

III Simposio Data Analytics

21 horas



**Sistemas Inteligentes para la Toma de
Decisiones
Fundamentos de Ingeniería de Datos**

Dr. Ing. Rodrigo Salas F. rodrigo.salas@uv.cl

- ▶ La habilidad para manipular y entender los datos es cada vez más crítico para el descubrimiento y la innovación.
- ▶ La ciencia de datos puede ayudar a conectar disciplinas, comunidades y los usuarios para proporcionar información más rica y profunda sobre los desafíos actuales y futuros.

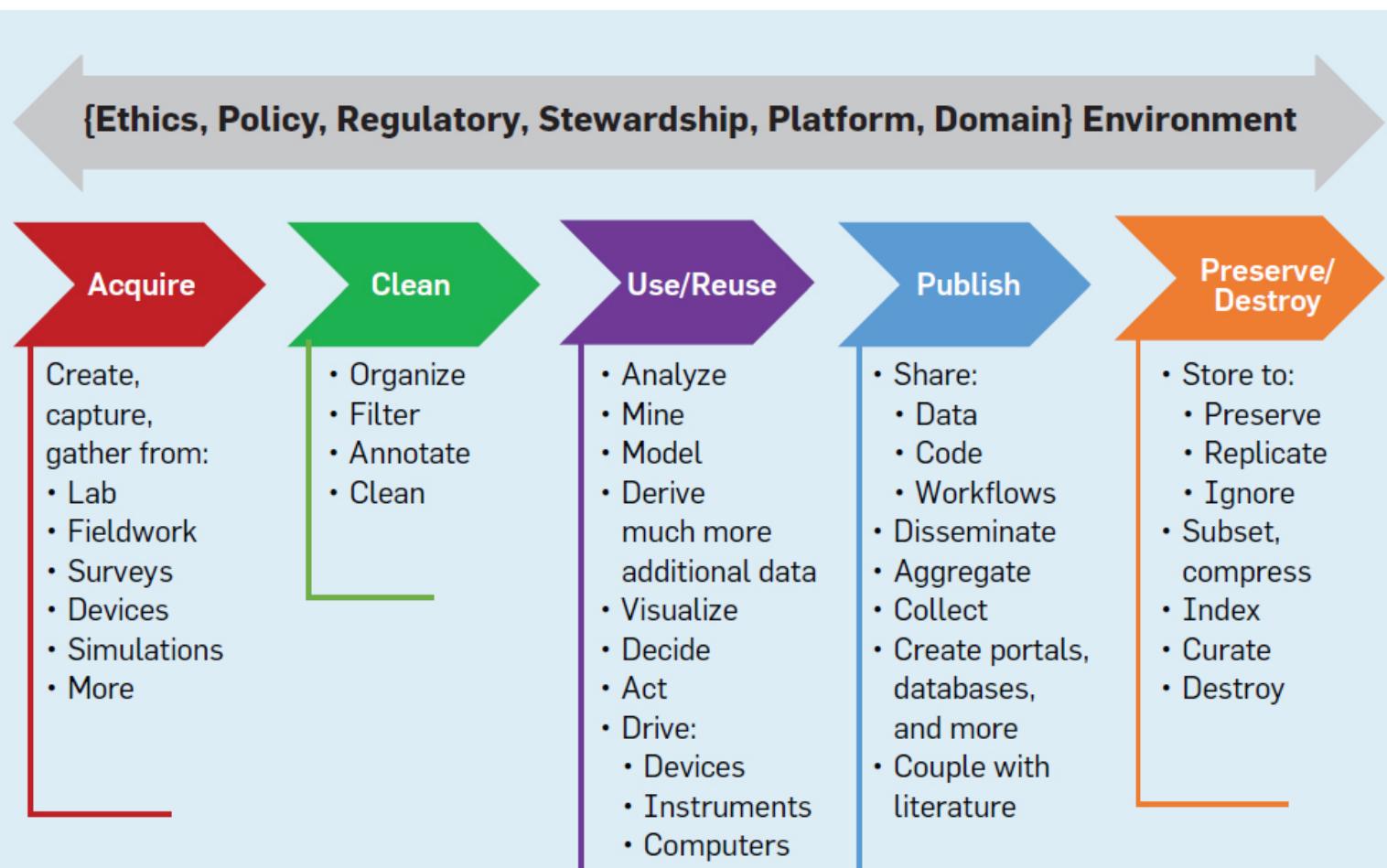
Data science promises new insights, helping transform information into knowledge that can drive science and industry.

- ▶ III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)

Ciencia de Datos

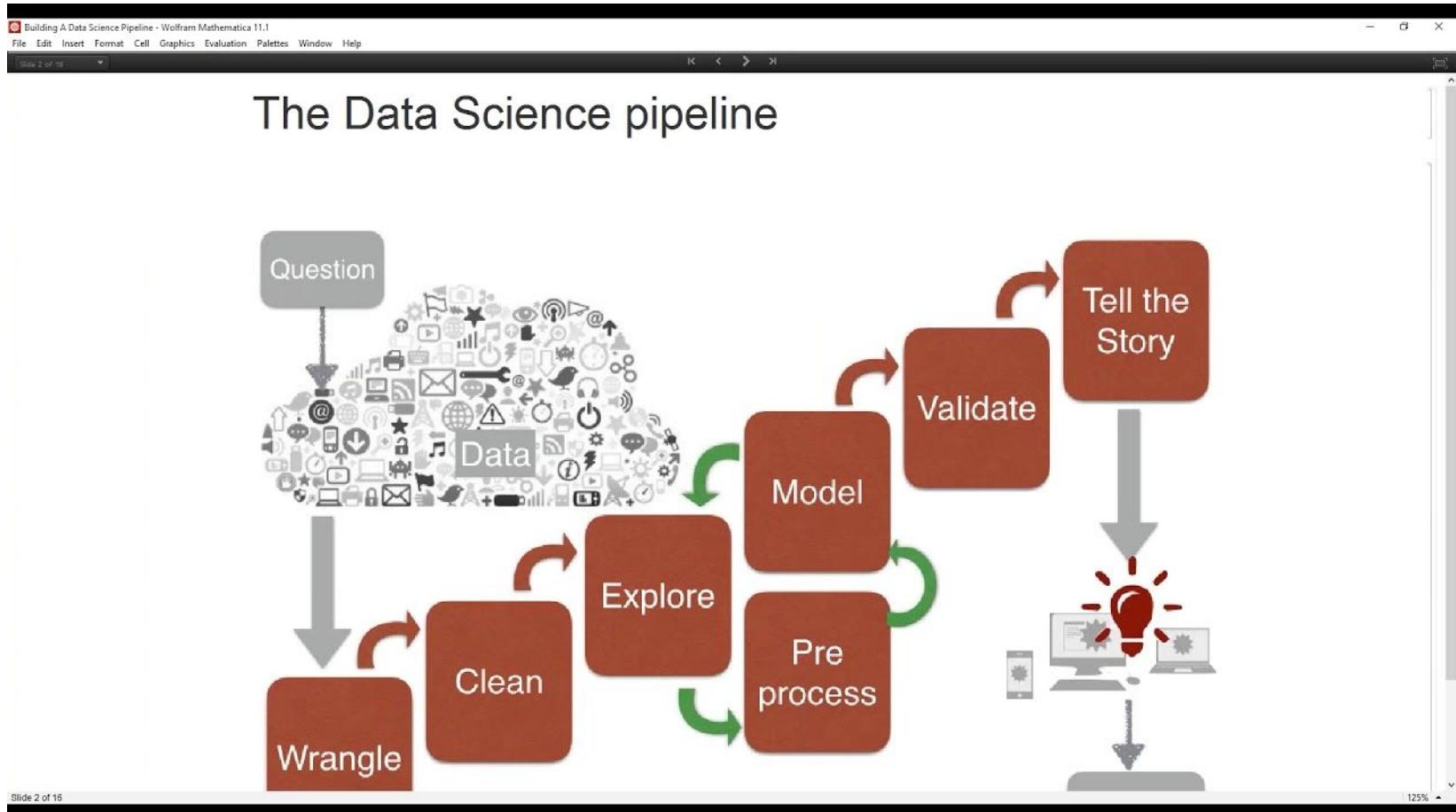
- ▶ La ciencia de los datos abarca un amplio conjunto de áreas, que incluyen la algorítmica centrada en los datos y el aprendizaje automático; la minería de datos y el uso de datos para el descubrimiento; recopilación, organización, administración y conservación de datos; Desafíos de privacidad y política asociada a los datos; y la pedagogía para apoyar la educación y capacitación de profesionales conocedores de datos.

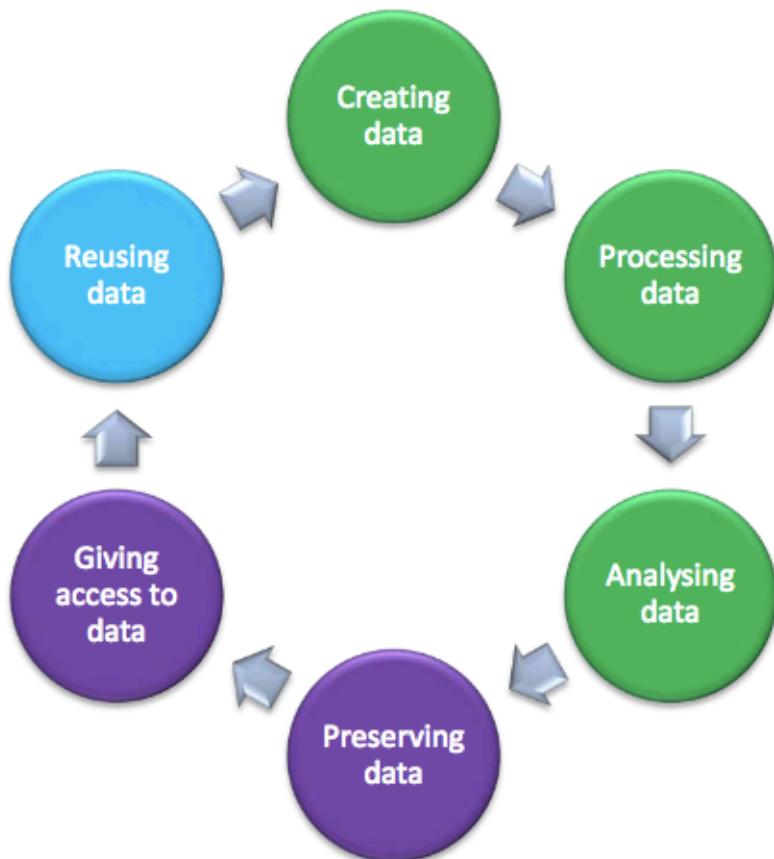
- ▶ III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)





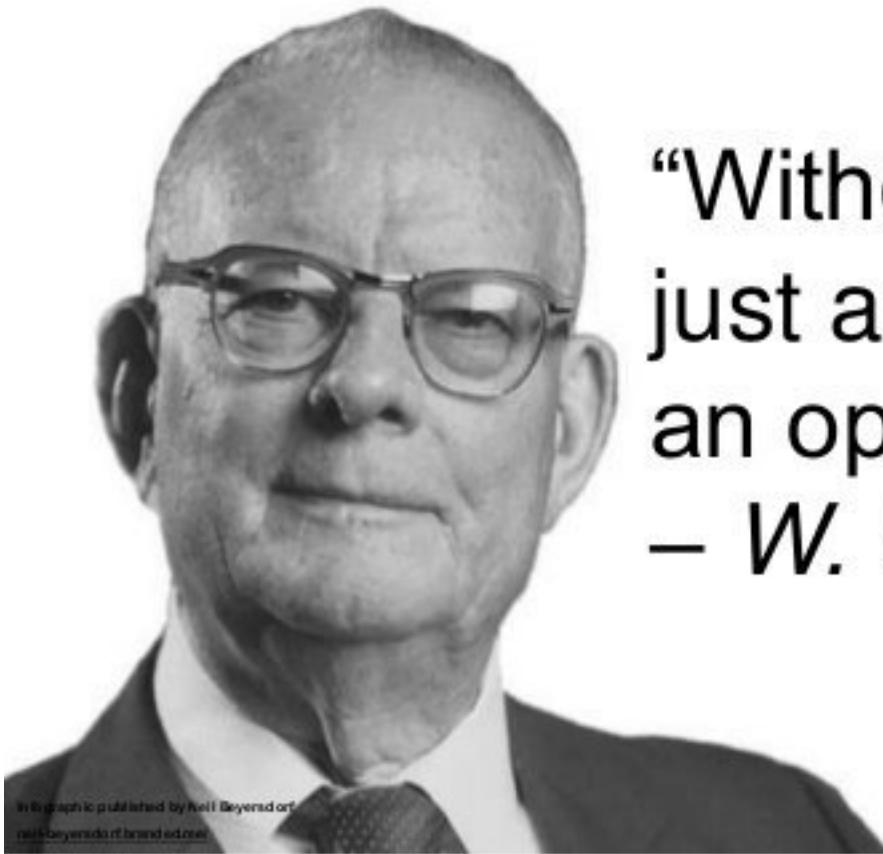
Pipeline de la Ciencia de Datos





- ▶ Los investigadores crean, procesan y analizan los datos.
- ▶ Los repositorios de datos tienen el rol de preservar y dar acceso a los datos
- ▶ Terceras personas reutilizarán los datos.

Pregunta de Investigación



“Without data you’re
just another person with
an opinion.”
– *W. Edwards Deming*

¿Qué es la pregunta de investigación?

- ▶ Una pregunta de investigación es la pregunta alrededor de la cual la investigación es realizada.
- ▶ Características:
 - ▶ claro: proporciona suficientes detalles que la audiencia puede entender fácilmente su propósito sin necesidad de una explicación adicional.
 - ▶ enfocado: es lo suficientemente estrecho para que pueda ser respondido a fondo en el espacio que permite la tarea de escritura.
 - ▶ conciso: se expresa en la menor cantidad de palabras posibles.
 - ▶ complejo: no se puede responder con un simple "sí" o "no", sino que requiere síntesis y análisis de ideas y fuentes antes de componer una respuesta.
 - ▶ discutible: sus posibles respuestas están abiertas a debate más que a hechos aceptados.

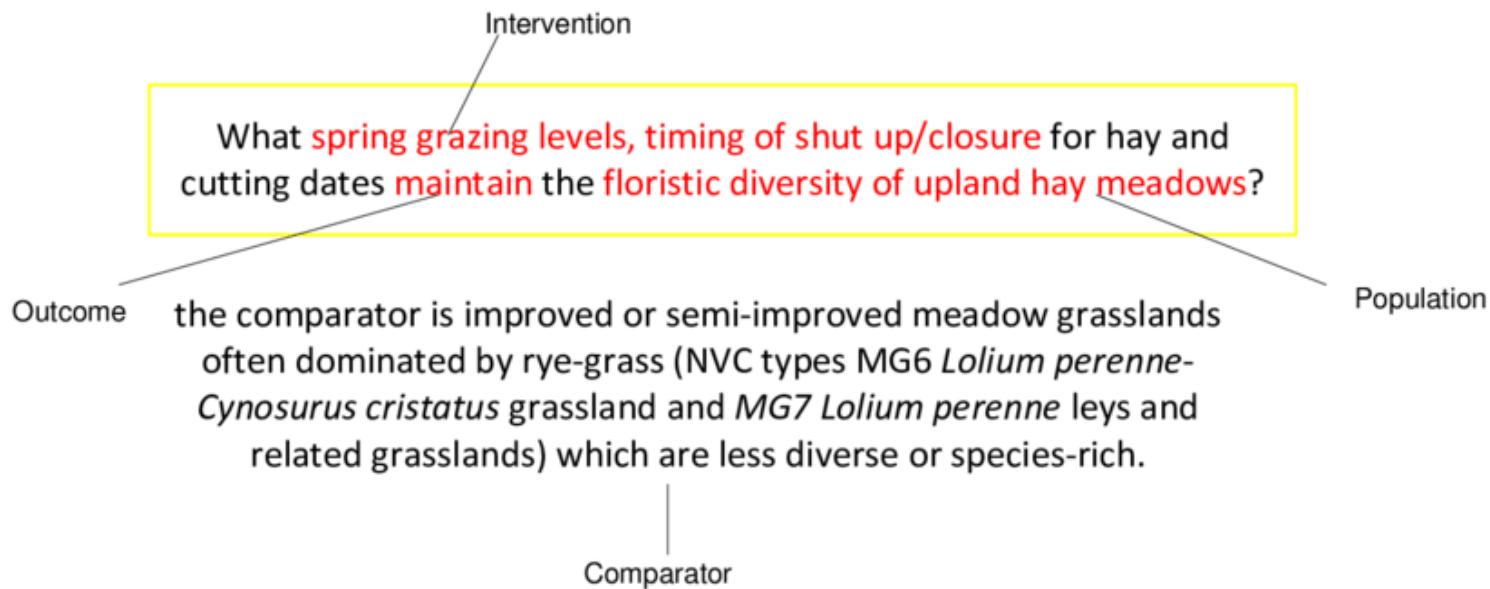
1. Elegir una temática general interesante
2. Realizar una investigación preliminar sobre la temática general seleccionada.
3. Tenga en consideración su audiencia
4. Comenzar a hacerse preguntas
5. Evaluar las preguntas para determinar si serían preguntas de investigación efectivas o si necesitarían más revisiones y refinamientos.
 1. ¿Es la pregunta de investigación clara?. Las preguntas de investigación deben ser lo más claras posible para que el escritor pueda dirigir su investigación de manera efectiva.
 2. ¿Está la pregunta de investigación bien enfocada?. Las preguntas de investigación deben ser lo suficientemente específicas para estar bien cubiertas en el espacio disponible.
 3. ¿Es tu pregunta de investigación compleja? Las preguntas de investigación no deben responderse con un simple “sí” o “no” o con datos fáciles de encontrar. Deberían, en cambio, requerir investigación y análisis por parte del escritor. A menudo comienzan con “¿Cómo?” o “¿Por qué?”.
6. Comenzar la investigación.

- ▶ **No está claro:** ¿Cómo deben los sitios de redes sociales abordar el daño que causan?
- ▶ **Claro:** ¿Qué medidas deben tomar los sitios de redes sociales como MySpace y Facebook para proteger la información personal y la privacidad de los usuarios?
- ▶ -----
- ▶ **Mal Enfocado:** ¿Cuál es el efecto sobre el medio ambiente del calentamiento global?
- ▶ -----
- ▶ **Enfocado:** ¿Cuál es el efecto más significativo del derretimiento glacial en la vida de los pingüinos en la Antártida?
- ▶ -----
- ▶ **Demasiado simple:** ¿cómo abordan los médicos la diabetes en los EE. UU.?
- ▶ **Apropiadamente complejo:** ¿Cuáles son los principales factores ambientales, de comportamiento y genéticos que predicen si los estadounidenses desarrollarán diabetes y cómo se pueden usar estos puntos comunes para ayudar a la comunidad médica a prevenir la enfermedad?
- ▶ -----
- ▶ <https://writingcenter.oumu.edu/guides/how-to-write-a-research-question>

- ▶ Características de una buena pregunta de investigación
 - ▶ Requiere que se realice un juicio
 - ▶ Permite un debate desde múltiples perspectivas
 - ▶ Permite respuestas que sintetizan múltiples perspectivas
 - ▶ Es simple - una idea en vez de varias
 - ▶ Se puede responder con los recursos disponibles
 - ▶ Debe ser interesante

ACRONIMO	DEFINICION
P	Población, pacientes, participantes o problema
I	Intervención, índice de test, factor de prognóstico, exposición
C	Comparación, control o test de referencia
O	Outcome o resultado esperado

Ejemplos



D. Stone (2013). Natural England Evidence Reviews: guidance on the development process and methods. D 10.13140/RG.2.1.3017.2004

Ejemplos

- Do adults with acute bronchitis who are treated with antibiotics note earlier improvement in clinical symptoms, compared to those who are given inhaled albuterol?
 - ✖ P - Patient population or problem
 - Adults with acute Bronchitis
 - ✖ I - Intervention
 - Antibiotics
 - ✖ C - Control or comparative intervention
 - Inhaled albuterol
 - ✖ O - Outcome
 - Earlier improvement in clinical sx's

EBM Question: In smokers with cough, does chest x-ray or chest CT have a better positive or negative predictive value for lung cancer?

P - Adult smokers with cough
I - Chest x-ray
C - Chest CT
O - Positive and negative predictive value for lung cancer

<https://slideplayer.com/slide/6925544/>

Definición de los Objetivos





Objetivos SMART



Specific (Específico)

- Los objetivos tienen que ser descritos **específicamente** de manera positiva.

- Claro el **Qué, Cuando y Cómo** para definir el **alcance**

Measurable (Medible)

- Los **logros** de los objetivos deben ser **medibles**.

- Que sea posible **cuantificar**, para poder **controlarlo**.

Attainable (Alcanzable)

- Debe ser **atractivo** para **el equipo** lograr los objetivos.

- Que podamos asignar **responsables**

Realistic (Realista)

- El objetivo debe ser **alcanzable de manera realista**.

- A la hora del **presupuesto** y de los recursos que **disponemos**

Time-bound (Oportuno)

- El objetivo tiene que establecerse dentro de un **marco de tiempo oportuno**.

- Definir el periodo de tiempo para completarlo



Manipulación de Datos



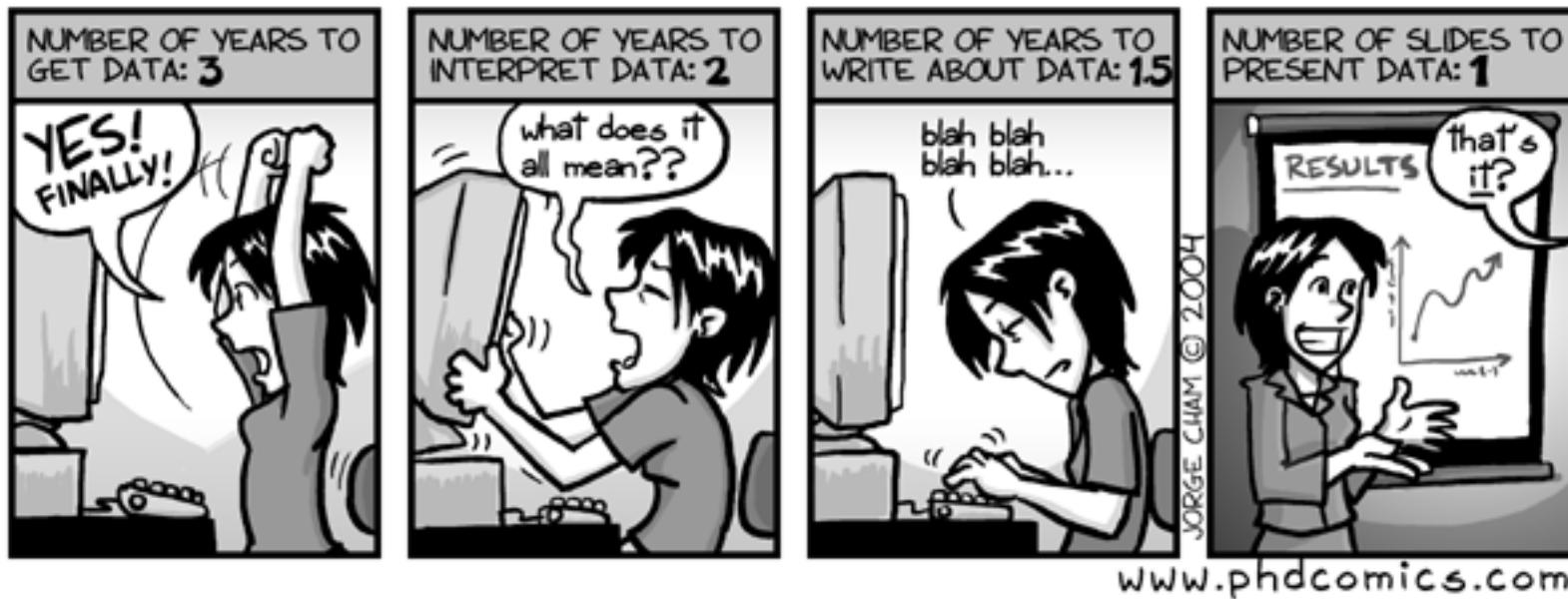
***"Torture the data, and it will
confess to anything."***

**Ronald Coase, winner of the
Nobel Prize in Economics**

Manipulación de datos

- ▶ Las actividades asociadas con la manipulación de los datos serían:
 - ▶ **Limpieza de datos:** remover los datos erróneos
 - ▶ **Edición de datos:** corregir los datos erróneos
 - ▶ **Imputación de datos:** procedimiento de sustitución de los datos faltantes
 - ▶ **Scraping de datos:** extraer una parte del conjunto de datos
 - ▶ **Wrangling y Munging los datos:** transformación de los datos a un formato que pueda ser utilizado
 - ▶ **Pre-procesamiento:** aplicación de una serie de técnicas con el fin de tener los datos filtrados y transformados.
 - ▶ **Fusión de datos:** proceso que consiste en integrar múltiples fuentes de datos.
 - ▶ **Integración de datos:** combinación de datos proveniente de diferentes fuentes de datos
- ▶ III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)

DATA: BY THE NUMBERS



Curación de Datos



I kind of have to be a master of
cleaning, extracting and trusting my
data before I do anything with it.

— *Scott Nicholson* —

AZ QUOTES

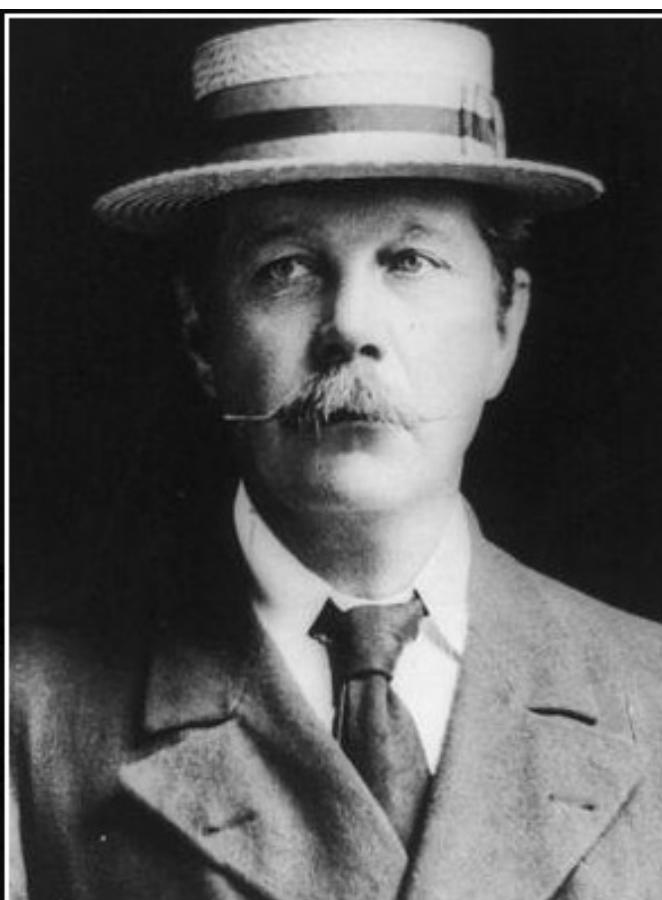
- ▶ **Curación de datos:** es la gestión activa y continua de los datos a lo largo de su ciclo de vida de vida de interés y utilidad para los estudios, la ciencia y la educación. La curación de datos permite el descubrimiento y recuperación de datos, mantiene la calidad de los datos, agrega valor y permite su reutilización a lo largo del tiempo a través de actividades que incluyen autenticación, archivo, administración, conservación y representación. (Universidad de Illinois)
- ▶ Se requiere de una curación activa para preservar la información de los datos.
- ▶ **Curación:** deriva del latín, de “curatio” y que es el producto de la suma de dos partes bien diferenciadas:
 - ▶ El verbo “curare”, que es sinónimo de “cuidar”.
 - ▶ El sufijo “-cion”, que se utiliza para indicar “acción y efecto”.
- ▶ III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)

- ▶ Asegurarse de que los datos estén bien descritos con metadatos utilizables por la máquina
 - ▶ • Asegurarse de que los datos estén completos, se expliquen por sí mismos y sean precisos (calidad)
- ▶ Proteger la confidencialidad y la privacidad al hacer que los datos estén disponibles (por ejemplo, eliminar identificadores, enclaves de datos virtuales)
- ▶ Tener en cuenta las necesidades de acceso y preservación a largo plazo.
- ▶ Identificar y / o crear repositorios digitales confiables para administrar datos a través del tiempo.

- ▶ Asignación de un DOI al conjunto de datos, por lo tanto estos tienen que ser:
 - ▶ Estables → no serán modificados
 - ▶ Completos → no van a ser actualizados
 - ▶ Permanentes → a través de la asignación de un DOI nos comprometemos a hacer que el conjunto de datos esté disponible para la posteridad.
 - ▶ Buena calidad → al asignar un DOI, le estamos dando el sello de aprobación diciendo que está completo y que todos los metadatos están disponibles.

Encontrando los patrones ocultos en los datos y sus tendencias





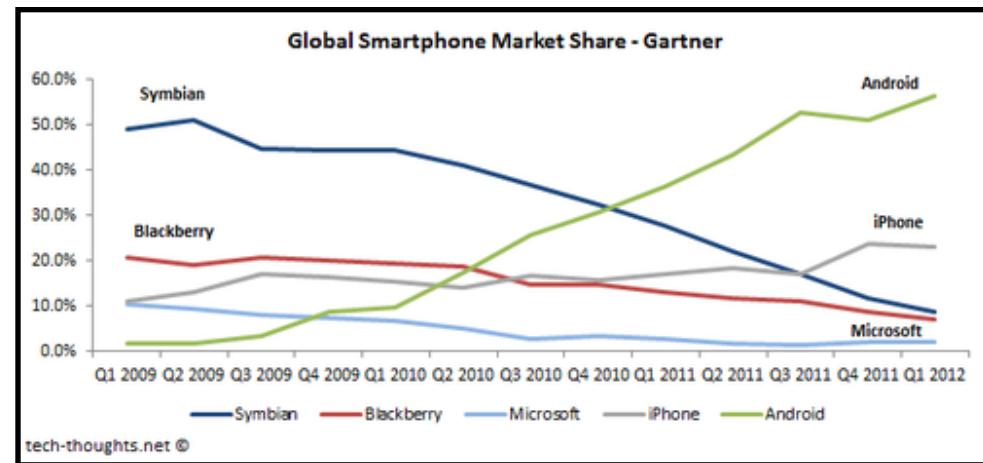
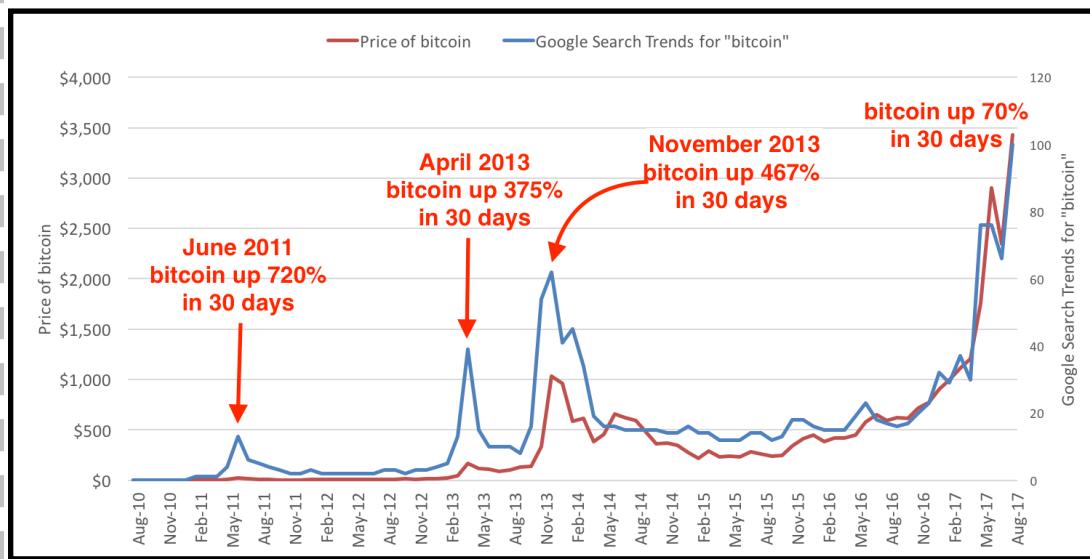
Data!data!data!" he cried impatiently. "I can't make bricks without clay.

— *Arthur Conan Doyle* —

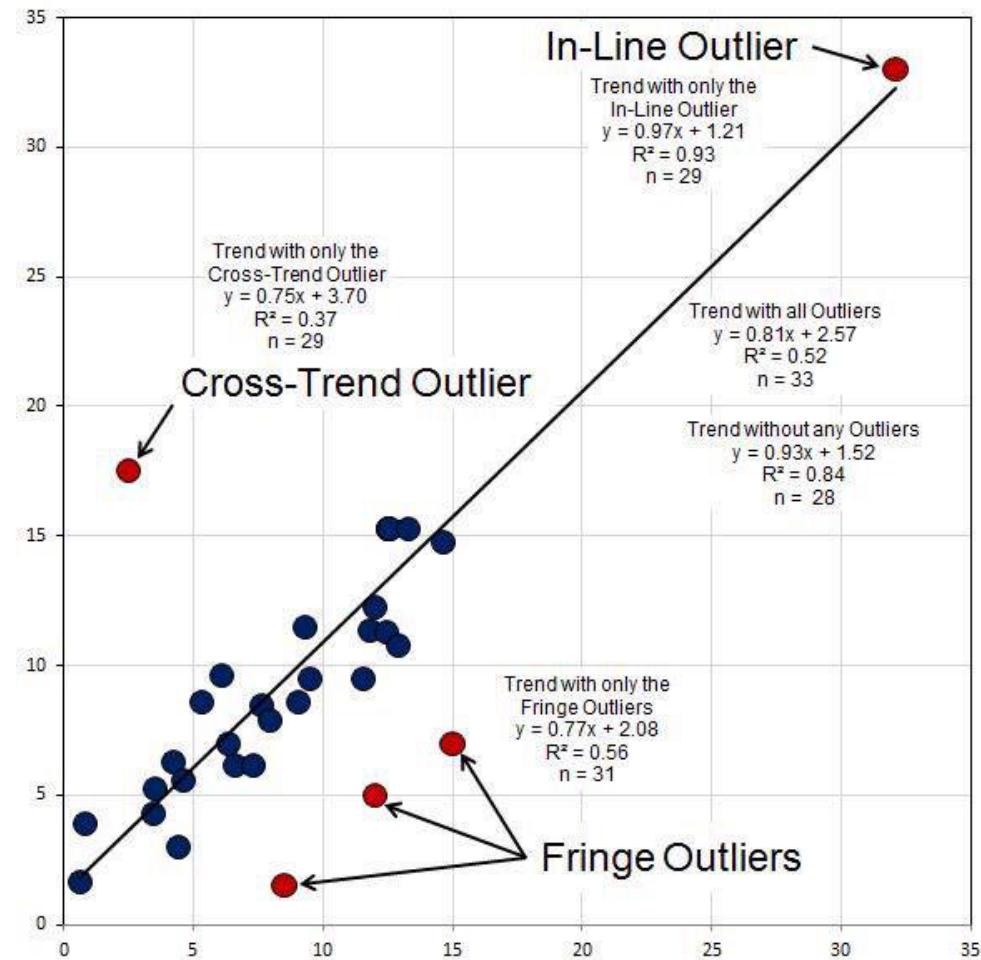
AZ QUOTES



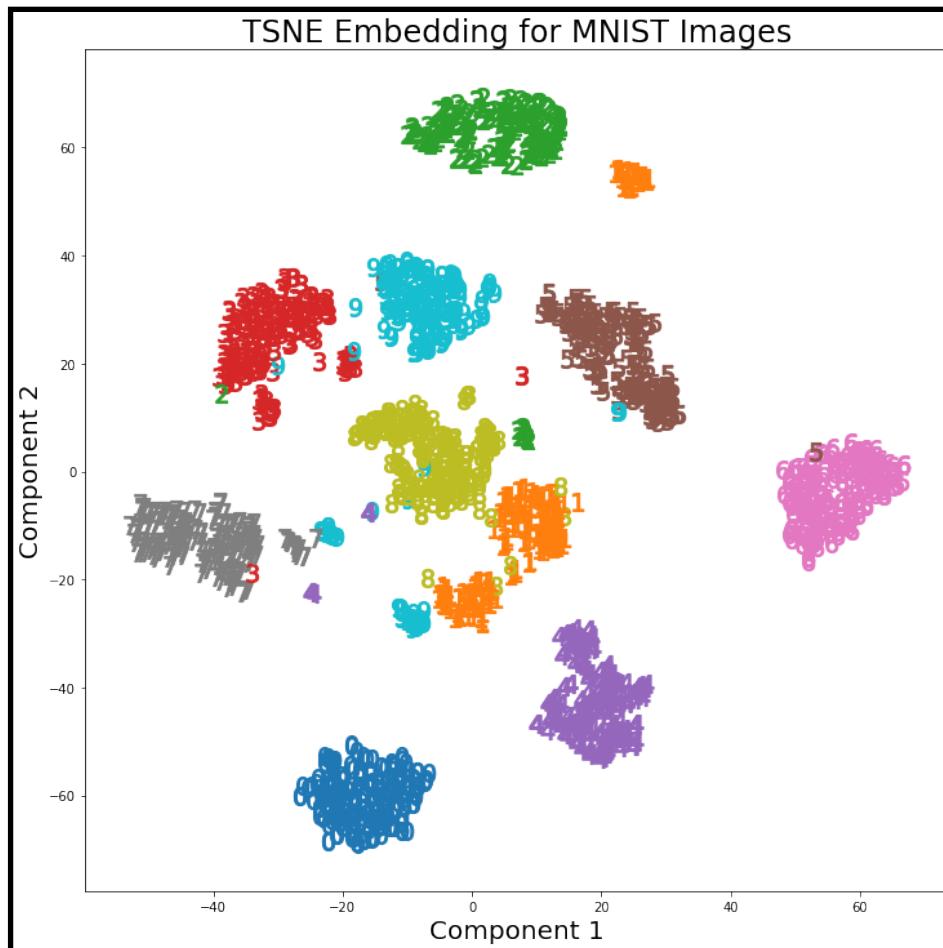
Tendencias



Outliers



High Dimensional Clusters



Narración de una historia con los datos



The W. EDWARDS
Deming
Institute



In God we trust, all
others must bring data.

*attribution disputed,
see source link

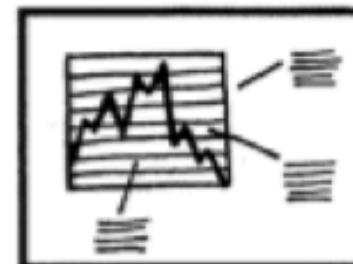
W. Edwards Deming

source: quotes.deming.org/3734

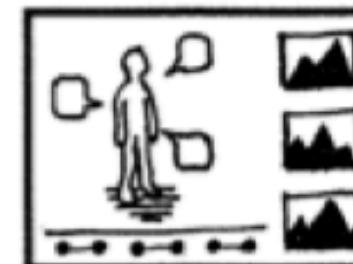
Seven Genres



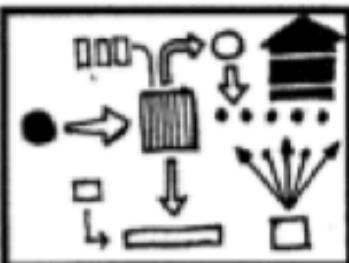
Magazine Style



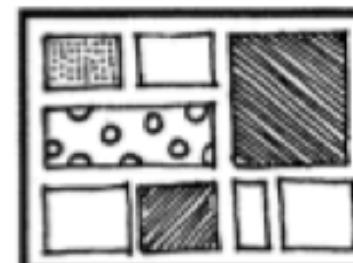
Annotated Chart



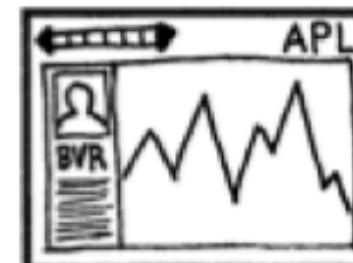
Partitioned Poster



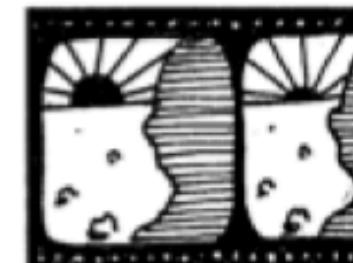
Flow Chart



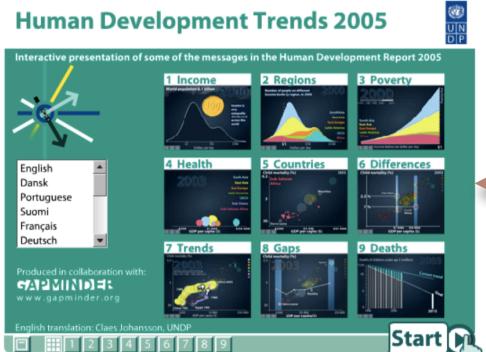
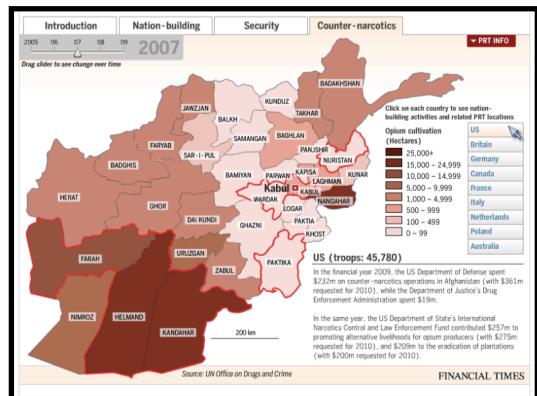
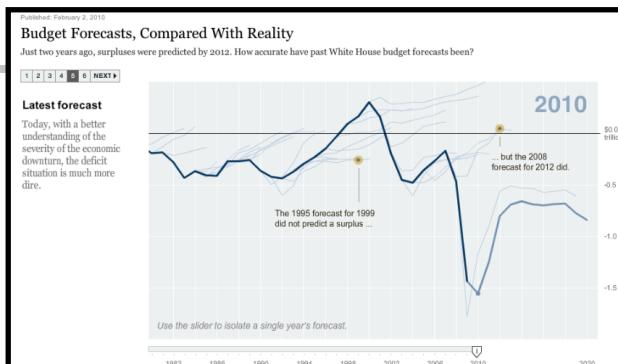
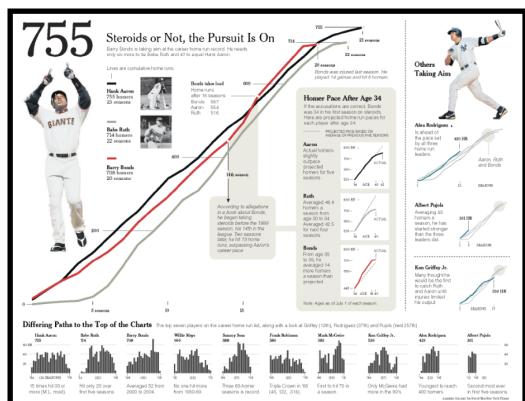
Comic Strip



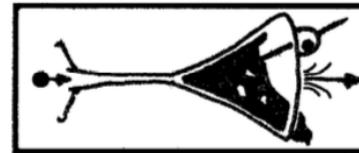
Slide Show



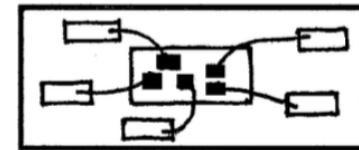
Film/Vide/Animation



- ▶ Frameworks narrativos:
 - ▶ Estructura visual (género): apoyar la historia
 - ▶ Interactivo: enganchar con la historia
 - ▶ Mensaje: contar la historia
- ▶ Enfoques:
 - ▶ Dirigidos por el autor: fuertemente ordenado, mensajes densos, muy poca interactividad
 - ▶ Dirigidos por el lector: débilmente ordenados, mensajes livianos, interactividad libre.
- ▶ Esquemas:
 - ▶ Estructura de copa de martini: prioriza el enfoque dirigido por el autor



- ▶ Historia drill-down: prioriza el enfoque dirigido por el lector

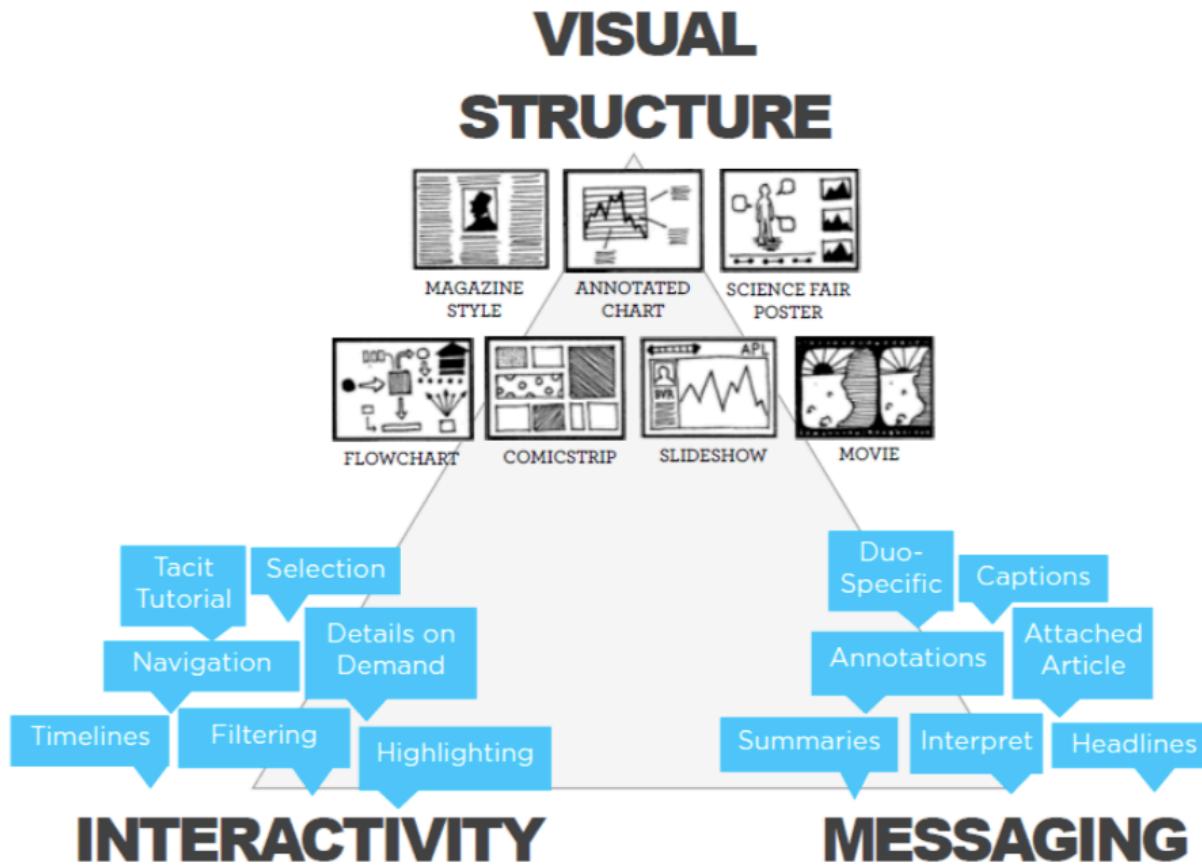


- ▶ Slide-show interactivos: promueve el diálogo entre ambos enfoques.



Definición de la historia y modelo

- ▶ Historia:
 - ▶ Secuencia ordenada de pasos
 - ▶ Cada paso contiene textos, imágenes, visualizaciones, videos, etc..
 - ▶ Caminos definidos a través de los pasos
 - ▶ Modelo periodista
 - ▶ Los periodistas recolectan la información a través de la investigación, entrevistas, etc... para reunir los hechos claves
 - ▶ Juntar los hechos claves (material básico) para producir la historia.
 - ▶ Modelo de Analista de Datos
 - ▶ Utilizar la visualización para la exploración y el análisis
 - ▶ Usar la visualización para la presentación (contar la historia) utilizando los resultados del análisis
 - ▶ Las herramientas usadas para el análisis podría no funcionar para la presentación.
- ▶ III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)



- ▶ Veena Mendiratta AS. A Storytelling with Data Visualization. ASA Workshop on Data Visualization Techniques.
- ▶ III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)

Design Space ... genres + interactivity + messaging

Author Driven

- strong order
- heavy messaging
- minimal interactivity

- clear story
- fast delivery
- author's message

Reader Driven

- weak order
- light messaging
- free interactivity

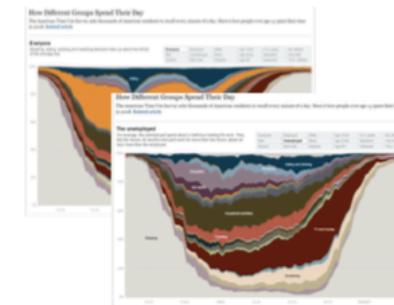
- query
- explore
- reader driven



martini glass



interactive
slide-show



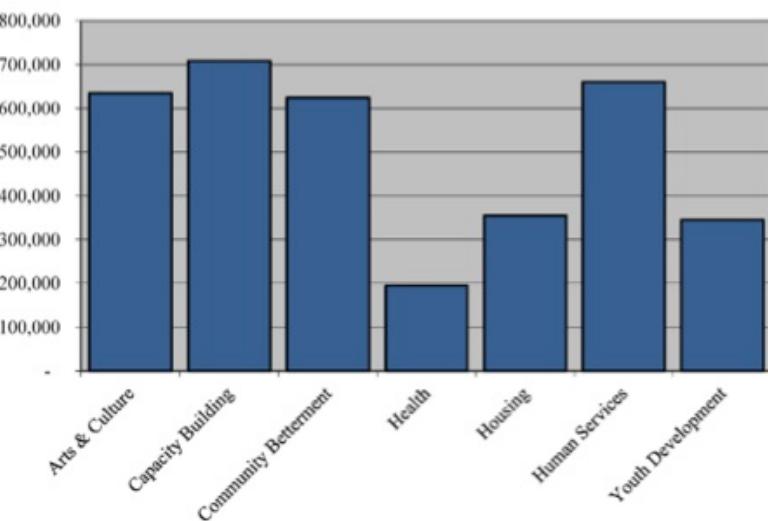
drill-down
story

- ▶ Veena Mendiratta AS. A Storytelling with Data Visualization. ASA Workshop on Data Visualization Techniques.
- ▶ III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)



Investment by area of impact

2006 - Present



We invest primarily in four areas

Since we began investing in 2006, **four areas have received more than \$600K each, accounting for 75% of total grantmaking activity**

Investment by Area of Impact

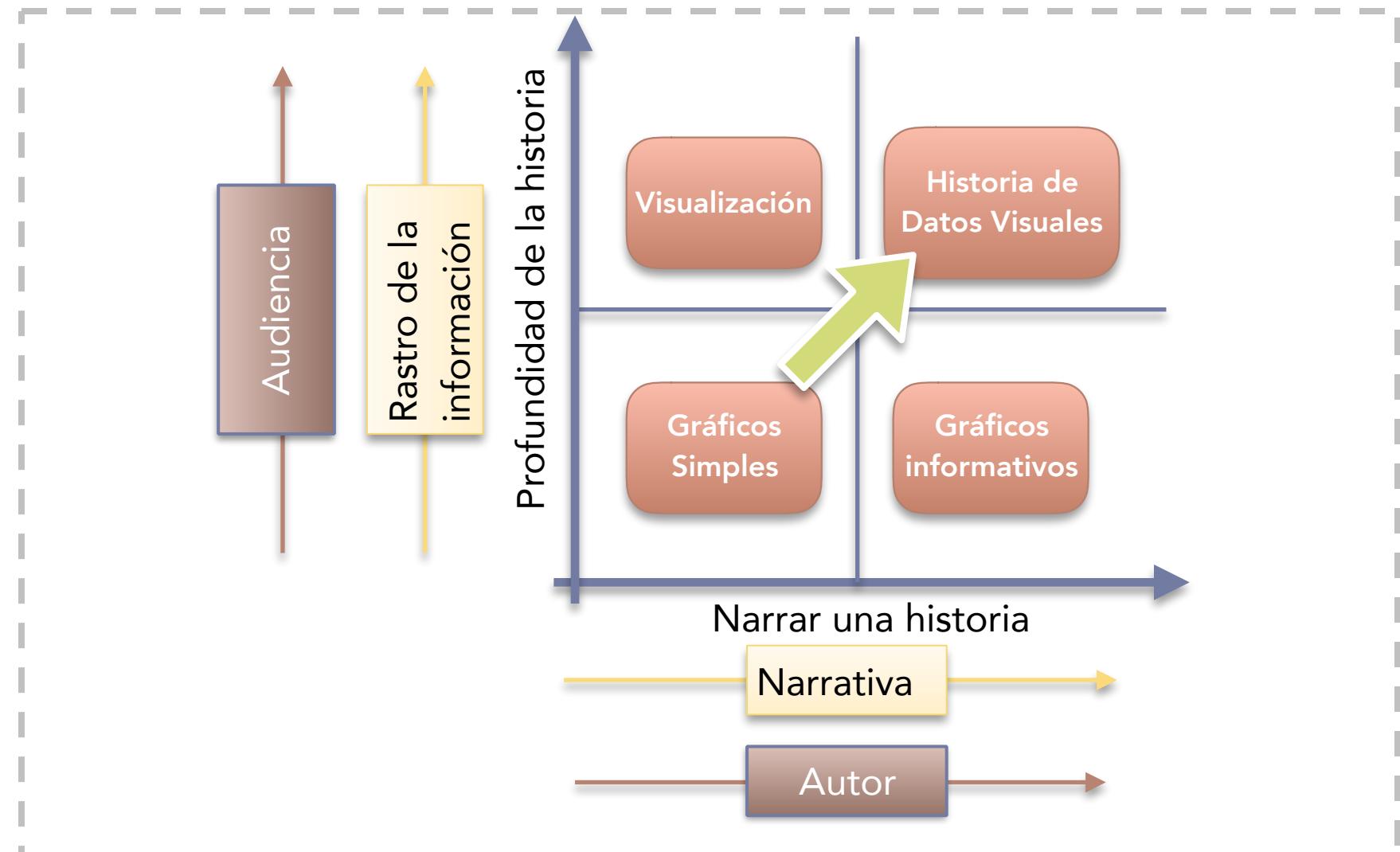
2006 - Present

Dollars in '000s

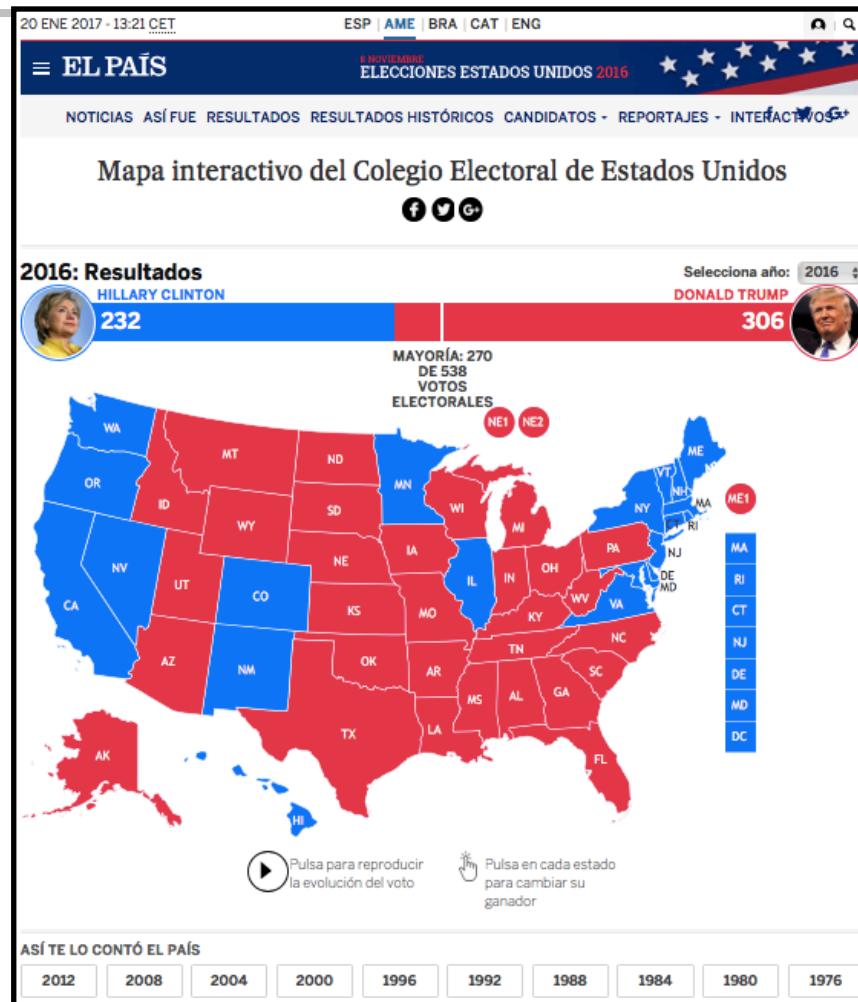
Capacity Building	\$710
Human Services	\$670
Arts & Culture	\$630
Community Betterment	\$620
Housing	\$360
Youth Development	\$340
Health	\$190

Cambios introducidos

- ▶ Reemplazar el título descriptivo por uno activo
- ▶ Agregar percepción con el texto
- ▶ Girar el gráfico
- ▶ Ordenar los datos
- ▶ Etiquetar el eje-x
- ▶ Eliminar sombreados y grillas innecesarias
- ▶ Reducir el ancho
- ▶ Utilizar colores estratégicamente

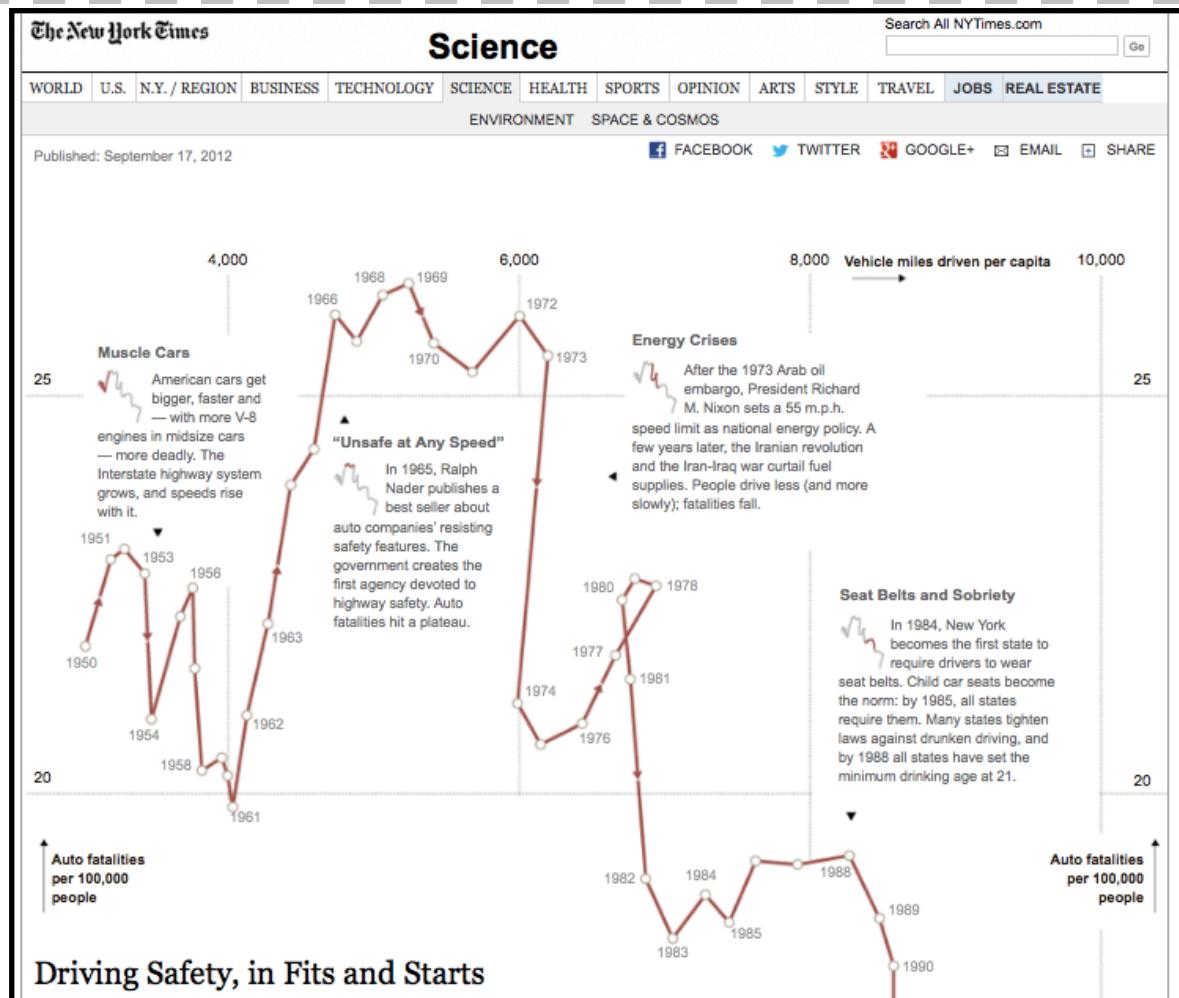


Mapa Interactivo de las Elecciones de Estados Unidos



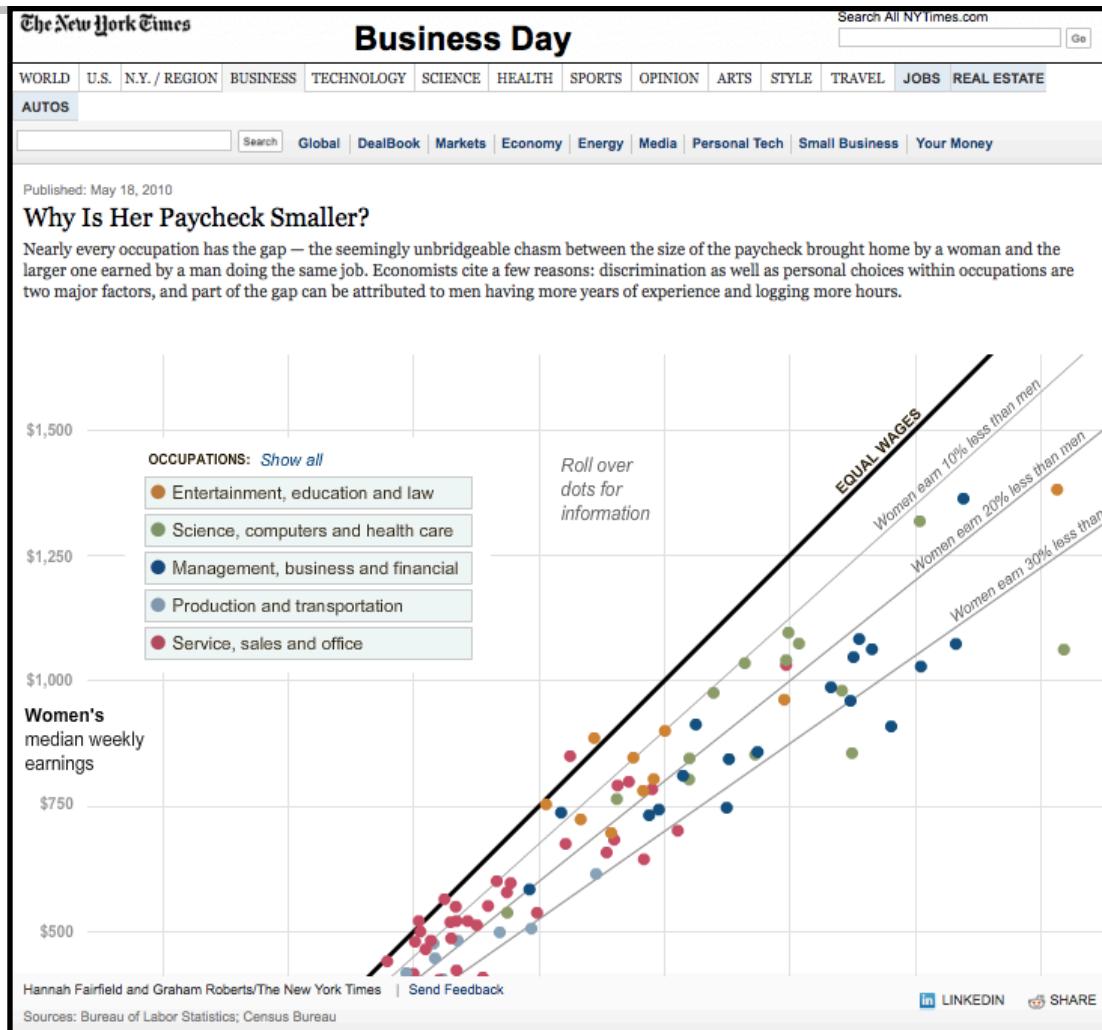
<https://elpais.com/especiales/2016/elecciones-eeuu/mapa-electoral/>

Construyendo una narrativa



<https://archive.nytimes.com/www.nytimes.com/interactive/2012/09/17/science/driving-safety-in-fits-and-starts.html>

Agregando Profundidad



https://archive.nytimes.com/www.nytimes.com/interactive/2009/03/01/business/20090301_WageGap.html?_r=0

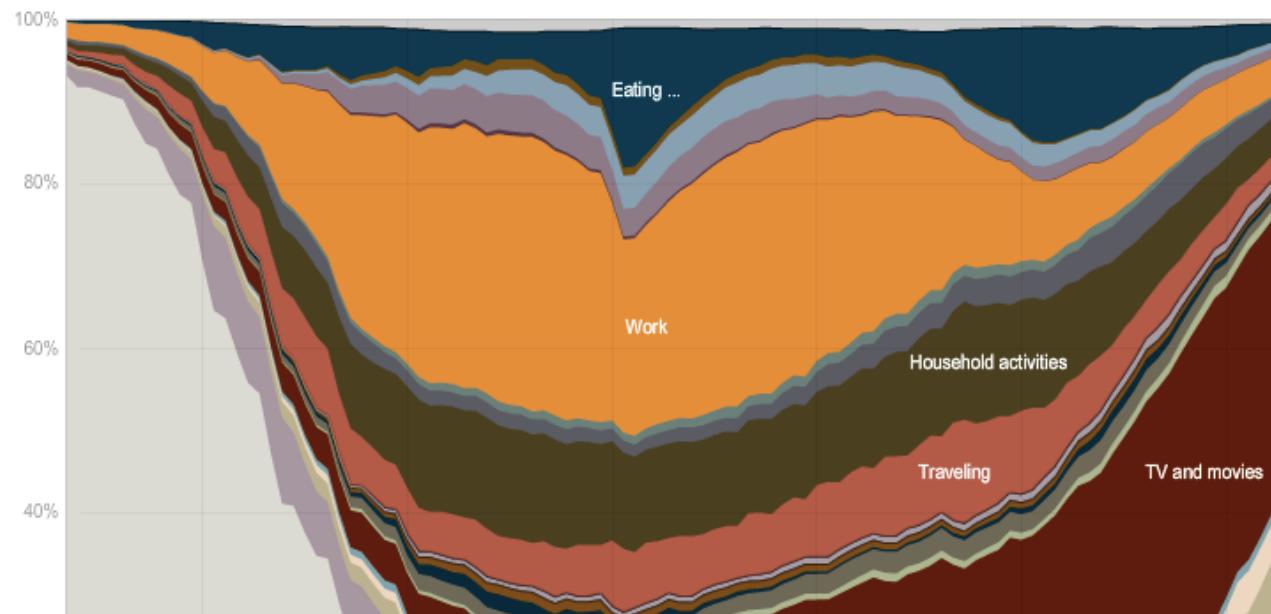
How Different Groups Spend Their Day

The American Time Use Survey asks thousands of American residents to recall every minute of a day. Here is how people over age 15 spent their time in 2008. [Related article](#)

Everyone

Sleeping, eating, working and watching television take up about two-thirds of the average day.

Everyone	Employed	White	Age
Men	Unemployed	Black	Age
Women	Not in lab...	Hispanic	Age



By SHAN CARTER, AMANDA COX, KEVIN QUEALY and AMY SCHOENFELD | [Send Feedback](#)

 [LINKEDIN](#)  [SHARE](#)

<https://archive.nytimes.com/www.nytimes.com/interactive/2009/07/31/business/20080801-metrics-graphic.html>

Resumen

- Understand the context – audience, data, takeaway/outcome
- Choose right display type – text, scatterplots, line charts, bar charts, ...
- Eliminate clutter – use Gestalt principles to cut if no information value
- Draw attention where you want to – preattentive attributes of color, size, ...
- Linear is better for storytelling
- Guide readers through the story –where to start, how to get back, reset
- Limit complexity initially, reveal as needed
- Cool and readability maybe at odds – recognize tradeoffs, tailor to audience

STRUCTURE

- Text good for storytelling
- State the point you want to make – don't leave the reader wondering
- Start with an interesting view
- Put numbers and facts in context
- Connect relevant text and graphics, e.g., see Figure 1
- Add summary/conclusions/"so what?"
- Labels and significant digits suggest what deserves attention

MESSAGING

- Show how the interactivity works, make it intuitive
- Limit interactivity to key elements – too much can distract from story

INTERACTIVITY

Referencias

- ▶ F. Berman, R. Rutenbar, B. Hailpern, et al. (2018) "Realizing the potential of Data Science". Communications of the ACM, vol. 61, no. 4. doi: 10.1145/3188721.
- ▶ E. Segel, J. Heer (2010). "Narrative Visualization: Telling Stories with Data". IEEE Transactions on Visualization and Computer Graphics, vol. 16 , no. 6. pp. 1139-1148, doi: 10.1109/TVCG.2010.179
- ▶ Lloyd-Williams, Michael (1997) "Discovering the hidden secrets in your data - the data mining approach to information" Information Research, 3(2) Available at: <http://informationr.net/ir/3-2/paper36.html>
- ▶ Fayyad, U. & Uthurasamy, R. (1996) "Data Mining and Knowledge Discovery in Databases" Communications of the ACM, 39(11), 24-26.
- ▶ Veena Mendiratta AS. A Storytelling with Data Visualization. ASA Workshop on Data Visualization Techniques.
- ▶
- ▶
- ▶
- ▶
- ▶ III Simposio de Data Analytics — Dr. Ing. Rodrigo Salas Fuentes (rodrigo.salas@uv.cl)