Interacting with Data

Léon Bottou

NEC Labs America

COS 424 – 2/2/2010

Summary

- Three short stories.

- Practical information about the course.

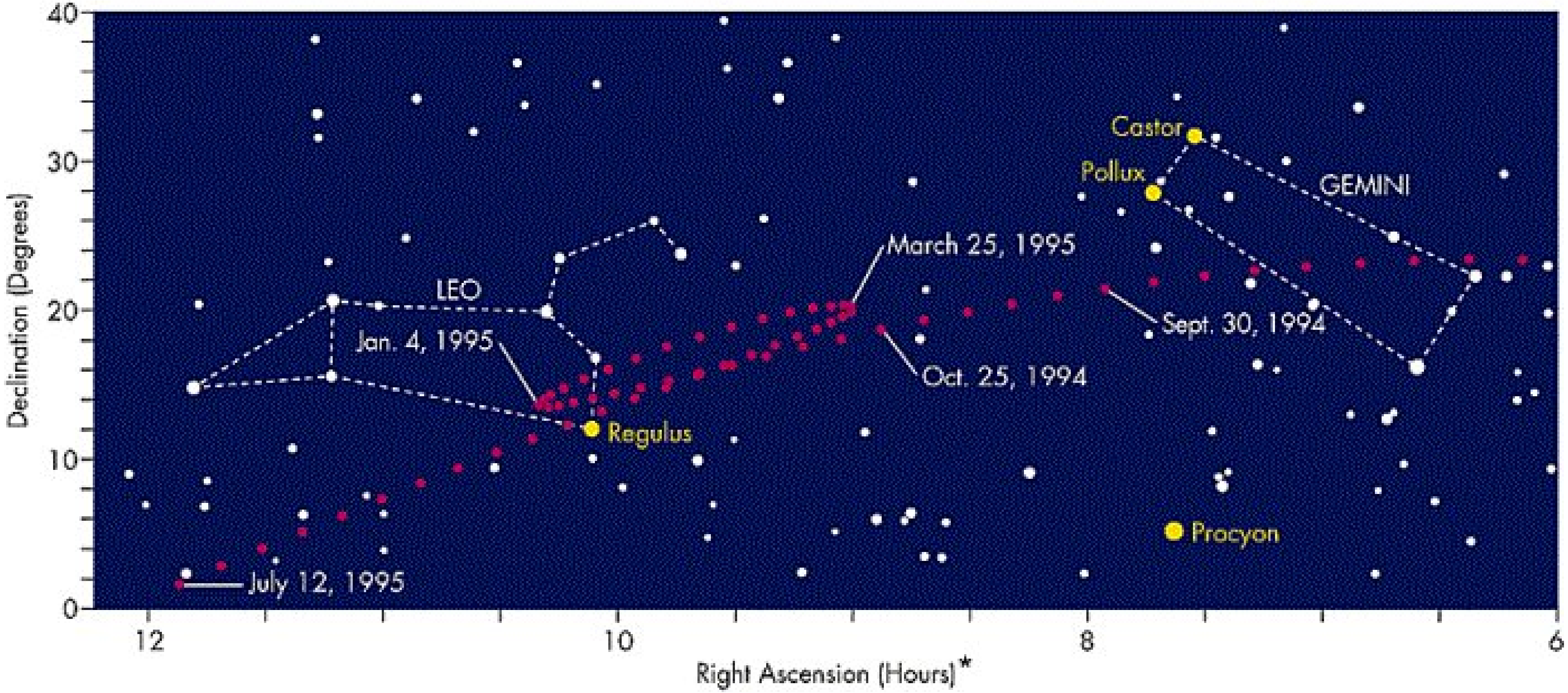
Story 1 – The orbit of Mars

Suppose you are ancient Greeks watching the sky.

- Stars move in unison. Like a big sphere.
- The Sun and the Moon follow nice trajectories relative to the stars. Like points sitting on interior spheres.
- The Planets are bizarre.
  - Mercury and Venus never go very far from the Sun.
  - Mars, Jupiter and Saturn follow very strange trajectories.

Story 1 – Retrograde Motion

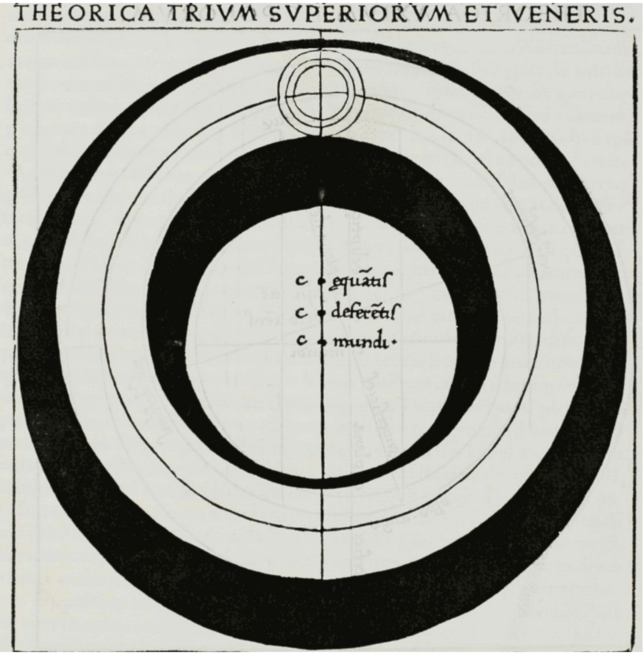
Mars makes really strange moves.

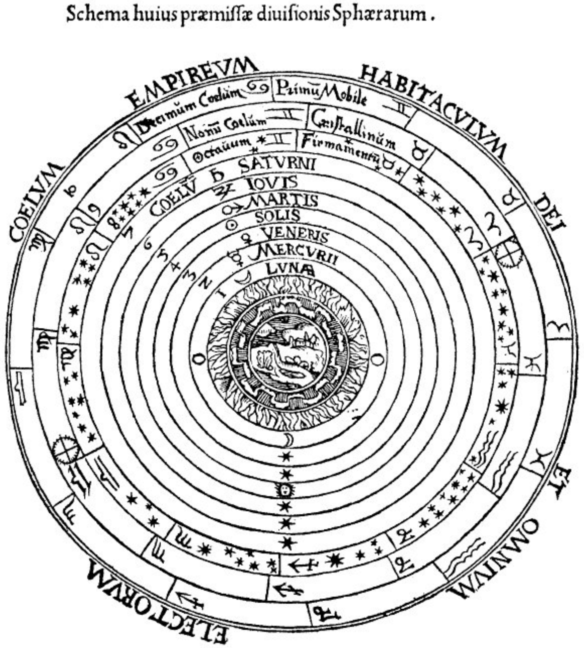


Jupiter and Saturn do the same, but that takes a lot longer.

Story 1 – Cycles and Epicycles

Aristotle (384-322BC), Ptolemy (90-168AD) : 53 to 55 spheres.

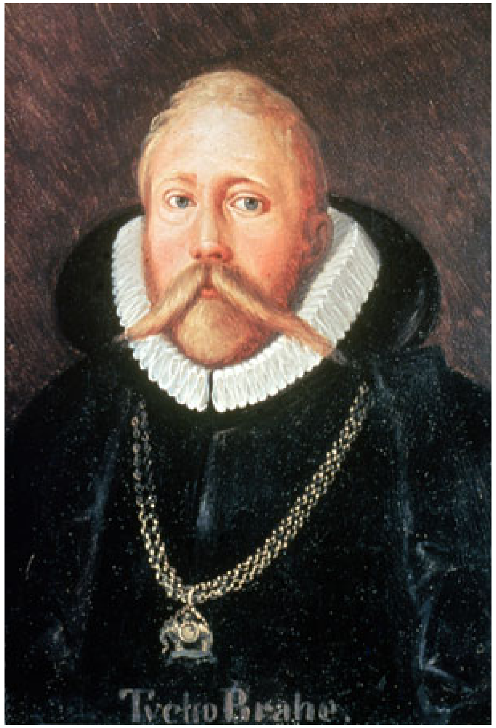




Copernicus : Puts the Sun in the center. Keeps the spheres.

Observation tables were not accurate enough to sort them out.

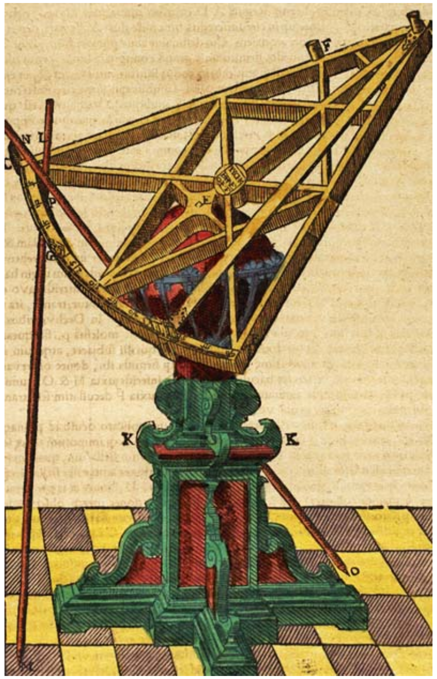
Story 1 – The Characters

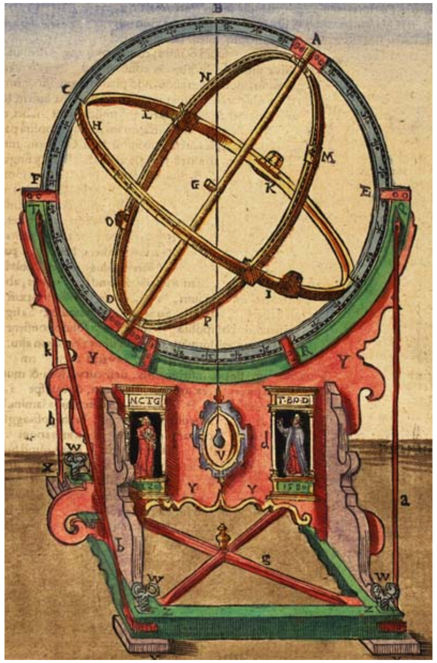


Tycho Brahe
1546-1601

Johannes Kepler
1571-1630

Story 1 – Tycho's Observatories





First in Uraniborg.

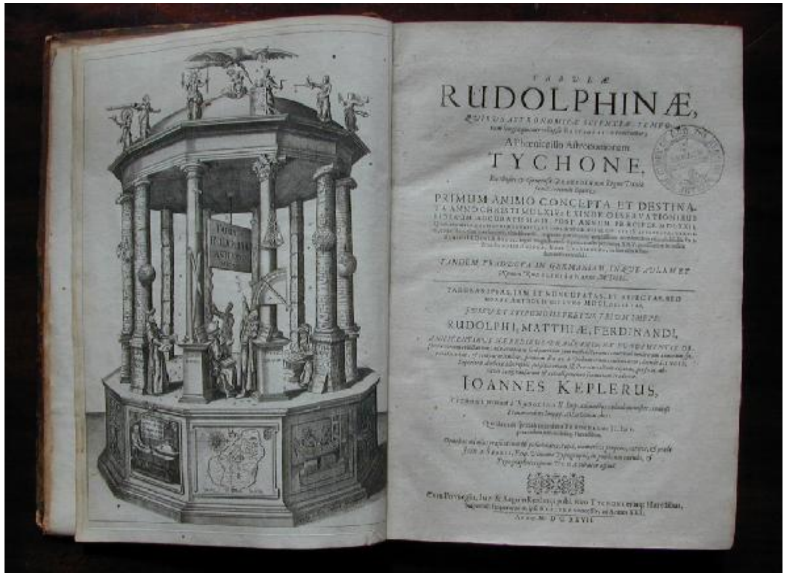
Then near Prague, thanks to a "grant" from emperor Rudolf II.

There he hires a bright young assistant named Johannes Kepler.

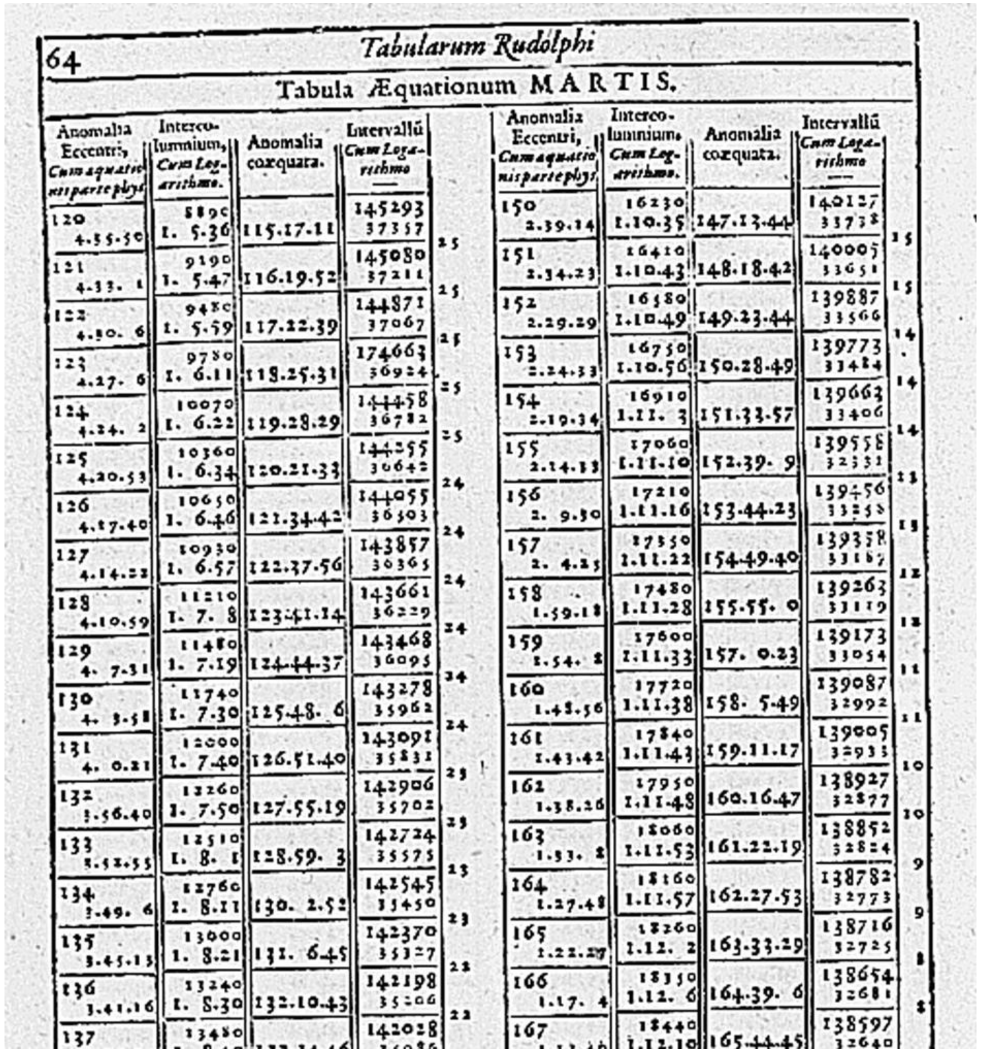
Without telescope, but with a modern approach to data collection:

- daily observation of 1000 stars and 7 planets,
- record positions $\pm 1'$ arc.

Story 1 – The Rudolphine Tables


TABULÆ
RUDOLPHINÆ,
TYCHONE,
PRIMUM ANIMO CONCEPTA ET DESTINA-
RUDOLPHI, MATTHIÆ, FERDINANDI,
IOANNES KEPLERUS,
ANNO M. DC. XXVII.

The *Tabulae Rudolphinae*
were finally published by
Kepler in 1627 under
emperor Ferdinand.

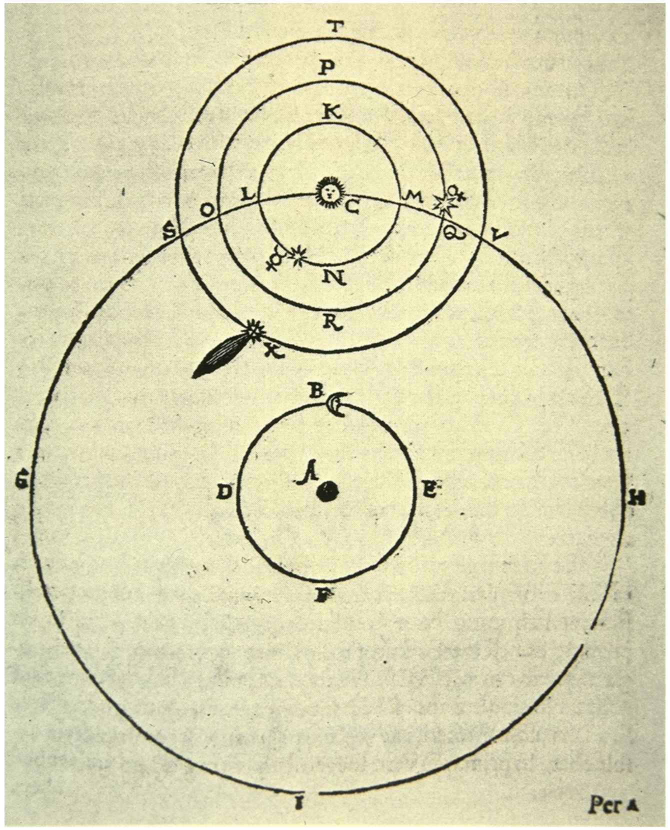


64 *Tabularum Rudolphi*

Tabula Æquationum MARTIS.

Anomalia Eccentri, *Cum æquationis parte physica*	Intercolumnium, *Cum Logarithmo.*	Anomalia coæquata.	Intervallū *Cum Logarithmo*		Anomalia Eccentri, *Cum æquationis parte physica*	Intercolumnium, *Cum Logarithmo.*	Anomalia coæquata.	Intervallū *Cum Logarithmo*	
120 4.55.50	8890 1. 5.36	115.17.11	145293 37357	25	150 2.39.14	16230 1.10.35	147.13.44	140127 33738	15
121 4.33. 1	9190 1. 5.47	116.19.52	145080 37211	25	151 2.34.23	16410 1.10.43	148.18.42	140005 33651	15
122 4.30. 6	9480 1. 5.59	117.22.39	144871 37067	25	152 2.29.29	16580 1.10.49	149.23.44	139887 33566	14
123 4.27. 6	9780 1. 6.11	118.25.31	174663 36924	25	153 2.24.33	16750 1.10.56	150.28.49	139773 33484	14
124 4.24. 2	10070 1. 6.22	119.28.29	144458 36782	25	154 2.19.34	16910 1.11. 3	151.33.57	139663 33406	14
125 4.20.53	10360 1. 6.34	120.21.33	144255 36642	24	155 2.14.33	17060 1.11.10	152.39. 9	139558 33331	13
126 4.17.40	10650 1. 6.46	121.34.42	144055 36503	24	156 2. 9.30	17210 1.11.16	153.44.23	139456 33258	13
127 4.14.22	10930 1. 6.57	122.37.56	143857 36365	24	157 2. 4.25	17350 1.11.22	154.49.40	139358 33187	12
128 4.10.59	11210 1. 7. 8	123.41.14	143661 36229	24	158 1.59.18	17480 1.11.28	155.55. 0	139263 33119	12
129 4. 7.31	11480 1. 7.19	124.44.37	143468 36095	24	159 1.54. 8	17600 1.11.33	157. 0.23	139173 33054	11
130 4. 3.58	11740 1. 7.30	125.48. 6	143278 35962	24	160 1.48.56	17720 1.11.38	158. 5.49	139087 32992	11
131 4. 0.21	12000 1. 7.40	126.51.40	143091 35831	23	161 1.43.42	17840 1.11.43	159.11.17	139005 32933	10
132 3.56.40	12260 1. 7.50	127.55.19	142906 35702	23	162 1.38.26	17950 1.11.48	160.16.47	138927 32877	10
133 3.52.55	12510 1. 8. 1	128.59. 3	142724 35575	23	163 1.33. 8	18060 1.11.53	161.22.19	138852 32824	9
134 3.49. 6	12760 1. 8.11	130. 2.52	142545 35450	23	164 1.27.48	18160 1.11.57	162.27.53	138782 32773	9
135 3.45.13	13000 1. 8.21	131. 6.45	142370 35327	22	165 1.22.27	18260 1.12. 2	163.33.29	138716 32725	8
136 3.41.16	13240 1. 8.30	132.10.43	142198 35206	22	166 1.17. 4	18350 1.12. 6	164.39. 6	138654 32681	8
137	13480		142028		167	18440 1.12.10	165.44.45	138597 32640	

Story 1 – The “War with Mars”



The Tychonic system

(Copernicus light)

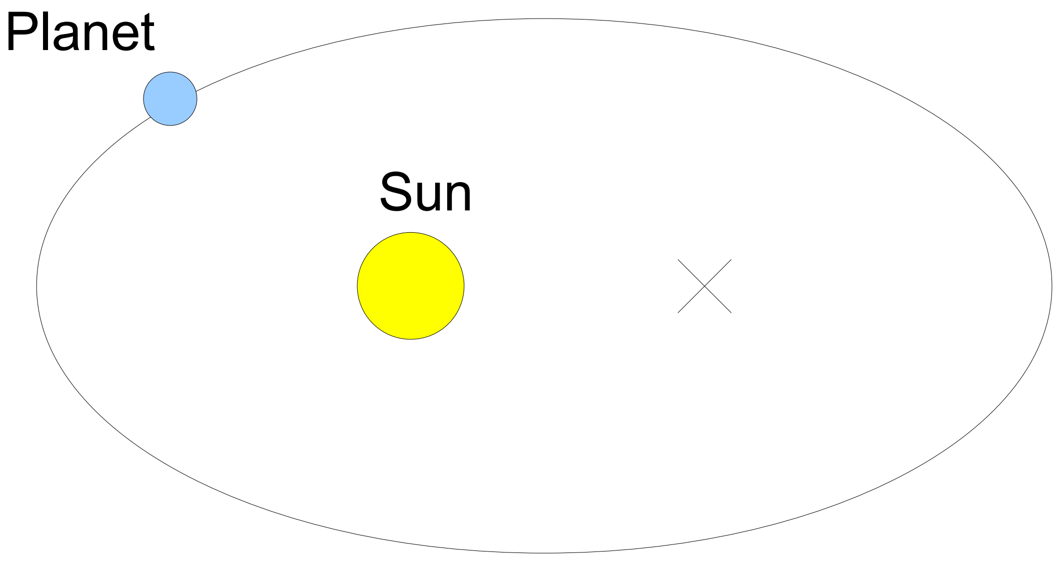
First model of the orbit of Mars.
under Tycho’s direction
– average discrepancy: 2’.
– maximal discrepancy: 8’.

Kepler still unhappy.
He wants to go Copernican.
Tycho does not like that.

Tycho died in 1601.

Story 1 – First law of Kepler (1605)

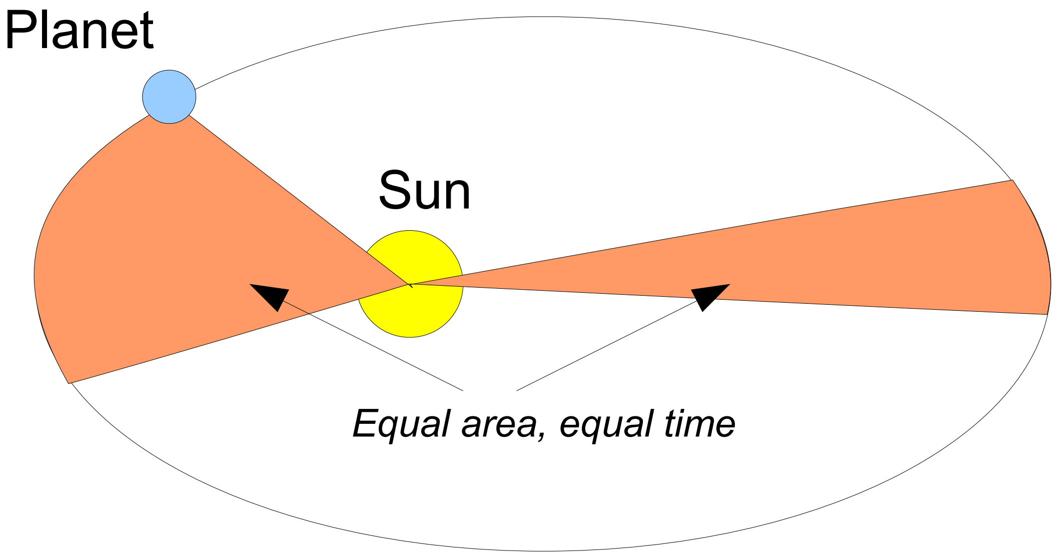
The orbits of the planets are ellipses
with the Sun at a focal point.



J. Kepler, *Astronomia nova*, 1609

Story 1 – Second law of Kepler (1609)

The line joining the planet to the Sun sweeps out equal areas in equal times as the planet travels around the ellipse.



J. Kepler, *Astronomia nova*, 1609

Story 1 – Third law of Kepler (1619)

The ratio of the squares of the revolutionary periods for two planets is equal to the ratio of the cubes of the length of their major axes.

$$\frac{P_{\mathrm{a}}^2}{P_{\mathrm{b}}^2} = \frac{R_{\mathrm{a}}^3}{R_{\mathrm{b}}^3}$$

J. Kepler, *Harmonices Mundi*, 1619

Story 1 – Validation

- Kepler had it mostly right in 1605.

- Galileo points a telescope to the sky. He observes the phases of Venus with a telescope in 1610 and concludes that Venus orbits the Sun.

- Newton publishes the *Principia* in 1687 and shows that the laws of Kepler (with a small correction) derive from his mechanics and from the idea of gravitation.

Story 1 – Epilogue

This is about the **foundation of the modern scientific approach**.

1. Get the best data you can.

2. Build models that fit the data as closely as possible.

3. Make sure you get external validation.
   – validate with testing data set aside from the beginning.
   – validate using different datasets for the same problem.
   – and more generally, build a convincing story...

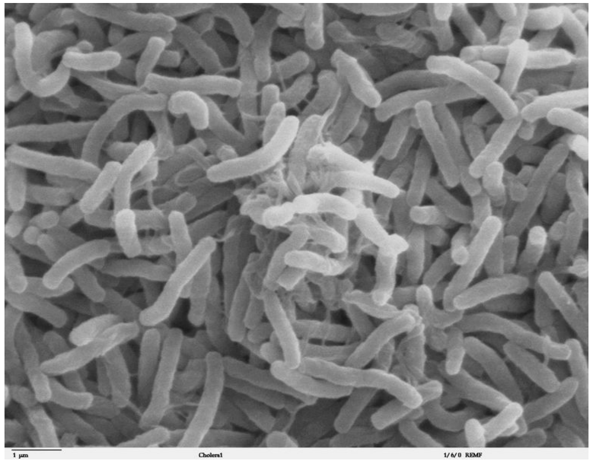
Story 2 – Cholera in London

London, 1854 :

- Industrial revolution.
- Two millions people.
- Insufficient sewage.
- Garbage removal problems.
- Little clean water.

Story 2 – The Characters

John Snow
1813-1858



Vibrio Cholerae
still around

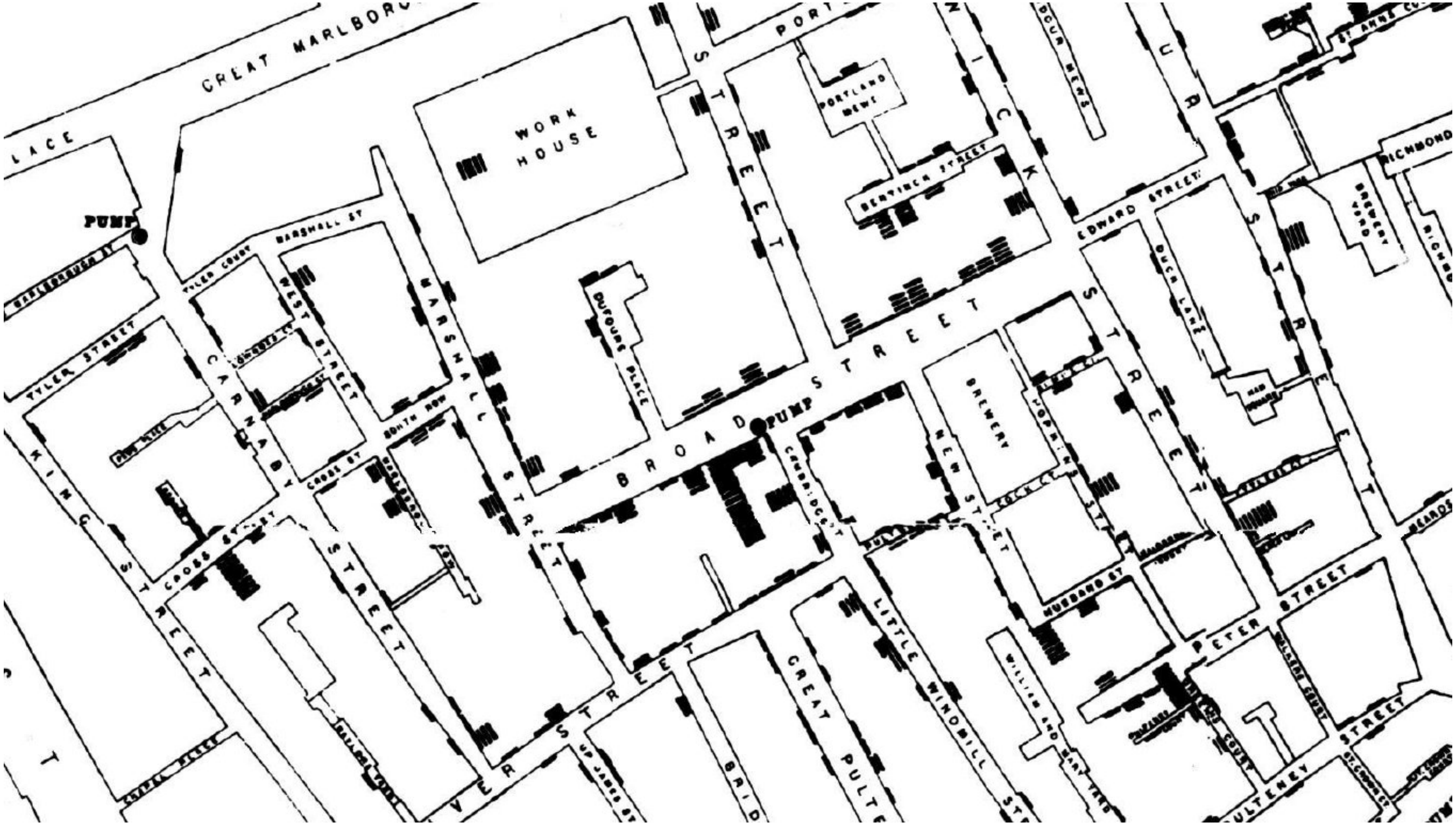
John Snow was a strong advocate of *hygiene* and *anesthesia*.

Story 2 – The Outbreak

> *The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and the adjoining streets, a few weeks ago. Within two hundred and fifty yards of the spot where Cambridge Street joins Broad Street, there were upwards of five hundred fatal attacks of cholera in ten days. The mortality in this limited area probably equals any that was ever caused in this country, even by the plague, and it was much more sudden, as the greater number of cases terminated in a few hours.*

John Snow, *On the mode of communication of cholera*, 1854.

Story 2 – The Map



Story 2 – The Broad Street Pump

> *On proceeding to the spot, I found that nearly all the deaths had taken place within a short distance of the [Broad Street] pump. There were only ten deaths in houses situated decidedly nearer to another street-pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pumps which were nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street. . .*

John Snow, *On the mode of communication of cholera*, 1854.

Story 2 – Epidemiology and Statistics

Snow uses simple statistics to confirm the role of impure water.

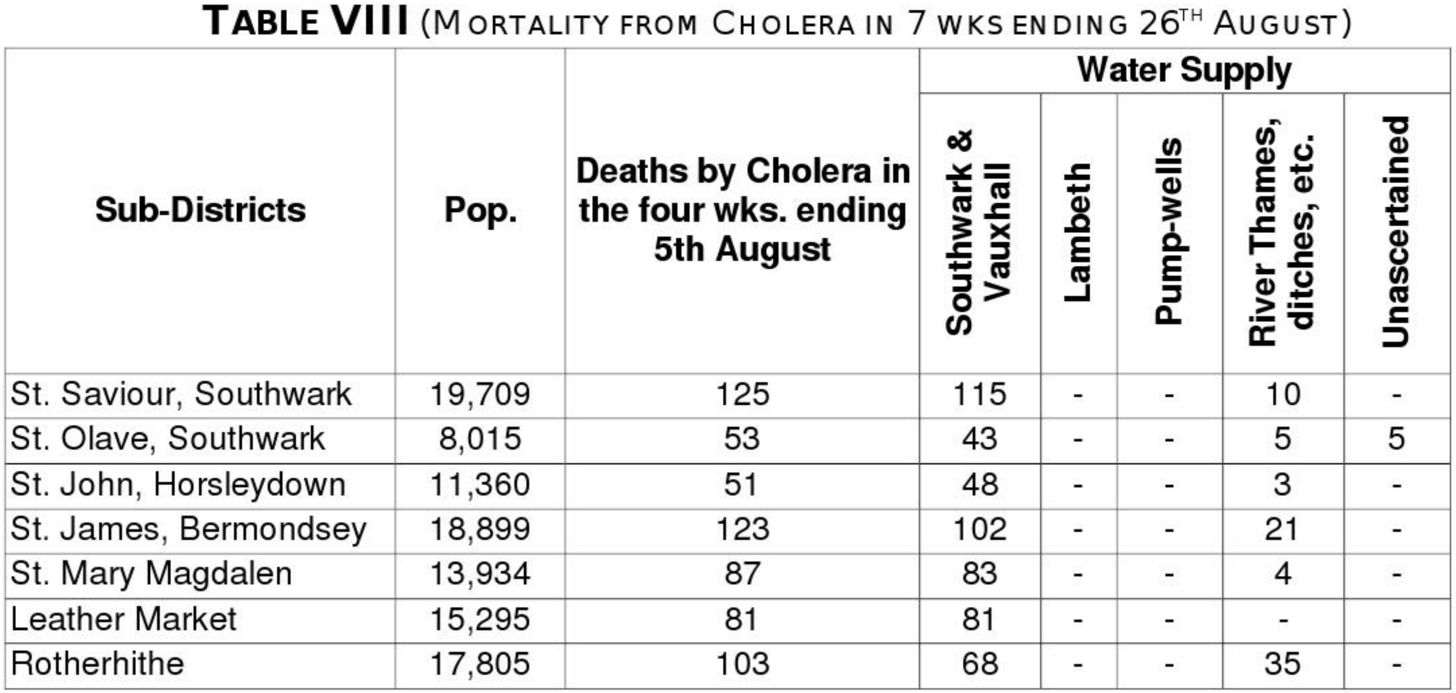


TABLE VIII (MORTALITY FROM CHOLERA IN 7 WKS ENDING 26TH AUGUST)

Sub-Districts	Pop.	Deaths by Cholera in the four wks. ending 5th August	Water Supply				
			Southwark & Vauxhall	Lambeth	Pump-wells	River Thames, ditches, etc.	Unascertained
St. Saviour, Southwark	19,709	125	115	-	-	10	-
St. Olave, Southwark	8,015	53	43	-	-	5	5
St. John, Horsleydown	11,360	51	48	-	-	3	-
St. James, Bermondsey	18,899	123	102	-	-	21	-
St. Mary Magdalen	13,934	87	83	-	-	4	-
Leather Market	15,295	81	81	-	-	-	-
Rotherhithe	17,805	103	68	-	-	35	-

Story 2 – Epidemiology and Statistics

Snow uses simple statistics to infirm competing hypotheses.

Table XIV *(partial)*

	No of Deaths	Ratio
Agents	12	1 in 40
Bricklayers and builders	14	1 in 39
Physicians, surgeons, . . .	16	1 in 265
Magistrates, barristers, . . .	13	1 in 375
Merchants	11	1 in 348
Footmen and men servants	25	1 in 1572

If cholera was propagated by effluvia from sick people,
– why are physicians less affected than their patients?
– why are men servants less affected than their masters?
– why are master brewers virtually immune?

Story 2 – Epilogue

This is again an example of the scientific method.
But there are important differences :

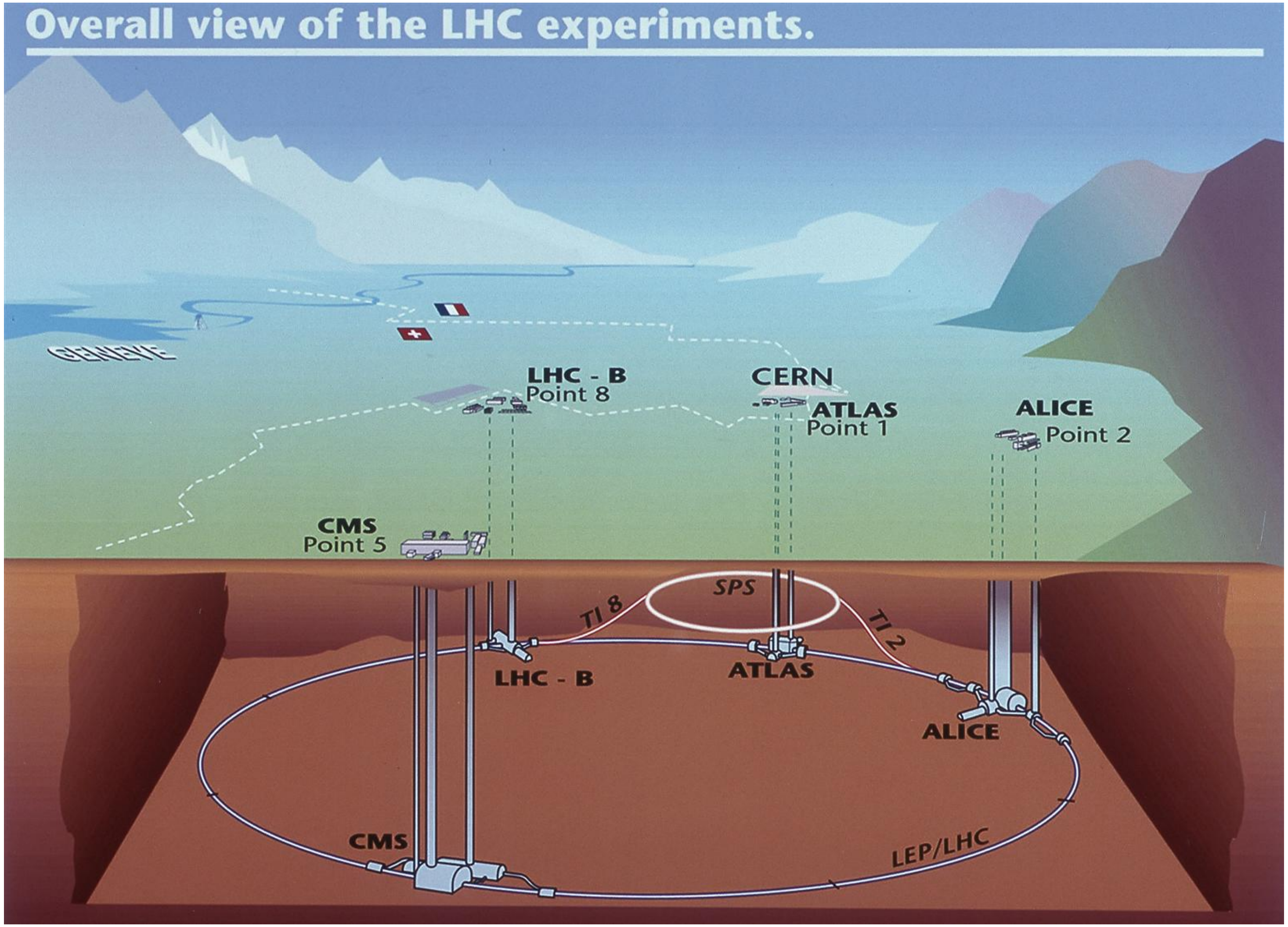
1. **Causality**: prediction versus intervention.
   - *What happens if we shut off the Broad St. pump?*
   - *What happens if we ensure that everyone gets clean water?*

   Nothing in John Snow's data tells us that directly.
   Only randomized experiments could tell.

2. **Noise and Calculus**: $A \Longrightarrow B$ does not mean $almost A \Longrightarrow almost B$.
   - Kepler's orbits are very accurate.
     We can use calculus to answer new questions.
     For instance, predicting eclipses and transit.
   - Simply counting the odds to get cholera ignores many factors.
     Noise accumulates quickly during calculus.
     We need direct evidence from the data to answer new questions.

Story 3 – Big Science

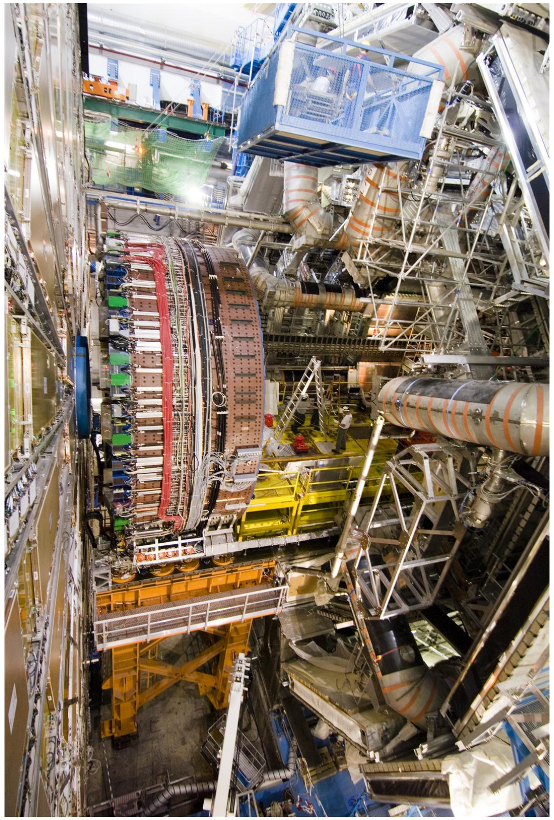
Story 3 – Large Hadron Collider

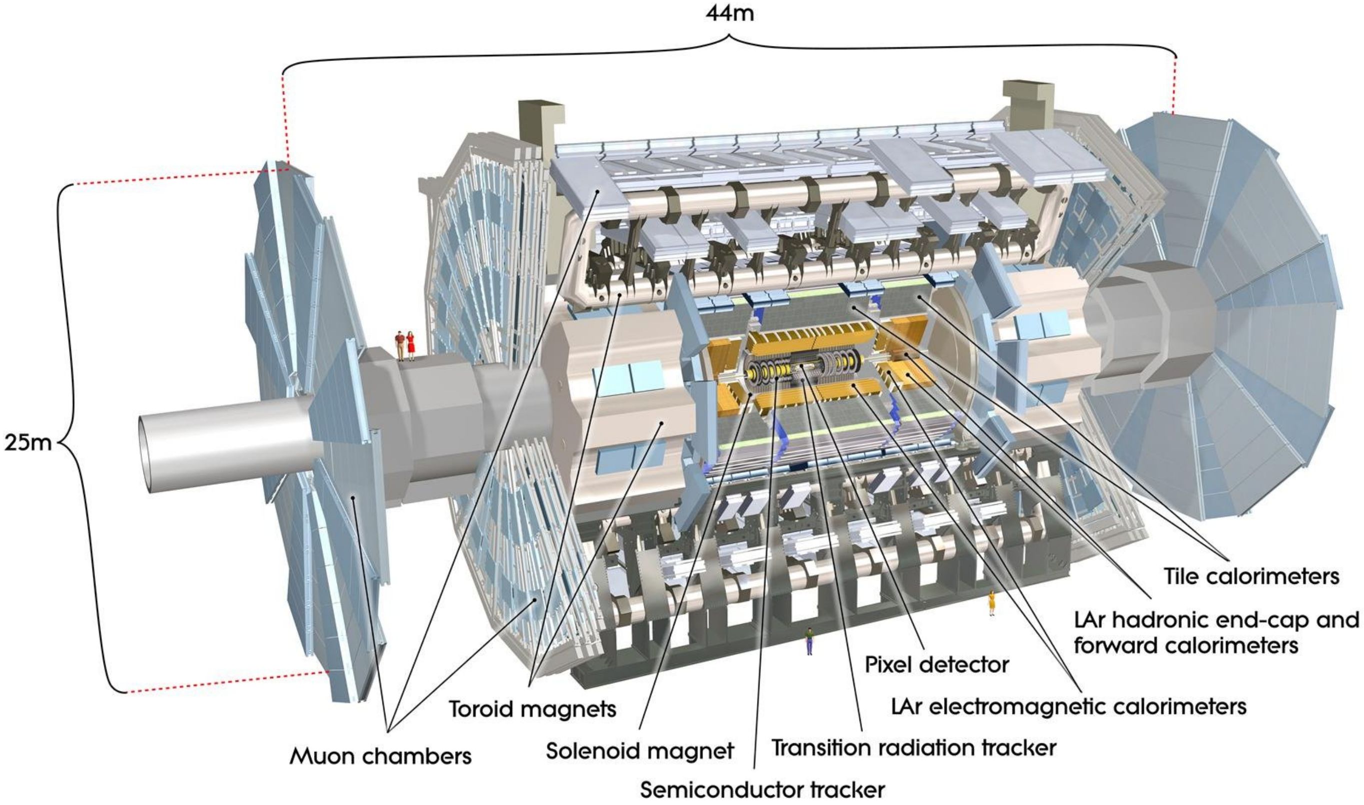


Story 3 – The Characters

The ATLAS collaboration. Many thanks to Kyle Cranmer, NYU

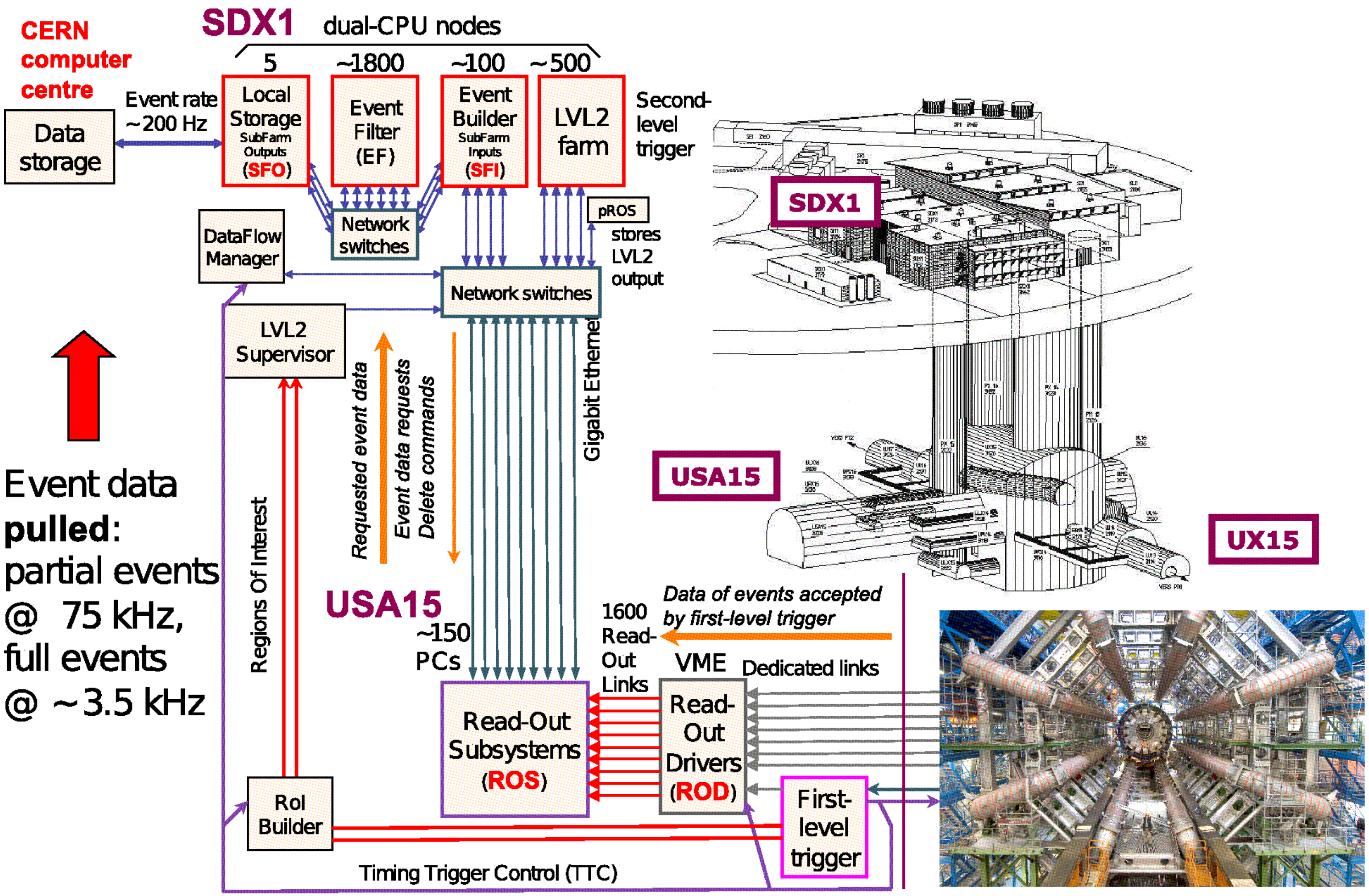
Story 3 – Atlas Overview




44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

– 40M events per second, 25MB per event.
– 1PB/s (reduced to 64TB/s after zero removal).

Story 3 – Atlas Triggers

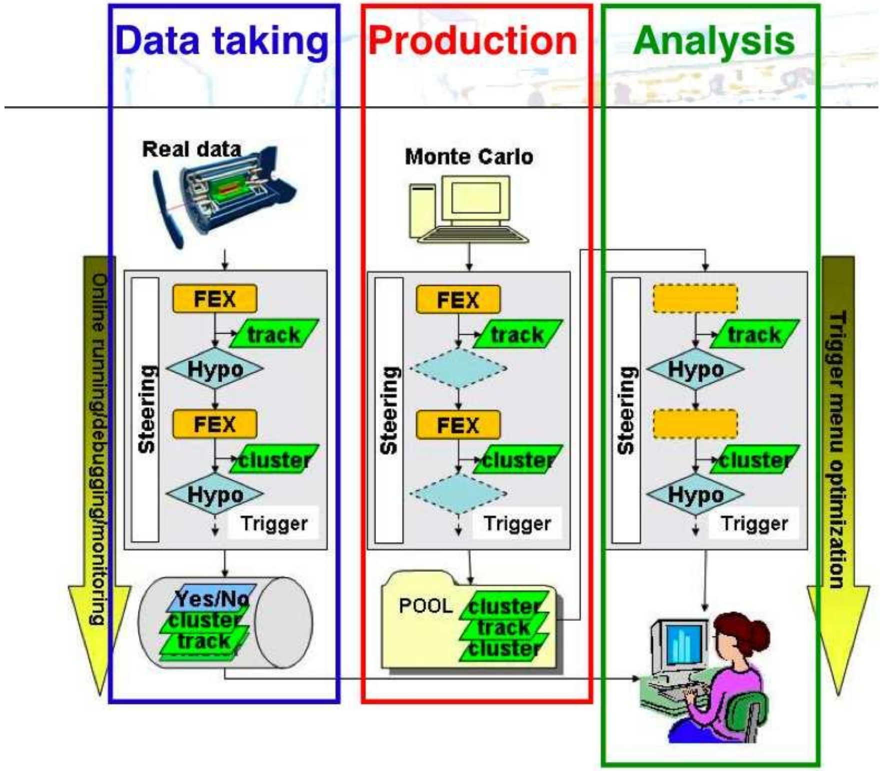


Event data **pulled**:
partial events @ 75 kHz,
full events @ ~3.5 kHz

Event data **pushed** @ 75 kHz,
1600 fragments of ~1 kByte each

Event data @ 40 MhZ
1PB/s (raw) or 64 GB/s (zcomp)

Story 3 – Atlas Analysis



Runs in ∼40 sites worldwide.

Compare statistics from
– observations,
– simulations.

Fine tune the triggers.

Cutting edge physics theories have many adjustable knobs.
They can fit almost any observation. How to validate?

Story 3 – Epilogue

Computers do not change the problems.

Computers change the scale of the problems.

but

What is a minor problem with small scale data

can turn into a formidable problem with large scale data.

The Course

Goals

– Learn selected theoretical tools.

– Learn selected practical approaches.

– Acquire experience with several kinds of data.

– Acquire the right attitude.

Topics

Classification, Clustering, Statistics,
Exploratory methods, Applications, . . .

See Also

– COS511 *Theoretical Machine Learning,* Rob Shapire.

– COS513 *Foundations of probabilistic modeling*, David Blei.

Details

People

- Professor: Léon Bottou `leon@bottou.org`
- TA: Sean Gerrish `sgerrish@cs.princeton.edu`

Web

- `http://www.cs.princeton.edu/courses/archive/spring10/cos424`
- Select [Assignments], [Administrivia].
- Add yourself to the course mailing list.
- Fill the brief survey.

Readings, Scribes

There is no perfect textbook for this class.

Scribe notes

Students will be asked to take scribe notes
which will be posted on the course web site.

Readings

Additional papers and book chapters will also be provided.
Mostly from three books which are on reserve.

– *The Elements of Statistical Learning*, Hastie et al.
– *Pattern Recognition and Machine Learning*, Bishop.
– *Principles of Data Mining*, Hand et al.

Homeworks

Four Homework Assignments

- Progressively acquire practical experience with data.
- Assignments will not be programming assignments.
- Computers as a tool rather than an end.

Software Tools

- There are many options, e.g. Matlab, R.
- Coding everything in C is a *very instructive* option.
- Many knowledgeable people recommend R.
- R tutorial on Tue 9/2, 11am, here, [Sean].

Project

Goals

Acquire experience on something fun.

Process

- Form groups of 2-3 students.
- Write a one page project proposal before 4/13.
- Feel free to discuss your proposal with me.
- Do the work.
- Write report before 5/01.
- Present poster on 5/04.

Next Lecture

Connecting the dots

with common sense and

a little bit of linear algebra.

Who wants to be scribe?