Albert Einstein: A Man for the Millenium?

John Stachel

Department of Physics & Center for Einstein Studies Boston University

Keywords: History of science, Philosophy of science, Special Relativity, General Relativity and Gravitation
PACS: 01.65.+g, 01.70.+w, 03.304+p, 03.70+k, 04.,11.30+Cp

THE LONG VIEW OF HISTORY

True story. Henry Kissinger was in China in 1972, laying the groundwork for President Nixon's visit. At a meeting with Chinese prime minister Chou En-Lai, Mr. Kissinger asked the prime minister if he believed whether the 1789 French Revolution benefited humanity. After mulling over the question for a few minutes, Chou En-Lai replied, "It's too early to tell." (J. Lau, cited from <www.yellowbridge.com/humor/chinaamerica.html>)

FIGURE 1. Kissinger – Chou-En-Lai

Chou En-Lai, heir to a 5000-year old civilization, was obviously trying to "put in his place" the upstart from the barely-200-years-old United States. Yet his answer contains a good deal of wisdom. Often, the historical evaluation of the significance of some important event can change long after the event occurred.

Euclid

Gauss

Bolyai

Lobachevski

FIGURE 2.

Here is an example from the history of science that is relevant to my topic: the case of Euclid, who flourished about 300 BCE. For over three millennia, if anyone asked the question "What was Euclid's major scientific contribution?", the answer was something like: "He codified *the* geometry of space." While I have placed emphasis on the singular, until the beginning of the 19th century there was no need to do so because there simply was no other geometry. As late as 1772, the renowned English philosopher David Hume wrote:

> Though there never were a circle or triangle in nature, the truths demonstrated by Euclid would forever retain their certainty and evidence. (*An Enquiry Concerning Human Understanding*, Section IV).

But, beginning with the work of Carl Friedrich Gauss, who coined the term "non-Euclidean geometry," it became clear that consistent alternative geometries could be developed that differed from Euclid's by negating his famous fifth or parallel postulate:

> If a straight line crossing two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if extended indefinitely, meet on that side on which are the angles less than the two right angles. (*Elements,* Book I)

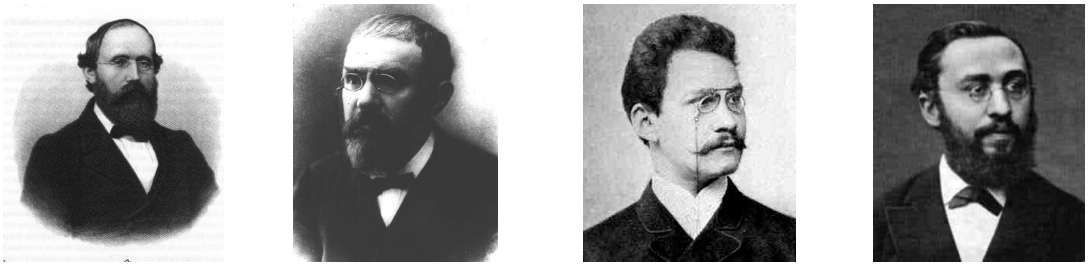
It is easier to formulate the alternatives to this postulate if we use this equivalent form:

> Given any straight line and a point not on it, there exists one and only on straight line that passes through the point and never intersects the first line, no matter how far it is extended.

Gauss considered a geometry in which there could be *more than one* such parallel line, but did not publish his results

> for I fear the cry of the Bœotians [i.e., the philistines] which would arise should I express my whole view on this matter (letter to Bessel, 1829).

Results similar to Gauss' were soon published by János Bolyai (1831) and Nikolai Lobachevski (1829). Some years later, Bernhard Riemann described a second non-Euclidean geometry, in which there are *no parallel lines* (1854 lecture, 1868 posthumous publication).



Riemann Poincaré Minkowski Lorentz

FIGURE 3.

As was soon discovered, two-dimensional Gauss-Bolyai-Lobachevski geometry can be interpreted as the geometry of a space of constant negative curvature (the surface of a hypersphere) embedded (locally) in three-dimensional Euclidean space; while two-dimensional Riemannian geometry can be interpreted as the geometry of a space of constant positive curvature (the surface of a sphere) embedded (globally) in three-dimensional Euclidean space. In both cases, "straight line" is to be interpreted as a geodesic (shortest curve between two points) of the surface.

In all three of these geometries, space is homogeneous and isotropic. Henri Poincaré developed what he called a "fourth geometry, as coherent as those of Euclid, Lobachevski and Riemann" (*Sur les hypothèses fondamentales de la géométrie*, 1887). The parallel postulate holds in this geometry and space is homogeneous; but it is no longer isotropic. In this two-dimensional version, the straight lines through any point fall into two classes; any line in one class can be "pseudo-rotated" into any other in the same class, but no

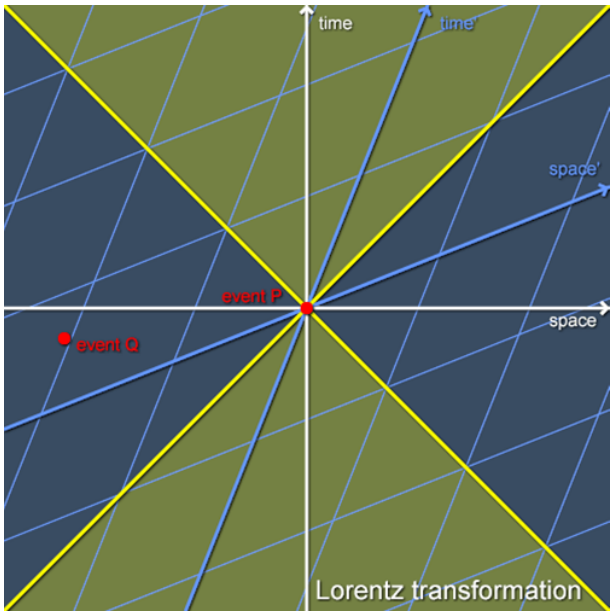


FIGURE 4. Minkowski space in 2 dimensions

"pseudo-rotation" can take a line of one class into a line of the other. The two classes are separated by a pair of straight lines, each of which is orthogonal to itself. With hindsight, we can see that this is a description of two-dimensional Minkowski space-time, the two classes consisting of the time-like and space-like lines, separated by a pair of null lines (see the next section); but no one seems to have realized this until long after the passing of Poincaré and Minkowski.

The four geometries mentioned were only the first of a host of new geometries invented since the floodgates were opened by Gauss. (Later, I shall discuss Weyl's definition of a geometry). Clearly, in the face of this profusion, the old answer to the question of Euclid's major contribution is unacceptable. A modern answer is given in *The Dictionary of Scientific Biography*:

> The *Elements* ... most remarkable feature is the arrangement of the matter so that one proposition follows on another in a strict logical order, with the minimum of assumptions and very little that is superfluous. ... The significance of Euclid's *Elements* in the history of thought is twofold. In the first place, it introduced into mathematical reasoning new standards of rigor which ... have been equaled again only in the past two centuries. In the second place, it marked a decisive step in the geometrization of mathematics ("Euclid," *DSB*, vol. IV).

Euclid's work has served as a model for many later attempts to logically organize other branches of science, and even philosophy (see Spinoza's *Ethica Ordine Geometrico Demonstrata* [*Ethics Demonstrated in Geometric Order*]

What About Einstein?

Having adopted the long view of history, we are ready to consider the question: "How will Einstein be viewed at the end of the next millennium?" In 3005 (assuming humanity survives until then- and given the current state of the world, this is a big assumption), what will physicists regard as his major contribution?

Today, just once century after 1905, we can already see a sifting out of certain items of his total œuvre as most significant. Were we to list *all* his accomplishments, the list would be long indeed. Such a list would certainly include:

- His estimate of molecular size based on the change in viscosity of a liquid when particles are suspended in it.
- His demonstration, based on the first theory of a stochastic process, that microscopic fluctuation phenomena can be observed in Brownian motion.
- His development of a new kinematics in the special theory of relativity, and the deduction from it of such remarkable features as:
  - The path dependence of proper time intervals (twin-paradox")
  - The equivalence of mass and energy ("$E = mc^2$").

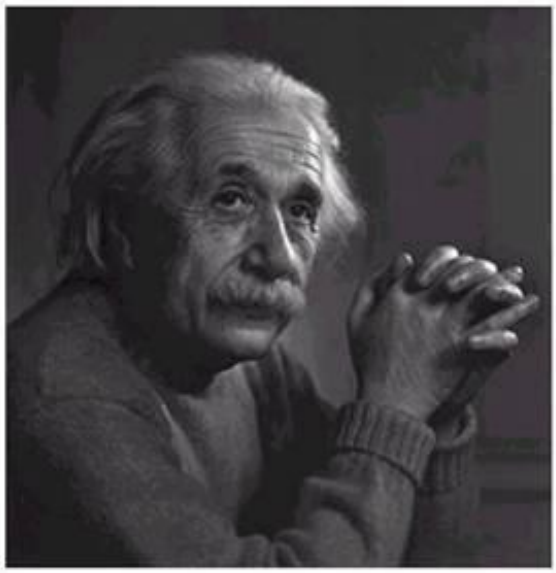


FIGURE 5. Einstein

- His development of general relativity, still the best theory of gravitation that we have.
- His proposal of the light quantum hypothesis, which developed into the theory of the photon, the first elementary particle to be given a quantum treatment.
- His quantum theory of solids, which provided the basis for explaining the anomalous low-temperature behavior of crystalline solids.
- His explanation of Planck's law based on the introduction of the A & B coefficients, which placed the concept of transition probabilities at the center of atomic physics.
- The Einstein-Podolsky-Rosen "paradox," which highlighted the nature of the quantum entanglement of two or more systems.
- His work on Bose-Einstein statistics, leading to his prediction of the existence of Bose-Einstein condensates, only recently confirmed.
- The Einstein-Infeld-Hoffmann derivation of the equations of motion of massive bodies from the field equations of general relativity – the list could go on indefinitely.

But from the perspective of 2005, most physicists would probably agree that a list of the works that did the most to change the domain of physics must include:

- The light quantum hypothesis and quantum theory of solids, which ultimately led to the fulfillment of Einstein's early prediction that neither classical mechanics (including its special-relativistic modifications) nor classical electrodynamics could survive the onslaught of the quantum of action. Of course the form taken by the fulfillment– non-relativistic quantum mechanics and (special-)relativistic quantum

field theory– left Einstein quite dissatisfied. This question is discussed in a later section.

- Special relativity (SR), which led to a realization that all of physics, including the future theory of elementary particles (and excluding only gravitation), would have to be reformulated in terms of representations of the Poincaré (or inhomogeneous Lorentz) group.
- General relativity (GR), which provides a theory of the inertio-gravitational field. It goes beyond the special theory by turning all space-time structures into dynamic fields. GR has survived 90 years of theoretical challenges and experimental tests, both local and astronomical, and forms the basis for current treatments of cosmology.

Given this 2005 perspective, perhaps it not excessive *hubris* to raise the question of how matters will look from the perspective of 3005. Clearly by then, most of the details mentioned above will have faded from sight; but I shall propose that Einstein's work on space-time structures provides clues suggesting a plausible guess about what will survive. However, before gazing into the crystal ball of prophecy, we need to look back at the history of the development of the concept of space-time in physics and discuss some philosophical controversies about its nature.

HISTORICAL SECTION

It has long been clear that space and time are intimately related in physics. Kinematics, the description of motion as change of place over time, involves both, as was already clear to Aristotle:

> Evidently time does not exist without a motion or change . . . it must be something belonging to a motion . . . "motion" in its most general and primary sense is change of place, which we call "locomotion" . . . time is continuous because a motion is continuous (Aristotle, *Physica*, ca. 350 BCE, Book IV).

But the modern concept of a union of space and time in one abstract space, now called space-time, only developed in the 20th century (some 18th century anticipations are discussed below). In addition to the development of the various geometries discussed in the previous section, a number of other developments contributed to our current concept of space-time and its diagrammatic representation. I shall single out a few other crucial developments. Underlying the possibility of any further developments was:

0. The ability to create symbolic representations.
1. Representation of spatial intervals, and later other, non-spatial (concrete) magnitudes, by lengths (one dimensional diagrams).
2. Representation of (abstract) time intervals by lengths;
3. Combination of one-dimensional representations of spatial and temporal intervals in a single, two-dimensional diagram;

4. The representation of motion in two- and three-dimensional diagrams and coordinatization of two- and three-dimensional Euclidean space using mutually perpendicular axes
5. Recognition that two such coordinatizations are related by a coordinate transformation representing a rotation (orthogonal transformation).
6. Generalization of the concept of space beyond its use to describe three-dimensional physical space to higher-dimensional spaces, in particular the concept of time as a fourth dimension.
7. Formulation of the concept of affine spaces of arbitrary dimension and their use to formulate the principle of inertia.
8. The four-dimensional generalization of Poincaré's fourth geometry.
9. Formulation of the concept of Riemannian spaces of variable curvature, both in the metric sense (geodesics, Gaussian curvature) and in the affine sense (parallel transport, affine curvature);

I shall briefly indicate, to the best of my knowledge, when each of these concepts was introduced.

Perhaps the most extraordinary step on the road to the concept of space-time was step 2, the representation of a time interval by a length, which has been called the spatial representation of time by Henri Bergson (see *Durée et simultanéité*, 1922) and the spatialization of time by Émile Meyerson (see *La déduction relativiste*, 1925). It was the culmination of several preceding developments.

FIGURE 6. Wounded bison attacking a man – c. 15,000 - 10,000 BC. Bison length: 43 in. (110 cm). Lascaux, France

FIGURE 7. Cuneiform tablet

0) Underlying the possibility of any further developments is the ability to create symbolic representations. Until this uniquely human faculty to create pictorial and other lasting symbolic representations developed, no further progress was possible in the production of shared abstract concepts. The earliest preserved pictorial representations are the cave drawings and paintings, which are at most about 40,000 years old. Any cave art much older than that would have deteriorated beyond recognition, so it is hard to say just when the human ability to create such representations arose.

The earliest surviving examples include both abstract symbolic and naturalistic representational elements. Implicit in such drawings, especially in the naturalistic ones, is the concept of representation of the spatial dimensions and relations of external objects by lines in the drawing. This already manifests a considerable power of abstraction.

1A) In some of the abstract drawings, one finds geometrical patterns involving straight lines. In more naturalistic drawings, one finds representations of straight objects, such as spears or arrows, by straight lines (Fig. 6).

1B) I shall now jump forty millennia to something at an even higher level of abstraction: the first representation of a non-length by a length: Jens Høyrup has cited the earliest preserved instances of such representations from ca. 1800 BCE in clay tablets of the Mesopotamian scribal school (Fig. 7). Here,

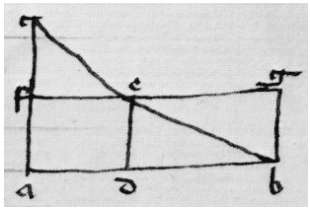
> a length is taken to represent something different from itself, viz. an area...[S]ince other (slightly later) Old Babylonian texts use lengths and widths to represent pure numbers, prices or complex arithmetical expressions, the step is real, no mere accident. . . . This is one of the great steps in the history of mathematics, one of the

very greatest, and whoever feels a chill when faced with intellectual progress should feel it here (see *Lengths, Widths, Surfaces/A Portrait of Old Babylonian Algebra and Its Kin*, 2002).

2) Only a millennium-and-a-half later is this method of representation applied to the concept of time: "The application of the concept of continuum to time and process does not appear to have taken place before the middle of the fourth century BC" (Hans-Joachim Waschkies, *Von Eudoxos zu Aristoteles/Das Fortwirken der Eudoxischen Proportionentheorie in der Aristotelischen Lehre vom Kontinuum, 1977)*.

It was Aristotle who, in his analysis of Zeno's paradoxes of motion, first represented a continuous time interval by a length. He utilized geometrical representations of both spatial and temporal intervals by one-dimensional line segments, placed parallel to one another, in order to compare the two (*Physica*, Book VI, Chapter 2). His example was followed by Archimedes, who employed similar diagrams (*On Spirals*, Propositions.I & II, ca. 225 BCE), an interesting case of a mathematician imitating a philosopher.

3A) Neither Aristotle, Archimedes, nor any of their successors for two millennia, combined the representations of temporal and spatial intervals into a single, two-dimensional diagram. The first to use time as part of a two-dimensional diagram appears to have been Nicholas Oresme in the mid-fourteenth century. To use somewhat anachronistic language, he plotted time against velocity, not distance (*Tractatus de configurationibus qualitatum et motuum*, ca. 1370; Marshall Clagett, ed. & transl., *Nicholas Oresme and the Medieval Geometry of Qualities and Motions*, 1968. For the claim that Giovanni di Casali preceded Oresme by a few years, see Marshall Clagett, *The Science of Mechanics in the Middle Ages*, 1961).



Nicholas Oresme (1321-1382)
"*Tractatus de configurationibus*"
The ordinate denotes the time, the abcissa denotes the velocity.

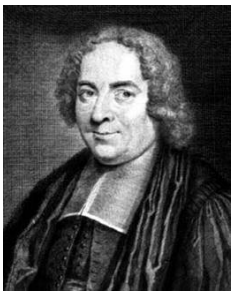
FIGURE 8.

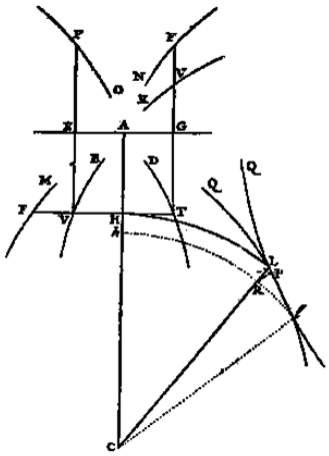
4A) René Descartes used two-dimensional spatial diagrams, with the points of a curve referred to two orthogonal spatial axes, (*La Geométrie*, 1637), and when he had to represent motion in three dimensions, he did so by projecting it onto several two-dimensional diagrams.

6A) The conjunction of time and space in a two-dimensional diagram took another half-century. However, Descartes did give a definition of *dimension* that justifies the spatial representation of *any* quantifiable property of a system:

By dimension, we understand nothing but the mode and reason, according to which some subject is considered to be measurable; so that not only length, breadth and depth are dimensions of a body, but in addition its gravity [i.e., weight] is a dimension, in accord with which subjects are weighed, its velocity is the dimension of motion, and an infinity of others of this type (*Regulae ad directionem ingenii* [*Rules for the direction of the understanding*], written between 1619-1628, first published in Dutch translation in 1684).

That for Descartes, time constitutes one such dimension is clear from the succeeding discussion.





Pierre Varignon (1654-1722) actually plotted position, time and velocity on the same diagram.

FIGURE 9.

3B) The first two-dimensional graphic representation of a one-dimensional motion, plotting distance and time as orthogonal coordinates, was done by Pierre Varignon near the turn of the 18th century (*Règle générale pour toutes sortes de mouvements de vitesse quelconques variées à discrétion*, 1698). Varignon actually plotted position, time, and velocity on the same diagram. Indeed, he first defined the concept of instantaneous velocity, and so may be said to have introduced extended configuration space as well. He saw the conceptual problem raised by this work:

Space and time being heterogeneous magnitudes, it is not properly they that are compared with each other in the relation called speed, but only the homogeneous magnitudes that express them; which here are, and will always be in what follows, either two lines, or two numbers, or two of any other homogeneous magnitudes that one wishes (*Des mouvements variés à volonté, comparés entre eux et avec les uniformes*, 1707).

About fifty years later, this problem was discussed in more detail by Jean le Rond d'Alembert:


Descartes


d'Alembert


Euler

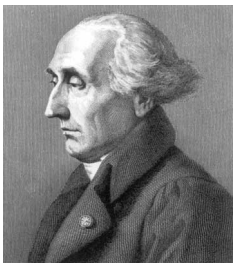

Lagrange

FIGURE 10.

> One cannot compare with each other two things of a different nature, such as space and time; but one can compare the relation of portions of time with that of the portions of the space traversed. By its nature, time flows uniformly and mechanics assumes this uniformity. In addition, without knowing time in itself and without having a precise measure of it, we cannot represent the relation of its parts more clearly that by that of portions of an indefinite straight line. Now, the analogy that exists between the parts of such a line and that of the space traversed by a body that moves in any sort of way, can always be expressed by an equation: one may thus imagine a curve, the *abcissae* of which represent the portions of time that have elapsed since the start of the motion, the corresponding ordinates representing the spaces traversed during these portions of time: the equation of this curve will express, not the relation of the times to the spaces, but, if one may so put it, the relation of the relation that the parts of time have to their unit, to that [relation] that the parts of space have to their unit (D'Alembert, *Traité de Dynamique*, 1743).

4B) Leonhard Euler appears to have been the first to use three-dimensional diagrams to represent motions, and to resolve forces and motions into their components along three mutually perpendicular axes (*Recherches sur le mouvement des corps célestes en général*, 1749; *Découverte d'un nouveau principe de mécanique*,1750)

5) Euler also realized that these three orthogonal axes could be chosen in many ways, each related to the other by a rigid rotation; and worked out the transformation between the two sets of Cartesian coordinates of a point (see, e.g., *Recherches sur la connoissance mécanique des corps*, 1758)

6B) D'Alembert was the first to discuss the concept of time as a fourth dimension:

> Above I said that it is not possible to conceive of more than three *dimensions*. A clever man of my acquaintance [d'Alembert himself?] believes that nevertheless one may regard duration as a fourth *dimension*, & that the product of time multiplied by solidity would be in some way a product of four *dimensions*. ("Dimension" in the *Encylopédie*, vol. 4, 1754)

Half a century later, Lagrange was less hesitant, affirming that:

> One may regard mechanics as a four-dimensional geometry, and mechanical analysis [i.e., analytical mechanics] as an extension of geometrical analysis (*Théorie des fonctions analytiques*, 1797)

6C) Euler used six coordinates to treat the six degrees of freedom of a rigid body –three for translation, three for rotation (*Theoria motus corporum solidorum seu rigidorum* [*Theory of the motions of solid or rigid bodies*], 1765). Lagrange introduced the idea of treating all the degrees of freedom of a mechanical system, however many, as abstract dimensions (*Mécanique analytique,* 1788). But he prided himself on having no diagrams in his book.

Diagrams involving such quantities as pressure and volume were introduced in the nineteenth century. As so often in thermodynamics, the idea was adapted from engineering practice: James Watt and John Southern used indicator diagrams to calculate the work done by a steam engine. Originally regarded as a trade secret, such diagrams were not published until 1822. Émile Clapeyron used such a diagram to represent the Carnot cycle (*Mémoire sur la puissance motrice de la chaleur,* 1834) and the idea passed into general usage in thermodynamics.

Mathematicians and physicists thus became accustomed to Descartes' idea (see 6A above) that any magnitude may be treated as a dimension, and any number of such magnitudes represented by an abstract space of higher dimension.

7) Hermann Grassmann abstracted the concept of parallelism from its metrical associations in Euclidean geometry and developed the concept of an affine geometry, and applied it to any number of dimensions (*Die lineale Ausdehnungslehre,* 1844, 2nd ed. 1878). He applied this concept to a number of problems in mechanics; but only much later was it realized that a four-dimensional affine space is the proper geometric setting for the law of inertia (Hermann Weyl, *Raum-Zeit-Materie,* 1918).

8) In 1905, Poincaré introduced the concept of a four-dimensional representation of the Lorentz transformations (*Sur la dynamique de l'électron*), quite independently of his earlier work in 1887 (see the opening section). In 1907, Hermann Minkowski realized that such a four-dimensional unification of space and time is particularly suited to the visualization of the Lorentz transformations (*Das Relativitätsprinzip*). He adapted a four-dimensional coordinate system similar to Poincaré's, and carried its geometrical interpretation further. He introduced the term *space-time* and (rather pretentiously) named special-relativistic space-time "*die Welt*" [*the universe* or *world*], leading to such terms as world point, world line and world tube. Like Poincaré, he represented the temporal coordinate by an imaginary number, so that Lorentz transformations could be interpreted geometrically as rotations in a four-dimensional (but complex) Euclidean space. It is more common now to use a real time coordinate and pseudo-rotations in a real but non-Euclidean space-time.

9) Riemann generalized Gauss' theory of surfaces of variable curvature to spaces of any number of dimensions that are locally flat (Euclidean) but globally non-flat, i.e. having a curvature that varies from point to point. What does curvature mean here? Riemann defined a fourth rank tensor, now called the Riemann curvature tensor, as a generalization of the Gaussian curvature (*Habilitationsschrift*, 1854, posthumously published in 1868).

The concept of parallel transport in such a Riemannian space was not introduced until 1916 by Tullio Levi-Civita (*Nozione di parallelismo in una varietà qualunque e con-*

sequente specificazione geometrica della curvatura Riemanniana, 1917), in response to the formulation of general relativity. Hermann Weyl (*Reine Infinitesimalgeometrie*, 1918) soon generalized his work by defining the concept of a non-flat affine geometry. Locally it is an affine-flat space, but globally non-flat in a new sense: In such a space, parallel transport of a vector around a closed curve results in a different vector, the affine curvature tensor being a measure of the difference This concept led to a deeper understanding of the relation between affine connection and inertio-gravitational field in both Newtonian and general relativity theory (see below).

Space-Time Structures

Before turning to S-R space-time, I shall discuss the concept of Galilei-Newtonian (G-N) space-time, the space-time associated with the Galileian law of inertia and Newtonian dynamics. Logically it comes before SR space-time, although in actuality its four-dimensional version was only developed afterwards. One reason for this is that the mathematical structure of G-N space-time crucially involves the concept of an affine space (see above).

There are two distinct types of space-time structure inherent in both G-N and S-R space-time:

1) The *chrono-geometrical structure*, which determines the behavior of (ideal) measuring rods (geometry) and clocks (chronometry); and
2) The *inertial structure*, which governs the behavior of free particles (i.e., particles subject to no external forces);
3) there are also *compatibility conditions* between the two.

The S-R inertial structure (and later the inertio-gravitational structure in GR) is represented mathematically by an affine connection. This connection is usually derived from the chronogeometry, which is represented mathematically by a pseudo-metric, and treated as secondary if mentioned at all. But I shall emphasize the connection for two reasons:

1) Much recent progress in GR has come from emphasis on its primary role in the most fruitful formulations of the theory in preparation for canonical quantization;
2) It illuminates the connection (pun intended) as well as the contrast between GR and Yang-Mills gauge theories.

Geometry

In both G-N and S-R space-times, the geometry of the relative space of each inertial frame is Euclidean; it can be measured with (ideal) measuring rods at rest in that frame, for example. The distance along any spatial path depends on the path taken, and the

shortest distance (geodesic) is along the *straightest path*. Mathematically an inertial frame is represented by a *fibration* of space-time; that is, a family of parallel time-like straight lines that fills the space-time, each line of which transvects the space-like hyperplanes of simultaneity (see below).

G-N Chronometry and Chrono-Geometry

In G-N kinematics, the chronometry is independent of the geometry: The time is *absolute* and *universal*, and space-time divides naturally into events that are *simultaneous* (i.e., occur at the same absolute time). Mathematically, the absolute time is represented by a *foliation* of space-time; that is, a family of parallel space-like hyper-planes of equal global time. A four-dimensional G-N *chrono-geometric* structure can be defined, which splits naturally into a unique chronometry (unique foliation) – time is absolute – and a three-parameter family of relative geometries (three-parameter family of fibrations) – space is relative to the choice of inertial frame.

S-R Chrono-Geometry

In contrast, S-R space and time are united in one absolute chrono-geometrical structure, represented mathematically by a flat four-dimensional pseudo-metric ("pseudo" because it has a non-definite signature, usually called Lorentzian), often called the Minkowski metric. This results in the existence at each point of space-time of a double null cone (consisting of a forward and backward cone) of events that have a null (i.e., zero) separation from the event in question. A *null separation* between two events is interpreted as the possibility of connecting them by a light signal (or any zero-rest mass particle); which way the signal can pass depends on which event is in the forward light cone of the other.

S-R Chronometry

Both *geometry* and *chronometry* are now *relative*. This results in a big difference between the two chronometries:

In G-N chronometry, as noted above, the absolute time along any path between two non-simultaneous events is independent of the path. All (ideal) clocks measure this absolute, universal time.

In S-R chronometry, the time along a path between any two events with a time-like separation (i.e., each one is within the forward or backward light cone of the other), usually called the *proper time, depends on the time-like path* taken between them. In this respect, S-R time is more like space, but there is still a big difference: The *longest time*

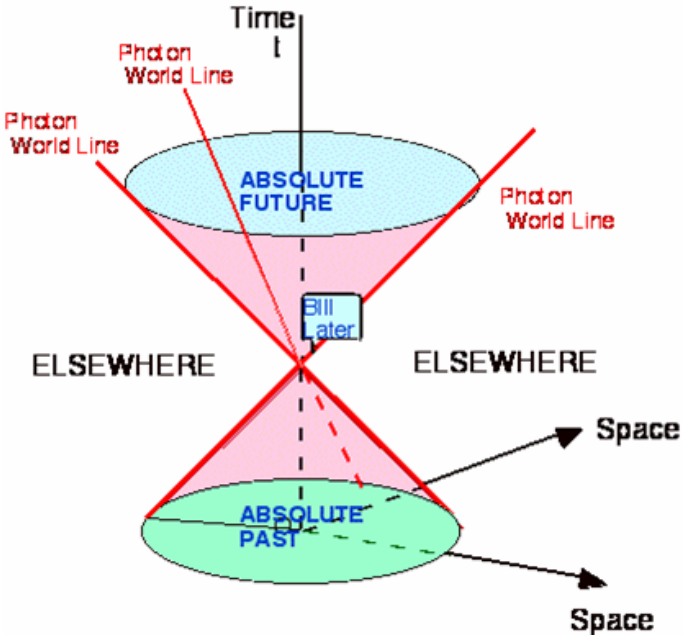


FIGURE 11. Minkowski space in 4 dimensions

interval between two events is along the *straightest path* between them (this observation is the essence of the "twin paradox").

Inertial Structure

The use of the terms *straightest* and *parallel* actually encroaches upon the domain of the second type of space-time structure: the *inertial structure*, which determines the motion of *freely-falling* (i.e., net force-free) structureless bodies ("particles").

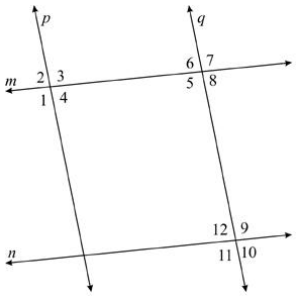
In both G-N and S-R space-times, such particles follow the time-like straightest inertial paths of space-time – straight lines for the flat space-times of both classical and special-relativistic physics, with the affine parameter coinciding with the absolute time in the first case, and with the proper time in the second. This is the mathematical expression of the law of inertia, common to both G-N and S-R space-times because it depends only on their common affine structure.

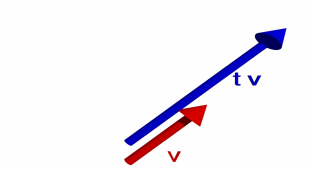
Affine Spaces

We shall only be concerned with torsion-free affine spaces.

Mathematically, to define the inertial structure, all we need is the concept of an affine space, for which parallelism and the ratio of parallel intervals are meaningful concepts. The affine structure defines the concept of parallelism for two vectors at neighboring

points of space-time. In an affine space, a curve is *straight* in the sense that its tangent vector always remains parallel to itself as it is parallel-transported along the curve.





Affine space: Parallel intervals Ratio of parallel vectors

FIGURE 12.

Compatibility Conditions

The two space-time structures – chrono-geometry and inertial structure – are *compatible* with each other. Mathematically this compatibility is expressed by the vanishing of the covariant derivative of the chrono-geometry. This has a number of kinematical consequences. For example:

1) The *extremal paths* (*geodesics*) as defined by the pseudo-metric (shortest paths for space-like curves, longest paths for time-like curves), coincide with the *straightest paths*, as defined by the inertial structure.
2) *Freely falling rods and clocks*, as defined by parallel transport with the inertial structure, continue to measure *proper space and time intervals* respectively, as defined by the chrono-geometry.

The Relativity Principle

Unless the relativity principle is taken into account, the combination of spatial and temporal dimensions in a single diagram shares a feature with the combination of such heterogeneous dimensions as pressure, volume and temperature in diagrams used to picture thermodynamic relations: While the three spatial coordinates in a given frame of reference (e.g. an inertial frame) can be mixed among themselves by rigid rotations (as noted above, a technique introduced in rigid body dynamics by Euler), the spatial and temporal coordinates of that frame can no more be mixed than can p, V and T.

Galilean Relativity

It follows from the laws of Newtonian mechanics, that no *mechanical* experiment can distinguish between any two inertial frames of reference. As long as it was believed that all physical phenomena could ultimately be reduced to mechanical interactions (the mechanical world view), this restriction seemed harmless.

Once this *Galilean relativity principle* is taken into account, the situation changes: Now, *the Galilei transformations*, which relate the Cartesian coordinates of an event with respect to two inertial frames of reference in relative motion with relative velocity $\mathbf{V}$, allow us to mix spatial and temporal coordinates:

$$\mathbf{r}' = \mathbf{r} - \mathbf{V}t,$$

where $\mathbf{r}'$, $\mathbf{r}$ are the Cartesian coordinate vectors relative to the origins of the respective inertial frames, and t is the absolute time (assuming the origins to coincide at time $t = 0$). Of course in classical (G-N) kinematics, time is absolute, and to emphasize this we must add

$$t' = t$$

to our transformation equations. Again one sees that space is relative to choice of an inertial frame of reference, but time remains universal and absolute.

Relativity Principle and Optics

With the rise of the wave theory of light and then Maxwell's explanation of light as a type of electromagnetic wave, the mechanical world view seemed to demand introduction of a mechanical medium-the ether- in which such waves would propagate. All attempts to detect the motion of the earth through the ether by optical or other electromagnetic phenomena failed.

The relativity principle seemed to apply to all these phenomena independently of the hypothetical ether. In 1874, Eleuthère Elie Nicolas Mascart formulated what we might call the Optical Principle of relativity, based on experimental tests of order (v/c), where v is the presumed velocity of the earth through the ether:

> No optical experiment can detect the motion of the earth through the ether. The earth's translational motion does not have a measurable influence on optical phenomena produced by a terrestrial source ... [T]hese phenomena do not provide us with a way to determine the absolute motion of a body and ... relative motions are the only ones that we are able to determine. (*Modifications qu'éprouve la lumière par suite du mouvement de la source lumineuse et mouvement de l'observateur (deuxième partie)*, 1874).

Yet optical and later electromagnetic theory predicted the existence of such effects.

To explain this apparent paradox, Hendrik Antoon Lorentz and then Poincaré introduced the concept of local time and the length-contraction hypothesis, which they interpreted as dynamical "compensations" for the expected effects of motion through the ether. They introduced what Poincaré named the Lorentz transformations from the unprimed coordinates in the ether frame to the primed coordinates in the moving frame:

$$\begin{aligned} \mathbf{r}' &= \gamma(\mathbf{r} - \mathbf{V}t) + (1-\gamma)[\mathbf{r} - (\mathbf{V}.\mathbf{r})\mathbf{V}/V^2] \\ t' &= \gamma[t - (\mathbf{V}.\mathbf{r})/c^2], \end{aligned}$$

where $\gamma = [1-(V/c)^2]^{1/2}$.

Neither Lorentz nor Poincaré realized the fundamental kinematical significance of these transformations; they interpreted them within the framework of N-G kinematics and the ether theory: Due to their motion through the ether, clocks *really* slow down and rigid rods *really* contract. There is a distinction between the "apparent," primed space and time coordinates of an event, as measured in a moving frame of reference by the slowed-down clocks and contracted measuring rods, and the "true," unprimed spatial and temporal coordinates as defined by clocks and rods at rest in the ether.

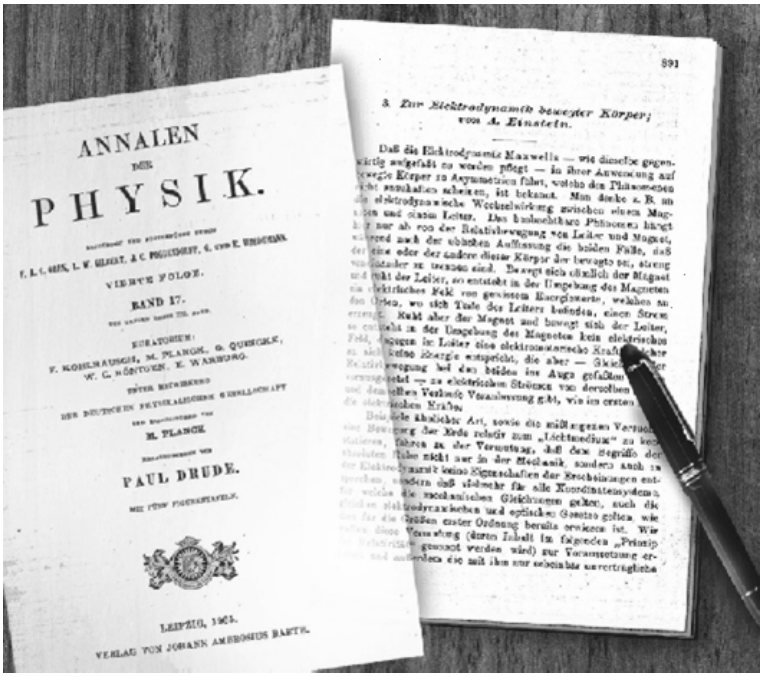

ANNALEN
DER
PHYSIK.
VIERTE FOLGE.
BAND 17.
PAUL DRUDE.
LEIPZIG, 1905.
VERLAG VON JOHANN AMBROSIUS BARTH.
891
3. Zur Elektrodynamik bewegter Körper;
von A. Einstein.

FIGURE 13. *"On the Electrodynamics of Moving Bodies"*

Special Relativity

It was Albert Einstein (*Zur Elektrodynamik bewegter Körper*, 1905) who first realized the need to replace such ideas, based on classical kinematics, with a new kinematics based on four key ideas:

- 1. Omit all reference to the hypothetical ether frame;
- 2. Take the failure of all attempts to detect absolute motion at face value, and postulate the relativity principle (all inertial frame of reference are equivalent) for all physical phenomena;
- 3. Add the well-tested postulate that the speed of light is independent of that of its source;
- 4. Combining 1, 2 and 3, one can derive the Lorentz transformations between any two inertial frames of reference. Interpret the measured spatial and temporal coordinates occurring in them as the "true" spatial and temporal coordinates of each inertial frame of reference; these transformations then form a group that does not single out any inertial frame.

The derivation of the Lorentz transformations requires that simultaneity of distant events be *defined* with respect to each inertial frame of reference in such a way as to make the speed of light the same in every inertial frame and independent of position and direction in that frame. It is important to realize that, if the concept of distant simultaneity is to be introduced at all, some definition always is needed. No physical result can depend on this definition; and it is even possible to dispense with such a definition. Bondi's K-calculus, for example, can treat the special theory without introducing such a definition (see, e.g., Hermann Bondi, *Relativity and Common Sense*, 1964).

In 1905 Einstein formulated his insights largely in ignorance of the most recent results of Lorentz and Poincaré, and treated space and time separately, rather than combined into space-time. But, since Einstein's new kinematics mixed both spatial *and* temporal coordinates in the transformation from one frame to another, the adoption of the space-time viewpoint, once suggested, was irresistible.

Global vs. Local Time: Newtonian Identity

The Newtonian absolute time is both *global* and *local*. It is:

> *Global*, because it can be used for defining distant simultaneity in each inertial frame, and even universal, because this definition will give the same result for two events, no matter in which inertial frame the definition is used.
>
> *Local*, because it provides the readings of any good clock along its world line, and absolute because the time difference read between any two events will be the same for all world lines.

Global vs. Local Time: Special-Relativistic Splitting

In SR, the global and local concepts of time, which coincide in Newtonian kinematics, split apart:

> *Global Time*: No matters of fact can depend on the definition of global time (see above); but various definitions may be useful in different contexts. For example, the retarded time along the light cones emanating from some world line (as utilized in the K-calculus) will give the same global time for all world lines passing through any one event, but different global times for the same event as defined by different but parallel world lines. Since it depends only on a single world line, this definition may be extended to general relativity.
>
> The Poincaré-Einstein convention is the most useful for an inertial frame of reference. It leads to different global times for the same event as defined in different inertial frames; but all world lines in the same inertial frame define the same global time for any event. Since it depends on distant parallelism that is independent of path, it cannot be extended to general relativity.
>
> *Local Time*: The concept of local time is now the proper time along any time-like world line. It is *absolute* (i.e., frame-independent) like the Newtonian time, but unlike it in being *path-dependent*. As noted above the local time is more like the spatial distance: a good clock is more like a good pedometer than previously thought.

The Moral of This Tale

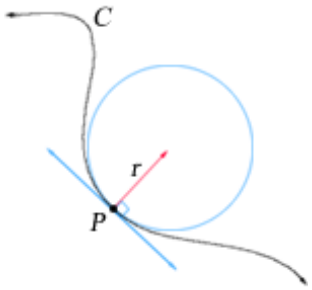
Loose talk about "space" and "time" being "relative" is just that, and often leads to serious philosophical misinterpretations. To sum up the moral again. In SR:

> The *global space* (fibration of S-T) and *global time* (foliation of S-T) are both relative, but *nothing physically significant* depends on them.
>
> *Local space* (integral along a space-like path) and *local time* (integral along a time-like path) are absolute, but both are *path dependent*

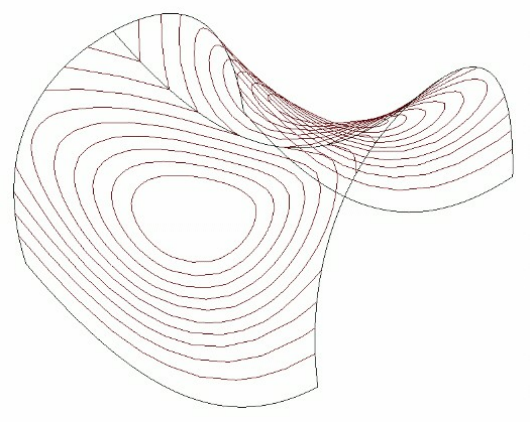
Curvature

In N-G and S-R space-times, both the chrono-geometrical and the inertial structures are *flat*, in the sense that:

> For *chronogeometry*, there is no *Gaussian curvature*, defined by the pseudo-metric and associated with any of the two-sections through a point.
>
> For *the inertial field*, there is no *affine curvature* associated with the parallel transport of a vector through space-time. Any vector parallel-transported around any closed curve coincides with itself when it returns to its starting point.



Osculating circle



Gaussian curvature

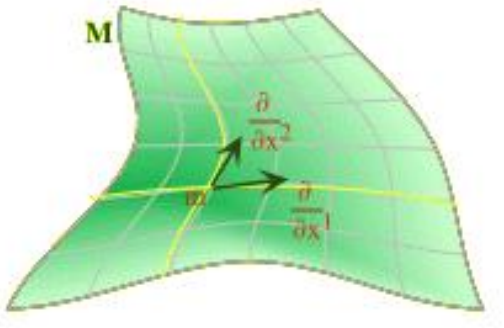
FIGURE 14.

Gaussian Curvature

The curvature of a plane curve at any of its points is the inverse of the radius of the osculating circle at that point.

At any point of a surface, each plane through that point intersects the surface in a plane curve. Take the maximum and minimum curvatures of these plane curves as the plane is varied. Their product is the Gaussian curvature of the surface at that point.
This definition depends on the embedding of the surface in Euclidean three-space. But Gauss showed that the Gaussian curvature is an intrinsic property of the surface. He proved that is can be expressed in terms of the metric components in the expression for the distance between two neighboring points of the surface in Gaussian (curvilinear) coordinates, the line element ds (*Theorema Egregium*):

$$ds^2 = g_{11}(dx^1)^2 + 2g_{12}(dx^1)(dx^2) + g_{22}(dx^2)^2.$$

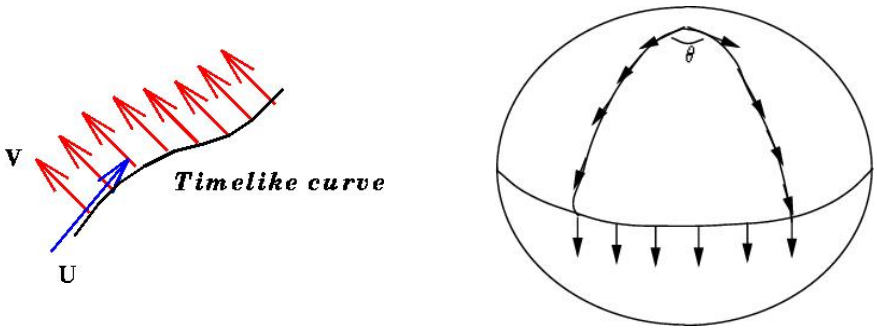


$$g_{11} = \frac{\partial}{\partial x^1} \cdot \frac{\partial}{\partial x^1} \qquad g_{12} = \frac{\partial}{\partial x^1} \cdot \frac{\partial}{\partial x^2}$$

$$g_{21} = \frac{\partial}{\partial x^2} \cdot \frac{\partial}{\partial x^1} \qquad g_{22} = \frac{\partial}{\partial x^2} \cdot \frac{\partial}{\partial x^2}$$

FIGURE 15. Line element

Locally, this line element expresses Euclidean geometry. It is just Pythagoras' Theorem expressed in curvilinear coordinates. The Riemann curvature tensor generalizes Gaussian curvature to a space of any number of dimensions.



Parallel transport Affine curvature

FIGURE 16.

Affine Curvature

Given a vector at any point in an affinely-connected space, the connection enables us to define the vector parallel to it at a neighboring point. By iterating this procedure, we may *parallel transport* a vector along any curve: What happens to a vector when it is transported parallel to itself around a closed curve? If there are any closed curves, for which the parallel-transported vector does not coincide with the original vector, one says the space is affinely curved. By taking a set of infinitesimal closed curves, one can define the components of the *affine curvature tensor*.

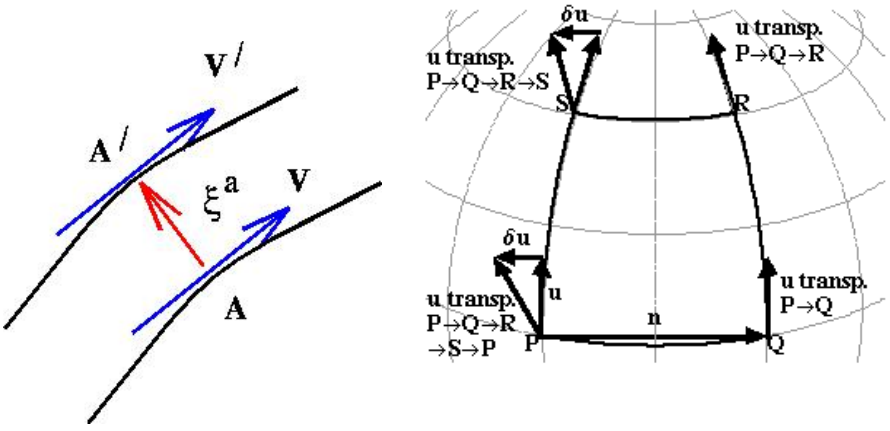


FIGURE 17. Geodesic deviation

Equation of geodesic deviation

The affine curvature tensor has another application that is especially important for its physical interpretation. Consider an infinitesimal displacement vector connecting two neighboring affinely straight lines (i.e., curves such that the tangent vector field along the curve is parallel transported into itself). The affine curvature tensor is a measure of how this displacement vector changes as a function of the affine parameter as we proceed along the two straight lines. If the displacement vector change is accelerated, then the affine curvature has a non vanishing component related to the direction of the lines and of the displacement vector.

Physically, an affine straight line corresponds to the path of a freely falling body, and the equation of geodesic deviation measures whether there is any relative acceleration between two nearby freely-falling bodies. Mathematically, the amount of such relative acceleration in various directions is a measure of the components of the affine curvature tensor; physically, it is a measure of the gravitational tidal forces.

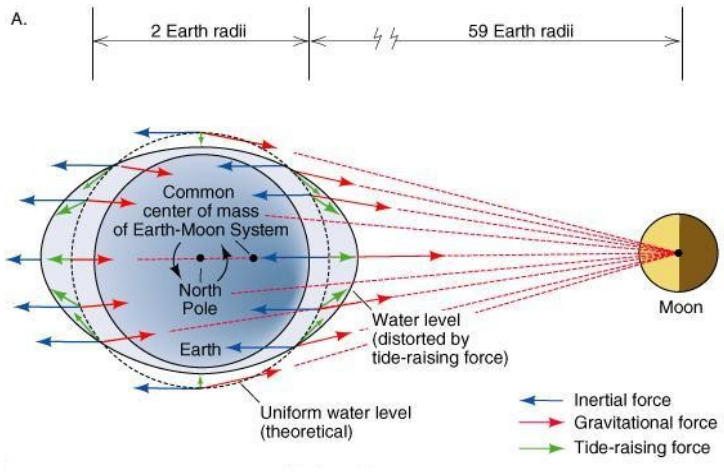


FIGURE 18. Tidal forces

Newtonian Gravitation

Special-relativistic space-time proved sufficient for the analysis of all physical phenomena, for which gravitation may be neglected. But it must be modified to include gravitation, because of:

The Equivalence Principle

Because inertial and gravitational mass are equal, there is no (unique) way to separate the effects of inertia and gravitation on a "freely-falling" body. Once this is understood, even at the Newtonian level, gravitation can no longer be treated as an external force acting on bodies, but must be regarded as a modification of the hitherto fixed inertial structure of space-time. This structure now becomes dynamical, an *inertio-gravitational* field.

While the inertial structures of both GN and SR space-times are associated with a flat affine connection, the inertio-gravitational field is associated with an affine structure that is no longer flat. The *affine curvature* associated with the Newtonian inertio-gravitational field describes *tidal gravitational forces*. This curvature obeys field equations that reduce to the field equations for the Newtonian gravitational "force" in any non-rotating frame of reference.

Although its symmetry group is enlarged to include all linearly-accelerated frames (this is the equivalence principle), the classical Newtonian chrono-geometrical structures are unmodified. The chrono-geometrical and inertio-gravitational structures remain compatible: ideal measuring rods and clocks still remain such in the presence of any inertio-gravitational field. But the compatibility conditions do not uniquely determine the inertio-gravitational field: Just enough freedom is left to introduce gravitational fields that reduce to the gradient of the Newtonian potential in non-rotating frames of reference.

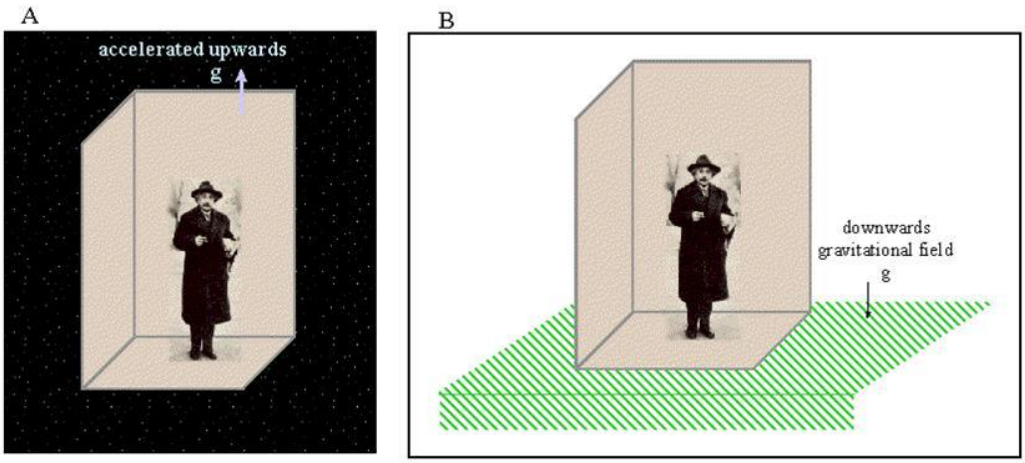


FIGURE 19. Take a ride on the Einstein elevator

General-Relativistic Space-Time

But special-relativistic chrono-geometry is no longer compatible with the dynamical inertio-gravitational field: The flat Minkowski metric is not compatible with the non-flat affine structure. To restore compatibility, the chrono-geometry must be modified: The pseudo-metric must become a non-flat, dynamical field that plays a dual role. In addition to determining the chrono-geometry, it also serves as the potentials that uniquely determine the inertio-gravitational field.

While the inertio-gravitational field traditionally was derived from the chrono-geometry, we favor the modern approach, which treats both as logically independent before the imposition of the field equations. One set of field equations then relates the inertio-gravitational field to all other matter and fields (the sources) by equating the contracted affine curvature tensor of the inertio-gravitational field to the stress-energy tensor of the sources. The other set of field equations are the compatibility conditions imposing the unique relation between chrono-geometry and inertio-gravitational field.

To succeed in formulating the special theory, Einstein had to attach physical significance to the coordinate system. To succeed in formulating general relativity, Einstein had to learn that coordinates have no inherent physical significance (see discusson below).

PHILOSOPHICAL SECTION

Two Concepts of Space: Absolute vs. Relational

Historically, since (at least) ancient Greek times, there has been a conflict between two views of the nature of space:

• The *absolute* concept: Space is a container, in which matter moves about. This view was espoused by Demokritos (and the Greek and Roman atomists):

> By convention are sweet and bitter, hot and cold, by convention is color; in truth are atoms and void" (Fr. 589, ca. 430 BCE. G. S. Kirk and J.E. Raven, *The Presocratic Philosophers*, 1957)

This is sometimes paraphrased as "Nothing exists but atoms and void. All else is mere opinion." Aristotle criticized the concept of a void:

> The believers in its reality present it to us as if it were some kind of receptacle or vessel, which may be regarded as full when it contains the bulk of which it is capable, and empty when it does not (*Physica*, Book VI).

• The *relational* concept: Space has no independent existence. It is just a certain set of positional relations between material entities. There cannot be a vacuum –the world is a *plenum*. Aristotle's doctrine is really a doctrine of *place* rather than *space*.

> The physicist must have a knowledge of Place ... because 'motion' in its most general and primary sense is change of place, which we call 'locomotion'... the

Demokritos Aristotle Voltaire Newton Leibniz

FIGURE 20.

> motions of simple bodies (fire, earth, and so forth) show not only that place is something but that place has some kind of power [*dunamin*] (*Physica*, Book IV).

Aristotelianism triumphed and atomism vanished from the Western philosophical tradition for almost two millennia. With its revival in early modern times and subsequent adoption by Newton, the conflict between the absolute and relational concepts was renewed in the 17th and 18th centuries in the battle between Newtonianism and Cartesianism (the philosophy of Rene Descartes). As Voltaire wittily observed:

> A Frenchman who arrives in London, will find philosophy, like everything else, very much changed there. He had left the world a plenum, and he now finds it a vacuum (*Lettres philosophiques.*, ca. 1778, “Letter XIV, On Descartes and Sir Isaac Newton”).

This time it was the absolute, Newtonian conception of space that triumphed in spite of the cogent arguments of Leibniz and Huygens against it:

> In fine, the better to resolve, if possible, every difficulty, he [Newton] proves, and even by experiments, that it is impossible there should be a plenum; and brings back the vacuum, which Aristotle and Descartes had banished from the world (*ibid.*, "Letter XV, On Attraction").

As Euler emphasized, absolute space seemed to be necessary if one wanted to use Newtonian dynamics (*Réflexions sur l’espace et le temps*, 1748).

Absolute versus Relational Concepts of Space and Time

Einstein summarized the situation in these words:

> Two concepts of space may be contrasted as follows:
> (a) space as positional quality of the world of material objects;
> (b) space as container of all material objects.
> In case (a), space without a material object is inconceivable. In case (b), a material object can only be conceived as existing in space; space then appears as a reality which in a certain sense is superior to the material world. ("Foreword" to Max Jammer, *Concepts of Space*, 1954).

Things versus Processes

The old emphasis on space and time favors the concept of *things*, which occupy regions of space at moments of time, but changing over (absolute) time. The new emphasis on space-time favors the concept of *processes*, which occupy regions of space-time, or even – with the development of the field concept – all of space-time. (*Events* are then defined as limiting case of processes, occupying vanishingly small regions of space-time.) We must now discuss the extension of our previous discussion of things in space to processes in space-time.

Two Concepts of Space-Time: Absolute vs. Relational

The *absolute*: Space-time is an independent container, in which processes take place. In addition to ponderable matter, such processes now include fields (e.g., the electromagnetic field) that may fill all space-time.

The *relational*: Space-time has no independent existence. It is just as certain set of relations between the elements of processes. There cannot be an empty space-time.

Pre-general Relativistic Situation

In the case of both Galilei-Newtonian and special-relativistic space-times, it was possible to hold either of these viewpoints although there were serious problems for the relational viewpoint:

1) The possible existence of regions of space-time that are devoid of all matter and fields.
2) The fact that the space-time structures remain the same, regardless of all the varying physical processes that can take place within them.
3) The fact that the space-time structures influence all physical processes (e.g., through the law of inertia), but are not influenced by them. This is a general problem with any fixed, background structures introduced into physics.

Fixed, Background Space-Time Structures

In the case of space-time, we refer to such structures as fixed, background space-time structures. Thus, we may sum up our previous discussion by saying that both GN and SR theories are based on fixed, background space-time structures.

Theories with background space-time structures have a kinematics that is logically prior to and independent of all dynamical physical theories. The slogan is: *Kinematics first,*

then dynamics! The background space-time is a stage, upon which various dynamical dramas can be enacted.

Such background space-time structures are essential features of all current quantum theories:

GN space-time is the stage for the quantum mechanics of non-relativistic quantum systems.

SR space-time (Minkowski space) is the stage for relativistic quantum field theories as all thoughtful workers on the subject recognize:

> The basic concept[s] of the theory are quantum fields defined on space-time, not particles. Space-time is assumed to be a four-dimensional real vector space with given metrical properties and Einstein causality, such that the Poincaré group (constituted by translations and Lorentz transformations) is implied as a symmetry group. This space-time structure fixed in advance – called Minkowski space – forms the register for recording physical events. The predictions of a relativistic quantum field theory on the outcome of scattering processes are of probabilistic nature, in this respect similar to those of (non-relativistic) quantum mechanics. However, a novel feature occurs: in these processes particles can be created and annihilated. The quantum fields, in terms of which the theory is constructed, are operators that depend on space-time and act on the space of physical state vectors.(Hans Günther Dosch, Volkhard F. Müller and Norman Sieroka, "*Quantum Field Theory, Its Concepts Viewed from a Semiotic Perspective*", 2004).

In short, both the formalism of quantum field theory and the measurement processes that test its predictions presuppose the SR (Minkowski) space-time structure.

The General-Relativistic Revolution - The Triumph of Relationalism

As discussed above, in the general theory of relativity, both the inertio- gravitational and the chrono-geometrical structures are dynamical fields. We speak of such theories, which are free of any background space-time structures, as background free: In a background-free theory, with no non-dynamical structures, kinematics and dynamics cannot be separated. The slogan is: *No Kinematics Without Dynamics!!!!*

The problems for the relational viewpoint discussed above now disappear:

1) There are no "empty" regions of space-time: Wherever there is space and time (chrono-geometric structure), there is always (at least) an inertio-gravitational field (affine structure).
2) The space-time structures are not independent of the processes taking place within them. Chrono-geometry and inertio-gravitation are dynamical fields, obeying field equations that couple them to each other and to all other physical processes.
3) Thus, there is now reciprocal interaction between space-time and other processes. Physical processes do not take place *in space-time*. Space-time is just *an aspect of the totality of physical processes*.

Rosenfeld Bronstein

FIGURE 21.

General relativity more-or-less forces one to adopt the relational viewpoint.

> On the basis of the general theory of relativity ... space as opposed to 'what fills space' ... has no separate existence. If we imagine the gravitational field ... to be removed, there does not remain a space of the type [of the Minkowski space of SR], but absolutely nothing, not even a 'topological space' [i.e., a manifold]... There is no such thing as an empty space, i.e., a space without field. Space-time does not claim existence on its own, but only as a structural quality of the field (Einstein,"Relativity and the Problem of Space," in *Relativity: The Special and the General Theory*, 1952 edition).

THE PROBLEM OF QUANTUM GRAVITY

The greatest challenge to theoretical physics today is: How to invent a theoretical structure that encompasses both Quantum Field Theory (background-dependent) and General Relativity (background-independent)? "That is the Question."

Quantizing General Relativity

In 1916, Einstein stated that general relativity would require a quantum version for the same reason that electromagnetism did: A gravitationally bound system would ultimately radiate away all its energy unless it was quantized.

The earliest attempts to apply the methods of QFT, developed by Heisenberg and Pauli in the late 1920s, to GR came in the early 1930s, first by Leon Rosenfeld (see John Stachel, "The Early History of Quantum Gravity," in Bala Iyer and Biplap Bhawal, eds, *Black Holes, Gravitational Radiation and the Universe*, 1998, pp. 525-534). The basic philosophy behind this work was that only technical difficulties (the non-linearity of the field equations) stand in the way of application of standard methods of QFT to GR, and the way to begin was by quantizing the linearized approximation to the field equations.

In the 1930s, only one physicist realized that such attempts raised profound conceptual problems due to the unique features of gravitation as compared to electromagnetism: Matvei Petrovich Bronstein (see Gennady Gorelik, "First Steps of Quantum Gravity and the Planck Values," *Studies in the history of general relativity* [*Einstein Studies*, vol. 3], 1992, pp. 364-379). He was the only serious contender with Lev Davidovich Landau for leadership of Soviet theoretical physics. Both were imprisoned during the Stalinist purges of the mid-1930s: Landau survived, Bronstein perished.

> In formal quantum electrodynamics, which does not take into consideration the structure of the elementary charge, there is no consideration limiting the increase of density .... With sufficiently high charge density in the test body, the measurement of the electrical field may be arbitrarily precise. In nature, there are probably limits to the density of the electrical charge... but formal quantum electrodynamics does not take these limits into account .... The quantum theory of gravitation represents a quite different case: it has to take into account the fact that the gravitational radius of the test body ... must be less than its linear dimensions ... The elimination of the logical inconsistencies connected with this requires a radical reconstruction of the theory, and in particular, the rejection of a Riemannian geometry dealing, as we see here, with values unobservable in principle, and perhaps also the rejection of our ordinary concepts of space and time, modifying them by some much deeper and nonevident concepts. *Wer's nicht glaubt, bezahlt einen Taler* ["*Let him who does not believe it pay a dollar*" – finale of a Grimm fable]. (Bronstein, *Quantentheorie schwacher Gravitationsfelder*, 1936)

There is no place here to say more about the history of quantum gravity (for the later history, see Carlo Rovelli, "Appendix B History" in *Quantum Gravity*, 2004).

But it is relevant to note that the conflict between those who see only technical problems in the application of existing techniques of QFT to general relativity and those who see profound conceptual issues in the reconciliation of quantum theory and general relativity, which started in the 1930s, continues to this day.

Background-Dependence versus Background-Independence

The first viewpoint is represented today mainly by people from the quantum field theory community. Their approach is basically to keep a background space-time (of however many dimensions), and somehow incorporate general relativity into the quantum formalism developed using this background structure. Currently, the strongest candidate put forward by advocates of this approach is string theory, or some variation or extension of it such as the elusive M-theory.

The second viewpoint is represented today mainly by people from the general relativity community. Their approach is to try to develop a background-independent formulation of quantum theory and apply it to general relativity. Currently, the strongest candidate put forward by advocates of this approach is loop quantum gravity (LQG), or some extension of it such as spin-foam theory. Without going into any details, I want to

emphasize the importance of the connection in the LQG program. In contrast to previous approaches to canonical quantization, such as geometrodynamics, which took the metric as primary, the most important achievements of LQG are based on taking a particular form of the connection as primary.

Being from this community myself, it is natural that I favor the background-independent approach, but have a number of critical reservations about how it is currently carried out (see John Stachel, "Structure, Individuality and Quantum Gravity," Steven French, Dean Rickles and Juha Saatsi, eds., *The Structural Foundations of Quantum Gravity*, to appear). But, as I shall emphasize, there are people in the string community who also favor this approach.

A New Formal Principle?

None of the current approaches has been completely successful in solving the basic problem of quantum gravity: the reconciliation of QFT with GR. In 1905, Einstein faced a similar situation in his attempts to reconcile Newtonian mechanics with Maxwell's electrodynamics. As he said much later:

> Gradually I despaired of the possibility of discovering the true laws by means of constructive efforts based on known facts. The longer and more desperately I tried, the more I came to the conviction that only the discovery of a universal formal principle could lead us to assured results (*Autobiographical Notes*, 1949)

Consideration of a striking common feature of QFT and GR has led me to propose a new formal principle that might serve as a guide in the further quest for a theory of quantum gravity, whatever direction(s) it may take.

From General Covariance to Permutation Invariance

What is the significance of the general covariance of the field equations of general relativity? If general covariance is given an active interpretation (as it should be – coordinate transformations can never have a direct physical significance), it requires invariance of the field equations under the diffeomorphism group acting on the underlying differentiable manifold M of space-time points. But what are diffeomorphisms? A little thought shows that they are just fancy permutations (automorphisms) of the homogeneous elements of M – permutations that are required to be continuous and differentiable because they act on the elements of a differentiable manifold – but permutations nevertheless.

I said above "the homogeneous elements of M", and this is an important part of the meaning of general covariance: the elements of M are not distinguished from each other unless and until some solution to the field equations is specified. Einstein's 1913 hole argument against general covariance was based on the tacit assumption that, just as in SR, the points of the space-time manifold could be individuated independently of the field, and it was only his realization in late 1915 that this assumption was untenable in

a background-independent theory that enabled him to justify his adoption of generally covariant field equations (see John Stachel, "Einstein's Search for General Covariance, 1912-1915" in *Einstein and the History of General Relativity*, 1989, pp. 63-100). Indeed it is this question of individuation that distinguishes algebra from geometry.

Geometry vs Algebra

A *geometry* consists of a set of elements, together with some relations between them, such that all of the elements are homogeneous under the group of automorphisms (permutations) that preserves all the relations. In such a case, since the relations are primary, one may speak of "The things (elements) between the relations"

Example: Euclidean plane geometry, a manifold homeomorphic to R^2, together with the group of translations and rotations acting on the points of the manifold.

An *algebra* consists of a set of elements, together with some relations between them, such that each element is individuated independently of the relations between it and the other elements. In such a case, since the elements are primary, one may speak of "The relations between things (elements)."

Example: The plane rotation group, each element of which is characterized by an angle.

A *representation* of an abstract space (geometry) is called *algebraic* if it characterizes the space by means of some *coordinatization* of its elements (points). A coordinatization is a one-one correspondence between the elements of an algebra and those of a geometry. Since any one coordinatization individuates the otherwise homogeneous elements of a geometry, the only way to keep them homogeneous is to demand invariance of any geometrically significant result under *all* admissible coordinatizations. These concepts of geometry, and coordinatization are due to Hermann Weyl (see *The Classical Groups*, 1939). This concept of algebra is due to I. R. Shafarevich (see *Basic Notions of Algebra*, 1997).

The coordinatization of a differentiable manifold is generally a local operation, since usually, no one coordinatization can cover the entire manifold. One must carefully distinguish between coordinate transformations (re-coordinatizations of the differentiable manifold), which are local, passive mathematical operations, needed to ensure that the points of the manifold remain homogeneous, and having no physical significance; and the diffeomorphisms of the manifold, which are global, active point transformations of great potential physical significance as we shall see.

Background-Independent Theories and Diffeomorphisms

In a background-independent theory, there are no non-dynamical relations to be preserved on the set of space-time points; so all possible permutations of the points of space-time are permissible. If one adds the demand that these permutations be continu-

ous (because space-time is a manifold) and differentiable (because it is a differentiable manifold), one gets the diffeomorphism group. As noted above, in GR the points of space-time have no inherent properties that individuate them. GR is a background independent theory, or in my terminology a "things-between relations" theory.

The Principle of Maximal Permutability

FIGURE 22.
S. MacLane

One can thus express the concept of general covariance in GR in the following form: The theory (GR) shall be invariant under all possible permutations of the basic entities of the theory (elements of space-time in GR) in the sense that any model of the theory (solution to the field equations of GR) shall be physically equivalent to any other model that results from it by such a permutation. In this form the principle for diffeomorphisms in general relativity can be both *generalized* and *abstracted*.

Generalization: "Generalization from cases refers to the way in which several specific prior results may be subsumed under a single more general theorem" (Saunders MacLane, *Mathematics, Form and Function*, 1986).

One can generalize the principle of diffeomorphism invariance from the pseudo-metric tensors and affine connections of general relativity to arbitrary geometric object fields, also called natural objects (see John Stachel and Mihaela Iftime, *Fibered Manifolds, Natural Bundles, Structured Sets, G-Sets and All That: The Hole Story From Space Time to Elementary Particles*, 2005; *The Hole Argument for Covariant Theories*, 2005)

Abstraction: "Abstraction by deletion ... One carefully omits parts of the data describing the mathematical concepts ... to obtain the more abstract concept" (Saunders MacLane, *ibid.*).

By dropping the assumptions of differentiability, one can extend the principle from theories based on differentiable manifolds to those based on *topological manifolds*; and by dropping the assumption of continuity the principle can be extended to theories based on *sets of discrete elements*.

Even if the concepts of space, time and space-time have to be greatly modified; or are themselves explained in terms of some more fundamental entities in some future theoretical advance, it is hard to believe that one would retreat from the relational to the absolute point of view concerning the fundamental entities, whatever their nature. This suggests adoption of the principle of *maximal permutability of the fundamental constituents* as a "universal formal principle" in Einstein's sense as a heuristic guide in the search for a theory of quantum gravity – and even beyond.

The heuristic force of this principle is reinforced by the observation that, like the points of space-time, the particles of non-relativistic QM and the field quanta of special-relativistic QFT also lack inherent individuality and hence obey the principle. They are only individuated (to the extent that they are) by some process (Feynman's word) or phenomenon (Bohr's word), in which they are involved. In any quantum system in non-relativistic QM, both the bosons and the fermions of any species can be arbitrarily permuted among themselves without changing the probability amplitude for any process; so, like the points of space-time, they are also "things between relations". And in QFT, the field quanta in any Fock space state are also completely indistinguishable.

A Background-Independent String Theory?

As currently constituted, string theory is based on a fixed, background space time structure on a manifold of some number of dimensions higher than four. Regardless of the details of particular models, it is clear that the principle of maximal permutability is violated: Only diffeomorphisms that are symmetries of the fixed, background structure (usually a flat pseudo-metric) are permissible permutations of the elements of the manifold.

FIGURE 23. B. Greene

Many string theorists are aware of this problem. Brian Greene recently presented an appealing vision of how a background-free string theory might look, but he emphasized how far string theorists still are from realizing this vision (*The Fabric of the Cosmos – Space, Time, and the Texture of Reality*, 2004):

> Since we speak of the "fabric" of spacetime, maybe spacetime is stitched out of strings much as a shirt is stitched out of thread. That is, much as joining numerous threads together in an appropriate pattern produces a shirt's fabric, maybe joining numerous strings together in an appropriate pattern produces what we commonly call spacetime's fabric. Matter, like you and me, would then amount to additional agglomerations of vibrating strings – like sonorous music played over a muted din, or an elaborate pattern embroidered on a plain piece of material – moving within the context stitched together by the strings of spacetime. ... [A]s yet no one has turned these words into a precise mathematical statement. As far as I can tell, the obstacles to doing so are far from trifling. .... [T]o make sense of this proposal, we would need a framework for describing strings that does not assume from the get-go that they are vibrating in a preexisting spacetime. We would need a fully spaceless and timeless formulation of string theory, in which spacetime emerges from the collective behavior of strings... Many researchers consider the development of a background-independent formulation to be the single greatest unsolved problem facing string theory.

Einstein's Greatest Contribution?

For reasons discussed above, I have been led to conjecture that, whatever form a future fundamental physical theory (such as some version of quantum gravity, or something even farther from our current conceptual framework) may take, there will be no absolute elements in it. Rather, its basic entities – whatever their nature – will be embedded in some discrete or continuous relational structure: The result will be a completely background-independent physics.

If I am proved right, then a millennium from now (assuming humanity still exists and has not relapsed into barbarism) then Einstein's greatest contribution to physics will be regarded as the development of the first, prototype background-independent physical theory!

As I indicated earlier, it is always dangerous to try to predict the future; still, I can draw wry comfort from the fact that – right or wrong – I shall not be around when my prophecy is finally tested. But I hope some of you may.

ACKNOWLEDGMENTS

I thank Lysiane Mornas for all the careful work put into the preparation of the written version of this paper.