Black Holes and General Relativity
(2020 Physics Nobel Prize)

Ghanashyam Date
Chennai Mathematical Institute

These notes grew out of the colloquium given in the Online Colloquium Series "Nobel Talks" of the Physics Dept. of BITS-Pilani on March 25, 2022 and the colloquium given at CMI on April 26, 2022.

THE LAUREATES AND THE CITATION

The Nobel Prize in Physics for 2020 was shared by **Roger Penrose**, **Reinhard Genzel** and **Andrea Ghez** in the proportions 1/2, 1/4 and 1/4 respectively.

credits:

"for the discovery that black hole formation is a robust prediction of the general theory of relativity"

"for the discovery of a supermassive compact object at the centre of our galaxy."

(https://www.nobelprize.org/prizes/physics/2020/summary/)

There are three distinct strands to the story of the discoveries: General relativity, Astrophysics and Astronomy. The table below gives an overview.

General Relativity	Astrophysics	Astronomy
Schwarzschild Singularity (1916)	Chandrasekhar Limit (1930-31)	Birth of Radio Astronomy (1933-39)
Cosmological Singularities (1922-1955) Raychaudhuri Equation (1955)	Tolman-Oppenheimer-Volkoff Limit, Datt-Oppenheimer-Snyder (1939)	⋮
Kruskal Extension (1960)	⋮	Discovery of C-273 Quasar (1959)
The Kerr Solution (1963)	...	Energy output and size of Quasar
Trapped Surfaces (1965)	...	Identification of Sagittarius A^* (1974)
Singularity Theorems (1965-1975-...)	...	Supermassive Compact Object(1980-2020)

Let us go through these developments.

General Relativity: Soon after Einstein published his field equations (Nov. 25, 1915), Schwarzschild published an exact solution to these complicated equations (Jan. 1916). It describes the space-time due to a point mass. Impressive as it was that an exact solution could be found so quickly and could play a role in the classic tests of general relativity, it had the "disturbing" feature of the metric component $g_{rr}(r)$ blowing up at $r = 2GM/c^2 =: R_S$ where M is the mass of the point mass. This is known as the "Schwarzschild singularity" and $R_S := 2GM/c^2$ is known as the Schwarzschild radius (or gravitational radius). It was also recognized that the solution describes the exterior geometry of a spherical, massive body of mass M and with radius $R > R_S$. Consequently, in application to solar system, the Schwarzschild singularity played no role. It is only with the Kruskal extension in 1960 that nature of the Schwarzschild singularity being a coordinate artifact was fully appreciated. (At the Schwarzschild radius, all curvatures are finite. The $r = 0$ is however a genuine curvature singularity.)

During the twenties to the mid-fifties, many more exact solutions to the Einstein field equations were found, notably modeling cosmological space-times. Invariably they exhibited curvature singularities. The big bang singularity of the Friedman-Lemaitre-Robertson-Walker

solution being the most familiar one. All these examples had a high degree of symmetry - the spherical symmetry in the case of the Schwarzschild and spatial homogeneity with/without isotropy in the case of cosmological models. Amal Kumar Raychaudhuri in 1955 considered if the singularity of the isotropic dust model would survive if isotropy is relaxed. He concluded in the positive and also gave the first form of the *Raychaudhuri equation*. His result is also considered as a first hint of a singularity theorem - that *singularities may not necessarily be due to high degree of symmetry*. A very readable account may be seen in the reference given at the end.

In 1963, Roy Kerr found another exact solution of the field equation without any matter source, which was not spherically symmetric - the *Kerr Solution*. This is now understood as representing a rotating black hole. This too had a singularity (and many other features!) at the centre. This solution still had the axisymmetry.

It was Roger Penrose, motivated partly by the discovery of a quasar (Quasi Stellar Radio Source), considered gravitational collapse without any symmetry assumptions. Examining first the case of spherical collapse, he identified "trapped" spherical surfaces and showed that these persist even if there are small deviations from spherical symmetry. He then shows that once trapped surfaces are formed, suitably identified singularity necessarily forms. This is the first singularity theorem with the essential ingredients.

The main outcome in this strand is the development of the more general and elaborate singularity theory. This theory invokes the *global structure of the space-times, Causality conditions, the Raychaudhuri equations, energy conditions and a suitable definition of a "singularity"*. We will discuss this shortly.

Astrophysics: Consider the second strand now. This is rooted in the astrophysical theory of stability of stars. The basic mechanism of stellar stability is that the gravitational collapse is halted by thermal pressure generated by non-gravitational means such as ordinary chemical heating, nuclear burning, quantum mechanical "degeneracy pressure" due to the Pauli exclusion principle etc.

Subrahamanyan Chandrasekhar, in 1930-31, noticed that in the stars known as white dwarfs, which are supported by electron degeneracy pressure, the electron motion is actually relativistic (electron speed comparable to the speed of light). This changes the relation between the mass-energy density and the pressure (polytropic index from 5/3 to 4/3). This results in an instability if the mass of the white dwarf exceeds the *Chandrasekhar limit* of about 1.4 times the solar mass (1 solar mass $\sim 2 \times 10^{33}$ gms.) What happens then? Well, all one can say is that the gravitational collapse ensues, possibly unstoppable. It may be noted

that white dwarfs have radii up to about 4000 km which is approximately 1000 times the corresponding Schwarzschild radius.

During 1938-39, it was recognized by Tolman, Oppenheimer and Volkoff that for neutron stars, stars in which the dominant matter consists of neutrons, the gravitational collapse could be halted by degeneracy pressure again as neutrons too obey the Pauli principle. They found that there is an analogue of the Chandrasekhar limit, that above about 2 - 3 solar masses (TOV limit), a neutron star too will become unstable and gravitational collapse will commence. Neutron stars have their radii around 10-15 km which is roughly 3 times the corresponding Schwarzschild radii.

To find out what happens in a non-stoppable collapse, Oppenheimer and Snyder consider the collapse of a spherically symmetric, dust in the Einstein equation and produced the time dependent solution - the Oppenheimer-Snyder solution - which showed that once the ball crosses the Schwarzschild radius, no light rays can emerge from its surface. The ball shrinks to zero size with infinite density, in finite time. Similar case was also considered and published by B. Datt from Kolkata, a year earlier in German and it was not noticed. It was only recently discovered.

Further studies of unstoppable gravitational collapse continued mainly to test the cosmic censorship hypothesis, and has also lead to the discovery of critical phenomena in gravitational collapse, by Choptuik around 1993.

The main lesson here is that astrophysical equilibrium states of gravitating bodies indicate an upper mass limit of up to about 3 solar masses, beyond which there is no known mechanism to halt the gravitational collapse.

(Radio) Astronomy: We now turn to the third strand.

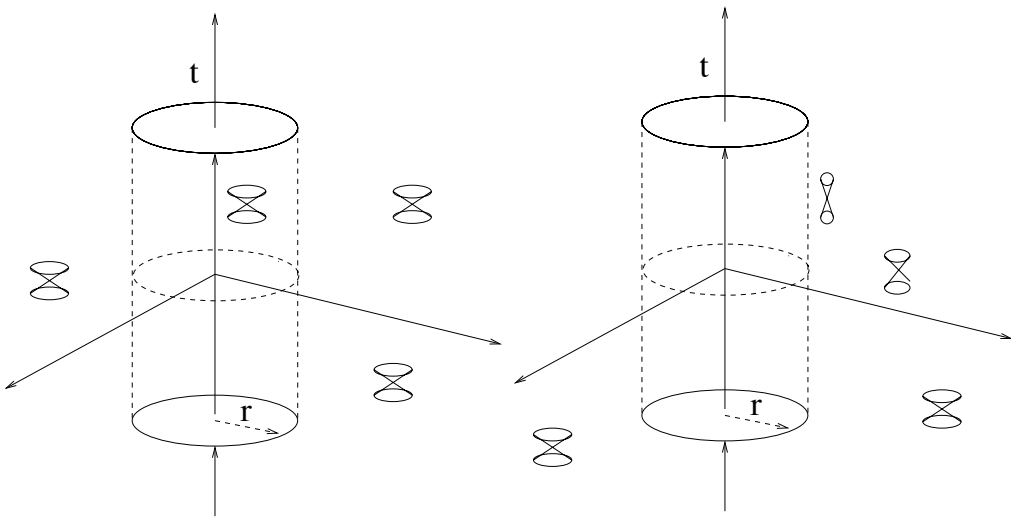
Karl Jansky, a radio engineer, detected the first ever extra-terrestrial radio source and managed to conclude that it was in the constellation of Sagittarius, near the centre of the Milky Way (1933). Subsequently, Grote Reber, another radio engineer, made the first map of the radio sky identifying Cygnus A and Cassiopeia A also as radio sources (1941-43). Thus was born the field of radio astronomy. After the world war II, radio astronomy picked up and had the first discovery of a Quasar, 3C-273, in 1959. The time scale of fluctuation of its intensity gives its size to be less than 1 light month and it is about 100 times brighter than the brightest galaxy in the galaxy cluster. The mechanism of such energetic radio emission cannot come from atomic/molecular but comes from synchrotron radiation (radiation from charges accelerated along curved paths). What could produce so much power from such a

small region? "Black hole" were a possible candidate. This was one of reasons that had prompted Penrose to examine if *general relativity accommodates such objects.*

In 1974, the radio source Sagittarius A, at center of our galaxy was resolved into a point like source Sagittarius A^* and attention focused on its study - its mass and its size. Its initial estimates of mass $\sim 10^6 M_\odot$ and size $\sim 10^8$ km, were arrived at from the spectral characteristics of the radio emission from its vicinity. The x-ray emission implied presence of accretion processes.

With Very Long Baseline Interferometry (VLBI), its position and proper motion could be tracked accurately while near infrared observations could track the motion of individual stars orbiting the central object. The size of the central region was still about a couple of thousand times the Schwarzschild radii and possible alternatives to a single compact object existed. These included: clusters of low mass stars, neutron stars and stellar mass black holes. The stability against collapse or dispersal gave a time scale of about a 10^5 years which ruled out this possibility. The possibility of a ball of heavy fermions such as sterile neutrinos, gravitinos held together by degeneracy pressure too was ruled out by 2002 and finally by Oct 2018 it was concluded that Sgr A* IS a black hole with a mass of $4.154 \times 10^6 M_\odot$.

Singularity Theory: Let us begin by visualizing a space-time. Here are the examples of the Minkowski space-time of special relativity and the Schwarzschild space-time.



Minkowski Spacetime

$$ds^2 = -dt^2 + dr^2 + r^2 d\Omega^2$$

Schwarzschild Spacetime

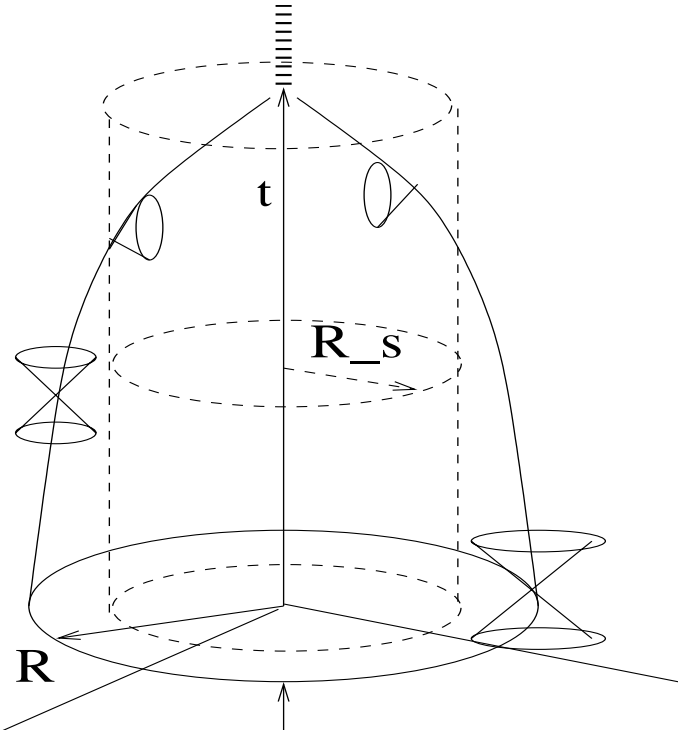
$$ds^2 = -\left(1 - \frac{R_s}{r}\right) dt^2 + \frac{1}{(1 - \frac{R_s}{r})} dr^2 + r^2 d\Omega^2$$

Notice the "uniform distribution" of identical light cones in the Minkowski space-time. By

contrast, the Schwarzschild space-time has light cones that get narrower as we come closer to the Schwarzschild radius. This is simply because $\frac{dt}{dr} = \pm(1 - R_S/r)^{-1}$ which increases as $r \to R_S$.

Next is the Oppenheimer-Snyder space-time.

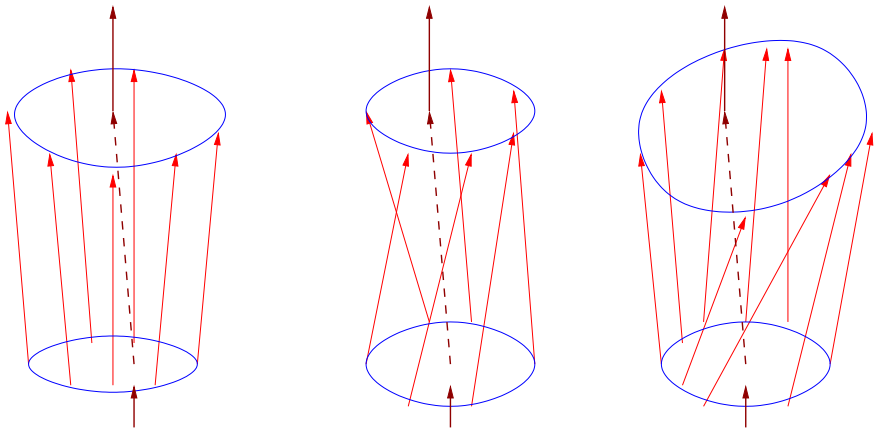
Here the light cones tilt inwards as the R_S is approached and align with the $r = R_S$ cylinder. No material particle or light can escape from the inner region. Thus we see that (a) a singularity at $r = 0$ does arise; (b) at the "Schwarzschild singularity", something peculiar does happen - the light cones tilt fully inward. Note also that this situation arises from a perfectly non-singular initial distribution. Of course it has been assumed that the density remains spatially constant all through the evolution.



Oppenheimer-Snyder Collapse Spacetime

Keeping such qualitative picture of spacetimes in mind, let us explore further properties.

Consider a bundle of time-like geodesics i.e. geodesics whose world line of a free falling particle always inside the future part of a local light cone. Single out one relative to which we track how the nearby geodesics move - come closer, go farther, keep circulating etc. The constant time cross-section of such bundles shows three types of distortions: *expansion* (scaling of its size), *twist* (rotation of the cross-section relative to the reference geodesic) and *shear* (differential movements of layers of the 3-dimensional cross-section).



Expansion Twist Shear

In any space-time, the evolution of bundle can be quantified in terms of these which satisfy well defined differential equations. Relevant for us is the equation satisfied by the expansion, denoted as θ:

$$\frac{d\theta}{d\tau} = -\frac{1}{3}\theta^2 + \text{twist}^2 - \text{shear}^2 - R_{\mu\nu}v^\mu v^\nu$$

This is the *Raychaudhuri equation.*

It follows that if $\theta = \theta_0 < 0$ for some τ_0 i.e. initially collapsing, and $R_{\mu\nu}v^\mu v^\nu \geq 0$ equivalent to strong-energy condition on stress tensor if Einstein equation is used is satisfied, then $\theta \to -\infty$ in a $\tau < \tau_0 + 3/|\theta_0|$.

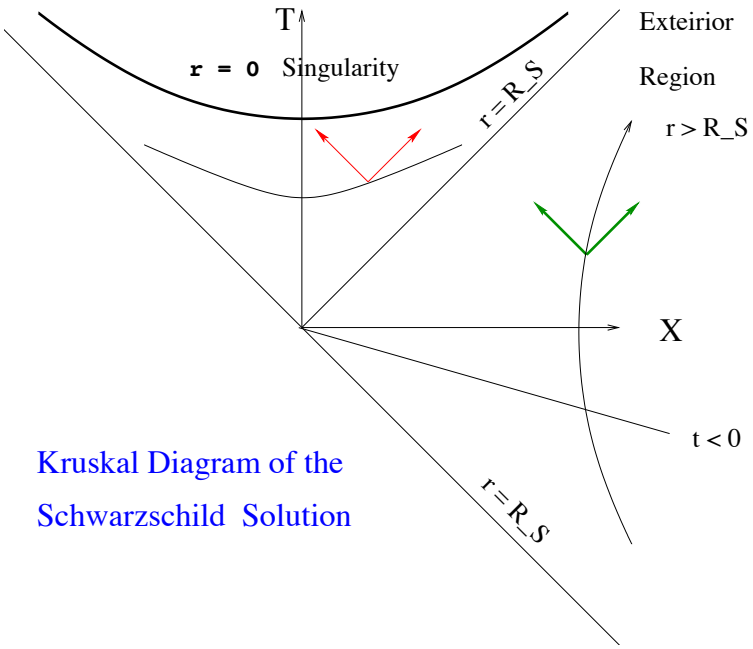
The bundle of geodesics is said to focus and the geodesic is said to have a conjugate point.

This is true independent of the Einstein equations.

Here, $\theta_0 < 0$ means the particles are initially collapsing. If the spacetime is a solution of the Einstein equation, then the condition on $R_{\mu\nu}$ translates into a condition on the stress tensor $T_{\mu\nu}$, which is known as the *strong energy condition.*

We have written the Raychaudhuri equation for a bundle of time-like geodesics. There is a corresponding equation for a bundle of null geodesics (light rays). The factor of $1/3 \to 1/2$ and the corresponding expansion, twist and shear need a little sophisticated definition. We are now ready to define Penrose's trapped surfaces.

Consider a 2 dimensional sphere in a spacetime and flash light rays in the outward and inward direction perpendicular to the surface. These constitute bundles of null geodesics for which the twist vanishes. These have expansions, $\theta_\pm$. The closed 2-surfaces for which both the expansions $\theta_\pm \leq 0$, are called *trapped surfaces.* Here is an example.



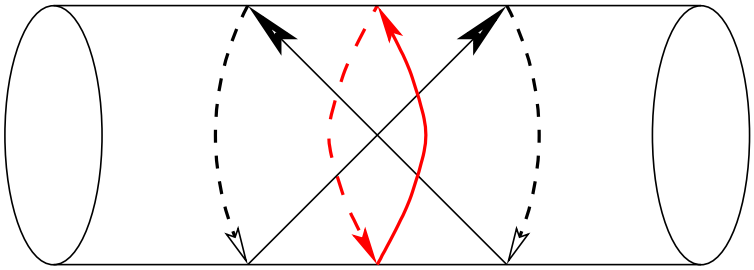
This representation of the Schwarzschild spacetime in terms of the Kruskal coordinates T, X, θ, ϕ. Each point in the diagram is a 2-sphere with coordinates θ, ϕ. Constant value of the previously used Schwarzschild coordinate r define curves in this Kruskal diagram which are displayed for various values of r. Remember that each point on these curves (except for $r = 0$) is 2-sphere. The 45 degree red and green lines denote the outward (towards increasing value of r) and inward (towards decreasing value of r) null geodesic bundles. The picture makes it obvious that for $r > R_S$, we have the usual expanding and contracting light spheres while for $r < R_S$, both light spheres are contracting (negative expansion) and thus constitute trapped spheres. Thus, indeed we have a solution of the Einstein equations having trapped surfaces.

The next important ingredient on the singularity theory is the *global structure of spacetime*. What is meant by this?

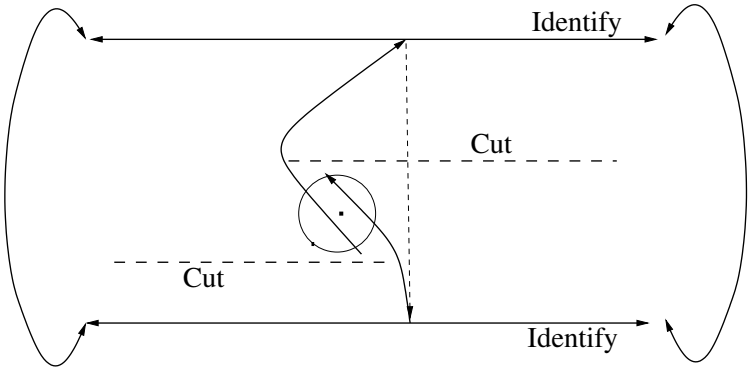
Einstein equations are partial differential equation. Thus their solutions are *local* solutions i.e. valid in a small region about a spacetime point. These can be extended to larger regions or local solution in different regions can be matched in the overlap of the regions etc. Such a process generates a *global solution*. Incidentally, this is also true for most other basic equations of physics. It is a property of the space-time metric that in a sufficiently small region we can always describe a solution with the light cone structure of the Minkowski spacetime. And in Minkowski spacetime, we know that we have a clear notion of future/past or "time orientation" and this is a pre-requisite for labeling of cause and effect. When we join many local solutions, we may end up loosing this property. Our global spacetime may

have close time-like curves! We meet our first challenge.

Here is a cartoon example of a two dimensional spacetime. The space axis is horizontal while the temporal direction is along the vertical but is wrapped around. At any point, we have the usual light cones. As they are extended, the lines wrap around and join as shown. And now you see how a closed time-like curve is possible! Every small portion (i.e. locally) our toy space-time looks like a flat Minkowski spacetime with no confusion about past/future), but the extended spacetime creates the ambiguity. The *causality condition* says that, such spacetimes are disallowed by fiat.



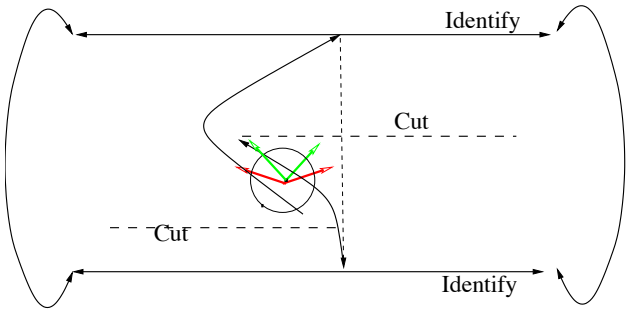
We are not done yet. We may have a spacetime which does not admit any closed time-like or null (causal in short) trajectories. But we may have some point(s) where a causal curve comes *arbitrarily close* to itself, even though it never connects exactly. Operationally, such a case will be difficult to decide if the causal curve is closed or not! Here is another cartoon example.



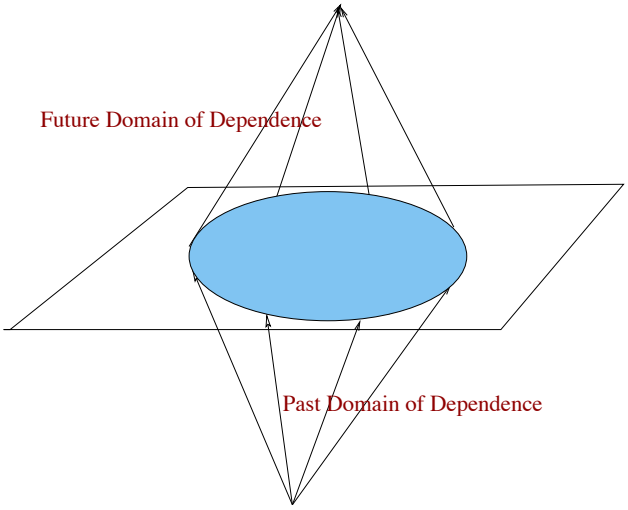
We have introduced two cuts in the previous spacetimes which prevent close causal curves, but permit a future directed causal curve to come arbitrarily close to itself. The *strong causality condition* says that such spacetimes are also to be disallowed.

One more step still. The light cone are a depiction of a metric. We may compare two different metrics and find that the light cone with reference to one is *wider* than that with reference to the other. This can be easily written down. In our example of the Schwarzschild spacetime,

we could also draw the Minkowski light cones which *are* wider than the Schwarzschild light cones as we come closer to R_S. Notice that causal directions in the narrower light cone are also causal directions in the wider light cones, but not conversely. Now, some of the space-like directions of the narrower light cone can be causal in the wider one and hence, a strongly causal spacetime can violate strong causality with a small change to a metric with wider light cones. This is depicted in the cartoon spacetime below. The red cones are the wider ones. Precluding such spacetimes - strongly causal spacetimes which cease to be so under widening of light cones - is called the *stable causality condition*. Stably causal spacetimes ensure that no causal pathologies arise.



Ok, we have restricted to spacetimes in which a causal chain can be constructed unambiguously between an 'effect' point and its 'cause' point. We can construct all possible causal chains for a given effect (i.e. consider all possible past directed causal curves from a given point a spacetime). Are all the causes so inferred accountable or accessible to our experimentation? We can change some of the causes a little bit and that will be registered at the effect point. But there may be some causes beyond our control which are also registered at the effect point. If so, then we cannot hope to have a definite prediction of effects from accounted causes. Consider the figure below.



The shaded portion of the planar surface is where we can introduce 'causes'. It is taken to be *acausal* i.e. no two points on it can be connected by a causal curve. Its future domain of dependence consists of those effect points which can have their causes registered on the shaded region. This is shown in the upper cone. Points outside this cone will have some of the past directed causal curves intersecting outside of the shaded portion. Likewise the past domain of dependence of the shaded region is the past cone. Points outside it will have some future directed causal curves missing the shaded region. If we can only confine ourselves to the full domain of dependence of some acausal surface, then we can have determinism. Thus we take as definition, *globally hyperbolic spacetimes* are those which are the full domain of dependence of some hypersurface (3-dimensional surface in a 4-dimensional spacetime). That hypersurface is called a *Cauchy surface*.

It so happens that globally hyperbolic spacetimes are free of all causal pathologies and support determinism. Thus now may now stipulate:

The physically admissible spacetimes are globally hyperbolic spacetimes which are solutions of Einstein equations with stress tensor satisfying the strong energy condition.

In everyone of these spacetimes, we have bundles of geodesics governed by the Raychaudhuri equation and also the *possibility* of trapped surfaces (Warning: this does not mean that all admissible solutions have a trapped surface, certainly they do not). There is a potential source of conflict.

In globally hyperbolic spacetimes, between any two causally connected points, there exist a curve with maximum proper time. This curve is time-like geodesic with *no* conjugate points on it. That is, if two spacetime points can be reached by causal means, we can always find a "freely falling" observer going from the earlier to the later event without any freely falling dust cloud closing on it.

We see the potential conflict. Raychaudhuri equation says conjugate points can exist and global hyperbolicity says no that cannot happen. Note that existence of conjugate point is conditional on an initially collapsing dust cloud. A trapped surface will precisely ensure that! What is the way out? The geodesic must terminate *before* the indicated conjugate point is reached. Though we have managed to get rid of all causal pathologies, have determinism, have Einstein equations with physically observed sources and yet we may have an endangered fate of an observer - an incomplete time-like geodesic. Thus we conclude, with Penrose, that *if a gravitational collapse has progressed far enough to have formed a trapped surface, then such a solution will have a singularity in the sense of an incomplete time-like geodesic.*

Note that there is no assertion of conditions for formation of trapped surfaces, but if they do form then singularity is inevitable. This conclusion does not use any special symmetry or otherwise stipulation (except those of admissible spacetimes), and hence is robust.

Note: There are many fine prints that have been glossed over, but the essential logic of a robust prediction of singularity if trapped surfaces form does not depend on these fine prints.

Let us briefly return to other, non-supermassive black holes. Observationally, these are suspected when astronomer notice a source which is faint in visible light but bright in the xray emission. Such a situation arises due to in-falling matter forming an accretion disk quite close to central region. Several such black holes have been identified.

Yet another method indicating black holes is based on the phenomenon of emission of gravitational waves. General relativity predicts that if the *quadrupole moment* (and/or higher) has a non-vanishing *second time derivative*, then such a mass distribution will loose energy by gravitational radiation. Two astrophysical bodies going around each other, thus loose energy and spiral towards each other and merge. The emitted gravitational wave form has a characteristic variation of amplitude with frequency - *a chirp waveform.* This contains the information about the masses and sizes of the coalescing bodies. Using the observed gravitational waveforms, the *Gravitational Wave Observatories* have detected over 80 binary *black hole mergers* with *masses ranging from about 5 to 150 solar masses.* This is a completely different mass range.

In summary:

Observationally, an object is a "black hole" if:

(a) the object is *compact* i.e. its radius is comparable to its Schwarzschild radius obtained from the estimate of its mass: $R_S := 2GM/c^2$;

(b) if the mass is larger than the maximum mass of the known stable, compact objects such as Neutron stars i.e. $\gtrsim 2 - 3M_\odot$.

In general relativity:, a "black hole" is identified as a *spatially compact region containing a trapped surface.* A precise definition of its size requires a suitable definition of a "horizon".

These two criteria are linked together because formation of a trapped surface guarantees a singularity, reflecting a run away collapse violating any conceivable upper mass limit. *And we can assert that,*

General relativity does admit black holes.

For further reading:

The book by Kip Thorn, *Black Holes and Time Warps*, W W Norton & Company, 1994, is very readable and informative.

For the more technical minded, the article by José M. M. Senovilla, David Garfinkle, *The 1965 Penrose singularity theorem* is available on the arxiv server at https://arxiv.org/pdf/1410.5226