# Swimming in a sea of data

Or

How I learned to stop worrying and love the data

Or

What is the difference between ‘data science’ and ‘data assimilation’

with sincere apologies for so many cliched phrases - I promise not to continue the trend...

Amit Apte
International Centre for Theoretical Sciences, Tata Institute of Fundamental Research (ICTS-TIFR), Bangalore, India

Summer School for Women in Mathematics and Statistics, ICTS-TIFR
22 May 2019

# Let us start with some data - from IC ‘Theoretical’ S!

## A weather station at ICTS

Quiz: Where on the ICTS campus is this located? (Hint: it is accessible only to authorised persons.)

Measurements taken: temperature, humidity, pressure, wind (speed and direction), rain, ... every 15 minutes, since March-2019 (not a very big dataset)

# Let us start with some data - from IC ‘Theoretical’ S!

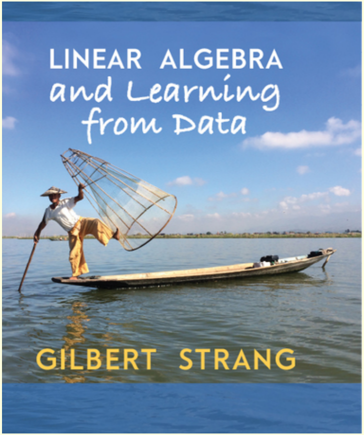
## A weather station at ICTS

So what do these “data” look like?

We will just plot a few things...

# We can plot data, but now what?

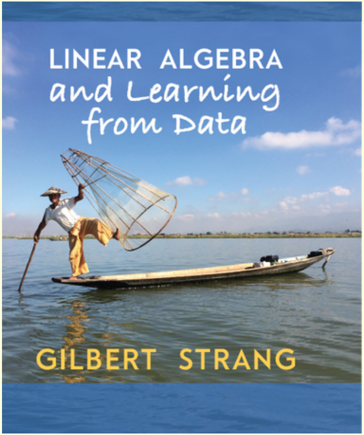
A little bit linear algebra helps to go a long way...



Recall: Measurements taken: temperature T, humidity H, pressure P, wind (speed and direction) V, rain R, ... every 15 minutes

# We can plot data, but now what?

A little bit linear algebra helps to go a long way...

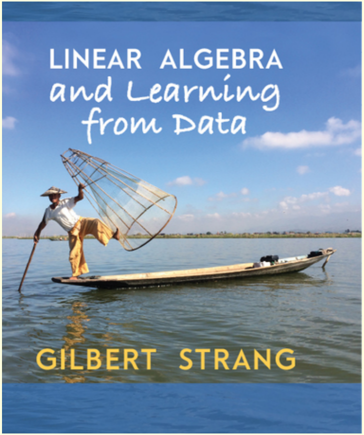


Recall: Measurements taken: temperature T, humidity H, pressure P, wind (speed and direction) V, rain R, ... every 15 minutes

$$\begin{array}{l}
T_1,\ T_2,\ T_3, \dots, T_d \\
H_1,\ H_2,\ H_3, \dots, H_d \\
P_1,\ P_2,\ P_3, \dots, P_d \\
V_1,\ V_2,\ V_3, \dots, V_d \\
R_1,\ R_2,\ R_3, \dots, R_d
\end{array}$$

# We can plot data, but now what?

A little bit linear algebra helps to go a long way...



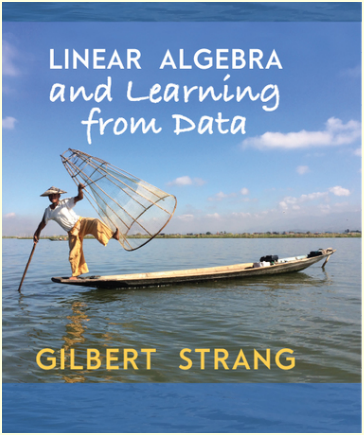
Recall: Measurements taken: temperature T, humidity H, pressure P, wind (speed and direction) V, rain R, ... every 15 minutes

Form a matrix $A =$

$$\begin{matrix} T_1, & T_2, & T_3, \ldots, T_d \\ H_1, & H_2, & H_3, \ldots, H_d \\ P_1, & P_2, & P_3, \ldots, P_d \\ V_1, & V_2, & V_3, \ldots, V_d \\ R_1, & R_2, & R_3, \ldots, R_d \end{matrix}$$

# We can plot data, but now what?

A little bit linear algebra helps to go a long way...



Recall: Measurements taken: temperature T, humidity H, pressure P, wind (speed and direction) V, rain R, ... every 15 minutes

Form a matrix $A =$

$$\begin{array}{l} T_1, \; T_2, \; T_3, \dots, T_d \\ H_1, \; H_2, \; H_3, \dots, H_d \\ P_1, \; P_2, \; P_3, \dots, P_d \\ V_1, \; V_2, \; V_3, \dots, V_d \\ R_1, \; R_2, \; R_3, \dots, R_d \end{array}$$

- each column: measurements at one time instant
- each row: one quantity (temperature, pressure, etc.) at all times

# We all live in a matrix

I am breaking my promise about cliche....

Recall the $n \times d$ "data matrix" $A =$

$T_1,\ T_2,\ T_3, \dots, T_d$
$H_1,\ H_2,\ H_3, \dots, H_d$
$P_1,\ P_2,\ P_3, \dots, P_d$
$V_1,\ V_2,\ V_3, \dots, V_d$
$R_1,\ R_2,\ R_3, \dots, R_d$

- each column: measurements at one time instant
- each row: one quantity (temperature, pressure, etc.) at all times
- Number of rows $n =$ number of variables measured
- Number of columns $d =$ number of times the measurements were taken

What do we do with this matrix?

# We all live in a matrix

I am breaking my promise about cliche....

Recall the $n \times d$ "data matrix" $A =$

$T_1,\ T_2,\ T_3, \ldots, T_d$
$H_1,\ H_2,\ H_3, \ldots, H_d$
$P_1,\ P_2,\ P_3, \ldots, P_d$
$V_1,\ V_2,\ V_3, \ldots, V_d$
$R_1,\ R_2,\ R_3, \ldots, R_d$

- each column: measurements at one time instant
- each row: one quantity (temperature, pressure, etc.) at all times
- Number of rows $n =$ number of variables measured
- Number of columns $d =$ number of times the measurements were taken

What do we do with this matrix?

Recall our notes: a change in temperature is somehow related to a change in humidity.

Can we quantify this using linear algebra?

# A little bit of linear algebra

keyword: 'Singular value decomposition'

- (column space) When $A$ is $n \times d$ matrix with columns $A = [a_1, a_2, \ldots, a_d]$, and $v = [v_1, v_2, \ldots, v_d]^t \in \mathbb{R}^d$ is an $d$-dimensional vector, then

**the vector $u = Av$ is a linear combination of columns of $A$:**

$$u = v_1 a_1 + v_2 a_2 + \cdots + v_d a_d \in \mathbb{R}^n$$

Q: what linear combinations of the data vectors $Av$ are "important"?

# A little bit of linear algebra

keyword: 'Singular value decomposition'

- (column space) When $A$ is $n \times d$ matrix with columns $A = [a_1, a_2, \dots, a_d]$, and $v = [v_1, v_2, \dots, v_d]^t \in \mathbb{R}^d$ is an $d$-dimensional vector, then

  **the vector $u = Av$ is a linear combination of columns of $A$:**

  $$u = v_1 a_1 + v_2 a_2 + \cdots + v_d a_d \in \mathbb{R}^n$$

- The equation $Ax = \lambda x$ does not make sense! (eigenvalues are not useful for data matrices.)

Q: what linear combinations of the data vectors $Av$ are "important"?

# A little bit of linear algebra

keyword: 'Singular value decomposition'

- (column space) When $A$ is $n \times d$ matrix with columns $A = [a_1, a_2, \ldots, a_d]$, and $v = [v_1, v_2, \ldots, v_d]^t \in \mathbb{R}^d$ is an $d$-dimensional vector, then

  **the vector $u = Av$ is a linear combination of columns of $A$:**

  $$u = v_1 a_1 + v_2 a_2 + \cdots + v_d a_d \in \mathbb{R}^n$$

- The equation $Ax = \lambda x$ does not make sense! (eigenvalues are not useful for data matrices.)
- But the equation $Av = \sigma u$ can make sense for $u \in$ column space!

Q: what linear combinations of the data vectors $Av$ are "important"?

# A little bit of linear algebra

## keyword: 'Singular value decomposition'

- (column space) When $A$ is $n \times d$ matrix with columns $A = [a_1, a_2, \ldots, a_d]$, and $v = [v_1, v_2, \ldots, v_d]^t \in \mathbb{R}^d$ is an $d$-dimensional vector, then

  **the vector $u = Av$ is a linear combination of columns of $A$:**

  $$u = v_1 a_1 + v_2 a_2 + \cdots + v_d a_d \in \mathbb{R}^n$$

- The equation $Ax = \lambda x$ does not make sense! (eigenvalues are not useful for data matrices.)
- But the equation $Av = \sigma u$ can make sense for $u \in$ column space!
- That introduces new concepts (to be defined precisely very soon):
  - $v$ as the right singular vectors and $u$ as the left singular vectors, replacing the eigenvectors
  - $\sigma$ as singular values replacing the eigenvalues

Q: what linear combinations of the data vectors $Av$ are "important"?

# Singular value decomposition is a calculus problem

- A calculus problem in linear algebra: maximize the ratio $\|Ax\|/\|x\|$ where $x \in \mathbb{R}^d$.

# Singular value decomposition is a calculus problem

- A calculus problem in linear algebra: maximize the ratio $\|Ax\|/\|x\|$ where $x \in \mathbb{R}^d$.
- Denote the maximum ratio by $\sigma_1$ and the vector that maximizes it as $v_1$, and denote $u_1 = Av_1/\sigma_1$.

# Singular value decomposition is a calculus problem

- A calculus problem in linear algebra: maximize the ratio $\|Ax\|/\|x\|$ where $x \in \mathbb{R}^d$.
- Denote the maximum ratio by $\sigma_1$ and the vector that maximizes it as $v_1$, and denote $u_1 = Av_1/\sigma_1$.
- Now, a problem combining linear algebra and calculus: maximize the ratio $\|Ax\|/\|x\|$ but for those $x \in \mathbb{R}^d$ which are perpendicular to $v_1$.

# Singular value decomposition is a calculus problem

- A calculus problem in linear algebra: maximize the ratio $\|Ax\|/\|x\|$ where $x \in \mathbb{R}^d$.
- Denote the maximum ratio by $\sigma_1$ and the vector that maximizes it as $v_1$, and denote $u_1 = Av_1/\sigma_1$.
- Now, a problem combining linear algebra and calculus: maximize the ratio $\|Ax\|/\|x\|$ but for those $x \in \mathbb{R}^d$ which are perpendicular to $v_1$.
- Denote the maximum ratio by

# Singular value decomposition is a calculus problem

- A calculus problem in linear algebra: maximize the ratio $\|Ax\|/\|x\|$ where $x \in \mathbb{R}^d$.
- Denote the maximum ratio by $\sigma_1$ and the vector that maximizes it as $v_1$, and denote $u_1 = Av_1/\sigma_1$.
- Now, a problem combining linear algebra and calculus: maximize the ratio $\|Ax\|/\|x\|$ but for those $x \in \mathbb{R}^d$ which are perpendicular to $v_1$.
- Denote the maximum ratio by $\sigma_2$ and the vector that maximizes it as $v_2$, and denote $u_2 = Av_2/\sigma_2$.
- ...

# Singular value decomposition is a calculus problem

- A calculus problem in linear algebra: maximize the ratio $\|Ax\|/\|x\|$ where $x \in \mathbb{R}^d$.
- Denote the maximum ratio by $\sigma_1$ and the vector that maximizes it as $v_1$, and denote $u_1 = Av_1/\sigma_1$.
- Now, a problem combining linear algebra and calculus: maximize the ratio $\|Ax\|/\|x\|$ but for those $x \in \mathbb{R}^d$ which are perpendicular to $v_1$.
- Denote the maximum ratio by $\sigma_2$ and the vector that maximizes it as $v_2$, and denote $u_2 = Av_2/\sigma_2$.
- ...
- maximize the ratio $\|Ax\|/\|x\|$ but for those $x \in \mathbb{R}^d$ which are perpendicular to $v_1, \ldots, v_{k-1}$ (prove: they form a subspace!) and denote the answer by $\sigma_k$ and $v_k$ and $u_k$.

# Singular value decomposition is a calculus problem

- A calculus problem in linear algebra: maximize the ratio $\|Ax\|/\|x\|$ where $x \in \mathbb{R}^d$.
- Denote the maximum ratio by $\sigma_1$ and the vector that maximizes it as $v_1$, and denote $u_1 = Av_1/\sigma_1$.
- Now, a problem combining linear algebra and calculus: maximize the ratio $\|Ax\|/\|x\|$ but for those $x \in \mathbb{R}^d$ which are perpendicular to $v_1$.
- Denote the maximum ratio by $\sigma_2$ and the vector that maximizes it as $v_2$, and denote $u_2 = Av_2/\sigma_2$.
- ...
- maximize the ratio $\|Ax\|/\|x\|$ but for those $x \in \mathbb{R}^d$ which are perpendicular to $v_1, \dots, v_{k-1}$ (prove: they form a subspace!) and denote the answer by $\sigma_k$ and $v_k$ and $u_k$.

Forming matrices $V = [v_1, v_2, ...]$ and $U = [u_1, u_2, ...]$ of all these vectors, we arrive at ...

# Singular value decomposition is a calculus problem

- A calculus problem in linear algebra: maximize the ratio $\|Ax\|/\|x\|$ where $x \in \mathbb{R}^d$.
- Denote the maximum ratio by $\sigma_1$ and the vector that maximizes it as $v_1$, and denote $u_1 = Av_1/\sigma_1$.
- Now, a problem combining linear algebra and calculus: maximize the ratio $\|Ax\|/\|x\|$ but for those $x \in \mathbb{R}^d$ which are perpendicular to $v_1$.
- Denote the maximum ratio by $\sigma_2$ and the vector that maximizes it as $v_2$, and denote $u_2 = Av_2/\sigma_2$.
- ...
- maximize the ratio $\|Ax\|/\|x\|$ but for those $x \in \mathbb{R}^d$ which are perpendicular to $v_1, \ldots, v_{k-1}$ (prove: they form a subspace!) and denote the answer by $\sigma_k$ and $v_k$ and $u_k$.

Forming matrices $V = [v_1, v_2, \ldots]$ and $U = [u_1, u_2, \ldots]$ of all these vectors, we arrive at ... (next slide)

# Matrix form of singular value decomposition

- A summary of previous slide (with a bit new terminology): Maximizing $\|Ax\|/\|x\|$ in successively "smaller" sub-spaces leads to three matrices:
  right singular vectors $V = [v_1, v_2, \dots]$
  left singular vectors $U = [u_1, u_2, \dots]$
  singular value $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \dots)$

# Matrix form of singular value decomposition

- A summary of previous slide (with a bit new terminology): Maximizing $\|Ax\|/\|x\|$ in successively "smaller" sub-spaces leads to three matrices:
  right singular vectors $V = [v_1, v_2, \dots]$
  left singular vectors $U = [u_1, u_2, \dots]$
  singular value $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \dots)$
- which satisfy that
  - $V$ is "orthogonal," i.e., $VV^t = I$ (homework: find dimension of the identity matrix)
  - and so is $U$, i.e., $U^t U = I$ (homework: find dimension of this identity matrix too!) and
  - the *sigma* values are positive and ordered $sigma_1 \geq \sigma_2 \dots ge0$ (homework: prove this!)

# Matrix form of singular value decomposition

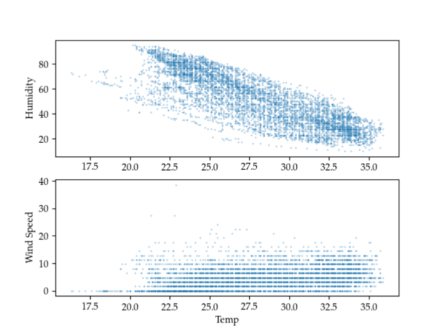
- A summary of previous slide (with a bit new terminology): Maximizing $\|Ax\|/\|x\|$ in successively "smaller" sub-spaces leads to three matrices:
  right singular vectors $V = [v_1, v_2, \dots]$
  left singular vectors $U = [u_1, u_2, \dots]$
  singular value $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \dots)$
- which satisfy that
  - $V$ is "orthogonal," i.e., $VV^t = I$ (homework: find dimension of the identity matrix)
  - and so is $U$, i.e., $U^tU = I$ (homework: find dimension of this identity matrix too!) and
  - the *sigma* values are positive and ordered $sigma_1 \geq \sigma_2 \dots ge0$ (homework: prove this!)
- and finally we get

### Singular value decomposition of $A$

$AV = U\Sigma$ which is equivalent to $A = U\Sigma V^t$.

# What do all these long calculations buy us?

Recall our notes: a change in temperature is somehow related to a change in humidity.



The SVD finds the line closest to the data points: The direction of the first singular vector $u_1$ is the direction of such a line!

## Eckart-Young theorem

If $B$ has rank $k$ then $\|A - A_k\| \leq \|A - B\|$ where $A_k = \sigma_1 u_1 v_1^t + \dots \sigma_k u_k v_k^t$.

# What is data science?

## Some opinions

[1]

# What is data science?

## Some opinions

- A redundant[1] expression

  like ‘past history’ or ‘Unexpected surprise’ or any of the 100s of phrases such as

  `https://en.oxforddictionaries.com/writing-help/avoiding-redundant-expressions`

[1] Science without data is bad science fiction, and use of data without science (i.e. without logically / mathematically consistent methods) is a swindle.

# What is data science?

## Some opinions

- A redundant[1] expression

  like 'past history' or 'Unexpected surprise' or any of the 100s of phrases such as

  `https://en.oxforddictionaries.com/writing-help/avoiding-redundant-expressions`

- A new name for an old subject called 'statistics'

  from wikipedia:

  - "Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data..."
  - "Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data..."

[1] Science without data is bad science fiction, and use of data without science (i.e. without logically / mathematically consistent methods) is a swindle.

# Some important ideas I did not talk about:

- Statistics and
- probability theory, which is required because

# Some important ideas I did not talk about:

- Statistics and
- probability theory, which is required because
- "observations" / "data" always contain errors, and

# Some important ideas I did not talk about:

- Statistics and
- probability theory, which is required because
- "observations" / "data" always contain errors, and
- the mathematical description of errors is in terms of statistics / probabilities of those errors

# Some important ideas I did not talk about:

- Statistics and
- probability theory, which is required because
- "observations" / "data" always contain errors, and
- the mathematical description of errors is in terms of statistics / probabilities of those errors
- The interplay between these "deterministic" (minimization, singular values, ...) ideas and "probabilistic ideas" (statistics, random variables, distributions, ...) is crucial for interpreting observations and data for scientific discoveries....

# Some important ideas I did not talk about:

- Statistics and
- probability theory, which is required because
- "observations" / "data" always contain errors, and
- the mathematical description of errors is in terms of statistics / probabilities of those errors
- The interplay between these "deterministic" (minimization, singular values, ...) ideas and "probabilistic ideas" (statistics, random variables, distributions, ...) is crucial for interpreting observations and data for scientific discoveries....

Finally: what do I work on? I did not talk about the "time evolution of the data"!! that leads me to data assimilation (not data science!)

# A few questions that data can help us answer!

- When will be the next total solar eclipse visible from Bangalore?
- What will be the closest approach of Halley's comet in next 200 years?
- How many times in the next minute will a double pendulum reach the lowest point? What will be the angle of a double pendulum after 5 min., 10 min., ...?
- What will be the total and regional monsoon rainfall in India 2019?
- When and how strong will be the next El Niño?
- What will be the extent of the Arctic sea-ice over next 50 years?
- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin tosses (Simpler?)?
- How many cars will pass in front of Eiffel tower in next 5 minutes?
- Who will be the prime minister in India 2029?
- How many quarks will be detected in a certain collision at the Large Hadron Colloider near Geneva?
- and many others ...

# A few questions that data can help us answer!

- When will be th e visible from Bangalore?

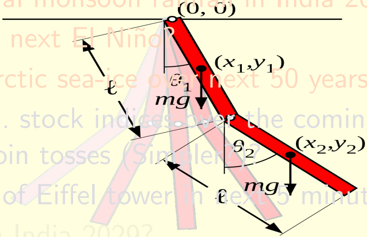
**deterministic, periodic, predictable**

- e closest approach of Halley's comet in next 200 years?
- How many times in the next minute will a double pendulum reach the lowest point? What will be the angle of a double pendulum after 5 min., 10 min., ...?
- What will be the total and regional monsoon rainfall in India 2019?
- When and how strong will be the next El Niño?
- What will be the extent of the Arctic sea-ice over next 50 years?
- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin tosses (Simpler?)?
- How many cars will pass in front of Eiffel tower in next 5 minutes?
- Who will be the prime minister in India 2029?
- How many quarks will be detected in a certain collision at the Large Hadron Colloider near Geneva?
- and many others ...

# A few questions that data can help us answer!

**deterministic, periodic, predictable**

- When will be th e visible from Bangalore?
- closest approach of Halley's comet in next 200 years?
- How many times in the next minute will a double pendulum reach the lowest point? What will be the angle of a double pendulum after 5 min., 10 min., ...?
- What will be the total and regional monsoon rainfall in India 2019?
- When and how strong will be the next El Niño?
- What will be the extent of the Arctic sea-ice over next 50 years?
- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin tosses (Simpler?)?
- How many cars will pass in front of Eiffel tower in next 5 minutes?
- Who will be the prime minister in India 2029?
- How many quarks will be detected in a certain collision at the Large Hadron Colloider near Geneva?
- and many others ...



Videos from https://youtu.be/wtnA6ouIu0U

# A few questions that data can help us answer!

- When will be th e visible from Bangalore?
  closest approach of Halley's comet in next 200 years?

**deterministic, periodic, predictable**

- How many times in th l a double pendulum reach the
  the angle of a double pendulum after 5 min., 10 min., ...?

**deterministic, chaotic, unpredictable**

- What will be the total and regional monsoon rainfall in India 2019?
- When and how strong will be the next El Niño?
- What will be the extent of the Arctic sea-ice over next 50 years?
- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin tosses (Simpler?)?
- How many cars will pass in front of Eiffel tower in next 5 minutes?
- Who will be the prime minister in India 2029?
- How many quarks will be detected in a certain collision at the Large Hadron Colloider near Geneva?
- and many others ...

# A few questions that data can help us answer!

- When will be th e visible from Bangalore?
  closest approach of Halley's comet in next 200 years?

**deterministic, periodic, predictable**

- How many times in th l a double pendulum reach the
  e the angle of a double pendulum after 5 min., 10 min., ...?

**deterministic, chaotic, unpredictable**

- What will be the total and regional monsoon rainfall in India 2019?
- When and how strong will be the next El Niño?
- What will be the extent of the Arctic sea-ice over next 50 years?
- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin tosses (Simpler?)?
- How many cars will pass in front of Eiffel tower in next 5 minutes?
- Who will be the prime minister in India 2029?
- How many quarks will be detected in a certain collision at the Large Hadron Colloider near Geneva?
- and many others ...

# A few questions that data can help us answer!

- When will be th e visible from Bangalore?

**deterministic, periodic, predictable**

- closest approach of Halley's comet in next 200 years?
- How many times in th l a double pendulum reach the e the angle of a double pendulum after 5 min., 10 min., ...?

**deterministic, chaotic, unpredictable**

- What will be the total nsoon rainfall in India 2019?

**deterministic(?), complex, multi-scale**

**"millions of double pendula!"**

- What will be the extent of th ice over next 50 years?
- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin tosses (Simpler?)?
- How many cars will pass in front of Eiffel tower in next 5 minutes?
- Who will be the prime minister in India 2029?
- How many quarks will be detected in a certain collision at the Large Hadron Colloider near Geneva?
- and many others ...

# A few questions that data can help us answer!

- When will be th… visible from Bangalore?

**deterministic, periodic, predictable**

- …closest approach of Halley's comet in next 200 years?
- How many times in th… l a double pendulum reach the … the angle of a double pendulum after 5 min., 10 min., …?

**deterministic, chaotic, unpredictable**

- What will be the total … nsoon rainfall in India 2019?
- What will be the extent of th… ice over next 50 years?

**deterministic(?), complex, multi-scale**

**“millions of double pendula!”**

- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin tosses (Simpler?)?
- How many cars will pass in front of Eiffel tower in next 5 minutes?
- Who will be the prime minister in India 2029?
- How many quarks will be detected in a certain collision at the Large Hadron Colloider near Geneva?
- and many others …

# A few questions that data can help us answer!

- When will be th visible from Bangalore?
- closest approach of Halley's comet in next 200 years?

**deterministic, periodic, predictable**

- How many times in th a double pendulum reach the angle of a double pendulum after 5 min., 10 min., ...?

**deterministic, chaotic, unpredictable**

- What will be the total nsoon rainfall in India 2019?
- What will be the extent of th ice over next 50 years?

**deterministic(?), complex, multi-scale**

**“millions of double pendula!”**

- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin tosses (Simpler?)?
- How many cars will pass in front of Eiffel tower in next 5 minutes?
- Who will be the prime minister in India 2029?
- How many quarks will be detected in a certain collision at the Large Hadron Colloider near Geneva?
- and many others ...

# A few questions that data can help us answer!

- When will be th … e visible from Bangalore? … approach of Halley's comet in next 200 years?

**deterministic, periodic, predictable**

- How many times in th … l a double pendulum reach the … angle of a double pendulum after 5 min., 10 min., …?

**deterministic, chaotic, unpredictable**

- What will be the total … nsoon rainfall in India 2019?
- What will be the extent of th … ice over next 50 years?

**deterministic(?), complex, multi-scale**

**"millions of double pendula!"**

- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin tosses (Simpler?)?
- How many cars will pass in front of Eiffel tower in next 5 minutes?
- Who will be the prime minister in India 2029?
- How many quarks will be detected in a certain collision at the Large Hadron Colloider near Geneva?
- and many others …

# A few questions that data can help us answer!

- When will be th e visible from Bangalore?
  closest approach of Halley's comet in next 200 years?

**deterministic, periodic, predictable**

- How many times in th l a double pendulum reach the
  the angle of a double pendulum after 5 min., 10 min., ...?

**deterministic, chaotic, unpredictable**

- What will be the total nsoon rainfall in India 2019?
- What will be the extent of th ice over next 50 years?

**deterministic(?), complex, multi-scale**

**"millions of double pendula!"**

- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin t Simpler?)?
- How many cars will pass i ext 5 minutes?
- Who will be t
- H rks w ted in a certain collision at the Large ron Colloider near Geneva?
- and many others ...

**stochastic, answers in probabilistic sense**

**"a million coin tosses!"**

# A few questions that data can help us answer!

- When will be th e visible from Bangalore? e closest approach of Halley's comet in next 200 years?

**deterministic, periodic, predictable**

- How many times in th l a double pendulum reach the e the angle of a double pendulum after 5 min., 10 min., ...?

**deterministic, chaotic, unpredictable**

- What will be the total nsoon rainfall in India 2019?
- What will be the extent of th ice over next 50 years?

**deterministic(?), complex, multi-scale**

**"millions of double pendula!"**

- What will be the three major U.S. stock indices over the coming week? What will be the next 5 coin to Simpler?)?
- How many cars will pass i xt 5 minutes?
- Who will be th
- H rks w ted in a certain collision at the Large on Colloider nea Geneva?

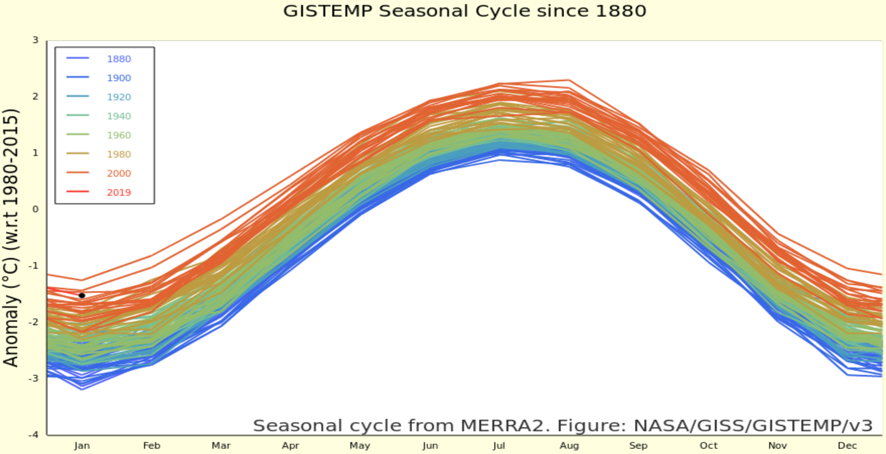
**stochastic, answers in probabilistic sense**

**"a million coin tosses!"**

- and many others ...

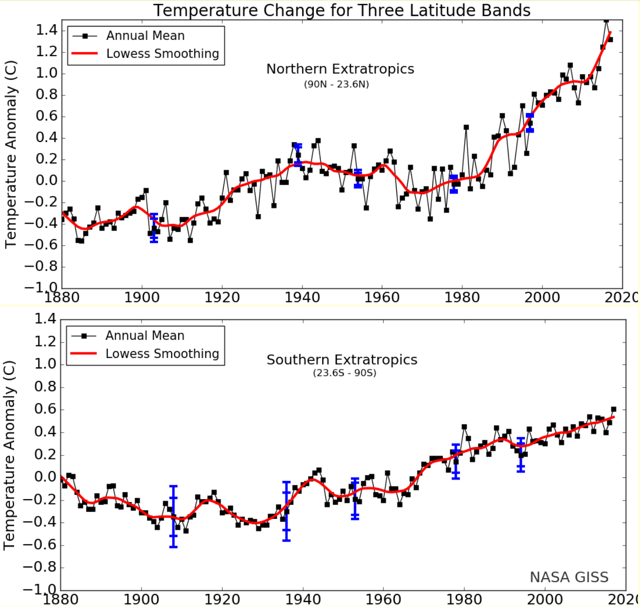
# Climate change is an important problem...

So what about changes in the climate? Or global warming? Here is what we know (temperature of the earth from 1880-2017)



`https://data.giss.nasa.gov/gistemp/graphs/`

# Climate change is an important problem...



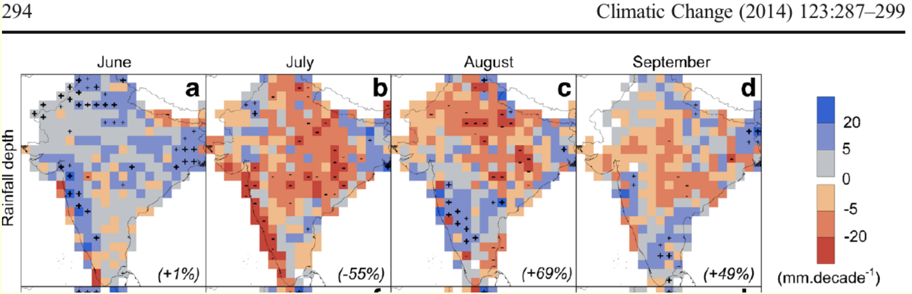
But, the global changes are not uniform: Northern extra-tropics have warmed more than the southern.

`https://data.giss.nasa.gov/`

`gistemp/graphs/`

# Climate change is an important problem...

Even more locally, changes are non-uniform:



294 Climatic Change (2014) 123:287–299

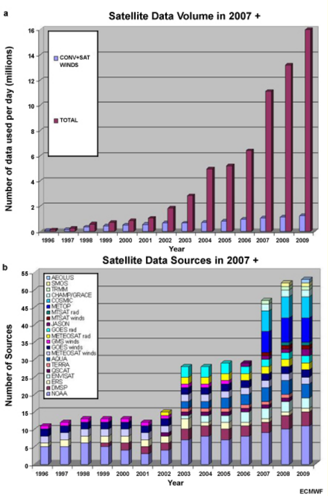
Rate of change of rainfall is more in one part than in other.
(Even the sign of change is different).

Lacombe G, McCartney M (2014) Uncovering consistencies in Indian rainfall trends observed over the last half century. Clim. Change 123(2): 287-299. http://dx.doi.org/10.1007/s10584-013-1036-5

# Data are the key to unravelling these complex mysteries

A remarkable change in the last 20-30 years: the amount of data is increasing "exponentially." Example from weather prediction:

- The first attempt at weather prediction used around 50-100 data points (Richardson 1920s)
- Next attempts: von Neumann, Charney, 1950s: a few KB (kilo=1000) of data
- 1970s - 1980s: a few 100 KB / a few MB (mega=1000 KB)
- Currently: 100s of MB / a few GB (giga=1000 MB)
- 2015-2020: a few TB (tera=1000GB)



ECMWF

# The other key is: scientific computations

Computing power has increased at the same rate as the availability of data.

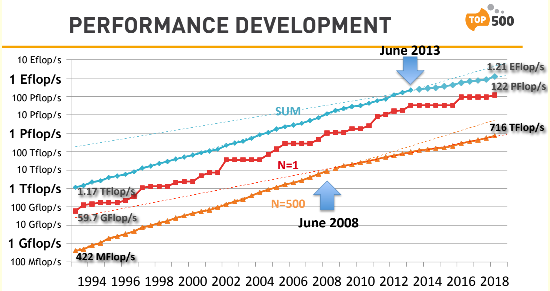
- The first attempt at weather prediction needed a few weeks of calculations by hand!! (Richardson 1920s)
- Next attempts: von Neumann, Charney, 1950s: a few Kilo-flop/s
- 1970s - 1980s: a few Mega-flop/s
- Currently: Peta-flop/s
- 2015-2020: Exa-flop/s

"32 x 2000 = 64000 computers [humans!!] would be needed to race the weather for the whole globe. That is a staggering figure." (Richardson 1922, p.219); image (c) Stephen Conlin 1986

# The other key is: scientific computations

Computing power has increased at the same rate as the availability of data.

- The first attempt at weather prediction needed a few weeks of calculations by hand!! (Richardson 1920s)
- Next attempts: von Neumann, Charney, 1950s: a few Kilo-flop/s
- 1970s - 1980s: a few Mega-flop/s
- Currently: Peta-flop/s
- 2015-2020: Exa-flop/s

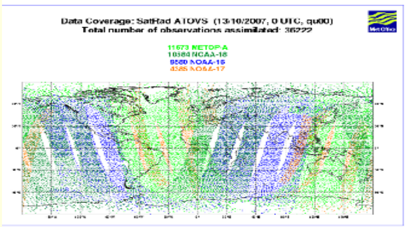


from https:

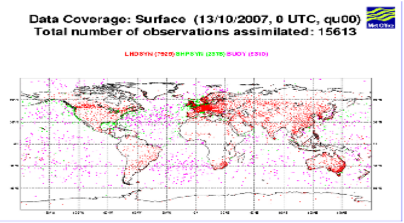
//www.top500.org/static/media/uploads/top500_ppt_201806.pdf

# Two specific areas that aim to combine data with models

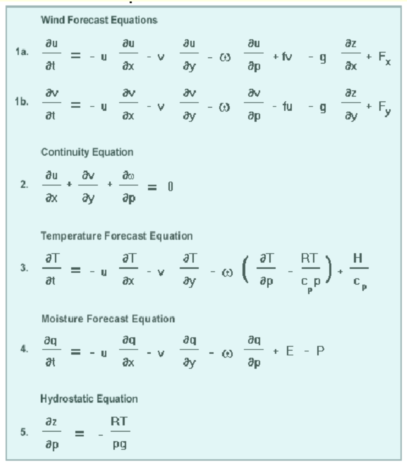
- Data assimilation: how do we use the observations, e.g. each dot on the left panel and more, with numerical models, e.g. equations shown on right?



Data Coverage: SatRad ATOVS (13/10/2007, 0 UTC, qu00)
Total number of observations assimilated: 36222



Data Coverage: Surface (13/10/2007, 0 UTC, qu00)
Total number of observations assimilated: 15613



Wind Forecast Equations

1a. $$\frac{\partial u}{\partial t} = - u \frac{\partial u}{\partial x} - v \frac{\partial u}{\partial y} - \omega \frac{\partial u}{\partial p} + fv - g \frac{\partial z}{\partial x} + F_x$$

1b. $$\frac{\partial v}{\partial t} = - u \frac{\partial v}{\partial x} - v \frac{\partial v}{\partial y} - \omega \frac{\partial v}{\partial p} - fu - g \frac{\partial z}{\partial y} + F_y$$

Continuity Equation

2. $$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial \omega}{\partial p} = 0$$

Temperature Forecast Equation

3. $$\frac{\partial T}{\partial t} = - u \frac{\partial T}{\partial x} - v \frac{\partial T}{\partial y} - \omega \left( \frac{\partial T}{\partial p} - \frac{RT}{c_p p} \right) + \frac{H}{c_p}$$

Moisture Forecast Equation

4. $$\frac{\partial q}{\partial t} = - u \frac{\partial q}{\partial x} - v \frac{\partial q}{\partial y} - \omega \frac{\partial q}{\partial p} + E - P$$

Hydrostatic Equation

5. $$\frac{\partial z}{\partial p} = - \frac{RT}{pg}$$

# Two specific areas that aim to combine data with models

- Markov random field model: can we find dominant patterns in Indian summer monsoon rainfall over last 100 years?

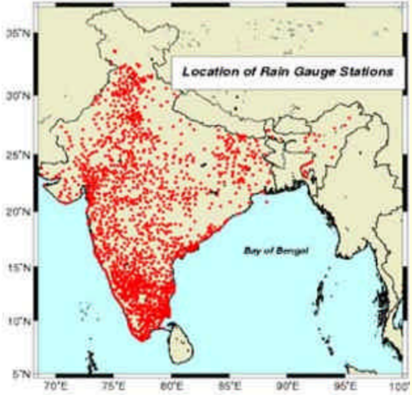


**Figure 1.** Location of 1803 rain gauge stations.

CURRENT SCIENCE, VOL. 91, NO. 3, 10 AUGUST 2006

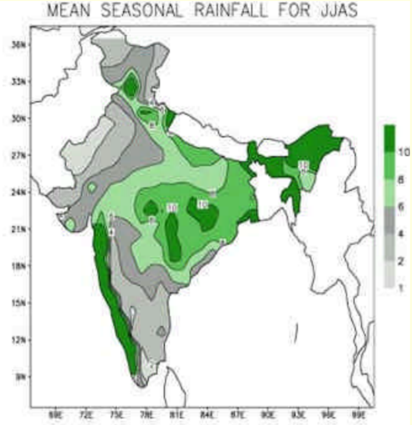


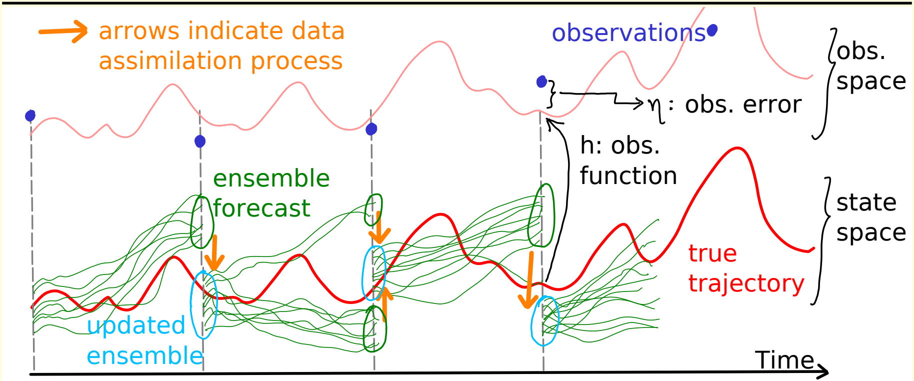
**Figure 3.** Spatial pattern of southwest monsoon seasonal (June t September) mean rainfall (mm/day).

# Data assimilation

## Combining data with observational models

The art of optimally incorporating

- partial and noisy observational data of a
- chaotic, nonlinear, complex dynamical system with an
- imperfect model (of the data and the system dynamics) to get an
- estimate and the associated uncertainty for the system state

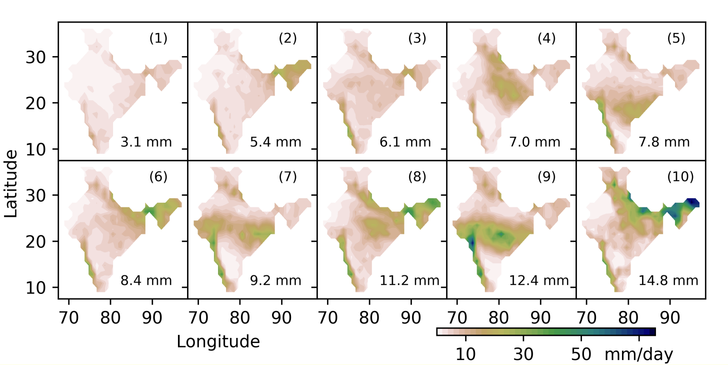


# Markov random fields

## Extracting patterns from data

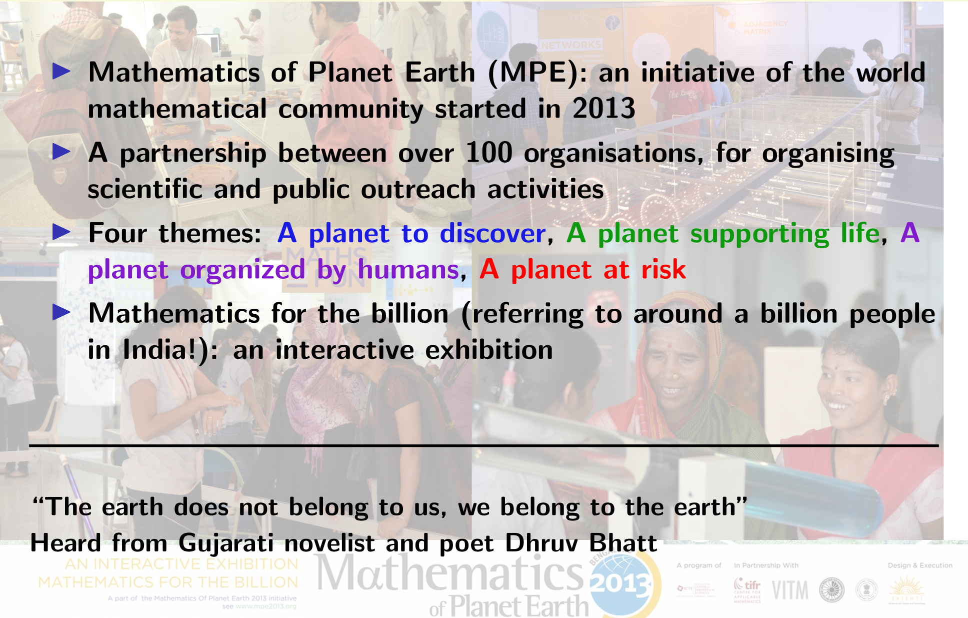
A probabilistic model that

- incorporates "domain knowledge" using probabilities, which are
- conditioned on observed rainfall data, with the aim of
- achieving clustering of locations and of days, and
- identifying dominant patterns in monsoon rainfall data



Common rainfall patterns for Indian summer monsoon (from `https://doi.org/10.1093/climsys/dzy009`)

# Mathematics: for the Planet Earth


NETWORKS
ADJACENCY MATRIX
MATHS FUN
AN INTERACTIVE EXHIBITION
MATHEMATICS FOR THE BILLION
A part of the Mathematics Of Planet Earth 2013 initiative
see www.mpe2013.org
Mathematics of Planet Earth
BENGALURU 2013
A program of
In Partnership With
tifr CENTRE FOR APPLICABLE MATHEMATICS
VITM
Design & Execution
SRISHTI

- **Mathematics of Planet Earth (MPE): an initiative of the world mathematical community started in 2013**
- **A partnership between over 100 organisations, for organising scientific and public outreach activities**
- **Four themes: A planet to discover, A planet supporting life, A planet organized by humans, A planet at risk**
- **Mathematics for the billion (referring to around a billion people in India!): an interactive exhibition**

---

**“The earth does not belong to us, we belong to the earth”**
**Heard from Gujarati novelist and poet Dhruv Bhatt**