# Theoretical Basics of Big Data Analytics and Real Time Computation Algorithms

*Spring 2016*

- What is **Big Data**?
  - Why is it **new and important**?
  - What **tools** do wee need to deal with it?
- Big collections of data:
  - Retail store inventory and transactions
  - Cell company records
  - Weather records

Databases: store and manage data

- **Digitized data** - routinely collected
  - Shopping transactions
  - Search queries
  - Internet traffic
  - Readings from sensors

  Often tossed away or stored and never used
- Recently: something special
  - New kinds of information
  - Was not anticipated
  - Emerges when BIG
- Valuable **hidden** information
  - not visible
  - has to be extracted,
  - processed
- **Big Data** is about:
  - New information
  - Specific tools

  Such definition is
  - is too vague
  - relies on the notion of *information*

# Information?

"Logic and Information" *Keith Devlin*

- What is *Iron?*
  - Go to Iron age
  - Ask Ironsmith
  - Examples: raw, processing, things
- Result: unsatisfactory
  - No frame of reference
  - Need to know: molecular structure...
- Now we are in **Information Age**
  - Know it exists
  - Have examples
  - Definitions in special cases
- Entering **Big Data (sub) Age**

# Examples of Big Data

"BIG DATA: A Revolution That Will Transform How We Live, Work, and Think"

*Viktor Mayer-Schönberger and Kenneth Cukier*

- **Target** - Detecting: a woman is pregnant
  - Two dozen products used as proxies
  - Estimate pregnancy stages, due date
  - Send relevant coupons
- **Correlation-based techniques**
  - Predict mechanical failures
  - Things break down *gradually*
  - Sensors + correlation analysis:
    - Whirling motor
    - Excessive heat
    - ...

- **UPS**
  - Replacing parts: 2-3years
    - Inefficient
  - Predictive analysis
    - Monitoring individual parts
  - Predictions made automatically
  - Based on:
    - Great number of cases
    - Correlation analysis
    - No complex models

  Modern cars
  - Lots of sensors
    - Temperature, Vibrations, Voltages...
  - Use *complex* models of prediction
  - Information tossed away - no learning
  - Imagine: transmitted, collected, and analyzed...

- **H1N1 Virus** 2009
  - Only hope: to slow its spread
  - Need to know: where it is
  - US Center for Diseases Control (CDC):
    * Doctors: to inform of new flu cases
    * Week or two out of date
    * $\Rightarrow$ Delays blinded health agencies

    Few weeks before H1B1:
    * Google: paper in “Nature”
    * Predict spread if winter flu by looking what people were *searching*
    * 3 Bil. queries a day

- **Google** “learning” technique
  - CDC data for 2003-2008
  - Correlations:
    - Search queries (50M most common)
    - Flu spread
  - Result:
    - Combination of 45 search terms
    - + Math model
    - = Strong correlation with official figures
  - So, in 2009 Google - more timely indicator

    No need in:
    - mouth swabs
    - contacting doctors...

    Instead: huge amount of data
    - Too Big
    - Too Noisy

- **Unexpected** data in existing collections:
  - Too Big
  - Too Noisy
- **Arranging new studies**
  Aspirin and orange juice vs Cancer
  - **Standard** way:
    * Specific tests
    * Time
    * Low confidence (small amount of data)
  - **Big Data** way:
    * Digitized med. records
    * Shopping transactions
    * Search queries
    * ... (lots of other data)

# Other Big Data challenges

- **LHC** Large Hadron Collider
  - **150 Mil.** sensors
  - **40 Mil.** times per second
  - Only **0.001%** saved
  - **25 PB** in 2012 (1PB=1000TB)
  - If all recorded:
    - **150 Mil. PB/year**
    - = **200** x all other sources in the world
- Modern **Aircraft**
  - **100,000** sensors
  - Only **3 GB** in an hour flight

Seems not Big, but

- Monitoring in **real time**
- **Combinations** of readings
- In **dynamics**
- Need to make very fast **predictions**

$\Rightarrow$ Big Data challenge

- **Digitized Media** Streaming
  - Large volumes
  - But: Nothing is hidden

  $\Rightarrow$ Not considered as Big Data

- **Information in Big Data**
  - Hidden
  - Requires special tools

  Analogy:

  - Rare mineral
  - Nuclear fusion energy

# Big Data Manipulations: Basic Steps

- **Extract** pieces of information (probably from distributed sources)
- **Unify** - transform to "canonical" form
  - Compact
  - Easy to handle
  - Contains sufficient information
- **Combine** pieces
- **Update** when new info arrives
- **Utilize** - Decision making

# Simplest Example

$x$ - object of interest (unknown value)
Observations:

$$y_i = x + \varepsilon_i, \qquad i = 1, \ldots, n$$

$\varepsilon_i$ - i.i.d. random values, $\mathrm{E}\varepsilon_i = 0$.
A good estimate of $x$:

$$\hat{x} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$y_1 \longrightarrow \hat{x}_1 = y_1$$

$$y_1, y_2 \longrightarrow \hat{x}_2 = \frac{y_1 + y_2}{2}$$

$$\vdots$$

$$y_1, \ldots, y_n \longrightarrow \hat{x}_n = \frac{y_1 + \cdots + y_n}{n}$$

# *Updating $\widehat{x}$*

$n$ : have $\widehat{x}_n$, get $y_{n+1}$

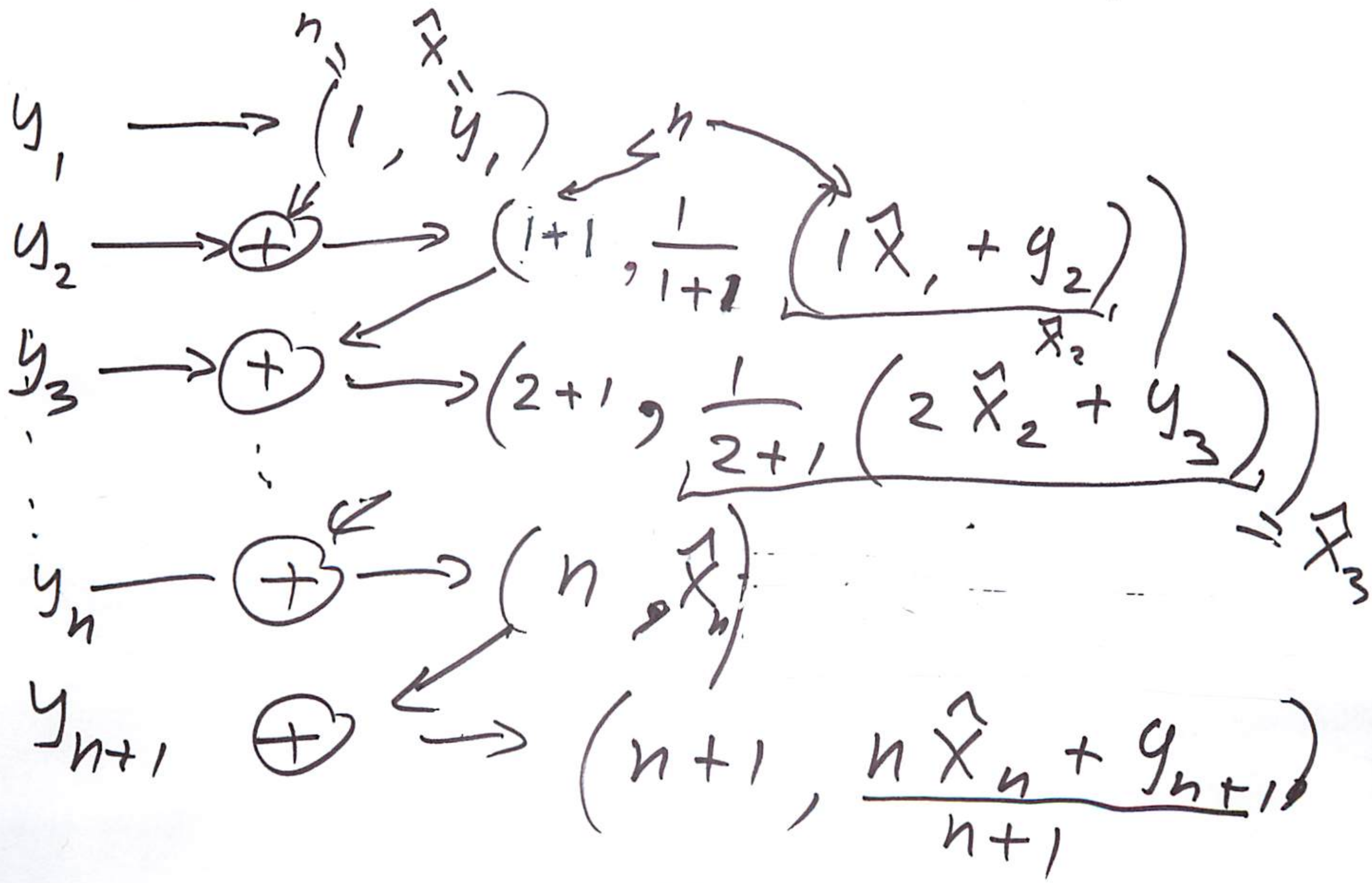
$$\widehat{x}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} y_i = \frac{1}{n+1}\left(\underbrace{\sum_{i=1}^{n} y_i}_{= n\hat{x}_n} + y_{n+1}\right)$$

$$= \frac{1}{n+1}\left(n\widehat{x}_n + y_{n+1}\right)$$

or

$$\widehat{x}_{n+1} = \widehat{x}_n \quad + \frac{1}{n+1}\left(y_{n+1} - \widehat{x}_n\right)$$

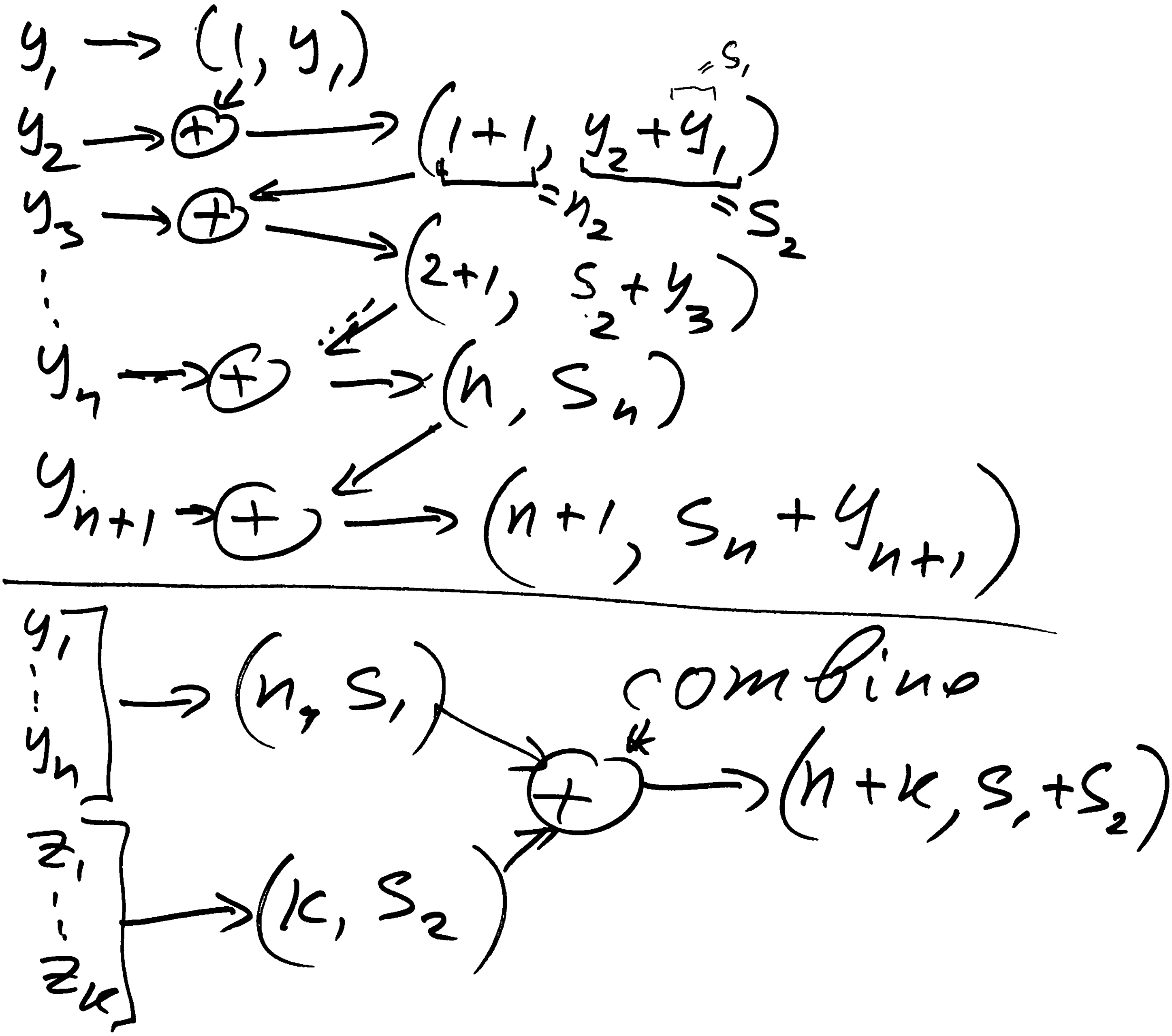
In addition to $\widehat{x}_n$ need to keep $n$.
"Explicit" form of information: $(n, \widehat{x}_n)$.



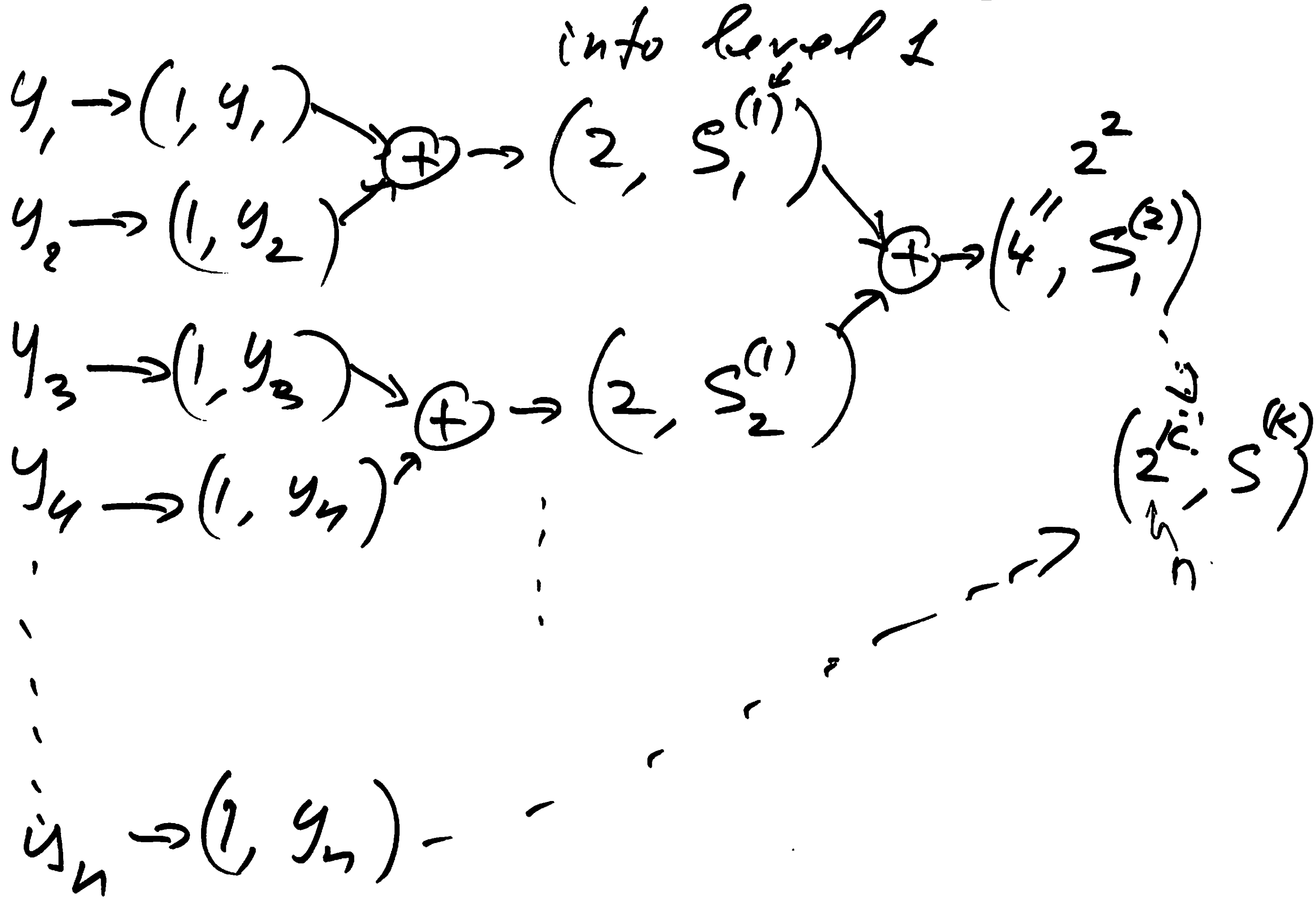
# *Updating "Canonical" Information* $(n, S)$

$n$ - number of readings, $S = \sum_{i=1}^{n} y_i$

$(n, S) \Rightarrow \hat{X} = S/n$



# Concurrent Combining




$n = 2^k \qquad k = \log_2 n$

if $n = 1000 \qquad k = 10$

$n = 1M \qquad k = 20$

$n = 1B \qquad k = 30$

# *Precision of $\widehat{x}$*

$$y_i = x + \varepsilon_i$$

$$\varepsilon_i \text{ - i.i.d.}, \quad \mathsf{E}\varepsilon_i = 0$$

$$\mathrm{Var}(\varepsilon_i) = \mathsf{E}(\varepsilon_i - \mathsf{E}\varepsilon_i)^2 = \mathsf{E}\varepsilon_i^2 = \underline{\underline{\sigma^2}}$$

$\widehat{x} = \frac{1}{n}\sum_{i=1}^n y_i$ - **unbiased** est. of $x$, i.e. $\mathsf{E}\widehat{x} = x$

$$\mathsf{E}\widehat{x} = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^n y_i\right) = \frac{1}{n}\mathsf{E}\sum_{i=1}^n (x + \varepsilon_i)$$

$$= \frac{1}{n}\sum_{i=1}^n x + \frac{1}{n}\underbrace{\mathsf{E}\sum_{i=1}^n \varepsilon_i}_{=\sum \underbrace{\mathsf{E}\varepsilon_i}_{=0}} = x$$

$$\mathrm{Var}(\widehat{x}) = \mathsf{E}(\widehat{x} - x)^2 = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^n (x + \varepsilon_i) - \underbrace{x}_{=\frac{1}{n}\sum x}\right)^2$$

$$= \mathsf{E}\frac{1}{n^2}\underbrace{\left(\sum_{i=1}^n (\cancel{x} + \varepsilon_i - \cancel{x})\right)^2}_{=\left(\sum \varepsilon_i\right)^2 = \sum_i \varepsilon_i \sum_j \varepsilon_j} = \frac{1}{n^2}\sum_{i,j=1}^n \mathsf{E}\varepsilon_i\varepsilon_j$$

$$= \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \underline{\underline{\frac{\sigma^2}{n}}}$$

$\mathsf{E}\varepsilon_i\varepsilon_j = 0$ if $i \neq j$

$\underline{\underline{\text{ind}}} \Rightarrow \mathsf{E}\varepsilon_i\varepsilon_j = \underbrace{\mathsf{E}\varepsilon_i}_{=0} \cdot \underbrace{\mathsf{E}\varepsilon_j}_{=0}$

$$\mathrm{Var}(\widehat{x}) = \frac{\sigma^2}{n} \to 0 \;\text{ as }\; n \to \infty.$$

If we collect canonical info $(n, S)$:

$$(n, S) \quad \Rightarrow \quad \widehat{x} = \frac{S}{n}, \quad \mathrm{Var}(\widehat{x}) = \frac{\sigma^2}{n}.$$

$(n, S)$ is sufficient to obtain $\widehat{x}$ and its variance, but only when $\sigma^2$ is known.

## *Suppose* $\sigma^2$ *is* **not** *known*

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \widehat{x})^2$$

- Unbiased estimate of $\sigma^2$.

$\hat{x} = S/n$

$$\sum_{i=1}^{n} (y_i - \widehat{x})^2 = \overbrace{\sum_{i=1}^{n} y_i^2}^{T} - 2 \overbrace{\sum_{i=1}^{n} y_i}^{=S} \cdot \widehat{x} + n\widehat{x}^2$$

$$= T - 2S\frac{S}{n} + n\left(\frac{S}{n}\right)^2 = T - \frac{S^2}{n}$$

$$T = \sum_{i=1}^{n} y_i^2$$

**New** *canonical information* $(n, S, T)$ :

$$n = \sum_{i=1}^{n} y_i^0, \qquad S = \sum_{i=1}^{n} y_i^1, \qquad T = \sum_{i=1}^{n} y_i^2$$

$$\widehat{\sigma^2} = \frac{1}{n-1}\left(T - \frac{S^2}{n}\right)$$

Estimate of the variance of $\widehat{x}$:

$$V = \widehat{\mathrm{Var}(\widehat{x})} = \frac{\widehat{\sigma^2}}{n} = \frac{1}{n(n-1)}\left(T - \frac{S^2}{n}\right)$$

$$(n, S, T) \quad \Rightarrow \quad \widehat{x} = \frac{S}{n}, \quad V = \frac{1}{n(n-1)}\left(T - \frac{S^2}{n}\right)$$

**Updating** can. info:

$$(n, S, T) \longrightarrow \oplus \longrightarrow (n+1, S+y, T+y^2)$$
$$y \nearrow$$

**Combining** can. info:

$$\begin{array}{l} (n_1, S_1, T_1) \searrow \\ \qquad\qquad\qquad \oplus \rightarrow (n_1+n_2, S_1+S_2, T_1+T_2) \\ (n_2, S_2, T_2) \nearrow \end{array}$$

## *Info in* **explicit** *form* $(n, \underline{\hat{x}}, \underline{V})$ :

Have $(n, \hat{x}_n, V_n)$, receive $y_{n+1}$

$$\hat{x}_{n+1} = \hat{x}_n + \frac{y_{n+1} - \hat{x}_n}{n+1}$$

$$\widehat{\sigma^2}_{n+1} = \underbrace{\frac{n-1}{n}\widehat{\sigma^2}_n}_{= V_n} + \frac{(y_{n+1} - \hat{x}_n)^2}{n+1}$$

$$V_{n+1} = \frac{\widehat{\sigma^2}_{n+1}}{n+1} = \underbrace{\frac{n-1}{n+1}V_n + \left(\frac{y_{n+1} - \hat{x}_n}{n+1}\right)^2}$$

**Updating** explicit info:

$$\begin{matrix} (n, \hat{x}, V) \searrow \\ y \nearrow \end{matrix} \oplus \rightarrow \left(n+1,\ \hat{x}_n + \frac{y_{n+1} - \hat{x}_n}{n+1},\ \ \right)$$

Explicit form for single observation:

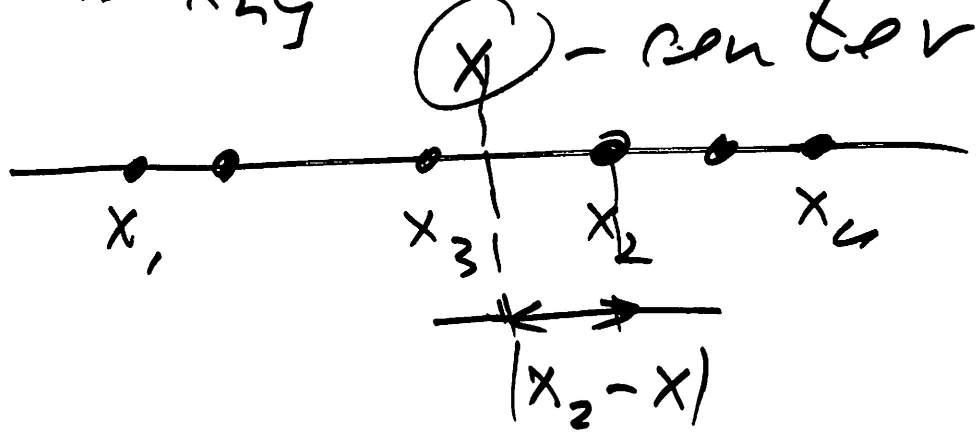
$$y \to (1, y, ?)$$

$$V = \underbrace{\frac{1}{n(n-1)}}_{=0} \sum_{i=1}^{n} (y_i - \hat{x})^2$$

$\Rightarrow$ Information in explicit form *may not exist*

# Center of a set of points

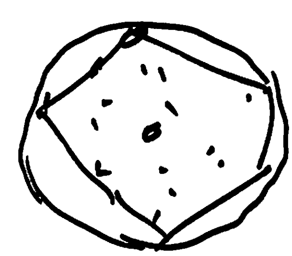
$\{x_1, \dots x_n\}$



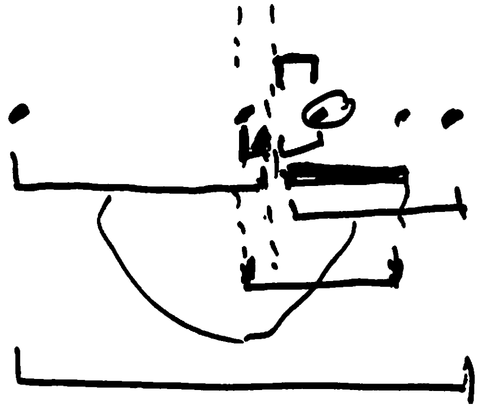

a) $\sum_{i=1}^{n} |x - x_i| \sim \min_{x}$

в) $\sum_{i=1}^{n} (x - x_i)^2 \sim \min_{x}$

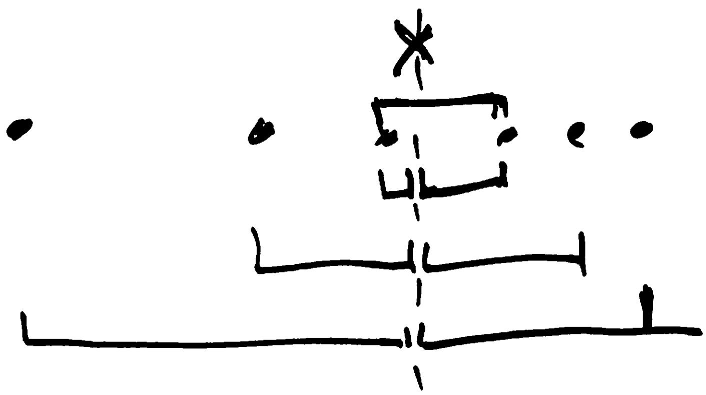
c) $\max_{i=1,..n} |x - x_i| \sim \min_{x}$



a.)



median (odd)



even
not unique

B) $$\sum_{i=1}^{n}(x-x_i)^2 = \sum_{i=1}^{n}\left((x-\bar{x})+(\bar{x}-x_i)\right)^2 =$$

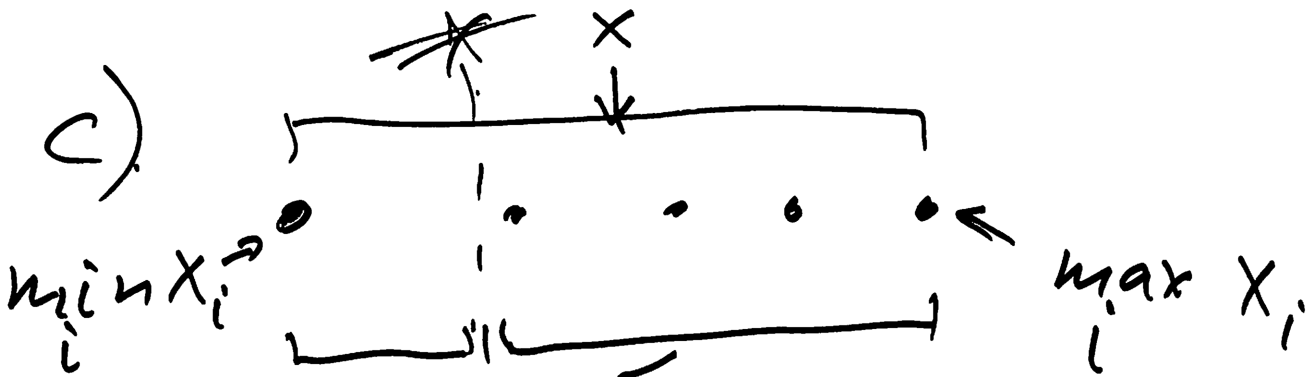
$$= \underline{n(x-\bar{x})^2} + 2\underbrace{\sum_{i=1}^{n}(x-\bar{x})(\bar{x}-x_i)}_{=0}$$

$$+ \underbrace{\sum_{i=1}^{n}(\bar{x}-x_i)^2}_{=\text{const}}$$

$$= n\underbrace{(x-\bar{x})^2}_{\text{min if } x=\bar{x}} + \sum_{i=1}^{n}(\bar{x}-x_i)^2$$

$\Rightarrow$ center is **mean**.

c)




$$\min_{x} \max_{i} |x - X_i|$$

$$x = \frac{1}{2}\left(\min_i X_i + \max_i X_i\right)$$

Call it "middle"