embeddable
- Section based; freestyle layout
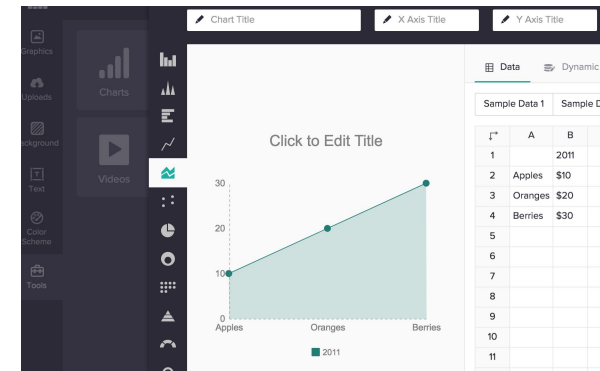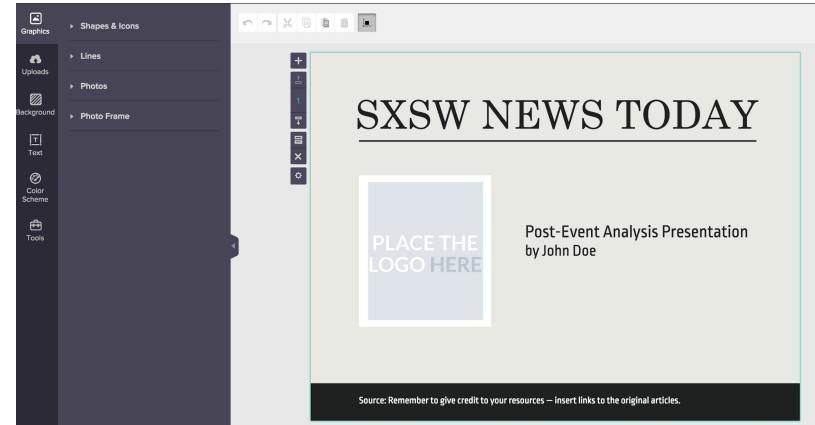
Price to Household Income Ratio (in year)

Amount (year)
- < 12.36
- 12.36 - 14.02
- 14.02 - 16.92
- 16.92 - 17.86
- > 17.86

https://magic.piktochart.com/output/18311094-dj
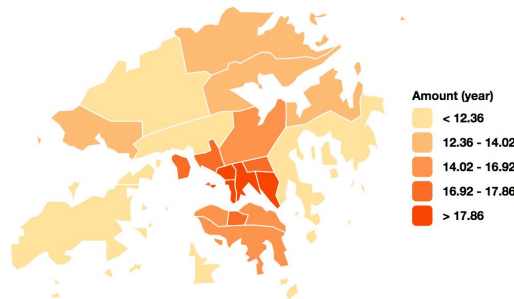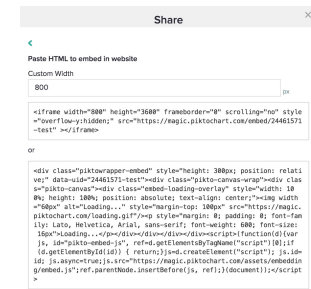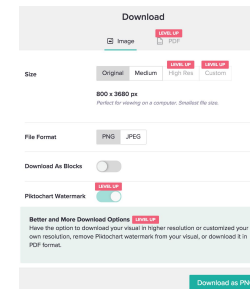
www.canva.com

- Freestyle layout
- Charting function is simpler
- Good for poster design
- Can export animated chart directly

plot.ly

- Designed for scientific visualisation community
- The abstraction is very similar to that of MATLAB or R
- Initially a commercial product
- Open-sourced in Nov 2015
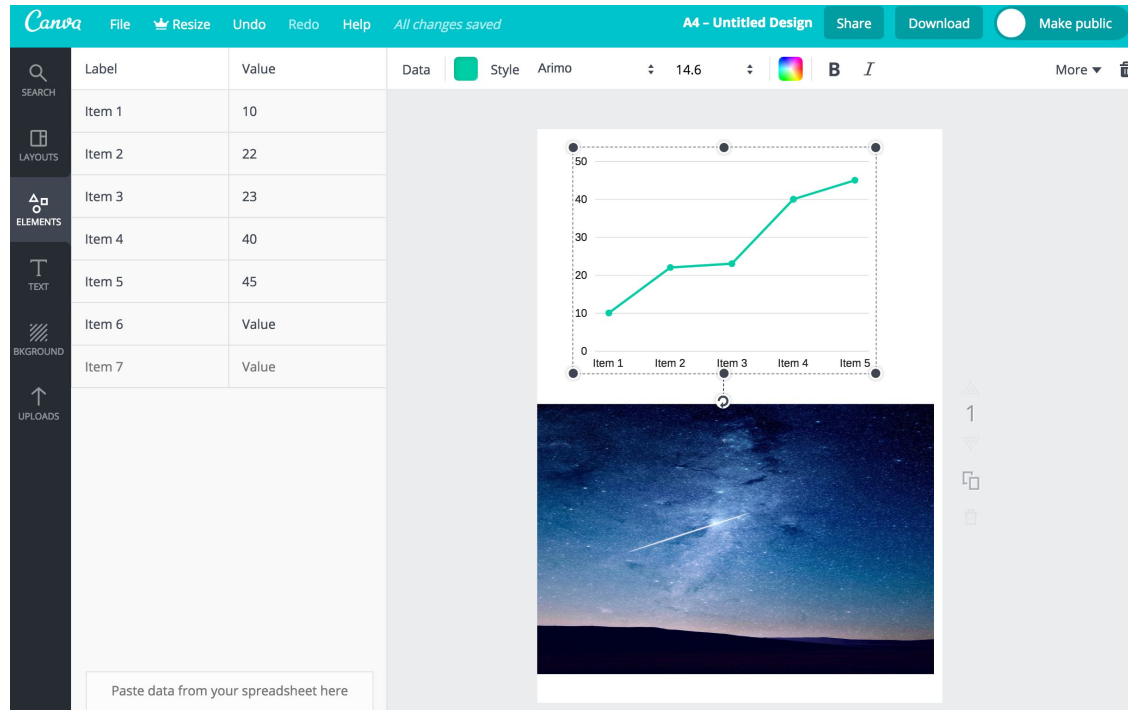
datawrapper.de

"By journalist, for journalist"

News of Oct 2017:

[Gregor Aisch](#) who worked for NYT Graphics before recently joined Datawrapper as CTO. [Lisa Charlotte Rost](#) who was a fellow in [NRP Visuals](#) also joined the team as a designer.

https://flourish.studio/

- Supported by Alberto Cairo
- Included in Google News Initiative
- 2 years, 10 people
- Support API for developers to extend the templates

海致BDP

Started from the Business Intelligence (BI) community

https://www.bdp.cn/home.html